RHEUMATOLOGY

Rheumatology 2017;56:255–262
doi:10.1093/rheumatology/kew385
Advance Access publication 14 November 2016

Original article

Analysis of correlation and causes for discrepancy between quantitative and semi-quantitative Doppler scores in synovitis in rheumatoid arthritis

Hamed Rezaei[1,2], Erik af Klint[2], Hilde Berner Hammer[3], Lene Terslev[4], Maria Antonietta D'Agostino[5], Yogan Kisten[1] and Laurent Arnaud[1]

Abstract

Objectives. The aim of this study was to evaluate the association between two semi-quantitative Doppler US scoring systems (SQS), and the quantitative scoring (QS) of Doppler pixel count.

Methods. Adult patients with RA and inadequate clinical response to anti-rheumatic therapy were examined with musculoskeletal US (MSUS). Dorsal MSUS of the wrists, MCP and MTP 2–5 were performed. MSUS images with sign of synovitis were collected and the QS was measured. Five assessors blinded to the QS evaluated the images independently, according to either SQS method. Association between QS and SQS was studied using correlations and multilevel models taking into account the clustering of ratings at the rater, patient and joint levels.

Results. Analysis of the 1190 ratings revealed a strong correlation ($\rho = 0.89$, $P < 0.0001$) and significant associations ($P < 0.0001$) between QS and SQS. Correlations between QS and SQS according to Szkudlarek *et al.* ($\rho = 0.87$, $P < 0.0001$) or Hammer *et al.* ($\rho = 0.91$, $P < 0.0001$) were similar. A total of 239 (20.1%) images were given a SQS grade that did not match that expected based on initial QS, using pre-defined cut-offs. Main explanations for discrepancies were different perceived region of interest (40.7%) and Doppler pixel count near cut-offs between SQS grades (32.3%).

Conclusion. We showed that both SQS methods correlated well with QS to assess synovitis, but SQS methods are intrinsically limited when the Doppler pixel count is close to the cut-offs between the SQS grades. Analysis discrepancies between these methods may help further revision of criteria used to assess disease activity with MSUS in RA.

Key words: rheumatoid arthritis, musculoskeletal ultrasound, scoring systems

CLINICAL SCIENCE

Rheumatology key messages

- Doppler semi-quantitative methods correlated well with quantitative Doppler pixel count to assess synovitis in RA.
- Doppler semi-quantitative methods are limited when Doppler pixel count is close to the cut-offs in RA.
- Analysis of discrepancies of the semi-quantitative methods may help revising the criteria for assessing the grade of synovitis.

[1]Department of Rheumatology, institute of medicine, Karolinska University Hospital, [2]Unit of Clinical Therapy Research, Inflammatory Diseases (ClinTRID), Karolinska Institutet, Stockholm, Sweden, [3]Department of Rheumatology, Diakonhjemmet Hospital, Oslo, Norway, [4]Center for Rheumatology and Spine Disease, Copenhagen, Denmark, Rigshospitalet and [5]APHP, Hôpital Ambroise Paré, Rheumatology Department, Boulogne-Billancourt, INSERM U1173, Laboratoire d'Excellence INFLAMEX, UFR Simone Veil, Versailles-Saint-Quentin University, Saint-Quentin en Yvelines

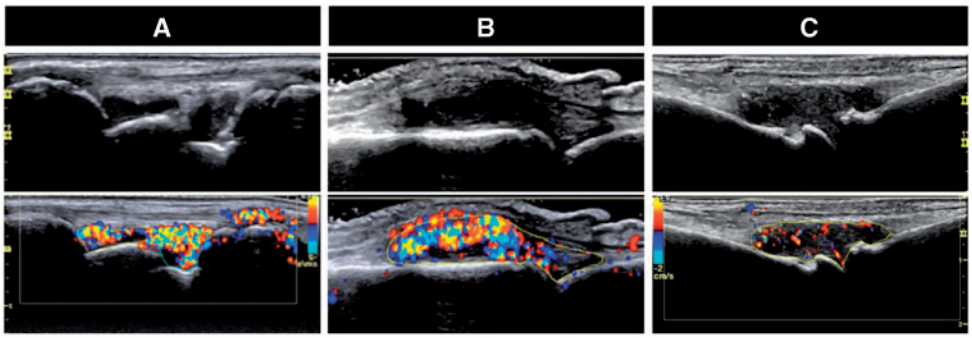
Submitted 31 May 2016; revised version accepted 19 September 2016

