ORIGINAL ARTICLE

Artificial intelligence algorithm improves radiologists' bone age assessment accuracy

Tien-Yu Chang[a], Ting Ywan Chou[b,c], I-An Jen[d], Yeong-Seng Yuh[e,f,*]

[a]Department of Radiology, Cheng-Hsin General Hospital, Taipei, Taiwan, ROC; [b]Department of Radiology, Cardinal Tien General Hospital, Taipei, Taiwan, ROC; [c]College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan, ROC; [d]Institute of Public Health, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, ROC; [e]Department of Pediatrics, Cheng-Hsin General Hospital, Taipei, Taiwan, ROC; [f]Department of Pediatrics, National Defense Medical Center, Taipei, Taiwan, ROC

Abstract

Background: Artificial intelligence (AI) algorithms can provide rapid and precise radiographic bone age (BA) assessment. This study assessed the effects of an AI algorithm on the BA assessment performance of radiologists, and evaluated how automation bias could affect radiologists.

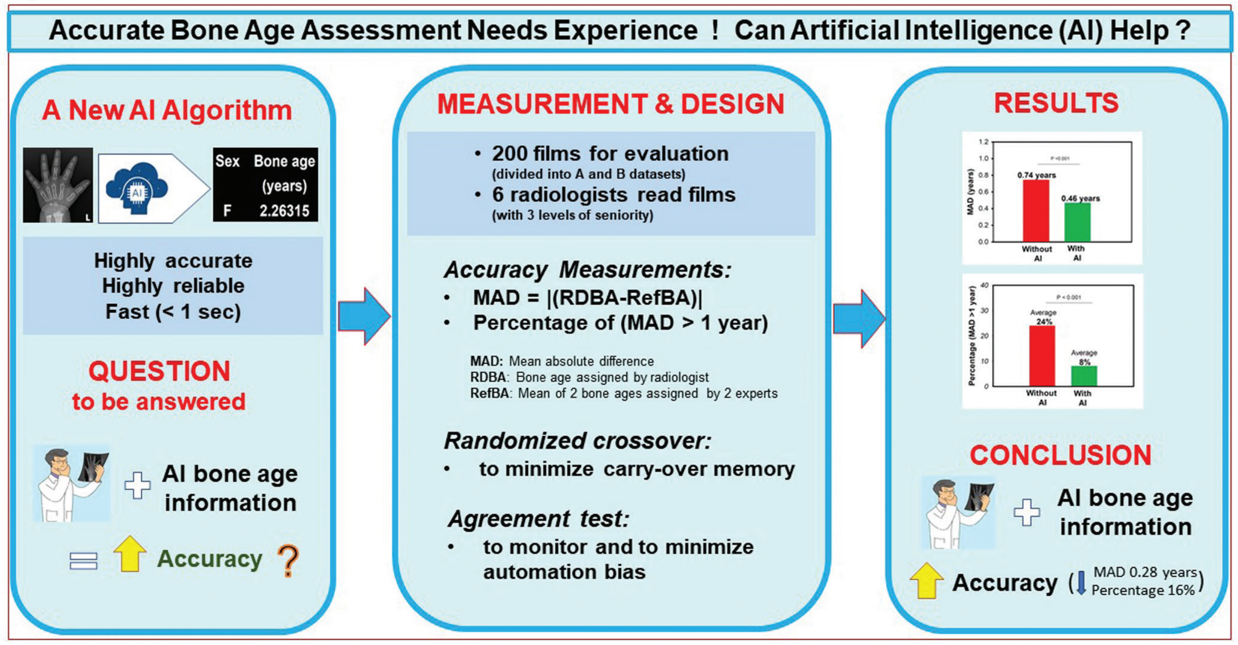
Methods: In this prospective randomized crossover study, six radiologists with varying levels of experience (senior, mid-level, and junior) assessed cases from a test set of 200 standard BA radiographs. The test set was equally divided into two subsets: datasets A and B. Each radiologist assessed BA independently without AI assistance (A– B–) and with AI assistance (A+ B+). We used the mean of assessments made by two experts as the ground truth for accuracy assessment; subsequently, we calculated the mean absolute difference (MAD) between the radiologists' BA predictions and ground-truth BA and evaluated the proportion of estimates for which the MAD exceeded one year. Additionally, we compared the radiologists' performance under conditions of early AI assistance with their performance under conditions of delayed AI assistance; the radiologists were allowed to reject AI interpretations.

Results: The overall accuracy of senior, mid-level, and junior radiologists improved significantly with AI assistance than without AI assistance (MAD: 0.74 vs 0.46 years, $p < 0.001$; proportion of assessments for which MAD exceeded 1 year: 24.0% vs 8.4%, $p < 0.001$). The proportion of improved BA predictions with AI assistance (16.8%) was significantly higher than that of less accurate predictions with AI assistance (2.3%; $p < 0.001$). No consistent timing effect was observed between conditions of early and delayed AI assistance. Most disagreements between radiologists and AI occurred over images for patients aged ≤8 years. Senior radiologists had more disagreements than other radiologists.

Conclusion: The AI algorithm improved the BA assessment accuracy of radiologists with varying experience levels. Automation bias was prone to affect less experienced radiologists.

