Development of CROMRIS (ChRonic nonbacterial Osteomyelitis MRI Scoring) Tool and Evaluation of its Interrater Reliability

Yongdong Zhao, MD, PhD[1,2], T. Shawn Sato, MD[3], Sabrina M. Nielsen, MSc[4,5], Meinrad Beer, MD[6], Mingqian Huang, MD[7], Ramesh S. Iyer, MD, MBA[8], Michael McGuire, MD[9], Anh-Vu Ngo, MD[8], Jeffrey P. Otjen, MD[8], Jyoti Panwar, MD, FRCR[10,11], Jennifer Stimec, MD[10], Mahesh Thapa, MD, MEd[8], Paolo Toma, MD[12], Angela Taneja, MD[13], Nancy E. Gove, PhD[2], Polly J. Ferguson, MD[14]

[1] Seattle Children's Hospital, Department of Pediatrics, University of Washington, Seattle, WA, USA

[2] Center for Clinical and Translational Research, Seattle Children's Research Institute, Seattle, WA, USA

[3] Department of Radiology, University of Iowa Carver College of Medicine, Iowa City, IA, USA

[4] Musculoskeletal Statistics Unit, The Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark

[5] The Rheumatology Research Unit, Department of Rheumatology, Odense University Hospital and University of Southern Denmark, Denmark

[6] Department of Diagnostic and Interventional Radiology, University Hospital of Ulm, Ulm, Germany

[7] Department of Radiology, Stony Brook University Hospital, Stony Brook, NY, USA

[8] Department of Radiology, Seattle Children's Hospital, University of Washington, Seattle, WA, USA

[9] Department of Radiology, Hackensack University Medical Center, Hackensack, NJ, USA

[10] Department of Medical Imaging, Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada

[11] Department of Radiology, Christian Medical College and Hospital, Vellore, India

[12] Department of Imaging, Bambino Gesù Children Hospital, IRCCS, Rome, Italy

[13] Pediatric Rheumatology, Children's Healthcare of Atlanta, Emory University, Atlanta, GA, USA

[14] Department of Pediatrics, University of Iowa Carver College of Medicine, Iowa City, IA, USA

Corresponding author:

Yongdong Zhao, MD, PhD

MA 7.110, 4800 Sand Point Way NE,

Seattle, WA 98105

Key words: Chronic nonbacterial osteomyelitis, chronic recurrent multifocal osteomyelitis, MRI, scoring tool, interrater reliability, CROMRIS

Running title: MRI scoring for CNO.

Total word count: 3,490

Abstract (word limit 250, current 249)

Background/Purpose: Serial MRI exams are often needed in chronic nonbacterial osteomyelitis (CNO) to determine the objective response to treatment. Our objectives were: 1) to develop a consensus-based MRI scoring tool for clinical and research use in CNO; 2) to evaluate interrater reliability and agreement using whole body (WB) MRI from children with CNO.

Method: Eleven pediatric radiologists discussed definitions and grading of signal intensity, size of signal abnormality within bone marrow and associated features on MRI through monthly conference calls and a consensus meeting, using a nominal group technique in July 2017. WB MRI scans from children with CNO were de-identified for training reading and an interrater reliability study. The reading by each radiologist was conducted in a randomized order. Interrater reliability for abnormal signal and severity were assessed using free-marginal kappa statistics.

Results: Radiologists reached a consensus on grading CNO-specific MRI findings and on describing bone units based on anatomy. A total of 45 sets of WB MRI scans, including four sets of non-CNO MRI exams, were selected for the final reading. The mean kappa of each category of bones was >0.7 with

majority >0.9 demonstrating substantial/almost perfect interrater reliability of readings among radiologists. The agreement on signal intensity and the size of signal abnormality within most commonly affected bones, femur and tibia was lower than that of other bones.

Conclusion: The CROMRIS tool, a comprehensive standardized scoring tool for MRI in children with CNO, was developed. Our interrater study demonstrated good interrater reliability and agreement of readings.

This study was approved by University of Iowa Internal Review Board, approval number 201609778.

INTRODUCTION

Chronic nonbacterial osteomyelitis (CNO) is a pediatric autoinflammatory bone disease challenging to physicians because of its occult nature and difficulty to assess disease activity. It is also known as chronic recurrent multifocal osteomyelitis (CRMO) and synovitis, acne, pustulosis, hyperostosis, and osteitis (SAPHO). Physical exam and traditional inflammatory markers are not sensitive metrics to monitor disease progression due to occasionally minimal or absent findings on physical examination, normal laboratory values, and lack of correlation between them(1). Radiographs are only 13-16% sensitive in detecting skeletal lesions in CNO(2) and bone scintigraphy was shown to be only 70% sensitive compared to magnetic resonance imaging (MRI)(3). The current gold standard imaging modality is whole body (WB) MRI(2,4,5), especially at the initial evaluation. However, the imaging findings of CNO can be non-specific and bone biopsy may be necessary.

