Skeletal Radiology (2019) 48:275–283
https://doi.org/10.1007/s00256-018-3033-2

SCIENTIFIC ARTICLE

Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability

Shahein H. Tajmir[1,2] · Hyunkwang Lee[1,3] · Randheer Shailam[1,2] · Heather I. Gale[4] · Jie C. Nguyen[5] · Sjirk J. Westra[1,2] · Ruth Lim[1,2] · Sehyo Yune[1,2] · Michael S. Gee[1,2] · Synho Do[1,2]

Received: 12 May 2018 / Revised: 10 July 2018 / Accepted: 20 July 2018 / Published online: 1 August 2018

Abstract

Objective Radiographic bone age assessment (BAA) is used in the evaluation of pediatric endocrine and metabolic disorders. We previously developed an automated artificial intelligence (AI) deep learning algorithm to perform BAA using convolutional neural networks. We compared the BAA performance of a cohort of pediatric radiologists with and without AI assistance.

Materials and methods Six board-certified, subspecialty trained pediatric radiologists interpreted 280 age- and gender-matched bone age radiographs ranging from 5 to 18 years. Three of those radiologists then performed BAA with AI assistance. Bone age accuracy and root mean squared error (RMSE) were used as measures of accuracy. Intraclass correlation coefficient evaluated inter-rater variation.

Results AI BAA accuracy was 68.2% overall and 98.6% within 1 year, and the mean six-reader cohort accuracy was 63.6 and 97.4% within 1 year. AI RMSE was 0.601 years, while mean single-reader RMSE was 0.661 years. Pooled RMSE decreased from 0.661 to 0.508 years, all individually decreasing with AI assistance. ICC without AI was 0.9914 and with AI was 0.9951.

Conclusions AI improves radiologist's bone age assessment by increasing accuracy and decreasing variability and RMSE. The utilization of AI by radiologists improves performance compared to AI alone, a radiologist alone, or a pooled cohort of experts. This suggests that AI may optimally be utilized as an adjunct to radiologist interpretation of imaging studies to improve performance.

Keywords Machine learning · Bone age · Augmented intelligence · Radiographs · Pediatric

Introduction

Machine learning has emerged as a powerful technique in computer science to teach computers to autonomously find patterns in data and now underlies many large-scale software products including Google Translate [1], Alexa speech recognition [2], and mastering the game of Go [3]. Intense research has focused on applying these techniques to medical applications with recent successes in detecting diabetic retinopathy [4] and detecting malignant melanomas with an accuracy rivaling that of board-certified dermatologists [5]. While there has been much discussion in the lay press about the role of machine learning in radiology [6–8], no direct assessment of the impact of a machine-learning algorithm on the performance of a cohort of radiologists has been performed.

Radiographic bone age assessment (BAA) is a central part of the clinical workup of pediatric endocrine and metabolic disorders, in which the patient's chronologic age is compared with their level of skeletal maturity based on a standardized reference. BAA in clinical practice is typically performed using either the Greulich and Pyle [9] or Tanner–Whitehouse [10] (TW2) methods by comparing a radiograph of the hand and wrist to an age-based atlas or determining age based on scoring specific radiographic features. In both cases, BAA is time-consuming and contains significant interrater variability among radiologists. BAA is an ideal application for automated image evaluation, as there is a single image—the left hand and wrist—and relatively standardized findings.

This work has been accepted for presentation at RSNA 2017 and awarded an RSNA Trainee Research Prize.

✉ Shahein H. Tajmir
shahein@stajmir.com

[1] Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

[2] Harvard Medical School, Boston, MA, USA

[3] Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA

[4] The Billings Clinic, Billings, MT, USA

[5] Children's Hospital of Philadelphia, Philadelphia, PA, USA

We have previously developed a fully automated, deep learning algorithm to perform bone age assessment (BAA) using convolutional neural networks (CNN) that achieved mean accuracies of 92.3% within 1 year for the female and male cohorts compared with radiology report reference [11]. As AI based interpretation tools enter radiology clinical practice, important unanswered questions include how these tools compare with radiologist performance and how they are best integrated into radiology practice.

The purpose of this study is to compare the performance of a deep learning-based BAA algorithm to a cohort of pediatric radiologists and evaluate the impact of the implementation of a deep learning-based BAA algorithm on radiologist accuracy and variability when performing BAA on a set of standardized cases with and without access to AI interpretation (Fig. 1).

Methods

Patients

IRB approval was obtained for this retrospective, HIPAA-compliant study. We constructed a balanced cohort with ten representative cases for each class and gender, representing 280 cases ranging from 5 to 18 years from a cohort of 8325 radiographs previously used to train a deep learning CNN. The CNN was trained and validated (85:15) again without these 280 cases. These radiographs were then interpreted by the CNN, creating a predicted bone age and attention map for each.

Patient characteristics and indications

Self-reported demographic data for the 280 test cases are presented in Table 1. The distribution of chronologic ages roughly matches that of the bone age- and gender-matched test bone age cohort—ten patients per class and gender (Appendix Table 5). The predominant indications for BAA were evaluation of short stature (92/280 = 33%), monitoring of growth hormone therapy (52/280 = 19%), precocious puberty (57/280 = 20%), and research (38/280 = 19%). Please see appendix Fig. 6 for a detailed list of indications and appendix Table 5 for a graph of chronological age distribution.

