Development of semiquantitative ultrasound scoring system to assess cartilage in rheumatoid arthritis

Peter Mandl1, Paul Studenic1, Emilio Filippucci2, Artur Bachta3, Marina Backhaus4, David Bong5, George A. W. Bruyn6, Paz Collado7, Nemanja Damjanov8, Christian Dejaco9,10, Andrea Delle-Sedie11, Eugenio De Miguel12, Christina Duftner13, Irina Gessl1, Marwin Gutierrez14, Hilde B. Hammer1, Cristina Hernandez-Diaz16, Annmaria Iagnocco17, Kei Ikeda18, David Kane19, Helen Keen20, Stephen Kelly21, Eszter Kovári22, Ingrid Möller5, Uffe Moller-Dohn23, Esperanza Naredo24, Juan C. Nieto25, Carlos Pineda14, Alex Platzer1, Ana Rodriguez26, Wolfgang A. Schmidt27, Gabriela Supp1, Marcin Szkudlarek28, Lene Terslev23, Ralf Thiele29, Richard J. Wakefield30,31, Daniel Windschall22, Maria-Antonieta D’Agostino33,34, Peter V. Balint35 and OMERACT Ultrasound Cartilage Task Force Group

Abstract

Objectives. To develop and test the reliability of a new semiquantitative scoring system for the assessment of cartilage changes by ultrasound in a web-based exercise as well as a patient exercise of patients with RA.

Methods. A taskforce of the Outcome Measures in Rheumatology Ultrasound Working Group performed a systematic literature review on the US assessment of cartilage in RA, followed by a Delphi survey on cartilage changes and a new semiquantitative US scoring system, and finally a web-based exercise as well as a patient exercise. For the web-based exercise, taskforce members scored a dataset of anonymized static images of MCP joints in RA patients and healthy controls, which also contained duplicate images. Subsequently, 12 taskforce members used the same US to score cartilage in MCP and proximal interphalangeal joints of six patients with RA in in a patient reliability exercise. Percentage agreement and prevalence of lesions were calculated, as intrareader reliability was assessed by weighted kappa and interreader reliability by Light’s kappa.

1Division of Rheumatology, Medical University of Vienna, Vienna, Austria, 2Department of Rheumatology, Università Politecnica delle Marche, Jesi, Ancona, Italy, 3Department of Rheumatology, Military Institute of Medicine, Warsaw, Poland, 4Department of Internal Medicine, Rheumatology and Clinical Immunology, Park-Klinik Weissensee Academic Hospital of the Charité, Berlin, Germany, 5Institute Pooal de Reumatologia, University of Barcelona, Barcelona, Spain, 6Department of Rheumatology, MC Groep, Lelystad, the Netherlands, 7Department of Rheumatology, Hospital Universitario Severo Ochoa, Madrid, Spain, 8University of Belgrade School of Medicine, Institute for Rheumatology, Belgrade, Serbia, 9Department of Rheumatology, Medical University of Graz, Graz, Austria, 10Department of Rheumatology, Hospital of Bruneck, Bruneck, Italy, 11Rheumatology Unit, University of Pisa, Pisa, Italy, 12Department of Rheumatology, Hospital Universitario La Paz, Madrid, Spain, 13Department of Internal Medicine, Clinical Division of Internal Medicine II, Medical University of Innsbruck/Tirol Kliniken, Innsbruck, Austria, 14Division of Musculoskeletal and Rheumatic Diseases, Instituto Nacional de Rehabilitación, México City, Mexico, 15Department of Rheumatology, Diakonhjemmet Hospital, Oslo, Norway, 16Laboratorio de Ultrasonido Musculosqueletico y Articular, Instituto Nacional de Rehabilitación, Mexico City, Mexico, 17Academic Rheumatology Center, University of Western Australia, Perth, Australia, 18Department of Rheumatology, Department of Rheumatology, Medical University of Vienna, Waehringer Guertel 18-20, 1090, Vienna, Austria, 19School of Medicine and Pharmacology Fiona Stanley Hospital Unit, University of Western Australia, Perth, Australia, 20Division of Rheumatology, Mile End Hospital, Barts Health NHS Trust, London, UK, 21School of PhD Studies, Semmelweis University, Budapest, Hungary, 22School of Rheumatology and Spine Diseases, Copenhagen University Hospital Glostrup, Copenhagen, Denmark, 23Department of Rheumatology, Bone and Joint Research Unit, Hospital Universitario Fundación Jiménez Díaz and Autónoma University, 24Department of Rheumatology, Hospital General Universitario Gregorio Marañon and Complutense University, 25Department of Rheumatology, Hospital Ramón y Cajal, Madrid, Spain, 26Medical Center for Rheumatology, Immanuel Krankenhaus Berlin, Berlin, Germany, 27Department of Rheumatology, Zealand’s University Hospital at Konge, Denmark, 28Department of Rheumatology, Department of Rheumatology, Azania and Dentistry, Leicester, USA, 29NIHR Leeds Biomedical Research Centre, Leeds Teaching Hospitals NHS Trust, 30Leeds Institute of Rheumatic and Musculoskeletal Medicine, Chapel Allerton Hospital, University of Leeds, Leids, UK, 31Department of Pediatrics, Asklepios Hospital Wiesentfeld, Wiesentfeld, Germany, 32Department of Rheumatology, APHP, Hopital Ambroise Paré, Paris, France, 33Department of Rheumatology, INSERM U1173, Laboratoire d’Excellence INFLAMEX, UFR Simone Veil, Versailles-Saint-Quentin University, Versailles, France and 343rd Rheumatology Department, National Institute of Rheumatology and Physiotherapy, Budapest, Hungary

