Original Article

Diagnostic performance of convolutional neural network-based Tanner-Whitehouse 3 bone age assessment system

Xue-Lian Zhou[1#], Er-Gang Wang[2,3#], Qiang Lin[4], Guan-Ping Dong[1], Wei Wu[1], Ke Huang[1], Can Lai[5], Gang Yu[6], Hai-Chun Zhou[5], Xiao-Hui Ma[5], Xuan Jia[5], Lei Shi[4], Yong-Sheng Zheng[4], Lan-Xuan Liu[4], Da Ha[4], Hao Ni[4], Jun Yang[4], Jun-Fen Fu[1]

[1]The Children's Hospital, Zhejiang University School of Medicine, Division of Endocrinology, National Clinical Research Center for Child Health, Hangzhou 310052, China; [2]Center for Genomics and Computational Biology, Duke University, Durham, NC, USA; [3]Department of Biomedical Engineering, Duke University, Durham, NC, USA; [4]Hangzhou YITU Healthcare Technology Co., Ltd, Hangzhou 310012, China; [5]The Children's Hospital, Zhejiang University School of Medicine, Division of Radiology, National Clinical Research Center for Child Health, Hangzhou 310052, China; [6]The Children's Hospital, Zhejiang University School of Medicine, Division of Information Science, National Clinical Research Center for Child Health, Hangzhou 310052, China

[#]These authors contributed equally to this work.

Correspondence to: Junfen Fu, PhD. The Children's Hospital, Zhejiang University School of Medicine, Division of Endocrinology, National Clinical Research Center for Child Health, 3333 Binsheng Road, Hangzhou 310052, China. Email: fjf68@zju.edu.cn.

Background: Bone age can reflect the true growth and development status of a child; thus, it plays a critical role in evaluating growth and endocrine disorders. This study established and validated an optimized Tanner-Whitehouse 3 artificial intelligence (TW3-AI) bone age assessment (BAA) system based on a convolutional neural network (CNN).

Methods: A data set of 9,059 clinical radiographs of the left hand was obtained from the picture archives and communication systems (PACS) between January 2012 and December 2016. Among these, 8,005/9,059 (88%) samples were treated as the training set for model implementation, 804/9,059 (9%) samples as the validation set for parameters optimization, and the remaining 250/9,059 (3%) samples were used to verify the accuracy and reliability of the model compared to that of 4 experienced endocrinologists and 2 experienced radiologists. The overall variation of TW3-metacarpophalangeal, radius, ulna and short bones (RUS) and TW3-Carpal bone score, as well as each bone (13 RUS + 7 Carpal) between reviewers and the AI, were compared by Bland-Altman (BA) chart and Kappa test, respectively. Furthermore, the time consumption between the model and reviewers was also compared.

Results: The performance of TW3-AI model was highly consistent with the reviewers' overall estimation, and the root mean square (RMS) was 0.50 years. The accuracy of the BAA of the TW3-AI model was better than the estimate of the reviewers. Further analysis revealed that human interpretations of the male capitate, hamate, the first distal and fifth middle phalanx and female capitate, the trapezoid, and the third and fifth middle phalanx, were most inconsistent. The average image processing time was 1.5±0.2 s in the TW3-AI model, which was significantly shorter than manual interpretation.

Conclusions: The diagnostic performance of CNN-based TW3 BAA was accurate and timesaving, and possesses better stability compared to diagnostics made by experienced experts.

Keywords: Artificial intelligence (AI); bone age; convolutional neural network (CNN); Tanner-Whitehouse 3 method (TW3 method)

Submitted Jul 25, 2019. Accepted for publication Feb 18, 2020.
doi: 10.21037/qims.2020.02.20
View this article at: http://dx.doi.org/10.21037/qims.2020.02.20

Introduction

Bone age, more so than chronological age, reflects the actual growth and development status of a child. The theory of skeletal physiological maturity was first proposed by Franz Boas (1), and since then, bone age was used to describe different stages of skeletal development. Bone age assessment (BAA) plays a pivotal role in confirming the diagnosis of endocrine diseases, predicting the adult height, and evaluating the efficacy of the treatment. Nevertheless, the basis of these evaluations requires an accurate, consistent, and stable assessment approach.

Greulich and Pyle (GP) (2) and Tanner-Whitehouse 3 (TW3) (3) are the most prevalently used BAA techniques. GP method compares the patient's radiographic information with the nearest standard radiograph in the atlas. Nevertheless, the degree of accuracy can only reach to half a year, and the doctor's subjectivity may cause significant variation between reviewers (4). However, due to its convenience and speed, approximately 76% of doctors worldwide still prefer the GP method (5). On the other hand, the TW3 method is based on a scoring system enabling the bone age estimation accuracy to be within a month. Specifically, the reviewer will firstly identify 20 bones [13 radius, ulna and short bones (RUS) + 7 Carpal], each with a categorized stage. Then, each stage is replaced by a score. Finally, a total score is calculated and transformed into the bone age. This method requires at least 20 minutes to complete the bone age evaluation for manual assessment. Although the TW3 method is more precise compared with the GP method, it is more complex and time-consuming. And, even when adopting the computer-aided detection (CAD) system, the rating for each bone still relies on a human interpretation that also imposes unavoidable inter- and intra-reviewer variability. New and advanced artificial intelligence (AI) techniques are urgently needed to aid the radiologist and clinicians in BAA.

Deep learning is a type of machine learning. When properly trained with a vast number of training samples, the algorithm can make accurate predictions for new input (6). In deep learning, a convolutional neural network (CNN) is a kind of feedforward neural networks with a deep structure that includes convolution or related calculations. It is widely used in image and video recognition, recommender systems, image classification, natural language processing, and medical image analysis (7). A CNN usually consists of an input and an output layer, as well as multiple hidden layers. The activation function is commonly a rectified linear unit (RELU) layer and includes pooling layers and fully connected layers. By using CNN, local information can be effectively utilized without manually selecting features. Also, through the sharing of perceptual fields, we can learn large scale images with small scale parameters. However, CNN is not suitable for long-distance logical reasoning, nor is it good at dealing cases with the feature of large shift or rotation.

Furthermore, the physical information of the features extracted by the convolutional layer is ambiguous. Due to the availability of big data in medical fields and enhanced computing power with graphics processing units (GPU), deep learning has been widely applied in medical applications, including the identification of brain tumors (8) and diabetic retinopathy in retinal fundus (9), early warning of lymph node metastases in breast cancer (10), and classification of skin cancer (11). Many attempts to automating BAA, such as BoneXpert, a system developed by Harvard Medical School and Stanford University, have been proposed over the past few years (12-14). Recently, AI models based on the GP method have been proven to possess great potential in making accurate and time-saving predictions (15). The following study aimed to establish a new large-scale, fully automated CNN-based, TW3 BAA system, and to compare the accuracy, stability, and efficiency of the model with experienced endocrinologists and radiologists.