Table 1 Self-reported patient ethnicities for patients in the evaluation cohort

Race	*N* (%)
White	188 (67%)
Hispanic	42 (15%)
Black	23 (8%)
Asian	12 (4%)
Middle East	7 (3%)
Southeast Asian	5 (2%)
Declined to respond	3 (1%)
Total	280 (100%)

Image processing and training

Our architecture first normalizes input images to have black backgrounds and a uniform size (512 × 512 pixels), then uses a preliminary segmentation CNN based on LeNet-5 with a 32 × 32 imaging patch size and stride of 4 to automatically segment the hand and remove extraneous data such as background artifacts, collimation, and annotation markers. The segmented and normalized image then enters the vision pipeline and has contrast-limited adaptive histogram equalization (CLAHE), denoising, and sharpening applied to enhance bony details, and finally is passed to the classification CNN for skeletal age classification. The classification CNN is based on an ImageNet pre-trained GoogLeNet fine-tuned on our train dataset by applying data augmentation with geometric (rotation, resizing, and shearing) and photometric (contrast and brightness) transformations to avoid overfitting. After holding out 280 images for testing, 15% of images were randomly selected for validation, and the remaining 6838 were utilized to train the CNN. All training was performed with a mini-batch stochastic gradient descent with a mini-batch size of 96 using

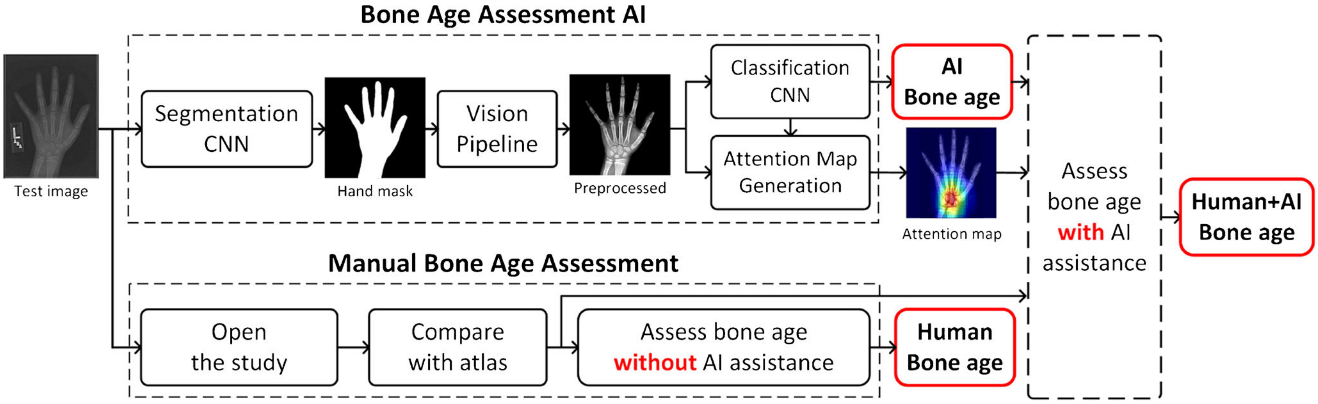


Fig. 1 Process flow map comparing CNN-based BAA and manual BAA

0.001 base learning rate, gamma of 0.1, and momentum term of 0.9, and weight decay of 0.005. The best CNN models were selected based on the validation loss values.

For visualization of the network as well as to provide AI interpretive information for the radiologists, attention maps were generated using the occlusion method [12]. This method iteratively slides a small patch across an image, presents the occluded images to the network, and creates two-dimensional attention maps based on the change in classification probability.

Image interpretation

Six board-certified, subspecialty trained pediatric radiologists from three academic medical centers with a mean of 13.8 years clinical experience (S.J.W. 21 years, R.L. 18 years, R.S. 16 years, M.S.G. 14 years, J. N. 9 years, and H.I.G. 5 years) interpreted the 280 test radiographs first using the GP atlas method. Three radiologists were randomly chosen to be presented the automated BAA results (including attention maps) and asked to report the BAA using the GP atlas and the additional AI information to test the effect of AI on BAA.

Reference standard

Bone age is ultimately a consensus evaluation with reference to a representative atlas, making it difficult to define a true gold standard reference. As a result, reference bone ages were determined using two different methods: (1) an independent reviewer and (2) a normalized mean. The independent reviewer was a radiologist (initials [removed for peer review]) who was not part of the six-member cohort. This reviewer had access to AI attention maps, AI bone age, individual rater cohort scores, and the clinical reports. The reviewer also compared all of these results to Gilsanz and Ratib's Digital Atlas of Skeletal Maturity [13], and selected the GP atlas timepoint closest to the GR BAA. The second method used the normalized mean by taking the mean of the six raters and selecting the closest GP time point.

