RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL ARTICLE

Scoring ultrasound synovitis in rheumatoid arthritis: a EULAR-OMERACT ultrasound taskforce-Part 2: reliability and application to multiple joints of a standardised consensus-based scoring system

Lene Terslev,[1] Esperanza Naredo,[2] Philippe Aegerter,[3] Richard J Wakefield,[4] Marina Backhaus,[5] Peter Balint,[6] George A W Bruyn,[7] Annamaria Iagnocco,[8] Sandrine Jousse-Joulin,[9] Wolfgang A Schmidt,[10] Marcin Szkudlarek,[11] Philip G Conaghan,[4] Emilio Filippucci,[12] Maria Antonietta D'Agostino[13,14]

To cite: Terslev L, Naredo E, Aegerter P, *et al*. Scoring ultrasound synovitis in rheumatoid arthritis: a EULAR-OMERACT ultrasound taskforce-Part 2: reliability and application to multiple joints of a standardised consensus-based scoring system. *RMD Open* 2017;**3**:e000427. doi:10.1136/rmdopen-2016-000427

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/rmdopen-2016-000427).

Received 23 December 2016
Revised 9 May 2017
Accepted 16 May 2017

Linked

► http://dx.doi.org/10.1136/rmdopen-2016-000428

CrossMark

For numbered affiliations see end of article.

Correspondence to
Professor Maria Antonietta D'Agostino, Rheumatology Department, Versailles Saint Quentin en Yvelines University, Ambroise Paré Hospital, APHP, 9 avenue Charles de Gaulle, 92100 Boulogne-Billancourt, France; maria-antonietta.dagostino@apr.aphp.fr

ABSTRACT

Objectives To test the reliability of new ultrasound (US) definitions and quantification of synovial hypertrophy (SH) and power Doppler (PD) signal, separately and in combination, in a range of joints in patients with rheumatoid arthritis (RA) using the European League Against Rheumatisms–Outcomes Measures in Rheumatology (EULAR-OMERACT) combined score for PD and SH.

Methods A stepwise approach was used: (1) scoring static images of metacarpophalangeal (MCP) joints in a web-based exercise and subsequently when scanning patients; (2) scoring static images of wrist, proximal interphalangeal joints, knee and metatarsophalangeal joints in a web-based exercise and subsequently when scanning patients using different acquisitions (standardised vs usual practice). For reliability, kappa coefficients (κ) were used.

Results Scoring MCP joints in static images showed substantial intraobserver variability but good to excellent interobserver reliability. In patients, intraobserver reliability was the same for the two acquisition methods. Interobserver reliability for SH (κ=0.87) and PD (κ=0.79) and the EULAR-OMERACT combined score (κ=0.86) were better when using a 'standardised' scan. For the other joints, the intraobserver reliability was excellent in static images for all scores (κ=0.8–0.97) and the interobserver reliability marginally lower. When using standardised scanning in patients, the intraobserver was good (κ=0.64 for SH and the EULAR-OMERACT combined score, 0.66 for PD) and the interobserver reliability was also good especially for PD (κ range=0.41–0.92).

Conclusion The EULAR-OMERACT score demonstrated moderate-good reliability in MCP joints using a standardised scan and is equally applicable in non-MCP joints. This scoring system should underpin improved reliability and consequently the responsiveness of US in RA clinical trials.

Key messages

What is already known about this subject?

► No consensus existed until now on a single ultrasound (US) scoring system for rheumatoid arthritis (RA) clinical trials.

What does this study add?

► A consensus-based US scoring system has been validated in multiple joints and has been shown to be highly reliable.

How might this impact on clinical practice?

► This highly reliable consensus-based scoring system should improve responsiveness and increase the uptake of US in RA clinical trials.

INTRODUCTION

Growing data suggest that ultrasound (US) is a valuable tool for assessing and classifying joint involvement and measuring disease activity based on the detection and scoring of synovitis in patients with rheumatoid arthritis (RA).[1] The benefit of US in the evaluation and monitoring of patients with RA is mainly based on its greater sensitivity in detecting synovitis compared with clinical examination.[2–4] Colour Doppler (CD) and power Doppler (PD) modes are able to detect pathological synovial blood flow, which reflects the inflammatory activity in the joint[5–7] and has predictive value in relation to radiographic progression of structural damage[8–10] and in relation to disease flare.[11–13] In addition, US-detected synovitis aids more

accurate early diagnosis of RA to enable earlier treatment.[14 15]

As RA clinical trials need objective and feasible methods for assessing inflammation response and with clinical practice focusing on tight control of disease activity, it has become imperative to improve the reliability of US in quantifying synovitis. Many scoring systems have been proposed, however, a recent literature review highlighted the lack of an expert-derived consensus.[16]

The Outcomes Measures in Rheumatology (OMERACT) US Working Group in collaboration with an US working party of the European League Against Rheumatisms (EULAR) conducted a series of US studies in order to understand possible reasons for a low agreement in detecting synovitis and to develop and validate an expert-derived consensus for scoring synovitis. The validation process, outlined in supplementary figure 1 in the supplementary online material, was carried out in a multistep approach (four steps) from 2005 to 2014. The first two steps are described in a companion paper[17] in which exercises in static images and in clinical setting revealed that the causes for the inconsistencies and the hampered reliability in scoring synovitis among rheumatologists from different European countries were related to several sources of variability such as the perception and weighting of the different US components (ie, synovial hypertrophy (SH), Doppler activity and also effusion) used for describing and grading the inflammatory process[17] as well as the differences in the US acquisition technique. Based on these discrepancies, the elementary components were redefined by Delphi consensus. It was agreed: (1) not to include effusion as an inflammatory component, as it was considered to be an inconsistent finding, frequently detected in healthy subjects or in inactive RA joints;[18 19] (2) to redefine synovitis based on SH and Doppler only and (3) to score them semiquantitatively (0–3) both separately and in combination using the novel EULAR-OMERACT combined score.[17] These steps were performed using metacarpophalangeal (MCP) joints as a model.