Keywords: Age determined by skeleton; Artificial intelligence; Radiologists

Graphical abstract



Accurate Bone Age Assessment Needs Experience ! Can Artificial Intelligence (AI) Help ?

A New AI Algorithm

Highly accurate
Highly reliable
Fast (< 1 sec)

QUESTION
to be answered

AI bone age information

Accuracy ?

MEASUREMENT & DESIGN

- 200 films for evaluation
(divided into A and B datasets)
- 6 radiologists read films
(with 3 levels of seniority)

Accuracy Measurements:

- MAD = |(RDBA-RefBA)|
- Percentage of (MAD > 1 year)

MAD: Mean absolute difference
RDBA: Bone age assigned by radiologist
RefBA: Mean of 2 bone ages assigned by 2 experts

Randomized crossover:

- to minimize carry-over memory

Agreement test:

- to monitor and to minimize automation bias

RESULTS

CONCLUSION

AI bone age information

Accuracy (MAD 0.28 years
Percentage 16%)

Chang TY et al. | Journal of the Chinese Medical Association (JCMA) | 2025; 88:(7).

Lay Summary: Bone age (BA) assessment is a measure of skeletal maturity and is used to predict a child's final adult height. The Greulich and Pyle plot method is a commonly used approach but is considered tedious, time-consuming, and requires experience. Four years ago, we developed an artificial intelligence (AI) model for BA assessment. In this study, we demonstrated that the accuracy of radiologists' BA assessments can be improved by providing BA information assigned by this AI model.

1. INTRODUCTION

Radiographic bone age (BA) assessment is central to the clinical evaluation of patients with pediatric endocrine and metabolic disorders; this assessment entails comparing a patient's chronological age with their level of skeletal maturity on the basis of a standardized reference.[1] In clinical practice, BA assessment is typically performed using either the Greulich and Pyle[2] method or the Tanner-Whitehouse 3 method[3]; the Greulich and Pyle method entails comparing radiographs of the left hand and wrist with an age-based atlas, and the Tanner-Whitehouse 3 method entails using the scores of specific radiographic features for assessment. However, both methods are time-consuming and involve substantial interrater variability among radiologists.[4] Automated image evaluation is ideal for BA assessment because it involves the use of only a single image and the findings are standardized.

With the rapid development of artificial intelligence (AI) algorithms and improvements in computer hardware, AI-based programs for facilitating diagnosis are being increasingly applied in medicine. One of the earliest medical applications of AI in medicine was assisting radiologists in assessing BA.[5–8] Studies have proposed AI-based programs for BA assessment and have reported that such programs could achieve satisfactory diagnostic performance.[9–12] Some studies have demonstrated that the predictions of AI-based programs for BA assessment are as accurate as those of experts.[13,14] We previously developed an automated convolutional neural network model for BA assessment.[15] In the past 2 years, we refined this model and achieved a satisfactory mean absolute difference value, as comparing the model's BA predictions and reference standard BA data from a test set of samples. While a number of AI algorithms exist, to our knowledge, there is no United States Food and Drug Administration-approved version in North America. Validation of the automated software tool is essential to fully implementing it. The integration of AI algorithms into clinical work will be accompanied by various articulated concerns, including issues related to their applicability and inherent psychological bias. Researchers must determine whether AI algorithms really improve radiologists' performance and how they can be best integrated into radiology practice. Automation bias is a known source of psychological tendency in human-machine interactions.[16] This bias represents the tendency for humans to favor the suggestions from automated decision-making system. Its implications regarding AI-aided BA reading remain unknown.

The main purpose of this study was to determine whether the use of an AI algorithm as a diagnostic aid for radiologists assessing the BA improved their accuracy, compared with the assessment without AI aid, using a randomized crossover design in a prospective laboratory setting. The second purpose of this study was to determine how automation bias could affect varying experienced radiologists by allowing them to be disagree with the AI interpretations.

2. METHODS

2.1. Model implementation

The AI algorithm was trained on a dataset that included 14 036 images from the Radiological Society of North America (RSNA) Bone Age Challenge and 2358 images of Taiwanese children from Cheng-Hsin General Hospital (CHGH). The image data were collected from radiographs taken in pediatric endocrine clinic of CHGH from October 1, 2010 to March 31, 2020. The hospital's institutional review board approved the collection and usage of these images for constructing an AI algorithm for BA assessment. Among the 2358 images, 1971 images were used for training set, 187 images were used of validation set and 200 images were used for test set. The BAs for those 2358 images from Taiwanese children were defined by the mean values from the BA readings of a senior pediatric radiologist and a senior pediatric endocrinologist. The RSNA image data were released by the RSNA in 2017 (https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017).[10] The Institutional Review Boards of Stanford University and the University of Colorado approved the curation and use of pediatric hand radiographs for developing machine learning methods. Our hospital's institutional review board provided approval for the use of the image data and also waived the need for informed consent [IRB (108)111A-86].

