Development and Validation of a Deep Learning Model for Scoring of Radiographic Finger Joint Destruction in Rheumatoid Arthritis

Toru Hirano1, Masayuki Nishide1, Naoki Nonaka2, Jun Seita2, Kosuke Ebina3, Kazuhiro Sakurada2, Atsushi Kumanogoh1

1Department of Respiratory Medicine and Clinical Immunology, Internal Medicine, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita, Osaka, Japan
2Medical Sciences Innovation Hub Program, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, Japan
3Department of Musculoskeletal Regenerative Medicine, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita, Osaka, Japan

Correspondence to: Toru Hirano, Associate Professor, Department of Respiratory Medicine and Clinical Immunology, Internal Medicine, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita, Osaka, 565-0871, Japan
E-mail: thirano@imed3.med.osaka-u.ac.jp
Abstract

Objectives: The purpose of this research was to develop a deep-learning model to assess radiographic finger joint destruction in rheumatoid arthritis (RA).

Methods: The model comprises two steps: joint-detection step and joint-evaluation step. Among 216 radiographs of 108 patients with RA, 186 radiographs were assigned to training/validation dataset and 30 were assigned to test dataset. In the training/validation dataset, images of proximal interphalangeal joints (PIP), interphalangeal joint of the thumb (IP) or metacarpophalangeal joints (MCP) were manually clipped and scored for joint space narrowing (JSN) and bone erosion by clinicians, and then these images were augmented. As a result, 11160 images were used to train and validate a deep convolutional neural network (CNN) for joint evaluation. Selected 3720 images were used to train machine learning for joint detection. These steps were combined as the assessment model for radiographic finger joint destruction. Performance of the model was examined using the test dataset, which were not included in the training/validation process, by comparing scores assigned by the model and clinicians.

Results: The model detected PIP, IP and MCP with a sensitivity of 95.3% and assigned scores for JSN and erosion. Accuracy (percentage of exact agreement) reached 49.3–65.4% for JSN and 70.6–74.1% for erosion. The correlation coefficients between scores by the model and clinicians per image were 0.72–0.88 for JSN and 0.54–0.75 for erosion.

Conclusion: Image processing with the trained CNN model is promising to assess radiographs in RA.

Keywords: rheumatoid arthritis, joint destruction, artificial intelligence

Key messages: The CNN-based deep-learning can be applied to develop a model for assessing hand radiographs. The model assesses joint space narrowing and bone erosion of the fingers of rheumatoid arthritis. This AI technology leads to more extensive and detailed evaluation of joints in future.
Introduction

Artificial intelligence (AI) is effectively used in a wide range of fields, including autonomous vehicles, translation, speech recognition, image processing, natural language processing, art and medicine [1]. In the medicine, processing of histopathological image or radiograph, mining of genomic data, screening for molecular target and analysis of clinical big data are considered suitable tasks for deep-learning in which multiple layers of neuron-like nodes mimic how human brains analyze information [2]. AI trained with deep-learning such as convolutional neural network (CNN), which introduces robustness to variations of images, can deliver an outstanding performance in classifying various images into significant categories [3]. Examples include grading of diabetic retinopathy [4], classification of skin cancer [5, 6], prediction of lung cancer mutations [7] and classification of interstitial lung disease [8].

The management of rheumatoid arthritis (RA) has progressed dramatically over the past several decades. With the use of numerous efficacious drugs, more than half of patients have achieved low disease activity or clinical remission, which include less joint pain, less joint swelling, lower serum inflammatory markers and better global health based on patients’ self-assessment. Disease activity and other factors such as progression of structural joint damage are considered to make treatment decisions [9]. For evaluating structural joint damage, radiograph has been the gold-standard.

Radiographic classification of RA or systemic arthritis was first proposed by Steinbrocker et al. [10]. Several methods for scoring radiographic joint damage in RA were proposed by Kellgren et al. [11], Sharp et al. [12, 13], Larsen et al. [14] or Genant [15]. Later, the Sharp/van der Heijde method for scoring radiographs of hands and feet was developed [16]. This method has been widely used especially in clinical studies. However, in clinical settings, this method is not commonly used because it requires a high level of skill and the differences between examiners are considerable.