Having established these basic steps,[17] the group moved to the second part of the validation process which is presented in this paper. The objectives were: (1) to evaluate the reliability of the EULAR-OMERACT combined score for grading synovitis in MCP joints, as well as the definition and quantification of SH and PD individually; (2) to test the reliability of a standardised consensus-based acquisition method compared with a 'usual practice' scanning method and (3) to evaluate the reliability of the new definitions for SH, PD and the EULAR-OMERACT combined score in non-MCP joints.

METHODS

Twelve US-experienced rheumatologists, who participated in the first part of the standardisation process,[17] participated in the following steps: (1) testing the validity of the new proposed definitions for scoring SH and PD separately and in combination (the EULAR-OMERACT PDUS score) on static images of MCP joints; (2) applying the same definitions and scoring systems in a real-time patient-based reliability exercise, by comparing a consensus-based scanning acquisition method previously obtained[17] to a 'usual practice' scanning method and (3) testing the reliability of the new definitions and of the EULAR-OMERACT combined score in non-MCP joints (wrist, proximal interphalangeal (PIP), knee and metatarsophalangeal (MTP)) in both reading static images and scanning patients.

In all the reliability exercises, the participants used both a semiquantitative (SQ) (0–3) and a binary score (yes/no). The definitions and the scoring systems used are presented in table 1.

All patients participating in the reliability exercises fulfilled the American College of Rheumatology classification criteria for RA[20] and were attending the rheumatology department of Ambroise Paré hospital in Boulogne-Billancourt (France).

Patients were selected based on the absence of joint deformities and the willingness to take part. The studies were conducted in accordance with the Declaration of Helsinki and each participant gave written informed consent.

Step 1. Web-based exercise

A set of high-quality US images of synovitis of MCP joints were selected from an anonymised register of patients with RA by two independent ultrasonographers (MADA and EN) in order to ensure inclusion of a broad range of synovitis severity. A random selection of images was shown twice in order to assess the intra-reader reliability.

Step 2. Patients-based exercise: scanning patients according to a different scanning approach

The experts performed a bilateral US scanning of the second–fifth MCP joints in eight different patients. The dorsal aspect of the joints was examined twice in two rounds over 2 days. In the first round, using a 'standardised acquisition method', the US examinations were performed using a longitudinal dorsal scan on the middle of the joint, first, in GS and then PD, to detect joint morphological abnormalities and synovial flow respectively. In the second round, a 'usual practice (free) acquisition approach' of the dorsal side of the MCP joint was used. In the standardised scan, the maximal grading was to be assessed in the midline. In the 'usual practice' scan method, the examiner recorded the maximal grading from any area of the joint.

Step 3. Testing the new definitions and the reliability of the EULAR-OMERACT combined score and of SH and PD individually in non-MCP joints

A set of high-quality US images of synovitis of wrist, PIP, knees and MTP joints from patients with RA was evaluated using images from the same register and applying the same approach as described in step 1.

Table 1 Definitions of severity grades (0–3) for each elementary component and for the EULAR-OMERACT combined score

Synovitis	SH (greyscale)	Doppler (PD)	Combined score* (greyscale SH + PD)
Grade 0 (normal)	No SH independently of the presence of effusion	No Doppler signal	No SH and no PD signal
Grade 1 (minimal)	Minimal hypoechoic SH* up to the level of the horizontal line connecting bone surfaces between the metacarpal head and the proximal phalanx	Up to three single Doppler spots OR up to one confluent spot and two single spots OR up to two confluent spots	Grade 1 hypoechoic SH and ≤ grade 1 PD signal
Grade 2 (moderate)	Moderate hypoechoic SH† extending beyond joint line but with the upper surface concave (curved downwards) or hypertrophy extending beyond the joint line but with the upper surface flat	>Grade 1 but ≤50% Doppler signals in the total greyscale background	Grade 2 hypoechoic SH and ≤ grade 2 PD signal; or grade 1 SH and a grade 2 PD signal
Grade 3 (severe)	Severe hypoechoic SH† with or without effusion extending beyond the joint line but with the upper surface convex (curved upwards)	>Grade 2 (>50% of the total greyscale background)	Grade 3 hypoechoic SH and ≤ grade 3 PD signal; or grade 1 or 2 SH and a grade 3 PD signal

*EULAR-OMERACT combined score.
†Independently of the presence of effusion.
EULAR-OMERACT, European League Against Rheumatisms–Outcomes Measures in Rheumatology; PD, power Doppler; SH, synovial hypertrophy.