Before training the algorithm, we converted each radiograph from Digital Imaging and Communications in Medicine (DICOM) format to Portable Network Graphics (PNG) format. Each of the radiographs contained images of the distal ulna, distal radius, carpal, metacarpal, and phalangeal bones and had a resolution of at least 1000 × 1000 pixels. The images were further downsized to 500 × 500 pixels by using the Python image library. The software architecture is presented in Fig. 1. The AI algorithm analyzed the hand images by examining the whole image of the hand and wrist. We defined the accuracy of this algorithm as the degree of agreement between its BA assessment results with the BA in an independent test set of 200 images of the bones of Taiwanese children. The 200 images, derived from 100 males and 100 females, have been used as a standard test set in our previous research.[17] This test set was independent of the training and validation sets; in this set, BA was defined as the average of ratings by a senior pediatric endocrinologist and a senior pediatric radiologist and served as the ground truth. We observed that the accuracy of our AI algorithm (MAD, 3.1 months) is superior to that of the set of winning algorithms in the RSNA challenge (MAD, 4.3 months).[10]

2.2. Ground-truth BA standard

Two trained and experienced reviewers, namely a pediatric radiologist and a pediatric endocrinologist with 39 and 36 years of experience in BA reading, respectively, used the Greulich and Pyle atlas to establish the ground-truth standard. The reliability as defined by intra-rater correlation coefficients for AI algorithm was 1.00, and was 0.993 for both experts. Both of them assessed all 200 standard cases through consensus. They were blinded to patient information, diagnosis, treatment, and previous BA reports and were provided only with the patient's sex information. No time limit was set for the assessment of the radiographs. In the event of a substantial disagreement, a discussion

*Address correspondence. Dr. Yeong-Seng Yuh, Department of Pediatrics, Cheng-Hsin General Hospital, 45, Cheng-Hsin Road, Taipei 112, Taiwan, ROC. E-mail address: yuhyeongseng@gmail.com (Y.-S. Yuh).

Conflicts of interest: The authors declare that they have no conflicts of interest related to the subject matter or materials discussed in this article.

Journal of Chinese Medical Association. (2025): 88: 530-537.

Received August 18, 2024; accepted March 7, 2025.

doi: 10.1097/JCMA.0000000000001248

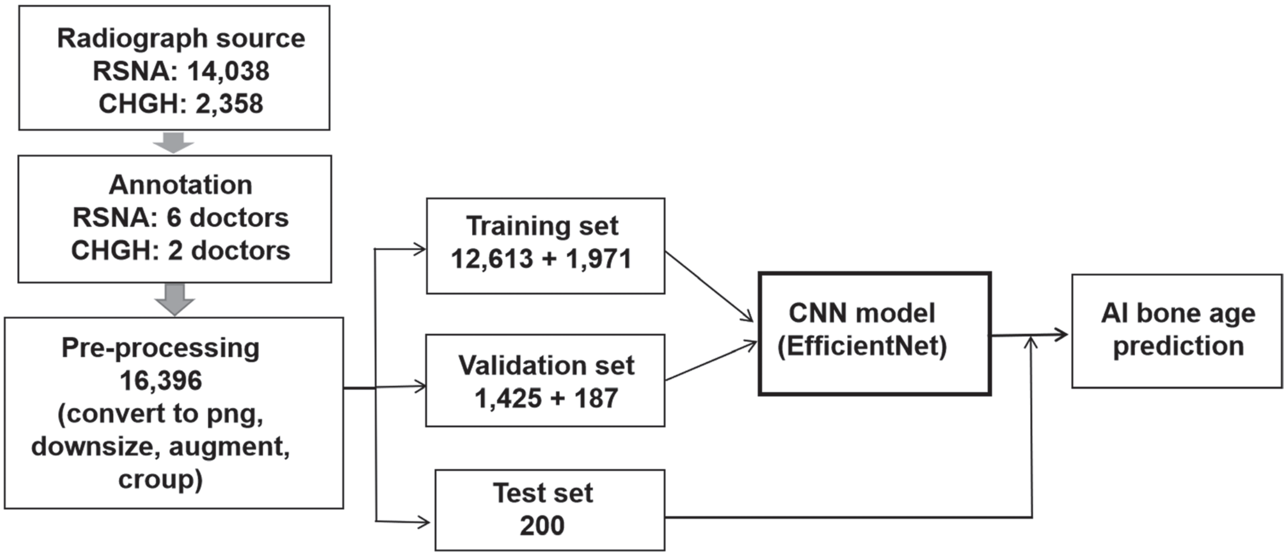


Fig. 1 Artificial intelligence algorithm architecture.

between the two reviewers was arranged. The mean value of the BAs assigned by the two experts was used as the ground-truth BA standard for subsequent assessments.

