




Reliability of the Pediatric Specific Musculoskeletal Ultrasound Scoring Systems for the Elbow, Wrist, and Finger Joints

Patricia Vega-Fernandez¹ , Ysabella Esteban¹, Edward Oberle², Jean-Philippe Proulx-Gauthier³, Matthew Clark⁴, Susan Shenoi⁵ , Akaluck Thatayatikom⁶, Heather Benham⁷, Emily J. Brunner⁸, Leandra Woolnough⁹, Michael Henrickson¹, Laura R. Pratt¹⁰, Deirdre De Ranieri¹¹ , Sarah Hoffmann¹², Ginger Janow¹³, Hulya Bukulmez¹⁴, Mekibib Altaye¹⁵, Amy Cassidy¹⁵, Tracy V. Ting¹, and Johannes Roth¹⁶, on behalf of CARRA JIA Ultrasound Workgroup

ABSTRACT. *Objective.* Musculoskeletal ultrasound (MSUS) is increasingly being used in the evaluation of pediatric musculoskeletal diseases. In order to provide objective assessments of arthritis, reliable MSUS scoring systems are needed. Recently, joint-specific scoring systems for arthritis of the pediatric elbow, wrist, and finger joints were proposed by the Childhood Arthritis and Rheumatology Research Alliance (CARRA) MSUS workgroup. This study aimed to assess the reliability of these scoring systems when used by sonographers with different levels of expertise.

Methods. Members of the CARRA MSUS workgroup attended training sessions for scoring the elbow, wrist, and finger. Subsequently, scoring exercises of B mode and power Doppler (PD) mode still images for each joint were performed. Interreader reliability was determined using 2-way single-score intraclass correlation coefficients (ICCs) for synovitis and Cohen κ for tenosynovitis.

Results. Seventeen pediatric rheumatologists with different levels of MSUS expertise (1–15 yrs) completed a 2-hour training session and calibration exercise for each joint. Excellent reliability (ICC > 0.75) was found after the first scoring exercise for all the finger and elbow views evaluated on B mode and PD mode, and for all of the wrist views on B mode. After a second training session and a scoring exercise, the wrist PD mode views reached excellent reliability as well.

Conclusion. The preliminary CARRA MSUS scoring systems for assessing arthritis of the pediatric elbow, wrist, and finger joints demonstrate excellent reliability among pediatric MSUS sonographers with different levels of expertise. With further validation, this reliable joint-specific scoring system could serve as a clinical tool and scientific outcome measure.

Key Indexing Terms: diagnostic imaging, juvenile idiopathic arthritis, ultrasonography

The authors wish to acknowledge Childhood Arthritis and Rheumatology Research Alliance (CARRA) and the ongoing Arthritis Foundation financial support of CARRA. This project was funded by a CARRA–Arthritis Foundation Small Grant. PVF was supported by the Center for Clinical and Translational Science and Training at the University of Cincinnati, which is funded by the National Institutes of Health (NIH) Clinical and Translational Science Award program (grant 2UL1TR001425-05A1 and KL2 [2KL2TR001426-05A1]); the National Institutes of Arthritis and Musculoskeletal Skin Diseases (award no. P30AR076316); and the Diversity and Health Disparities Award, which is funded by Cincinnati Children's Hospital Medical. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

¹P. Vega-Fernandez, MD, MSc, Y. Esteban, MD, M. Henrickson, MD, MPH, T.V. Ting, MD, MSc, Department of Pediatrics, University of Cincinnati, Division of Rheumatology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA; ²E. Oberle, MD, Nationwide Children's Hospital, Columbus, Ohio, USA; ³J.P. Proulx-Gauthier, MD, FRCPC, Department of Pediatrics, CHU de Québec-Université Laval, Québec City, Québec, Canada; ⁴M. Clark, MD, Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, USA; ⁵S. Shenoi, MBBS, MS, Seattle Children's Hospital and Research Center University of Washington, Seattle, Washington,