CNO can affect virtually any bone, and there is no uniform approach to assess all bones identically. Previously, CNO lesions on MRI were reported by the number of active lesions(5–10) and their anatomical locations. Recently, detailed scoring systems were reported (11,12). WB MRI has the dual advantages of greater sensitivity and lack of ionizing radiation when compared to skeletal scintigraphy(3), and is more commonly used in pediatric rheumatology across the world(2,4,13,14). Standardized reporting of each imaging characteristic across all

bones within patients with CNO is critical in establishing imaging outcome measurements in CNO for future studies. Our objective is to develop a practical and consensus-based MRI scoring tool for clinical and research use in CNO. Furthermore, interrater agreement and reliability will be evaluated using WB MRI from children with CNO.

MATERIALS AND METHODS

The development of the ChRonic nonbacterial Osteomyelitis MRI Scoring (CROMRIS) tool consisted of 3 steps: 1) Literature review of previously reported MRI scoring tools of CNO, 2) Initial development of a standardized MRI scoring tool for CNO, 3) Consensus meeting. Subsequently, the interrater agreement and reliability were assessed.

We did a literature review on previously reported MRI scoring tools of CNO as preparation for the meetings. The results of the review were detailed in the result section and presented at the conference call meetings and consensus conference. Members of an international CNO MSK radiologist working group initiated the process to develop a standardized MRI scoring tool for CNO at the Society of Pediatric Radiology (SPR) annual conference in Vancouver, Canada, in 2015. Since the first meeting, eleven pediatric radiologists, each with at least five years of experience reading musculoskeletal and CNO MRI from seven different pediatric hospitals in North America and Europe, were identified through soliciting pediatric radiologists within the CNO work group. Group members

discussed definitions and grading of signal intensity, size of signal abnormality within bone marrow and surrounding tissue, physis damage and vertebral compression on MRI through monthly conference calls. Representative MRI images (STIR sequence except that skull used T2 sequence from 1.5T or 3T scanner) of active bone inflammation were assembled by members using a separate set of images to establish an atlas to illustrate the proposed scoring system.

Consensus meeting

At the face-to-face conference (Seattle, July 2017), seven radiologists and two pediatric rheumatologists (YZ, PJF) attended the meeting. The facilitators, YZ and PJF) participated in the discussion but were not eligible to vote. Nominal group technique was used to achieve consensus (defined as ≥70% agreement within the group) on all questions considered during the meeting.

Interrater agreement and reliability

The interrater agreement and reliability study was approved by the institutional review board (IRB) from Iowa Children's Hospital (# 201609778). Written informed consent was waived due to the retrospective nature and usage of anonymized images. A total of eighty-two sets of pre-existing WB MRI scans (STIR sequence with 3-4 mm thickness from 1.5T or 3T scanner) between January 2013 and August 2016 from children with CNO or other diseases at the University of Iowa Children's Hospital were used for training reading and for

assessing interrater agreement and reliability. A video tutorial was produced for training and interrater calibration exercise. Nine sets of MRI examinations were used for the training reading to improve familiarity of the tool before reliability study. Of the total 82 sets of MRIs, four sets from subjects older than 18, nine sets for training and one set from a patient with leukemia were excluded. Among the remaining 68 sets of MRI from 45 patients (19 patients had MRIs at more than one time point), 45 sets of MRI examinations from 45 patients (limit one set per patient and the set at the beginning of the disease course if more than one set is available), including four sets of MRI studies from non-CNO patients, were used for assessing interrater agreement and reliability for each radiologist to read in a randomized order. Controls were included in the analyses to ensure variability in the sample. Data were recorded with detailed scoring form (**Supplement 1**). There was no gold standard defined for comparisons.

Statistical analysis

For the interrater agreement and reliability study, descriptive analysis was performed to assess the prevalence of abnormalities at each site defined as agreement among >70% of the radiologists. Data were presented combining similar types of bones per patient. Absolute agreement for each site was defined as the proportion of patients for whom the ratings were the same for all 11 radiologists.

We assessed interrater reliability, i.e., how well the persons can be distinguished from each other despite measurement errors, using the free-marginal kappa

statistic described by Brennan and Prediger(15). The free-marginal kappa statistic is recommended when raters are not instructed about the number of observations that should be assigned to each category(15) and when the distribution of ratings is highly skewed(16). The kappa coefficients were interpreted according to Landis and Koch(17). Mean kappa (and range) was calculated by categories of bones: The spine, complex bone, flat bones, hand/foot, and long bones. The long bones were further divided into proximal epiphysis, proximal metaphysis, diaphysis, distal metaphysis and distal epiphysis. All analyses were conducted using R version 3.5.1(18).

RESULTS

Literature review

Search was conducted in PubMed using the following MeSH terms: (SAPHO[All Fields] OR "chronic recurrent multifocal osteomyelitis"[All Fields] OR "chronic nonbacterial osteomyelitis"[All Fields] OR "non-bacterial osteitis"[All Fields]) AND ("magnetic resonance imaging"[MeSH Terms] OR ("magnetic"[All Fields] AND "resonance"[All Fields] AND "imaging"[All Fields]) OR "magnetic resonance imaging"[All Fields] OR "mri"[All Fields]) AND (Score[All Fields] OR scoring[All Fields]). Five peer-reviewed publications were identified and one (4) was excluded because it did not mention a scoring system. A total of three separate tools were reported in the remaining four eligible articles. Two reported an MRI score system for the osteitis lesions ranged from zero to two points and the