2.3. Assessment performance of radiologists with and without AI assistance

As mentioned, we evaluated influence of the implementation of the AI algorithm on radiologists' BA assessment accuracy. Six radiologists with different seniority were recruited. The seniority of radiologists was determined by their service time in the radiology department (two senior radiologists, with more than 15 years; two mid-level radiologists, with 6 to 10 years; two junior radiologists, with <3 years of experience). Specifically, each radiologist independently assessed 200 identified standard radiographs by using the Greulich and Pyle method either with (+) or without (–) the aid of AI information. Each of the radiologists was blinded to the others' results and was only informed about each candidate's sex. The 200 standard films were evenly divided into two subsets: datasets A and B. Each subset contained bone images from 50 male and 50 female individuals aged one to 19 years. A randomized crossover design with four periods (each with a 4-week duration) was arranged to minimize anticipation and carryover effects. The research design is illustrated in Fig. 2. At the beginning of each month, radiologists received a packet on a universal serial bus (USB) 3.0 flash drive containing 100 images and an Excel file. In the dataset containing AI information (AI+), the BAs assigned by the AI were provided in the Excel file, and there were two columns of assignment markers indicating the radiologist's comments (agree or disagree) on the AI information. Each radiologist who made an assessment with the aid of AI information could agree or disagree with the AI's predictions. The BA estimated by radiologist was used to compared with the ground-truth BA, no matter radiologists agreed or disagreed with the AI estimated BA. In the dataset without AI information (AI–), there were no AI BA information and comment column. In the image number column of Excel file, each number was connected to the corresponding image. Click the image number and the radiograph image would be displayed on the screen. Radiologists could assess BA and filled the data into corresponding cells in Excel columns. The order of images in Excel was randomized each month to minimize carryover effects. BA data were collected at the end of each month. The effect of timing for providing AI information was evaluated by comparing the statistically *p* values (derived from paired *t* test for with or without AI assistance BA difference) between providing AI assistance immediately or with a delay.

2.4. Statistical analysis

Several statistical methods were used to compare the AI algorithm's predictions with the human reviewers' assessment results. Lin's[18] concordance correlation coefficient (CCC) is a measure of the degree of conformity of bivariate pairs of observations to a standard; it was used to evaluate the accuracy and precision of AI algorithm. Bland-Altman plots were used to calculate the mean and 95% CI of the difference between the ground-truth BA and AI algorithm's predictions.[19] Student's *t* test was used to compare ground-truth BA between dataset A and dataset B. We assessed the normality of continuous variables by using skewness and kurtosis tests. The MAD was used for accuracy evaluation; it was calculated as the mean of the absolute value of the difference between the radiologist-estimated BA and the ground-truth BA. The proportion of estimates for which the MAD exceeded one year was calculated, and the proportions of cases of improved accuracy (decreased MAD) and decreased accuracy (increased MAD) were recorded. McNemar's test was used to compare these proportions. Statistical differences were considered significant at $p < 0.05$. All statistical analyses were performed using SPSS software (v.22.0; SPSS Inc., Chicago, IL), and Bland-Altman plots were created using SigmaPlot (version 12.5; Systat Software Inc., San Jose, CA).

3. RESULTS

3.1. Ground-truth BA distribution of A/B datasets and model performance

There was no significant statistical difference between datasets A and B ($p > 0.05$). Fig. 3 displays a box plot of the age distributions for datasets A and B. Fig. 4 presents a Bland-Altman plot of the difference between the AI algorithm's BA predictions and ground-truth BA. The mean difference was –0.04 years, with 95% limits of agreement ranging from –0.41 to +1.33 years. Moreover, the

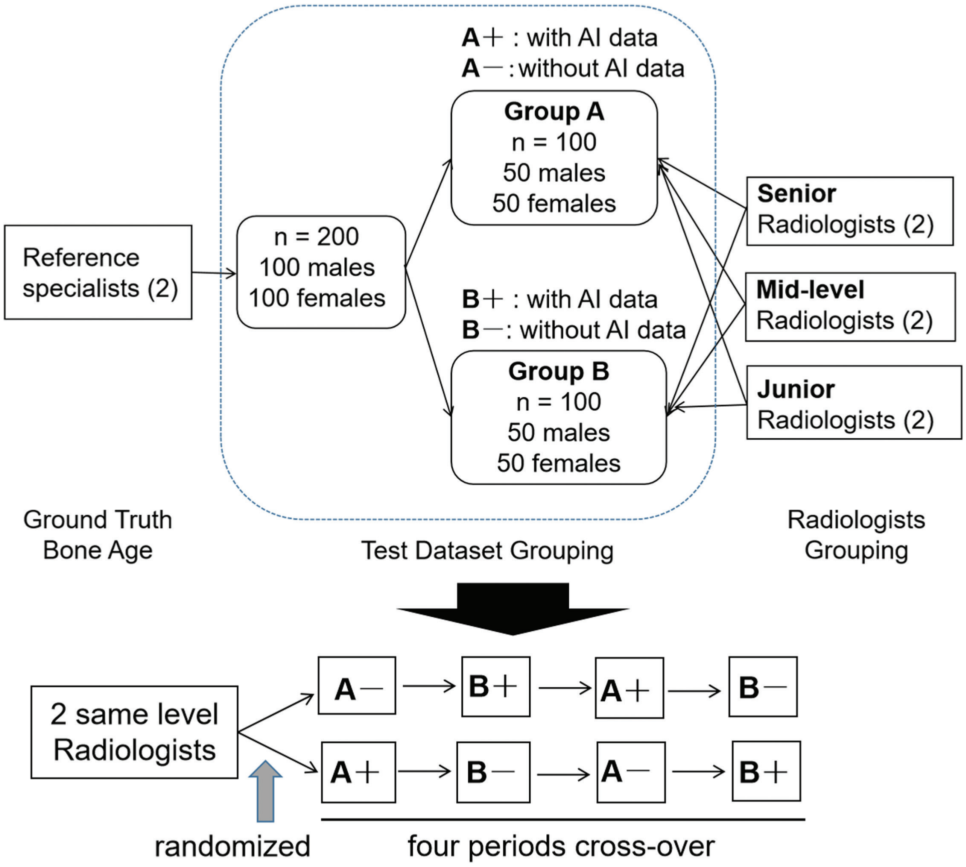


Fig. 2 Randomized crossover study design.