Submitted 12 January 2019; accepted 16 March 2019

Correspondence to: Peter Mandl, Division of Rheumatology, Medical University of Vienna, Währinger Gürtel 18-20, 1090, Vienna, Austria. E-mail: peter.mandl@meduniwien.ac.at
Introduction

Joint damage in patients with RA commonly implies the loss of hyaline cartilage and peri-articular erosive changes [1]. It has been shown that loss of cartilage in RA may be more clearly associated with irreversible physical disability than bony damage and therapy directed solely against the erosive process does not ensure the reduction of cartilage loss [2, 3]. Particular attention should therefore be given to early detection and therapeutic interference with cartilage destruction, an early key event of disease pathogenesis [2]. The assessment of cartilage and bone damage in RA has traditionally relied on radiographic assessment in which joint space narrowing has served as a surrogate marker of cartilage loss. The most widely used measure of cartilage damage is the Sharp score and its modifications [4, 5]. Although joint space narrowing is an accepted surrogate marker for cartilage loss, it lacks precision particularly in non-weight-bearing joints and discernment of the relative contributions of damage to cartilage and other soft tissue structures within the joint space narrowing score is not possible [6]. Recently, musculoskeletal ultrasonography (MSUS) has been suggested as a reliable and reproducible tool for the assessment of cartilage in RA in the small joints of the hand [7–9]. A scoring system for cartilage involvement has recently been validated and added to the OMERACT Rheumatoid Arthritis Magnetic Resonance Imaging Score [10, 11].

This study reports on the work of the OMERACT Ultrasound Working Group (USWG), which focused on application of the metric properties of MSUS for detecting and evaluating cartilage damage in RA. The main objectives of the study were to develop standardized definitions for the appearance of normal hyaline cartilage on MSUS, its assessment, elementary lesions for assessing hyaline cartilage change and the grading of such changes, and to test the reliability of a consensual semiquantitative scoring system for the assessment of cartilage changes in the MCP joints by ultrasound (US) in a web-based exercise as well as a patient exercise of patients with RA. Secondary objectives included the testing of the impact of US machines on reliability and the testing of the semiquantitative scoring system on PIP joints.

Methods

Thirty-four international rheumatologist experts in MSUS from 17 countries (Australia, Austria, Denmark, France, Germany, Hungary, Ireland, Italy, Japan, Mexico, Netherlands, Norway, Poland, Serbia, Spain, UK and USA) formed a taskforce within the OMERACT USWG in 2015. The experts agreed upon a sequence of tasks according to the OMERACT filter 2.0 for US studies [12]. As a first step, a systematic literature review (SLR) was performed on studies addressing the sonographic assessment of cartilage in patients with RA. Based on the information obtained from the SLR, the steering committee of the taskforce (P.M., E.F., M.A.D.A. and P.V.B.) formulated statements, including a semiquantitative scoring system, which were agreed upon by the experts in a Delphi exercise. This was followed by testing the reliability of the scoring system first in a web-based exercise on images collected by the experts, followed by a patient-exercise. The ethics committee of the Medical University of Vienna approved the study, which was conducted according to the guidelines of the Declaration of Helsinki, and informed patient consent was obtained.