In this research, we attempt to develop a model for scoring of radiographic finger joint destruction in RA. Our model comprises two steps as shown in Figure 1A. The first step is a detection of joints by machine learning (cascade classifier using Haar-like features). The second step is a scoring of joint destruction by deep-learning (CNN), which comprises convolutional layers, pooling layers and fully connected layers. CNN processes input image as
2-dimentional matrix data and gives output as numerical values or probability of categories. CNN is currently considered as the most efficient algorithm for image processing. We examine the performance of our model by comparing scores assigned by the model and those by rheumatologists using radiographs that are not included in the CNN training/validation process.

Methods
Patients and images
Digital Anterior-Posterior (AP) radiographs of front bilateral hands of 108 patients with RA, were collected retrospectively. Patients were diagnosed as RA according to the 1987 Rheumatoid Arthritis Classification by the American Rheumatism Association [17] or the 2010 American College of Rheumatology (ACR) – European League Against Rheumatism Collaborative Initiative (EULAR) Classification Criteria [18]. All patients were treated at Osaka University Hospital and were enrolled in the institute’s cohort of RA patients. Clinical information such as sex, age, disease duration and clinical laboratory values were collected from the medical charts. This research was approved by the ethics committee of Osaka University Hospital and was conducted in accordance with the Declaration of Helsinki.

Among 216 radiographs of the 108 patients, we used 186 radiographs of the 93 patients for training/validation, and 30 radiographs of the 15 patients for test of the trained model (Figure 1B). From the 186 radiographs, areas of proximal interphalangeal joint (PIP), interphalangeal joint of the thumb (IP) or metacarpophalangeal joint (MCP) were manually clipped with the use of a graphical software, and 1860 clipped images were generated. The images were gray-scaled with the resolution ranging from 40 x 40 pixels to 80 x 80 pixels. The degree of joint destruction was scored by the consensus among 2 rheumatologists with 10 or 15 years of experience in rheumatology according to the Sharp/van der Heijde method [19]. The scores consist of joint space narrowing (JSN) score and bone erosion score. Briefly, JSN score is defined as follows: 0, no joint space narrowing; 1, focal or doubtful; 2, generalized, >50% of the original joint space left; 3, generalized, <50% of the original joint space left or subluxation; and 4, bony ankylosis or complete luxation. Erosion score is defined as follows: 0, no erosion; 1, discrete; 2, larger, <50% of the joint surface; 3, extend over the middle of the bone; and 5, complete collapse. The erosion score of a single joint is calculated as the sum of each score in the joint with a maximum score of 5.
Detection of joints by machine learning

The first step of the machine learning was to detect finger joints. The finger joints were detected by cascade classifier using Haar-like features [20]. The classifier was trained to detect finger joints such as PIP, IP and MCP. The clipped 1860 images from 186 radiographs were augmented by horizontal flipping, and a total of 3720 images was used to train the classifier (Figure 1B). For training the classifier and applying joint detection, Open Source Computer Vision Library, Open-CV (version 3.4, Intel Corporation, Santa Clara, CA, USA) was used.

Scoring of joint destruction by CNN

The second step of the machine learning was to assign JSN score and erosion score to each joint detected at the previous step. The collected original radiographs were split into three sets, namely, train, validation and test dataset, followed by the standard practice in machine learning (Figure 1B). Each dataset contained 146, 40 and 30 radiographs respectively. Radiographs contained in the train dataset were used to tune the parameters of the CNN and radiographs contained in the validation dataset were used to monitor the performance of the CNN model during the training process. After the training process, radiographs in the test dataset were used to evaluate the trained CNN model by comparing scores assigned by the model and clinicians. In the training process, radiographs in the train and validation dataset were augmented by horizontal flipping and/or rotation (+10° or −10°). As a result, we obtained a total of 8760 and 2400 images of PIP, IP or MCP for the train and validation dataset respectively. Subsequently, obtained images were resized to 48 x 48 pixels and offered to the CNN model. The CNN model comprises 2 convolution layers (filter size: 3, padding size: 1, stride: 1, activation function: rectified linear unit (ReLU)), 2 pooling layers (filter size: 2, stride: 2, max pooling) and 3 fully connected layers with one hidden layer (512 nodes, activation function: ReLU, rate of dropout: 0.5) (Figure 1C). The loss function was set to softmax cross entropy, and the optimization algorithm was set to adaptive moment estimation (Adam) [21]. Batchsize of the training was set to 512. Batch normalization was introduced in the CNN for erosion score based on the preliminary experiment [22]. Output was given as probability of each JSN class or each erosion class, and the class with the highest probability was determined. The Open Source Library for Neural Networks, Chainer (version 5.1, Preferred Networks, Tokyo, Japan) was used for an implementation of the CNN model [23].
Test of the model
To test the performance of the trained model, we assessed the consistency of judgements between the model and two clinicians. One of the two clinicians (clinician 2) was officially trained for scoring of joint destructions. Thirty radiographs in the test dataset were used (Figure 1B). The numbers of each JSN class or erosion class assigned by the model or clinicians were counted, and the distributions of score were compared. Percentage of exact agreement (PEA) which is identical to accuracy, and percentage of close agreement (PCA) which is within 1.0 score difference at the joint level were assessed. Sensitivity and specificity (score 0 versus ≥1) were also assessed. Total score of a radiograph for JSN or erosion was calculated as the sum of each JSN score or erosion score of PIP, IP and MCP. Correlations between total scores assigned by the model and those by clinicians were assessed using Pearson’s correlation coefficients.

