European Radiology (2023) 33:3544–3556
https://doi.org/10.1007/s00330-022-09330-0

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

The uncovered biases and errors in clinical determination of bone age by using deep learning models

Mei Bai[1] · Liangxin Gao[2] · Min Ji[1] · Jianbang Ge[2] · Lingyun Huang[2] · HaoChen Qiao[3] · Jing Xiao[2] · Xiaotian Chen[4] · Bin Yang[1] · Yingqi Sun[1] · Minjie Zhang[1] · Wenjie Zhang[5] · Feihong Luo[6] · Haowei Yang[1] · Haibing Mei[7] · Zhongwei Qiao[1]

Received: 12 June 2022 / Revised: 13 October 2022 / Accepted: 28 November 2022 / Published online: 20 December 2022

Abstract

Objectives To evaluate AI biases and errors in estimating bone age (BA) by comparing AI and radiologists' clinical determinations of BA.

Methods We established three deep learning models from a Chinese private dataset (CHNm), an American public dataset (USAm), and a joint dataset combining the above two datasets (JOIm). The test data CHNt (n = 1246) were labeled by ten senior pediatric radiologists. The effects of data site differences, interpretation bias, and interobserver variability on BA assessment were evaluated. The differences between the AI models' and radiologists' clinical determinations of BA (normal, advanced, and delayed BA groups by using the Brush data) were evaluated by the chi-square test and Kappa values. The heatmaps of CHNm-CHNt were generated by using Grad-CAM.

Results We obtained an MAD value of 0.42 years on CHNm-CHNt; this result indicated an appropriate accuracy for the whole group but did not indicate an accurate estimation of individual BA because with a kappa value of 0.714, the agreement between AI and human clinical determinations of BA was significantly different. The features of the heatmaps were not fully consistent with the human vision on the X-ray films. Variable performance in BA estimation by different AI models and the disagreement between AI and radiologists' clinical determinations of BA may be caused by data biases, including patients' sex and age, institutions, and radiologists.

Conclusions The deep learning models outperform external validation in predicting BA on both internal and joint datasets. However, the biases and errors in the models' clinical determinations of child development should be carefully considered.

Key Points

- *With a kappa value of 0.714, clinical determinations of bone age by using AI did not accord well with clinical determinations by radiologists.*
- *Several biases, including patients' sex and age, institutions, and radiologists, may cause variable performance by AI bone age models and disagreement between AI and radiologists' clinical determinations of bone age.*
- *AI heatmaps of bone age were not fully consistent with human vision on X-ray films.*

Keywords Deep learning · Child development · X-ray film · Radiologists · Computers

✉ Min Ji
ilovexray_349@163.com

✉ Zhongwei Qiao
zqiao@fudan.edu.cn

[1] Department of Radiology, Children's Hospital of Fudan University, No 399, Wan Yuan Road, Minhang District, Shanghai 201102, China

[2] Ping An Technology, Shenzhen, China

[3] School of Public Health, Yale University, New Haven, USA

[4] Department of Clinical epidemiology, Children's Hospital of Fudan University, Shanghai, China

[5] Information Technology Center, Children's Hospital of Fudan University, Shanghai, China

[6] Department of Endocrinology, Children's Hospital of Fudan University, Shanghai, China

[7] Department of Radiology, Ningbo Women and Children's Hospital, Ningbo, China

Abbreviations

AD	Absolute difference of bone age values between AI and radiologists
BA	Bone age
BN	Batch normalization
Brush data	The variability of skeletal age in the Brush Foundation Study of Human Growth and Development, led by Professor T. Wingate Todd
CA	Chronological age
CHNm	Bone age model from a Chinese private dataset (11226 images from our hospital in 2018)
CHNt	Chinese test dataset (1246 images from our hospital in 2018)
JOIm	Joint model with combined data from the CHN model and the USA model
JOIt	Joint test dataset from China and America data
LOA	Limits of agreement
MAD	Mean absolute difference of bone age values between AI and radiologists
MAE Loss	Mean absolute error loss function
MSE Loss	Mean square error loss function
PACS	Picture Archiving and Communication System
RELU	Rectified linear unit
RMSE	Root mean square error
SD	Standard deviation
USAm	Bone age model from an American public dataset (9607 images from the 2017 RSNA Pediatric Bone Age Machine Learning Challenge)
USAt	American test dataset (1060 images from the 2017 RSNA Pediatric Bone Age Machine Learning Challenge)

Introduction

Bone age (BA) assessment is an interpretation of skeletal maturity from the X-ray of the left hand. The BA value estimated is a doctor's reference in children's health care or other circumstances, e.g., forensic analysis and sports medicine [1]. Radiologists usually make a BA report based on the Greulich-Pyle (G&P) atlas; this method is one of the popular methods for BA assessment [2]. In clinical practice, BA assessment includes the BA value and clinical determination of BA. The patient's BA value is assigned by best matching his left hand and wrist radiograph with a reference standard image from the G&P atlas. The clinical determinations based on BA value include advanced, normal, and delayed skeletal development. The Brush data are used to define the skeletal development condition. Some studies have suggested that AI has a potential advantage over humans in BA assessment because BA is a quantitative value and therefore is an ideal target for automated image evaluation [3, 4].

Deep learning, known as a subtype of machine learning, has shown high accuracy in performing different tasks for medical image analysis [5]. In recent years, many novel approaches based on deep learning have been utilized for BA assessment [6]. The Radiological Society of North America (RSNA) Pediatric Bone Age Machine Learning Challenge [7] was launched at the 2017 RSNA Annual Meeting, and with a low mean absolute difference (MAD) ranging from 4.265 to 4.907 months for the 10 best teams, the result of the challenge demonstrated the success of machine learning in BA assessment [8].

Generally, BA assessment is affected by ethics, region, economic status, and nutrition. Deep learning model training using image data from various settings or patient populations may be able to mitigate the generalization problem [9]. The problem of generalization is that a model trained in some situations cannot make the same accurate prediction in new ones. However, at this point, few papers have addressed the generalization of BA models by comparing single and joint data sources (institutions). Few studies have evaluated the factors, including patients, radiologists, and clinical determination, on the effect of AI models. Few papers on using deep learning for BA assessment have addressed the Brush data. These papers evaluate the performance of BA models in terms of MAD values but rarely evaluate the differences between human and machine clinical determinations of BA by using Brush data.

In this study, we established three AI models: (1) the USA model (USAm) from the publicly available RSNA dataset, (2) the CHN model (CHNm) from the dataset of the National Children's Medical Centre in China, and (3) the JOI model (JOIm) from the mixed dataset from the above two models. This study aimed to evaluate AI performance in assessing BA and the effects of patient sex and age, data site differences, interpretation bias, and interobserver variability on AI performance. We further assessed the agreement between AI and radiologists' clinical determinations of BA. Because AI estimations of BA are a black box [10], we used the heatmaps to observe the AI vision on the X-ray films compared with the human behaviors in medical procedures.

Methods

The workflow chart in this study is shown in Fig. 1. The steps included model design, statistics, and heatmap generation.

