

The importance of sonographer experience and machine quality with regards to the role of musculoskeletal ultrasound in routine care of rheumatoid arthritis patients

L. Brulhart¹, H.-R. Ziswiler²,
G. Tamborrini³, P. Zufferey⁴

on behalf of the physicians
participating in the SONAR/SCQM
programmes

¹Rheumatology HUG, Geneva,
Switzerland;

²OsteoRheuma Bern, Bern Switzerland;

³Rheumatology and Ultrasound
Department, Bethesda Hospital Basel,
Switzerland, 4RH/DAL, CHUV, Lausanne,
Switzerland.

Laure Brulhart, MD
Hans- Rudolf Ziswiler, MD
Giorgio Tamborrini, MD
Pascal Zufferey, MD

Please address correspondence to:
Laure Brulhart,
Division of Rheumatology,
Department of Medicine Specialty,
University Hospital of Geneva,
Avenue Beau-Séjour 26,
CH-1211 Geneva 14, Switzerland.
E-mail: laure.brulhart@hcuge.ch

Received on March 1, 2014; accepted in
revised form on June 25, 2014.

© Copyright CLINICAL AND
EXPERIMENTAL RHEUMATOLOGY 2015.

Key words: musculoskeletal
ultrasound, rheumatoid arthritis,
scoring

ABSTRACT

Objective. *Regarding recent progress, musculoskeletal ultrasound (US) will probably soon be integrated in standard care of patient with rheumatoid arthritis (RA). However, in daily care, quality of US machines and level of experience of sonographers are varied. We conducted a study to assess reproducibility and feasibility of an US scoring for RA, including US devices of different quality and rheumatologist with various levels of expertise in US as it would be in daily care.*

Methods. *The Swiss Sonography in Arthritis and Rheumatism (SONAR) group has developed a semi-quantitative score using OMERACT criteria for synovitis and erosion in RA. The score was taught to 108 rheumatologists trained in US. One year after the last workshop, 19 rheumatologists participated in the study. Scans were performed on 6 US machines ranging from low to high quality, each with a different patient. Weighted kappa was calculated for each pair of readers.*

Results. *Overall, the agreement was fair to moderate. Quality of device, experience of the sonographers and practice of the score before the study improved substantially the agreement. Agreement assessed on higher quality machine, among sonographers with good experience in US increased to substantial (median kappa for B-mode and Doppler: 0.64 and 0.41 for erosion).*

Conclusions. *This study demonstrated feasibility and reproducibility of the Swiss US SONAR score for RA. Our results confirmed importance of the quality of US machine and the training of sonographers for the implementation of US scoring in the routine daily care of RA.*

Introduction

Given recent developments, musculoskeletal ultrasound (US) will probably soon be integrated into the standard care of patients with rheumatoid arthritis (RA). Different US scores have been developed and proposed for the evaluation of RA (1-3). They have been shown to correlate with disease activity and response to treatment (2, 4, 5). Although regarded as operator-dependent, studies

addressing inter-observer agreement had shown moderate to good agreement (6, 7). Most studies assessing reproducibility of US were carried out between experts or between two sonographers trained together. For routine care, feasibility and inter-reader agreement need to be assessed among specialists with broad range of experience and practice in US.

In Switzerland, US is part of daily clinical practice of many rheumatologists working in private office or hospitals. Thus, US practice in Switzerland offers the opportunity to assess an US score for the management of RA under daily practice conditions.

The objective of this study was to evaluate the feasibility and inter-reader agreement of an US score for RA among rheumatologists with various levels of expertise in US reflecting the reality in daily routine care.