Correspondence to: Hamed Rezaei, Department of Rheumatology, Institute of Medicine, Karolinska University Hospital and Karolinska institutet, SE-171 76 Stockholm, Sweden. E-mail: hamed.rezaei@karolinska.se

Introduction

Different clinical scoring systems are used to monitor disease activity in RA [1–3]. However, subclinical activity may be observed despite clinical remission, and lead to radiographic progression [4–7]. Musculoskeletal US (MSUS), particularly Doppler US, is increasingly used for the evaluation of synovial vascularization, and therefore plays a major role in disease monitoring in RA [8–14]. The amount of Doppler signals measured in the synovial

Fig. 1 Region of interest (ROI) for both SQS and QS was characterized according to the OMERACT definition for synovitis (an anechoic or hypoechoic region over the dorsum of the joint, visible in longitudinal planes which may exhibit positive colour Doppler signals or not)



(**A**) Dorsal longitudinal radial scan of the wrist; (**B**) dorsal longitudinal scan of the MCP joint; (**C**) dorsal longitudinal scan of the MTP joint.

tissue, that is, the region of interest (ROI) according to either semi-quantitative scoring (SQS) or quantitative scoring (QS) methods [15–19], is used as a surrogate measure of synovial inflammation in RA joints. SQS methods visually score the amount of Doppler signals (i.e. the colour fraction), in the synovial area using 4-grade ordinal ranking systems. Specifically, Szkudlarek *et al.* [18], used a scale based on the proportion of Doppler signal (grade 0: no flow; grade 1: single vessel signals; grade 2: confluent vessel signals in less than half the area of the synovium; grade 3: confluent vessel signals in more than half the area of the synovium) whereas Hammer *et al.* [15] have suggested to score the amount of Doppler signals in the synovium into none, mild, moderate and major after visual comparison with a reference atlas. None of the SQS has been chosen by consensus so far [20], and this creates some heterogeneity in MSUS scoring in clinical practice and research.

Conversely, a major strength of the QS system is that it allows objective measurement of Doppler pixels (i.e. the colour fraction in the synovial area) using a continuous numeric scale [7, 19] that is thought to reflect the inflammatory activity in each synovial area. While SQS and QS scoring systems have separately shown high reliability when applied to the same patient cohort [21], only limited knowledge exists as to the degree to which these scoring systems agree upon each other.

The aim of this study was therefore to evaluate the association between both SQS systems described by Szkudlarek *et al.* and Hammer *et al.* [15, 18] and the QS system described by Qvistgaard *et al.* [19] in RA patients with active disease. Also, we sought to elucidate the reasons for potential discrepancies between SQS and QS assessments, in order to better understand the intrinsic limitations of these methods, and provide a rational basis for their future revision.

Methods

Study design and patients

Adult patients fulfilling the 1987 ACR classification criteria [22] and/or the 2010 ACR/EULAR classifications criteria [23] for RA, with inadequate clinical response to DMARDs and/or first anti TNF therapy were examined with MSUS between 2012 and 2015, as part of a clinical evaluation upon entering a prospective observational study with add-on therapy. Clinical data including DAS28 [3] and its components, status for ACPA and RF were obtained for each patient. The study was approved by the regional ethics committee (Karolinska Institutet, Stockholm, Sweden). All patients received oral and written information prior to inclusion, and provided written consent.

Musculoskeletal US scanning

MSUS scanning was performed by one participant (H.R.) using the same settings on the same US equipment (General Electric LOGIQ E9 unit, Wauwatosa, WI, USA) with a linear array transducer. The brightness modulation frequency was set at 15 MHz. The colour Doppler US setting was adjusted according to recommendations [24] and was as follows: frequency of 10 MHz, pulse repetition frequency of 0.5 KHz, wall filter of 69 Hz and the gain level for colour Doppler US was just below the noise level.