Results
Patients and images
The characteristics of the patients are shown in Table 1. Among the patients, 90 (83.3%) were female. The median and the interquartile range of age were 64.9 [53.5, 72.6] years old, and those of disease duration were 12.2 [6.4, 17.6] years. All participants were diagnosed as RA. Seropositivity of anti-cyclic citrullinated peptide antibody (ACPA) was 67.6%. In the training/validation dataset, the distribution of the 1860 clipped images by joint was 744 for PIP, 186 for IP and 930 for MCP. The JSN score and the erosion score assigned by clinicians are summarized in Table 2. Scores for intercarpal joints were not summarized because the model trained by the machine learning could not detect many of them.

Detection of joints
Figure 2 shows representative images processed by the model, which detected finger joints and then assigned scores of joint destructions. Finger joints such as PIP, IP and MCP were identified as red rectangles by the model. Figure 2A shows the whole hand image processed by the model. In this image, 4 joints of PIP, one of IP and 5 of MCP were correctly detected. DIP joints, some intercarpal joints and wrist were also detected, however many intercarpal joints were not identified correctly.
Scoring of joint destruction

In Figure 2, the number on the upper-left corner of the rectangle indicates JSN score assigned by the model (yellow letter) and that on the lower-right indicates erosion score assigned by the model (blue letter). In Figure 2B, an enlarged image shows joints with JSN score 0, 2, 3 or 4, and erosion score 0, 4, or 5. In Figure 2C, another enlarged image shows joints with JSN score 0, 2 or 4, and erosion score 0, 3 or 5. The accuracy (PEA) of scoring during the training/validation process of the CNN increased continuously with epoch, which is the number of repetitions of the training (Figure 3A for JSN and 3C for erosion). The loss, which is the discrepancy between the score assigned by the CNN and the score determined by clinicians, decreased with epoch (Figure 3B for JSN and 3D for erosion). The training of the CNN was stopped at epoch 40 for JSN and at epoch 110 for erosion, when the values of loss were minimum respectively. In the validation dataset, the accuracy (PEA) of JSN score reached 60.6% (Figure 3A, blue line) and that of erosion score reached 72.6% (Figure 3C, blue line).

Test of the model

The rate of joint detection by the trained model reached 95.3% (286/300). In detail, 98.3% of PIP (118/120), 86.7% of IP (26/30) and 94.0% of MCP (142/150) were correctly detected by the model. Joint detection failed in 2 joints of PIP, 4 of IP and 8 of MCP. These areas often contained severely impaired bone alignment or luxation. The distributions of score by the model and clinicians were shown in Figure 3E for JSN and 3G for erosion. Consistency of scores by the model and clinicians are summarized in Tables 3. PEA (accuracy) between the model and two clinicians were 49.3–65.4% for JSN and 70.6–74.1% for erosion. PCA were 64.0–85.3% for JSN and 84.3% for erosion. Sensitivity and specificity (score 0 versus ≥1) were 88.0–94.2% and 52.0–74.8% for JSN, and 34.8–42.4% and 88.2–89.4% for erosion (Supplementary Tables S1 and S2, available at Rheumatology Advances in Practice online). Scatter plots of total score per radiograph were shown in Figure 3F for JSN and 3H for erosion. The correlation coefficients between scores by the model and two clinicians were 0.72–0.88 for JSN and 0.54–0.75 for erosion.