Materials and methods

The Swiss Sonography in Arthritis and Rheumatism (SONAR) group has developed a semi-quantitative score for RA (<http://www.sonar-group.ch>) using published OMERACT criteria for synovitis and erosion (8). The score includes the assessment of 22 joints: bilateral metacarpophalangeal (MCP) and proximal interphalangeal (PIP) joints 2 to 5, wrist, olecranon fossa and supra-patellar recess. Both synovitis, defined as synovial thickening or effusion, in B-mode and Doppler activity were scored for each joint on a semi-quantitative scale from 0 to 3 (Fig. 1-2) according to the OMERACT recommendations (9). Due to better sensitivity, the longitudinal volar aspect was chosen for synovitis detection in MCP and PIP and the dorsal aspect for Doppler (10, 11). Erosions were recorded in MCP 2 and 5, PIP 2 to 5 and ulnar styloid and were graded on a semi-quantitative scale (0: no erosion, 1: small erosion, 2: medium erosion, 3: large or multiple erosions).

During two workshops with theory and practice, the scoring method was taught to 108 Swiss rheumatologists trained in US. Subsequently, these rheumatologists were encouraged to practice the score in their offices. About 9 months

Competing interests: none declared.

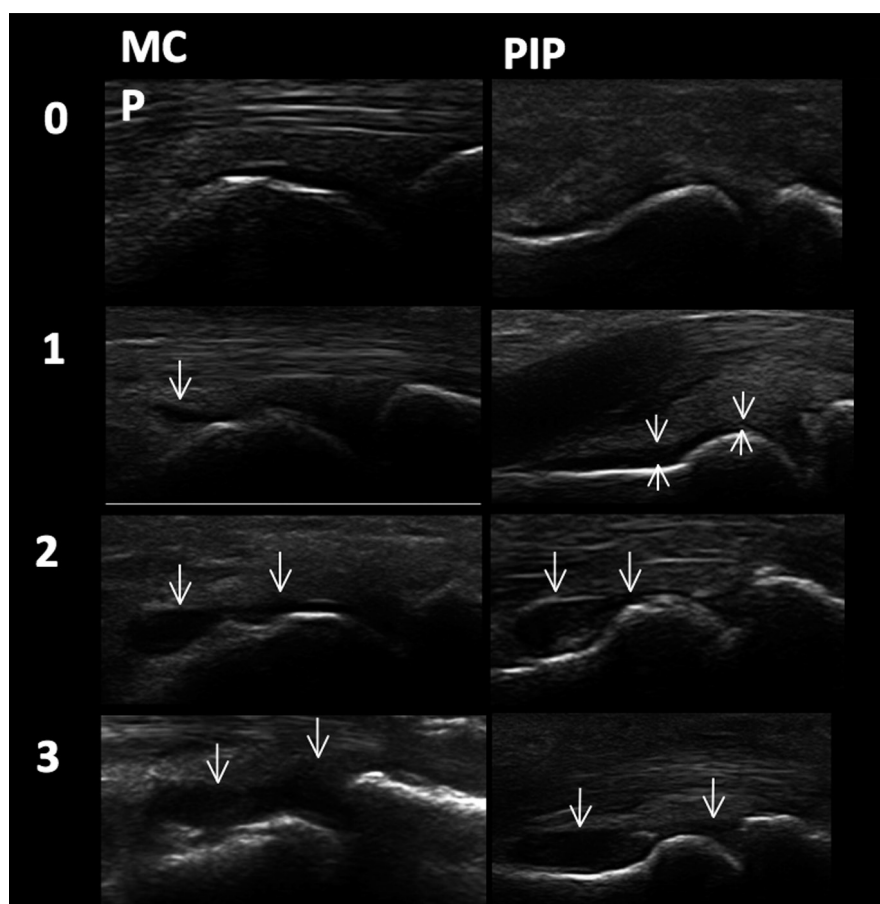


Fig. 1. Semi-quantitative score for synovitis in B-mode. 0: No well visualised hypo or anechoic swelling of the synovium.

1: Hypo or anechoic swelling of the synovium thicker than cartilage, not reaching the bone level of the metacarpal or proximal phalangeal head. 2: Hypo or anechoic bloated swelling of the synovium reaching the level of the bone proximally but not distally. 3: Hypo or anechoic bloated swelling of the synovium reaching the phalanx.