After the exercise on static images, the experts performed bilateral US scanning of the wrist, PIP,[2–5] knee and MTP[1–5] joints in six different patients twice in two rounds over 2 days (first day wrist and PIP joints, second day knee and MTP joints), using predefined joint positions as follows:

- *Wrist joints* (ie, radiocarpal and midcarpal joints were evaluated as a single site): palms facing down and wrist positioned flat on the examining table, as neutral as possible but relaxed; shoulder and elbow relaxed; elbow rested on the table. Scanning at the level of the radio-lunate joint.
- *PIP joints*: palms facing down and wrist positioned flat on the examining table, as neutral as possible but relaxed, scanning on the dorsal midline aspect.
- *Knee joints* (ie, suprapatellar and parapatellar recesses were scored as a single site): knee 30° flexed and scanning on suprapatellar midline for the suprapatellar recess; knee extended and scanning the parapatellar areas using the retinacula as a landmark for the parapatellar medial and lateral recesses. Doppler signal was recorded only in the medial and lateral parapatellar recesses.
- *MTP joints*: foot placed resting (with knee 30° flexed) over its plantar aspect. Scanning recorded on the dorsal midline aspect.

For all examinations, identical ESAOTE Technos MPX (Genoa, Italy) US machines with an 8–14 MHz linear array transducer were used with identical PD settings (frequency of 10.1 MHz, pulse repetition frequency of 750 Hz and Doppler gain of 50–53 dB). Each patient was assigned to one machine and the sonographers then rotated from one machine to the next in a predefined sequence with 10 min allocated for scanning and recording the findings on a standard score sheet. Participants were blinded to the patients' clinical details (ie, presence or not of active disease).[17]

STATISTICAL ANALYSIS

The intraobserver and interobserver reliability of scoring static and dynamic images were assessed according to weighted Kappa coefficients (κ) relying on absolute differences and in order to take into account the magnitude of discrepancy between categories. Intraobserver coefficients were evaluated on pairs of measures performed by the same sonographer at each site, while interobserver coefficients were exclusively based on the first measure of those pairs. Interobserver reliability was studied by calculating the mean κ for all pairs (ie, Light's κ).[21] Kappa values were evaluated according to Landis and Koch.[22] Percentage of observed agreement (ie, percentage of observations that obtained the same score) and prevalence of the observed lesions were also calculated. Statistical analysis was performed using the R software (http://www.r-project.org/).

RESULTS

Step 1. Testing the definition and reliability of the EULAR-OMERACT combined score on static images

Thirty-six images of MCP joints were scored. Table 2 shows the observed agreement, prevalence and κ values results. The agreement was good for the novel definitions of synovitis components (SH and PD) both separately and in combination (EULAR-OMERACT combined score) with the best obtained for PD alone.

Table 2 Prevalence of observed lesions, observed agreement, intraobserver and interobserver reliabilities when scoring synovitis in metacarpophalangeal joints on static images

	Intraobserver				Interobserver			
	Prevalence % (min–max)	Observed agreement (min–max)	Kappa range (min–max) (y/n) + 95% CI	Kappa range (min–max) (0–3) + 95% CI	Prevalence % (mean)	Observed agreement (mean)	Mean kappa (y/n) + 95% CI	Mean kappa (0–3) + 95% CI
SH	Grade 0: 9–34 Grade 1: 11–33 Grade 2: 8–47 Grade 3: 38–67	0.31–0.94	−0.01–0.94 (−0.31–0.31 to 0.82–1)	0.06–0.96 (−0.42–0.27 to 0.9–1)	Grade 0: 22 Grade 1: 21 Grade 2: 35 Grade 3: 23	0.76	0.57 (0.35 to 0.57)	0.78 (0.54 to 0.78)
PD	Grade 0: 59–71 Grade 1: 4–15 Grade 2: 11–9 Grade 3: 5–14	0.5–1	−0.03–1.0 (−0.34–0.29 to 0.77–1)	0.01–1.0 (−0.28–0.26 to 1–1)	Grade 0: 66 Grade 1: 10 Grade 2: 15 Grade 3: 9	0.88	0.97 (0.8 to 0.97)	0.98 (0.80 to 0.98)
Combined score* (SH+PD)	Grade 0: 13–70 Grade 1: 4–60 Grade 2: 8–45 Grade 3: 6–67	0.39–1	−0.01–1.0 (−0.24–0.4 to 1–1)	0.06–1.0 (−0.0–0.4 to 1–1)	Grade 0: 32 Grade 1: 20 Grade 2: 28 Grade 3: 20	0.81	0.80 (0.51 to 0.80)	0.79 (0.57 to 0.79)

*EULAR-OMERACT combined score.
EULAR-OMERACT, European League Against Rheumatisms–Outcomes Measures in Rheumatology; PD, power Doppler; SH, synovial hypertrophy.

Surprisingly, the intraobserver reliability showed a great variability between the 12 sonographers for all parameters. Similar results were seen for the binary score. The interobserver reliability was good to excellent for the SQ score of SH, PD and the EULAR-OMERACT combined score. For the binary score, the reliability was good to excellent for PD and the EULAR-OMERACT combined score, but only moderate for SH (table 2). When comparing the interobserver reliability for the SQ score with the binary score for PD and the EULAR-OMERACT combined score, reliability showed almost identical κ values—the highest κ values were seen for the PD score (SQ PD score: κ=0. 98 and binary PD score: κ=0.97). For SH, the binary score was considerably lower (κ=0.57) than the SQ score (κ=0.78).