The dorsal wrist was scanned longitudinally in central and radial positions. Both radiocarpal and midcarpal joints were included in both positions and were evaluated as two separate joints (Fig. 1). Second to fifth metacarpophalangeal (MCP 2-5) and metatarsophalangeal (MTP 2-5) joints were also scanned in dorsal longitudinal positions in all patients (Fig. 1). Including the ulnar aspect of the wrist joint was avoided for measurement of the colour Doppler pixel count as this position should have demanded separate tracing of distal radio-ulnar joint, the

joints between os triquetrum and os hamatum; and between os hamatum and the fourth and fifth metacarpal bones. Additionally, there is not direct articulation between the distal ulnar and the carpal bones which is separated by the triangular disc.

Generation of study images and colour Doppler US scoring

All joints with grade 1 or more synovial hypertrophy in grey-scale and/or positive colour Doppler activity in the synovial area were selected. The ROI for both SQS and QS was defined in accordance with the OMERACT definition [25] of synovial hypertrophy and synovitis (Fig. 1). For each selected joint, a 4 s US film clip was recorded on the US machine. Based on each US film clip, the image with the highest colour Doppler activity and no colour Doppler noise was recorded and computerized measurement of colour Doppler pixel count (QS) was performed by one participant (H.R.), according to Qvistgaard *et al.* [19].

All recorded images were then distributed to each of the five participants, who performed SQS scoring according to either Szkudlarek *et al.* [18] (three participants: H.R., L.T. and E.K.) or Hammer *et al.* [15] (two participants: H.B.H. and Y.K.). All participants were blinded to each other, as well as to the QS assessments. HR performed the SQS >1 year after he measured the QS, and thus was also considered blinded.

Statistical analysis

Descriptive data are presented as counts (percentage), medians (range) or as means (S.D.) and their 95% CI. Agreement between the five participants regarding the SQS assessments was performed using a two-way mixed consistency intra-class correlation coefficient (ICC).

In order to assess the association between SQS and QS, we built a linear mixed model considering QS as the dependent variable and SQS as the fixed effect (unadjusted model). Then, multilevel hierarchical models were constructed to take into account the clustering of scorings within each participant (model 1), within each participant and each patient (model 2) and finally within each participant, patient and joint (with wrist, MCP and MTP joints considered separately) (model 3).

Then, we built generalized linear mixed models using a multinomial distribution with logit link function considering QS as the explanatory variable, and SQS grades as the dependent variable, and the same multilevel clustering of ratings as in models 1, 2 and 3, respectively. The predicted values generated by each of these models were saved, and receiver operating characteristic (ROC) curves were built from the values, to determine the sensitivity and specificity of various QS cut-off values to distinguish between SQS grades.

Finally, we defined discrepant cases as those for which the SQS given by a participant differed from the theoretical grade calculated from the QS, using the following rules: QS of 0% for grade 0, QS between 1 and 9% for grade 1, between 10 and 49% for grade 2 and >50% for grade 3. These discrepant cases were extracted, and each participant was asked to re-grade his/her own discrepant cases, blinded to the initial SQS grading and original QS, and to provide an explanation for the discrepancy. Then, discrepant images and explanations provided were reviewed in consensus by four authors (H.R., E.K., Y.K. and L.A.), and classified into a limited number of categories. A new measurement of the colour Doppler pixel count from the recorded images was performed on the same US machine for all discrepant images by the same participant (H.R.) as for the initial evaluation of the QS. The test-retest reproducibility of QS was very good [ICC: 0.97 (95% CI: 0.96–0.98)], with a mean (S.D.) difference between both QS values of 1.59 ± 7.67 points on the 0% to 100% scale.

All tests were bilateral and p-values <0.05 were considered statistically significant. The statistical analyses were performed with the IBM SPSS Statistics 23 software (IBM Corp., Armonk, New York, USA).

