



## ORIGINAL RESEARCH

# Automated ultrasound system ARTHUR V.2.0 with AI analysis DIANA V.2.0 matches expert rheumatologist in hand joint assessment of rheumatoid arthritis patients

Bill Aplin Frederiksen,<sup>1</sup> Hilde Berner Hammer ,<sup>2</sup> Lene Terslev,<sup>3</sup> Mads Ammitzbøll-Danielsen,<sup>3</sup> Thiusius Rajeeth Savarimuthu,<sup>4</sup> Anders Bossel Holst Weber,<sup>5</sup> Søren Andreas Just <sup>1</sup>

**To cite:** Frederiksen BA, Hammer HB, Terslev L, *et al.* Automated ultrasound system ARTHUR V.2.0 with AI analysis DIANA V.2.0 matches expert rheumatologist in hand joint assessment of rheumatoid arthritis patients. *RMD Open* 2025;**11**:e005805. doi:10.1136/rmdopen-2025-005805

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/rmdopen-2025-005805>).

Received 17 April 2025  
Accepted 10 July 2025



© Author(s) (or their employer(s)) 2025. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

For numbered affiliations see end of article.

## Correspondence to

Dr Søren Andreas Just;  
[soreen.andreas.just@rsyd.dk](mailto:soreen.andreas.just@rsyd.dk)

## ABSTRACT

**Objective** To evaluate the agreement and repeatability of an automated robotic ultrasound system (ARTHUR V.2.0) combined with an AI model (DIANA V.2.0) in assessing synovial hypertrophy (SH) and Doppler activity in rheumatoid arthritis (RA) patients, using an expert rheumatologist's assessment as the reference standard. **Methods** 30 RA patients underwent two consecutive ARTHUR V.2.0 scans and rheumatologist assessment of 22 hand joints, with the rheumatologist blinded to the automated system's results. Images were scored for SH and Doppler by DIANA V.2.0 using the EULAR-OMERACT scale (0–3). The agreement was evaluated by weighted Cohen's kappa, percent exact agreement (PEA), percent close agreement (PCA) and binary outcomes using Global OMERACT-EULAR Synovitis Scoring (healthy  $\leq 1$  vs diseased  $\geq 2$ ). Comparisons included intra-robot repeatability and agreement with the expert rheumatologist and a blinded independent assessor.

**Results** ARTHUR successfully scanned 564 out of 660 joints, corresponding to an overall success rate of 85.5%. Intra-robot agreement for SH: PEA 63.0%, PCA 93.0%, binary 90.5% and for Doppler, PEA 74.8%, PCA 93.7%, binary 88.1% and kappa values of 0.54 and 0.49. Agreement between ARTHUR+DIANA and the rheumatologist: SH (PEA 57.9%, PCA 92.9%, binary 87.3%, kappa 0.38); Doppler (PEA 77.3%, PCA 94.2%, binary 91.2%, kappa 0.44) and with the independent assessor: SH (PEA 49.0%, PCA 91.2%, binary 80.0%, kappa 0.39); Doppler (PEA 62.6%, PCA 94.4%, binary 88.1%, kappa 0.48).

**Conclusions** ARTHUR V.2.0 and DIANA V.2.0 demonstrated repeatability on par with intra-expert agreement reported in the literature and showed encouraging agreement with human assessors, though further refinement is needed to optimise performance across specific joints.

## INTRODUCTION

Musculoskeletal ultrasound is a sensitive and dynamic tool for assessing synovial

## WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ Musculoskeletal ultrasound improves the early detection and monitoring of rheumatoid arthritis but is highly operator dependent.
- ⇒ Automated ultrasound systems offer potential but require validation of both scanning and interpretation performance.

## WHAT THIS STUDY ADDS

- ⇒ A fully automated robotic ultrasound system (ARTHUR V.2.0) combined with AI-based scoring (DIANA V.2.0) achieved repeatability comparable to expert rheumatologists.
- ⇒ The system showed encouraging agreement with expert grading, though further refinement is needed to optimise performance across joint types.

## HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ⇒ Automation of scanning and scoring may help overcome workforce limitations in rheumatology, supporting earlier and more consistent assessment of inflammatory arthritis.
- ⇒ These findings support further clinical trials and health economic evaluations to guide future implementation and policy adoption.

inflammation in rheumatoid arthritis (RA), particularly through evaluation of synovial hypertrophy (SH) and Doppler activity.<sup>1–3</sup> To promote consistency in interpretation, the EULAR-OMERACT collaboration has developed validated scoring systems for SH, Doppler signal and combined synovitis assessment.<sup>4</sup> These systems are adopted in research and increasingly used in clinical practice.

Despite this progress, ultrasound remains heavily dependent on operator expertise for

both image acquisition and interpretation, contributing to variability in clinical decision-making.<sup>5</sup> Furthermore, access to timely musculoskeletal ultrasound is constrained by a growing shortage of rheumatologists and prolonged waiting times for specialist evaluation in many healthcare systems.<sup>6</sup>

Recent technological developments have introduced a fully automated approach to joint ultrasound assessment. ARTHUR V.2.0 is a CE-marked robotic system capable of autonomously acquiring standardised scans of small joints and is well tolerated by patients.<sup>7</sup> DIANA V.2.0, also CE-marked, is an AI model trained to assess ultrasound images according to EULAR-OMERACT grading criteria and has previously demonstrated expert-level performance in image interpretation.<sup>8</sup>

While DIANA has been validated independently, the combined performance of ARTHUR and DIANA—automating the complete workflow from scanning to scoring—has not yet been systematically evaluated. This study investigates the metric properties of the ARTHUR+DIANA pipeline in a cohort of patients with established, clinically active RA. We explore the agreement between ARTHUR+DIANA and a rheumatological expert scoring, assessing intra- and inter-rater reliability, and repeatability. Such an evaluation is essential before considering broader clinical implementation of fully automated ultrasound solutions.