Step 2. Testing the definition and reliability of the EULAR-OMERACT combined score in patients

No major differences were recorded in the intraobserver reliability when scanning in a patient-based exercise for the 'standardised scan' and 'usual practice scan' (slightly better for the standardised scan) for all synovitis components (SH and PD) and the EULAR-OMERACT combined score, and for both binary and SQ grading (table 3).

However, interobserver reliability for both SQ and binary scores for all components was better when using the standardised scan approach (table 4). The κ values were good for SQ PD (κ=0.79) but excellent for SH and the EULAR-OMERACT combined score (κ=0.87 and 0.86, respectively). The SQ score performed slightly better than the binary score for PD (SQ score: κ=0.79; binary κ=0.76) and the EULAR-OMERACT combined score (SQ score: κ=0.86; binary κ=0.85).

Only the PD grading for both the binary score and the SQ score had better interobserver reliability in static images than when scanning patients with a standardised scan (tables 2 and 4).

Step 3. Testing the definition and reliability of the EULAR-OMERACT combined score in other joints

In the web-based exercise on static images, 100 images of wrist, PIP, knee and MTP joints were included representing a broad range of different degrees of synovitis. Table 5 shows the observed agreement, prevalence and reliability of the different degrees of SH, PD and the EULAR-OMERACT combined score. When scoring static images, the intraobserver reliability was good to excellent for SH and PD (better for SQ grading than binary) and excellent for the EULAR-OMERACT combined P score (κ=0.84). The interobserver reliability was good for all components (better for binary score than SQ grading) and best for PD (binary=0.88 and SQ=0.86). Table 6 shows the inter-reader reliability for the synovitis components and the EULAR-OMERACT combined score according to the different joints. The inter-reader reliability for the EULAR-OMERACT combined score in the wrist was κ=0.61, for the PIPs κ=0.75, for knees κ=0.55 and for the MTPs κ=0.58.

When evaluating the EULAR-OMERACT combined score in patients, the intraobserver reliability was good with almost identical values for the binary and SQ scores for all single components and in combination, ranging from 0.64 to 0.66 (table 5). The interobserver reliability was moderate to good for all components (0.43–0.61) and best for the SQ PD score (0.61) (table 5).

Supplementary figure 2 (online file) shows image examples on the EULAR-OMERACT combined score applied to PIP, MTP, knee and wrist joints.

Following the results of this multistep project, the group agreed on the following procedures for scoring synovitis by US: (1) The presence of a hypoechoic SH is mandatory for defining the presence of an US-detected synovitis and for grading Doppler activity. (2) Grading synovitis, at joint level, should be performed by using the SQ EULAR-OMERACT score (based on the combined presence of both GS SH and Doppler (table 1)).

(3) If different areas of severity are present in the same joint, the final severity grade is given by the area with the maximum of severity. (4) The acquisition and grading of synovitis by US should be performed by using a dorsal approach. (5) A standardised scan, with the position of the probe in the midline, should be recommended in the case of multicentre clinical trials using US, although it might underestimate the real inflammatory activity of the joint.

DISCUSSION

Over the last 10 years, the EULAR-OMERACT US group has worked on standardising the US detection, acquisition and grading of synovitis in patients with RA using a stepwise approach. In the first step, the group developed: (1) new definitions of the elementary components and a novel scoring system based on the grade of severity of SH and PD both separately and in combination: the EULAR-OMERACT combined score; and (2) a standardised image acquisition technique.[17] In the second part of this multistep validation process, the reliability of these new definitions and of the scoring system for SH and PD separately, and EULAR-OMERACT combined was tested in static images, then in patients on MCP and non-MCP joints. The participation of the same multinational team in every step of the validation process added value to the consistency of the results.

The new definitions for grading SH and PD independently and combined (EULAR-OMERACT combined score) considerably improved the reliability when scoring both static images and patients. In these studies, PD was chosen as the optimal Doppler modality for the particular US machines used for depicting inflammation, but PD may be substituted by CD in the presented scoring system when working with machines where CD is more sensitive than PD.[23] The interobserver reliability for the EULAR-OMERACT combined score in static images was good to excellent. In patients, the intraobserver reliability of the EULAR-OMERACT combined score showed some

Table 3 Prevalence of observed lesions, observed agreement and intraobserver reliability when scoring synovitis in MCP joints in patients with RA: standardised scan acquisition versus free scan acquisition