highest score among lesions was used to indicate disease severity in SAPHO(19,20). Bone marrow edema, bone erosions or synovitis (with or without joint effusion) were ascertained. The presence of only one finding was scored 1 point and two or more findings 2 points. A second tool used a semi-quantitative approach to evaluate the characteristics of CNO lesions from MRI in children(11). A comprehensive grading system for the evaluation of the extent of bone edema and soft tissue inflammation, as well as the presence or absence of periosteal reaction, hyperostosis, physeal damage and vertebral compression were reported (11). A third tool, radiologic index for WB MRI in patients with NBO (RINBO), defined the size of active lesions by the absolute measurements and clustered the number of active lesions into 3 categories as unifocal, paucifocal (2-4 lesions) and multifocal (5 or more lesions)(12). Soft tissue inflammation, periosteal reaction and hyperostosis were classified as extramedullary findings and spinal involvement was distinguished between active with abnormal STIR signal and chronic with deformation. Surrounding soft tissue inflammation was not included. Points were assessed for each of four parameters of interest (number of radiologic active lesions [RAL], maximum size of RAL, extramedullary affection and spine involvement) with a maximum score of 10.

Typical WB MRI protocols include coronal images of the entire body, and sagittal images of entire spine, acquired with fluid-sensitive sequence (short tau inversion recovery [STIR]; turbo inversion recovery magnitude [TIRM]; or fat saturation) without contrast. Axial sequences of the pelvis and knees, and sagittal images of

the ankles and feet, were also included in one of the centers (Iowa) that enhanced lesion identification in commonly affected sites. This protocol was therefore adopted by the group with consensus. T1-weighted images have been used to confirm findings from fluid-sensitive sequence in CNO. It was considered optional as it adds scanning time. Diffusion-weighted imaging (DWI) was reported (21), however, not routinely performed in participating institutions. A recent study did not show difference in sensitivity of differentiating CNO lesions between STIR sequence alone and combining T1-weighted, DWI and STIR sequences(22). Thus, T1-weighted and DWI sequences were not included in the scoring system but use of DWI should be reconsidered when more data is available on its use in CNO. Detailed discussion based on the reported scoring tools led to the newly developed tool.

The consensus process of final CROMRIS tool

At the face-to-face conference (Seattle, July 2017), consensus defined as ≥70% agreement within the group (23) was reached on all questions considered during the meeting. The complete atlas developed following the consensus meeting includes evaluation of 20 sites using four different variables (**Supplement 2**).

Inclusion and definition of various characteristics of MRI findings in CNO

As presented in **Table 1**, hyperintensity of bone marrow was defined as increased STIR signal within bone marrow compared to the nearby normal marrow, as per the interpreting radiologist's assessment. Terminology of "bone

edema" was discussed and replaced by "bone marrow hyperintensity" with consensus for scientific clarity and the uncertainty of pathology. Linear metaphyseal lines caused by bisphosphonate were included in the atlas to avoid misinterpretation as bone marrow hyperintensity. Periosteal reaction was deemed difficult to confirm by MRI whereas soft tissue inflammation was readily detectable. Thus, "hyperintensity of surrounding tissue" was included with consensus to report the presumed inflammation within soft tissue and periosteum. "Hyperostosis" was a common term used in radiograph though identifiable on MRI as "bony expansion". Thus the latter term was adopted by the group. Vertebral compression and joint effusion were included. Growth plate irregularity was discussed and voted not suitable for assessment in MRI with consensus. Kyphosis and limb hypertrophy were assessable in WB MRI and thus included in this tool. Leg length discrepancy cannot be assessed reliably in MRI and thus was voted not to be included as part this tool. None of the above parameters was assigned as acute or chronic at this stage because a prospective longitudinal study is required to distinguish among them.

Grading scale of parameters and definition of bone units

In general, signal intensity of bone marrow was graded with three levels: absent, less than fluid signal, and similar to fluid signal. Confidence level of identifying abnormal signal was also recorded as low, medium, or high. The size of signal intensity within each unit/segment was graded using relative measurement because of various body size and bone size in affected patients. Small was

defined as <25% of estimated volume, medium as 25-50% of estimated volume and large as >50% of estimated volume. When imaging was inadequate for a confident estimate of the size, “unable to estimate the size” was recorded. The following parameters were graded as present or absent: signal hyperintensity of surrounding tissue (soft tissue/periosteum), bony expansion, continuity of signal abnormality between diaphysis and adjacent segment in long bones, hypertrophy of limbs, signal intensity of posterior and/or lateral elements in spine, and kyphosis of entire spine. Vertebral compression was graded as normal, presence of some height loss, or plana (defined as complete flattening of a vertebral body).

The division of bone units and segments was discussed, and the consensus was to follow anatomical divisions in complex bones and group bones into one unit in less commonly affected sites (hands and fore/mid foot) and less well visualized sites. Long bones were divided into the following five segments anatomically: proximal epiphysis, proximal metaphysis, diaphysis, distal metaphysis, and distal epiphysis. The spine was graded as individual vertebra from cervical to lumbar region. However, in addition to the grading of anterior vertebral body, abnormal signals within “lateral and posterior elements” including pedicles, lamina and posterior processes were also reported. Based on existing literature, the prevalence of signal hyperintensity within metatarsal bones is less common than in the talus and calcaneus(24). Therefore, the consensus was to grade any signal hyperintensity within metatarsal bones as abnormal, and only signal hyperintensity with confluence in talus or calcaneus as abnormal.