USA; ⁶A. Thatayatikom, MD, AdventHealth for Children, Orlando, Florida, USA; ⁷H. Benham, DNP, APRN, Scottish Rite for Children Dallas, Dallas, Texas, USA; ⁸E.J. Brunner, DO, Geisinger Medical Center, Danville, Pennsylvania, USA; ⁹L. Woolnough, MD, MSCS, Department of Pediatrics, UFHealth, Gainesville, Florida, USA; ¹⁰L.R. Pratt, MD, University of Nebraska Medical Center, Omaha, Nebraska, USA; ¹¹D. De Ranieri, MD, Department of Pediatrics, Northwestern Feinberg School of Medicine, Division of Rheumatology, Ann and Robert H. Lurie Children's Hospital, Chicago, Illinois, USA; ¹²S. Hoffmann, MD, Children's Hospital of Richmond, Virginia, USA; ¹³G. Janow, MD, MPH, Joseph M. Sanzari Children's Hospital, Hackensack, New Jersey, USA; ¹⁴H. Bukulmez, MD, Department of Pediatrics, Division of Pediatric Rheumatology, Metro Health Medical System, Case Western Reserve University, Cleveland, Ohio, USA; ¹⁵M. Altaye, PhD, A. Cassidy, PhD, Department of Pediatrics, University of Cincinnati, Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA; ¹⁶J. Roth, MD, PhD, FRCPC, Children's Hospital Eastern Ontario, Ottawa, Ontario, Canada. Address correspondence to Dr. P. Vega-Fernandez, 3333 Burnet Ave, Cincinnati, OH 45229, USA. Email: patricia.vegafernandez@cchmc.org. Accepted for publication September 13, 2022.

Juvenile idiopathic arthritis (JIA) is a significant cause of morbidity worldwide.¹ Persistent joint inflammation can lead to functional limitations and lower health-related quality of life.¹ Clinical evaluation of JIA disease activity includes the assessment of active joint count, physician global assessment (PGA), parent/patient global assessment, presence and duration of morning stiffness, and biologic markers of inflammation.² While these variables are included in validated composite outcome measures such as the Juvenile Arthritis Disease Activity Score (JADAS),³ recent guidelines acknowledge the need for further standardization of the patient/parent global assessment and PGA.² The PGA can have poor interrater reliability among providers, particularly in patients with low disease activity or inactive disease.⁴ In addition, the reliability of active joint count is limited⁵ and cannot always adequately identify joints with synovitis.⁶

Musculoskeletal ultrasound (MSUS) is increasingly being used in children.⁷ It is well tolerated, readily available, and relatively inexpensive compared to other imaging modalities. Normal age-related findings and definitions of pediatric synovitis on MSUS have been developed.⁸⁻¹¹ MSUS can provide point of care information including the identification of subclinical disease.^{6,12,13} In order to provide objective assessments of arthritis, reliable scoring systems are necessary. They exist for rheumatoid arthritis,^{14,15} but in light of the unique sonographic features of the pediatric joint, specific MSUS scoring systems for JIA are needed. Our group recently proposed a joint-specific MSUS scoring system for the assessment of arthritis of the pediatric elbow, wrist, and finger joints that demonstrated excellent reliability when used by experienced ultrasonographers (> 7 years of experience).¹⁶ As MSUS use increases in pediatric rheumatology, reliable scoring systems for different levels of experience are needed. The objective of this study was to assess the interreader reliability of a B mode and power Doppler (PD) mode scoring system for arthritis of the pediatric elbow, wrist, and finger¹⁶ among sonographers with different levels of experience.