Data acquisition

Our ethics committee approved this retrospective study and waived the requirement for informed consent. After excluding

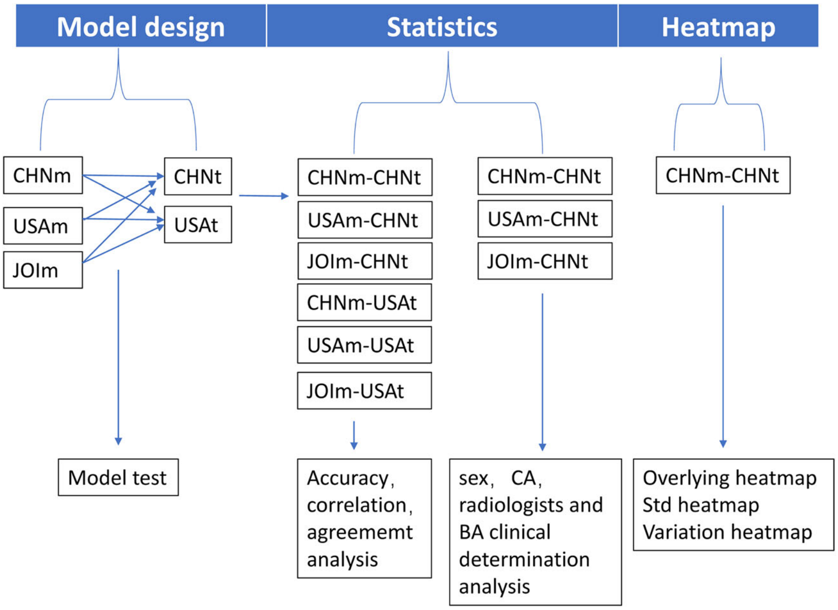

Fig. 1 The workflow chart in this study. There were three steps, including deep learning model design, statistical evaluation of the performance of AI models, and heatmap generation and explanation. Three AI models were generated by using data from China (CHNm), America (USAm), and both China and America (JOIm). Two test datasets were from China (CHNt) or America (CHNt). The performance of AI models (CHNm, USAm, and JOIm) was evaluated with some parameters, including the mean absolute difference (MAD) and Bland–Altman plots. Based on the clinical determination of BA with the Brush data rule, the sensitivity and specificity of three models detecting abnormalities (advanced and delayed development) were calculated. The effects of sex, chronological age (CA), radiologists, and population were analyzed. The heatmaps were shown to help clarify AI decisions

abnormal images and reports, 12,472 radiographs of the left hand and wrist were originally retrieved from our children's hospital between July and September 2018 and used for our deep-learning model (CHNm) and test data (CHNt). All DICOM left hand and wrist radiographs, radiology reports, radiologists' names, and the sex and chronological age (CA) of patients were exported from the Picture Archiving and Communication System. Images were labeled by BA value; BA values were extracted from the radiology reports. All radiology reports were provided by pediatric radiologists with more than 10 years of experience with reference to the paper-based Greulich-Pyle atlas (second edition) [2]. Ten senior pediatric radiologists took part in evaluations using CHNt (n = 1246). Their years of working experience in interpreting and reporting radiographs were 37 years (for D1, who reviewed 524 images), 33 years (for D2, who reviewed 127 images), 20 years (for D3, who reviewed 2 images), 18 years (for D4 and D5, who reviewed 109 and 69 images, respectively), 16 years (for D6 and D7, who reviewed 60 and 214 images, respectively), 15 years (for D8, who reviewed 82 images), 11 years (for D9, who reviewed 23 images), and 10 years (for D10, who reviewed 36 images). To evaluate the effect of interobserver variability on CHNm-CHNt performance, disputed cases outside the 95% limits of agreement (LOAs) of the difference between the AI and radiology BA reports were rerated by another radiologist with 10 years of experience, and then we took the average of the reports and rerated the BA of disputed cases as a new manual BA.

A total of 10,667 images from the RSNA dataset [7] were used as the USA model (USAm) and USA test data (USAt) after excluding some images with additional artifacts and missing parts of hands. We further mixed data from CNHm and USAm together to implement the third AI model (JOIm). Image numbers and demographic data for all datasets are shown in Table 1.

Data preprocessing

The first task of the preprocessing pipeline was to extract the hand bone region in the X-ray radiographs. To automatically generate the hand mask, the U-Net [11] network architecture originally suggested for image segmentation was employed. We manually annotated 200 hand bone masks by using an online annotation service as the training dataset. In the training phase, we used the optimized Dice loss function as the target of optimizing the segmentation network.

Second, we aligned the important region of the hands into a common coordinate space. Therefore, we detected the coordinates of several specific key points of a hand for this purpose. ResNet [12] network was used as the feature extraction backbone network for extracting location information. The output was 6 coordinates corresponding to three sets of key points: tip of the distal phalanx of the third finger, tip of the distal phalanx of the thumb, and center of the capitate (Fig. 2). We used the mean square error loss function to train our landmark detection model.

Table 1 Summary information for three BA models with data from China (CHNm), America (USAm) and joint (JOIm), and two test datasets from China (CHNt) and America (USAt)

Variable	No. of images in males/total (%)	Report bone age, year, median [IQR]	Chronological age, year, median [IQR]
Training dataset			
China	3570/9980 (35.77)	10.0 [8.0–11.5]	9.0 [8.0–11.0]
America	4599/8547 (53.81)	11.0 [8.0–13.0]	/
Joint	8169/18,527 (44.09)	10.0 [8.0–12.5]	/
Valid dataset			
China	434/1246 (34.83)	10.0 [8.0–11.0]	9.0 [8.0–11.0]
America	558/1060 (52.64)	11.5 [9.0–13.5]	/
Joint	992/2306 (43.02)	10.0 [8.0–12.0]	/
Test dataset			
China	435/1246 (34.91)	10.0 [8.5–11.5]	9.0 [8.0–11.0]
America	576/1060 (54.34)	11.0 [8.0–13.0]	

IQR interquartile range

Bone age AI model

We randomly divided the preprocessed data into a training set, validation set, and test set at a ratio of 8:1:1. The attention module CBAM was integrated into the inception V3 network to improve the network's feature extraction capability (Fig. 3). Then, we concatenated the feature with the patient's sex, which was encoded and mapped to [0, 1] before inputting