	Standardised scan				Free scan			
	Prevalence % (min–max)	Observed agreement (min–max)	Kappa (min–max) + 95% CI		Prevalence % (min–max)	Observed agreement (min–max)	Kappa (min–max) + 95% CI	
			Binary (y/n)	Semiquantitative (0–3)			Binary (y/n)	Semiquantitative (0–3)
SH	Grade 0: 0–80 Grade 1: 5–45 Grade 2: 0–50 Grade 3: 0–15	0.40–1	0.0–1 (–0.21–0.3 to 0.6–1)	0.34–1 (0.32–0.45 to 0.9–1)	Grade 0: 0–55 Grade 1: 5–55 Grade 2: 5–65 Grade 3: 0–5	0.50–0.95	0.0–1 (–0.01–0.03 to 0.6–1)	0.44–0.94 (0.21–0.44 to 0.87–0.97)
PD	Grade 0: 55–95 Grade 1: 0–30 Grade 2: 0–10 Grade 3: 0–5	0.65–1	0.22–1 (0.21–0.3 to 0.8–1)	0.26–1 (0.25–0.29 to 0.8–1)	Grade 0: 55–90 Grade 1: 0–25 Grade 2: 0–15 Grade 3: 0–20	0.65–1	0.38–1 (0.36–0.3 to 0.81–1)	0.29–1 (–0.21–0.3 to 0.6–1)
Combined score* (SH+PD)	Grade 0: 0–80 Grade 1: 10–45 Grade 2: 0–50 Grade 3: 0–15	0.40–1	0.39–1 (0.34–0.43 to 0.7–1)	0.32–1 (0.29–0.34 to 0.8–1)	Grade 0: 0–60 Grade 1: 5–43 Grade 2: 0–70 Grade 3: 0–20	0.50–0.95	0.0–1 (–0.11–0.3 to 0.92–1)	0.40–0.93 (0.40–0.52 to 0.76–0.95)

*EULAR-OMERACT combined score.
EULAR-OMERACT, European League Against Rheumatisms–Outcomes Measures in Rheumatology; MCP, metacarpophalangeal; PD, power Doppler; RA, rheumatoid arthritis; SH, synovial hypertrophy.

Table 4 Prevalence of observed lesions, observed agreement and interobserver reliability when scoring synovitis in MCP joints in patients with RA: standardised scan versus free hand scan

	Standardised scan					Free scan				
	Prevalence % (mean)	Binary (yes–no) mean		Semiquantitative (0–3) mean		Prevalence (mean)	Binary (yes–no) mean		Semiquantitative (0–3) mean	
		Observed agreement	Kappa + 95% CI	Observed agreement	Kappa + 95% CI		Observed agreement	Kappa + 95% CI	Observed agreement	Kappa + 95% CI
SH	Grade 0: 47 Grade 1: 28 Grade 2: 22 Grade 3: 3	0.63	0.87 (0.86 to 0.9)	0.46	0.87 (0.86 to 0.89)	Grade 0: 33 Grade 1: 30 Grade 2: 33 Grade 3 : 4	0.63	0.69 (0.66 to 0.7)	0.41	0.71 (0.69 to 0.73)
PD	Grade 0: 84 Grade 1: 8 Grade 2: 6 Grade 3: 1	0.88	0.76 (0.73 to 0.79)	0.83	0.79 (0.76 to 0.8)	Grade 0: 79 Grade 1: 12 Grade 2: 8 Grade 3: 1	0.84	0.70 (0.68 to 0.72)	0.78	0.75 (0.72 to 0.77)
Combined score* (SH+PD)	Grade 0: 47 Grade 1: 29 Grade 2: 21 Grade 3: 3	0.62	0.85 (0.84 to 0.87)	0.45	0.86 (0.8 to 0.89)	Grade 0: 35 Grade 1: 28 Grade 2: 31 Grade 3 : 5	0.64	0.74 (0.73 to 0.76)	0.40	0.73 (0.72 to 0.76)

MCP, metacarpophalangeal; PD, power Doppler; RA, rheumatoid arthritis; SH, synovial hypertrophy.
*EULAR-OMERACT combined score.
EULAR-OMERACT, European League Against Rheumatisms–Outcomes Measures in Rheumatology

Table 5 Prevalence of observed lesions, observed agreement and intraobserver and interobserver reliabilities when scoring synovitis in non-MCP joints on static images and in patients

Component	Static images					Patients				
	Intra		Inter			Intra		Inter		
	Kappa range (min–max) (95% CI*)	Observed agreement (min–max)	Mean kappa (95% CI)	Observed agreement Mean	Prevalence % (mean)	Kappa range (min–max) (95% CI)	Observed agreement (min–max)	Mean kappa (95% CI)	Observed agreement Mean	Prevalence %(mean)
SH score (0–3)	0.73–0.9 (0.75–0.91 to 0.72–0.94)	0.6–0.95	0.6 (0.35 to 0.59)	0.57	Grade 0 : 18% Grade 1 : 23.9% Grade 2 : 31.6% Grade 3 : 26.6%	0.4–0.82 (0.16–0.38 to 0.60–0.83)	0.54–0.88	0.51 (0.24 to 0.38)	0.59	Grade 0 : 60.4 Grade 1 : 20.8 Grade 2 : 14.7 Grade 3: 4.1
SH detection (yes/no)	0.64–0.84 (0.63–0.88 to 0.73–0.86)		0.68 (0.54 to 0.78)			0.47–0.84 (0.36–0.63 to 0.77–1)		0.43 (0.36 to 0.49)		
PD score (0–3)	0.9–1 (0.94–0.99 to 1–1)	0.88–1	0.86 (0.72 to 0.86)	0.84	Grade 0 : 24.1% Grade 1 : 27.4% Grade 2 : 29.2% Grade 3 : 25.4%	0.48–0.88 (0.42–0.71 to 0.83–0.98)	0.76–0.94	0.61 (0.42 to 0.61)	0.79	Grade 0 : 80.4 Grade 1 : 10.2% Grade 2 : 8.2 Grade 3 : 1.2
PD detection (yes/no)	0.9–1 (0.94–0.99 to 1–1)		0.88 (0.80 to 0.95)			0.47–0.9 (0.3–0.61 to 0.77–1)		0.53 (0.43 to 0.60)		
Combined score* (0–3)	0.75–1 (0.73–1 to 0.96–1)	0.69–1	0.65 (0.41 to 0.65)	0.63	Grade 0 : 14.9% Grade 1 : 14.9% Grade 2 : 34.2% Grade 3: 28%	0.42–0.81 (0.44–0.66 to 0.72–0.94)	0.54–0.88	0.52 (0.36 to 0.52	0.60	Grade 0 : 61.4 Grade 1 : 19.1 Grade 2 : 15.2 Grade 3: 4.3
Combined score* (SH+PD) (yes/no)	0.75–0.82 (0.71–0.87 to 0.72–0.84)		0.6 (0.5 to 0.58)			0.45–0.84 (0.32–0.58 to 0.72–0.94)		0.51 (0.36 to 0.54)		