## METHODS

### Study design

30 RA patients from Svendborg Hospital's Rheumatology section were included. Patients were consecutively recruited from the outpatient rheumatology clinic at Odense University Hospital, Svendborg, Denmark. Eligible participants were adults (≥18 years) with a diagnosis of RA according to the 2010 American College of Rheumatology/European Alliance of Associations for Rheumatology criteria, and with clinically active disease in at least one hand joint. Exclusion criteria included severe joint deformities and the inability to provide informed consent. Patients with severe hand joint deformities were excluded because such anatomical alterations would interfere with ARTHUR's current ability to perform standardised, reproducible sweeps across joint surfaces. Furthermore, joints with severe deformities fall outside the scope of the validated EULAR-OMERACT ultrasound scoring systems that were also used as the basis for training DIANA. Of 34 screened patients, 2 were excluded due to severe deformities and 2 declined participation.

All patients had metacarpophalangeal (MCP) 1–5, interphalangeal joint (IP) 1, proximal interphalangeal (PIP) 2–5 and wrist (radiocarpal-intercarpal (RCIC)) joints on both hands, scanned two times by ARTHUR V.2.0 (figure 1A) and then by the rheumatologist (figure 1B). The same rheumatologist performed all ultrasound assessments. The rheumatologist performing the manual

ultrasound and scoring was blinded to the image selection and output generated by ARTHUR and DIANA. To assess repeatability, ARTHUR V.2.0 performed two independent scans of each hand. Between scans, the patient's hands were completely repositioned on the scanning platform to simulate separate scanning sessions and ensure realistic variation in joint orientation and transducer contact.

The ARTHUR V.2.0 system is a fully automated robotic ultrasound platform, also being CE-certified (MDR class IIa), designed to perform standardised sweeps over hand joints. It uses real-time image analysis to identify optimal acquisition frames, which are subsequently passed to DIANA for interpretation.

DIANA V.2.0 is a convolutional neural network-based image analysis model developed and CE-certified by ROPCA under MDR as a class IIa device. The model was trained and validated on over 10 000 ultrasound images of hand joints annotated and segmented by expert rheumatologists using the EULAR-OMERACT grading system. These images were acquired from clinical trials from hospitals in Denmark and included data primarily from two different ultrasound scanner models, General Electric Logiq 9 and 10 systems. No images from the present study were included in the training or validation of DIANA V.2.0.

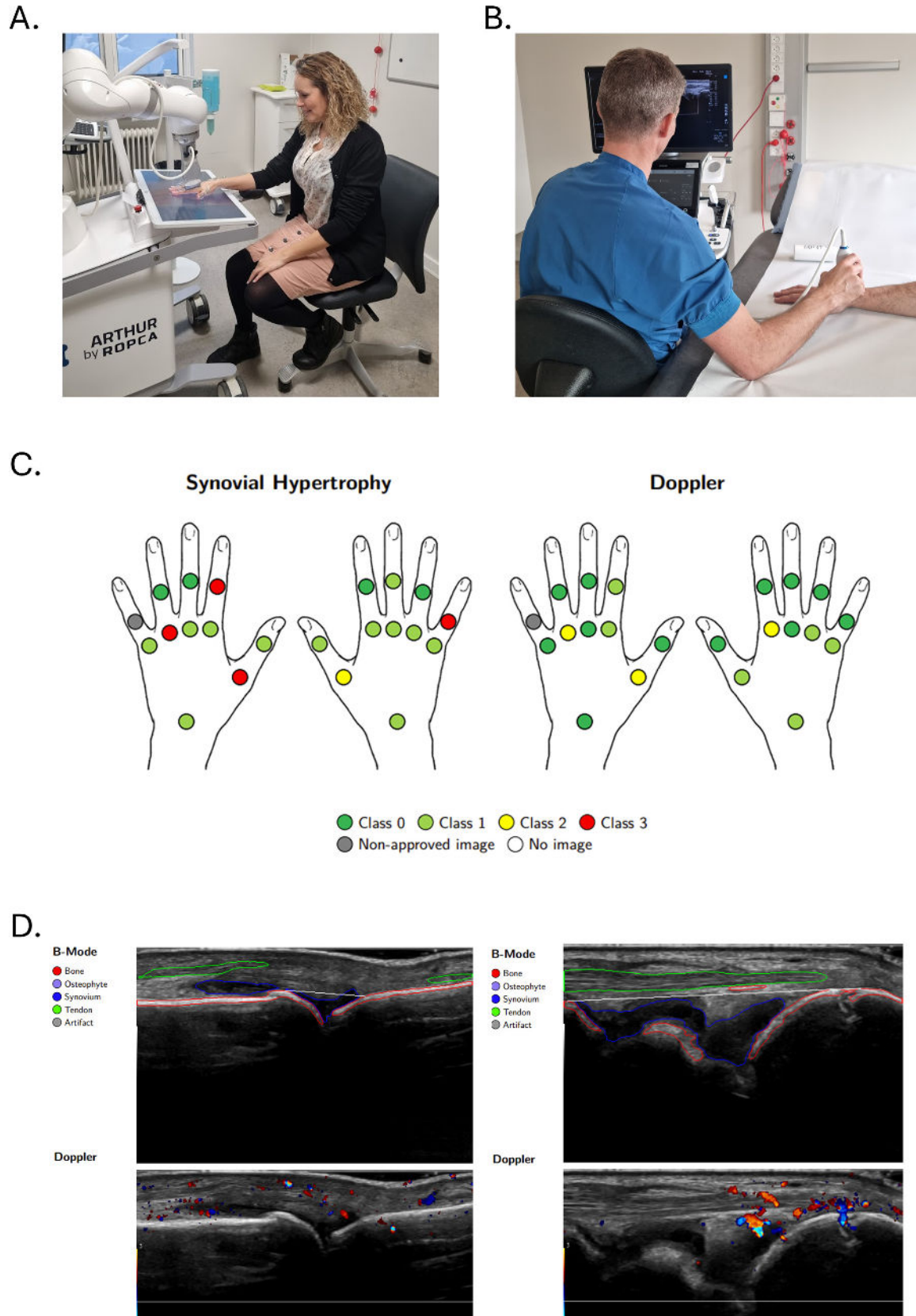
Both ARTHUR and the rheumatologist used a General Electric Logiq 10, R3, ultrasound scanner with an ML6-15 ultrasound probe to obtain the images. Identical settings were applied by ARTHUR and the rheumatologist with the Doppler signal gain set to a sensitivity just below the noise level, Doppler frequency 10.3 MHz, pulse repetition frequency 0.8 and the wall filter 86 Hz.