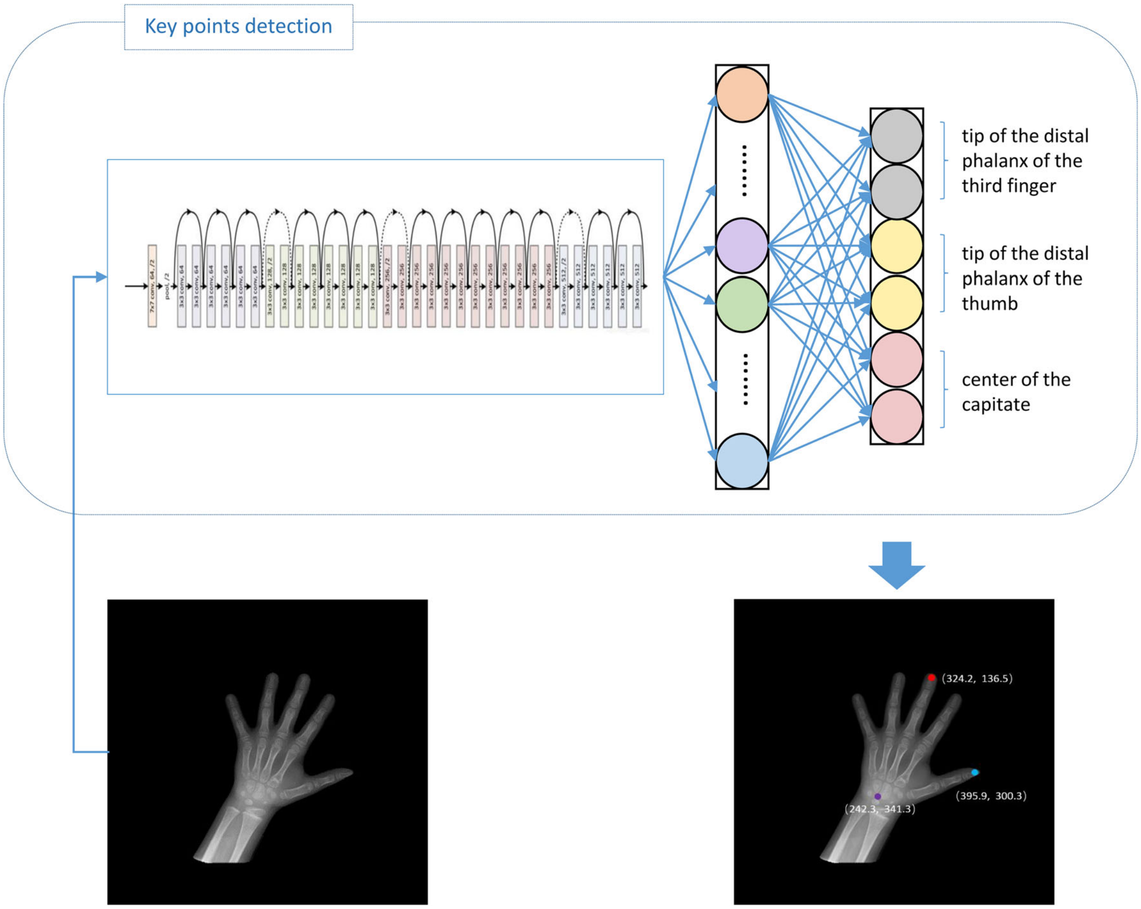

Fig. 2 Key points detection model. The U-Net network architecture was employed for image segmentation. The optimized Dice loss function is the target of optimizing the segmentation network. ResNet was used as the feature extraction backbone network. Three regions of the hand were aligned into a common coordinate space. The output was 6 coordinates, including the tip of the distal phalanx of the third finger, tip of the distal phalanx of the thumb, and center of the capitate

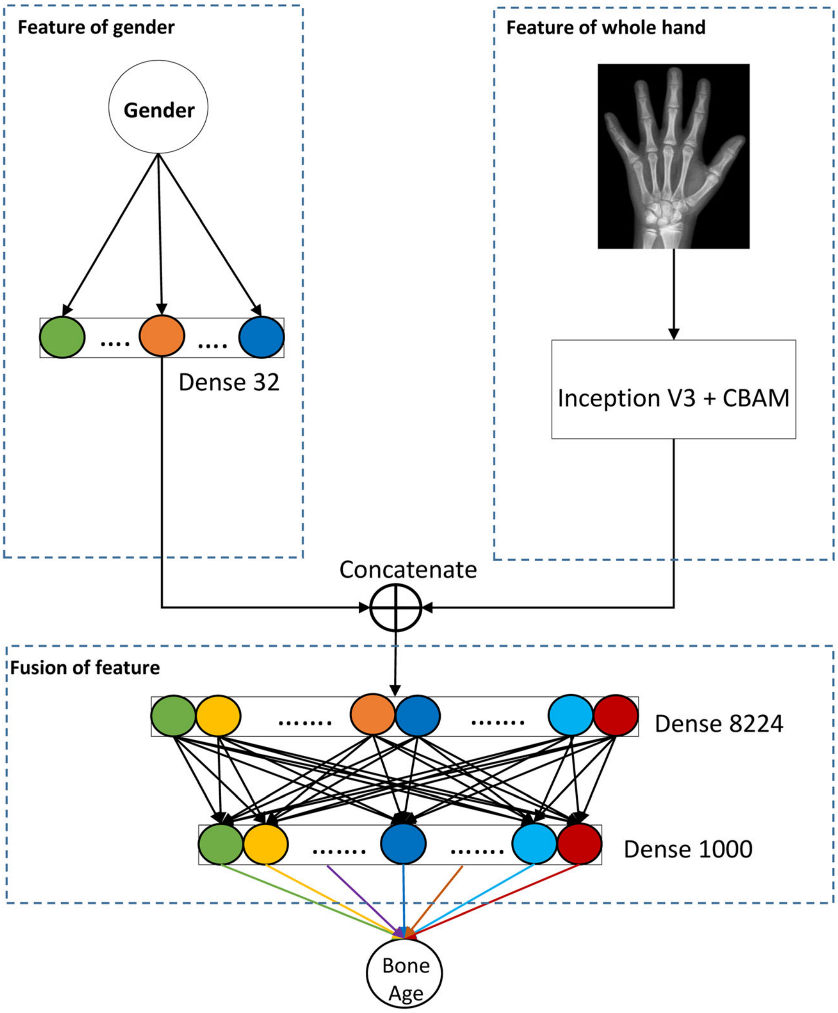

Fig. 3 Bone age assessment model. The attention module CBAM and inception V3 network were integrated into the model. The patient's sex was encoded and mapped to [0, 1]. Then, two fully connected layers were used to regress out the BA. After each convolution, batch normalization (BN) and the rectified linear unit (RELU) were applied. Dropout was used in the fully connected layer at a rate of 0.5

the network, where 0 means male and 1 means female. Finally, we used two fully connected layers to further regress out the BA. After each convolution, batch normalization (BN) [13] and a rectified linear unit (RELU) [14] were applied. Dropout was used in the fully connected layer at a rate of 0.5.

In the training process, we employed the mean absolute error loss function as an optimization goal to train the BA regression model. Adam [15] updated the weight with a learning rate of 0.01 in the initialization phase and gradually decayed the learning rate as the epoch increased to obtain a better convergence effect.

Regional heatmap activation

We utilized the Grad-CAM [16] method to generate heatmaps to determine which part of an image was locally significant for fine-grained classification, as we rarely learned from existing clinical programs (e.g., G&P [2] and TW3 [17]). To utilize Grad-CAM, we extracted the feature from the last convolution layer of the network.

BA clinical determination

BA values from radiology reports and AI models tested on CHNt were analyzed by using Brush data to classify BA diagnoses [2]. The Brush data were used to classify a BA as normal (if the BA was limited to the range of ± 2 SD of the CA), delayed (if the BA was lower than −2 SD of the CA), or advanced (if the BA was higher than + 2 SD of the CA). Brush data reflect the variability of BA and are widely accepted as clinical determinations of BA. We classified BA diagnoses into three groups and performed a statistical analysis comparing human and AI estimations of BA in terms of Kappa values.

