RHEUMATOLOGY

Rheumatology 2022;61:SI65–SI72
doi:10.1093/rheumatology/keab556
Advance Access publication 15 July 2021

Original article

Real-time *vs* static scoring in musculoskeletal ultrasonography in patients with inflammatory hand osteoarthritis

Lotte A. van de Stadt [1], Féline P. B. Kroon[1,2], Frits R. Rosendaal[3], Desirée van der Heijde[1], Monique Reijnierse[4], Naghmeh Riyazi[5], Ragnhild de Slegte[6], Jendé van Zeben[7], Cornelia F. Allaart[1], Margreet Kloppenburg[1,3] and Marion C. Kortekaas[1,8]

Abstract

Objectives. Agreement between real-time and static ultrasonography has not been studied in musculoskeletal diseases. We studied this agreement in inflammatory hand OA.

Methods. Ultrasonography was performed blinded to clinical information of 30 joints of 75 patients with hand OA, treated with prednisolone in a randomized placebo-controlled double-blind trial. Images were scored real-time at acquisition and stored images were scored static (paired in known chronological order) for inflammatory features and osteophytes (score 0–3). Agreement between methods was studied at joint level with quadratic weighted kappa. At patient level intra-class correlations (ICC) of sum scores and change in sum-scores (delta baseline—week 6) were calculated. Responsiveness of scoring methods was analysed with generalized estimating equations (GEE) with treatment as independent and ultrasonography findings as dependent variable.

Results. Agreement at baseline was good to excellent at joint level (kappa 0.72–0.88) and moderate to excellent at patient level (ICC 0.58–0.91). Agreement for change in sum scores was poor to fair for synovial thickening and effusion (ICC 0.18 and 0.34, respectively), while excellent for Doppler signal (ICC 0.80). Real-time ultrasonography discriminated between prednisolone and placebo with a mean between-group difference of synovial thickening of −2.5 (95% CI: −4.7, −0.3). Static ultrasonography did not show a decrease in synovial thickening.

Conclusion. While cross-sectional agreement between real-time and static ultrasonography is good, static ultrasonography measurement of synovial thickening did not show responsiveness to prednisone therapy while real-time ultrasonography did. Therefore, when ultrasonography is used in clinical trials, real-time dynamic scoring should remain the standard for now.

Key words: ultrasonography, reliability, static, inflammatory rheumatic diseases, hand OA

CLINICAL SCIENCE

[1]Department of Rheumatology, Leiden University Medical Center, Leiden, [2]Department of Rheumatology, Zuyderland Medical Center, Heerlen, [3]Department of Clinical Epidemiology, Leiden University Medical Center, [4]Department of Radiology, Leiden University Medical Center, Leiden, [5]Department of Rheumatology, Haga Hospital, The Hague, [6]Department of Radiology, Reade, Amsterdam, [7]Department of Rheumatology, Sint Franciscus Vlietland Groep, Rotterdam and [8]Department of Rheumatology, Flevoziekenhuis, Almere, Netherlands

Submitted 18 February 2021; accepted 7 July 2021

Correspondence to: Lotte A. van de Stadt, Leids Universitair Medisch Centrum, Stafcentrum Reumatologie, C1-R, Postbus 9600, 2300RC Leiden, Netherlands. E-mail: lavdstadt@lumc.nl

Introduction

In the past few decades, ultrasonography has undergone a dramatic evolution, acquiring a role in diagnosis and assessment of disease activity in clinical practice in multiple rheumatic musculoskeletal diseases. In hand OA, the role of inflammation has long been underestimated. However, high resolution imaging modalities including ultrasonography changed this view [1–3]. The advantage of ultrasonography over other imaging modalities is that it is inexpensive, non-invasive and without contraindications,

Rheumatology key messages

- Cross-sectional agreement between real-time and static ultrasonography scoring in hand osteoarthritis is good.
- Responsiveness to prednisone therapy was shown with real-time scoring but not with static scoring.
- Real-time scoring of ultrasound images should remain the standard for clinical hand osteoarthritis trials.

and it allows the investigation of multiple features of both soft and bony tissue during the same assessment in a timely manner. It is useful to investigate inflammatory features in large as well as small joints and has been found in many studies to be both reliable and valid in the hands of experienced ultrasonographists [4–6]. Moreover, we showed that ultrasonography was responsive to treatment and could therefore be of use in clinical trials [7].