Statistical analysis

Quantitative variables are presented as means with ranges. Accuracies were reported as the exact same result or accuracy within 1 year. Bone age accuracy and root mean squared error (RMSE) were used as measures of accuracy. ^{2} was used to test for exact accuracy statistical significance, and two-tailed *t* test was used for RMSE statistical significance. Intraclass correlation coefficient (ICC) based on two-way random average measures was chosen to evaluate inter-rater variation amongst the radiologists without and with AI as a measure of variability. Statistical differences were considered significant at $p < 0.05$.

Experimental environment

All experiments were run on a Devbox (NVIDIA Corp, Santa Clara, CA, USA) containing four TITAN X GPUs with 12GB of memory per GPU [22], and on Nvidia deep learning frameworks, including Nvidia-Caffe (0.16.1) and Nvidia DIGITS (5.1). Excel 360 and MedCalc version 17.9 were used for statistical analysis.

Results

AI and cohort accuracies when compared to the independent reviewer reference

AI RMSE was 0.548 years and mean single-reader RMSE was 0.704 years, ranging from 0.544–0.902. AI exact accuracy was 73.2 and 98.9% within 1 year. Mean single-reader accuracy was 62.6%, ranging from 50.7–74.6% (Table 2). For context, the original clinical reports had an exact accuracy of 68.6% and an RMSE of 0.633 years when compared to the independent reviewer reference.

Table 2 Accuracies and root mean square error (RMSE) for the machine learning algorithm and individual raters when compared to both the independent reviewer and normalized mean references

Rater	Independent reviewer reference			Normalized means		
	Exact accuracy *N* (%)	Accuracy within 1 year *N* (%)	RMSE (years)	Exact accuracy *N* (%)	Accuracy within 1 year *N* (%)	RMSE (years)
AI	205 (73.2%)	277 (98.9%)	0.548	191 (68.2%)	276 (98.6%)	0.601
Rater 1	203 (72.5%)	278 (99.3%)	0.544	204 (72.9%)	278 (99.3%)	0.541
Rater 2	159 (56.8%)	269 (96.1%)	0.765	173 (61.8%)	273 (97.5%)	0.689
Rater 3	142 (50.7%)	260 (92.9%)	0.862	156 (55.7%)	270 (96.4%)	0.754
Rater 4	192 (68.6%)	275 (98.2%)	0.607	193 (68.9%)	279 (99.6%)	0.567
Rater 5	147 (52.5%)	255 (91.1%)	0.902	149 (53.2%)	259 (92.5%)	0.843
Rater 6	209 (74.6%)	276 (98.9%)	0.544	194 (69.3%)	278 (99.3%)	0.573

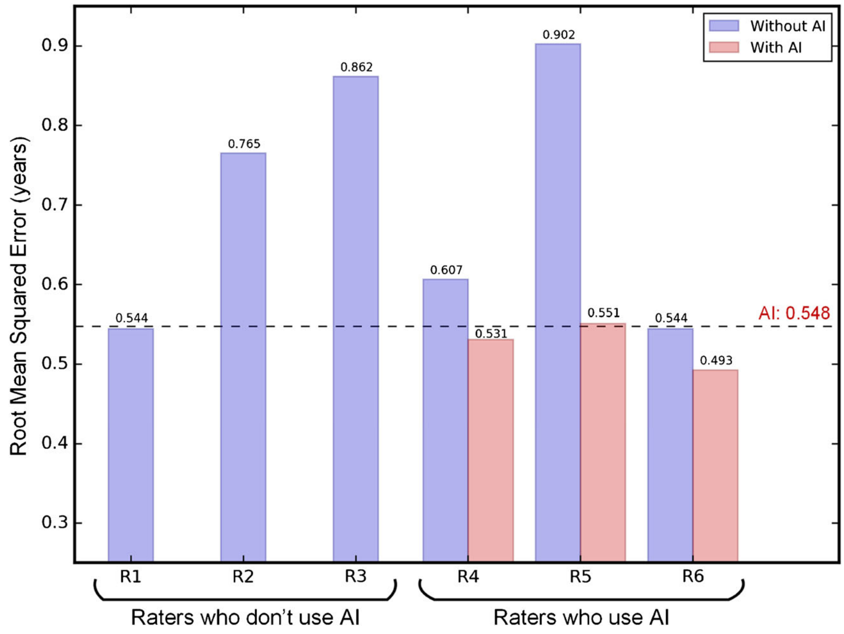


Fig. 2 Individual reader root mean square error in years without AI assistance. AI RMSE was 0.548 years