First step: systematic literature review

An SLR was performed in the PubMed and Embase databases using the search terms: cartilage AND RA AND (US OR ultrasonography). Both original articles and reviews, as well as abstracts presented at the 2010–2016 ACR and EULAR scientific meetings were included. Titles, abstracts and full reports of articles identified were systematically screened and verified by PM and PVB with regard to inclusion and exclusion criteria. Studies published in English up to November 2016, on the use of MSUS for the imaging of cartilage in adult (≥18 years) patients with RA were included. Data with a particular focus on definitions, scanning technique, scoring of cartilage, cartilage changes and cartilage loss were extracted using a standardized template that was specifically designed for the review. The results of the SLR were used by the steering committee to develop statements for the Delphi process.
Second step: Delphi exercise

A written Delphi questionnaire was constructed on the basis of data collected from the SLR and sent to the participating experts. It consisted of nine statements/items, including definitions for the appearance of normal hyaline cartilage on MSUS, its assessment, elementary lesions for assessing hyaline cartilage change and the grading of such changes. The panel was asked to rate each item using a level of agreement or disagreement for each statement according to a five-point Likert scale [13], which was graded as follows: 1, strongly disagree; 2, disagree; 3, neither agree nor disagree; 4, agree; 5, strongly agree.

Group agreement was defined as total cumulative agreement >75% (with a score of 4–5). Only when sentences achieved a score >75%, did we consider that the group had reached a consensus and that the statement was defined as appropriate. The answers from each Delphi questionnaire were summarized with mean scores by a facilitator (P.M.) and re-sent with a revised questionnaire to the panel for the next round, until agreement was reached for all statements.

Third step: web-based intra- and interreader reliability exercise

Taskforce members were instructed to acquire MSUS images of MCP joints 2–5 of healthy subjects and patients with RA using a joint position of ~90 degrees of flexion, which exposes the largest accessible area of hyaline cartilage in the MCP joints [9] using the standardized dorsal longitudinal midline and transverse scans, according to guidelines set forth in a recent review on the sonographic imaging of cartilage, in particular ensuring an insonation angle of 90 degrees [14]. The MSUS equipment used for acquiring the images included the following US units: General Electric Logiq S8, P9 and E9 (GE Medical Systems, Ultrasound and Primary Care Diagnostics, Wauwatosa, WI, USA); ESAOTE Mylab XVG, 25, 70, Class C and Twice (ESAOTE S.p.A. Genoa, Italy); Siemens Acuson Antares and 2000 (Siemens Healthcare GmbH, Erlangen, Germany); Philips Epic7 (Phillips Medical Systems, Andover, MA, USA); Hitachi-Aloka Avius and Ascendus (Hitachi Medical Corporation, Tokyo, Japan). After a collection period of 1 month, the images were sent by e-mail to a facilitator (P.M.). A randomly selected group of 25 images were displayed twice in order to assess intrareader reliability. This was sent to the participants, asking them to read each image and grade the metacarpal cartilage based on the semiquantitative scoring system, which was agreed upon in the Delphi exercise.

Fourth step: patient-based intra- and inter-observer reliability exercise

Twelve taskforce members (A.D.-S., A.I., C.De., C.Du., D.B., E.F., G.A.W.B., H.B.H., H.K., I.M., M.-A.D.’A., P.V.B.) participated in a patient-based intra- and interobserver reliability exercise. During this meeting, MCP and PIP joints 2–5 of six patients with RA were assessed twice on the same day by all experts using US machines.

General Electric Logiq E9, S8 and e (GE Medical Systems, Ultrasound and Primary Care Diagnostics, Wauwatosa, WI, USA) equipped with high-frequency transducers (L8–18i-RS ranging from 8–18 MHz and L10–22-RS ranging from 10–22 MHz) with presets calibrated for the appropriate assessment of cartilage. Participants assessed metacarpal cartilage on the dorsal aspect of the respective joints according to recent guidelines [14], ensuring an insonation angle of 90 degrees and utilizing either the standardized dorsal longitudinal midline scan using the so-called flick-view position (in full possible flexion) or the freehand or dynamic technique whereby the joint position remained the same as during the standardized scan; however, the sonographer was at liberty to shift the transducer and use both longitudinal and transverse planes to assess the entire. Cartilage was scored by the semiquantitative scoring system agreed upon in the Delphi process. Cartilage in the PIP joint was examined on the dorsal aspect using only the dynamic technique outlined above. Two of the patients were examined on the same machine in the morning and the afternoon session, and four patients were examined on different machines, in order to evaluate inter-machine variability.