However, ultrasonography is investigator dependent, which may lead to issues with inter-rater reliability, which is variable depending on the modality and investigated disease but ranges from moderate to excellent [8–11]. Good inter-rater reliability is particularly important for the use of ultrasound outcome measures in multicentre clinical trials in which suboptimal inter-rater reliability is a concern. It has been proposed that a central experienced and dedicated rater, who reads and scores all stored static ultrasonography images of a study afterwards ('static scoring' as opposed to 'real-time scoring': the usual real-time, dynamic scoring by the ultrasonographist at the moment of the ultrasound examination at every site independently), could potentially avoid inter-rater reliability issues as well as differences in experience between raters. This way, variability caused by multiple raters, could be minimized and this could enhance the value and applicability of the use of ultrasonography in multicentre clinical trials. Another advantage of static scoring is the possibility of scoring all time points simultaneously in known chronological order, as opposed to scoring all time points separately and independently. This method has been shown to increase sensitivity for traditional radiographs [12]. Improved sensitivity could improve detection of low-grade inflammation and subtle changes over time. However, agreement between real-time scoring and static scoring has not been investigated in rheumatic musculoskeletal diseases.

In order to investigate the agreement of static and real-time scoring, data of the Hand OA Prednisolone Efficacy (HOPE) study were analysed. The HOPE study is a study in hand OA in which significant improvement of clinical as well as ultrasonography outcome measures in prednisolone treated patients, in comparison to placebo treated patients, occurred [7]. Ultrasound images obtained during the study were stored during assessment, enabling static scoring afterwards, which made this an ideal study to investigate agreement between the two scoring methods.

Methods

Patients

Hand OA patients were recruited for the HOPE study from outpatient clinics at two sites in the Netherlands. Patients enrolled in the Leiden University Medical Center (LUMC) were included for the present analysis [7]. Patients fulfilled the American College of Rheumatology hand OA criteria [13] and showed signs of inflammation ($\geq$1 DIP/PIP with soft swelling or erythema, $\geq$1 PIP/DIP with a positive power Doppler signal or synovial thickening of grade $\geq$2 on ultrasonography and finger pain of $\geq$30 mm on a 100 mm visual analogue scale with a flare-up during a 48 h washout of NSAIDs [paracetamol if NSAIDs were contraindicated]). Visits were scheduled at baseline, week 2, 6 and 14. Patients were randomly assigned to the treatment or placebo group. Patients self-administered 10 mg prednisolone or placebo daily for 6 weeks, after which medication was tapered to cessation in 2 weeks. The study was approved by medical ethics committees at the LUMC and Zuyderland Medical Center, conducted in accordance with Good Clinical Practice guidelines and adhered to the Declaration of Helsinki. All patients provided written informed consent. Patient partners were involved in development and execution of the study and providing information to patients.

For the full study protocol, see the original publication [7].

Ultrasonography

Ultrasound was performed with a GE Logiq E9 (GE Healthcare, Chicago, IL, USA) with a 6–15 MHz linear array transducer at baseline, week 6 and week 14. The DIP/PIP 2–5, IP-1, MCP 1–5 and CMC-1 joints of both hands were assessed for synovial thickening, effusion, Doppler signal and osteophytes (all scores ranged from 0 to 3 per joint, 30 joints per patient, joints with prothesis or arthodesis excluded) [14]. Synovial thickening was defined as abnormal hypoechoic (relative to subdermal fat, but sometimes may be isoechoic or hyperechoic) intra-articular tissue that is non-displaceable and poorly compressible [15]. Doppler signal was defined as a colour Doppler signal within a region of grey scale synovitis. The Doppler settings were optimized with the help of the application specialist of the manufacturer. The settings were saved after optimization and kept the same throughout the entire study for all patients. Hand joints were scanned on the dorsal side in the longitudinal plane with the hand lying flat. The transverse plane was used to confirm findings where necessary [15].

Ultrasonography during the study ('real-time scoring') was performed by a trained ultrasonographer (F.B.P.K.) and a senior ultrasonographer (M.C.K.) with more than 10 years of experience in hand OA ultrasound research. The two ultrasonographers scored the live images together (simultaneously) in real-time with use of a scoring

atlas (ISBN/EAN 978-90-827311-0-1) [16]. A consensus score for each joint was recorded. Representative images (at least two per joint) were stored during scanning for later use. For the 'static scoring', the saved static images were scored by one ultrasonographer (M.C.K.) with a minimum of 6 months between real-time and static scoring. For each patient, images of each visit were scored simultaneously in known chronological order. Of 10 randomly selected patients, images were scored twice at separate time points for static ICC. The ultrasonographers were blinded to all patients' data and treatment at all times. During real-time scoring, access to previous scoring data was not allowed. Moreover, patients were instructed not to reveal any information concerning clinical experience or signs. Although the ultrasonographers were not blinded for visit number, recollection of earlier scores was thought to be highly unlikely, due to the lack of clinical information as well as the fact that in between visits, different ultrasound exams of the hands of different OA patients were performed by the two ultrasonographers. Physical examination was performed by trained independent research nurses.