Methods

Data collection

The institutional review boards of our hospital approved the study. A total of 9,059 left-hand radiographs were obtained from our hospital between January 2012 and December 2016. All images were drawn from the picture archives and communication systems (PACS). The radiology reports included the patient's accession numbers, chronological age, sex, and bone age. *Figure 1* shows the age distribution of the data sets, and *Table 1* summarizes the average chronological age and the estimated bone age of each set.

Among the radiographs, 8,005/9,059 (88%) were randomly selected for the training set, 804/9,059 (9%) were used for validation, and the remaining 250/9,059 (3%) were used for the test set. The training set was used to optimize the model parameters, and the validation set was used to tune hyper-parameters to optimize the model.

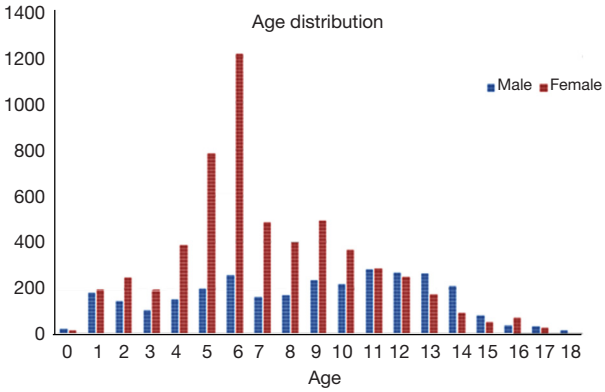

Figure 1 The age distribution of the data sets. A total of 9,059 patients were enrolled in this study; the chronologic age distributions of the patients were 0–18 years old for males and 0–17 years old for females, and the average age was 7.8±3.8 years old.

Table 1 The data set distribution and average chronological age and bone age

Variables	Image no. of males	Image no. of females	Total no.	Average chronological age	Average bone age (TW3-Carpal)	Average bone age (TW3-RUS)
Training set	2,813	5,192	8,005	7.7±3.8	6.9±3.2	7.6±3.5
Validation set	268	536	804	7.9±3.9	7.2±3.4	7.1±2.9
Test set	125	125	250	9.3±4.4	7.5±3.7	8.3±4.6

TW3, Tanner-Whitehouse 3; RUS, radius, ulna and short bones.

Data annotations

The radiograph annotation team included more than 100 professional radiologists and endocrinologists from Children's Hospital in Fujian and Zhejiang Province. During the annotation process, each expert evaluated the same image at least 3 times, and the mode value was chosen as his/her final annotation result. Five different reviewers estimated the rank (from A to I) of each hand bone. The result was accepted as the estimated Ground Truth (eGT) only when the same result was obtained from at least 3 reviewers. Otherwise, a re-grade was needed.

Data pre-processing

Before training the model, each radiograph was first converted from Digital Imaging and Communications in Medicine (DICOM) to a portable network graphic (PNG) file format. The original images were further compressed to 256×256 pixels. The original training data set [8,005] was further expanded more robustly to train the model into more than 100,000 samples by rotating, shifting, and scaling the original images. The augmented parameters and selected value ranges are summarized in *Table 2*. The full implementation pipeline is shown in *Figure 2*.

Model implementation

The primary components of the model included an alignment module and a later classification module. The two modules were built on the same backbone, known as a deep residual network (ResNet), which is a deep CNN with 50 layers and about 3.6×10^9 floating point operations (FLOPS). The model was built according to an open-source machine learning library (TensorFlow version 1.4.1; Google, Mountain View, CA, USA). Training of layers was performed by stochastic gradient descent in batches of 20 images per step, using an Adam Optimizer with a learning rate of 0.001. Training on all categories was run for 80,000 iterations since the training of the final layers for all classes had converged by then. After 80,000 iterations through the entire dataset, the training was stopped due to the absence of further improvement in both accuracy (*Figure S1A*) and sigmoid loss (*Figure S1B*).

Table 2 The augmented parameters and selected value range

Augmented parameters	Random rotation (degree)	Random shift (pixel)	Random scaling	Augment factor
Value range	U (–15, 15)	U (–5, 5)	U (0.9, 1.1)	10

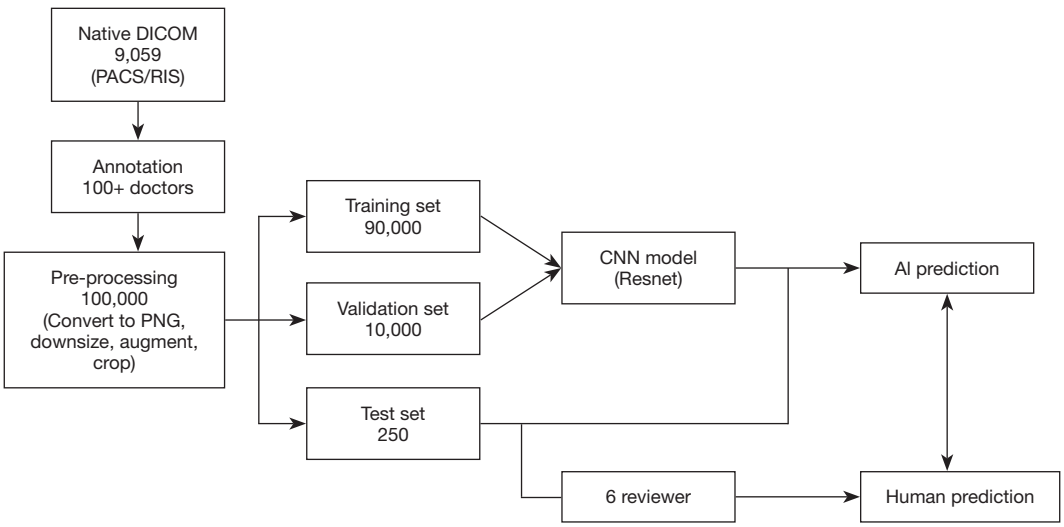

Figure 2 Data flow diagram from raw DICOM images, and automated CNN-based TW3 BAA *vs.* manual human assessment. DICOM, Digital Imaging and Communications in Medicine; CNN, convolutional neural network; TW3, Tanner-Whitehouse 3; BAA, bone age assessment; PACS, picture archives and communication systems; RIS, radiology information system; PNG, portable network graphic; AI, artificial intelligence.

Left-hand radiographs were used as input data, and the alignment module was trained to directly regress all the 20 ossification center regions of TW3. *Figure S2* depicts a total of 13 ossification center regions and 1 carpal bone region, which were fitted by the regression algorithm.

In the classification module, the relevant 20 bones (13 RUS, 7 Carpal) were labeled by the 59 localized points inferred from the alignment module, which were then cropped and impute into the same CNN. The relevant 20 ossification center regions (13 RUS, 7 Carpal) inferred from the alignment module were then passed through a classification network for the labeling of their ossification levels. The classification module used a softmax layer to output multi-classification ranks ranging from A to I for each of the bones. Finally, the ranks for all concerned bones were sent into a TW3-RUS/TW3-Carpal calculator, summed to get the respective final score and cross-referenced with the skeletal maturity table. The output of the classifier was an estimated bone age, according to TW3-RUS and TW3-Carpal method. The scores corresponding to different ranks of bone development are summarized in *Table S1* (for males) and *Table S2* (for females). The Python code (version 3.7.3) implementing the deep CNN and simulation algorithms can be found online at https://github.com/bmehighday/bone-age-algorIthm.