Statistical analysis

Bland–Altman plots and 95% LOA (mean ± 1.96 SD) were created to illustrate the BA difference between AI estimations of BA and reported BAs. Pearson correlation analysis was used to assess the correlation of AI-determined BAs and reported BAs. The performances of all AI models were evaluated in terms of mean values, the standard deviation (SD), the MAD, and the root mean square error (RMSE) of differences between the AI determinations of BA and reported BAs. The accuracies of all AI models were assessed using percentages of cases with the values of the absolute difference between AI and reported BAs within 0.5 years, 1 year, and 2 years.

When data were not distributed normally, nonparametric alternatives were used for comparing two or multiple groups (different sexes, CAs, radiologists, and clinical determinations), i.e., the Mann–Whitney *U* test or Kruskal–Wallis test, respectively. The classifications of BA diagnoses were analyzed by the chi-square test. The agreement of human and AI estimations was calculated by kappa values.

The statistical analyses were performed using SPSS 17.0 (SPSS Inc.). Differences were considered significant at $p < .05$. The figures were drawn using SPSS and GraphPad Prism v 5.0 software (GraphPad Software Inc.).

Results

Performance of deep learning models

The differences between BA estimations by three AI models tested on CHNt and USAt and radiologists' reports of BA values are shown in the Bland–Altman plot (Fig. 4a, b, c, d, e and f) with mean bias and 95% LOA. The percentages of scattered dots outside the 95% LOA were lowest on CHNm-USAt with 4.0% (42/1060) in Fig. 4 d but highest on USAm-USAt with 6.1% (65/1060) in Fig. 4e. The limits from the upper to the lower line of the 95% LOA were narrowest on CHNm-CHNt in Fig. 4a (−1.129 to 1.058 years) but broadest on CHNm-USAt in Fig. 4d (−2.285 to 1.664 years).

BAs determined by AI on CHNm-CHNt, USAm-USAt, JOIm-CHNt, and JOIm-USAt (all $r = 0.98$) showed a stronger correlation (linear) with reported BAs than BAs determined by AI on USAm-CHNt ($r = 0.96$) and CHNm-USAt ($r = 0.95$).

The results of the internal validation (CHNm-CHNt, USAm-USAt, JOIm-CHNt, JOIm-USAt) and the external validation (CHNm-USAt, USAm-CHNt) were analyzed to evaluate AI performance. The internal validation is that the training and test datasets are from the same institution. The external validation is that the training and test datasets are from the two separate institutions. Table 2 shows the summary statistics of the accuracy of CHNm, USAm, and JOIm tested on CHNt and USAt. In terms of the MAD, RMSE, and accuracy (with percentage) of the difference between AI and reported BAs within 0.5 years, 1 year, and 2 years, CHNm-CHNt outperformed USAm-CHNt, whereas USAm-USAt outperformed CHNm-USAt. Fortunately, better performances of JOIm validated on the two test datasets were obtained, and they were similar to internal validations separately. This bias may come from some factors, as we show in the following.

The distribution of the absolute differences (ADs) between reported BAs and BAs determined by three AI models tested on CHNt with patients' sex and CA, radiologists, and clinical classifications of BA diagnoses are shown by box plots (Fig. 5 a, b, c and d). All AD values were calculated as medians because of the non-normal distribution. In Fig. 5a and Table 2 (rows 3–5), the medians of AD values among females were lower than those among males on CHNm-USAt (0.66/0.83 years, $p < .001$) and USAm-CHNt (0.50/0.75 years, $p < .001$), higher than those among males on JOIm-USAt (0.45/0.31 years), but similar to those among males in other models ($p > .05$). This result indicated that sex has a varied effect on the accuracy of BA estimation. Figure 5b shows that the values were larger for USAm than for CHNm and JOIm for extremely small CAs (2–5 years), small for middle CAs (6–14 years), and large for extremely large CAs (15–17 years) in all three models. The image numbers were very small (< 40 cases) for the small and large CA groups but relatively large for the middle CA group. In Fig. 5c, USAm had a larger BA difference regarding all radiologists ($p = .023$) than CHNm and JOIm (both $p > .05$). The medians of AD values of AI and reported BAs for all radiologists ranged from 0.25 to 0.42 years on CHNm, 0.42 to 0.96 years on USAm, and 0.25 to 0.79 years on JOIm. In Fig. 5d, the normal group presented a smaller BA difference than the advanced and delayed groups. The highest performance was observed for CHNm-CHNt in the normal group (0.39 years), and the lowest performance was observed for USAm-CHNt in the delayed group (1.02 years).

The effect of interobserver variability of disputed cases on CHNm performance was analyzed.

Sixty-nine disputed cases were outside the 95% LOA (> 1.022 years, < −1.166 years) of the difference between AI-determined BA on CHNm-CHNt and reported BA in Fig. 4a and were rerated. The SD of the difference between the rerated BA and reported BA was 0.88 years. The average rerated BA and reported BA as a new manual BA was 10.73 years, which was closer to the mean AI-determined BA (10.70 years) than the reported BA (10.59 years). The proportion of disputed cases where the new manual BA agreed better with the CHNm-CHNt AI-determined BA was 65.2% (45/69). This finding is interesting and lead us to consider whether AI would outperform individual doctors in estimating BA.

Regional heatmap

The values of hot spots in the heatmaps of every radiograph can be transformed into the range of 0–1. We intentionally separated the values into groups according to patients' sex and age in accordance with the GP atlas, and in every group, we obtained the average value to better show each group characteristic. In Fig. 6, the first row shows a typical X-ray film.