Ultrasonography sum scores were calculated as total sum of scores of 30 scanned joints.

Statistical analysis

All statistical analyses were performed using STATA version 16 (StataCorp LLP, College Station, TX, USA). For reliability of static *vs* real-time scoring at joint level, agreement between these scores was calculated by quadratic weighted kappa-statistics for ordinal data. As sensitivity analysis, intra-class correlation taking into account clustering of joints within a patient (mixed effect model, absolute agreement) were performed. At patient level, agreement between different of sum scores and change in sum scores (delta baseline − week 6) was measured with ICC (mixed effect model, absolute agreement). Values were interpreted as follows: 0–0.20: poor agreement; 0.21– 0.40: fair agreement; 0.41–0.60: moderate agreement; 0.61–0.80: good agreement; 0.81–1.00: excellent agreement.

Responsiveness of ultrasonography findings after treatment was analysed with generalized estimating equations on patient level and on joint level, with robust standard errors and the working correlation structure specified as exchangeable. Data from baseline and 6 weeks were used. The independent variables were treatment group, visit number (categorical), interaction between treatment group and visit number and the baseline value of the dependent variable (continuous).

Results

Patient characteristics

Of 92 hand OA patients included in the HOPE trial, 75 were enrolled in the LUMC where static ultrasonography images were available for analysis. Patients were a mean of 63.3 (91) years old, 79% female, with a mean body mass index of 28 (4.8) kg/m^2. They had a median disease duration of 4.0 (Inter Quartile Range (IQR) 0.2–9.3) years and 69% suffer from erosive disease. Mean KL sum score was 37 (16) and mean visual analogue scale finger pain 52 (20) mm. The patients of LUMC did not differ in characteristics from patients enrolled in other centres, except for visual analogue scale pain fingers, which was lower in patients enrolled in the LUMC [mean (S.D.): 52 (20) *vs* 64 (18)] (other data not shown). Of patients enrolled in the LUMC, 39 patients (52%) were treated with prednisolone and 36 (48%) were treated with placebo. Patient characteristics were well-balanced between treatment groups (Supplementary Table S1, available at *Rheumatology* online).

Ultrasound

All patients had signs of synovial thickening and osteophytes as assessed by real-time ultrasonography, and almost all had signs of effusion (99%) or a positive Doppler signal (95%) in at least one joint. Median number of affected joints (score >0) was low for Doppler signal (median 2 out of 30 joints), but higher for effusion, synovial thickening and osteophytes (median 8, 12 and 21 out of 30 joints, respectively). Total ultrasonography sum score for osteophytes was high [mean 45 (12)], whereas sum score was low for positive Doppler signal [mean 5.9 (4.4)], with sum scores for synovial thickening and effusion in between [mean 16 (6.3) and 11 (6.0) respectively]. These differences were also reflected by a high percentage (72%) of total joints scanned that showed osteophytes *vs* a lower percentage showing synovial thickening, effusion or a positive Doppler signal (Supplementary Table S2, available at *Rheumatology* online).

Reliability

Static vs real-time scoring at baseline on joint level

At joint level on baseline and 6 weeks, real-time and static showed good to excellent agreement for all ultrasonography modalities. Intra-class correlation taking into account clustering of joints within patients showed similar results (Table 1). Intra-reader agreement of static scores was good to excellent [0.73 (95% CI: 0.69, 0.77) for synovial thickening, 0.80 (95% CI: 0.78, 0.83) for effusion, 0.89 (95% CI: 0.85, 0.92) for osteophytes and 0.87 (95% CI: 0.86, 0.88) for Doppler signal].

Static vs real-time scoring at patient level

Sum scores for 30 joints were calculated for analysis at patient level. ICCs were assessed at baseline and after treatment at week 6. Sum scores of real-time and static scores were comparable, with static scores slightly higher than real-time scores (Fig. 1). Sum scores for synovial thickening showed moderate agreement at baseline and 6 weeks, whereas for effusion, osteophytes and Doppler score agreement was good to excellent.