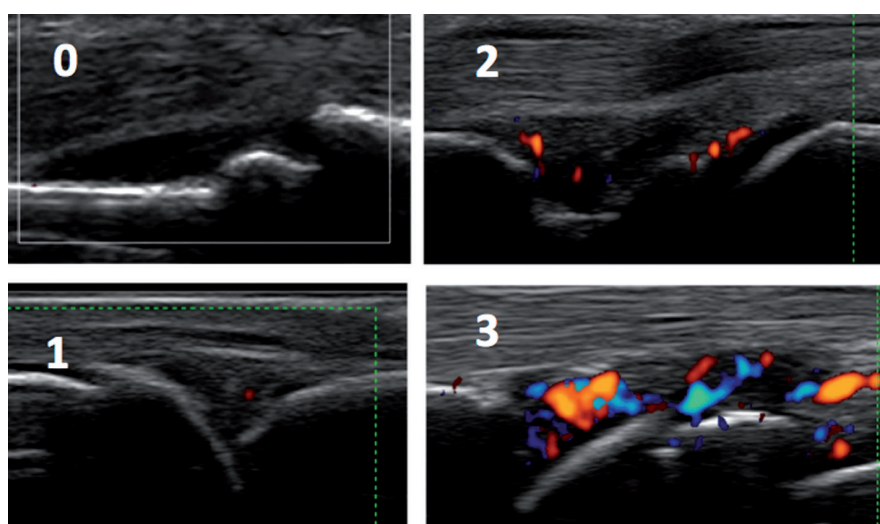


Fig. 2. Semi-quantitative score for Doppler activity. 0: non-Doppler activity.

1: Few points, non-confluent. 2: confluent, less than 50% of the synovium. 3: confluent, equal to or more than 50% of the synovium

after the last workshop, 19 rheumatologists who had participated in the SONAR educational courses were

randomly selected and invited for the inter-observer exercise presented in this study.

Ultrasound devices

Examinations were performed using six different ultrasound machines, each with a different patient: two SonoSite M-Turbo equipped with a 6–13 MHz linear array transducer, two Esaote MyLab25 equipped with a 6–18 MHz linear array transducer, one GE Voluson E8 equipped with a 6–15-D Matrix linear array probe and one GE Voluson i equipped with a 4.5–16.5 MHz linear array transducer.

Statistics

Statistics were performed using STATA version 11 for Windows (STATA Corp, Texas, USA). To measure the level of agreement, we computed weighted 2 kappa statistics (kappa) for each pair of assessors who had scored the same patient. Kappa scores between 0.81 and 1 were considered as very good agreement, 0.61 to 0.8 as substantial agreement, 0.41 to 0.60 as moderate agreement, 0.21 to 0.4 as fair agreements and, 0.00 to 0.2 as little agreement and below 0 as poor agreement (12). Due to the lack of precision of the semi-quantitative score, the kappa was calculated according to presence or absence for erosion. One assessor clearly out of range for B-mode scan was excluded from analysis.

Results

Nineteen assessors have performed 29 scans, each patient having been evaluated by 4 to 6 readers. Eight of the 29 examinations were not completed. Five out of the 8 incomplete examinations were performed by rheumatologists who had practiced the score less than 5 times prior to the study. All missing data were related to erosion scoring except 3 cases with missing elbows scoring. Median time requirement for score completion was 20 minutes (interquartile range (IQR): 20 to 30 minutes).

Sonographers

All sonographers were rheumatologists and all except two had attended the basic education course in musculoskeletal ultrasound of the Swiss Society of Ultrasound in Medicine (SGUM). Median number of years of experience since the end of training was 4 (ranging from 2

Table I. Influence of previous experience with the score, level of expertise in ultrasound and quality of the device on inter-reader agreement.