Impact of pairing AI with radiologists

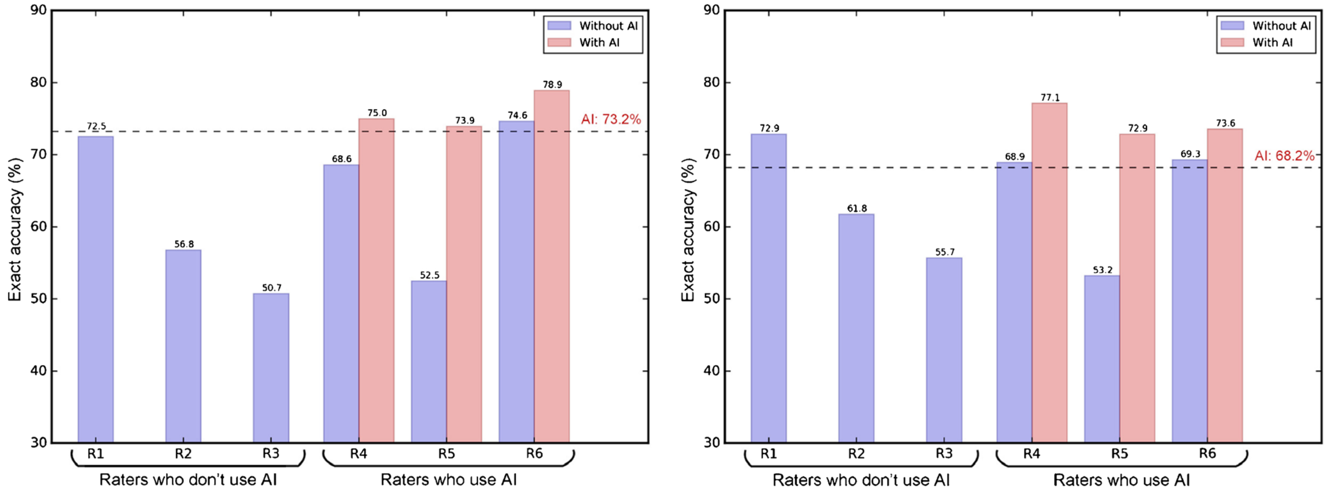
Radiologists who utilized AI had pooled RMSE decrease from 0.684 to 0.525 years, all individually decreasing—0.607 to 0.531 for rater 4, 0.902 to 0.551 for rater 5, 0.544 to 0.493 for rater 6 (Fig. 2). Combined AI and radiologist interpretation resulted in higher accuracy than AI alone or the six-reader cohort mean.

Effect persistence with an alternative measure of ground truth

Similar improvements in accuracy and RMSE persisted when normalized cohort mean rating was used as the reference (Fig. 3). Radiologists paired with AI had increases in accuracy and RMSE (Table 3).

Interrater variation

Intraclass coefficients ICC(2,k) were calculated amongst the three radiologists exposed to AI. ICC without AI was 0.9914 (95% CI 0.9894 to 0.9930) and with AI was 0.9951 (95% CI 0.9940 to 0.9960). For comparison, ICC among the three radiologists who were not exposed to AI was 0.9908 (95% CI 0.9888 to 0.9925), similar to the other three radiologists without AI, but worse than the AI-assisted radiologists. Bland–Altman plots were generated and revealed decreased spread of ratings and decreased limits of agreement when paired with AI (Fig. 4).



(a) Referee standard as ground truth

(b) Normalized means as ground truth

Fig. 3 Individual reader exact and within 1 year accuracy in years without AI assistance. **a** Accuracy when compared to the independent reviewer reference. **b** Accuracy when compared to normalized mean reference

Table 3 Effect of AI on reader performance

	Independent reviewer reference						Normalized means					
	Rater 4		Rater 5		Rater 6		Rater 4		Rater 5		Rater 6	
	(−) AI	(+) AI	(−) AI	(+) AI	(−) AI	(+) AI	(−) AI	(+) AI	(−) AI	(+) AI	(−) AI	(+) AI
Exact accuracy	68.6%	75.0%	52.5%	73.9%	74.6%	78.9%	68.9%	77.1%	53.2%	72.9%	69.3%	73.6%
X^2	$p = 0.091$		$\boldsymbol{p = 0.0001}$		$p = 0.230$		$\boldsymbol{p = 0.029}$		**p = 0.0001**		$p = 0.2617$	
RMSE (years)	0.607	0.531	0.902	0.551	0.544	0.493	0.567	0.478	0.843	0.521	0.573	0.524
t test p value	$\boldsymbol{p = 0.0005}$		$\boldsymbol{p < 10^{-11}}$		$\boldsymbol{p = 0.0051}$		$\boldsymbol{p = 0.00003}$		$\boldsymbol{p = 10^{-10}}$		**p = 0.005**	
Limit of agreement (95%)	±1.17	±1.01	±1.73	±1.08	±1.03	±0.93	±1.12	±0.95	±1.59	±1.02	±1.12	±1.03

Exact accuracy, root mean square error in years, and 95% limit of agreement with and without AI assistance. Bolded values are statistically significant