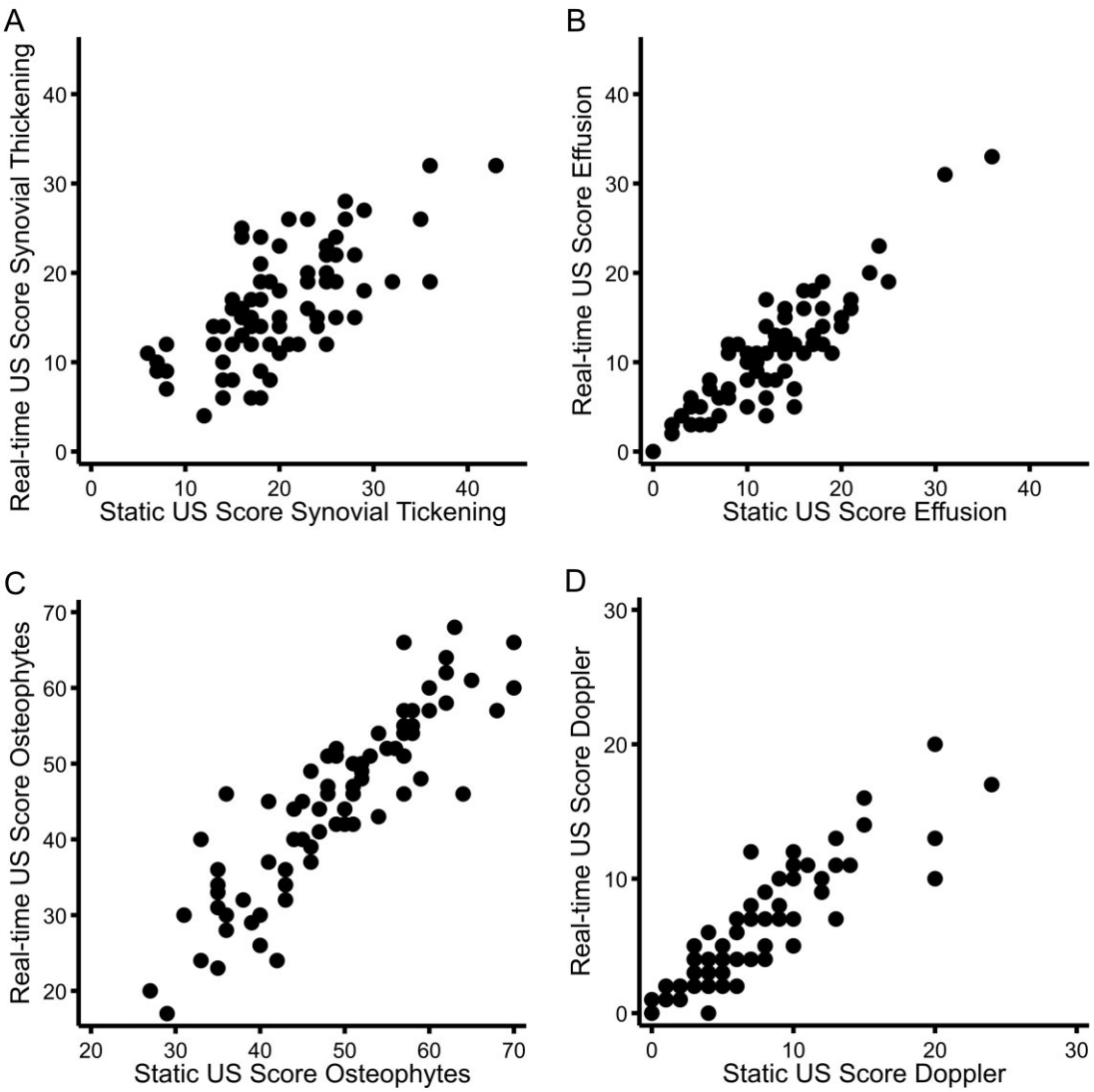
However, after calculation of the change in sum scores between baseline and 6 weeks, agreement

Table 1 Agreement between real-time and static ultrasonography scores at joint level on baseline

Ultrasonography modality	Baseline		Week 6	
	Quadratic weighted kappa (95% CI)	ICC (95% CI)	Quadratic weighted kappa (95% CI)	ICC (95% CI)
Synovial thickening	0.72 (0.69, 0.74)	0.73 (0.69, 0.77)	0.69 (0.66, 0.71)	0.70 (0.16, 0.35)
Effusion	0.78 (0.75, 0.81)	0.80 (0.78, 0.83)	0.77 (0.74, 0.80)	0.78 (0.76, 0.80)
Osteophytes	0.88 (0.87, 0.89)	0.89 (0.85, 0.92)	0.86 (0.92, 0.90)	0.89 (0.84, 0.92)
Doppler signal	0.87 (0.84, 0.90)	0.87 (0.86, 0.88)	0.88 (0.86, 0.89)	0.87 (0.86, 0.88)