Discussion

In this research, we demonstrated how a deep-learning model was trained in order to assess radiographic finger joint destruction in RA. Disease durations of the patients were relatively long (median 12.2 years), and 67.6% of the patients were seropositive for ACPA, which is a
strong predictor for radiographic progression in RA [24–26]. Thus, patients enrolled in this study were predisposed to have joint destruction. Finger joints such as PIP, IP and MCP were correctly detected by the model with a sensitivity of 95.3%. Intercarpal joints were tended to be ignored by the model, probably because images of intercarpal joints were not offered to the machine learning and the structures of these areas were too complex to detect for the current model. In addition, some joints with severely impaired alignments or luxation were ignored. PEA (accuracy) of JSN score reached 49.3–65.4%, and that of erosion score reached 70.6–74.1%. PCA of JSN score reached 64.0–85.3%, and that of erosion score reached 84.3%. Percentage of agreement of JSN score for PIP/IP were obviously low (Table 3). As machine learning was conducted using images of PIP, IP and MCP together, it may be difficult for the model to judge the differences between joints. As shown in Figures 3G, the distribution of erosion score seems comparable. However, as shown in Figure 3E, the model and the clinician 1 judged too much for score 0, 2 and 3 compared to clinician 2. The correlation coefficients between scores by the model and two clinicians were 0.72–0.88 for JSN score and 0.54–0.75 for erosion score. In the previous report, correlation coefficients between readings of multiple observers, who were radiologists or rheumatologists, were 0.585–0.947 for JSN score and 0.529–0.962 for erosion score [13]. Sensitivity and specificity were 88.0–94.2% and 52.0–74.8% for JSN, and 34.8–42.4% and 88.2–89.4% for erosion. From these results, the most problematic thing of our model was underdiagnosis of erosions and overdiagnosis of JSN.

Thus, image processing technique for hand radiographs using the CNN may be used in the evaluation of joint destruction in RA. Assessment by the model takes less than 1 second per image (average 0.63 second). This is obviously faster than the time humans need to assess. Although echography and MRI examinations are increasingly used for the assessment of joint damage, radiograph keeps a unique value by providing a comprehensive or panoramic view of joints. Automated assessment of radiographs with a deep-learning algorithm would be of great value in many clinical situations. Moreover, novel radiographic findings about joint destruction may be discovered. Recently, many studies using deep-learning or CNN for assessing joints or bones have been reported. These include diagnosis of hip osteoarthritis [27], bone age assessment [28, 29], fracture detection [30] and assessment of knees [31–33].

This research has some limitations. We used the labelled data with scores assessed by the consensus of 2 rheumatologists. Validity and generalizability of the model would be improved
if greater number of images with accurate scoring results were offered to the CNN. Next, our model frequently failed to detect intercarpal joints, which are often impaired in RA. For clinical application, it is needed to include these areas in the assessment by the model. In this study, it was difficult to identify each area of intercarpal joints by the model, and the model assessing these areas could not be developed. Thirdly, sensitivity for erosions was obviously low (34.8–42.4%), indicating oversight for erosions by the model. Additionally, PEA for JSN in PIP/IP was low (24.3–58.3%) and the specificity was also low (52.0–74.8%), indicating overestimation of JSN by the model. To overcome these problems, larger number of data should be added, and the structure of the network or the parameter of the machine learning need to be considered. We examined several settings of parameters such as batchsize (64, 128, 256 or 512), number of epoch (maximum epoch 200), optimization algorithm (Adam, AdaDelta, SGD, RMSprop), and introduction of batch normalization or dropout. Combinations of these settings or parameters are enormous, and further study to optimize them is needed for better performance.

In this study, we introduced a deep-learning model using CNN, which assesses finer joint destruction of RA. This model provides a partial assessment among many joints that can be destroyed in RA. However, to the best of our knowledge, CNN-based deep-learning model has not been applied to automated assessment of radiographic joint destruction in RA. The introduction of AI is useful for prevention of oversight, reduction of time and effort, health survey and assessment by both specialists and non-specialists. In addition, this methodology can be applied to the other areas of joint, such as elbow, shoulder, hip, knee, foot or spine, and to the other disorders such as osteoporosis, fracture or bone tumors. We conclude that image processing with the trained CNN model is promising to assess radiographic finger destruction in RA.