The experienced rheumatologist (BAF), an ultrasound teacher and a lecturer both at national and EULAR levels, has more than 10 years of experience in musculoskeletal ultrasound.

### Quality assessment and scoring of disease activity

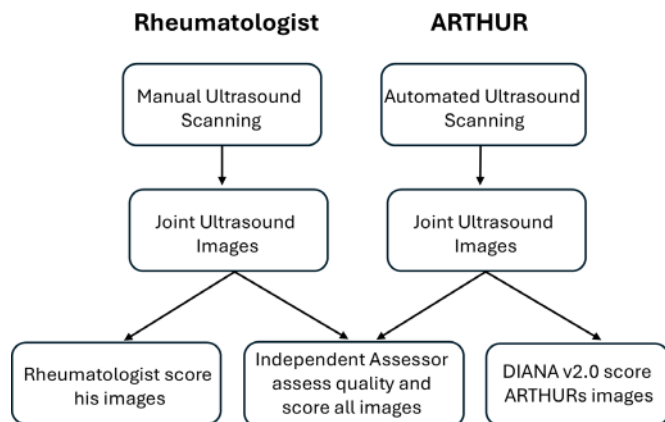
For an overview of the evaluation flow, see figure 2. The rheumatologist performed and interpreted all manual ultrasound examinations. This assessment was used as the reference standard for the primary outcome comparisons.

An independent expert assessor (HBH), another rheumatologist with over 20 years of experience in musculoskeletal ultrasound and a lecturer at both national and EULAR levels, reviewed all stored images from both the rheumatologist and ARTHUR V.2.0. The independent assessor (IA) assessed image quality and graded SH, Doppler activity and the combined synovitis score (Global OMERACT-EULAR Synovitis Scoring (GLOESS)) was calculated according to EULAR-OMERACT standards<sup>9</sup> (see figure 2). For each joint, the IA evaluated two sets of images: (1) those obtained by the rheumatologist and (2) those obtained by ARTHUR. To reflect a comprehensive assessment and avoid potential false negatives due to variability in image quality or acquisition angle, the highest



**Figure 1** (A) Patient scanned by ARTHUR. (B) The rheumatologist is performing an ultrasound scan on a patient. (C) Example of the scan result from the PDF report from DIANA V.2.0 (not related to patients on images A and B). (D) Example of how DIANA V.2.0 segments ultrasound images and presents them in the PDF report, so the clinician can see the reasoning behind a given disease activity score. The blue marking is synovium and cartilage, the red marking is bone and the green tendon. The white line over the joint is to discriminate between grade 0 and 1 (synovium below) and grade 2 and 3 synovial hypertrophy (synovium over the line).





**Figure 2** Flowchart of image acquisition and scoring procedure. Ultrasound images were acquired by both the rheumatologist and the ARTHUR V.2.0 system. The independent assessor reviewed and scored all images blinded to the source. For each joint, the highest score assigned by the independent assessor across the two image sets was used for comparison with the corresponding assessment by the rheumatologist or DIANA.

grade from either image set was assigned as the final IA score for that joint. This conservative approach aimed to maximise sensitivity in detecting pathology, acknowledging the clinical relevance of not overlooking inflammatory changes.

Only joints successfully scanned and approved by the IA were included in the main analysis. Joints with insufficient image quality, as determined by the IA, were excluded. As the analysis was predefined to focus on valid image pairs, no imputation or maximum disagreement was applied to missing data. To explore whether the exclusion of low-quality images could introduce a systematic bias related to disease activity, we collected the rheumatologist's disease assessments in these joints.

DIANA V.2.0 evaluated ARTHUR's images and created a rapport showing why a given score was obtained (figure 1C,D). The rheumatologist and the IA scored the ultrasound images using the CVAT (Computer Vision Annotation Tool) platform. In this study, CVAT was used as a structured manual scoring interface, allowing side-by-side viewing of static images and Doppler clips. All image grading was performed manually according to the EULAR-OMERACT scoring system, and no automated or AI-assisted functions were used within the platform. The IA could choose not to grade an image if it is of low quality.

### Patient and public involvement

Patients or the public were not involved in the design, conduct, reporting or dissemination of this research.

### Statistics

The primary comparison in this study was between ARTHUR+DIANA and the expert rheumatologist, whose grading of manually acquired images served as the reference standard (ground truth). Secondary comparisons

were made between ARTHUR+DIANA and the IA, who graded all images (from both the rheumatologist and ARTHUR) independently and blinded to source and AI outputs.