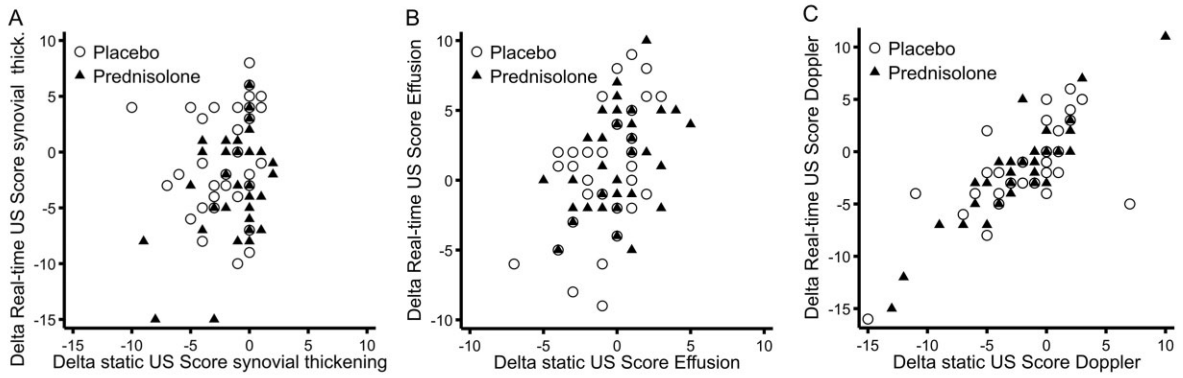
$n = 75$ patients, 30 joints per patient. ICC: intra-class correlation coefficient linear mixed model (random patient, fixed rating), absolute agreement.

Fig. 1 Real-time *vs* static ultrasonography sum scores at baseline



Correlation for real-time scoring (*y*-axes) with static scoring (*x*-axes) at baseline of synovial thickening (**A**), effusion (**B**), osteophytes (**C**) and Doppler signal (**D**). US: ultrasonography.

Fig. 2 Real-time *vs* static ultrasonography delta sum scores



Correlation of real-time scoring (*y*-axes) with static scoring (*x*-axes) of the change in sum scores (delta) between baseline and week 6 score for synovial thickening (**A**), effusion (**B**) and Doppler (**C**). US: ultrasonography.

Table 2 Agreement on patient level

Ultrasonography modality	ICC (95% CI)		
	Baseline	Week 6	Delta week 6 − baseline
Synovial thickening	0.59 (0.26, 0.76)	0.58 (0.24, 0.77)	0.18 (0, 0.40)
Effusion	0.84 (0.66, 0.92)	0.84 (0.75, 0.89)	0.34 (0.12, 0.53)
Osteophytes	0.82 (0.50, 0.92)	0.78 (0.56, 0.88)	ND
Doppler signal	0.86 (0.75, 0.92)	0.91 (0.85, 0.94)	0.80 (0.70, 0.87)

Delta osteophytes: mostly 0, ICC cannot be determined. ICC: intra-class correlation coefficient linear mixed model (random patient, fixed rating), absolute agreement.

dropped to poor for synovial thickening and fair for effusion. Agreement for change in Doppler scores remained excellent (Fig. 2 and Table 2). Change in osteophytes over 6 weeks was not evaluated, since the change was as expected mostly 0.

Responsiveness static ultrasonography scoring on prednisolone treatment

Previously we have shown that synovial thickening decreased by treatment with prednisolone as measured with real-time ultrasonography, while there was no change in the placebo group [7]. The same analysis was repeated for the 75 patients included in the present study and results were similar to the total group of patients analysed in the original paper. Patients treated with prednisolone showed a reduced synovial thickening score [mean between group difference −2.5 (95% CI: −4.7, −0.3) points] as compared with patients treated with placebo who showed stable scores

(Table 3). For the other features no decrease in ultrasonography scores was seen. Analysis at joint level showed similar results (data not shown).

However, when static scoring was used to assess response of synovial thickening on treatment, no difference was seen between treatment groups [mean difference 0.8 (95% CI: −0.4, 1.9)]

(Table 3 and Fig. 3). The other modalities also failed to show a response, but these results were comparable to the results of real-time scoring. Analysis at joint level gave the same results (data not shown).