AI performance variation based on patient ethnicity

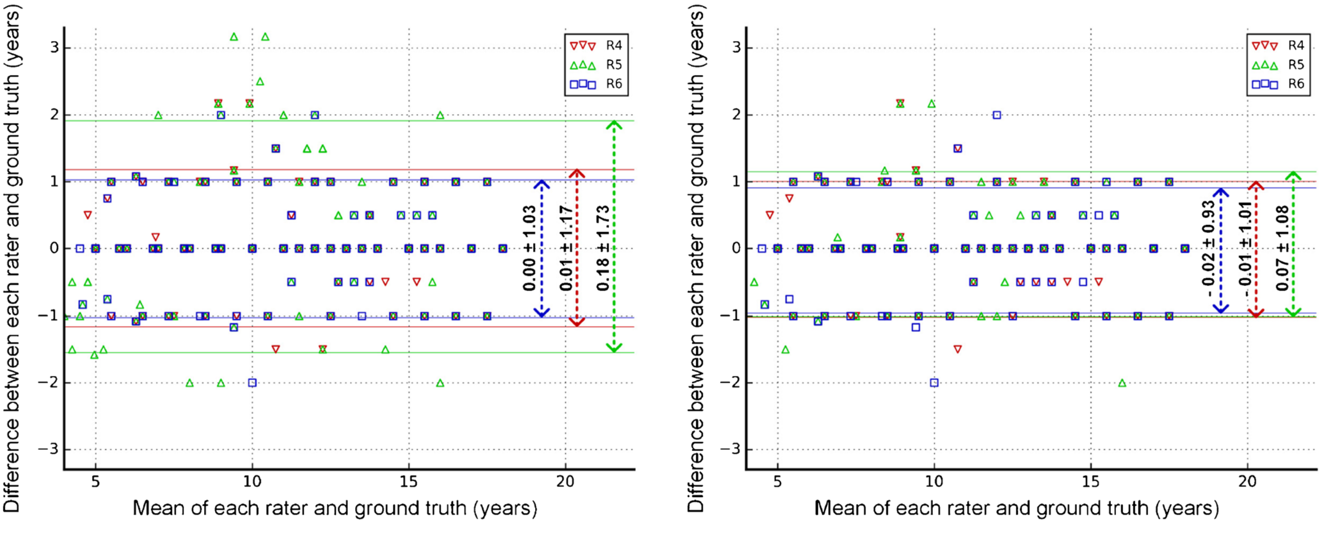
Performance of the AI algorithm was evaluated based on the self-reported ethnicity/race of the patients. AI RMSE of the combined cohort was 0.548 years, with 0.551 years in Caucasian children and 0.542 years in non-Caucasian children (Table 4; $p = 0.891$).

Discussion

Machine learning-derived approaches have great potential for application in medicine, allowing rapid and scalable systems to perform complex analysis of medical data [14]. While most work has focused on applications of computer vision to natural images, these techniques can also be applied to medical images such as detection of malignant skin lesions [5] or diabetic retinopathy screening [15]. Recent work has demonstrated systems to detect tuberculosis on chest radiographs [16], stage and predict prognosis of COPD [17], and identify anatomic structures on abdominal CT [18]. These techniques have also been applied outside imaging by using a machine learning model to perform automated stratification of indeterminate breast lesions into surgical and observation groups, avoiding surgery in 30% of cases [19].

Fully automated BAA for use in the clinical setting has been a goal in computer vision and radiology research dating back to at least 1989 [20]. While most prior approaches have utilized hand-crafted features extracted from regions of interest [21], our approach utilizes transfer learning with a pre-trained CNN to automatically extract key features from all bones present in the hand and wrist, without the limitations imposed by hand-crafted features.

One of the challenges in BAA study design is the inherent variability in radiologist clinical interpretation of bone age radiographs, which makes selection of an appropriate reference standard difficult. For our study, we chose two different reference standards: (1) an independent radiologist reviewer and (2)



(a) Raters without AI

(b) Raters with AI

Fig. 4 Bland–Altman plots for the three raters who utilized AI. Axes are reported in years. Use of AI was associated with increased ICC and decreased variability in BAA assessment

Table 4 Accuracies and RMSE for the machine learning algorithm when comparing Caucasian versus non-Caucasian children

	Exact accuracy *N* (%)	Accuracy within 1 year *N* (%)	RMSE (years)
All	205 (73.2%)	277 (98.9%)	0.548
Caucasian	140 (74.5%)	185 (100.0%)	0.551[a]
Non-Caucasian	65 (70.7%)	92 (100.0%)	0.542[a]