Statistical analysis

A mean paired inter-observer difference was calculated for each reviewer pair to compare the performance of the human reviewers to paired inter-observer. The overall performance of the model was assessed by comparing the root mean square (RMS) and mean values. RMS was calculated as the square root of the sum of the squares of the paired differences, and the mean was calculated as the average of the paired differences. To assess the overall agreement between reviewers along with the agreement between the model and each reviewer, 95% confidence limits of agreement were calculated. Bland-Altman (BA) plot was used to show the consistency between the model and reviewers. Individual bone agreements were performed by Fleiss' kappa statistics (*Table 3*). Statistical significance

was determined by using paired *t*-tests for comparing mean values and F-tests for comparing variances (i.e., RMS). A value with P value <0.05 was considered statistically significant. All the statistical analyses were conducted by R statistical software, version 3.3.2 (R Foundation).

Results

In total, 8,809 images were obtained to train and validate the model, and another independent 250 images were used to test it. The chronologic age distributions of the patients were 0–18 years old for males and 0–17 years old for females, and the average age was 7.8±3.8 years old (*Figure 1*). The data set, mean chronologic age, and bone age are shown in *Table 1*, and the male-to-female ratio of the training and validation set was 3,081/8,809:5,728/8,809 (35%:65%), and 125/250:125/250 (50%:50%) for the test set.

Table 3 The interpretation of κ

κ	Interpretation
<0	Poor agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

The efficiency of TW3-AI model

We compared the time consumption for BAA between the TW3-AI model and the endocrinologists in the test set. The average processing time for the TW3-AI model was 1.5±0.2 s, which was significantly shorter than the average time (525.6±55.5 s) needed for endocrinologists or radiologists to assess bone age according to the TW3 rule.

The diagnostic performance of the TW3-AI model

The accuracy of the diagnostic performance of the TW3-AI model was evaluated in the test set. *Tables 4,5* shows the statistical difference of BAA by the TW3-AI model and reviewers. The average RMS of the model is 0.50 years, which is not significantly different from the average RMS of the 6 reviewers, which means that the performance of the model was not inferior to manual assessment. Meanwhile, the model's RMS was significantly lower than reviewer 5 in both TW3-Carpal and TW3-RUS (P<0.05), but not lower than the other reviewers.

The BA plot shows the difference between the model and the mean of the 6 reviewers in TW3-Carpal (*Figure 3A*) and TW3-RUS (*Figure 3B*), which demonstrates a high consistency between the model and reviewers. However, when we compared the model with each reviewer individually, the BA plot showed a poor consistency between the model and reviewer 5 in TW3-Carpal (*Figure S3A*) and TW3-RUS (*Figure S3B*). The agreement between BAA made by the model and by the reviewers is shown in *Figure 4A* (TW3-Carpal) and

Table 4 The statistical differences of BAA between reviewers and the TW3-Carpal system

Variables	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Mean
Mean							
Reviewer	–0.50	–0.49	0.13	–0.36	0.85	0.36	0.00
Model	–0.33	–0.33	–0.23	–0.31	–0.11	–0.19	–0.25
P value	<0.01	<0.01	<0.01	0.09	<0.01	<0.01	–
RMS							
Reviewer	0.74	0.80	0.91	0.78	1.18	0.93	0.89
Model	0.54	0.56	0.56	0.58	0.36	0.37	0.50
P value	0.92	0.99	0.91	0.80	<0.01	0.11	–

BAA, bone age assessment; TW3, Tanner-Whitehouse 3; RMS, root mean square.

Table 5 The statistical differences of BAA between reviewers and the TW3-RUS system

Variables	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Mean
Mean							
Reviewer	–0.66	–0.78	0.32	0.18	0.70	0.25	0.00
Model	–0.3	–0.32	–0.13	–0.16	–0.07	–0.14	–0.19
P value	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	–
RMS							
Reviewer	0.73	0.85	0.91	0.78	1.15	1.03	0.91
Model	0.57	0.57	0.57	0.58	0.35	0.38	0.50
P value	0.85	0.65	0.71	0.94	0.011	0.11	–

BAA, bone age assessment; TW3, Tanner-Whitehouse 3; RUS, radius, ulna and short bones; RMS, root mean square.

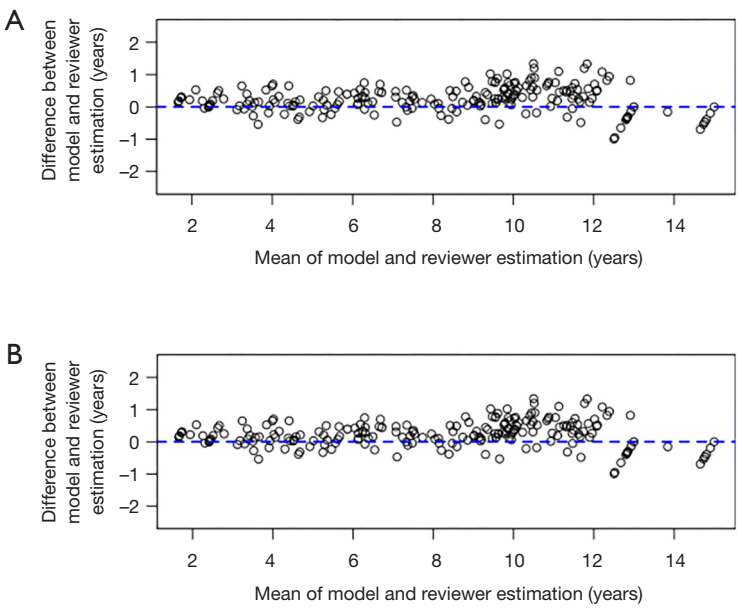

Figure 3 The difference between the model and reviewers. (A) BA plot showing the difference of bone age estimates between the mean of 6 human reviewers and the TW3-AI model (TW3-Carpal); (B) BA plot showing the difference of bone age estimates between the mean of 6 human reviewers and the TW3-AI model (TW3-RUS). BA, Bland-Altman; TW3, Tanner-Whitehouse 3; AI, artificial intelligence; RUS, radius, ulna and short bones.

Figure 4B (TW3-RUS). All assessments are within the 95% confidence limits of agreement between each other.

High variability of reviewer interpretation of individual bones

Kappa-test was used to evaluate the consistency both between the TW3-AI model and reviewers, and between reviewers. *Table 3* shows the interpretation of κ. The overall consistency between the model and reviewers is better than the between reviewers. Further analysis revealed that for experienced endocrinologists and radiologists, their interpretations were most variable in the male capitate and hamate, the female capitate and trapezoid in TW3-Carpal (*Figure 5A*). The bones with the highest estimation variation in TW3-RUS were the male first distal and fifth middle phalanx, the female third phalanx, and the fifth middle phalanx (*Figure 5B*).