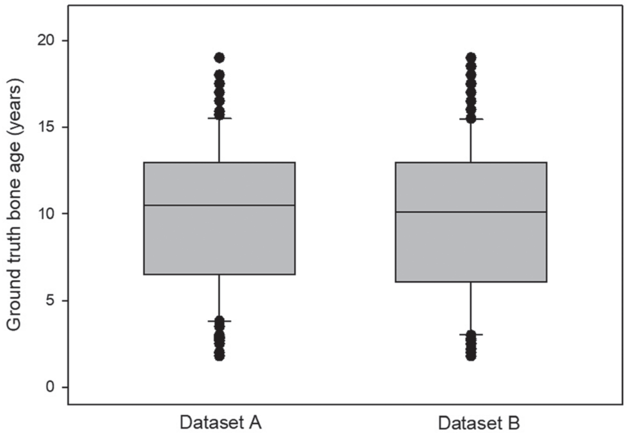


Fig. 3 Box plot for ground-truth bone age distribution of A and B datasets.

CCC between the AI algorithm's BA predictions and ground-truth BA was 0.997. The MAD between the AI algorithm's BA predictions and ground-truth BA was 0.261 ± 0.249 years.

3.2. Assessment performance of radiologists with and without AI assistance

As illustrated in Fig. 5, the radiologists' assessment accuracy improved significantly ($p < 0.001$) with AI assistance (MAD, 0.46 years) when compared with that without AI assistance (MAD, 0.74 years). The individual data of each radiologist are presented in Table 1. The proportions of assessments for which the MAD exceeded one year are presented in Table 2. These proportions were significantly lower with AI assistance than without such assistance ($p < 0.001$). The proportions of assessments with improved or decreased BA prediction accuracy with AI assistance are listed in Table 3. All the radiologists were affected by AI interpretation, either improved or worsened their BA assignments. The proportions of assessments with improved BA prediction accuracy were significantly higher than those of

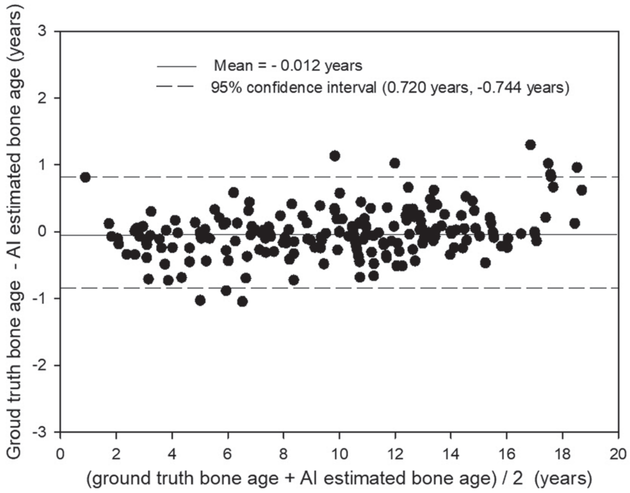


Fig. 4 Bland-Altman scatter plot of comparisons between AI-estimated bone age and ground-truth bone age. AI = artificial intelligence.

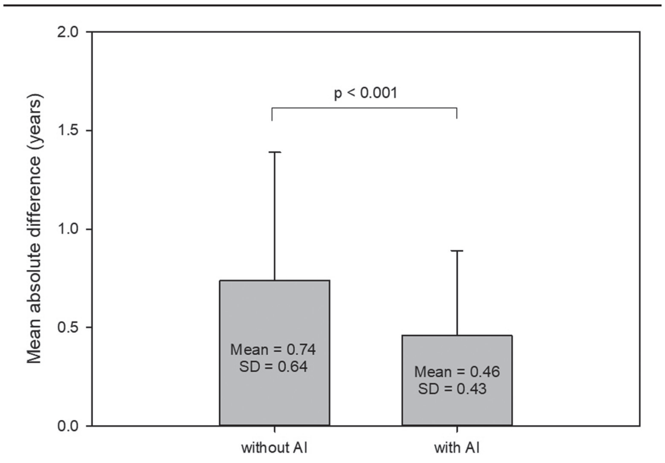


Fig. 5 Overall accuracy comparison between bone age assessments with and without AI assistance. AI = artificial intelligence.

assessments with decreased BA prediction accuracy ($p < 0.001$). Junior radiologists achieved higher improvement number than mid-level and senior radiologists.

3.3. Effects of providing AI interpretations immediately or with a delay on radiologist performance

We further evaluated whether the timing of providing AI assistance (immediately or after a delay) would affect the radiologists' performance. The evaluation results are presented in Table 4. In general, there was no universal influence of timing for providing AI assistance. For S1, M2, J1, and J2 radiologists, both initial providing and delay providing showed statistically significant improvement with AI assistance. S2 radiologist showed a significant difference only in the session when AI aid was provided with a delay. However, M1 radiologist showed significant difference only in the session when AI aid was provided initially.