*EULAR-OMERACT combined score.
PD, power Doppler signal; SH, synovial hypertrophy.

Table 6 Interobserver reliability results for detection and scoring of synovitis components and for the combined PD US score according to the different joints in static images and in patients

	Static images				Patients							
	Wrist	PIP	Knee	MTP	Wrist		PIP		Knee		MTP	
	Mean kappa	Mean kappa	Mean kappa	Mean kappa	Mean observed agreement	Mean kappa	Mean observed agreement	Mean kappa	Mean observed agreement	Mean kappa	Mean observed agreement	Mean kappa
SH score (0–3)	0.51	0.69	0.60	0.54	0.5	0.60	0.7	0.52	0.51	0.5	0.76	0.50
SH detection (yes/no)	0.35	0.74	0.81	0.64		0.23		0.2		0.42		0.34
PD score (0–3)	0.84	0.89	0.85	0.86	0.82	0.67	0.95	0.1	0.53	0.49	0.96	0.54
PD detection (yes/no)	0.91	0.92	0.83	0.89		0.42		0.1		0.41		0.48
Combined score* (0–3)	0.61	0.75	0.55	0.58	0.87	0.67	0.7	0.53	0.67	0.54	0.72	0.50
Combined score* (SH+PD) (yes/no)	0.61	NA	0.62	NA		0.43		0.3		0.41		0.39

*EULAR-OMERACT combined score.
MTP, metatarsophalangeal; NA, non-applicable; PD, power Doppler signal; PIP, proximal interphalangeal; SH, synovial hypertrophy.

variability, probably due to the initial difficulty to apply the new definition in 'real life' scanning. However, the interobserver reliability was good to excellent.

The number of patients involved in the two patient-based reliability exercises can be seen as a limitation as they are in the lower range of the sample sizes usually used in imaging studies.[24] However, as several joints were scanned in each patient (two times 8 joints in the first exercise and 22 joints per patient in the second exercise), which in these exercises are seen as independent contributors, and as several examiners participated, the results can be seen as robust enough for supporting the reliability of the scoring.

By using a stepwise process involving discussion and agreement, we were able to evaluate the real impact of the scanning technique. A standardised approach with the probe in a longitudinal plane on the dorsal aspect and in the midline of the joint was found to improve the reliability as compared with a 'usual practice' scanning approach. This provides further evidence supporting the concept that guidelines for image acquisition are needed and that the dorsal aspect of the joint with the probe in the midline is recommended to improve reliability when assessing small joints[25] in multicentre trials though a free scan may detect more accurately the real amount of inflammation in a joint (as stated in the procedures for scanning (3)).

In the first steps of the validation process, the MCP joint was used as a model for evaluating the scoring systems. However, as the final goal of this process was to produce a generalisable instrument to be used in global assessment of disease activity, other joint areas were subsequently incorporated.[26–28] The grading of SH and PD independently and the combined EULAR-OMERACT score were therefore evaluated in other commonly affected joints such as wrist, PIP, knee and MTP joints. Both intra-reader and inter-reader reliabilities were good in static images but lower in patients compared with the reliability in MCP joints and though a standardised scanning approach was applied in these joints, further training in these joint regions may improve the reliability.

Regarding the US assessment of synovitis, it is important to emphasise that although a binary scoring system may be more reliable than a SQ grading, it is optimal only in monitoring patients for whom synovitis completely disappears during treatment. A binary score does not have the sensitivity to detect partial improvement and relies on the ability of the treatment to leave no residual inflammation, making it unsuitable for monitoring treatment effects in longstanding RA, where a complete disappearance of the SH can be hampered by the copresence of osteoarthritis or when complete remission has not been obtained. Considering that a number of studies have reported the presence of minimal abnormalities in both GS and Doppler in healthy controls,[18 19 29] the development of standardised recommendations for scoring synovitis is a major step forward in the development of the concept of a minimal detectable US synovitis

and defining the threshold of normality. The use of a consensus process, based on the analysis of disagreement and exploration of factors affecting reliability, represents a major advantage of this programme of work. The inclusion of clear definitions of each synovial component and the application of a standardised scanning approach has ensured a highly robust process.