Total scores as reported by RINBO (12) were not recommended because our first step was to describe and grade lesions from each individual bone unit reliably. Future studies will be needed to determine the exact weight of each characteristic using a much larger representative cohort.

Interrater agreement and reliability

The 45 subjects were mainly females with a median age of 11 (interquartile range [IQR], 9-15) and a median disease duration of approximately 3.3 years (**Table 2**). About 80% of WB MRIs were collected with additional axial images of pelvis and knees, and sagittal images of ankles/feet, in addition to the coronal plane images of the entire body and sagittal sequences of entire spine, as done in 20% of subjects. The 11 raters were mainly from USA with median years of experience of 7 years (IQR, 6-10) (**Supplement 3**).

Lower extremities were most commonly affected bones by CNO with abnormal bone marrow signal (**Figure 1**). Upper extremities, including humerus, radius and hand, were reported at 2-9% presence among these patients. Among spine, thoracic spine is the most commonly affected site. Pelvic bones, clavicle and mandible were well represented. Lesions in cervical spine, manubrium/sternum, rib, scapula, skull, ulna were absent within this cohort. Hyperintensity within surrounding tissue was detected adjacent to tibia, femur, fibula, foot, humerus, periacetabulum, clavicle, and mandible. Bony expansion was present only in

femur, humerus, clavicle and mandible. Vertebral compression was mostly present in the thoracic spine. Detailed data from individual bone units (i.e., left femur, right mandible) are available in **Supplement 4**.

The signal intensity of bone marrow hyperintensity had low absolute agreements (<60%) in more commonly affected bones such as femur, tibia, fore/mid foot, hindfoot, and clavicle (**Figure 2A**). The majority of less-commonly affected bones, including the spine, pelvis, hands, scapula, patella and radius, had near or greater than 80% of absolute agreements**.** The presence of hyperintensity within surrounding tissue and bony expansion agreed very well (>80%) in all bones (**Figure 2B, 2C**). Detailed data from individual bone unit were available in **Supplement 5**. Most segments of femur and tibia had lower agreement for the size of bone marrow hyperintensity compared to other long bones (**Figure 3A**). Among other bones, all had good absolute agreement (>80%) except for the clavicle, mandible, fore/mid foot and hindfoot (**Figure 3B**). The severity of vertebral compression assessed by radiologists has shown excellent absolute agreement in all patients (**Figure 3C**).

The mean kappa of each category was >0.7 with majority >0.9 demonstrating substantial/almost perfect reliability (**Table 3**). The lowest kappa coefficient was observed in bone marrow hyperintensity for tibia (right, 0.60, 95%CI, 0.49-0.71) and the corresponding absolute agreement was only 29% (**Supplement 6**). Spine, complex bones (pelvis), flat bones had higher agreements in bone marrow

hyperintensity than hands/foot and long bones. The signal size of bone marrow hyperintensity within each category agreed perfectly though hands/foot and proximal/distal metaphysis of long bones had the lowest kappa scores. The reliability of presence of hyperintensity within surrounding tissue, presence of bony expansion and vertebral compression were all almost perfect. Detailed data from individual bone units are available in **Supplement 6**. Joint effusion data showed excellent agreement (**Supplement 6**). Most of low- and medium-confidence readings were from more commonly affected sites such as femur, tibia, and foot (**Supplement 7**).

DISCUSSION

This is the first consensus-based MRI scoring tool for children with CNO and the first comprehensive assessment of interrater reliability of such a tool. Our tool includes the most commonly described characteristics seen in children with CNO from MRI and the grading system can be used as a potential research tool after further development and validation. An atlas and training video were developed that may guide radiologists who are less familiar or less experienced in reporting MRI from these affected children.

We have further defined these parameters and developed a semi-quantitative scoring system as an assessment tool for longitudinal studies to measure the response to treatments. Comparing to RINBO system (12), our scoring tool included bone marrow hyperintensity (bone edema), size of bone lesion,

vertebral compression, bony expansion (hyperostosis). Several key differences between these two tools are: 1) periosteal reaction was deemed not reliable by our group in consensus so not included in current tool; 2) the size of lesion was reported in current tool as relative to the bone which is more appropriate for pediatric population; 3) a total score was not proposed because further studies are needed to determine the weight of each variable.

Defining the minimum abnormal signal is challenging because of individual scoring variation, as suggested by the low absolute agreement of signal hyperintensity in the commonly affected bones (tibia and femur). Therefore, we used a predefined 70% agreement as a threshold to determine if a “true” abnormal signal existed in bone marrow. Based on this principle, we found similar distribution of lesions among entire skeleton as previous reports(9,10,25,26). Abnormal signal within surrounding tissue and bony expansion were present at most long bones, but was uncommonly seen in the clavicle and mandible.

In addition, the absolute agreement of the intensity of signal abnormality was poor in commonly affected sites suggested that individual radiologist differ in their assessing of various levels either due to inadequate calibration/training or inherent challenge from defined classification. Most low- and medium-confidence readings were from commonly affected sites. These results suggested that adding mandated calibration exercise with a special focus on less conspicuous

lesion might improve the interrater agreement. In contrast, the absolute agreements of abnormal signal in surrounding tissue and bony expansion were >80% except for tibia. Although the prevalence of these findings were less common than that of bone marrow hyperintensity, it was likely that these features were more distinguishable by radiologists and thus more agreeable among radiologists.