		> 5 past scores (n:7)	< 5 past scores (n:12)	Performing ≥100 US/year (n:27)	Performing <100 US/year (n:7)	High device quality [#] (n:37)	low device quality ^{##} (n:20)
Grayscale	Median :	0.54	0.39	0.64	0.19	0.55	0.33
	IQR:	0.53 – 0.72	0.15 – 0.52	0.53 – 0.72	0.12 – 0.42	0.42 – 0.7	0.14 – 0.42
Doppler	Median :	0.85	0.07	0.63	0.00	0.5	0
	IQR :	0.71 – 0.89	0.0 – 0.32	0.44 – 0.78	-0.04 – 0.21	0.33 – 0.77	-0.07 – 0.18
Erosion	Median :	missing data	0.32	0.41	0.12	0.37	0.14
	IQR:		0.11 – 0.43	0.25 – 0.58	0.11 – 0.15	0.14 – 0.53	0.09 – 0.4

IQR: interquartile range.

^{*}Defined as having completed US training and performing at least 100 US per year. [#]Esaote MyLab25, GE Voluson E8 or GE Voluson i. ^{##}SonoSite M-Turbo.

to 12). Two rheumatologists were still in training, working in university hospital under the supervision of members of the SONAR group. Ten assessors worked in private office, 2 in public hospitals and 7 in university hospitals. Six performed over 200 US per year and three performed less than 50 US per year. Four rheumatologists had not practiced the score after having attended the SONAR educational courses. Among others, median number of score performed prior to the study was 5 (ranging from 2 to 30).

Patients

All 6 patients had RA as judged by a rheumatologist. Five were female with a median age of 65.5 years old (4–71). Five out of 6 were either seropositive or erosive. All patients had active dis-

ease with a median DAS28ESR of 4.8 (3.4–5.9).

Inter-reader agreement

Overall, the agreement was fair to moderate. The mean kappa for B-mode was 0.45 (standard deviation (SD) 0.22) and 0.37 (SD 0.33) for Doppler. Median kappa for erosion was 0.33 (IQR: 0.1–0.51). Both the level of experience in US and previous practice of the score improved the agreement substantially (Table I).

Moreover, the agreement was greatly influenced by the quality of US machines. Median kappa was 0.55 (IQR: 0.42–0.7) for B-mode when assessed on a good quality device compared with 0.33 (IQR: 0.14–0.42) on a low quality device.

When assessed on a high quality device and among sonographers who performed at least 100 examinations per year, the intra-reader agreement increased to substantial. Median kappa was 0.64 (IQR: 0.53 to 0.72) for B-mode, 0.64 (IQR: 0.44–0.78) for Doppler and 0.42 (IQR: 0.22–0.58) for erosions. Lower kappa values were obtained when one of the assessors had practiced the score less than five times before the study.

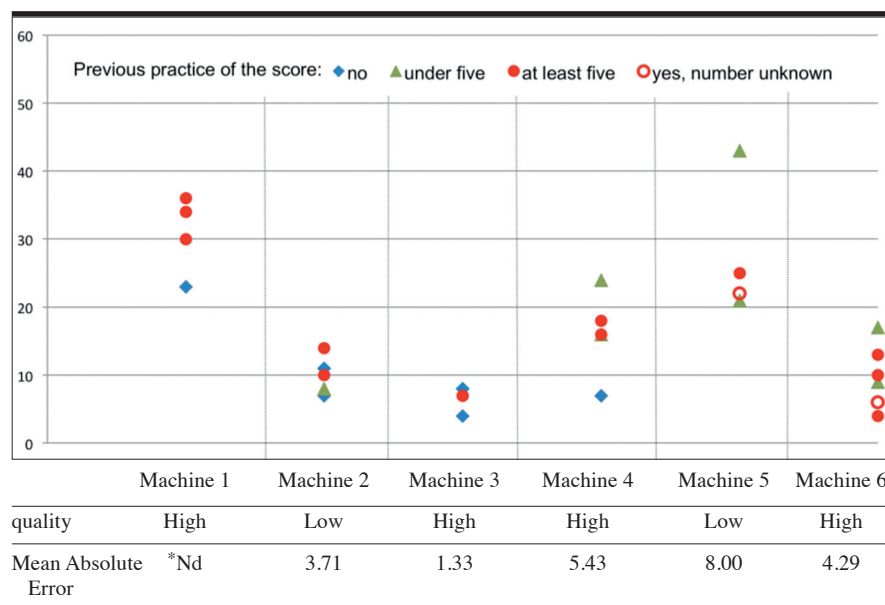
For the calculation of the mean absolute error, examination done by a tutor of the SONAR group was defined as the reference value for each machine. For B-mode, the mean absolute error was 4.64 (SD: 3.82) (Fig. 3). Due to relatively low Doppler activity and number of erosions, the mean absolute error was not calculated for Doppler and erosions.

