Concise report

Head-to-head comparison of quantitative and semi-quantitative ultrasound scoring systems for rheumatoid arthritis: reliability, agreement and construct validity

Lene Terslev1,2, Karen Ellegaard1, Robin Christensen1, Marcin Szkudlarek3, Wolfgang A. Schmidt4, Peter S. Jensen1, Henning Bliddal1 and Søren Torp-Pedersen1,5

Abstract

Objective. To evaluate the reliability and agreement of semi-quantitative scoring (SQS) and quantitative scoring (QS) systems. To compare the two types of scoring system and investigate the construct validity for both scoring systems.

Methods. A total of 46 RA patients (median disease duration of 6.5 years) were enrolled in the study. They were investigated with colour Doppler ultrasound using the central position of the wrist. Disease activity score based on 28 joints (DAS-28) was determined for all patients using CRP. Two participants trained in the SQS system and two in the QS system evaluated the 46 anonymized images. All images were scored twice by each of the two assessors in order to assess both intra- and inter-reader reliability.

Results. The reliability for the two systems were 0.964 for the QS, and 0.817 for the SQS, with a comparable inter-reader agreement for both scoring systems; 95% limits of agreement for the QS being between −7.7% and +6.7% on the colour fraction scale (0–100%), whereas SQS was between −0.8 and +0.8 on the ordinal scale from 0 to 3. There was a direct but non-linear relationship between the two modalities (Spearman’s \(r\) = 0.73) and critical conceptual issues in the agreement between the scoring systems were revealed. The construct validity was poor for both systems with only a weak correlation to CRP.

Conclusion. High reliability and good agreement of both scoring systems were found when applied to the same patient cohort. Different scoring systems appear to be highly correlated.

Key words: Doppler, scoring systems, rheumatoid arthritis, construct validity, agreement.

Introduction

One of the key features of RA is synovitis characterized by synovial hypertrophy and hyperaemia. Though Doppler activity is not mandatory to the OMERACT definition of synovial hypertrophy [1], most scoring systems deal with the presence or absence of Doppler as a measure of synovial hyperaemia. Doppler US is used increasingly to monitor treatment efficacy and failure in patients with rheumatic diseases by scoring the changes over time. Different types of scoring system have been proposed; those of interest are the semi-quantitative scoring (SQS)
systems and the quantitative scoring (QS) systems. They all grade the amount of colour pixels in the region of interest (ROI), which is the synovial hypertrophy.

Several SQS systems have been presented, but none has been chosen by consensus [2]. They all grade the amount of Doppler using a Likert scale with four response categories from 0 to 3. Only the scoring systems that describe the different grades in detail (how much of the synovium is covered in colour pixels in the different grades [3, 4]) ensure a high degree of reproducibility and exportability. The QS systems assess the colour pixels in the synovium either in the longitudinal plane alone [5] or as a sum for both the longitudinal and transverse plane for each joint [6, 7]. They all have in common that the obtained images need post-processing on a computer. The quantification takes more time, but has the potential to be more objective and—it is anticipated—more sensitive to change than a four-point ordinal scale.

Though both types of scoring system are used extensively, little knowledge exists as to the degree to which these scoring systems are directly associated with each other (i.e. agreement) and with the clinical status in the same group of patients (i.e. construct validity). A direct association between the two types of scoring system will allow for optimal use of research data for implementation in clinical practice. The most frequently used scoring systems are the SQS system by Szkuhdarek et al. [4] and the QS system suggested by Qvistgaard et al. [5]. Since both scoring systems operate with colour fractions (percentage of the ROI covered by coloured pixels), a direct comparison is possible.

The aim of this cross-sectional study in RA patients was to (i) evaluate the reliability according to the OMERACT filter [8] and agreement of the SQS [4] and the QS systems [5]; (ii) compare the two types of scoring system with each other; and (iii) investigate the construct validity for both the scoring systems.