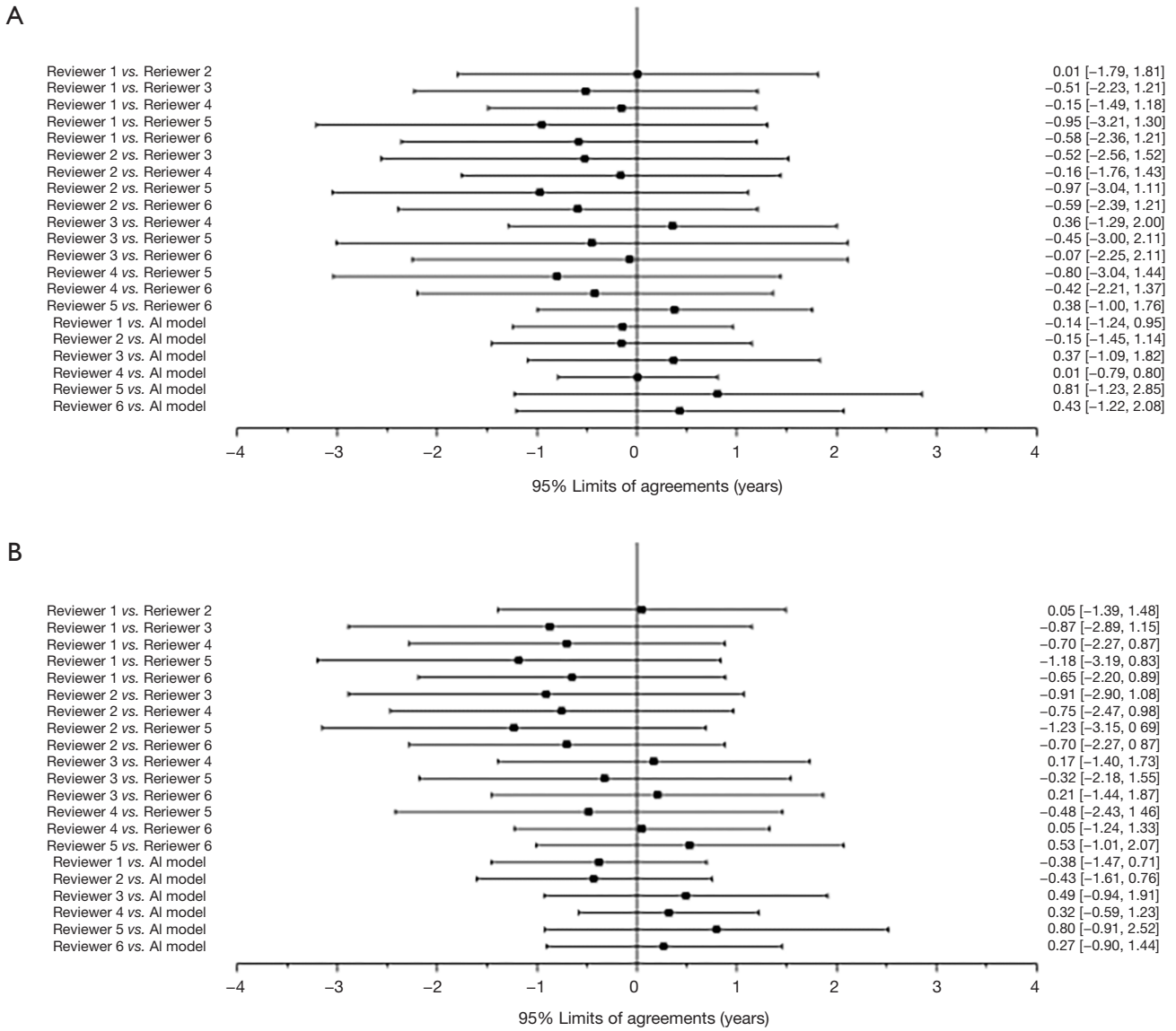

Figure 4 The agreement between BAA made by the model and reviewers. (A) Ninety-five percent limits of agreement between bone age estimates of human reviewers and TW3-AI model (TW3-Carpal); (B) 95% limits of agreement between bone age estimates of human reviewers and TW3-AI model (TW3-RUS). TW3, Tanner-Whitehouse 3; AI, artificial intelligence; RUS, radius, ulna and short bones.

The variation between reviewers in the assessment of these bones was further investigated in this study. Firstly, in TW3-Carpal, different reviewer interpretations occurred in rank B, E, F, and G when assessing the male capitate, while for the male hamate, rank B, F, G, and H were easily misestimated (*Figure S4A*). Similarly, for female samples, reviewers mostly misinterpreted rank C, E, and G in the capitate and rank B, E, F, and G in the trapezoid (*Figure S4B*). Secondly, in TW3-RUS, reviewers mostly misinterpreted rank C, D, and E in the male first distal phalanx, and rank B, E, and H in the fifth middle phalanx (*Figure S4C*). For female samples, rank B, E, and F in the fifth phalanx, and rank C, E, and F in the third middle phalanx were the most misinterpreted ranks (*Figure S4D*).