3.4. Radiologist disagreements with AI predictions

In reviewing the dataset with AI assistance, the radiologists occasionally rejected the AI predictions, noting their disagreement after the assessment. The disagreement rates for these assessments are presented in Table 5. The number of disagreements from senior radiologists (S1 and S2) was higher than that from mid-level radiologists (M1 and M2) and junior radiologists (J1 and J2). J1 radiologist, who achieved the highest proportion of improvement (as shown in Table 3), assigned the least disagreement in Table 5. Most disagreements of radiologists occurred concerning images from children aged ≤8 years.

4. DISCUSSION

Rapid advancements in AI can improve diagnostic accuracy in radiology. In this prospective randomized crossover study, we compared the BA assessment results of six radiologists with or without AI assistance. Our results demonstrate that the BA assessment accuracy was significantly improved with AI assistance. We also evaluate the influence of automation bias in this

Table 1

Performance of individual radiologists with and without AI assistance

	Comparison of MAD between with and without AI information				
	With AI information		Without AI information		
Radiologist	Mean, y	SD	Mean, y	SD	p
S1	0.3539	0.3211	0.7016	0.6690	<0.001
S2	0.5323	0.5180	0.7684	0.6694	<0.001
M1	0.4634	0.4009	0.5541	0.4849	0.017
M2	0.5136	0.4432	0.6921	0.5668	<0.001
J1	0.2992	0.2691	0.7908	0.7454	<0.001
J2	0.5984	0.5020	0.9087	0.6290	<0.001

MAD = mean absolute difference.

Table 2

Proportions at which the absolute difference exceeded 1 y

Radiologist	S1	S2	M1	M2	J1	J2	Total
AI (–)	20%	27.5%	18%	23.5%	27.0%	34%	24%
AI (+)	2%	16.5%	7.5%	10.5%	0.5%	13.5%	8.4%
p	<0.001	0.001	<0.001	<0.001	<0.001	<0.001	<0.001

AI = artificial intelligence; AI (–) = without assistance of artificial intelligence; AI (+) = with assistance of artificial intelligence.

Table 3

Changes in prediction accuracy with AI assistance

Radiologist	S1	S2	M1	M2	J1	J2	Total ($p < 0.001$)
Improved (≥1 y)	39	19	26	20	49	48	201 (16.8%)
Worsened (≥1 y)	4	1	12	2	1	1	28 (2.3%)

AI = artificial intelligence.

Table 4

Effects of providing AI assistance initially or with a delay

	Comparison between MAD with and without AI information									
	AI information first					AI information later				
	AI (+)		AI (–)			AI (+)		AI (–)		
Radiologist	Mean, y	SD	Mean, y	SD	p	Mean, y	SD	Mean, y	SD	p
S1	0.3293	0.3146	0.5386	0.5344	0.001	0.3784	0.3271	0.8663	0.7478	<0.001
S2	0.5897	0.5184	0.6247	0.5107	0.224	0.4750	0.5137	0.9120	0.8250	<0.001
M1	0.4410	0.4105	0.5460	0.5016	0.032	0.4857	0.3919	0.5623	0.4001	0.190
M2	0.5702	0.4240	0.7497	0.5743	0.001	0.4410	0.4105	0.5460	0.5016	0.001
J1	0.2663	0.2584	0.8690	0.8216	<0.001	0.3321	0.2744	0.7125	0.6553	<0.001
J2	0.6397	0.5579	0.9203	0.6717	0.001	0.5570	0.4379	0.8970	0.5865	<0.001

AI = artificial intelligence; AI (–) = without AI information; AI (+) = with AI information; MAD = mean absolute difference.

Table 5

Proportion of disagreements with AI interpretations for various age categories

Age	S1	S2	M1	M2	J1	J2	Total
≤8 y	25	46	17	33	13	13	147 (12.2%)
8-15 y	30	17	5	21	4	13	90 (7.5%)
≥15 y	10	2	1	1	1	1	16 (1.5%)
All age	65	65	23	55	18	27	253 (21.2%)

AI = artificial intelligence.

study. In the past, several studies have highlighted the superior accuracy and efficiency of AI algorithm for BA assessment when comparing with manual approaches.[12,20–25] Only one study has examined the automation bias of providing AI assistance.[12] The present study allow radiologists to disagree with AI estimates and was designed to monitor automation bias. To the best of our knowledge, this is the first prospective randomized crossover study to examine automation bias.

Four requirements must be met for a computer-aided diagnosis (CAD) method to be successful in clinical practice: (1) CAD must improve radiologists' performance; (2) CAD must save time; (3) CAD must be seamlessly integrated into the workflow; and (4) CAD must not impose liability concerns, and any incremental costs must be negligible or reimbursed.[26] For the first requirement, our study revealed that the MAD observed without AI assistance was 0.74 years, which improved to 0.46 years with AI assistance; additionally, the proportion of predictions with age differences of >1 year decreased from 33.2% to 14.3%. These findings are consistent with those of a previous multicenter prospective randomized controlled study,[12] which revealed that the accuracy of BA assessment by pediatric radiologists improved with AI assistance; specifically, the initial MAD observed without AI assistance was 0.5 years, which improved to 0.45 years with AI assistance, and the proportion of assessments with age differences of >1 year decreased from 13% to 9.3%. Therefore, the aforementioned study and the present study provide strong evidence that AI interpretations can improve radiologists' performance.