At the joint level, inter-rater agreement was assessed using percent exact agreement (PEA), percent close agreement (PCA; defined as a score difference of  $\pm 1$ ), binary agreement (normal (grade 0–1) vs abnormal (grade 2–3)), sensitivity and specificity. Weighted Cohen's kappa (ordinal weights) was calculated and interpreted according to Landis and Koch: <0.20 poor, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good and 0.81–1.00 very good agreement.

Intra-robot agreement (repeatability) was evaluated by comparing ARTHUR+DIANA's first and second scans using the same metrics. The second ARTHUR scan was used in all subsequent comparisons with the rheumatologist and the IA.

At the patient level, agreement and sensitivity were assessed by dichotomising disease status: the presence of  $\geq 1$  joint with disease activity (binary GLOESS  $\geq 2$ ) in both hands was classified as disease-positive; absence in all joints was classified as disease-negative.

All analyses were conducted using Stata V.18.0, excluding joints with missing or rejected evaluations.

## RESULTS

### Study population

Patient characteristics are shown in table 1.

### Image quality assessment

ARTHUR scanned 660 hand joints two times. All images were assessed by the IA, resulting in a scanning success rate of 85.45%, the highest scores for MCP 2–4 and PIP 2–4 (table 2). The rheumatologist scanned 660 joints one time with a success rate of 100%.

To evaluate potential selection bias, we compared the rheumatologist's disease activity scores in these excluded joints. As shown in online supplemental table 1, there was no consistent indication that failed scans occurred more frequently in joints with higher disease activity.

**Table 1** Baseline characteristics of the study population

Characteristic	Value
Number of patients	30
Age (years, mean $\pm$ SD)	64.9 $\pm$ 9.6
Gender (male/female)	7/23
Disease duration (years, mean $\pm$ SD)	10.8 $\pm$ 10.9
DAS28-CRP (mean $\pm$ SD)	3.9 $\pm$ 1.3
On treatment with DMARDs (%)	73.3
On treatment with biologics (%)	36.7
DAS28-CRP, Disease Activity Score of 28 joints with C Reactive Protein; DMARDs, disease modifying antirheumatic drugs.	

**Table 2** ARTHUR joint scanning success rate by joint type

Joint type	Left-hand joint images obtained by ARTHUR (%)	Right-hand joint images obtained by ARTHUR (%)	Images thereafter declined by assessor (%)	Total successful joint images obtained, after assessor evaluation (%)
MCP1	86.67	96.67	10.81	81.67
MCP 2	100.00	96.67	5.08	93.33
MCP 3	100.00	100.00	1.67	98.33
MCP 4	96.67	86.67	5.45	86.67
MCP 5	90.00	76.67	6.00	78.33
IP 1	80.00	73.33	13.04	66.67
PIP 2	96.67	100.00	1.69	96.67
PIP 3	90.00	100.00	0.00	95.00
PIP 4	100.00	96.67	5.08	93.33
PIP 5	90.00	83.33	9.62	78.33
Wrist RCIC	90.00	83.33	17.31	71.67
<b>Overall</b>	<b>92.73</b>	<b>90.30</b>	<b>6.62</b>	<b>85.45</b>

IP 1, interphalangeal joint of thumb; MCP, metacarpophalangeal joints; PIP, proximal interphalangeal joints; RCIC, radiocarpal-intercarpal joint.

### Intra-robot agreement

The intra-rater agreement between the two ARTHUR scans is shown in [table 3](#).

ARTHUR V.2.0 repeatability showed a PEA of 63% for SH and 75% for Doppler assessment, with corresponding PCA of 93% and 94%. On a healthy versus inflamed joint level for SH, it was 88% and 91% for Doppler. For the GLOESS, combining SH and Doppler scores, PEA was 60%, PCA 92% and healthy versus inflamed joint of 85% agreement.

### ARTHUR and DIANA versus the rheumatologist

In [table 4](#), a comparison between ARTHUR V.2.0+DIANA V.2.0 and the rheumatologist (the ground truth) is presented.

The direct comparison between ARTHUR and the rheumatologist shows, for SH assessment, a PEA of 58% and PCA of 93%, with Doppler PEA of 77% and PCA of 93%. On the healthy versus inflamed joint assessment, it is 87% for SH and 91% for Doppler. For the GLOESS assessment, the agreement was similar, with a PEA of 56%, PCA 91% and healthy versus inflamed 85%. All kappa values showed fair to moderate agreement.

The agreement between the rheumatologist and ARTHUR+DIANA for each joint is shown in [table 5](#). To evaluate the classification performance of ARTHUR+DIANA against the rheumatologist's and the IA scoring, confusion matrices for SH and Doppler activity were constructed and are provided in online supplemental tables 2–5.

The highest agreement is seen on PIP 2, 3 and 4 for both SH, Doppler and GLOESS, while lower agreement was seen on MCP 1, 4, 5 and PIP 5, IP 1 and RCIC.

### Comparison with the IA

The rheumatologist and ARTHUR+DIANA showed comparable performance against the IA ([table 6](#)).

Comparison with the IA shows very similar results for both SH, Doppler and GLOESS, for both rheumatologist and ARTHUR+DIANA. For more detailed data on the distribution of SH and Doppler grading for the rheumatologist, ARTHUR+DIANA and the IA, respectively, can be seen in online supplemental tables 6 and 7.

The agreement and sensitivity between the IA and the performing rheumatologist and ARTHUR+DIANA are shown in [table 7](#).