Discussion

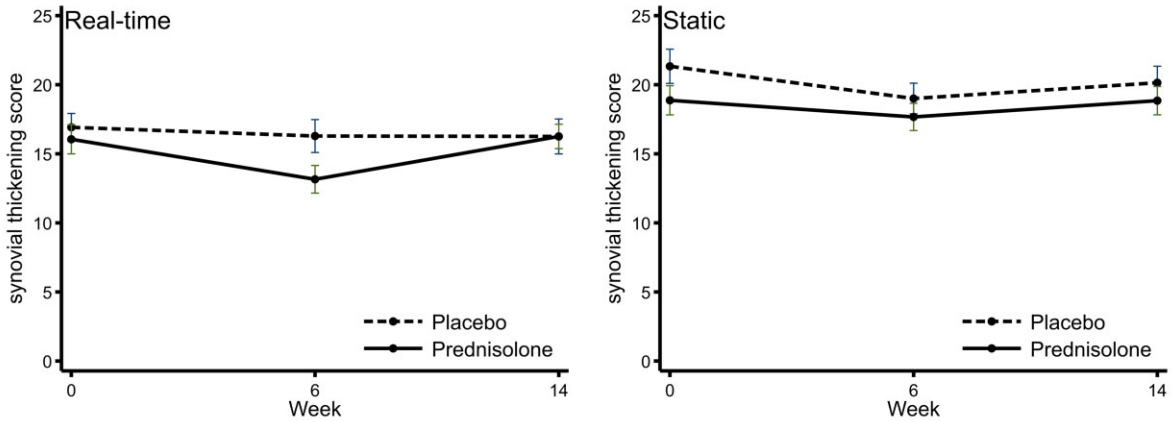
Ultrasonography enables researchers to assess large numbers of small joints in a limited amount of time, making it an ideal tool to assess rheumatic musculoskeletal diseases, such as RA or hand OA, in which multiple small joints are affected. Traditionally, ultrasonography is scored real-time during the assessment. Scoring of static stored images would facilitate using a central reader in multicentre trials. Although the use of static images might result in loss of some information, it is possible to score visits paired and with known chronological order. This method increases sensitivity without loss of specificity in radiographic scoring and could thereby increase accuracy, compensating for loss of information of static scoring. We aimed to assess the agreement of inflammatory ultrasound features as well as responsiveness to

Table 3 Ultrasonography sum scores at baseline and after 6 weeks of treatment with prednisolone or placebo

US modality	US mode	Mean (s.d.) at baseline		Mean change (s.d.) between baseline and week 6			
		Prednisolone ($n = 39$)	Placebo ($n = 36$)	Prednisolone ($n = 39$)	Placebo ($n = 35$)	Adjusted mean between-group difference (95% CI)[a]	*P*-value
Synovial thickening	Real-time	16.0 (6.6)	16.9 (6.0)	−2.9 (4.9)	−0.4 (4.7)	−2.5 (−4.7 to −0.3)	0.02
	Static	18.9 (6.6)	21.3 (7.4)	−1.2 (2.4)	−1.9 (2.5)	0.8 (−0.4, 1.9)	0.19
Effusion	Real-time	11.5 (6.7)	10.1 (5.1)	1.6 (3.6)	0.9 (4.5)	0.8 (−1.1, 2.6)	0.40
	Static	13.2 (7.1)	12.1 (5.8)	0.1 (2.1)	−0.8 (2.2)	0.8 (−0.2, −1.8)	0.11
Doppler signal	Real-time	5.1 (4.2)	6.9 (4.4)	−1.7 (4.5)	−1.3 (4.3)	−0.3 (−2.2, 1.7)	0.78
	Static	6.0 (4.4)	7.9 (5.6)	−2.2 (4.0)	−1.9 (4.4)	−0.2 (−2.1, 1.7)	0.81

[a]Generalized estimating equations adjusted for baseline values. US: ultrasonography.

Fig. 3 Ultrasonography synovial thickening score of patients treated with prednisolone *vs* placebo



Synovial thickening score over time for prednisolone treated patients (solid line) and placebo treated patients (dashed line) as measured by real-time scoring and static scoring.

treatment between a real-time and a static scoring method. For static scoring a method was used that was expected to reflect clinical practice in multicentre trials, i.e. static scoring was performed by one central reader, with visits paired and in known chronological order.

Agreement between real-time and static scoring methods of synovial thickening, effusion, osteophytes and Doppler signal was good to excellent for both scoring methods on joint level as well as for sum scores on patient level at all time points, except for synovial thickening at patient level in which agreement was lower. However, when agreement was assessed for the change in sum scores between baseline and week 6, ICC for Doppler remained excellent, while the ICC dropped considerably for both effusion and synovial thickening. ICC for osteophytes was not analysed, since no change in osteophyte scores was expected during the 6-week time period.

Subsequently, when the responsiveness to treatment was studied for the inflammatory features using both scorings methods, only real-time scoring of synovial thickening demonstrated a response in the prednisolone treated group compared with placebo, while static scoring was unable to show any difference. The reduction in synovial thickening was in line with improvement in other outcome measurements, suggesting the results with real-time scoring are robust. This implies that dynamic real-time scoring is superior to static scoring in assessing subtle differences induced by treatment.

There are a few explanations why real-time scoring could be superior to static scoring as performed in this study. The first major difference between real-time and static scoring is the physical presence of the patient while the scoring takes place. The underlying anatomy or appearance of a joint may, for instance, be of guidance in real-time scoring, a feature lacking in static scoring. Since patients were instructed to refrain from commenting on their improvement, we do not think this was a major issue although subtle body language

cannot be ruled out. Also, in real-time scoring the entire joint can be assessed, while in static scoring only a numbered amount of previously selected and stored images are available for scoring. This might cause loss of information and therefore loss of accuracy. It may also cause a more systematic difference because the images chosen to be stored could be the images with the most extreme pathology, whereas in real-time scoring, the whole joint is taken into account, which may also include more subtle pathology and scores being averaged over the joint subconsciously. Indeed, this could be the explanation for the slightly higher static scores as compared with the real-time scores we found. Scoring stored video-clips instead of static images might resolve some of these issues.