Table 1 Baseline characteristics

Baseline characteristics (number of patients: 37)	
Age, mean (S.D.), years	56 (14.4)
Disease duration, mean (S.D.), years	8.9 (9.5)
Female sex, *n* (%)	25 (67.6)
Current smokers, *n* (%)	10 (27)
ACPA positivity, *n* (%)	30 (81.1)
RF positivity, *n* (%)	25 (67.6)
DAS28, mean (S.D.)	4.6 (1.3)
ESR, mean (S.D.)	21 (17.2)
CRP, mean (S.D.)	7.24 (10.23)
Swollen joint count, mean (S.D.)	6.0 (3.9)
Tender joint count, mean (S.D.)	9.3 (6.1)

Results

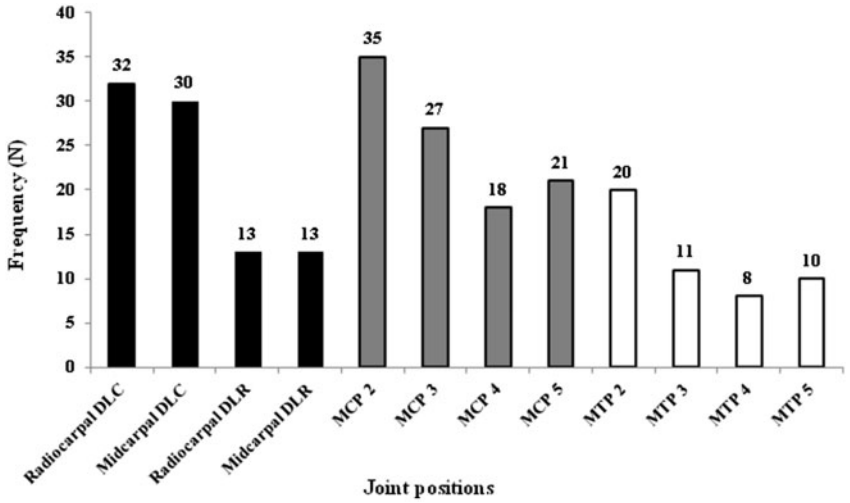
Patient characteristics

This study included 37 adult RA patients (Table 1). All patients had at least moderate disease activity according to the DAS28, and ongoing anti rheumatic treatment with stable dosage during the last 3 months before inclusion. Among them, 33 patients (89.2%) were treated with methotrexate at a mean (S.D.) dosage of 17.9 (3.7) mg, 13 patients (35.1%) received oral corticosteroids at a mean (S.D.) dosage of prednisone of 5.4 (2.9) mg and 6 patients (16.2%) were treated with anti-TNF therapies in combination with DMARDs.

Number of study images recorded for each joint

Out of a total of 888 joints assessed, 238 (26.8%) had grade 1 or more synovial hypertrophy in grey-scale and/or positive colour Doppler activity in the synovial area, and were therefore recorded for further SQS and QS assessments. The total number of images recorded for each joint

FIG. 2 Distribution of 238 joints with a sign of active synovitis



Number of patients: 37. DLC: dorsal longitudinal central; DLR: dorsal longitudinal radial.

was as follows: MCP joints, n = 101 (42.4%); wrist, n = 88 (37.0%); MTP joints: 49 (20.6%) (Fig. 2).

Agreement of SQS between participants

Each of the five participants performed SQS assessment of the 238 recorded images, yielding a total of 1190 joint ratings. The distribution of SQS grades given by participants was as follows: grade 0 in 103 ratings (8.7%), grade 1 in 230 (19.3%), grade 2 in 469 (39.4%) and grade 3 in 388 (32.6%), independently of which scoring system was used. Excellent reliability of the SQS scoring was observed, with very high agreement between participants for SQS assessments [ICC: 0.97 (95% CI: 0.96, 0.97)].

Associations between QS and SQS

Considering the 1190 joint ratings, we observed a strong correlation between the QS and SQS assessments as a whole [$\rho = 0.89$ (95% CI: 0.87, 0.90), $P < 0.0001$]. Further, the correlations between QS and SQS as described by either Szkudlarek *et al.* [18] [$\rho = 0.87$ (95% CI: 0.86, 0.89)], or Hammer *et al.* [15] [$\rho = 0.91$ (95% CI: 0.89, 0.92)] were similar.

Due to the highly correlated structure of the dataset (data not shown), we estimated the association between QS and SQS, as well as the mean QS for each SQS grade (Table 2), using multilevel hierarchical models taking into account the clustering of data within each participant (model 1), within both participants and patients (model 2), as well as within patients, participants, and joints (model 3). We found a significant association between QS and SQS in all these models ($P < 0.0001$, for all). Similarly, we confirmed a significant association ($P < 0.0001$, for all) between QS and SQS as described by either Szkudlarek *et al.* [18] or Hammer *et al.* [15].