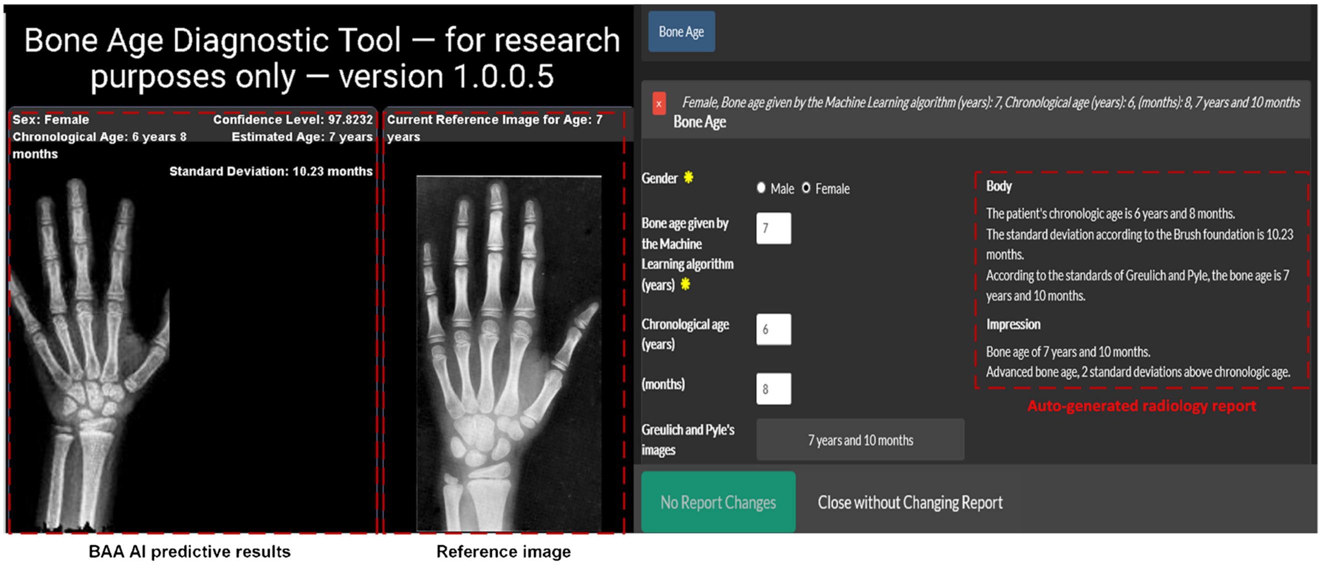
There were 188 Caucasian and 92 non-Caucasian children in the test dataset

[a] Two-tailed, two-sample equal variance *t* test, $p = 0.891$

a normalized mean cohort value from the six pediatric radiologists. We believe that a cohort-based reference standard is the most valid reference that best reflects the range of BAA in clinical practice. The six pediatric radiologists spanned three large academic medical centers and enabled a robust assessment of BAA intrinsic variation (measured as RMSE). Our six-radiologist mean cohort RMSE for BAA without AI was 0.661 years, comparable to previously published RMSE values—ranging from a mean RMSE of 0.96 years in a British cohort [22] and 0.59 years in the ATLAS dataset [23] to 0.51 ± 0.44 years [24] in a recent analysis in a Korean cohort. Thus we believe that our baseline radiologist BAA performance can be considered consistent with standard clinical radiologist performance.

An important result of our work is that AI BAA performance is at a level comparable to pediatric radiologists, similar to recently reported work by Larson et al. [8]. AI achieved an RMSE of 0.601 years, which was not significantly different from the cohort mean RMSE and is comparable with the previously reported values of RMSE intrinsic to BAA. In addition, AI had comparable BAA accuracy compared to the pediatric radiologist reader cohort, with no significant difference in exact (68.2 and 63.6%) or within 1 year (98.6 and 97.4%) accuracies, respectively. The slight but not significantly increased accuracy achieved by AI compared to the reader cohort could reflect a small degree of overfitting given that four of the six raters also provided interpretations for the initial training dataset.

Another goal of our study was to assess the impact of AI on pediatric radiologist BAA performance. To do this, we asked pediatric radiologists to interpret bone age radiographs before and after access to AI input. Our results also show that access to AI improves the accuracy and decreases the variability of subspecialty-trained pediatric radiologists BAA. Among the three radiologists who were paired with AI, the mean RMSE decreased from 0.661 to 0.508 years. Mean accuracy increased from 63.8 to 74.5% when compared to AI accuracy of 68.2%. All individual radiologist+AI RMSEs statistically decreased below that of AI or the radiologists alone, while accuracy statistically increased for two out of the three (Table 3). Importantly, the improvement that AI provides for pediatric radiologist BAA accuracy and variability is observed when compared with two different reference standards (both an independent reviewer and normalized cohort mean). Our study design (six independent evaluations of 280 standardized cases) accounts for the fact that a true reference BAA standard in clinical practice should incorporate both accuracy as well as intrinsic variation among different radiologists. These results build on recent data by Kim et al. [25] demonstrating that AI can help trainees improve their


Bone Age Diagnostic Tool — for research purposes only — version 1.0.0.5

Sex: Female
Chronological Age: 6 years 8 months
Confidence Level: 97.8232
Estimated Age: 7 years
Standard Deviation: 10.23 months

Current Reference Image for Age: 7 years

Female, Bone age given by the Machine Learning algorithm (years): 7, Chronological age (years): 6, (months): 8, 7 years and 10 months
Bone Age

Gender ✱ ● Male ○ Female
Bone age given by the Machine Learning algorithm (years) ✱ 7
Chronological age (years) 6
(months) 8
Greulich and Pyle's images 7 years and 10 months

Body
The patient's chronologic age is 6 years and 8 months.
The standard deviation according to the Brush foundation is 10.23 months.
According to the standards of Greulich and Pyle, the bone age is 7 years and 10 months.
Impression
Bone age of 7 years and 10 months.
Advanced bone age, 2 standard deviations above chronologic age.

Bone Age
x
Auto-generated radiology report
No Report Changes
Close without Changing Report
BAA AI predictive results
Reference image

Fig. 5 Screenshot of the bone age assessment tool with its outputs visually emphasized. This tool is directly embedded into PACS for use at point of care

accuracy and interpretation speed when paired with a neural network-based bone age classifier.