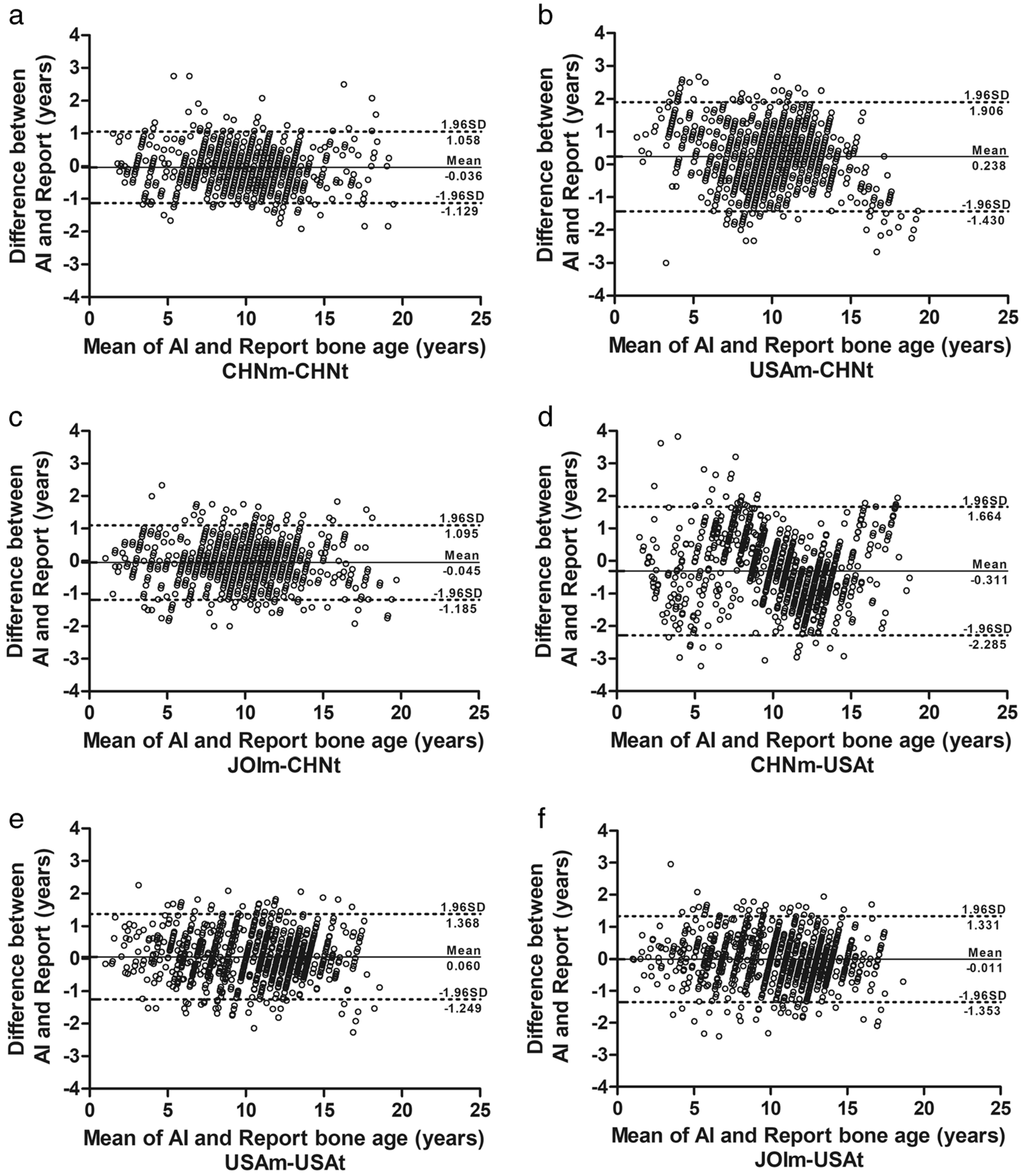

Fig. 4 Bland–Altman plot showing the difference between AI and reported BAs. **a** CHNm-CHNt. **b** USAm-CHNt. **c** JOIm-CHNt. **d** CHNm-USAt. **e** USAm-USAt. **f** JOIm-USAt. The solid line represents the mean difference, and the dotted lines represent the 95% LoA. The percentages of scattered dots outside the 95% LOA were lowest on CHNm-USAt with 4.0% (42/1060) in d but highest on USAm-USAt with 6.1% (65/1060) in e. The limits from the upper to the lower line of the 95% LOA were narrowest on CHNm-CHNt in a (−1.129 to 1.058 years) but broadest on CHNm-USAt in d (−2.285 to 1.664 years)

The second row shows the heatmap pictures, the third shows the SD values, and the fourth shows the variation values.

These heatmaps showed the AI vision, which may explain the black box of AI. For younger children, such as the 5-year-old male group (column 1 in Fig. 6), the heatmap focused more on the phalanxes and less on the carpals, but radiologists focused more on the carpals according to the GP atlas. For older children, such as the 14-year-old male group (column 3 in Fig. 6), the heatmap focused moderately more on the carpals, but radiologists focused more on the metacarpals and the radius. For the hands of 18-year-old boys, most AI hot spots were shown on the carpal area. For children in the middle-age group, for example, for 8-year-old boys, AI focused more on the phalanxes and moderately more on the carpals, but radiologists focused more on the phalanxes, metacarpals, and carpals. This result indicated that AI and humans might focus on

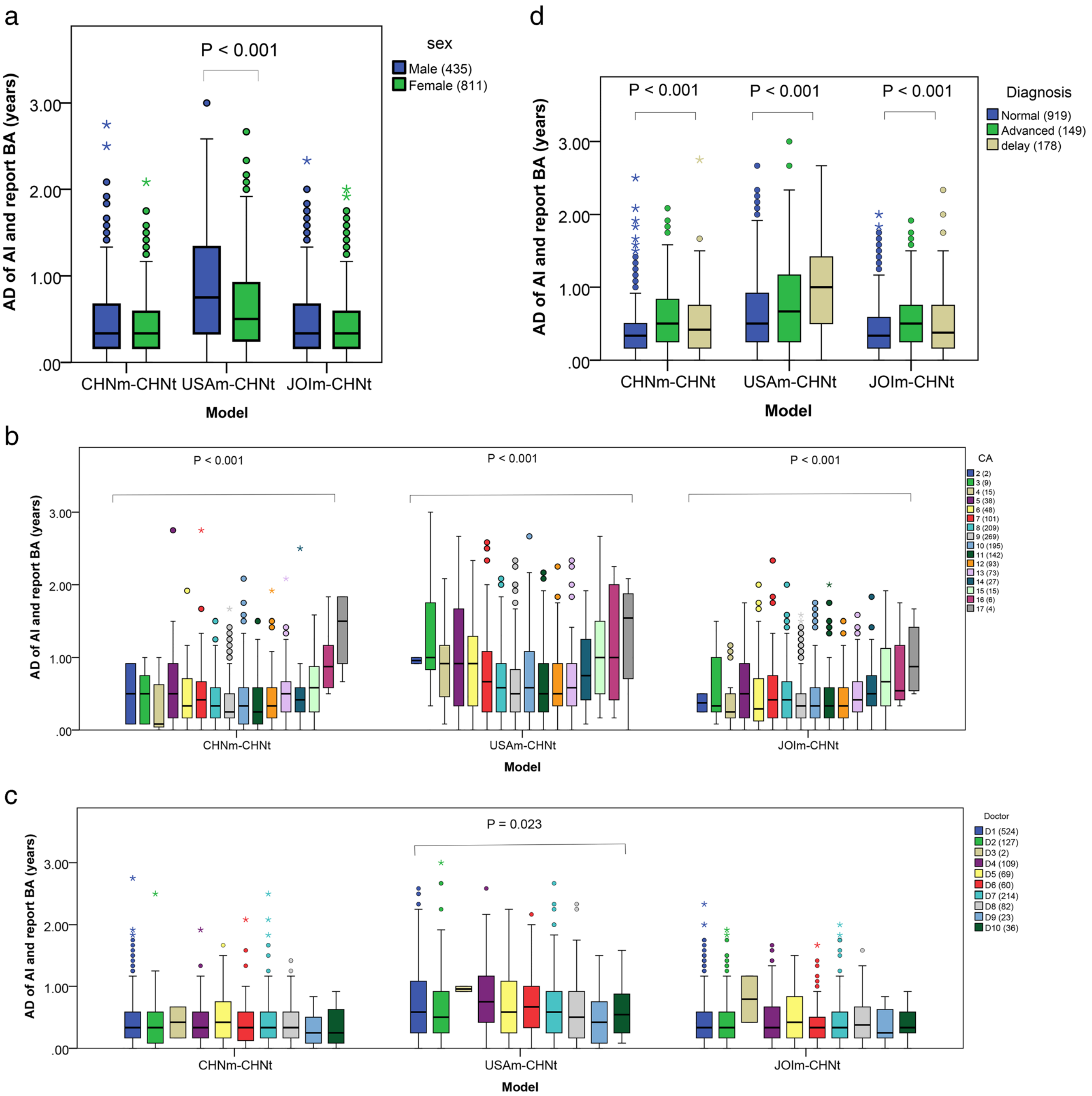

Fig. 5 Box plot showing the median and quartiles of BA absolute difference (AD) distribution of three AI models tested on CHNt and reported between (**a**) different sexes, (**b**) CAs, (**c**) radiologists, and (**d**) BA diagnoses. *p* values are shown when the differences in BA AD values between AI and reports between different groups on each model were significant. The difference between sexes was analyzed using the Mann–Whitney *U* test. The differences among different CA groups, different radiologists, and different BA diagnoses were analyzed using the Kruskal–Wallis test. The numbers of images for each factor are shown in brackets

different regions on the hand bone when making a decision on BA estimation.