Though the current study did not address possible intermachine variability. This may be a problem in clinical practice best solved by using the best equipment giving the patient optimal evaluation.[23] In multicentre trials, the quality of the machines may be different but is almost equivalent as a prerequisite. In addition, the patient is examined always with the same equipment and same settings minimising the variability.

In our study, the reliability of the EULAR-OMERACT combined score was comparable to that of the elementary synovitis components. This has important implications in multicentre studies as both components can be equally reliable in monitoring RA depending on the joint size (ie, Doppler mode is less sensitive for deep anatomic areas) and the Doppler sensitivity of the US machine.[23] Furthermore, the participation of several sonographers from different countries confirms the applicability of the proposed scoring system to multicentre clinical trials and in daily practice.

In conclusion, using an expert-derived consensus process, the EULAR-OMERACT group have developed a standardised EULAR-OMERACT combined scoring system taking both PD and SH components into account in the evaluation of synovitis of multiple RA joints, which is highly reliable when applied in scanning patients. The reliability was further improved when a standardised scanning procedure was used. The application of the proposed EULAR-OMERACT combined score and the new definition of synovitis based on the presence of SH and PD, as well as a standardised scanning approach for synovitis in RA, will ensure a greater degree of homogeneity and comparability in future US studies and facilitate the development of a Global EULAR-OMERACT Synovitis Scoring system at patient level for monitoring RA activity in clinical trials and routine care. The group is currently working on establishing an optimal reduced joint set for scoring synovitis in patients with RA using the EULAR-OMERACT combined score.

Author affiliations
[1]Rheumatology department, Centre for Rheumatology and Spine Diseases, Rigshospitalet-Glostrup, Copenhagen, Denmark
[2]Rheumatology and Joint and Bone Research Unit, Hospital Universitario Fundacion Jimenez Diaz, Madrid, Spain
[3]Public Health Department, APHP, Hôpital Ambroise Paré, Boulogne-Billancourt; INSERM U1173, Versailles-Saint-Quentin University, Montigny le Bretonneaux, France
[4]Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds and NIHR Leeds Musculoskeletal Biomedical Research Unit, Leeds, UK
[5]Department of Internal Medicine, Rheumatology and Clinical Immunology, Park-Klinik Weissensee, Berlin, Germany
[6]National Institute of Rheumatology and Physiotherapy, Budapest, Hungary
[7]Department of Rheumatology, MC Groep Hospitals, Lelystad, the Netherlands
[8]Rheumatology Unit, Università di Torino, Torino, Italy
[9]Department of Rheumatology, CHRU de Brest, Brest, France
[10]Medical Centre for Rheumatology, Immanuel Krankenhaus, Buch, Berlin, Germany
[11]Department of Rheumatology, Zealand's University Hospital at Kàge, Copenaghen, Denmark
[12]Clinica Reumatologica, Università Politecnica delle Marche, Jesi, Ancona, Italy
[13]Rheumatology Department, APHP, Ambroise Paré Hospital, Boulogne-Billancourt, France
[14]INSERM U1173, Laboratoire d'Excellence INFLAMEX, UFR Simone Veil, Versailles-Saint-Quentin University, Montigny le Bretonneaux, France

Contributors MADA designed the study. All authors contributed to the acquisition of data and have read and revised the manuscript. PA and MADA performed all statistical analysis and interpretation of data. LT and MADA drafted the manuscript.

Funding PGC is supported in part by the National Institute for Health Research (NIHR) Leeds Musculoskeletal Biomedical Research Unit.

Disclaimer The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR nor the Department of Health.

Competing interests None declared.

Patient consent Detail has been removed from this case description/these case descriptions to ensure anonymity. The editors and reviewers have seen the detailed information available and are satisfied that the information backs up the case the authors are making.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement There are no unpublished data available.

REFERENCES

1. Colebatch AN, Edwards CJ, Østergaard M, *et al*. EULAR recommendations for the use of imaging of the joints in the clinical management of rheumatoid arthritis. *Ann Rheum Dis* 2013;72:804–14.
2. Backhaus M, Burmester GR, Sandrock D, *et al*. Prospective two year follow up study comparing novel and conventional imaging procedures in patients with arthritic finger joints. *Ann Rheum Dis* 2002;61:895–904.
3. Wakefield RJ, Green MJ, Marzo-Ortega H, *et al*. Should oligoarthritis be reclassified? ultrasound reveals a high prevalence of subclinical disease. *Ann Rheum Dis* 2004;63:382–5.
4. Naredo E, Bonilla G, Gamero F, *et al*. Assessment of inflammatory activity in rheumatoid arthritis: a comparative study of clinical evaluation with grey scale and power doppler ultrasonography. *Ann Rheum Dis* 2005;64:375–81.
5. Szkudlarek M, Court-Payen M, Strandberg C, *et al*. Power Doppler ultrasonography for assessment of synovitis in the metacarpophalangeal joints of patients with rheumatoid arthritis: a comparison with dynamic magnetic resonance imaging. *Arthritis Rheum* 2001;44:2018–23.
6. Terslev L, Torp-Pedersen S, Savnik A, *et al*. Doppler ultrasound and magnetic resonance imaging of synovial inflammation of the hand in rheumatoid arthritis: a comparative study. *Arthritis Rheum* 2003;48:2434–41.
7. Andersen M, Ellegaard K, Hebsgaard JB, *et al*. Ultrasound colour Doppler is associated with synovial pathology in biopsies from hand joints in rheumatoid arthritis patients: a cross-sectional study. *Ann Rheum Dis* 2014;73:678–83.
8. Naredo E, Möller I, Cruz A, *et al*. Power Doppler ultrasonographic monitoring of response to anti-tumor necrosis factor therapy in patients with rheumatoid arthritis. *Arthritis Rheum* 2008;58:2248–56.
9. Brown AK, Conaghan PG, Karim Z, *et al*. An explanation for the apparent dissociation between clinical remission and continued structural deterioration in rheumatoid arthritis. *Arthritis Rheum* 2008;58:2958–67.