Statistical analysis

Intraobserver reliability was assessed by weighted kappa and interobserver reliability was assessed by Light’s kappa. Kappa values were interpreted as follows: values of: 0–0.20 represent slight; 0.21–0.40 fair; 0.41–0.60 moderate; 0.61–0.80 good and >0.80 excellent reliability [15]. Additionally, 95%CI were calculated. Percentage of observed agreement (i.e. percentage of observations that obtained the same score) and prevalence of the observed lesions were also calculated. Statistical analyses were performed using R and STATA. The ethics committee of the Medical University of Vienna approved the study, which was conducted according to the guidelines of the Declaration of Helsinki. Each patient gave written informed consent to participate.

Results

Systematic literature review

A total of 198 articles were identified of which finally nine studies reporting on original research could be included in the SLR [7–9, 16–21]. The flowchart of the review process is included in Supplementary Fig. S1, available at Rheumatology online. Data extracted from the included studies were used to formulate statements for the Delphi process and were also shared with the participants for the patient exercise. Key data from studies selected for final review are summarized in Table 1.

Delphi exercise

A total of 27 experts were invited of whom 24 (89%) participated in both the first and second, round of the Delphi exercise. For seven out of nine statements, agreement was >75%, and for two statements <75% after the first round. Wording was improved based on experts’
<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects</th>
<th>Sample size</th>
<th>Joint scanned</th>
<th>Scoring method used</th>
<th>Definition used for cartilage damage</th>
<th>Reliability</th>
<th>Correlation with other method</th>
</tr>
</thead>
<tbody>
<tr>
<td>[16]</td>
<td>Healthy, RA, OA</td>
<td>30/48/60</td>
<td>knee</td>
<td>Discrete measurement &amp; binary grading</td>
<td>Blurring, irregularity</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>[17]</td>
<td>Healthy/RA</td>
<td>20/20</td>
<td>MCP 2-3</td>
<td>Binary grading</td>
<td>Loss of definition, indistinctness</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>[7]</td>
<td>Healthy, RA, OA, unclassified arthritis</td>
<td>34/48/18/24</td>
<td>MCP 2-5, PIP 2-5</td>
<td>Discrete measure</td>
<td>Thinning of cartilage</td>
<td>ICC for cartilage thickness (bilateral sum score): 0.844 (interobserver) 0.928 (intraobserver)</td>
<td>Correlation between MSUS score and radiographic JSN for both hands: adjusted $r^2$: 0.513 ($P&lt;0.001$) and JSW of the same finger joints: adjusted $r^2$: 0.635 ($P&lt;0.001$)</td>
</tr>
<tr>
<td>[8]</td>
<td>RA</td>
<td>20</td>
<td>MCP 2-3</td>
<td>Semiquantitative score 0-4, adapted from [22]</td>
<td>● 1=loss of the sharpness of the superficial margin of the hyaline cartilage; ● 2=partial thickness defect of the cartilage layer; ● 3=full thickness defect of the cartilage layer with a normal subchondral bone profile; ● 4=complete loss of the cartilage layer and subchondral bone involvement.</td>
<td>Weighted kappa: 0.672, 0.537 &amp; 0.832 (interobserver reliability for dorsal, lateral and volar quadrant) Weighted kappa values for total additive scores per joint: 0.729 &amp; 0.733, respectively, for detection and for semiquantitative assessment of cartilage damage</td>
<td>NA</td>
</tr>
<tr>
<td>[18]</td>
<td>RA</td>
<td>100</td>
<td>knee</td>
<td>Binary grading based on definition adopted from [23]</td>
<td>Thickening or thinning of cartilage layer, loss of definition of chondro-synovial margin and pitting of the articular surface</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>[19]</td>
<td>RA</td>
<td>30</td>
<td>knee</td>
<td>Discrete measurement &amp; binary grading</td>
<td>Irregularity, loss of clarity</td>
<td>NA</td>
<td>Agreement between MRI &amp; MSUS on cartilage morphology, kappa: 0.658, 0.851 (medial &amp; lateral)</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects</th>
<th>Sample size</th>
<th>Joint scanned</th>
<th>Scoring method used</th>
<th>Definition used for cartilage damage</th>
<th>Reliability</th>
<th>Correlation with other method</th>
</tr>
</thead>
<tbody>
<tr>
<td>[20]</td>
<td>RA</td>
<td>52</td>
<td>hip</td>
<td>Binary grading</td>
<td>Loss of sharpness of the cartilage margins, loss of homogeneity of the cartilage layer, cartilage thinning (focal or extending to the entire cartilaginous layer)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>[9]</td>
<td>Healthy⁶/RA</td>
<td>3/5/35</td>
<td>MCP 2-5</td>
<td>Discrete measure</td>
<td>Thinning of cartilage</td>
<td>ICC</td>
<td>ICC between sono- graphic and anatomic cartilage thickness: 0.73</td>
</tr>
</tbody>
</table>
| [21]  | RA#      | 15          | MTP 2-5       | Semiquantitative score 0-6, adapted from [25] | • 1=Blurred margin or partial lack of clarity, without thickness change  
• 2=Blurred margin and partial lack of clarity, without thickness change  
• 3=Blurred margin and complete lack of clarity  
• 4=Difficult-to-define margin and completely opaque band;  
• 5=Marked thickness change;  
• 6=No visualized cartilage band | Correlation between in vivo & in vitro MSUS (before and after operation): $r=0.74 \ (P < 0.01)$ | Correlation between MSUS grading & histologic grading: $r=0.67-0.83 \ (P < 0.01)$ |