METHODS

Seventeen pediatric rheumatology providers who are members of the Childhood Arthritis and Rheumatology Research Alliance (CARRA) MSUS workgroup participated. All providers had prior formal training in pediatric MSUS with 1 to 15 years of subsequent clinical experience in pediatric MSUS. For the analysis, participants were divided into 2 groups: an expert group, defined as participants with > 5 years of experience in MSUS and > 10 MSUS studies per week in children (n = 5); and a nonexpert group (n = 12). This study was approved by the Cincinnati Children's Hospital Medical Center Institutional Review Board (approval no. 2018-7939). Written assent and consent to participate was obtained from all children whose images were used in the scoring exercises. Written informed consent for publication was obtained from all the study participants.

For each of the joints (elbow, wrist, and finger), participants received an initial 2-hour online virtual training session from an expert who had contributed to the preliminary CARRA scoring system (PVE, EO, JR).¹⁶ The training session included reviews of normal sonoanatomy, pathologic findings in JIA, the preliminary CARRA semiquantitative scoring system, and case-based examples of scoring including pitfalls. Participants then took part in a calibration exercise using still images. Through a subsequent debrief, any remaining questions were addressed.

The preliminary CARRA scoring system for the elbow, wrist, and finger joints consists of a semiquantitative grading from 0 to 3 (0 = normal or no pathology, 3 = severe pathology) for both B mode and PD mode images; tenosynovitis is assessed through a binary system with 0 (no pathology) or 1 (presence of pathology) in B mode and PD mode (Supplementary Table S1-S4, available with the online version of this article).¹⁶ Anonymized MSUS images of children aged 2 to 17 years were used, and the age of the patient was available to the participants.

Scoring exercises of both B mode and PD mode were completed for each joint. Interreader reliability was estimated using 2-way single-score intraclass correlation coefficients (ICC), a validated statistical measure of interreader reliability when variables in a study are rated by multiple coders.¹⁷ An ICC was considered excellent for values of 0.75 to 1.00, good 0.60 to 0.74, fair 0.40 to 0.59, and poor < 0.40.¹⁸ As a nominal variable, agreement for the scoring system of the extensor tendons of the wrist in transverse view was assessed using Cohen κ coefficient. κ values from 0.0 to 0.2 indicate slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement, and 0.81 to 1.0 almost perfect or perfect agreement.¹⁷ For those views that did not attain excellent reliability (ICC) or moderate agreement (κ) for all participants for the lower end of the 95% CI, the participants underwent a subsequent round of calibration and scoring exercises using a different set of B mode and PD mode images. The statistical analysis software used was SAS version 9.4 (SAS Institute).

RESULTS

A total of 300 still images were used for the first round of calibration and scoring exercises. These images were obtained in children aged 2 to 18 years, distributed equally across this age range, and including a broad number of images for each of the grade 0 to 3 categories (Supplementary Table S5, available with the online version of this article). Interreader reliability results for the entire group of raters are shown in the Table. In general, experts and nonexperts combined demonstrated excellent interreader agreement for all B mode and PD mode views of the elbow and finger joints as well as the distal radioulnar and midline radiocarpal views of the wrist. For all participants, the view of the radiocarpal joint in ulnar probe position reached excellent reliability in B mode but only good reliability in PD mode for the lower limit of the CI (excellent for the ICC itself). Images of the extensor tendons of the wrists demonstrated moderate agreement in B mode (κ criteria; see Methods section) but only fair agreement in PD mode for all participants. Given the overall excellent reliability for synovitis and moderate agreement for tenosynovitis in B mode, instead of proposing a modification of the scoring system, the PD mode scoring was repeated for the ulnar view of the radiocarpal joint and the extensor tendons following a second training session with special focus on the distinction of physiologic and pathologic findings. Regardless of level of expertise, excellent interrater reliability of 0.96 (95% CI 0.94-0.97) and moderate agreement of 0.72 (95% CI 0.61-0.83) were obtained, respectively, following the second training and scoring exercise. Separate results for the expert and nonexpert group are shown in Supplementary Table S6 and S7.

DISCUSSION

This study demonstrated excellent reliability of the preliminary semiquantitative CARRA MSUS scoring system for the pediatric elbow, wrist, and finger joints for providers with different levels of expertise. In addition, advanced MSUS concepts were

Table. Interreader reliability exercise for the pediatric elbow, wrist, and finger joints.