**Table 3** Repeatability metrics of ARTHUR V.2.0 and DIANA V.2.0

Metric	SH (0–3)	Doppler (0–3)	GLOESS (0–3)
PEA ( $\pm 95\%$ CI)	63.05 (58.94 to 67.02)	74.78 (71.01 to 78.29)	60.25 (56.10 to 64.28)
PCA ( $\pm 95\%$ CI)	92.99 (90.58 to 94.95)	93.70 (91.38 to 95.55)	92.12 (89.60 to 94.19)
Binary ( $\pm 95\%$ CI)	88.09 (85.15 to 90.63)	90.54 (87.84 to 92.82)	85.29 (82.11 to 88.09)
Kappa ( $\pm 95\%$ CI)	0.54 (0.47 to 0.60)	0.49 (0.40 to 0.58)	0.55 (0.48 to 0.61)

GLOESS, Global OMERACT-EULAR Synovitis Scoring system; PCA, percent close agreement; PEA, percent exact agreement; SH, synovial hypertrophy.

**Table 4** Agreement between ARTHUR+DIANA and rheumatologist assessments

Metric	SH (0–3)	Doppler (0–3)	GLOESS (0–3)
PEA ( $\pm 95\%$ CI)	57.95 (53.90 to 61.92)	77.28 (73.72 to 80.57)	56.13 (52.06 to 60.13)
PCA ( $\pm 95\%$ CI)	92.88 (90.53 to 94.80)	94.20 (92.02 to 95.92)	91.23 (88.68 to 93.36)
Binary ( $\pm 95\%$ CI)	87.25 (84.33 to 89.81)	91.21 (88.66 to 93.35)	84.77 (81.65 to 87.54)
Kappa ( $\pm 95\%$ CI)	0.38 (0.31 to 0.45)	0.44 (0.35 to 0.54)	0.40 (0.33 to 0.47)

Binary agreement is defined as healthy (GLOESS $\leq 1$ ) versus diseased (GLOESS $\geq 2$ ).

GLOESS, Global OMERACT-EULAR Synovitis Score; PCA, percent close agreement; PEA, percent exact agreement; SH, synovial hypertrophy.

Comparing rheumatologist and ARTHUR individually with the IA joint image assessments shows similar results across SH, Doppler and GLOESS. The distribution of respectively SH and Doppler grading at a patient level for the rheumatologist, DIANA and the IA can be seen in online supplemental tables 8 and 9.

## DISCUSSION

ARTHUR V.2.0 demonstrated an overall success rate of 85% in producing IA-approved joint ultrasound images. Particularly for the clinically critical MCP 2–4 and PIP 2–4 joints in RA, an average success rate exceeding 95% was achieved. In contrast, lower success rates were observed for joints such as the IP 1 and RC/IC, likely reflecting anatomical complexity and positioning challenges inherent to these areas. Several factors can influence automated scanning performance, including patient positioning, joint deformities and imaging artefacts. Furthermore, the IA excluded an additional 5% of successfully scanned joints due to image quality concerns not detected by ARTHUR's internal quality control. While this highlights the need for continued refinement of the quality assessment algorithm, the system already produces interpretable scans

in the vast majority of relevant joints. Ongoing development will target these limitations to improve robustness and alignment with expert-level standards. The exclusion due to insufficient scan quality could theoretically bias the results. An exploratory analysis (online supplemental table 1) was performed, but no systematic relationship between scan failure and disease activity was found. However, improving the robustness and success rate of automated scanning remains a key focus in further development. Current research focuses on optimising scanning protocols and refining ARTHUR's system to enhance performance across all joint types.

Proper probe pressure and gel coupling are critical for acquiring high-quality joint ultrasound images, especially in small joints. During the development of ARTHUR, multiple optimisation cycles were conducted with active RA patients to ensure the system applies minimal, standardised pressure that does not suppress vascular signal. Additionally, ARTHUR provides clear instructions to the patient on how to apply sufficient gel to the hand before scanning. Moreover, we have previously shown that being scanned by ARTHUR is not associated with greater discomfort than manual scanning.<sup>7</sup>

**Table 5** Performance of ARTHUR+DIANA compared with the rheumatologist by joint type

Joint	SH			Doppler			GLOESS		
	PEA	PCA	Binary	PEA	PCA	Binary	PEA	PCA	Binary
MCP 1	50.9	80.0	74.5	78.2	90.9	89.1	49.1	80.0	72.7
MCP 2	62.7	98.3	88.1	78.0	96.6	93.2	61.0	98.3	88.1
MCP 3	61.7	96.7	86.7	85.0	96.7	96.7	61.7	95.0	88.3
MCP 4	50.9	94.5	92.7	69.1	90.9	87.3	40.0	89.1	80.0
MCP 5	50.0	94.0	86.0	68.0	90.0	86.0	40.0	94.0	76.0
PIP 2	67.8	96.6	91.5	88.1	96.6	94.9	69.5	96.6	93.2
PIP 3	73.7	98.2	96.5	80.7	96.5	96.5	75.4	96.5	96.5
PIP 4	72.9	94.9	91.5	88.1	96.6	96.6	74.6	94.9	94.9
PIP 5	51.9	96.2	94.2	88.2	96.1	96.1	51.9	96.2	94.2
IP 1	52.2	91.3	89.1	60.9	93.5	89.1	52.2	89.1	87.0
RC/IC	36.5	78.8	67.3	59.6	90.4	75.0	34.6	71.2	57.7

GLOESS, Global OMERACT-EULAR Synovitis Scoring system; IP 1, interphalangeal joint of thumb; MCP, metacarpophalangeal joints; PCA, percent close agreement; PEA, percent exact agreement; PIP, proximal interphalangeal joints; RCIC, radiocarpal-intercarpal joint; SH, synovial hypertrophy.