BA clinical determination

The distributions of BA diagnosis of 1246 test images from CHNt are shown in Table 3. As indicated by the *p* values, compared with the radiology report, the clinical classifications of CHNm and JOIm (both $p > .05$) showed no difference; however, those of USAm ($p < .001$) showed a difference. The results of CHNm and JOIm in evaluating BA would seem to be perfect. We further individually matched the 1246 subjects and showed that the Kappa values were 0.714 on CHNm, 0.716 on JOIm, and 0.53 on USAm ($p < .001$), as shown in

Table 2 Summary statistics of the difference between AI and reported BAs

Variable	CHNm		USAm		JOIm	
	CHNt	USAt	CHNt	USAt	CHNt	USAt
Mean (SD), year	−0.04 (0.56)	−0.31 (1.01)	0.24 (0.85)	0.06 (0.67)	−0.05 (0.58)	−0.01 (0.68)
MAD, y	0.42	0.85	0.70	0.52	0.44	0.52
Male, mean/median	0.46/0.33	0.93/0.83	0.86/0.75	0.51/0.39	0.48/0.33	0.50/0.31
Female, mean/median	0.40/0.33	0.76/0.66	0.62/0.50	0.52/0.42	0.43/0.33	0.55/0.45
p values	0.363	< .001*	< .001*	0.262	0.375	0.018*
RMSE, year	0.56	1.05	0.88	0.67	0.58	0.68
Male	0.64	1.13	1.05	0.67	0.64	0.67
Female	0.51	0.95	0.78	0.67	0.55	0.70
#Accuracy within 0.5 years,%	71.19	34.15	46.63	59.47	69.58	57.64
Male	57.80	29.34	33.26	60.7	59.17	60.07
Female	72.82	39.88	50.80	58.68	69.74	54.75
#Accuracy within 1 year, %	94.06	65.09	76.16	85.75	92.30	85.94
Male	88.30	60.07	61.01	85.76	86.47	86.98
Female	95.69	71.07	82.29	85.74	94.10	84.71
#Accuracy within 2 years, %	99.52	95.47	97.59	99.53	99.92	99.43
Male	98.62	94.10	94.50	99.48	99.08	99.48
Female	99.63	97.11	97.79	99.59	99.63	99.28

*It means significant difference of MAD values in male and female with Mann–Whitney *U* test

# Accuracy: percentage of difference between AI BA and report BA

SD standard deviation, *MAD* mean absolute difference

Table 4. Our results indicated that there was no high agreement between these models and radiologists.

Discussion

In this study, we evaluated the performance of bone age deep learning models established by using our hospital clinical data and RSNA data. In our study, the best MAD value was 0.42 years on CHNm-CHNt. This finding was as accurate and precise as those of previous studies, in which the MADs ranged from 0.38 to 0.64 [18–24]. However, the performance of BA models was worse with external validation (CHNm-USAt and USAm-CHNt) than with internal validation (CHNm-CHNt and USAm-USAt). The results were consistent with the studies of Larson [19] and Koita [23].

An interesting finding of this study is the heatmaps of hands. We found that AI heatmaps were not fully consistent with human focusing areas based on the GP atlas. This finding indicated on what areas and how AI focuses to some extent; this focus has not been described in previous studies. We are not the first to report heatmaps on the AI BA model [21]. However, we showed a different heatmap from previous findings, with a heatmap generated by the whole hand, not by several partial hand regions, as shown in the previous study.

Regarding the clinical determination of BA in the normal, advanced, and delayed groups, our study and Larson [19] both found that the chi-square test showed no difference between AI and human clinical determinations of BA. However, further analysis of kappa values indicated no high agreement (kappa 0.714) between AI and humans. This result questions the performance of AI in BA diagnosis. Moreover, USAm tested on external data showed the worst agreement between AI and humans, with the broadest limit on the Bland–Altman figure and lowest kappa values (0.53). This result further reflects the generalization problem when AI faces external and new data.

The effects of patients, institutions, and radiologists on AI performance were also assessed. We observed several biases between AI and radiologists; these biases include children's sex and age, institutions, and radiologists. Apparently, the data biases mentioned above may cause variable AI performance by the three BA models and disagreement in BA diagnosis between AI and radiologists.

As the GP atlas shows, males and females of the same age have different BA atlases. Children of the same age have different characteristics of the BA atlas, and the same characteristics may belong to different ages. Lee et al found that a higher level of MAD errors is seen for the female cohort in a sex-aware model; this finding may suggest that a relatively higher growth rate of the female cohort causes greater deviation from the nominal growth trajectory for individual subjects [25]. Compared to