Joint		Exercise 1, n = 300 Images		Exercise 2, n = 28 images	
		ICC (95% CI) ^a			
		B Mode	PD Mode	B Mode	PD Mode
Elbow	Anterior humeroradial and humeroulnar joint recesses in longitudinal view	0.97 (0.96-0.98)	0.89 (0.84-0.92)	N/A	N/A
	Posterior humeroulnar joint recess in longitudinal view	0.96 (0.94-0.97)	0.90 (0.86-0.92)	N/A	N/A
Wrist	Distal radioulnar joint recess in transverse view	0.94 (0.92-0.96)	0.96 (0.94-0.97)	N/A	N/A
	Dorsal radiocarpal joint recess in midline longitudinal view	0.93 (0.91-0.94)	0.97 (0.96-0.98)	N/A	N/A
	Dorsal radiocarpal joint recess in ulnar longitudinal view	0.87 (0.81-0.93)	0.83 (0.61-0.91)	N/A	0.96 (0.94-0.97)
	Extensor tendons in transverse view ^b	0.67 (0.53-0.81)	0.51 (0.37-0.65)	N/A	0.72 (0.61-0.83)
Finger	MCP dorsal joint recess in longitudinal view	0.93 (0.91-0.97)	0.97 (0.96-0.98)	N/A	N/A
	MCP volar joint recess in longitudinal view	0.93 (0.91-0.94)	0.87 (0.84-0.90)	N/A	N/A
	PIP volar long joint recess in longitudinal view	0.96 (0.95-0.97)	0.89 (0.86-0.91)	N/A	N/A
	PIP dorsal long joint recess in longitudinal view	0.96 (0.94-0.97)	0.95 (0.93-0.97)	N/A	N/A

^a Intraclass correlation coefficient (ICC) was based on a 2-way random effects model for a single measure. ICC was defined as excellent 0.75-1.00, good 0.60-0.74, fair 0.40-0.59, and poor < 0.40.¹⁸ ^b κ values of 0.0-0.2 indicate slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1.0 almost perfect or perfect agreement.¹⁷ ICC: intraclass correlation coefficient; MCP: metacarpophalangeal joint; N/A: not applicable; PD: power Doppler; PIP: proximal interphalangeal joint.

successfully taught in a virtual format. By demonstrating the reliability of this semiquantitative measurement instrument, our study supports the potential use of MSUS scoring systems as an objective outcome measure at bedside and in further research studies.

Several MSUS scoring systems have been published in recent years.^{16,19-22} Only a few of these scoring systems have been evaluated with sonographers of variable experience. The pediatric MSUS scoring system of the knee proposed by the CARRA group was tested in pediatric rheumatology providers from the CARRA JIA ultrasound (US) working group (n = 16) with < 1 to 10 years of US experience. This exercise demonstrated good to excellent reliability for B mode and PD mode views.²² Most recently, Rossi-Semerano et al²³ reported the reliability of the Outcome Measures in Rheumatology (OMERACT) pediatric US synovitis scoring system among 13 pediatric ultrasonographers of diverse subspecialty backgrounds: 9 rheumatologists, 2 pediatricians, and 2 radiologists with varying degrees of experience. This group used a total of 75 images to evaluate the reliability of the most representative view of the wrist, elbow, metacarpophalangeal (MCP) II, knees, and ankle joints. For the scoring systems of the MCP II, wrist, and elbow, they found fair to good reliability for the B mode and excellent reliability for PD mode. However, the scoring system used was not joint-specific, which the authors acknowledged.²³ Our study used a larger sample of normal and pathologic images (n = 300), including all views recommended in the evaluation of JIA and involved 17 pediatric rheumatology sonographers with different levels of expertise. Our joint-specific scoring system is based on 1-plane view per area, using the most representative view to capture pathology, with any abnormal findings confirmed in a second plane.