References
1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
2. Webb S. Deep learning for biology. Nature 2018;554:555–7.
3. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep Learning: A Primer for Radiologists. Radiographics 2017;37:2113–31.
4. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316:2402-10.

5. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.

6. Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol 2019;180:373-81.

7. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559-67.

8. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir Med 2018;6:837-45.

9. Smolen JS, Landewé R, Bijlsma J, Burmester G, Chatzidionysiou K, Dougados M, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2016 update. Ann Rheum Dis 2017;76:960-77.

10. Steinbrocker O, Traeger CH, Batterman RC. Therapeutic criteria in rheumatoid arthritis. J Am Med Assoc 1949;140:659-62.

11. Kellgren JH. Radiological signs of rheumatoid arthritis: a study of observer differences in the reading of hand films. Ann Rheum Dis 1956;15:55-60.

12. Sharp JT, Lidsky MD, Collins LC, Moreland J. Methods of scoring the progression of radiologic changes in rheumatoid arthritis. Correlation of radiologic, clinical and laboratory abnormalities. Arthritis Rheum 1971;14:706-20.

13. Sharp JT, Bluhm GB, Brook A, Brower AC, Corbett M, Decker JL, et al. Reproducibility of multiple-observer scoring of radiologic abnormalities in the hands and wrists of patients with rheumatoid arthritis. Arthritis Rheum 1985;28:16-24.

14. Larsen A, Dale K, Eck M. Radiographic evaluation of rheumatoid arthritis and related conditions by standard reference films. Acta Radiol Diagn (Stockh) 1977;18:481-91.
15. Genant HK. Methods of assessing radiographic change in rheumatoid arthritis Am J Med 1983;75:35-47.

16. van der Heijde DM, van Riel PL, Nuver-Zwart IH, Gribnau FW, vad de Putte LB, et al. Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis. Lancet 1989;1:1036-8.

17. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.

18. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO 3rd, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62:2569-81.

19. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 1999;26:743-5.

20. Viola P, Jones M. Rapid Object Detection using a Boosted Cascade of Simple Features. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2001;I-511-8.

21. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv 2014;1412.6980.

22. Loffe S and Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv 2015;1502.03167.

23. Tokui S, Oono K, Hido S, Clayton J. Chainer: a next-generation open source framework for deep learning. In Proceedings of Workshop on Machine Learning Systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS) 2015;5:1-6.

24 Meyer O, Labarre C, Dougados M, Goupille P, Cantagrel A, Dubois A, et al. Anticitrullinated protein/peptide antibody assays in early rheumatoid arthritis for predicting five year radiographic damage. Ann Rheum Dis 2003;62:120-6.

25. Vencovský J, Macháček S, Sedová L, Kafková J, Gatterová J, Pesáková V, et al. Autoantibodies can be prognostic markers of an erosive disease in early rheumatoid arthritis. Ann Rheum Dis 2003;62:427-30.

26. Koga T, Okada A, Fukuda T, Hidaka T, Ishii T, Ueki Y, et al. Anti-citrullinated peptide antibodies are the strongest predictor of clinically relevant radiographic progression in
rheumatoid arthritis patients achieving remission or low disease activity: A post hoc analysis of a nationwide cohort in Japan. PLoS One 2017;12:e0175281.

27. Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PLoS One 2017;12:e0178992.

28. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal 2017;36:41–51.

29. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. Radiology 2018;287:313–22.

30. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 2018;115:11591–6.

31. Liu F, Zhou Z, Samsonov A, Blankenbaker D, Larison W, Kanarek A, et al. Deep Learning Approach for Evaluating Knee MR Images: Achieving High Diagnostic Performance for Cartilage Lesion Detection. Radiology 2018;289:160–9.

32. Tack A, Mukhopadhyay A, Zachow S. Knee menisci segmentation using convolutional neural networks: data from the Osteoarthritis Initiative. Osteoarthritis Cartilage 2018;26:680–8.

33. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med 2018;15:e1002699.

Acknowledgements
We thank Eiryo Kawakami, Haruhiko Koseki, Koiti Asizaki, Takashi Morizono, Tatsuhiko Tsunoda, Ichirou Miki, Toshitada Takemori for planning the research protocol and suggesting constructive opinions, and Eri Itotagawa for collecting the clinical information. TH designed the study, constructed the machine learning model and wrote the initial draft of the manuscript. MN, NN and KE contributed to analysis and interpretation of data. JS, KS and AK designed the research concept and critically reviewed the manuscript.