Kappa analysis showed moderate to substantial agreement on the MRI size readings of most commonly affected bones (tibia and femur). When grouped into large categories such as long bones, or spine, the agreement significantly increased which was likely due to the relatively fewer abnormal signals. Hands and feet were scored as region by grouping multiple bones and the size of bone marrow hyperintensity may not be estimated well enough. It explained why the agreement of this variable is the lowest among all categories. Similarly, the signal size of bone marrow hyperintensity of proximal and distal metaphyses of long bones was also the least agreed due to the difficulty of clearly identifying the border/definition of this segment within long bones. These are very helpful observations that will allow further improvements of our scoring system. Future studies will aim to answer the following questions: 1) is this scoring tool sensitive to change of clinical disease activities in CNO from a longitudinal study and what is the intrarater reliability; 2) what is the interrater reliability of this tool in a validation cohort; 3) how to integrate scores from each body site as a total score for disease activity on a whole-body level and whether this score can differentiate

CNO patients from non-CNO patients.

There were limitations of our study. Firstly, even with our large sample size of subjects, some bone sites were not well represented for interrater study. Future studies using a different subset of MRIs with enriched prevalence of signal abnormalities in less commonly affected sites and inadequately scanned area (i.e. upper extremities) are needed to validate our findings. Secondly, joint effusion was not adequately scored but due to its complexity and less weight in managing these patients, we decided that this should be a separate effort. Thirdly, there was no gold standard of the abnormal signals identified by radiologists for this study. A more objective approach of identifying signal threshold is needed and may be accomplished through machine learning by creating a consensus reading result. Fourthly, lower agreement and reliability may have been obtained due to unequal familiarity of the tool despite the training. Fifthly, even with radiologists from 7 centers, this consensus may not be completely representative. Lastly, the correlation of abnormal signals on MRI and the actual pathology from CNO was not confirmed. Therefore, a longitudinal study with detailed clinical characterization in children with CNO and healthy children may shed light on the clinical significance of these parameters. Nevertheless, we accomplished a comprehensive MRI scoring tool for CNO with a consensus from experienced radiologists across 7 centers and 2 continents and showed excellent reliability and agreements in each category of bones and moderate to substantial reliability and agreements in readings from individual

bones.

CONCLUSION

The CROMRIS tool, a comprehensive standardized scoring tool for MRI in children with CNO, was developed. Our interrater study demonstrated good interrater reliability and agreement of readings from a group of radiologists. Because CNO is a rare disease and collaborative research is needed in this field, a consensus-based system, as the CROMRIS tool, representing experienced radiologists from different centers and countries, will likely be adopted by future studies. This tool can be validated in a prospective study and may become a key element of disease activity assessment in CNO.

Author contribution

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Zhao had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design: Ferguson, Sato, Zhao.

Analysis and interpretation of data: Gove, Nielsen, Zhao.

Critical review of the manuscript: all authors.

ACKNOWLEDGEMENT

The authors would like to thank other physician participants who contributed in initial conference calls: Nancy Chauvin, Kirsten Ecklund, and Andrea Doria. This study was funded by a CARRA-Arthritis Foundation small grant. Polly Ferguson is supported by R01AR059703 from NIH/NIAMS. Yongdong Zhao is supported by Clinical Research Scholar Program from Seattle Children's Research Institute, Bristol-Myers Squibb. The Parker Institute, Bispebjerg and Frederiksberg Hospital (Sabrina M. Nielsen) is supported by a core grant from the Oak Foundation (OCAY-13-309).

REFERENCES

1. Zhao Y, Ferguson PJ. Chronic Nonbacterial Osteomyelitis and Chronic Recurrent Multifocal Osteomyelitis in Children. Pediatr Clin North Am 2018;65:783-800.
2. Fritz J, Tzaribatchev N, Claussen CD, Carrino JA, Horger MS. Chronic recurrent multifocal osteomyelitis: comparison of whole-body MR imaging with radiography and correlation with clinical and laboratory data. Radiology 2009;252:842-51.
3. Morbach H, Schneider P, Schwarz T, Hofmann C, Raab P, Neubauer H, et al. Comparison of magnetic resonance imaging and Technetium-labelled methylene diphosphonate bone scintigraphy in the initial assessment of chronic non-bacterial osteomyelitis of childhood and adolescents. Clin Exp Rheumatol 2012;30:578-82.

4. Voit AM, Arnoldi AP, Douis H, Bleisteiner F, Jansson MK, Reiser MF, et al. Whole-body magnetic resonance imaging in chronic recurrent multifocal osteomyelitis: Clinical longterm assessment may underestimate activity. J Rheumatol 2015;42:1455-62.

5. Roderick M, Shah R, Finn A, Ramanan A V. Efficacy of pamidronate therapy in children with chronic non-bacterial osteitis: disease activity assessment by whole body magnetic resonance imaging. Rheumatology (Oxford) 2014;53:1973-6.

6. Beck C, Morbach H, Beer M, Stenzel M, Tappe D, Gattenlöhner S, et al. Chronic nonbacterial osteomyelitis in childhood: prospective follow-up during the first year of anti-inflammatory treatment. Arthritis Res Ther 2010;12:R74.