Since all views reached excellent reliability¹⁸ after the second training session, the differences in the ICC noted after the first scoring exercise may have resulted from variability in the

experience level of the participants interpreting MSUS rather than intrinsic issues with the scoring system. The lower ICC levels and 95% CI range in the MCP and proximal interphalangeal (PIP) volar views in PD mode for the expert group was a function of the number of positives rating being small. Reliability coefficients assume an equal distribution of positive and negative findings.²⁴ The cut-offs used for differentiating the various levels of agreement (poor to excellent) were based on Cicchetti.¹⁸ Other authors (Koo and Li²⁵) have proposed slightly higher cut-offs. It is important to note that no universal agreement exists on how to define these levels. The ICC ranges between 0.00 and 1.00, with values closer to 1.00 representing stronger reliability. Given that most of our ICC values are 0.90 or above, with the lower end of the range being very close to it, the results suggest very good reliability independent of the specific cut-offs used. We also considered the lower end of the 95% CI for the decision on whether agreement was good enough; we did not base this decision simply on the ICC value itself.

The value of training sessions, which include a review of the definitions of key MSUS findings in healthy children and in children with arthritis, normal sonoanatomy, the scoring system, sonographic images with pathology, and a calibration session, was demonstrated. Participant feedback following the first and second training exercises noted that careful review of the normal spectrum of sonographic findings related to the degree of skeletal maturation was most helpful. Given the in-person meeting limitations imposed by the coronavirus disease 2019 (COVID-19) pandemic, this project was conducted in an online virtual setting. A major benefit of the change in format included the opportunity for the recording of the sessions and the possibility to reach out to a larger group. The session recordings were used by participants unable to attend the virtual meetings. The excellent reliability reached in this project supports the use of an online virtual format as an effective method for pediatric MSUS training, including training for scoring exercises.

Additional exercises of the reliability of the proposed MSUS scoring systems in real-time US images among different patient age groups, for instance, through patient-based exercise, may follow. Future studies will also need to assess the construct and predictive validity of this preliminary MSUS scoring system.

In conclusion, a novel MSUS scoring system for B mode and PD mode of the pediatric elbow, wrist, and finger showed excellent reliability among pediatric rheumatology ultrasonographers with varying levels of expertise. This was supported by an in-depth virtual training format. This joint-specific scoring system for pediatric arthritis could serve as a clinical and scientific outcome measure, following further refinement and validation.

ONLINE SUPPLEMENT

Supplementary material accompanies the online version of this article.