Materials and methods

Patients and clinical outcome measures

A cohort of patients fulfilling the ACR 1987 criteria for RA [9] was investigated before commencing treatment with an anti-TNF drug (adalimumab, etanercept or infliximab) in the period 2002–06. Included in this retrospective study were only patients with disease activity in the wrist joint with the presence of both synovial hypertrophy and colour Doppler activity. For each patient one anonymized image, from the most active wrist joint, was obtained and distributed to the investigators on a CD, in randomized order. DAS based on a 28-joint count (DAS-28) using CRP was obtained for all patients.

US assessment

Scanning was performed with a Siemens Acuson Sequoia (Mountain View, CA, USA) using a linear array transducer with 14 MHz centre frequency. Colour Doppler was chosen over power Doppler, as it was more sensitive on that machine. The same colour Doppler pre-set was used for all examinations and no re-adjustment of Doppler parameters was performed. The machine settings were adjusted as recommended [10]. The patient was placed opposite the investigator with the hand in a prone and relaxed position.

The dorsal wrist was scanned longitudinally and care was taken not to exert pressure with the transducer during the examination. For the purpose of this study’s objective, only the images from the central position were used (i.e. primary outcome for the US scoring) as the probe position here is found to be the most reproducible position with little variation in acquisition in our experience.

Both the radiocarpal joint and the intercarpal joint were included and evaluated as one joint. The scan plane in the central position with the most colour Doppler activity was identified and saved. This image acquisition technique has previously shown an intra-class correlation coefficient (ICC) value of 0.77 when assessing the test–retest reliability [11], suggesting excellent (ICC > 0.75) reliability [12].

Image evaluation

Two of the participants (M.S. and W.A.S.) trained in the SQS system performed the scoring according to the Szkudlarek scoring system [4] and two participants (K.E. and P.S.J.) trained in the QS system performed the scoring according to Qvistgaard et al. [5]. All images were scored twice to assess intra-reader reliability.

Statistical analysis

For each US modality, we calculated the single-measure ICC using a three-way random effects model (with a factor for patient, assessor and test). Agreement between the US assessors for each of the methods was evaluated as recommended by Bland and Altman [13], by constructing the 95% limits of agreement. The agreement between the two scores was done informally by plotting the empirical estimates reported as the median value, with the measure of dispersion being the 5th and 95th percentiles from the empirical data distribution (capturing 90% of all the data) using all the repeated measures. To assess the differences in the two methods’ construct validity, the two scoring systems were correlated using Spearman’s correlation coefficient (r) to DAS-28 (using CRP), patient global and CRP at baseline. It was decided a priori that the US measure with the greatest correlation with these established constructs would have the greatest construct validity per se.

Results

Patient characteristics

The study sample included 46 RA patients (80% females) with Doppler activity in at least one wrist joint, with a mean (s.d.) age of 57 (14) years. The median disease duration was 6.4 years (range 0.1–34.2 years), with an average DAS-28 of 5.1 (range 1.4–7.8). The patients had a median number of swollen and tender joints of 8 and 11 (28 joints assessed), respectively. Their median global VAS was 70 mm (range 6–100); the median CRP level was 12.5 mg/l (range 1.0–90.0), and the median HAQ was 1.3 (range 0.0–2.6).
The reliability and agreement for the individual scoring systems

Overall these patients had a median (minimum–maximum) Doppler activity of 19.5% (0.0–70.4%) according to the QS system and 2% (0–3) by the SQS system on the ordinal scale from 0 to 3. According to the reliability measures, both US methods had excellent reliability. The ICC for the QS was higher (ICC = 0.964) than the observed reliability for the SQS (ICC = 0.817). The inter-rater agreement was also comparable for both scoring systems, with 95% limits of agreement for the QS being between −7.7% and +6.7% on the colour fraction scale (from 0% to 100%), whereas SQS was between −0.8 and +0.8 on the ordinal scale from 0 to 3. Thus the true colour fraction (QS) may reasonably well be between 7.7% below and 6.7% above the average approximation, whereas the SQS implies that the truth would be between 0.8 below and 0.8 above the average approximation.