10. Døhn UM, Ejbjerg B, Boonen A, *et al*. No overall progression and occasional repair of erosions despite persistent inflammation in adalimumab-treated rheumatoid arthritis patients: results from a longitudinal comparative MRI, ultrasonography, CT and radiography study. *Ann Rheum Dis* 2011;70:252–8.
11. Scirè CA, Montecucco C, Codullo V, *et al*. Ultrasonographic evaluation of joint involvement in early rheumatoid arthritis in clinical remission: power doppler signal predicts short-term relapse. *Rheumatology* 2009;48:1092–7.
12. Saleem B, Brown AK, Quinn M, *et al*. Can flare be predicted in DMARD treated RA patients in remission, and is it important? A cohort study. *Ann Rheum Dis* 2012;71:1316–21.
13. Foltz V, Gandjbakhch F, Etchepare F, *et al*. Power Doppler ultrasound, but not low-field magnetic resonance imaging, predicts relapse and radiographic disease progression in rheumatoid arthritis patients with low levels of disease activity. *Arthritis Rheum* 2012;64:67–76.
14. Minowa K, Ogasawara M, Murayama G, *et al*. Predictive grade of ultrasound synovitis for diagnosing rheumatoid arthritis in clinical practice and the possible difference between patients with and without seropositivity. *Mod Rheumatol* 2016;26:1–6.
15. Nakagomi D, Ikeda K, Okubo A, *et al*. Ultrasound can improve the accuracy of the 2010 American College of Rheumatology/European League against rheumatism classification criteria for rheumatoid arthritis to predict the requirement for methotrexate treatment. *Arthritis Rheum* 2013;65:890–8.
16. Mandl P, Naredo E, Wakefield RJ, *et al*. A systematic literature review analysis of ultrasound joint count and scoring systems to assess synovitis in rheumatoid arthritis according to the OMERACT filter. *J Rheumatol* 2011;38:2055–62.
17. D'Agostino MA, Terslev L, Aegerter P, *et al*. EULAR-OMERACT Ultrasound Taskforce: development of a standardized synovitis Scoring System in rheumatoid Arthritis. *(accompanying paper) Submitted*.
18. Witt M, Mueller F, Nigg A, *et al*. Relevance of grade 1 gray-scale ultrasound findings in wrists and small joints to the assessment of subclinical synovitis in rheumatoid arthritis. *Arthritis Rheum* 2013;65:1694–701.
19. Padovano I, Costantino F, Breban M, *et al*. Prevalence of ultrasound synovial inflammatory findings in healthy subjects. *Ann Rheum Dis*. In Press. 2016;75:1819–23.
20. Arnett FC, Edworthy SM, Bloch DA, *et al*. The American Rheumatism Association 1987 revised criteria for the classification of Rheumatoid arthritis. *Arthritis Rheum* 1988;31:315–24.
21. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychol Bull* 1971;76:365–77.
22. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977;33:363–74.
23. Torp-Pedersen S, Christensen R, Szkudlarek M, *et al*. Power and color Doppler ultrasound settings for inflammatory flow: impact on scoring of disease activity in patients with rheumatoid arthritis. *Arthritis Rheumatol* 2015;67:386–95.
24. Obuchowski NA. How many observers are needed in clinical studies of medical imaging? *AJR Am J Roentgenol* 2004;182:867–9.
25. Witt MN, Mueller F, Weinert P, *et al*. Ultrasound of synovitis in rheumatoid arthritis: advantages of the dorsal over the palmar approach to finger joints. *J Rheumatol* 2014;41:422–8.
26. Backhaus M, Ohrndorf S, Kellner H, *et al*. Evaluation of a novel 7-joint ultrasound score in daily rheumatologic practice: a pilot project. *Arthritis Rheum* 2009;61:1194–201.
27. Naredo E, Rodríguez M, Campos C, *et al*. Validity, reproducibility, and responsiveness of a twelve-joint simplified power doppler ultrasonographic assessment of joint inflammation in rheumatoid arthritis. *Arthritis Rheum* 2008;59:515–22.
28. Perricone C, Ceccarelli F, Modesti M, *et al*. The 6-joint ultrasonographic assessment: a valid, sensitive-to-change and feasible method for evaluating joint inflammation in RA. *Rheumatology* 2012;51:866–73.
29. Millot F, Clavel G, Etchepare F, *et al*. Investigators of the french early Arthritis Cohort ESPOIR. musculoskeletal ultrasonography in healthy subjects and ultrasound criteria for early arthritis (the ESPOIR cohort). *J Rheumatol* 2011;38:613–20.