Much attention has been focused on the potential of AI to replace humans in performing complex visual tasks, including radiologic interpretation [6, 7]. Our results indicate that performance is optimized when AI is deployed in conjunction with radiologist interpretation. Further studies are needed to elucidate the ways in which AI and radiologist image interpretation synergizes, but it is likely that AI is more helpful in cases that are not easily mapped to a specific timepoint. In this way, perhaps AI can be used in other areas of radiology as a time-saving tool to allow radiologists to spend more time on challenging cases.

In addition, our dataset included an ethnically diverse patient population, allowing us to compare AI performance across different ethnic groups. Our results show that AI demonstrates similar BAA performance across different ethnicities, providing good evidence of its generalizability.

Our system was directly embedded into both PACS and our computerized dictation system to aid interpretation and reduce burdens to use (Fig. 5). The system consists of a webapp that allows the radiologist to view the AI BAA prediction, easily scroll through the reference Greulich and Pyle images, and make the final determination while generating a structured report with Brush foundation standard deviations. The system saves time by avoiding table lookups and transcription errors while also keeping the radiologist focused on the images rather than distracting their attention to the atlas or the reporting system.

Strengths of our system include a diverse population and multiple experienced readers to provide a robust ground truth. Limitations of our system include a single site as the source of the training dataset and the intrinsic use of BAA in patients with suspected disease. As our experimental design specifically tried to determine the impact of AI's interpretation on radiologist accuracy and agreement, our study design required immediate interpretation with and without AI, precluding time measurements to compare interpretation acceleration. Additionally, our retrospective design precludes evaluating the impact of higher accuracy on subsequent patient care. Further investigations should utilize multi-site training data and normal healthy patients, while preserving the ability to measure time-savings and the impact of improved BAA on subsequent clinical care as well as assessing whether improvement over time is consistent.

Conclusions

AI performs similarly to practicing pediatric radiologists for BAA. The utilization of AI by radiologists improves performance compared to AI alone, a radiologist alone, or a pooled cohort of experts. This suggests that AI may optimally be utilized as an adjunct to radiologist interpretation of imaging studies, suggesting a model for how AI may best be utilized in radiology.

Compliance with ethical standards

Conflict of interest None

Appendices

Appendix 1

Table 5 Detailed indications for each of the 280 patients in the test cohort

Indication	*N*
Chronic steroids	1
Chronic steroids/cystic fibrosis	1
Chronic steroids/Duchenne muscular dystrophy	1
Congenital adrenal hyperplasia	2
Congenital anorchia	1
Congenital hypomagnesemia, hypercalcemia syndrome	1
Hypertension	1
Kallmann syndrome	2
Klinefelter syndrome	1
Leg length discrepancy	4
Malignancy/astrocytoma	1
Malignancy/brain tumor	1
Malignancy/craniopharyngioma	1
Malignancy/medulloblastoma	2
McCune Albright	1
Noonan syndrome on GH	1
Prader–Willi syndrome	1
Precocious puberty	56
Precocious puberty/NF1	1
Research	15
Research (anorexia)	1
Research (autism bone mass)	1
Research (depression)	1
Research (peak bone mass study)	20
Scoliosis	14
Secondary amenorrhea	1
Short stature	79
Short stature (GH treatment)	52
Short stature (mitochondrial disorder)	1
Short stature (Turners and GH)	1
Short stature (Turners)	1
Short stature Prader–Willi syndrome	2
Short stature, congenital adrenal hyperplasia	2
Short stature, delayed puberty	2
Short stature/IBD	1
Short stature/NF1	1
Short stature/Prader–Willi syndrome/GH treatment	2
Turner syndrome	3
Total	280