⁶Including cadaver; ⁷scans were performed pre-operatively as well as in vitro on the resected metatarsal heads; ICC: intraclass correlation coefficient; JSN: joint space narrowing; JSW: joint space width; MSUS: musculoskeletal ultrasound; NA: not assessed.
comments and all statements were presented in the second round of the exercise, in which all statements achieved agreement. Table 2 shows the final statements and their agreement. Among others, the participants agreed upon the definition of normal hyaline cartilage on US as well as on the elementary lesion of cartilage damage: blurring of the outer margin and/or the subchondral margin under orthogonal insonation, focal or diffuse thinning of the hyaline cartilage layer as well as the incomplete or complete loss of homogeneity of the echostucture. Based on these definitions, a semiquantitative scoring system ranging from 0–2 (grade 0, normal cartilage; grade 1, minimal change: focal thinning or incomplete loss of cartilage; grade 2, severe change: diffuse thinning or complete loss of cartilage) was formulated (Fig. 1). In addition, the participants agreed upon a statement on the quantitative assessment of cartilage, taking into consideration the recommendations from a recent review on the pitfalls of cartilage measurement on US, in particular the need for orthogonal insonation, inclusion of the outer margin in the measurement and correction for the higher speed of sound in hyaline cartilage as compared to soft tissue.

**Web-based intra- and interreader reliability exercise**

A total of 17 taskforce members sent 20 anonymized images each of MCP joints 2-5 acquired from healthy subjects and patients with RA, both in the longitudinal

**Table 2** Statements and final agreement after the second round of Delphi exercise

<table>
<thead>
<tr>
<th>Category</th>
<th>Statement</th>
<th>Round</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSUS definition of normal hyaline cartilage</td>
<td>Normal hyaline cartilage has a homogeneous anechoic or hypoechoic echostructure, parallel to the echogenic bony cortex, is delineated by a sharp subchondral margin, and possesses a sharp outer margin, when the cartilage is insonated orthogonally.</td>
<td>1</td>
<td>88%</td>
</tr>
<tr>
<td>Assessment of hyaline cartilage by MSUS</td>
<td>Hyaline cartilage must be assessed using orthogonal insonation (MSUS beam falling perpendicular to the hyaline cartilage surface). The joint should be positioned to expose the largest accessible area of hyaline cartilage. The entire area of cartilage accessible within the acoustic window should be scanned, in both longitudinal and transverse planes. The optimization of settings, including the position of the joint at the baseline examination as well as the maintenance of such settings for possible follow-up examination(s) is mandatory for the purpose of monitoring in clinical practice. Whenever possible, anatomical landmarks should be identified and utilized to ensure that follow-up assessments are conducted at the appropriate locations.</td>
<td>1</td>
<td>96%</td>
</tr>
<tr>
<td>Elementary MSUS lesions of cartilage change</td>
<td>Blurring of the outer margin and/or the subchondral margin under orthogonal insonation. Focal or diffuse thinning of the hyaline cartilage layer, Incomplete or complete loss of homogeneity of the echostructure.</td>
<td>1</td>
<td>84%</td>
</tr>
<tr>
<td>Grading MSUS cartilage changes</td>
<td>For the purpose of grading changes, hyaline cartilage should be assessed in both longitudinal and transverse planes.</td>
<td>2</td>
<td>94%</td>
</tr>
<tr>
<td>Semiquantitative grading of cartilage</td>
<td>A 3-grade semiquantitative scoring system (i.e. grade 0, normal cartilage; grade 1, minimal change: focal thinning or incomplete loss of cartilage; grade 2, severe change: diffuse thinning or complete loss of cartilage can be used to grade hyaline cartilage change in RA.</td>
<td>1</td>
<td>80%</td>
</tr>
<tr>
<td>Quantitative grading of cartilage</td>
<td>Hyaline cartilage thickness can be measured using the largest distance between the subchondral and outer margins, and if possible including the outer, but not the subchondral margin by the caliper tool. For monitoring purposes, the cartilage thickness measurement using the calipers is sufficient and does not need to be corrected, when correlating with anatomical/histological or other imaging measurement, the data obtained using the calipers should be corrected for the higher speed of sound in hyaline cartilage as compared to soft tissue.</td>
<td>2</td>
<td>77%</td>
</tr>
</tbody>
</table>