Discussion

Our group successfully established a CNN-based TW3-

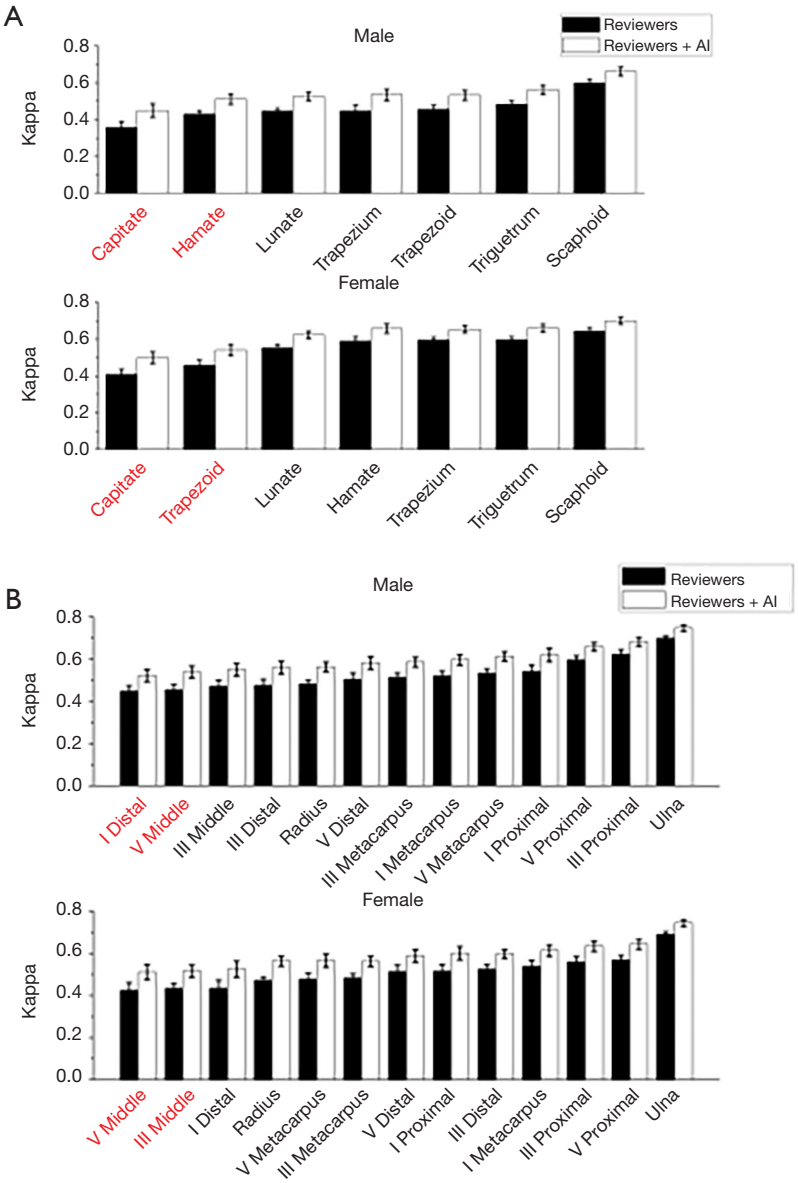

Figure 5 The variability of reviewer interpretation of individual bones. (A) The male capitate and hamate and the female capitate, and trapezoid were the most variable bones interpreted by reviewers according to TW3-Carpal; (B) the first distal and fifth middle phalanx for male, and the third and fifth middle phalanx for females are the most variable bones interpreted by reviewers according to TW3-RUS. TW3, Tanner-Whitehouse 3; AI, artificial intelligence; RUS, radius, ulna and short bones.

AI BAA system, which was developed from a training set size of 8,005 and a validation set size of 804 clinical hand radiographs. The BAA of the model was nearly real-time. After continuous optimization, the model reached an accuracy within 95% of the confidence limits of agreement compared with that of the experts' assessment. In a head-to-head comparison, the consistency between the TW3-AI model and reviewers is better than that between reviewers. We concluded that our TW3-AI model performed similarly to experienced endocrinologists and radiologists in terms of the accuracy of BAA, with better stability than manual interpretations.

BAA is a crucial tool in pediatric clinics, which can be used to evaluate the current status of children's growth

and development, to tell the future growth potential and predicted adult height, and to inform the efficacy of the treatment for the diseases and includes characteristics like short stature, congenital adrenal hyperplasia, and precocious puberty. Consequently, an accurate, consistent, and stable BAA is the prerequisite for clinical endocrine work. However, in the endocrinology department of our hospital, there are more than 47,000 outpatients annually which challenges our endocrinologists to consistently, accurately, and rapidly assess bone age. Thus, the development of a machine learning model would solve many of the problems that pediatricians face every day.

To the best of our knowledge, our model is the first AI model based on the TW3 rule for BAA. Developed on a TW3 scoring system, which rates for both carpal and RUS bones from A to I, our model is different from most of the previous works based on the extraction of morphological features just from carpal or RUS bones (16-18). In this study, the age distribution of the included patients covered the infancy to late adolescence stages, and both carpal and RUS bones were used to train and validate the model. As a result, our model had high accuracy and stability in BAA and was applicable not only to young children but also to older teenagers.

Further analysis revealed that the bones that were most variably interpreted by reviewers were the capitate, hamate, the first distal and fifth middle phalanx of male patients; and the capitate, trapezoid, and the third and fifth middle phalanx of female patients. The underlying reason for the variation was that the grading and scoring of these bones are subjective, so inter- and intra-reviewer variation is inevitable. By strengthening the learning of these individual bones and ranks scoring, we could improve the accuracy and consistency of BAA in clinicians, which in turn can be used as a reference for future model optimization for clinicians in the evaluation of bone age.

Much work has been done on refining an automatic system to evaluate bone age. HANDX and CASAS systems were the earliest attempts for automatic BAA (19,20). Both systems were based on feature extraction of hand bones, and they showed better consistency than manual assessment. Nevertheless, they are more time-consuming than the manual evaluation of bone age (16). Recently, Harvard Medical School and Stanford University School of Medicine individually developed an automated deep learning system for BAA; both models are based on feature extraction (13,14). Our model was based on the TW3 rule, which is recognized as the most objective method used to evaluate bone age. It also demonstrated superior stability in BAA, as the RMS was 0.50 years in both TW3-Carpal and TW3-RUS, which was smaller compared to the RMS of 0.67 years in the model developed by Stanford University.

There are several limitations to our model. Firstly, similar to the previous works in BAA, there is no gold standard for bone age evaluation, because the inter- and intra-reviewer variations are inevitable (21,22). Previous studies have reported a standard error of the inter-reviewer variation from 0.45 to 0.83 years (standard deviation of 0.64 to 1.17 years) (12,23). In this study, the RMS of inter-reviewer variation was 0.72 years, which is comparable with previous research, while the RMS between the model and reviewer was 0.50 years, which showed superior stability compared to the manual assessment. Secondly, all the images were obtained from our hospital, suggesting more images should be collected from other medical centers to reduce the bias. Thirdly, our model cannot detect certain diseases that human specialists might identify when analyzing the images, such as rickets, hypochondroplasia, and other congenital syndromes (22). Nonetheless, we believe that with the development of medically oriented machine learning techniques, the advantages of AI-model in BAA will become increasingly apparent.

In summary, we developed an automated CNN-based TW3-AI model that can estimate bone age with similar accuracy and superior stability compared to manual assessment. The highly accurate and efficient TW3-AI model will spare clinicians from the tedious clinical viewing process, and thoroughly improve the level of diagnosis and treatment for children's endocrine diseases.

Acknowledgments

We are incredibly grateful to all the patients who took part in this study, along with the whole team, including the technicians, engineers, clerical workers, research scientists, and nurses.

Funding: This study was supported by the National Key Research and Development Program of China (No. 2016YFC1305301), the National Natural Science Foundation of China (No. 81570759 and 81270938), the Fundamental Research Funds for the Central Universities (2017XZZX001-01), the Research Fund of Zhejiang Major Medical and Health Science and Technology & National Ministry of Health (WKJ-ZJ-1804), the Public Welfare Technology Application Research Program of Zhejiang Provincial Science and Technology Project

(2016C33130), and the Zhejiang Province Natural Sciences Foundation Zhejiang Society for Mathematical Medicine (LSZ19H070001).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: The institutional review boards of our hospital approved the study (No. 2016-IRB-018). All patients provided written informed consent.