Sensitivity and specificity of predefined QS cut-off values to distinguish between SQS grades

Using the multilevel hierarchical models, we found that a QS cut-off value of 10% yielded a sensitivity of 96–99% and a specificity of 77–85% to distinguish between grades 1 and 2 (Table 3). Similarly, using a QS cut-off value of 50% yielded of sensitivity of 98–99% and a specificity of 91–92% to distinguish between grades 2 and 3 (Table 3).

Number and location of discrepancies between SQS as graded by participants and SQS derived from QS

In 239 (20.1%) of the 1190 joint ratings, participants gave a SQS grade that did not match that expected based on our suggested QS cut-off values (Table 3). Four of these 239 discrepant joint ratings (3.4%) were misclassified by all 5 participants, 14 (11.8%) by 4 participants, 20 (16.8%) by 3 participants, 22 (18.5%) by 2 participants and 59 (49.6%) by a single participant. The proportion of discrepant ratings was significantly higher ($P = 0.045$) in the wrist ($n = 105$, 23.9%) compared with the MCPs ($n = 90$, 17.2%) and the MTPs ($n = 44$, 18.0%).

Explanations for discrepancies between SQS as graded by participants and SQS derived from QS

Each participant was then asked to grade again his/her own discrepant images according to the same SQS systems, blinded to the initial SQS and QS assessments, and to provide a short explanation for the discrepancy. These new SQS grades were the same as the initial ones in 117 (48.9%) of the 239 discrepant images, differed by 1 grade (e.g. grade 2 instead of grade 3) in 120 images (50.2%), and by 2 grades (e.g. grade 1 instead of grade 3) in 2 images (0.8%). Valuable explanations allowing us to understand the reasons for discrepancy could be provided for 130 (54.4%) of the 239 discrepant images and are shown in Table 4. The main reasons were a

Table 2 Mean colour Doppler pixel counts for each colour Doppler semi-quantitative score

Semi quantitative score	Colour Doppler pixel count, %			
	Raw data Mean (95% CI)	Model 1 Mean (95% CI)	Model 2 Mean (95% CI)	Model 3 Mean (95% CI)
Grade 0	0.40 (0[a], 3.17)	0.22 (0[a], 2.98)	2.43 (0[a], 6.22)	4.93 (0.92, 8.94)
Grade 1	7.64 (5.78, 9.49)	7.49 (5.64, 9.34)	11.32 (8.18, 14.46)	14.22 (10.81, 17.63)
Grade 2	29.83 (28.53, 31.13)	29.79 (28.49, 31.09)	29.56 (26.70, 32.41)	30.90 (27.76, 34.05)
Grade 3	69.24 (67.82, 70.67)	69.42 (68.00, 70.84)	63.92 (60.92, 66.93)	60.89 (57.58, 64.20)

Raw data: unadjusted; model 1: adjusted for multiple raters; model 2: Clustered by patients, adjusted for multiple raters; model 3: clustered by joints and patients, adjusted for multiple raters.
[a]Negative values generated by the model were replaced by 0.

Table 3 Sensitivity and specificity of predefined quantitative scoring cut-off values in different statistical models

Cut-offs between grades[a]	Sensitivity and specificity in the different statistical models							
	Raw data		Model 1		Model 2		Model 3	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Grades 0–1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Grades 1–2	0.94	0.72	0.99	0.85	0.99	0.82	0.96	0.77
Grades 2–3	0.90	0.84	0.99	0.92	0.99	0.91	0.98	0.92

Raw data: unadjusted; model 1: adjusted for multiple raters; model 2: Clustered by patients, adjusted for multiple raters; model 3: clustered by joints and patients, adjusted for multiple raters.
[a]Cut-off between grade 0 (G0) and G1: 0%; between G1 and G2: 10%; between G2 and G3: 50%.

disagreement on what should be considered the outline of the ROI (40.7%), and the amount of Doppler signal being very close to one of the cut-offs between SQS grades with participants being unsure of which SQS grade should be attributed (32.3%).