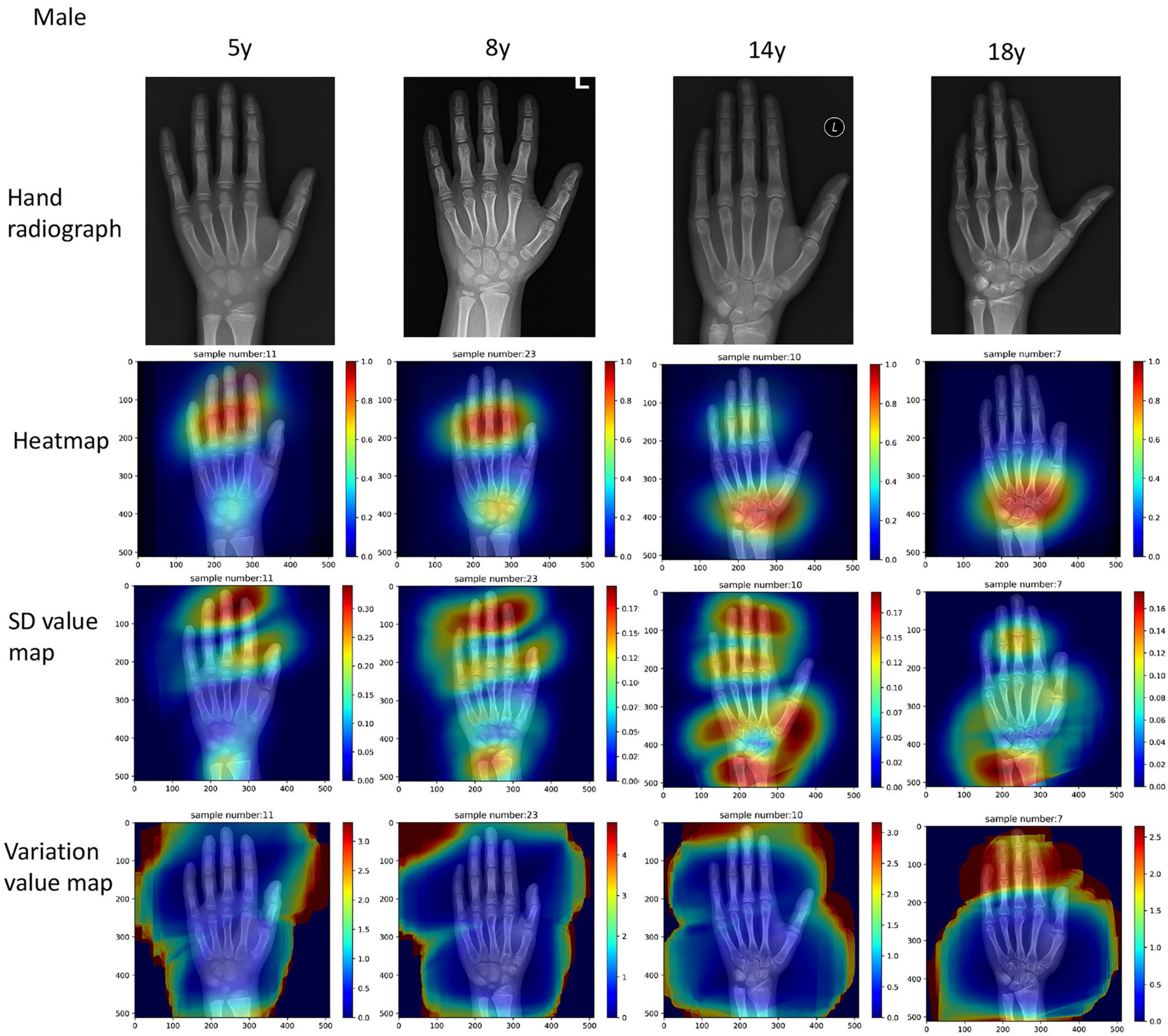

Fig. 6 Samples of hand radiograph (row 1), heatmap (row 2), standard deviation (SD) value (row 3), and variation value map (row 4) of 5 years, 8 years, 14 years, and 18 years in males. These attention heatmaps show the AI vision. For younger children, such as the 5-year-old male group, the heatmap focused more on phalanxes, but radiologists focused more on carpals according to the GP atlas. For older children, such as the 14-year-old male group, the heatmap focused more on carpals, but radiologists focused more on the metacarpal and radius

Lee's finding, our results are varied. We observed a significant difference in MAD values between males and females for the three test datasets (CHNm-USAt, USAm-CHNt, and JOIm-USAt). These results are shown in Table 2 in the results section ("*" means $p < .05$). A lower MAD indicates higher AI performance on the females of CHNm-USAt, the females of USAm-

Table 3 The clinical classifications of bone age diagnosis by using CHNt (test dataset from China, N=1246) for three models and radiology reports

Classifications	Radiology report	CHNm-CHNt	USAm-CHNt	JOIm-CHNt
Advanced	149	117	144	118
Normal	919	955	986	954
Delayed	178	174	116	174
p values[#]	/	.10	< .001	.12

[#] The difference of the clinical classifications between each AI model and report with X^2 test

Table 4 The kappa values of clinical classifications of bone age diagnosis of CHNt (test dataset from China, $N = 1246$) for three models and radiology reports

Deep learning BA models	Radiology report BA			Kappa value (95% CI)	*p* value
	Advanced	Normal	Delayed		
CHNm-CHNt				0.714 (0.671–0.757)	< .001
Advanced, no. (%)	99 (7.95)	18 (1.44)	0		
Normal, no. (%)	50 (4.01)	865 (69.42)	40 (3.21)		
Delayed, no. (%)	0	36 (2.89)	138 (11.08)		
USAm-CHNt				0.530 (0.477–0.583)	< .001
Advanced, no. (%)	96 (7.78)	48 (3.85)	0		
Normal, no. (%)	52 (4.17)	839 (67.34)	95 (7.62)		
Delayed, no. (%)	1 (0.08)	32 (2.57)	83 (6.66)		
JOIm-CHNt				0.716 (0.673–0.759)	< .001
Advanced, no. (%)	99 (7.95)	19 (1.52)	0		
Normal, no. (%)	50 (4.01)	865 (69.42)	39 (3.13)		
Delayed, no. (%)	0	35 (2.81)	139 (11.16)		

CI confidence interval

CHNt, and the males of JOIm-USAt. A higher MAD indicates lower performance on the males of CHNm-USAt, the males of USAm-CHNt, and the females of JOIm-USAt. The capability of AI performance is not consistent with the sex classification.

The worse performance of BA models with external validation implies that the AI model is not fitted across different sites [9]. This institutional bias is due mainly to the different physical characteristics of the population from different institutions. For example, some studies demonstrated the application of the GP atlas to assess bone age in children of diverse ethnicities [26] and indicated cross-racial growth differences between Asian and White children [27]. Our results indicated better validation in internal datasets but poor validation across "institutions," mainly because of the different populations. Our JOIm implemented with combined data from China and America showed better performance. Mutusa et al joined private and public data and built a BA model [24] and an increasing number of datasets from different institutions to solve the generalization problem across institutions.

Our results also showed larger differences between AI and radiology reports in the abnormal BA group than in the normal BA group possibly because skeletal maturation inconsistencies in carpals and tubular bones in disease conditions make interpreting BA with reference to the GP atlas, which was based on a typical child's bone structure, more difficult. These conditions include growth hormone deficiency [28], obesity [29], and chronic renal insufficiency [30].

Our study has three limitations. First, the population distribution in this study is a limitation and challenge. The distribution of males and females was not even, nor was the age distribution. For example, the chronological ages of the CHN training dataset showed an approximate Gaussian distribution. The 10-year-old population accounted for 18.5% and was the peak. However, the numbers of younger and older children in this dataset were lower. This age distribution would affect the accuracy of the model. Second, the "labeling rule" is another limitation of our study. The test data CHNt ($n = 1246$) were labeled by ten senior pediatric radiologists. The interobserver variability would be a limitation. This variation could be decreased by averaging two or more reads. In our study, 69 disputed data were rerated, and the average was closer to AI than to either of the reads. This finding indicates that rerating the BA of radiographs may help improve the AI determination of BAs. The result is supported by Van Rijn [3] and Mutasa [24]. Third, we think the clinical determination of bone age may be a limitation and value in this study. The human–machine comparative performance aspect of this study was presented in terms of not only the MAD value but also the classification of child development. The clinical determination of bone age is based only on the BA value from the wrist and hand radiographs, chronological age, and SD value from Brush data. More clinical information from pediatricians is needed to evaluate pediatric growth conditions.

In the future, our study will focus on assessing pediatric growth conditions, not just bone age assessments. Because pediatric growth conditions are not assessed just by bone age, some clinical history and lab data are very useful. Classification of BA diagnosis was made only by Brush data and radiology reports, which may not be accurate for assessing child development without clinical data and other physical examinations. The data science Venn diagram by Drew Conway [31] indicated that there was a danger zone when big data were mined by using domain knowledge and hacking skills. Our future study will focus on the collection of clinical data of BA diagnosis and visual saliency maps to provide "explicability" for AI to improve the accuracy of BA prediction.