**Table 6** Performance metrics comparing rheumatologist and ARTHUR with DIANA, versus the independent assessor (IA)

Metric	Rheumatologist vs IA	ARTHUR+DIANA vs IA
SH (0–3)	Kappa: 0.46 (0.41 to 0.52)	Kappa: 0.39 (0.32 to 0.45)
	PEA: 51.52 (47.63 to 55.39)	PEA: 49.01 (44.95 to 53.07)
	PCA: 94.09 (92.01 to 95.76)	PCA: 91.23 (88.68 to 93.36)
	Sensitivity: 29.11 (22.17 to 36.86)	Sensitivity: 37.84 (30.00 to 46.17)
	Specificity: 99.80 (98.90 to 99.99)	Specificity: 93.64 (90.99 to 95.70)
	Binary agreement: 82.88 (79.78 to 85.68)	Binary agreement: 79.97 (76.55 to 83.09)
Doppler (0–3)	Kappa: 0.45 (0.38 to 0.52)	Kappa: 0.48 (0.41 to 0.55)
	PEA: 63.73 (59.93 to 67.41)	PEA: 62.58 (58.59 to 66.46)
	PCA: 94.08 (92.00 to 95.76)	PCA: 94.37 (92.22 to 96.07)
	Sensitivity: 34.44 (24.74 to 45.20)	Sensitivity: 41.86 (31.30 to 52.99)
	Specificity: 99.65 (98.74 to 99.96)	Specificity: 95.75 (93.64 to 97.32)
	Binary agreement: 90.74 (88.27 to 92.85)	Binary agreement: 88.08 (85.22 to 90.56)
GLOESS (0–3)	Kappa: 0.48 (0.42 to 0.53)	Kappa: 0.42 (0.35 to 0.49)
	PEA: 50.15 (46.27 to 54.03)	PEA: 47.68 (43.64 to 51.75)
	PCA: 91.97 (89.63 to 93.93)	PCA: 90.40 (87.76 to 92.63)
	Sensitivity: 30.48 (23.97 to 37.62)	Sensitivity: 42.04 (34.66 to 49.70)
	Specificity: 99.79 (98.83 to 99.99)	Specificity: 92.76 (89.88 to 95.03)
	Binary agreement: 80.15 (76.90 to 83.13)	Binary agreement: 77.98 (74.46 to 81.22)
GLOESS, Global OMERACT-EULAR Synovitis Scoring system; PCA, percent close agreement; PEA, percent exact agreement; SH, synovial hypertrophy.		

The IA excluded 44 joints due to insufficient image quality, which were not flagged by ARTHUR's internal quality assessment system. While these joints were still processed by DIANA, they were not included in the main analysis. Although ARTHUR incorporates an automated image quality control module, this finding suggests room for further refinement to better align with expert-level standards. Continued development of the quality control algorithm remains a key focus to ensure robust and reliable input for AI interpretation.

While robotic systems might be expected to exhibit near-perfect reproducibility due to their programmed nature, several factors can introduce variability in automated ultrasound scans. Minor variations in patient positioning between scans, even with positioning aids, can affect image acquisition. Additionally, the dynamic

nature of soft tissues and the presence of artefacts, such as patient movement, gel placement or probe pressure, can influence image quality and subsequent analysis.

Despite these challenges, ARTHUR V.2.0+DIANA V.2.0 demonstrated moderate repeatability, with a PEA of 63% for SH and 75% for Doppler, and PCA values of 93% and 94%, respectively. Binary agreement was 88% for SH and 91% for Doppler. While these PCA values suggest consistent performance, they should be interpreted with caution, as agreement within  $\pm 1$  on a 4-point scale can occur relatively easily by chance. The kappa values reflected moderate reproducibility, consistent with intraobserver variability previously reported among experienced rheumatologists.<sup>5</sup> Future improvements should focus on enhancing the precision and robustness of repeated assessments. Currently, ARTHUR employs

**Table 7** Agreement and sensitivity on a patient level (disease activity/remission)

Metric	Rheumatologist vs IA	ARTHUR+DIANA vs IA
SH	Patient agreement: 53.33 (34.33 to 71.66)	Patient agreement: 86.67 (69.28 to 96.24)
	Sensitivity: 53.33 (34.33 to 71.66)	Sensitivity: 86.67 (69.28 to 96.24)
Doppler	Patient agreement: 66.67 (47.19 to 82.71)	Patient agreement: 83.33 (65.28 to 96.36)
	Sensitivity: 56.52 (34.49 to 76.81)	Sensitivity: 91.30 (71.96 to 98.93)
GLOESS	Patient agreement: 60.00 (40.60 to 77.34)	Patient agreement: 86.67 (69.28 to 96.24)
	Sensitivity: 60.00 (40.60 to 77.34)	Sensitivity: 86.67 (69.28 to 96.24)
GLOESS, Global OMERACT-EULAR Synovitis Scoring system; IA, independent assessor; SH, synovial hypertrophy.		