Discussion

We aimed to investigate whether two semi-quantitative colour Doppler assessment methods were associated with a fully quantitative colour Doppler assessment of inflammatory joints in RA patients, and to understand the reasons for potential discrepancies between these methods.

In line with the study by Terslev *et al.* [21], we observed a strong correlation between SQS and SQ assessments. Further, we confirmed this association using multilevel models that were able to take into account the clustering of ratings within each participant, patient and joint. This is crucial as disease activity is highly correlated within the different positions used to assess each joint, as well as between the different joints of a given patient.

We found that both the correlations and strength of the associations between the QS and either SQS systems were very similar. This is important as there is still a lack of consensus regarding the most appropriate semi-quantitative scoring system that should be preferred at the joint level [20].

In order to further understand the limitations of SQS assessments, we performed an in-depth review of all images for which we observed a discrepancy between the SQS given by the participants and that calculated from the QS, using pre-specified cut-off values. Even if it was not the aim of this study to define new cut-off values between the different SQS grades, we found that these pre-specified QS cut-off values had a very good sensitivity and specificity to distinguish between the different SQS grades, across all models (Table 3). In addition we were able to confirm that the cut-off value (50%) between grades 2 and 3 as suggested by Szkudlarek *et al.* [18] has a very good sensitivity and specificity (Table 3) when applied to our study data.

Using these cut-off values, we found that SQS assessments were either under- or over-estimated by at least one of the participants in about 20% of the joint ratings, when compared with SQS derived from QS measurements. The main reason leading to discrepancies was that the participants did not agree with the ROI used for the initial QS measurement, despite the fact that these ROI were defined in accordance with the OMERACT definition of synovitis [25]. This might underline the need for further consensus on how to define the ROI at the joint

Table 4 Causes of discrepancies

Causes of discrepancy	*n* (%)
Different perceived ROI (The rater does not agree with the ROI used for the initial Doppler % measurement)	53 (40.7)
Borderline Doppler % very close to a cut-off (After review of the images, the rater report that he/she has misclassified the grade because the actual Doppler % was very close to 10% or 50%)	42 (32.3)
Conflict between grading according to visual % and signals (Discrepancy between the theoretical grading related to the presence of inflammatory signals and the Doppler pixel count visually perceived)	19 (14.6)
Error in measurement or transcription of initial Doppler pixel count (A new measure of the Doppler pixel count shows that the value initially entered in the database was not correct)	13 (10.0)
Paratenonitis (The rater does not agree as how to handle a paratenonitis)	10 (7.7)
Blood vessel initially misinterpreted as an inflammatory signal (After review of the image, the rater agrees that he/she has initially misinterpreted a small blood vessel as an inflammatory signal)	6 (4.6)
Artefacts (Presence of artefacts in the image, making the rating more difficult)	5 (3.8)
Capsule asymmetry (Asymmetry in the capsule, leading to a mistake in the grading)	2 (1.5)

Analysis of the cause of the discrepancies between actual *vs* theoretical grading based on Doppler pixel count. The participants were able to provide a valuable explanation for 130 (54.4%) of the 239 discrepant cases (no explanation could be provided for 46 discrepant cases and one author did not provide explanations for the discrepancies).

level, especially in the wrist which is the joint in which we observed the highest proportion of discrepancies. Also, our data show that it remains to be agreed, ideally by consensus, upon how paratenonitis [26] should be considered when one is tracing the ROI. On the other hand, it is worthy to underline the presence of normal vessels in the joint capsule, particularly visible in the wrist. These vessels can be interpreted as pathological, and belonging to the synovial membrane in both static images and film clips. The real position of those vessels can be appreciated only during a live examination.

The second reason for misclassification was the fact that the amount of Doppler signals was in a borderline zone, that is, between two consecutive scores in the SQS systems. This is an important finding, showing that both SQS systems used in this study did not perform well when the synovial colour fraction was around 10 and 50%, therefore leading to either an over- or under-estimation of the actual disease activity. This may be seen as an important limitation of the SQS methods, and conversely as a strength of the QS method as the score in the QS are based on the pixels calculated with dedicated software [27], in contrast to the SQS where it is made in the mind of the examiner and therefore hardly as accurate.