The integration of AI into automation in medical care offer renewed optimism for the diagnostic aid in radiology. When it performs well, automation can reduce error and improve decision performance. It also, however, has the potential to introduce new type of errors. Automation bias happens when users become over reliant on AI support, which reduces vigilance in information seeking and processing.[27] To monitor the effect of automation bias in our study, we allowed the radiologists to reject the AI interpretation. If the radiologists were influenced by automation bias, they would have persuaded themselves to accept AI predictions without adequate justification. Our study revealed that junior radiologists were more easily influenced by AI interpretations and generally accepted them. Senior radiologists had worse accuracy than middle-level radiologists, but had more disagreements with AI interpretations. The majority of disagreements among radiologists (147/253) fell into the ≤8 years category, and among them, 79/147 were in the range of ≤3 years old. This finding is consistent with past research. According to a survey by the American Society of Pediatric Radiology, radiologists have less confidence in assessing BA in young children, especially those under 3 years old.[28] Additionally, we examined the number of improved and less accurate predictions with AI assistance in our study; the results indicate that although the radiologists experienced a higher proportion of improvement in assessments (16.8%) than getting less accurate (2.3%) with AI assistance. Therefore, although automation bias might have influenced the judgment of radiologists, the total influence was positive toward an accurate direction.

Our AI algorithm takes <1 second to make a single BA prediction. However, when AI is used as an auxiliary tool, a radiologist may spend considerably more time judging the accuracy of the AI-provided interpretation, the length of which depends on the radiologist's experience and trust in the accuracy of AI. Although we did not measure the time saved with AI assistance, we asked radiologists on our follow-up questionnaire whether they believed AI assistance saved time. Four radiologists agreed that AI assistance saved time, and two indicated no difference in time spent with AI assistance. Whether AI algorithm really saves time remains inconclusive. Another concern is that if the potential benefits of reduction in interpretation time of AI might increase the risk of automation bias. When radiologists try to reduce time, they are prone to accept AI predictions without adequate scrutiny or even persuade themselves to accept AI predictions without adequate justification. This dilemma remains poorly understood. We need to test the AI algorithm in real clinical workflow in our future study in order to answer these questions.

Whether the timing of providing AI information will affect the automation bias was evaluated in this study. Radiologists evaluated two datasets in two independent sessions with or without AI assistance; two sessions were spaced 4-week apart. Our results reveal that the effect of timing for providing AI information varied. For example, one of the radiologists (S2) showed no improvement when the AI information was provided in the early session, whereas another radiologist (M1) showed no improvement when the AI information was provided in the delay session. When the results of early session and delay session were pooled together, the effect of timing order should have been neutralized. Indeed, with AI assistance, the combined accuracy of early and delay sessions were improved significantly for both of S2 and M1 radiologists. Although the performance of radiologists of all experience levels improved with AI assistance, the performance of junior radiologists exhibited the highest level of improvement. A possible reason for this finding is that BA assessment is a meticulous task; junior radiologists require several years of experience with the assistance of senior physicians before they can evaluate BA independently, leading them to over reliant on AI assistance.

The task for BA assessment is particularly well suited to be performed by AI because of relatively well-defined nature of the assessment of BA and relatively consistency and simplicity of the digital radiographs of hand. It is also a relatively tedious, repetitive and time-consuming job from a clinical perspective that makes it a good candidate for clinical implementation. AI poses challenges to the future of radiology, raising questions about whether young doctors will be less inclined to train as radiologists and whether AI is dangerous.[29] Consensus exists in the radiology community that AI will not replace radiologists but that radiologists who use AI will replace those who do not.[30] In the case of BA assessment, AI models are unlikely to be used without a radiologist's input; this is because AI models cannot detect radiographs subtle bony abnormalities on radiographs, such as fracture or congenital malformation. In the future clinical applications, an optimal blend of human and AI-based information processing would be a good choice.

This study has several limitations. First, the ground truth in the study was established with inputs from only two experts. Although two experts were noted to have several years of experience in BA assessment, they may not represent the optimal golden standard. Our previous study revealed that the intrarater correlation and interrater correlation of the same experts were 0.993 and 0.992, respectively.[17] Therefore, they are qualified to represent expert manual BA assessment of current clinical practice. Second, the study's single-center design may limit its generalizability. The training and experience of radiologists may be different in other hospitals, and if the study is performed in other population the results may vary. Therefore, it is important for external validation to confirm the finding in different population. In the future, more in-depth studies, such as a multicenter study, should be conducted to address this limitation. Third, the study adopted a prospective laboratory experimental design, which may not reflect the complexity of performance in clinical settings. Future studies in clinical settings are required to validate these results. Fourth, the study did not assess the intra-rater consistency of the radiologists. Further study is needed to evaluate whether this new AI algorithm will also improve the reproducibility in repetitive measurement of radiologists. Fifth, we used a combined databank to construct our AI algorithm. The drawback of combined bank is that the ethnic background may be different from the local users. Therefore, the current AI model is not perfect. The applicability of AI model in different age groups should be further carefully evaluated. We are currently planning to cooperate with other hospitals to increase our local databank and to refine the AI algorithm.