References

1. Tanner JM. A history of the study of human growth. London: Cambridge University Press, 1981.
2. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. Stanford: Stanford University Press, 1959:238-393.
3. Tanner JM, Healy MJR, Cameron N, Goldstein H. Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 Method). London: W.B. Saunders, 2001.
4. Roche AF, Rohmann CG, French NY, Dávila GH. Effect of training on replicability of assessments of skeletal maturity (Greulich-Pyle). Am J Roentgenol Radium Ther Nucl Med 1970;108:511-5.
5. De Sanctis V, Soliman AT, Di Maio S, Bedair S. Are the new automated methods for bone age estimation advantageous over the manual approaches? Pediatr Endocrinol Rev 2014;12:200-5.
6. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, Kim N. Deep learning in medical imaging: general overview. Korean J Radiol 2017;18:570-84.
7. Wang S, Wang R, Zhang S, Li R, Fu Y, Sun X, Li Y, Sun X, Jiang X, Guo X, Zhou X, Chang J, Peng W. 3D convolutional neural network for differentiating pre-invasive lesions from invasive adenocarcinomas appearing as ground-glass nodules with diameters ≤3 cm using HRCT. Quant Imaging Med Surg 2018;8:491-9.
8. Charron O, Lallement A, Jarnet D, Noblet V, Clavier JB, Meyer P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. Comput Biol Med 2018;95:43-54.
9. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.
10. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM; the CAMELYON16 Consortium, Hermsen M, Manson QF, Balkenhol M, Geessink O, Stathonikos N, van Dijk MC, Bult P, Beca F, Beck AH, Wang D, Khosla A, Gargeya R, Irshad H, Zhong A, Dou Q, Li Q, Chen H, Lin HJ, Heng PA, Haß C, Bruni E, Wong Q, Halici U, Öner MÜ, Cetin-Atalay R, Berseth M, Khvatkov V, Vylegzhanin A, Kraus O, Shaban M, Rajpoot N, Awan R, Sirinukunwattana K, Qaiser T, Tsang YW, Tellez D, Annuscheit J, Hufnagl P, Valkonen M, Kartasalo K, Latonen L, Ruusuvuori P, Liimatainen K, Albarqouni S, Mungal B, George A, Demirci S, Navab N, Watanabe S, Seno S, Takenaka Y, Matsuda H, Ahmady Phoulady H, Kovalev V, Kalinovsky A, Liauchuk V, Bueno G, Fernandez-Carrobles MM, Serrano I, Deniz O, Racoceanu D, Venâncio R. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199-210.
11. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
12. Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging 2009;28:52-66.
13. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, Choy G, Do S. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017;30:427-41.
14. 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.
15. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, Pan I, Pereira LA, Sousa RT,

Abdala N, Kitamura FC, Thodberg HH, Chen L, Shih G, Andriole K, Kohli MD, Erickson BJ, Flanders AE. The RSNA pediatric bone age machine learning challenge. Radiology 2019;290:498-503.
16. Seok J, Hyun B, Kasa-Vubu J, Girard A. Automated classification system for bone age X-ray images. Seoul: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2012:208-13.
17. Zhang A, Gertych A, Liu BJ. Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones. Comput Med Imaging Graph 2007;31:299-310.
18. Somkantha K, Theera-Umpon N, Auephanwiriyakul S. Bone age assessment in young children using automatic carpal bone feature extraction and support vector regression. J Digit Imaging 2011;24:1044-58.
19. Michael DJ, Nelson AC. HANDX: a model-based system for automatic segmentation of bones from digital hand radiographs. IEEE Trans Med Imaging 1989;8:64-9.
20. Pietka E, McNitt-Gray MF, Kuo ML, Huang HK. Computer-assisted phalangeal analysis in skeletal age assessment. IEEE Trans Med Imaging 1991;10:616-20.
21. Kim JR, Lee YS, Yu J. Assessment of bone age in prepubertal healthy Korean children: comparison among the Korean standard bone age chart, Greulich-Pyle method, and Tanner-Whitehouse method. Korean J Radiol 2015;16:201-5.
22. van Rijn RR, Thodberg HH. Bone age assessment: automated techniques coming of age? Acta Radiol 2013;54:1024-9.
23. Bull RK, Edwards PD, Kemp PM, Fry S, Hughes IA. Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods. Arch Dis Child 1999;81:172-3.

Cite this article as: Zhou XL, Wang EG, Lin Q, Dong GP, Wu W, Huang K, Lai C, Yu G, Zhou HC, Ma XH, Jia X, Shi L, Zheng YS, Liu LX, Ha D, Ni H, Yang J, Fu JF. Diagnostic performance of convolutional neural network-based Tanner-Whitehouse 3 bone age assessment system. Quant Imaging Med Surg 2020;10(3):657-667. doi: 10.21037/qims.2020.02.20

Supplementary

Table S1 The scores corresponding to different grades of bone development (male)

Model	Grade score bone	A	B	C	D	E	F	G	H	I
TW3-RUS	I metacarpus	0	6	9	14	21	26	36	49	67
	III metacarpus	0	4	5	9	12	19	31	43	52
	V metacarpus	0	4	6	9	14	18	29	43	52
	I proximal phalanx	0	7	8	11	17	26	38	52	67
	III proximal phalanx	0	4	4	9	15	23	31	40	53
	V proximal phalanx	0	4	5	9	15	21	30	39	51
	I distal phalanx	0	5	6	11	17	26	38	46	66
	III distal phalanx	0	4	6	8	13	18	28	34	49
	V distal phalanx	0	5	6	9	13	18	27	34	48
	III middle phalanx	0	4	6	9	15	22	32	43	52
	V middle phalanx	0	6	7	9	15	23	32	42	49
	Radius	0	16	21	30	39	59	87	138	213
	Ulna	0	27	30	32	40	58	107	181	–
TW3-Carpal	Triquetrum	0	10	13	28	57	84	102	124	–
	Lunate	0	14	22	39	58	84	101	120	–
	Scaphoid	0	26	36	52	71	85	100	116	–
	Trapezium	0	23	31	46	66	83	95	108	117
	Trapezoid	0	27	32	42	51	77	93	115	–
	Hamate	0	73	75	79	100	128	159	181	194
	Capitate	0	100	104	106	113	133	160	214	–

TW3, Tanner-Whitehouse 3; RUS, radius, ulna and short bones.

Table S2 The scores corresponding to different grades of bone development (female)

Model	Grade score bone	A	B	C	D	E	F	G	H	I
TW3-RUS	I metacarpus	0	8	12	18	24	31	43	53	67
	III metacarpus	0	5	8	12	16	23	37	47	53
	V metacarpus	0	6	9	12	17	23	35	48	52
	I proximal phalanx	0	9	11	14	20	31	44	56	67
	III proximal phalanx	0	5	7	12	19	27	37	44	54
	V proximal phalanx	0	6	7	12	18	26	35	42	51
	I distal phalanx	0	7	9	15	22	33	48	51	68
	III distal phalanx	0	7	8	11	15	22	33	37	49
	V distal phalanx	0	7	8	11	15	22	32	36	47
	III middle phalanx	0	6	8	12	18	27	36	45	52
	V middle phalanx	0	7	8	12	18	28	35	43	49
	Radius	0	23	30	44	56	78	114	160	218
	Ulna	0	30	33	37	45	74	118	173	–
TW3-Carpal	Triquetrum	0	11	16	31	56	80	104	126	–
	Lunate	0	16	24	40	59	84	106	122	–
	Scaphoid	0	24	35	51	71	88	104	118	–
	Trapezium	0	20	27	42	60	80	95	111	119
	Trapezoid	0	21	30	43	53	77	97	118	–
	Hamate	0	72	74	78	102	131	161	183	194
	Capitate	0	84	88	91	99	121	149	203	–

TW3, Tanner-Whitehouse 3; RUS, radius, ulna and short bones.