As previously shown [20], the two main factors that may influence the sensitivity of synovitis detection by Doppler US in RA, and consequently the grade of intra-articular Doppler signals are: equipment-dependent factors and operator-dependent factors. In this study, a single MSUS operator (H.R.) assessed all patients with the same MSUS settings, on the same US equipment. This, along with the relatively large total number of images assessed, the high agreement rate between five highly experienced MSUS raters, and the use of multilevel models that were able to take into account the highly correlated structure of the dataset, can be considered important strengths of this study. Despite very high test-retest reproducibility, the fact that the QS was performed by a single participant and then used as a gold standard may paradoxically limit the generalizability of our findings. Another limitation is that valuable explanations could be provided for only 54% of the discrepant cases. As there is no definition for semi-quantitative scoring of the volar aspect of the MCP joints in the both SQS systems [15, 18] used in this study, we could not include this aspect for SQS or colour Doppler pixel count.

Altogether, we have shown that two semi-quantitative methods for assessing synovitis correlated similarly well with a fully quantitative Doppler assessment of inflammatory joints in RA patients. This suggests that any of these methods can be used to assess synovial activity in RA. However, the QS system, due to its scale structure, is more sensitive to change than SQS systems, which is an important metrological property to follow-up disease activity [28, 29]. Furthermore, some important unanswered questions regarding the global quantification of synovitis by Doppler technique remain, including the minimal activity that should be considered pathological in a single joint [20], and consensus on how to trace the ROI for each joint area, particularly in the wrist [30].

Acknowledgements

We would like to thank all the participating patients, rheumatologists and nurses who made this study possible. Author contributions: H.R., E.K., H.B.H., M.A.D.A. and L.A. contributed to the conception and design of this study. H.R. and E.K. designed the musculoskeletal US performance. H.R. performed all the musculoskeletal US scanning and quantitative colour Doppler pixel count. All authors except L.A. performed semi quantitative scoring. L.A. and H.R. performed the statistical analysis and interpreted the data together with other authors. Classifying of discrepant images and provided explanation was performed by H.R., L.A., E.K. and Y.K. H.R. and L.A. drafted the article and all co-authors revised it critically for important intellectual content and approved the final version.

Funding: No specific funding was received from any bodies in the public, commercial or not-for-profit sectors to carry out the work described in this manuscript.

Disclosure statement: H.B.H. has received fees for speaking and/or consulting from AbbVie, BMS, Pfizer, UCB, Roche, MSD and Novartis. L.T. has received speakers' honoraria from UCB, Roche, AbbVie, Pfizer, General Electric, Janssen, MSD. R.H. has received lecture honoraria from General Electric and AbbVie. E.a.K. has received lecture honoraria from AbbVie. All other authors have declared no conflicts of interest.

References

1 Aletaha D, Smolen JS. The definition and measurement of disease modification in inflammatory rheumatic diseases. Rheum Dis Clin North Am 2006;32:9–44.

2 Aletaha D, Smolen JS. The Simplified Disease Activity Index (SDAI) and Clinical Disease Activity Index (CDAI) to monitor patients in standard clinical care. Best Pract Res Clin Rheumatol 2007;21:663–75.

3 Prevoo ML, van 't Hof MA, Kuper HH *et al*. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44–8.

4 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.

5 Brown AK, Quinn MA, Karim Z *et al*. Presence of significant synovitis in rheumatoid arthritis patients with disease-modifying antirheumatic drug-induced clinical remission: evidence from an imaging study may explain structural progression. Arthritis Rheum 2006;54:3761–73.

6 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.

7 Taylor PC, Steuer A, Gruber J *et al*. Comparison of ultrasonographic assessment of synovitis and joint vascularity with radiographic evaluation in a randomized, placebo-controlled study of infliximab therapy in early rheumatoid arthritis. Arthritis Rheum 2004;50:1107–16.

8 Kane D, Balint PV, Sturrock RD. Ultrasonography is superior to clinical examination in the detection and localization of knee joint effusion in rheumatoid arthritis. J Rheumatol 2003;30:966–71.