Discussion

This is the first study to assess feasibility and inter-reader agreement of an US score for standard care of patients with RA taking into account quality of US machines and experience of the assessors.

The SONAR score was feasible with a median time required to be performed of 20 minutes.

Inter-reader agreement was moderate for B-mode and fair for Doppler and erosion assessment. Reproducibility of the clinical examination was not tested in this study. However, clinical assessment, which is still the gold standard for RA management, had a fair to moderate inter-observer agreement as well in a recent publication (13). The reproducibility of the SONAR score



*Nd: no expert had scanned on the machine 1.

Fig. 3. score in B-mode for each assessor and mean absolute error.

was substantially improved when performed on a good quality US machine and assessed among rheumatologist with higher experience in US. Under these conditions, reproducibility was moderate to substantial (median kappa: 0.64 for B-mode and Doppler). These results were similar to those of two recent studies addressing reproducibility of US among sonographers of various level of experience (14, 15). Chavez-Lopez *et al.* founded a kappa value of 0.48 for synovial hypertrophy, 0.47 for Doppler and 0.63 for erosions (15). The inter-observer reliability of the US score has shown moderate to substantial kappa values (14). D'Agostino *et al.* did not find much influence of the machine type on the reliability. However, the inter-observer exercise was conducted among a group of experts which does not reflect the daily care practice (16). In accordance with our results, previous studies have already shown that consensus guidelines and training improve US reliability 7. Moreover, our results drew attention to the need to standardise US examination for inclusion in standard care. Similarly, in the diagnosis of primary Sjögren's syndrome, despite of a satisfactory intra-observer agreement for minor salivary gland biopsy interpretation, Tavoni *and al.* have observed significant discrepancies emphasising the need of a greater standardisation of the procedure (17). The lower agreement for erosion assessment in our study may be explained once again by the low quality of some of the machines, by a lack of precision in the definition of the semi-quantitative erosion score and number of missing data among assessor who had already practiced the score. The main limitation of our study was the small number of US score done on each patient. A broad range of results was expected because of the difference in the level of experience with US, especially in applying the score, and the different quality of devices. Less experienced assessors tended to scan on lower quality machine and they had less often practiced the score prior to

the study, overestimating perhaps effect of the experience on reproducibility. As some rheumatologists still scan one low quality machine in private office we have decided to include one of them (SonoSite equipped with a 6–13 MHz linear array transducer) in our evaluation. Our results confirm that this type of machine does not have enough resolution for synovitis assessment on small joints and emphasise the importance of device for RA scoring. The impact of disease activity on the reproducibility could not be assessed due to the small differences in disease activity between patients.

In conclusion, this is the first study to demonstrate feasibility and reproducibility of the Swiss US SONAR score for RA by operators with various levels of expertise in US reflecting the real life situation in routine daily care. Our results confirm that attention needs to be paid on the quality of the US device and the training of sonographers before US scoring can be endorsed in general clinical practice for RA management.

Acknowledgement

We would like to thank Thomas Gerber, Andreas Krebs, and Stephan Mariacher-Gehler, members of the SONAR group, and all the Swiss rheumatologists who participated in the study.