Funding: This work was supported by the research program “Hub for Predictive and Preventative Precision Medicine Driven by Big Data” conducted RIKEN, the Japan Science
and Technology Agency (JST) and Osaka University, Council for Science, Technology and Innovation (CSTI), cross-ministerial Strategic Innovation Promotion Program (SIP), and “Innovation AI Hospital System” (Funding Agency: National Institute of Biomedical Innovation, Health and Nutrition (NIBIOHN)).

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

Figure legends

Figure 1. Flow of machine learning. (A) The first step of the machine learning is a detection of finger joints, and the second step is a scoring of joint destruction. These steps are combined as the assessment model for radiographic finger joint destruction. (B) The first step used 3720 images for machine learning (*). The second step used 8760 images derived from 146 radiographs for training the CNN (train dataset), and 2400 images derived from 40 radiographs for validation during the training process (validation dataset). Thirty radiographs were used for testing the performance of the model (test dataset). (C) The network of the CNN consists of 2 convolution layers, 2 pooling layers, and 3 fully connected layers.

Figure 2. A representative image processed by the model. (A) A whole hand image processed by the model. The red rectangle indicates joints such as proximal interphalangeal joints (PIP), interphalangeal joint of the thumb (IP) or metacarpophalangeal joints (MCP). The Number on the upper-left on the rectangle indicates JSN score (yellow letter) and that on lower-right indicates erosion score (blue letter). (B) An enlarged image shows the joints with JSN score 0, 2, 3 or 4, and those with erosion score 0, 4, or 5. (C) Another enlarged image shows the joints with JSN score 0, 2 or 4, and those with erosion score 0, 3 or 5. JSN: joint space narrowing.

Figure 3. Test of the model. (A, B) The accuracy, identical to percentage of exact agreement (PEA), and the loss of JSN score during the process for training dataset (red line) and validation dataset (blue line). (C, D) The accuracy (PEA) and the loss of erosion score for training dataset (red line) and validation dataset (blue line). (E) Distribution of JSN score assigned by the model (black bar) and by clinicians (light and dark gray bars). (F) Correlation of JSN score between the model and clinicians. (G) Distribution of erosion score assigned by the model (black bar)
and by clinicians (light and dark gray bars). (H) Correlation of erosion score between the model and clinicians. JSN: joint space narrowing.
Table 1. Characteristics of the patients

|                                | Total   | Train/validation | Test    |
|--------------------------------|---------|------------------|---------|
| N                              | 108     | 93               | 15      |
| Sex (female/male)              | 90 / 18 | 77 / 16          | 13 / 2  |
| Age (year)                     | 64.9 [53.5, 72.6] | 64.9 [53.4, 72.6] | 64.2 [56.8, 76.0] |
| Disease duration (year)        | 12.2 [6.4, 17.6] | 12.3 [6.8, 18.6] | 9.4 [0.7, 14.1] |
| Class I / II / III / IV        | 39 / 56 / 13 / 0 | 35 / 46 / 12 / 0 | 4 / 10 / 1 / 0 |
| Stage I / II / III / IV        | 28 / 19 / 29 / 32 | 24 / 16 / 26 / 27 | 4 / 3 / 3 / 5 |
| ACPA positive                  | 73 (67.6%) | 63 (67.7%)      | 10 (66.7%) |
| N of radiographs               | 216     | 186              | 30      |

Values in age and disease duration are described as median and interquartile range. Other values were number in each category. N: number; ACPA: anti-cyclic citrullinated peptide antibody.
Table 2. Scoring of joint destruction on train/validation dataset

| Score | JSN score | Erosion score |
|-------|-----------|---------------|
|       | PIP/IP    | MCP           | PIP/IP | MCP |
| 0     | 128       | 510           | 644    | 761 |
| 1     | 58        | 28            | 74     | 22  |
| 2     | 356       | 184           | 93     | 33  |
| 3     | 326       | 127           | 28     | 28  |
| 4     | 62        | 81            | 22     | 15  |
| 5     | N.D.      | N.D.          | 69     | 71  |