REFERENCES

1. Palman J, Shoop-Worrall S, Hyrich K, McDonagh JE. Update on the epidemiology, risk factors and disease outcomes of juvenile idiopathic arthritis. *Best Pract Res Clin Rheumatol* 2018;32:206-22.
2. Ringold S, Angeles-Han ST, Beukelman T, et al. 2019 American College of Rheumatology/Arthritis Foundation guideline for the treatment of juvenile idiopathic arthritis: therapeutic approaches for non-systemic polyarthritis, sacroiliitis, and enthesitis. *Arthritis Rheumatol* 2019;71:846-63.
3. Consolaro A, Ruperto N, Bazso A, et al. Development and validation of a composite disease activity score for juvenile idiopathic arthritis. *Arthritis Rheum* 2009;61:658-66.
4. Taylor J, Giannini EH, Lovell DJ, Huang B, Morgan EM. Lack of concordance in interrater scoring of the provider's global assessment of children with juvenile idiopathic arthritis with low disease activity. *Arthritis Care Res* 2018;70:162-6.
5. Guzman J, Burgos-Vargas R, Duarte-Salazar C, Gomez-Mora P. Reliability of the articular examination in children with juvenile rheumatoid arthritis: interobserver agreement and sources of disagreement. *J Rheumatol* 1995;22:2331-6.
6. Vega-Fernandez P, Oberle E, Henrickson M, et al. Correlation of subclinical synovitis with juvenile idiopathic arthritis outcome measurements [abstract]. *Arthritis Rheumatol* 2021;73 Suppl 9.
7. Windschall D, Malattia C. Ultrasound imaging in paediatric rheumatology. *Best Pract Res Clin Rheumatol* 2020;34:101570.
8. Collado P, Vojinovic J, Nieto JC, et al; Omeract Ultrasound Pediatric Group. Toward standardized musculoskeletal ultrasound in pediatric rheumatology: normal age-related ultrasound findings. *Arthritis Care Res* 2016;68:348-56.
9. Roth J, Jousse-Joulin S, Magni-Manzoni S, et al; Outcome Measures in Rheumatology Ultrasound Group. Definitions for the sonographic features of joints in healthy children. *Arthritis Care Res* 2015;67:136-42.
10. Roth J, Ravagnani V, Backhaus M, et al; OMERACT Ultrasound Group. Preliminary definitions for the sonographic features of synovitis in children. *Arthritis Care Res* 2017;69:1217-23.
11. Collado P, Windschall D, Vojinovic J, et al. Amendment of the OMERACT ultrasound definitions of joints' features in healthy children when using the Doppler technique. *Pediatr Rheumatol Online J* 2018;16:23.
12. Janow GL, Panghaal V, Trinh A, Badger D, Levin TL, Ilowite NT. Detection of active disease in juvenile idiopathic arthritis: sensitivity and specificity of the physical examination vs ultrasound. *J Rheumatol* 2011;38:2671-4.
13. De Lucia O, Ravagnani V, Pregnotato F, et al. Baseline ultrasound examination as possible predictor of relapse in patients affected by juvenile idiopathic arthritis (JIA). *Ann Rheum Dis* 2018;77:1426-31.
14. 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.
15. Bruyn GAW, Siddle HJ, Hanova P, et al. Ultrasound of subtalar joint synovitis in patients with rheumatoid arthritis: results of an OMERACT reliability exercise using consensual definitions. *J Rheumatol* 2019;46:351-9.
16. Vega-Fernandez P, Ting TV, Oberle EJ, et al; CARRA Musculoskeletal Ultrasound Workgroup. Musculoskeletal Ultrasound in Childhood Arthritis Limited Examination: a comprehensive, reliable, time-efficient assessment of synovitis. *Arthritis Care Res* 2021 Jul 30 (Epub ahead of print).
17. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol* 2012;8:23-34.
18. Cicchetti DV. Multiple comparison methods: establishing guidelines for their valid application in neuropsychological research. *J Clin Exp Neuropsychol* 1994;16:155-61.
19. Collado P, Naredo E, Calvo C, et al; ECO-JIA Study Group. Reduced joint assessment vs comprehensive assessment for ultrasound detection of synovitis in juvenile idiopathic arthritis. *Rheumatology* 2013;52:1477-84.
20. Vojinovic J, Magni-Manzoni S, Collado P, et al. SAT0636 Ultrasonography definitions for synovitis grading in children: the Omeract pediatric ultrasound task force [abstract]. *Ann Rheum Dis* 2017;76:1015.
21. Sande NK, Boyesen P, Aga AB, et al. Development and reliability of a novel ultrasonographic joint-specific scoring system for synovitis with reference atlas for patients with juvenile idiopathic arthritis. *RMD Open* 2021;7:e001581.
22. Ting TV, Vega-Fernandez P, Oberle EJ, et al; Childhood Arthritis and Rheumatology Research Alliance Juvenile Idiopathic Arthritis Ultrasound Workgroup. Novel ultrasound image acquisition protocol and scoring system for the pediatric knee. *Arthritis Care Res* 2019;71:977-85.
23. Rossi-Semerano L, Breton S, Semerano L, et al. Application of the OMERACT synovitis ultrasound scoring system in juvenile idiopathic arthritis: a multicenter reliability exercise. *Rheumatology* 2021;60:3579-87.
24. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257-68.
25. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155-63.