MSUS: musculoskeletal ultrasound.
and transverse scans. The conveners (E.F., P.M. and P.V.B.) reviewed the total number of 340 images for quality and created a dataset of 123 images, consisting of 73 individual images as well as duplicates of 25 randomly selected images. The dataset was sent to participants who graded each image using the semiquantitative grading system agreed upon in the Delphi exercise. The kappa values for intrareader reliability of the web-based exercise were 0.87 (95% CI 0.83–0.92) and for interreader reliability 0.64 (95% CI 0.63–0.64).

Patient intra- and interreader reliability exercise
In the patient exercise, 4 out of 6 patients were women; mean age was 64 (range: 52–67) years, mean disease duration was 15 (range: 4–31) years, 83% (5/6) of patients were rheumatoid factor and anti-citrullinated peptide antibody positive. The observed prevalence of grades of cartilage damage for both the MCP and the PIP joints are listed in Table 3.

The results of the reliability of the semiquantitative scoring system are summarized in Table 4. The intrareader agreement was 84.2% (range: 64.9–100) and 76.2% (range: 57.7–95.3) for the MCP joints (standardized and dynamic, respectively) and 57.1% (range: 23.7–92.9) for the PIP joints. Kappa values for intrareader reliability were 0.78 (95% CI 0.74–0.82) for the standardized scan of the MCP joints, 0.83 (95% CI 0.80–0.86) for the dynamic scan of the MCP joints and 0.66 (95% CI 0.60–0.71) for the PIP joints. The interreader agreement was 62.7% (range 28.1–79.1) and 64.3% (range: 45.8–80.2) for the MCP joints (standardized and dynamic, respectively) and 44.1% (range: 18.7–69.8) for the PIP joints. Kappa values for interreader reliability were 0.44 (95% CI 0.38–0.51) for the standardized scan of the MCP joints, 0.48 (95% CI 0.41–0.54) for the dynamic scan of the MCP joints and 0.17 (95% CI 0.13–0.21) for the PIP joints.

The estimates for each individual joint are listed in Table 5. Based on the kappa values, no individual joint
Impact of different US machines on reliability

Table 4 presents overall prevalence, intra- and interreader agreement and intra- and interreader reliability estimates for each joint group and scanning method.

Table 5 provides intra- and interreader reliability estimates for each joint.

Discussion

The main objective of this study was to develop definitions for cartilage damage in RA and test the reliability of a semiquantitative scoring system. Standardization of changes and validated scoring system would facilitate the dissemination of this technique in daily practice and allow adequately trained sonographers to participate in multicentre research studies aiming to assess cartilage changes.

This is the first reliability study of a sonographic scoring of cartilage abnormalities in RA that was developed according to the OMERACT framework. The inclusion of different US machines in both the web-based and patient-based exercise corresponds to the real-life application of US in routine clinical practice and multicentre studies, and also allowed us to demonstrate that the use of different vs same US machines indeed has an impact on reliability.