continuous AI-driven analysis of real-time ultrasound images during joint sweeps to determine optimal probe positioning, returning to the point of highest image quality for acquisition of SH and Doppler still images and a Doppler video clip. Comparing these real-time sequences with previously recorded data from the same patient may further improve reproducibility by refining spatial targeting. Initial work in this direction is ongoing, but further validation and regulatory approval will be necessary before such functionality can be integrated into clinical workflows. At the patient level, the rheumatologist identified SH (SH $\geq$ grade 1) in at least one hand joint in 50% of patients, whereas the IA reported SH in all patients (100%), and ARTHUR+DIANA detected SH in 86.7%. This discrepancy may reflect differences in scoring thresholds for mild synovitis, as well as the use of the highest grade across the two image sets. This may partly reflect the limitations of scoring based on a single static image rather than dynamic, real-time scanning. As the assessor noted, “when only a single image is available, even minor anisotropy can be misinterpreted as effusion—something that could be clarified during live scanning but cannot be adjusted for when only one frame is available”. Notably, ARTHUR+DIANA tended to assign higher SH grades than the rheumatologist: only 44.2% of joints were classified as grade 0 by ARTHUR+DIANA, compared with 57.3% by the rheumatologist, while grade 3 was recorded in 6.0% versus 1.2%, respectively (online supplemental table 6). This pattern may indicate increased sensitivity of the automated system, but also a possible tendency to up-score borderline findings. Further refinements to DIANA’s SH scoring model are currently in development to improve its alignment with expert human interpretation.

In the current study, a Doppler and/or SH score over 1 (GLOESS over 1), was determined as inflammation. This was done as previous studies have indicated that the presence of Doppler activity with a score of 1 may be seen in normal joints, suggesting a higher PD cut-off of  $\geq 2$  as a sign of pathology.<sup>10–12</sup> Other studies have defined SH  $< 2$  and a Doppler of 0 as healthy. A future study will look at the impact of the selection criteria on already assessed RA and arthralgia patients.

The interpretation of agreement metrics must be considered in light of the data distribution. In our dataset, Doppler activity scores were highly skewed, with approximately 95% of joints scored as grade 0 or 1. This skewness inflates the apparent agreement on binary classification (healthy vs inflamed), as simply classifying most joints as normal would result in high binary accuracy by default.

This likely contributes to the discrepancy observed between binary agreement values ( $> 90\%$ ) and the more moderate kappa values ( $\sim 0.45$ ), which adjust for chance agreement and are therefore more conservative and robust in imbalanced datasets. For this reason, we included both unweighted and weighted agreement metrics (PEA, PCA, binary and kappa) to provide a

more comprehensive assessment of diagnostic consistency. These findings highlight the importance of using multiple complementary metrics when evaluating agreement between human and AI-based assessments in clinical imaging.

DIANA V.2.0 alone has previously been shown to perform at an expert rheumatologist’s level when compared with multiple specialists in the assessment of SH, Doppler activity and osteophytes.<sup>8</sup> Here, an extra layer of complexity is added as ARTHUR V.2.0 autonomously guides the patient and acquires the joint images and then sends them to DIANA V.2.0. In a study by Hammer *et al*, comparing five rheumatology ultrasound raters reported median (range) percentages of PEA for SH/Doppler assessments were 73.1 (70.3–80.6)/83.7 (76.7–87.6) and for PCA 98.1 (96.2–99.7)/98.0 (96.8–98.4).<sup>13</sup> So ARTHUR’s performance versus the rheumatologist in this trial is comparable to these results among human experts.

Due to the differences between ratings among human experts, it can be difficult to modify the AI due to differences among experts. Therefore, comparing the entire system with multiple experts who scan the same patients will be necessary in the future, as was done in the assessment of DIANA V.2.0.<sup>8</sup> Many important learnings have been taken from this study. If we compare the confusion matrices for SH and Doppler between the rheumatologist and ARTHUR/DIANA, we see DIANA scores SH higher than the human expert. This can be due to an overestimation on DIANA’s part, but can also be influenced by the image selected by ARTHUR in the sweep over the joint. If we look at table 5 we can see that compared with the rheumatologist, the joints with the lower PEA agreement are RCIC, MCP 4, 5 and IP 1, although they all have satisfactory PCA and binary agreement, across the SH, Doppler and GLOESS domains. Coming efforts will work on fine-tuning the assessment by DIANA of these joints, in combination with improving the image selection in the sweep function.

This study has several limitations. First, only joints successfully scanned by ARTHUR were included in the primary analysis, which may overestimate system performance and introduce selection bias. Although we assessed whether failed scans were associated with disease activity, further work is needed to reduce scan failure rates. Second, comparisons were made to a single expert rheumatologist, despite known variability in ultrasound scoring between experienced raters; this limits the generalisability of the observed agreement levels. Further ultrasound findings were not correlated with clinical joint inflammation assessed by palpation. Finally, patients with severe joint deformities were excluded due to current technical constraints in robotic scanning and because such joints fall outside the validated scope of the EULAR-OMERACT scoring systems. Future system development should address these challenges. An important limitation of the current AI model is the use of a combined score for the wrist joint (radiocarpal/intercarpal), in line with



the EULAR–OMERACT synovitis grading recommendations. While this is appropriate for standardised scoring exercises, future clinical implementation would benefit from separate scoring of individual wrist compartments, as these may carry distinct clinical relevance.

Importantly, this study was conducted in RA patients with established, longstanding disease. Additional studies in patients with early RA or undifferentiated arthritis are needed to evaluate system performance in these clinically critical populations.

Interestingly, despite all included patients having clinically verified arthritis in at least one joint, the expert rheumatologist identified inflammatory changes on ultrasound in only ~50% of cases. This may indicate a systematic underestimation of pathology, which could partially explain the modest agreement observed between ARTHUR+DIANA and the expert. Potential factors include differences in individual scoring thresholds, despite the use of EULAR-OMERACT definitions, and technical variation such as probe pressure. In contrast, the IA identified inflammatory pathology in a higher proportion of cases, highlighting the variability even among experienced readers and underscoring the need for prestudy calibration and consensus scoring to improve consistency.