Association between the two scoring systems and construct validity

Fig. 1 illustrates the association between the median score for the two systems, as well as the empirical data distribution across all the repeated measures. As quantified by Spearman’s correlation coefficient ($r = 0.73$ (95% CI 0.49, 0.96; $n = 46$)), it is reasonable to assume a substantial relationship between the two modalities. There was a non-linear association between the measures, with 90% of the empirical data dispersion suggesting a pragmatic association (with 90% certainty). For example, 0 on the SQS would also equal 0 on the QS; a score of 1 on the SQS corresponds to a QS between 0% and 9%; an SQS of 2 was between 4% and 45% on the QS; the highest category of the SQS (category 3) corresponded to a QS value $>34\%$ per se—implying critical issues in the agreement between measures when colour fraction is between 4% and 45%. In order to assess the construct validity of these US scores, we tested whether they were associated with the clinical findings. The construct validity of US measurements is presented in Table 1. The QS and SQS both had a weak correlation with CRP ($P = 0.042$ and 0.037, respectively). There was no correlation with DAS-28 for either of the scoring systems nor any correlation with the other single components of the DAS-28 apart from CRP.
Both types of scoring system have separately been validated in treatment trials [3, 6, 14–17] and validated with MRI as a reference method [18–20]. However, no information exists on the reliability and association of two well-defined scoring systems applied to the same patient population. Since the Szkudlarek score [4] and the Qvistgaard score [5] operate with colour fractions, a direct comparison is possible.

How reliable are the scoring systems?

One of the most important issues for the application of a scoring system is the reliability. In the present study, only frequent users of the two scoring systems participated. We found excellent reliability for both types of scoring system applied on exactly the same cohort of patients—based on repeated use of the same image for each of the 46 patients—which is in agreement with the previous findings [4, 5, 21, 22]. The reliability was, however, better for the QS than the SQS, possibly explained by the more objective method implied in the QS approach (the pixels are calculated on a computer with dedicated software, in contrast to the SQS where it is made in the mind of the examiner and therefore hardly as accurate [4]).

How is the association between the scoring systems when applied to the same population of RA patients?

The two scoring systems had a good association ($r = 0.73$) indicating a high degree of comparability on images sampled in a cross-sectional study. In a recent study evaluating different scoring systems (quantitative and semi-quantitative), during treatment all were sensitive to change, but no direct correlation between them was established [21]. The QS system gives the true value for the SQS system’s grades 2 and 3 and could therefore be seen as the gold standard for the SQS system. From the data shown in Fig. 1, it is clearly demonstrated that the difficulty for the SQ score is to differentiate between grades 2 and 3 when the Doppler activity is close to 50% on either side. Even scores as low as 34% may be graded as 3 (statistically seen—not observed), which may be somewhat problematic when applying the SQS system for longitudinal follow-up. Though the two evaluated scoring systems show comparability in a cross-sectional study, a comparison of their sensitivity to change still needs to be performed on the same cohort of patients.

How are the scoring systems associated with clinical and laboratory evaluation (construct validity)?

The construct validity of the scoring systems seems to vary with the size and number of joints evaluated, as well as positions evaluated in the joint [2]. In the present study we found that both scoring systems only correlate with the CRP ($P = 0.042$ and 0.037, respectively), but not with DAS-28 or any other single component. This is probably because only a single position in a single joint was evaluated. In a study applying three positions in the wrist and in studies with multi-joint assessments, a correlation with DAS-28 has been demonstrated [23, 24].

Conclusion

High reliability and good agreement between the scoring systems indicate that findings in studies using different scoring systems reflect the same disease entity, although sensitivity to change, correlation with clinical improvement and minimum important difference need to be established. The construct validity of the scoring systems was poor, with only a weak correlation with CRP independent of the scoring system.

Rheumatology key message

- There is high agreement and reliability between two different scoring systems for RA.

Funding: This study was supported by the Oak Foundation.

Disclosure statement: The authors have declared no conflicts of interest.
References