7. Hospach T, Langendoerfer M, von Kalle T, Maier J, Dannecker GE. Spinal involvement in chronic recurrent multifocal osteomyelitis (CRMO) in childhood and effect of pamidronate. Eur J Pediatr 2010;169:1105-11.

8. Miettunen PM, Wei X, Kaura D, Reslan WA, Aguirre AN, Kellner JD. Dramatic pain relief and resolution of bone inflammation following pamidronate in 9 pediatric patients with persistent chronic recurrent multifocal osteomyelitis (CRMO). Pediatr Rheumatol Online J 2009;7:2.

9. Wipff J, Costantino F, Lemelle I, Pajot C, Duquesne A, Lorrot M, et al. A large national cohort of French patients with chronic recurrent multifocal osteitis. Arthritis Rheumatol (Hoboken, NJ) 2015;67:1128-37.

10. Jansson A, Renner ED, Ramser J, Mayer A, Haban M, Meindl A, et al.

Classification of non-bacterial osteitis: Retrospective study of clinical, immunological and genetic aspects in 89 patients. Rheumatology 2007;46:154-60.

11. Zhao Y, Chauvin NA, Jaramillo D, Burnham JM. Aggressive Therapy Reduces Disease Activity without Skeletal Damage Progression in Chronic Nonbacterial Osteomyelitis. J Rheumatol 2015;42:1245-51.
12. Arnoldi AP, Schlett CL, Douis H, Geyer LL, Voit AM, Bleisteiner F, et al. Whole-body MRI in patients with Non-bacterial Osteitis: Radiological findings and correlation with clinical data. Eur Radiol 2017;27:2391-9.
13. Guérin-Pfyffer S, Guillaume-Czitrom S, Tammam S, Koné-Paut I. Evaluation of chronic recurrent multifocal osteitis in children by whole-body magnetic resonance imaging. Joint Bone Spine 2012;79:616-20.
14. Darge K, Jaramillo D, Siegel MJ. Whole-body MRI in children: current status and future applications. Eur J Radiol 2008;68:289-98.
15. Brennan RL, Prediger DJ. Coefficient Kappa: some uses, misuses, and alternatives. Educ Psychol Meas 1981;41:687-99.
16. Quarfoot D, Levine RA. How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution? Am Stat 2016;70:373-84.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
18. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013
19. Assmann G, Kueck O, Kirchhoff T, Rosenthal H, Voswinkel J,

Pfreundschuh M, et al. Efficacy of antibiotic therapy for SAPHO syndrome is lost after its discontinuation: an interventional study. Arthritis Res Ther 2009;11:R140.

20. Jung J, Molinger M, Kohn D, Schreiber M, Pfreundschuh M, Assmann G. Injection into Sternocostoclavicular Joints in Patients with SAPHO Syndrome. Semin Arthritis Rheum 2012;42:266-70.
21. Leclair N, Thörmer G, Sorge I, Ritter L, Schuster V, Hirsch FW. Whole-Body Diffusion-Weighted Imaging in Chronic Recurrent Multifocal Osteomyelitis in Children. PLoS One 2016;11:e0147523. doi: 10.1371/journal.pone.0147523.
22. Merlini L, Carpentier M, Ferrey S, Anooshiravani M, Poletti P, Hanquinet S. Whole-body MRI in children: Would a 3D STIR sequence alone be sufficient for investigating common paediatric conditions? A comparative study. Eur J Radiol 2017;88:155-62.
23. Orbai A, Wit M De, Mease P, Shea JA, Gossec L, Leung YY, et al. International patient and physician consensus on a psoriatic arthritis core outcome set for clinical trials. Ann Rheum Dis 2017;76:673-80.
24. Shabshin N, Schweitzer ME, Morrison WB, Carrino JA, Keller MS, Grissom LE. High-signal T2 changes of the bone marrow of the foot and ankle in children: Red marrow or traumatic changes? Pediatr Radiol 2006;36:670-6.
25. Borzutzky A, Stern S, Reiff A, Zurakowski D, Steinberg E a, Dedeoglu F, et al. Pediatric Chronic Nonbacterial Osteomyelitis. Pediatrics 2012;130:e1190-7.

26. Roderick MR, Shah R, Rogers V, Finn A, Ramanan A V. Chronic recurrent multifocal osteomyelitis (CRMO) – advancing the diagnosis. Pediatr Rheumatol Online J 2016;14:47.

Figure legends

Figure 1, Mean prevalence of bone lesions based on >70% agreement among 11 radiologists on the presence of bone marrow hyperintensity (HI), surrounding tissue hyperintensity, bony expansion within entire skeleton.

Figure 2, Mean absolute agreement of bone lesions among all 11 radiologists on the signal intensity of bone marrow hyperintensity (BMH) (A), presence of surrounding tissue inflammation (B), presence of bony expansion (C).

Figure 3, Mean absolute agreement of the signal size of bone marrow hyperintensity (BMH) in each segment of long bones (A) and other bones (B). Absolute agreement of the severity of vertebral compression (C).