The successful implementation of any new medical technology depends on acceptance by patients, physicians and hospital administrators. Previous research has demonstrated high patient acceptance of ARTHUR's assessments in RA.<sup>7</sup> Further investigation is needed to ensure widespread clinical adoption. While demonstrating performance equivalence to rheumatologists is essential, understanding how the system can be effectively integrated into routine care, whether by saving time, improving diagnostic accuracy or enhancing patient outcomes, will be crucial for its clinical and economic viability.

Further clinical research is now focused on further developing and validating ARTHUR V.2.0 and DIANA V.2.0 in larger, multi-centre studies with diverse patient and rheumatologist populations, including the EU Horizon-funded project AutoPIX.

In conclusion, the combination of fully automated ultrasound acquisition (ARTHUR V.2.0) and AI-based scoring (DIANA V.2.0) represents a strong step toward standardised, operator-independent ultrasound assessment. The system demonstrated repeatability on par with intra-expert agreement reported in the literature and showed encouraging agreement with human assessors, though further refinement is needed to optimise performance across specific joints.

#### Author affiliations

<sup>1</sup>Section of Rheumatology, Department of Medicine, Svendborg Hospital - Odense University Hospital, Svendborg, Denmark

<sup>2</sup>Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY), Diakonhjemmet Hospital, Oslo, Norway

<sup>3</sup>Center for Rheumatology and Spine Diseases, Rigshospitalet Glostrup, Glostrup, Denmark

<sup>4</sup>The Maersk Mc-Kinney Møller Institute, Syddansk Universitet, Odense, Denmark

<sup>5</sup>ROPCA, Odense, Denmark

**Acknowledgements** Thank you to all patients participating in this study and all personnel at the Section of Rheumatology, Svendborg Hospital for helping.

**Contributors** SAJ conceived and led the study and is the guarantor. BAF and SAJ coordinated data collection and study execution. HBH, LT and MA-D contributed to study design and clinical interpretation. TRS and ABHW developed and supported the AI and robotic systems. BAF, SAJ and ABHW performed the data analysis. SAJ drafted the initial manuscript. All authors critically revised the manuscript for intellectual content and approved the final version. During the preparation of this work, the author(s) used ChatGPT to improve readability and language in parts of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article. This statement is also written in the manuscript.

**Funding** The author(s) declare financial support was received for the research, authorship and/or publication of the article. SAJ is supported by a grant from the Region of Southern Denmark (21/17499).

**Competing interests** SAJ and TRS are cofounders of Ropca Aps, developing AI and producing the automated ultrasound scanning system called ARTHUR. ARTHUR's AI cannot currently assess osteophyte severity. AC is a full-time employee of Ropca Aps.

**Patient consent for publication** Not applicable.

**Ethics approval** This study involves human participants. All patients signed informed consent, and all trials were approved by the National Ethical Board, VMK, in compliance with Danish law (Case number: 2400033). Participants gave informed consent to participate in the study before taking part.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** No data are available. The data supporting this study cannot be shared with third parties, as data sharing was not included in the patient consent forms approved by the national ethics committee (VMK).

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

#### ORCID iDs

Hilde Berner Hammer <http://orcid.org/0000-0001-7317-8991>

Søren Andreas Just <http://orcid.org/0000-0002-3946-5919>

#### REFERENCES

- 1 Aletaha D, Smolen JS. Diagnosis and Management of Rheumatoid Arthritis: A Review. *JAMA* 2018;320:1360–72.
- 2 Nam JL, D'Agostino MA. Role of ultrasound imaging in individuals at risk of RA. *Best Pract Res Clin Rheumatol* 2017;31:71–9.
- 3 Combe B, Landewe R, Daïen CI, et al. 2016 update of the EULAR recommendations for the management of early arthritis. *Ann Rheum Dis* 2017;76:948–59.
- 4 D'Agostino M-A, 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.
- 5 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.
- 6 Battafarano DF, Ditmyer M, Bolster MB, et al. 2015 American College of Rheumatology Workforce Study: Supply and Demand Projections of Adult Rheumatology Workforce, 2015–2030. *Arthritis Care Res (Hoboken)* 2018;70:617–26.

- 7 Frederiksen BA, Schousboe M, Terslev L, *et al.* Ultrasound joint examination by an automated system versus by a rheumatologist: from a patient perspective. *Adv Rheumatol* 2022;62:30.
- 8 Aplin Frederiksen B, Berner Hammer H, Schultz Overgaard B, *et al.* Performance of an Artificial Intelligence Model Compared to Multiple Human Experts in Scoring Synovitis Severity and Osteophyte Severity on Joint Ultrasound Images [abstract]. *Arthritis Rheumatol* 2024;76.
- 9 Ventura-Ríos L, Hernández-Díaz C, Ferrusquia-Toriz D, *et al.* Reliability of ultrasound grading traditional score and new global OMERACT-EULAR score system (GLOESS): results from an inter- and intra-reading exercise by rheumatologists. *Clin Rheumatol* 2017;36:2799–804.
- 10 Padovano I, Costantino F, Breban M, *et al.* Prevalence of ultrasound synovial inflammatory findings in healthy subjects. *Ann Rheum Dis* 2016;75:1819–23.
- 11 Terslev L, Torp-Pedersen S, Qvistgaard E, *et al.* Doppler ultrasound findings in healthy wrists and finger joints. *Ann Rheum Dis* 2004;63:644–8.
- 12 Kitchen J, Kane D. Greyscale and power Doppler ultrasonographic evaluation of normal synovial joints: correlation with pro- and anti-inflammatory cytokines and angiogenic factors. *Rheumatology (Sunnyvale)* 2015;54:458–62.
- 13 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.