Although an OMERACT taskforce on hand OA reported good agreement on definitions of cartilage damage in hand OA [26], recent attempts at developing a semiquantitative scoring system in hand OA have found only moderate intrareader and fair interreader reliability [27]. It was suggested that the poor reliability, in particular of the two intermediate scores (scores 1 and 2 on a 0–3 scale) may be explained by the fact that the proposed definitions could not help to sufficiently distinguish between intermediate grades. The SLR revealed a single study by Filippucci et al. who have performed a single-center interreader reliability study on MCP 2–3 joints in RA patients using two experienced rheumatologists [8] and reported substantial reliability for a 0–4 semiquantitative scoring system. In the Delphi exercise, the taskforce opted for a simpler semiquantitative scoring system of 0–2. Using this system, we found substantial to excellent intrareader reliability and moderate to substantial interreader reliability (web-based and patient-based exercise respectively) in the MCP joints of RA patients. The dynamic or freehand

Table 4: Overall prevalence, intra- and interreader agreement and intra- and interreader reliability

<table>
<thead>
<tr>
<th>Joint group and scanning method</th>
<th>Grade (0–2)</th>
<th>Prevalence range in %</th>
<th>Intrareader agreement in mean (range), %</th>
<th>Intrareader reliability kappa (%95CI)</th>
<th>Interreader agreement in mean (range) %</th>
<th>Interreader reliability kappa (%95CI)</th>
<th>Interreader reliability kappa (%95CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCP Standard</td>
<td>0</td>
<td>9.3–54.2</td>
<td>84.2 (64.9–100)</td>
<td>0.78 (0.74–0.82)</td>
<td>62.7 (28.1–79.1)</td>
<td>0.44 (0.38–0.51)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>12.5–55.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>14.6–43.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCP Dynamic</td>
<td>0</td>
<td>18.7–53.1</td>
<td>76.2 (57.7–95.3)</td>
<td>0.83 (0.80–0.86)</td>
<td>64.3 (45.8–80.2)</td>
<td>0.48 (0.41–0.54)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8.7–48.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>20.8–42.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIP Dynamic</td>
<td>0</td>
<td>3.1–61.4</td>
<td>57.1 (23.7–92.9)</td>
<td>0.66 (0.60–0.71)</td>
<td>44.1 (18.7–69.8)</td>
<td>0.17 (0.13–0.21)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>23.9–67.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>8.3–56.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MCP: metacarpophalangeal joint; PIP: proximal interphalangeal joint.

Table 5: Intra- and interreader reliability estimates for each joint

<table>
<thead>
<tr>
<th>MCP</th>
<th>Scanning method</th>
<th>Intrareader reliability kappa (%95CI)</th>
<th>Interreader reliability kappa (%95CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Standard</td>
<td>0.86 (0.74–0.90)</td>
<td>0.47 (0.39–0.56)</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>0.83 (0.77–0.88)</td>
<td>0.51 (0.41–0.59)</td>
</tr>
<tr>
<td>3</td>
<td>Standard</td>
<td>0.85 (0.77–0.90)</td>
<td>0.47 (0.34–0.59)</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>0.86 (0.80–0.91)</td>
<td>0.57 (0.48–0.70)</td>
</tr>
<tr>
<td>4</td>
<td>Standard</td>
<td>0.78 (0.70–0.85)</td>
<td>0.33 (0.22–0.50)</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>0.67 (0.53–0.75)</td>
<td>0.34 (0.22–0.52)</td>
</tr>
<tr>
<td>5</td>
<td>Standard</td>
<td>0.80 (0.73–0.86)</td>
<td>0.44 (0.31–0.58)</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>0.69 (0.59–0.78)</td>
<td>0.41 (0.28–0.57)</td>
</tr>
<tr>
<td>PIP</td>
<td>2</td>
<td>0.72 (0.62–0.79)</td>
<td>0.14 (0.09–0.19)</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>0.72 (0.61–0.81)</td>
<td>0.26 (0.17–0.36)</td>
</tr>
<tr>
<td>4</td>
<td>Dynamic</td>
<td>0.68 (0.57–0.77)</td>
<td>0.19 (0.13–0.26)</td>
</tr>
<tr>
<td>5</td>
<td>Dynamic</td>
<td>0.50 (0.31–0.64)</td>
<td>0.09 (0.04–0.16)</td>
</tr>
</tbody>
</table>

MCP: metacarpophalangeal joint; PIP: proximal interphalangeal joint.

could be selected that performed better than the others, although overall the MCPs and PIPs joints performed worse as compared with MCP 2–4 and PIP 2–4, respectively. Finally, the estimates for each individual patient are listed in Supplementary Table S1, available at Rheumatology online. Kappa values for intra- and interreader reliability varied consistently, with higher overall estimates for patients 2 and 4, who both had relatively longer disease duration, as compared with the other patients.