References

1. NAREDO E, RODRIGUEZ 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.
2. 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.
3. OHRNDORF S, BACKHAUS M: Advances in sonographic scoring of rheumatoid arthritis. *Ann Rheum Dis* 2013; 72 (Suppl. 2): ii69-75.
4. DOUGADOS M, JOUSSE-JOULIN S, MISTRETTA F *et al.*: Evaluation of several ultrasonography scoring systems for synovitis and comparison to clinical examination: results from a prospective multicentre study of rheumatoid arthritis. *Ann Rheum Dis* 2010; 69: 828-33.
5. NAREDO E, MOLLER I, CRUZ A, CARMONA L, GARRIDO J: Power Doppler ultrasonographic monitoring of response to anti-tumor necrosis factor therapy in patients with rheumatoid arthritis. *Arthritis Rheum* 2008; 58: 2248-56.
6. SCHEEL AK, SCHMIDT WA, HERMANN KG *et al.*: Interobserver reliability of rheumatologists performing musculoskeletal ultrasonography: results from a EULAR "Train the trainers" course. *Ann Rheum Dis* 2005; 64: 1043-9.
7. NAREDO E, MOLLER I, MORAGUES C *et al.*: Interobserver reliability in musculoskeletal ultrasonography: results from a "Teach the Teachers" rheumatologist course. *Ann Rheum Dis* 2006; 65: 14-9.
8. WAKEFIELD RJ, BALINT PV, SZKUDLAREK M *et al.*: Musculoskeletal ultrasound including definitions for ultrasonographic pathology. *J Rheumatol* 2005; 32: 2485-7.
9. NAREDO E, WAKEFIELD RJ, IAGNOCCO A *et al.*: The OMERACT ultrasound task force--status and perspectives. *J Rheumatol* 2011; 38: 2063-7.
10. VLAD V, BERGHEA F, LIBIANU S *et al.*: Ultrasound in rheumatoid arthritis - volar versus dorsal synovitis evaluation and scoring. *BMC Musculoskelet Disord* 2011; 12: 124.
11. SCHEEL AK, HERMANN KG, KAHLER E *et al.*: A novel ultrasonographic synovitis scoring system suitable for analyzing finger joint inflammation in rheumatoid arthritis. *Arthritis Rheum* 2005; 52: 733-43.
12. LANDIS JR, KOCH GG: The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-74.
13. SALAFFI F, FILIPPUCCI E, CAROTTI M *et al.*: Inter-observer agreement of standard joint counts in early rheumatoid arthritis: a comparison with grey scale ultrasonography--a preliminary study. *Rheumatology (Oxford)* 2008; 47: 54-8.
14. OHRNDORF S, FISCHER IU, KELLNER H *et al.*: Reliability of the novel 7-joint ultrasound score: results from an inter- and intraobserver study performed by rheumatologists. *Arthritis Care Res (Hoboken)* 2012; 64: 1238-43.
15. CHAVEZ-LOPEZ MA, HERNANDEZ-DIAZ C, MOYA C *et al.*: Inter- and intra-observer agreement of high-resolution ultrasonography and power Doppler in assessment of joint inflammation and bone erosions in patients with rheumatoid arthritis. *Rheumatol Int* 2013; 33: 173-7.
16. D'AGOSTINO MA WR, BACKHAUS M, BALINT PV *et al.*: Combined evaluation of influence of sonographer and machine type on the reliability of power Doppler ultrasonography (PDUS) for detecting, scoring and scanning synovitis in rheumatoid arthritis patients: results of an intermachine reliability exercise. *Ann Rheum Dis* 2008; 67: 421.
17. TAVONI AG, BALDINI C, BENCIVELLI W *et al.*: Minor salivary gland biopsy and Sjögren's syndrome: comparative analysis of biopsies among different Italian rheumatologic centers. *Clin Exp Rheumatol* 2012; 30: 929-33.