JSN: joint space narrowing; PIP: proximal interphalangeal joint; IP: interphalangeal joint of the thumb; MCP: metacarpophalangeal joint, N.D.: not defined.
Table 3. Consistency of scores by the model and clinicians

| Evaluator       | Index | Total | PIP/IP | MCP  |
|-----------------|-------|-------|--------|------|
| **For JSN**     |       |       |        |      |
| Model vs. Clinician 1 | PEA   | 65.4% | 58.3%  | 72.5%|
|                  | PCA   | 85.3% | 84.0%  | 86.6%|
| Model vs. Clinician 2 | PEA   | 49.3% | 24.3%  | 74.6%|
|                  | PCA   | 64.0% | 43.1%  | 85.2%|
| Clinician 1 vs. 2 | PEA   | 55.5% | 36.7%  | 74.5%|
|                  | PCA   | 67.6% | 52.7%  | 82.6%|
| **For erosion** |       |       |        |      |
| Model vs. Clinician 1 | PEA   | 74.1% | 66.0%  | 82.4%|
|                  | PCA   | 84.3% | 81.9%  | 86.6%|
| Model vs. Clinician 2 | PEA   | 70.6% | 65.2%  | 76.1%|
|                  | PCA   | 84.3% | 81.3%  | 87.3%|
| Clinician 1 vs. 2 | PEA   | 70.6% | 66.0%  | 75.2%|
|                  | PCA   | 88.0% | 88.7%  | 87.2%|

Totally 286 joints were assessed. Fourteen joints were not identified by the model. PEA is the ratio of exact agreement, and PCA is the ratio of close agreement (within 1.0 score difference) among evaluators. PEA: Percentage of exact agreement; PCA: percentage of close agreement.
Figure 1

A) Evaluation system for finger joint destruction

B) Machine Learning

Step 1: Score detection by Machine Learning model
Step 2: Scoring of destruction by Deep Learning (CNN) model

Input (48 x 48 pixels)

Convolution layer 1
- filter size: 3, padding size: 1, stride: 1
- activation function: ReLU

Pooling layer 1
- filter size: 2, stride: 2
- max pooling

Convolution layer 2
- filter size: 3, padding size: 1, stride: 1
- activation function: ReLU

Pooling layer 2
- filter size: 2, stride: 2
- max pooling

Fully connected layer 1

Fully connected layer 2
- activation function: ReLU
- rate of dropout: 0.5

Output (probability of category)

338x190mm (300 x 300 DPI)
Figure 3

A. Accuracy of JSN score
B. Loss of JSN score
C. Accuracy of erosion score
D. Loss of erosion score

E. Number of patients by JSN score
F. Number of patients by erosion score

338x190mm (300 x 300 DPI)
SUPPLEMENTARY MATERIAL

Supplementary Table S1. Cross-table for JSN or Non-JSN joints.

Supplementary Table 1. Cross-table for JSN or Non-JSN joints

| Model          | JSN (score ≥1) | Non-JSN (score 0) |
|----------------|----------------|-------------------|
| Clinician 1    | 147            | 20                |
| JSN (score ≥1) |                |                   |
| Non-JSN (score 0) | 30           | 89                |
| Clinician 2    | 81             | 5                 |
| JSN (score ≥1) |                |                   |
| Non-JSN (score 0) | 96           | 104               |

Totally 286 joints were assessed. Fourteen joints were not identified by the model. Sensitivity for JSN are 88.0% (147/167) for the judgement of clinician 1, and 94.2% (81/86) for that of clinician 2. Specificity are 74.8% (89/119) and 52.0% (104/200). JSN: Joint space narrowing.
Supplementary Table S2. Cross-table for Erosive or Non-erosive joints.

|            | Erosive (score ≥1) | Non-erosive (score 0) |
|------------|---------------------|------------------------|
| Clinician 1| Erosive (score ≥1)  | 25                     | 34                      |
|            | Non-erosive (score 0)| 24                     | 203                     |
| Clinician 2| Erosive (score ≥1)  | 23                     | 43                      |
|            | Non-erosive (score 0)| 26                     | 194                     |

Totally 286 joints were assessed. Fourteen joints were not identified by the model. Sensitivity for erosion are 42.4% (25/59) for the judgement of clinician 1, and 34.8% (23/66) for that of clinician 2. Specificity are 89.4% (203/227) and 88.2% (194/220).