The deep learning models outperformed external validation in predicting BA on both internal and joint datasets. However, the AI models' clinical determinations of bone age were not in high agreement with the clinical determinations by radiologists. Several factors, including patients' sex and age, institutions, and radiologists, contributed to the bias of AI performance. The heatmaps of bone age were useful in clarifying how AI made decisions.

Acknowledgements Thanks for the relevant work go to these people: Chenjie Ye (Children's Hospital of Fudan University), Haoting Feng (GE Company), Xiaolin Ge (Children's Hospital of Fudan University), Qi Chen (Children's Hospital of Fudan University), Chuqing Pan (Ping An International Smart City, China), Zhenzhong Xie (Ping An International Smart City, China), Haozhe Si (University of llinois at Urbana-Champaign).

Funding This study has received funding by Shanghai Municipal Health Commission (No. 2019SY038. Establishment of bone age AI model for normal children in Shanghai and prediction of bone age for common endocrine diseases in children. PI: Zhongwei Qiao).

Declarations

Guarantor The scientific guarantor of this publication is Zhongwei Qiao, who is the director of department of radiology in Children's Hospital of Fudan University.

Conflict of interest The authors of this manuscript declare relationships with Ping An Technology. Liangxin Gao, Jianbang Ge, Lingyun Huang, and Jing Xiao are staff members of Ping An Technology, who implement the bone age models. Patients' data for the bone age models did not include patients' private information, but only images, sexes, and chronological age. These authors declare that there is no conflict of interest regarding the publication of this paper. The other authors (from our hospital) report no competing interests.

Statistics and biometry Xiaotian Chen kindly provided statistical advice for this manuscript, who is a staff of department of clinical epidemiology in the Children's Hospital of Fudan University.

Informed consent Written informed consent was waived by the ethics committee of the Children's Hospital of Fudan University.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- retrospective
- diagnostic or prognostic study
- performed at one institution

References

1. Creo AL, Schwenk WF 2nd (2017) Bone age: a handy tool for pediatric providers. Pediatrics 140(6):e20171486
2. Greulich WW, Pyle SI (1959) Radiographic atlas of skeletal development of the hand and wrist, 2nd edn. Stanford University Press, Stanford, California
3. Van Rijn RR, Thodberg HH (2013) Bone age assessment: automated techniques coming of age? Acta Radiol 54:1024–1029
4. Lee H, Tajmir S, Lee J et al (2017) Fully automated deep learning system for bone age assessment. J Digit Imaging 30:427–441
5. Summers RM (2018) Deep learning lends a hand to pediatric radiology. Radiology 287:323–325
6. Nadeem MW, Goh HG, Ali A, Hussain M, Khan MA, Ponnusamy VAP (2020) Bone age assessment empowered with deep learning: a survey, open research challenges and future directions. Diagnostics (Basel) 10:781
7. Halabi SS, Prevedello LM, Kalpathy-Cramer J et al (2018) The RSNA pediatric bone age machine learning challenge. Radiology 290:498–503
8. Siegel EL (2018) What can we learn from the RSNA pediatric bone age machine learning challenge? Radiology 290:504–505
9. Yasaka K, Abe O (2018) Deep learning and artificial intelligence in radiology: current applications and future directions. PLoS Med 15: e1002707
10. Choy G, Khalilzadeh O, Michalski M et al (2018) Current applications and future impact of machine learning in radiology. Radiology 288:318–328
11. Ronneberger O, Fischer P, Brox T (2015) U-Net convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 234-241. Available via https://link.springer.com/content/pdf/10.1007/978-3-319-24574-4_28.pdf. https://doi.org/10.48550/arXiv.1505.04597
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for Image Recognition2016 IEEE Conference on computer vision and pattern recognition (CVPR), pp 770-778. Available via https://arxiv.org/abs/1512.03385. https://doi.org/10.48550/arXiv.1512.03385
13. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shiftProceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. JMLR.org, Lille, France, pp 448–456. Available via https://arxiv.org/abs/1502.03167. https://doi.org/10.48550/arXiv.1502.03167
14. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machinesProceedings of the 27th International Conference on International Conference on Machine Learning. Omnipress, Haifa, Israel, pp 807–814. Available via http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=33D549C0E5AD0F9F9593A3FDE2309E35?doi=10.1.1.165.6419&rep=rep1&type=pdf: https://doi.org/10.5555/3104322.3104425
15. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego. Available via https://arxiv.org/abs/1412.6980v9. https://doi.org/10.48550/arXiv.1412.6980
16. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128:336–359
17. Tanner JG, Healy MJR, Goldstein H, Cameron N (2001) Assessment of skeletal maturity and prediction of adult height: TW3 method. W.B Saunders Company, London, United Kindom
18. Kim JR, Shim WH, Yoon HM et al (2017) Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. AJR Am J Roentgenol 209:1374–1380
19. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP (2017) Performance of a deep-learning neural network

model in assessing skeletal maturity on pediatric hand radiographs. Radiology 287:313–322

20. Tong C, Liang B, Li J, Zheng Z (2018) A deep automated skeletal bone age assessment model with heterogeneous features learning. J Med Syst 42:249
21. Ren X, Li T, Yang X et al (2019) Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. IEEE J Biomed Health 23:2030–2038
22. Tajmir SH, Lee H, Shailam R et al (2019) Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. Skeletal Radiol 48:275–283
23. Koitka S, Kim MS, Qu M, Fischer A, Friedrich CM, Nensa F (2020) Mimicking the radiologists' workflow: estimating pediatric hand bone age with stacked deep neural networks. Med Image Anal 64:101743
24. Mutasa S, Chang PD, Ruzal-Shapiro C, Ayyala R (2018) MABAL: a novel deep-learning architecture for machine-assisted bone age labeling. J Digit Imaging 31:513–519
25. Lee JH, Kim YJ, Kim KG (2020) Bone age estimation using deep learning and hand X-ray images. Biomed Eng Lett 10:323–331
26. Ontell FK, Ivanovic M, Ablin DS, Barlow TW (1996) Bone age in children of diverse ethnicity. AJR Am J Roentgenol 167:1395–1398
27. Zhang A, Sayre JW, Vachon L, Liu BJ, Huang HK (2009) Racial differences in growth patterns of children assessed on the basis of bone age. Radiology 250:228–235
28. Hernandez R, Poznanski AK, Kelch RP, Kuhns LR (1977) Hand radiographic measurements in growth hormone deficiency before and after treatment. AJR Am J Roentgenol 129:487–492
29. Polito C, Di Toro A, Collini R, Cimmaruta E, D'Alfonso C, Del Giudice G (1995) Advanced RUS and normal carpal bone age in childhood obesity. Int J Obes Relat Metab Disord 19:506–507
30. Polito C, Greco N, Opallo A, Cimmaruta E, La Manna A (1994) Alternate-day steroids affect carpal maturation more than radius, ulna and short bones. Pediatr Nephrol 8:480–482
31. Convay D (2010) The data science Venn diagram. Available via http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.