DIR

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.242999

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Automatic bone age assessment: a Turkish population study

Samet Öztürk[1]
Murat Yüce[2]
Gül Gizem Pamuk[3]
Candan Varlık[3]
Ahmet Tan Cimilli[3]
Musa Atay[3]

[1]Esenler Obstetrics & Gynecology and Pediatrics Hospital, Clinic of Radiology, İstanbul, Türkiye

[2]Icahn School of Medicine at Mount Sinai Biomedical Engineering and Imaging Institute, New York, USA

[3]University of Health Sciences Türkiye, Bağcılar Training and Research Hospital, Clinic of Radiology, İstanbul, Türkiye

PURPOSE

Established methods for bone age assessment (BAA), such as the Greulich and Pyle atlas, suffer from variability due to population differences and observer discrepancies. Although automated BAA offers speed and consistency, limited research exists on its performance across different populations using deep learning. This study examines deep learning algorithms on the Turkish population to enhance bone age models by understanding demographic influences.

METHODS

We analyzed reports from Bağcılar Hospital's Health Information Management System between April 2012 and September 2023 using "bone age" as a keyword. Patient images were re-evaluated by an experienced radiologist and anonymized. A total of 2,730 hand radiographs from Bağcılar Hospital (Turkish population), 12,572 from the Radiological Society of North America (RSNA), and 6,185 from the Radiological Hand Pose Estimation (RHPE) public datasets were collected, along with corresponding bone ages and gender information. A random set of 546 radiographs (273 from Bağcılar, 273 from public datasets) was initially randomly split for an internal test set with bone age stratification; the remaining data were used for training and validation. BAAs were generated using a modified InceptionV3 model on 500 × 500-pixel images, selecting the model with the lowest mean absolute error (MAE) on the validation set.

RESULTS

Three models were trained and tested based on dataset origin: Bağcılar (Turkish), public (RSNA–RHPE), and a Combined model. Internal test set predictions of the Combined model estimated bone age within less than 6, 12, 18, and 24 months at rates of 44%, 73%, 87%, and 94%, respectively. The MAE was 9.2 months in the overall internal test set, 7 months on the public test set, and 11.5 months on the Bağcılar internal test data. The Bağcılar-only model had an MAE of 12.7 months on the Bağcılar internal test data. Despite less training data, there was no significant difference between the combined and Bağcılar models on the Bağcılar dataset ($P > 0.05$). The public model showed an MAE of 16.5 months on the Bağcılar dataset, significantly worse than the other models ($P < 0.05$).

CONCLUSION

We developed an automatic BAA model including the Turkish population, one of the few such studies using deep learning. Despite challenges from population differences and data heterogeneity, these models can be effectively used in various clinical settings. Model accuracy can improve over time with cumulative data, and publicly available datasets may further refine them. Our approach enables more accurate and efficient BAAs, supporting healthcare professionals where traditional methods are time-consuming and variable.

CLINICAL SIGNIFICANCE

The developed automated BAA model for the Turkish population offers a reliable and efficient alternative to traditional methods. By utilizing deep learning with diverse datasets from Bağcılar Hospital and publicly available sources, the model minimizes assessment time and reduces variability. This advancement enhances clinical decision-making, supports standardized BAA practices, and improves patient care in various healthcare settings.

KEYWORDS

Bone age assessment, deep learning, artificial intelligence, convolutional neural network, InceptionV3

Corresponding author: Samet Öztürk

E-mail: drozturksamet@gmail.com

Received 15 October 2024; revision requested 26 November 2024; last revision received 13 January 2025; accepted 22 January 2025

Epub: 17.03.2025

DOI: 10.4274/dir.2025.242999

You may cite this article as: Öztürk S, Yüce M, Pamuk GG, Varlık C, Cimilli AT, Atay M. Automatic bone age assessment: a Turkish population study. *Diagn Interv Radiol*. 17 March 2025 DOI: 10.4274/dir.2025.242999 [Epub Ahead of Print].