Supplement 1. Blank recording form of CNO MRI scoring

Supplement 2. Complete atlas of representative CNO bone lesions from each individual bone

Supplement 3. Reader's characteristics (n=11)

Supplement 4. Detailed data of the prevalence of abnormal bone marrow hyperintensity, surrounding tissue hyperintensity, and bony expansion from individual bone unit

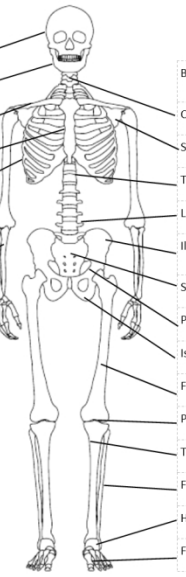
Supplement 5. Detailed data of the absolute agreement of abnormal bone marrow hyperintensity, surrounding tissue hyperintensity, and bony expansion from individual bone unit

Supplement 6. Kappa results of signal intensity of bone marrow hyperintensity, presence of surrounding tissue hyperintensity, presence of bony expansion,

signal size of bone marrow hyperintensity, presence of vertebral compression and joint effusion

Supplement 7. Descriptive data of readings with various confidence levels

Bony expansion	Surrounding tissue HI	Bone marrow HI	Bone
0%	0%	0%	Skull
2%	2%	2%	Mandible
2%	2%	9%	Clavicle
0%	0%	0%	Manubrium/ sternum
0%	0%	0%	Rib
2%	2%	9%	Humerus
0%	0%	7%	Radius
0%	0%	0%	Ulna
0%	0%	2%	Hand



Bone	Bone marrow HI	Surrounding tissue HI	Bony expansion
C spine	0%	0%	0%
Scapula	0%	0%	0%
T spine	7%	0%	0%
L spine	2%	0%	0%
Ilium	4%	0%	0%
Sacrum	9%	0%	0%
Periacetabulum	9%	2%	0%
Ischium/pubis	2%	0%	0%
Femur	40%	9%	7%
Patella	2%	0%	0%
Tibia	58%	11%	0%
Fibula	18%	7%	0%
Hindfoot	22%	0%	0%
Fore/mid foot	22%	2%	0%

Figure 1

279x215mm (150 x 150 DPI)

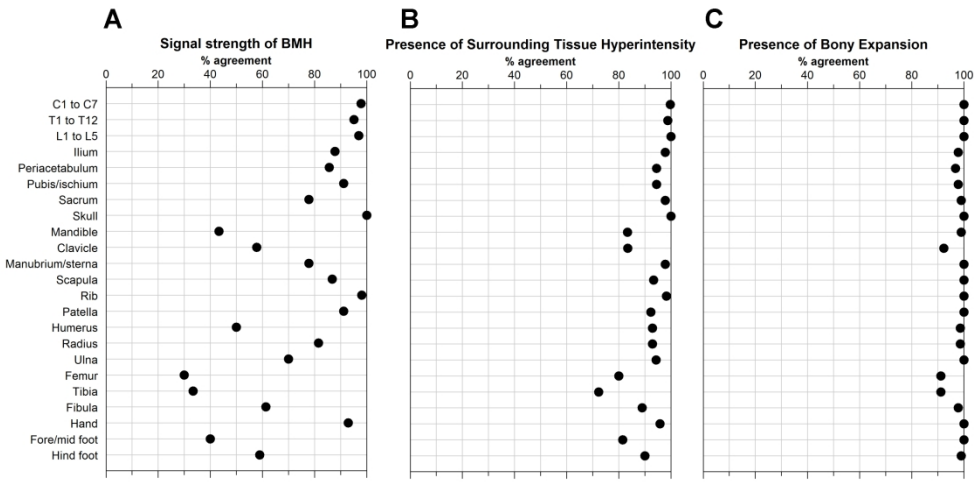


Figure 2.