Impact of different US machines on reliability

Intrareader reliability was better for examinations that took place on the same machine as compared with those performed on different machines (kappa values: 0.73 (95%CI 0.63–0.81) vs 0.59 (95%CI 0.51–0.66) for the standard scan of the MCP joints; 0.64 (95%CI 0.53–0.73) vs 0.52 (95%CI 0.43–0.59) for the dynamic scan of the MCP joints and 0.59 (95%CI 0.49–0.69) vs 0.48 (95%CI 0.40–0.56) for the PIP joints).

Discussion

The main objective of this study was to develop definitions for cartilage damage in RA and test the reliability of a semiquantitative scoring system. Standardization of changes and validated scoring system would facilitate the dissemination of this technique in daily practice and allow adequately trained sonographers to participate in multicentre research studies aiming to assess cartilage changes.

This is the first reliability study of a sonographic scoring of cartilage abnormalities in RA that was developed according to the OMERACT framework. The inclusion of different US machines in both the web-based and patient-based exercise corresponds to the real-life application of US in routine clinical practice and multicentre studies, and also allowed us to demonstrate that the use of different vs same US machines indeed has an impact on reliability.

Although an OMERACT taskforce on hand OA reported good agreement on definitions of cartilage damage in hand OA [26], recent attempts at developing a semiquantitative scoring system in hand OA have found only moderate intrareader and fair interreader reliability [27]. It was suggested that the poor reliability, in particular of the two intermediate scores (scores 1 and 2 on a 0–3 scale) may be explained by the fact that the proposed definitions could not help to sufficiently distinguish between intermediate grades. The SLR revealed a single study by Filippucci et al. who have performed a single-center interreader reliability study on MCP 2–3 joints in RA patients using two experienced rheumatologists [8] and reported substantial reliability for a 0–4 semiquantitative scoring system. In the Delphi exercise, the taskforce opted for a simpler semiquantitative scoring system of 0–2. Using this system, we found substantial to excellent intrareader reliability and moderate to substantial interreader reliability (web-based and patient-based exercise respectively) in the MCP joints of RA patients. The dynamic or freehand
scanning of the MCP joints was found to be slightly superior to the standardized view. In addition, we could also confirm that using the same machine in both the morning and afternoon round leads to improved intraobserver reliability.

At the same time, by opting for a scoring system between 0–2, while this may be reliable and useful for assessing focal or severe cartilage damage at single timepoints, based on the above-mentioned experiences in scoring it may potentially be less discriminant in studies investigating progression apart from those conducted on patients with very early disease. In addition to the semi-quantitative scoring system, the taskforce also agreed on a statement on quantitative grading, which may provide a more accurate evaluation of cartilage, albeit it would likely be less feasible in a multicentre study. The latter definition is in line with recent recommendations that highlighted the pitfalls of US measurement of cartilage [14].

An additional limitation of our study could be the absence of PIP images in the web-based reliability exercise, which may explain the moderate intrareader reliability and only slight interobserver reliability in the PIP joints. The latter results may also reflect technical problems associated with the visualization of cartilage in this joint, which may require further modification of the scanning technique (e.g. utilizing palmar transverse scans to visualize cartilage). Although the number of patients seems very low, the number of examined structures in total was quite high (n = 96). In addition, the number of readers was also quite high (n = 12). According to several reports [28, 29] focusing on improving variability of reliability studies, it is important either to have an adequate number of patients or of readers. In these studies, 6–8 patients or 10–14 readers are recommended as adequate sample sizes. The number of patients utilized in our study is in the range used in previous reliability exercises on US [30, 31]. Although we took care to include patients in the patient exercise that conform to an average RA population with regard to distribution of age and sex and included patients with different disease duration, due to the low patient number, which is usual for such exercises, we cannot rule out a patient selection bias, which may have affected the results.

Based on the present study, the OMERACT USWG recommends the use of the presently described semiquantitative MSUS score for assessing cartilage pathology in the MCP joints of patients with RA. Further testing of this scoring system in the MCP joints of other RA cohorts in addition to joints where cartilage can be visualized (e.g. knee, metatarsophalangeal, tibiotalar, etc.) and assessment of sensitivity of change in longitudinal studies is required before the scoring system can be recommended as an outcome measure to be used in clinical trials.

Acknowledgements


Funding: This work was supported by an unrestricted grant from UCB Pharma GmbH.

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

Supplementary data

Supplementary data are available at Rheumatology online.

References


15 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.