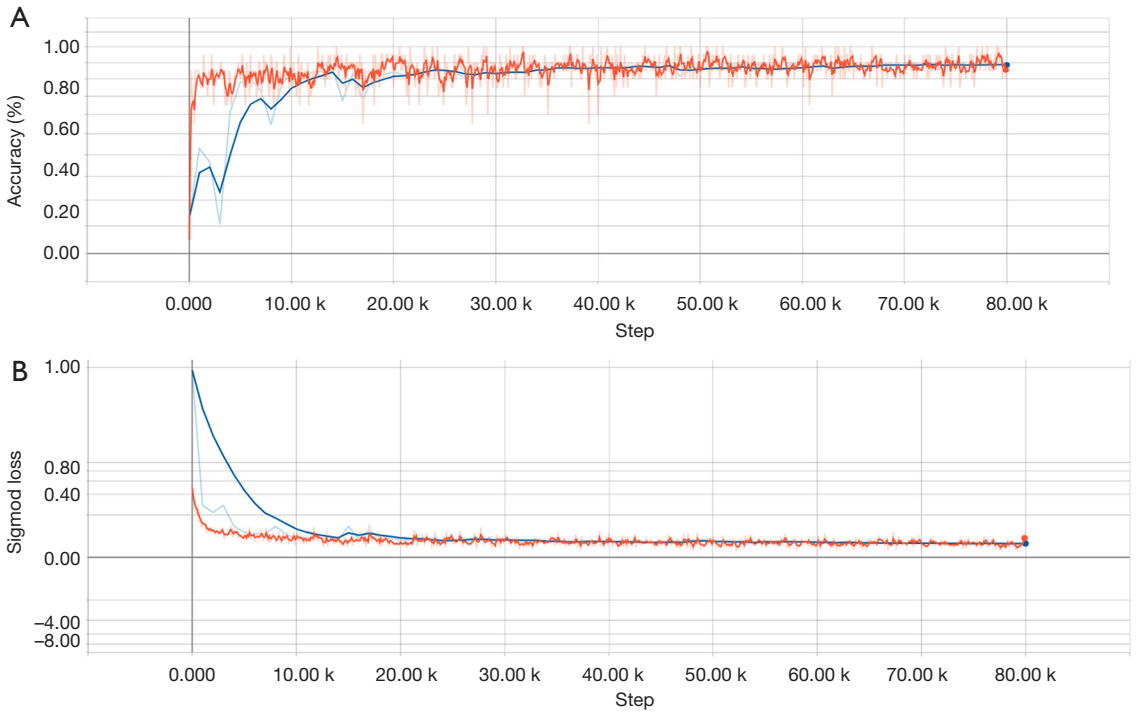

Figure S1 Plot shows performance in the training and validation data sets using Tensor Board. (A) The accuracy was plotted against the training step, and (B) sigmoid loss was plotted against the training step during the training of the multi-class classifier throughout 80,000 iterations. Plots were normalized with a smoothing factor of 0.6 to visualize the trends. Training dataset, orange; validation dataset, blue.

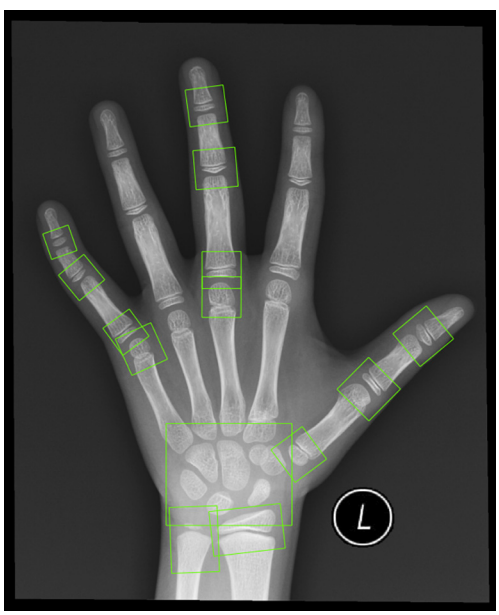

Figure S2 A total of 13 ossification center regions and 1 carpal bone region which were fitted by the regression algorithm.