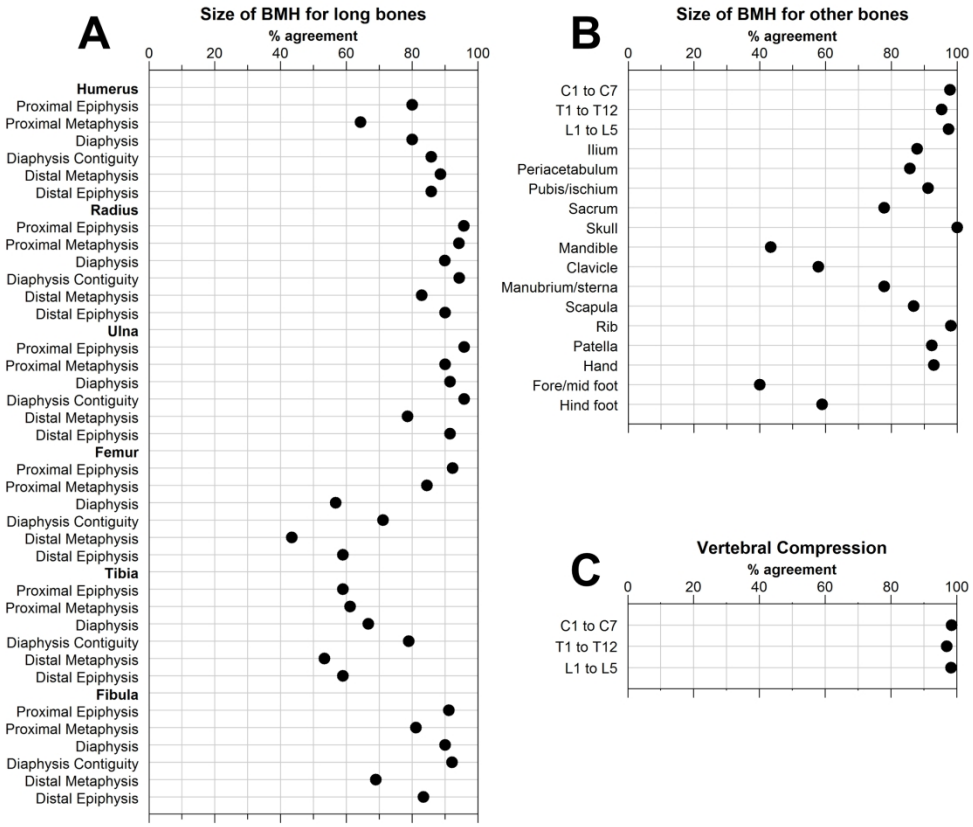


Figure 3.

219x190mm (300 x 300 DPI)

Table 1. Definition of parameters in CROMRIS tool

Parameter	Definition
Hyperintensity within bone marrow	Presence of abnormal signal intensity within bone marrow on a fluid-sensitive sequence* with confluence pattern, per radiologist's discretion
Hyperintensity within surrounding tissue	Presence of abnormal signal intensity other than a normal luminal structure (i.e., bladder, intestine, cerebrospinal fluid space, vasculature) within surrounding tissue on a fluid-sensitive sequence*
Bony expansion	Enlarged bone contour that is greater than expected
Joint effusion	More than physiological amount of joint fluid within a joint space
Vertebral compression	Decreased height comparing to the adjacent vertebra
Limb hypertrophy	Abnormally increased size of the limb comparing to contralateral side
Kyphosis	An exaggerated outward curve of spine on sagittal view

*STIR, fat saturation or TIRM

Table 2. Patient characteristics*

Variables	**Patients (n=45)** Median (IQR) or number (frequency)
Age, years	11 (9-15)
Height, cm†	148 (134-167)
Weight, kg†	41 (31-65)
Female, n (%)	31 (69)
CNO diagnosis‡, n (%)	41 (91)
Duration of disease†, months	40 (16-56)
Basic WB MRI§, n (%)	10 (22)
Complete WB MRI¶, n (%)	35 (78)

CNO, chronic nonbacterial osteomyelitis; IQR, interquartile range; WB, whole body.

*Data are presented as medians (IQR) unless stated otherwise.

† Data were only available for 44 patients on height and weight, and for 41 patients on duration of disease.

‡Other 4 patients had recurrent fever (1), juvenile idiopathic arthritis (1) and unknown conditions (2) at the time of MRI.

§Basic WB MRI protocol includes STIR sequence of coronal plane of entire body (upper extremities excluded) in 4-5 stations, sagittal plane of entire spine in 2 stations.

¶Complete WB MRI protocol added STIR sequence of axial plane of pelvis and knees, coronal plane of upper extremities as well as sagittal plane of ankles to the basic WB MRI protocol.

Table 3. Kappa results of MRI readings among radiologists for each category of bones

	Signal intensity of BMH	**Signal size of BMH**	**Presence of Surrounding Tissue Hyperintensity**	**Presence of Bony Expansion**	**Vertebral Compression**
	Mean (range)	*Mean (range)*	*Mean (range)*	*Mean (range)*	*Mean (range)*
Spine	0.98 (0.96-0.99)	0.98 (0.96-0.99)	0.99 (0.98-0.99)	NA*	0.98 (0.97-0.99)
Complex bone	0.93 (0.89-0.98)	0.94 (0.91-0.96)	0.97 (0.96-0.98)	0.98 (0.97-0.99)	-
Flat bone	0.93 (0.72-0.99)	0.95 (0.87-0.99)	0.97 (0.91-0.99)	0.96 (0.93-0.99)	-
Hand/foot	0.80 (0.63-0.94)	0.83 (0.67-0.97)	0.94 (0.92-0.98)	0.99 (0.99-0.99)	-
Long bone	0.75 (0.60-0.90)	-	0.92 (0.76-0.99)	0.97 (0.93-0.99)	-
Proximal Epiphysis	-	0.95 (0.87-0.99)	-	-	-
Proximal Metaphysis	-	0.89 (0.78-0.96)	-	-	-
Diaphysis	-	0.94 (0.89-0.97)	-	-	-
Diaphysis Contiguity	-	0.91 (0.80-0.98)	-	-	-
Distal Metaphysis	-	0.81 (0.65-0.96)	-	-	-
Distal Epiphysis	-	0.92 (0.85-0.96)	-	-	-

BMH: bone marrow hyperintensity.

* Kappa could not be calculated due to full agreement for all subsets within the spine.

Spine includes cervical, thoracic and lumbar vertebrae. Complex bone refers to pelvis which was divided into ilium, periacetabulum, pubis/ischium and sacrum on each side. Flat and irregular bones included skull, mandible, clavicle, sterna/manubrium, ribs, patella and scapula. Hands were graded as one unit including phalanges, metacarpal and carpal bones on each side. Feet were divided into Fore/mid foot and hind foot. Fore/mid foot included phalanges, metatarsal and tarsal bones. Hind foot included talus and calcaneus.