Another possibility is that the method of scoring the static images, with known chronological order, might contribute to loss of accuracy. This scoring method for static images was chosen based on results of earlier studies that compared blinded (single reading or paired without knowledge of the chronological order) and unblinded (paired in known chronological order) scoring of radiographs in RA, which found that unblinded scoring increase sensitivity with higher progression rates, less signal-to-noise ratio and without serious overestimation of non-relevant findings [12, 17, 18]. By using this method in static scoring, we hoped to improve accuracy, but our results were the opposite. This may be partly due to expectation bias, caused by scoring with known chronologic order. In the aforementioned studies concerning blinding or unblinding of scan sequence, progression of structural damage outcomes was investigated. The studies were all performed with 'one-way' pathology, i.e. radiographic progression, in which improvement was extremely rare over the relatively short period of follow-up. Whether scoring with known sequence is also more sensitive without loss of specificity if pathology can change in both directions has not been studied.

A third aspect could be that relevant inflammatory features in hand OA are scarce, as reflected in the low number of joints per patient with pathology and small mean change over time, implying diagnostic accuracy is all the more important. As shown by van Tuyl *et al*. [17], the advantages of scoring with known chronological order are diminished when differences over time are smaller. If this advantage is lost in our study due to small changes over time, loss of accuracy due to another factor (such as the absence of the patients) may gain more weight in the overall outcome.

Furthermore, our results show that agreement between real-time and static scoring is good, but not perfect. Some random noise signals are scored and these could contribute to the difference between static and real-time scoring. Less than perfect agreement is to be expected and might be acceptable when one time point is considered. However, when change of pathology (subtracting scores) is taken into account, differences between scoring methods at both time points could increase the total disagreement over these time points. This may be a cause of the lower agreement for the change in sum scores.

Lastly, the fact that real time scoring was done by two ultrasonographers in consensus, while static scoring for logistical reasons was done by one reader could have been of influence. However, the single reader of static images was one of the two ultrasonographers that scored real-time. Since this reader is very experienced in research in hand OA ultrasound, we believe this effect to be minimal.

Altogether, these factors might explain why in this study, dynamic, real-time scoring of synovial thickening did detect a treatment effect, while static scoring did not. In other fields, such as oncology, real-time scoring has also shown more precision with a higher specificity than static scoring [19].

In measurements over time, the only feature for which agreement remained good was Doppler scoring; for osteophytes no change could be expected. Unfortunately, in this study we could not show responsiveness to treatment for Doppler with real-time scoring. Especially in diseases in which Doppler signal is a prominent feature in active synovitis, such as RA, static scoring of Doppler signal might be considered for further study.

In conclusion, real-time ultrasonography scoring and static scoring of stored images show good to excellent agreement when one time point is assessed. However, agreement of change in sum scores for synovial thickening and effusion between dynamic, real-time scoring and static scoring with known chronological order was reduced to slight–poor. Because real-time scoring of synovial thickening was responsive to therapy with prednisolone, and static scoring as performed in this study was not, dynamic real-time scoring should remain the standard for clinical trials in hand OA.

Acknowledgements

The authors thank all patients for their participation in the HOPE study, and the rheumatologists' inclusion of patients in the HOPE study. We also thank research nurses B.A.M.J. van Schie-Geyer and S. Wongsodihardjo, and technicians J.C. Kwekkeboom and E.I.H. van der Voort for their contributions.

Funding: This research was funded by the Dutch arthritis foundation: grant number HOPE 4-1-303

Disclosure statement: L.A.vd.S., F.P.B.K., F.R.R., N.R., R.d.S., J.vZ., C.F.A. and M.C.K. have declared no conflicts of interest. D.vd.H. reports personal fees from AbbVie, Amgen, Astellas, AstraZeneca, Bayer, BMS, Boehringer Ingelheim, Celgene, Cyxone, Daiichi, Eisai, Eli-Lilly, Galapagos, Gilead, Glaxo-Smith-Kline, Janssen, Merck, Novartis, Pfizer, Regeneron, Roche, Sanofi, Takeda, UCB Pharma, outside the submitted work; and is a Director of Imaging Rheumatology bv. M.K. reports grants from Dutch Arthritis Association, during the conduct of the study; and fees for consultancy from Abbvie, Pfizer, Levicept, GlaxoSmithKline, Merck-Serono, Kiniksa,

Flexion, Galapagos, Jansen, CHDR; and local investigator of industry-driven trial from Abbvie; and from Wolters Kluwer (UptoDate), Springer Verlag (Reumatologie en klinische immunologie); and grants from Pfizer and IMI-APPROACH, outside the submitted work. M.R. reports grants from Dutch Arthritis Foundation, outside the submitted work.