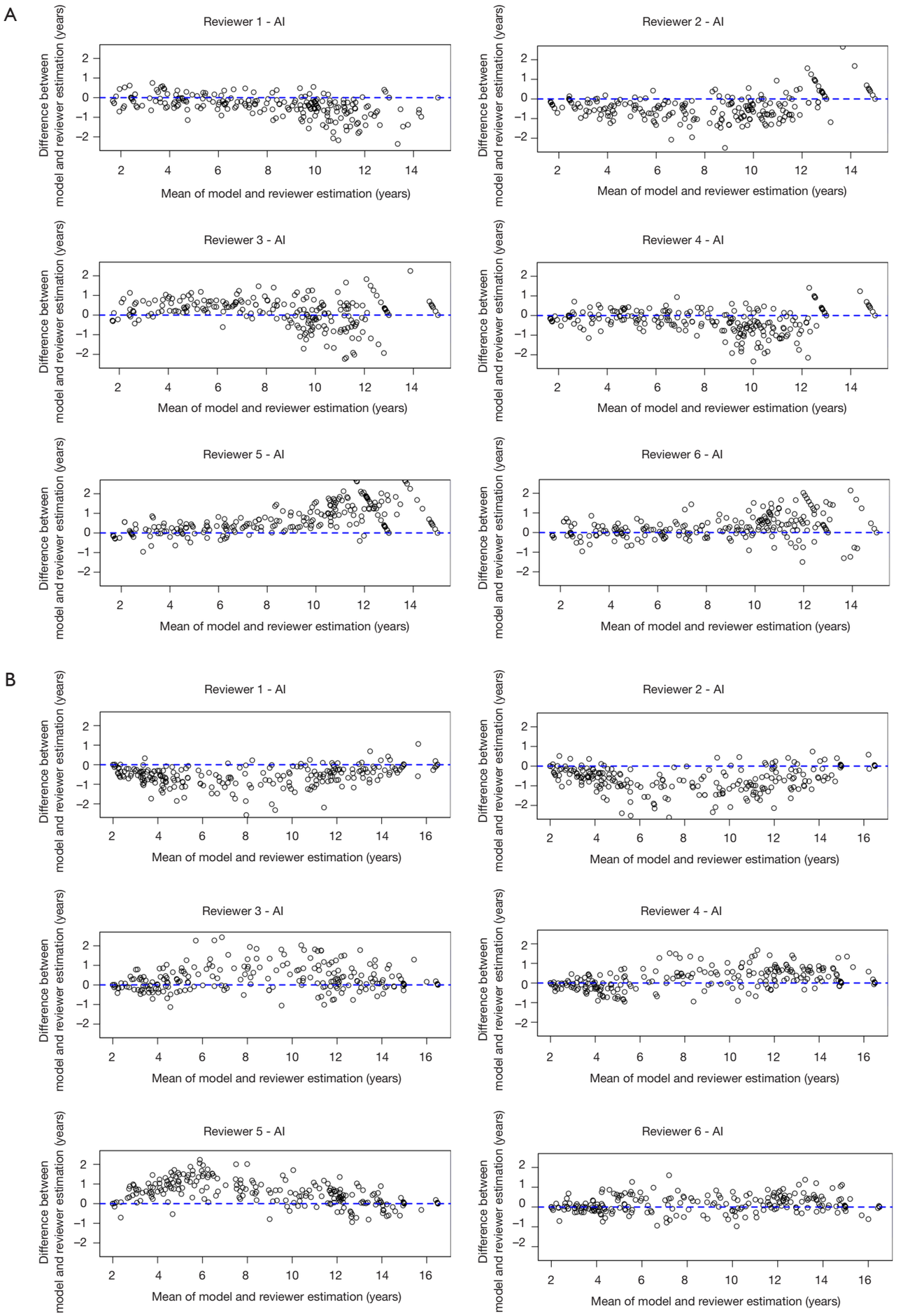

Figure S3 The difference of BAA between each reviewer and TW3-AI model. (A) BA plot showing the difference of bone age estimates between each reviewer and TW3-AI model (TW3-Carpal), which shows a poor consistency between the model and reviewer 5; (B) BA plot showing the difference of bone age estimates between each reviewer and TW3-AI model (TW3-RUS) shows a poor consistency between the model and reviewer 5. BA, Bland-Altman; TW3, Tanner-Whitehouse 3; AI, artificial intelligence; RUS, radius, ulna and short bones.

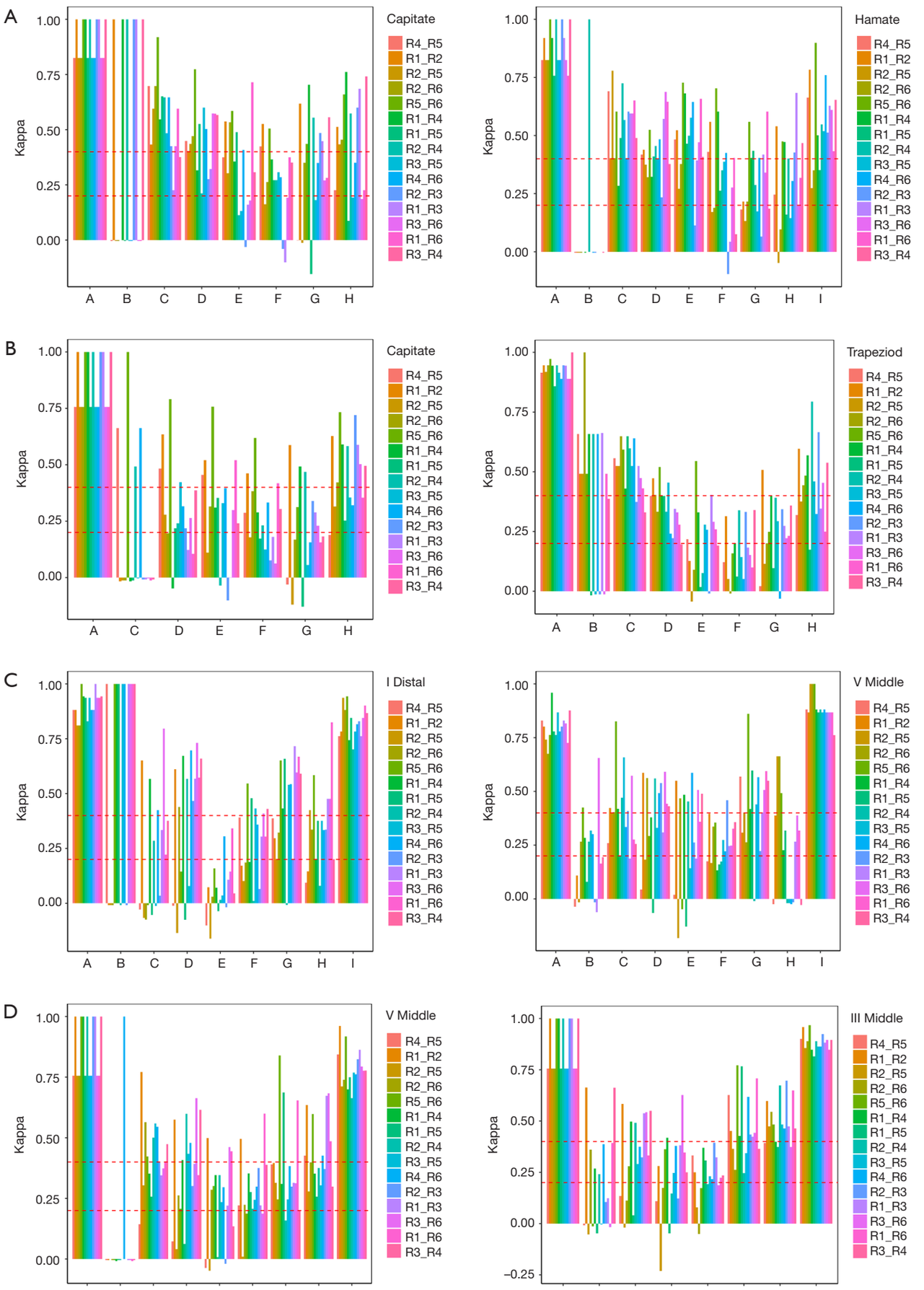

Figure S4 A, B, C, D, E, F, G, H, and I in the abscissa represents the level of bone maturity. (A) Rank B, E, F, and G in the capitate and rank B, F, G, and H in the hamate are the most easily misestimated by human reviewers for males according to TW3-Carpal; (B) rank C, E, and G in the capitate, and rank B, E, F, and G in the trapezoid are the most easily misestimated by human reviewers for females, according to TW3-Carpal; (C) rank C, D, and E in the male first distal phalanx, and rank B, E, and H in the fifth middle phalanx are the most easily misestimated by human reviewers for males according to TW3-RUS; (D) rank B, E, and F in the fifth phalanx, and rank C, E, and F in the third middle phalanx are the most easily misestimated by human reviewers for females according to TW3-RUS. TW3, Tanner-Whitehouse 3; RUS, radius, ulna and short bones.