Appendix 2

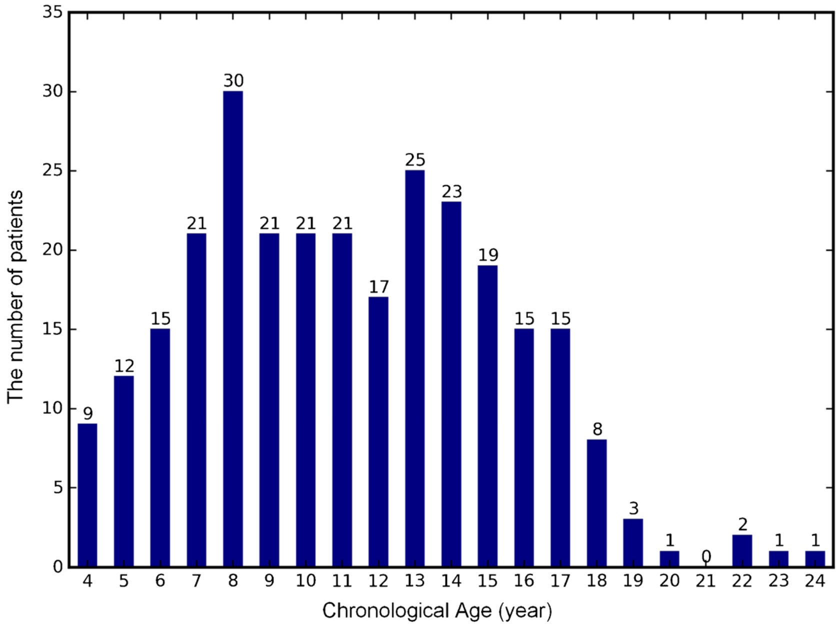


Fig. 6 Chronological age distribution of 280 patients in the test cohort

References

1. Johnson M, Schuster M, Le QV, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv [csCL]*. 2016. http://arxiv.org/abs/1611.04558.
2. Maas R, Rastrow A, Goehner K, Tiwari G, Joseph S, Hoffmeister B. Domain-specific utterance end-point detection for speech recognition. In: Interspeech 2017. 2017. https://doi.org/10.21437/interspeech.2017-1673.
3. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. Nature. 2017;550(7676):354–9. https://doi.org/10.1038/nature24270.
4. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–10. https://doi.org/10.1001/jama.2016.17216.
5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8. https://doi.org/10.1038/nature21056.
6. Lewis-Kraus G. The Great A.I. Awakening. The New York Times. https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html. Published December 14, 2016. Accessed 23 Oct 2017.
7. Mukherjee S. A.I. Versus M.D. The New Yorker. https://www.newyorker.com/magazine/2017/04/03/ai-versus-md. Published March 27, 2017. Accessed 23 Oct 2017.
8. 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. 2017;170236. https://doi.org/10.1148/radiol.2017170236.
9. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. Am J Med Sci. 1959;238(3):393. https://doi.org/10.1097/00000441-195909000-00030.
10. Ehrenberg ASC. J R Stat Soc Ser C Appl Stat. 1977;26(1):80. https://doi.org/10.2307/2346874.
11. Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging. 2017. https://doi.org/10.1007/s10278-017-9955-8.
12. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer vision – ECCV 2014. Lecture Notes in Computer Science. Springer International Publishing; 2014. p. 818-833. https://doi.org/10.1007/978-3-319-10590-1_53.
13. Gilsanz V, Ratib O. Hand bone age: a digital atlas of skeletal maturity. Berlin Heidelberg: Springer; 2011. https://doi.org/10.1007/978-3-642-23762-1.
14. Abuzaghleh O, Barkana BD, Faezipour M. Noninvasive real-time automated skin lesion analysis system for melanoma early detection and prevention. IEEE J Transl Eng Health Med. 2015;3:2900310. https://doi.org/10.1109/JTEHM.2015.2419612.
15. van Grinsven MJJP, van Ginneken B, Hoyng CB, Theelen T, Sanchez CI. Fast convolutional neural network training using selective data sampling: application to hemorrhage detection in color fundus images. IEEE Trans Med Imaging. 2016;35(5):1273–84. https://doi.org/10.1109/TMI.2016.2526689.
16. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017:162326. https://doi.org/10.1148/radiol.2017162326.
17. González G, Ash SY, Vegas Sanchez-Ferrero G, et al. Disease staging and prognosis in smokers using deep learning in chest computed tomography. Am J Respir Crit Care Med. 2017. https://doi.org/10.1164/rccm.201705-0860OC.
18. Lee H, Troschel FM, Tajmir S, et al. Pixel-level deep segmentation: artificial intelligence quantifies muscle on computed tomography for body morphometric analysis. J Digit Imaging. 2017. https://doi.org/10.1007/s10278-017-9988-z.

19. Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. Radiology. 2017:170549. https://doi.org/10.1148/radiol.2017170549.
20. Michael DJ, Nelson AC. HANDX: a model-based system for automatic segmentation of bones from digital hand radiographs. IEEE Trans Med Imaging. 1989;8(1):64–9. https://doi.org/10.1109/42.20363.
21. Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging. 2009;28(1):52–66. https://doi.org/10.1109/TMI.2008.926067.
22. King DG, Steventon DM, O'Sullivan MP, et al. Reproducibility of bone ages when performed by radiology registrars: an audit of Tanner and Whitehouse II versus Greulich and Pyle methods. Br J Radiol. 1994;67(801):848–51. https://doi.org/10.1259/0007-1285-67-801-848.
23. Cao F, Huang HK, Pietka E, Gilsanz V. Digital hand atlas and web-based bone age assessment: system design and implementation. Comput Med Imaging Graph. 2000;24(5):297–307. http://www.ncbi.nlm.nih.gov/pubmed/10940607
24. Kim SY, Oh YJ, Shin JY, Rhie YJ, Lee KH. Comparison of the Greulich-Pyle and Tanner Whitehouse (TW3) methods in bone age assessment. J Korean Soc Pediatr Endocrinol. 2008;13(1):50–5. https://www.koreamed.org/SearchBasic.php?RID=0113JKSPE/2008.13.1.50&DT=1
25. Kim JR, Shim WH, Yoon HM, et al. Computerized bone age estimation using deep learning-based program: evaluation of the accuracy and efficiency. AJR Am J Roentgenol. 2017:1-7. https://doi.org/10.2214/AJR.17.18224.