Data availability statement

All data relevant to the study are included in the article or uploaded as supplementary information, available at *Rheumatology* online.

Supplementary data

Supplementary data are available at *Rheumatology* online.

References

1 Kortekaas MC, Kwok WY, Reijnierse M *et al.* Pain in hand osteoarthritis is associated with inflammation: the value of ultrasound. Ann Rheum Dis 2010;69:1367–9.

2 Mancarella L, Addimanda O, Pelotti P *et al.* Ultrasound detected inflammation is associated with the development of new bone erosions in hand osteoarthritis: a longitudinal study over 3.9 years. Osteoarthritis Cartilage 2015;23:1925–32.

3 van Beest S, Damman W, Liu R *et al.* In finger osteoarthritis, change in synovitis is associated with change in pain on a joint-level; a longitudinal magnetic resonance imaging study. Osteoarthritis Cartilage 2019; 27:1048–56.

4 Mathiessen A, Slatkowsky-Christensen B, Kvien TK, Hammer HB, Haugen IK. Ultrasound-detected inflammation predicts radiographic progression in hand osteoarthritis after 5 years. Ann Rheum Dis 2016;75:825–30.

5 Kortekaas MC, Kwok WY, Reijnierse M, Huizinga TW, Kloppenburg M. In erosive hand osteoarthritis more inflammatory signs on ultrasound are found than in the rest of hand osteoarthritis. Ann Rheum Dis 2013;72:930–4.

6 Kortekaas MC, Kwok WY, Reijnierse M, Kloppenburg M. Inflammatory ultrasound features show independent associations with progression of structural damage after over 2 years of follow-up in patients with hand osteoarthritis. Ann Rheum Dis 2015;74:1720–4.

7 Kroon FPB, Kortekaas MC, Boonen A *et al.* Results of a 6-week treatment with 10 mg prednisolone in patients with hand osteoarthritis (HOPE): a double-blind, randomised, placebo-controlled trial. Lancet 2019;394:1993–2001.

8 D'Agostino MA, Terslev L, Aegerter P *et al.* Scoring ultrasound synovitis in rheumatoid arthritis: a EULAR-OMERACT ultrasound taskforce—Part 1: definition and development of a standardised, consensus-based scoring system. RMD Open 2017;3:e000428.

9 Hammer HB, Iagnocco A, Mathiessen A *et al.* Global ultrasound assessment of structural lesions in osteoarthritis: a reliability study by the OMERACT ultrasonography group on scoring cartilage and osteophytes in finger joints. Ann Rheum Dis 2016;75: 402–7.

10 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.

11 Zabotti A, Filippou G, Canzoni M *et al.*; OMERACT Ultrasound Task Force members. OMERACT agreement and reliability study of ultrasonographic elementary lesions in osteoarthritis of the foot. RMD Open 2019;5: e000795.

12 van Der Heijde D, Boonen A, Boers M, Kostense P, van Der Linden S. Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. Rheumatology (Oxford) 1999;38: 1213–20.

13 Altman R, Alarcon G, Appelrouth D *et al.* The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hand. Arthritis Rheum 1990;33:1601–10.

14 Keen HI, Lavie F, Wakefield RJ *et al.* The development of a preliminary ultrasonographic scoring system for features of hand osteoarthritis. Ann Rheum Dis 2008;67: 651–5.

15 Bruyn GA, Iagnocco A, Naredo E *et al.*; OMERACT Ultrasound Working Group. OMERACT definitions for ultrasonographic pathologies and elementary lesions of rheumatic disorders 15 years on. J Rheumatol 2019;46: 1388–93.

16 Kortekaas MC. LUMC hand osteoarthritis ultrasound scoring atlas. Leiden, the Netherlands: Marion Kortekaas [ebook], 2017.

17 van Tuyl LHD, van der Heijde D, Knol DL, Boers M. Chronological reading of radiographs in rheumatoid arthritis increases efficiency and does not lead to bias. Ann Rheum Dis 2014;73:391–5.

18 Roemer FW, Hunter DJ, Crema MD *et al.* An illustrative overview of semi-quantitative MRI scoring of knee osteoarthritis: lessons learned from longitudinal observational studies. Osteoarthritis Cartilage 2016;24:274–89.

19 Van Holsbeke C, Yazbek J, Holland TK *et al.* Real-time ultrasound vs. evaluation of static images in the preoperative assessment of adnexal masses. Ultrasound Obstet Gynecol 2008;32:828–31.