9 Karim Z, Wakefield RJ, Conaghan PG *et al*. The impact of ultrasonography on diagnosis and management of patients with musculoskeletal conditions. Arthritis Rheum 2001;44:2932–3.

10 Matsos M, Harish S, Zia P *et al*. Ultrasound of the hands and feet for rheumatological disorders: influence on clinical diagnostic confidence and patient management. Skeletal Radiol 2009;38:1049–54.

11 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.

12 Rezaei H, Torp-Pedersen S, Af Klint E *et al*. Diagnostic utility of musculoskeletal ultrasound in patients with suspected arthritis inverted question mark a probabilistic approach. Arthritis Res Ther 2014;16:448.

13 Szkudlarek M, Klarlund M, Narvestad E *et al*. Ultrasonography of the metacarpophalangeal and proximal interphalangeal joints in rheumatoid arthritis: a comparison with magnetic resonance imaging, conventional radiography and clinical examination. Arthritis Res Ther 2006;8:R52.

14 Szkudlarek M, Narvestad E, Klarlund M *et al*. Ultrasonography of the metatarsophalangeal joints in rheumatoid arthritis: comparison with magnetic resonance imaging, conventional radiography, and clinical examination. Arthritis Rheum 2004;50:2103–12.

15 Hammer HB, Bolton-King P, Bakkeheim V *et al*. Examination of intra and interrater reliability with a new ultrasonographic reference atlas for scoring of synovitis in patients with rheumatoid arthritis. Ann Rheum Dis 2011;70:1995–8.

16 Ohrndorf S, Backhaus M. Advances in sonographic scoring of rheumatoid arthritis. Ann Rheum Dis 2013;72(Suppl 2):ii69–75.

17 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.

18 Szkudlarek M, Court-Payen M, Jacobsen S *et al*. Interobserver agreement in ultrasonography of the finger and toe joints in rheumatoid arthritis. Arthritis Rheum 2003;48:955–62.

19 Qvistgaard E, Rogind H, Torp-Pedersen S *et al*. Quantitative ultrasonography in rheumatoid arthritis: evaluation of inflammation by Doppler technique. Ann Rheum Dis 2001;60:690–3.

20 Mandl P, Naredo E, Wakefield RJ, Conaghan PG, D'Agostino MA. 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.

21 Terslev L, Ellegaard K, Christensen R *et al*. Head-to-head comparison of quantitative and semi-quantitative ultrasound scoring systems for rheumatoid arthritis: reliability, agreement and construct validity. Rheumatology 2012;51:2034–8.

22 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.

23 Aletaha D, Neogi T, Silman AJ *et al*. Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62:2569–81.

24 Torp-Pedersen ST, Terslev L. Settings and artefacts relevant in colour/power Doppler ultrasound in rheumatology. Ann Rheum Dis 2008;67:143–9.

25 Wakefield RJ, Balint PV, Szkudlarek M *et al*. Musculoskeletal ultrasound including definitions for ultrasonographic pathology. J Rheumatol 2005;32:2485–7.

26 Gutierrez M, Filippucci E, Salaffi F, Di Geso L, Grassi W. Differential diagnosis between rheumatoid arthritis and psoriatic arthritis: the value of ultrasound findings at metacarpophalangeal joints level. Ann Rheum Dis 2011;70:1111–4.

27 Schmidt WA, Schonau V, Reiche BE *et al*. Grading of ultrasound Doppler signals in synovitis: does it need an update? Rheumatology 2015;54:1897–903.

28 Ellegaard K, Christensen R, Torp-Pedersen S *et al*. Ultrasound Doppler measurements predict success of treatment with anti-TNF-α drug in patients with rheumatoid arthritis: a prospective cohort study. Rheumatology 2011;50:506–12.

29 Terslev L, Torp-Pedersen S, Qvistgaard E, Danneskiold-Samsoe B, Bliddal H. Estimation of inflammation by Doppler ultrasound: quantitative changes after intra-articular treatment in rheumatoid arthritis. Ann Rheum Dis 2003;62:1049–53.

30 Ellegaard K, Torp-Pedersen S, Lund H *et al*. Quantification of colour Doppler activity in the wrist in patients with rheumatoid arthritis–the reliability of different methods for image selection and evaluation. Ultraschall Med 2008;29:393–8.