In conclusion, AI assistance increased the accuracy of BA predictions made by radiologists with different levels of experience while allowing radiologists to maintain their own independent judgment. Radiologists reading BAs were prone to automation bias when being supported by an AI algorithm, especially in less experienced radiologists. This effect must be considered to ensure safe application and accurate diagnostic performance when combining human users and AI.

ACKNOWLEDGMENTS

The study was supported by a research grant from the Cheng Hsin General Hospital, Taiwan, ROC [CHGH 112-(N)06].

We acknowledge all the participants who were involved in this study.

REFERENCES

1. Creo AL, Schwenk WF. Bone age: a handy tool for pediatric providers. *Pediatrics* 2017;140:e201714.
2. Greulich WW, Pyle SI, editors. *Radiographic atlas of skeletal development of the hand and wrist*. 2nd ed. Standford, CA: Standford University Press; 1959.
3. Tanner J, Healy M, Goldstein H, Cameron N, editors. *Assessment of skeletal maturity and prediction of adult height (TW3 method)*. London: WB Saunders; 2001.
4. Satoh M. Bone age: assessment methods and clinical applications. *Clin Pediatr Endocrinol* 2015;24:143–52.
5. Rubin DA. Assessing bone age: a paradigm for the next generation of artificial intelligence in radiology. *Radiology* 2021;301:700–1.
6. Rijn RV, Todberg H. Bone age assessment: automated techniques coming of age? *Acta Radiol* 2013;54:1024–9.
7. Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J. Bone age assessment with various machine learning techniques: a systematic literature review and meta-analysis. *PLoS One* 2019;14:e0220242.
8. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, et al. Fully automated deep learning system for bone age assessment. *J Digit Imaging* 2017;30:427–41.
9. Larson DB, Chen MCC, 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.
10. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, et al. The RSNA pediatric bone age machine learning challenge. *Radiology* 2019;290:498–503.
11. Nadeem MW, Nadeem MW, Goh HG, Ali A, Hussain M, Ponnusamy VA. Bone age assessment empowered with deep learning: a survey, open research challenges and future directions. *Diagnostics (Basel)* 2020;10:781.
12. Eng DK, Khandwala NB, Long J, Fefferman NR, Lala SV, Strubel NA, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. *Radiology* 2021;301:692–9.
13. Tajmir SH, Lee H, Shailam R, Gale HI, Nguyen JC, Westra SJ, et al. Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. *Skeletal Radiol* 2019;48:275–83.
14. Martin DD, Calder AD, Ranke MB, Binder G, Thodberg HH. Accuracy and self-validation of automated bone age determination. *Sci Rep* 2022;12:1–12.
15. Peng CT, Chan YK, Yuh YS, Yu SS. Applying convolutional neural network in automatic assessment of bone age using multi-stage and cross-category strategy. *Appl Sci* 2022;12:12798.
16. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology* 2023;307:e222176.

17. Yuh YS, Chou TY, Chow JC. Applicability of the Greulich and Pyle bone age standards to Taiwanese children: a Taipei experience. *J Chin Med Assoc* 2022;85:767–73.
18. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45:255–68.
19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;327:307–10.
20. Cheng CF, Huang ETC, Kuo JT, Liao KYK, Tsai FJ. Report of clinical bone age assessment using deep learning for an Asian population in Taiwan. *Biomedicine (Taipei)* 2021;11:50–8.
21. Bowden JJ, Bowden SA, Ruess L, Adler BH, Hu H, Krishnamurthy R, et al. Validation of automated bone age analysis from hand radiographs in a North American pediatric population. *Pediatr Radiol* 2022;52:1347–55.
22. Lea WWI, Hong SJ, Nam HK, Kang WY, Yang ZP, Noh EJ. External validation of deep learning-based bone-age software: a preliminary study with real world data. *Sci Rep* 2022;12:1232.
23. Kim PH, Yoon HM, Kim JR, Hwang JY, Choi JH, Hwang J, et al. Bone age assessment using artificial intelligence in Korean pediatric population: a comparison of deep-learning models trained with healthy chronological and Greulich-Pyle ages as labels. *Korean J Radiol* 2023;24:1151–63.
24. Wang X, Zhou B, Gong P, Zhang T, Mo Y, Tang J, et al. Artificial intelligence–assisted bone age assessment to improve the accuracy and consistency of physicians with different levels of experience. *Front Pediatr* 2022;10:818061.
25. Liu Y, Ouyang L, Wu W, Zhou X, Huang K, Wang Z, et al. Validation of an established TW3 artificial intelligence bone age assessment system: a prospective, multicenter, confirmatory study. *Quant Imag Med Surg* 2024;14:144.
26. Van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology* 2011;261:719–32.
27. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017;24:423–31.
28. Breen MA, Tsai A, Stamm A, Kleinman PK. Bone age assessment practices in infants and older children among Society for Pediatric Radiology members. *Pediatr Radiol* 2016;46:1269–74.
29. Gallix B, Chong J. Artificial intelligence in radiology: who's afraid of the big bad wolf? *Eur Radiol* 2019;29:1637–9.
30. Langlotz CP. Will artificial intelligence replace radiologists? *Radiol Artif Intell* 2019;1:e190058.