Children's growth is characterized by non-linear progression, typically advancing in a sequential manner. Although metrics such as height and weight are useful for monitoring growth, bone development often provides the closest approximation to chronological age. The Greulich and Pyle (GP) and Tanner and Whitehouse (TW) methods are commonly employed for bone age assessment (BAA).[1,2] However, these methods rely on the expertise of radiologists and are subject to interpretation biases.[3] To address this, automatic BAA models have been developed, offering enhanced accuracy, repeatability, and efficiency.[4] Our study aims to evaluate the performance of deep learning algorithms within the Turkish population and enhance model efficacy at a population level. Additionally, we seek to demonstrate that establishing a model "from scratch" is feasible for a medium-sized hospital without relying on funds, grants, or dedicated commercial software.

Methods

Ethics approval

Approval was granted by the Non-Interventional Clinical Research Ethics Committee of University of Health Sciences Türkiye, Bağcılar Training and Research Hospital, with the ethics committee decision numbered 2023/09/08/051 and dated September 22, 2023. Informed consent was waived due to the retrospective nature of the study. All procedures in the present study involving human participants were performed in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Main points

- Population-specific deep learning model: We developed an automated model using the YOLOv8m architecture for hand detection and modified InceptionV3 for bone age assessment (BAA) tailored for the Turkish population by integrating Bağcılar Hospital, Radiological Society of North America, and Radiological Hand Pose Estimation datasets.
- Improved accuracy with combined data: The Combined model achieved a mean absolute error of 9.2 months and a 96% correlation with the reference standard, outperforming our single-source models.
- Clinical application and future prospects: This provides a consistent and efficient BAA tool, reducing radiologists' workload and variability. We aim to enhance accuracy with more diverse data and validate the model through broader clinical studies.

Data collection and dataset creation

Wrist and hand radiographs, bone age reports, and gender information for patients aged 0–18 years were collected from the Picture Archiving and Communication System without interpretational hindrances. A total of 2,933 radiographs conforming to Turkish standards were acquired from hospital records. Patients aged >18 years, images with severe artifacts or inappropriate field of view, and reports without BAA were excluded; 2,730 X-rays were found to be eligible (Figure 1). While integrating X-rays from Bağcılar into the dataset, evaluations by S.Ö. (who had 6 years of radiology experience) were compared with the clinical reading report. When the difference was ≤6 months, the report was deemed accurate. In cases where the difference was >6 months, A.T.C. (who had 32 years of radiology experience) and S.Ö. reevaluated images together, and a reference standard was obtained with a consensus decision. Additionally, two different open-source public datasets were incorporated [Radiological Society of North America (RSNA): https://www.rsna.org/rsnai/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017 and Radiological Hand Pose Estimation (RHPE): https://www.kaggle.com/datasets/ipythonx/rhpe-bone-age] with the filtering age range set to 0–18 years, resulting in a hybrid dataset sourced from various devices, vendors, and populations.[5,6] After filtering, the RSNA dataset consisted of 12,572 labeled radiographs, while the RHPE dataset included 6,185 labeled radiographs, and these datasets were further concatenated with the Bağcılar dataset. From the combination of all three datasets, an internal test dataset (n = 546) was created by randomly selecting 10% of Bağcılar data (n = 273) and an equal amount of public data (n = 273). The remaining data were used to create three distinct training and validation splits (Bağcılar, Public, and Combined), maintaining a 9:1 training-to-validation ratio (Figure 1). Bone age-based stratification was applied during the random splitting of each dataset using the train_test_split function from the scikit-learn Python library.

Model structure

In 2017, an RSNA BAA competition was held. The structure of the models used by the competitors and their error rates were published by Halabi et al.[5] The winner of the competition was a commercial company that profited from this work. The authors do not have any collaboration, partnership, or

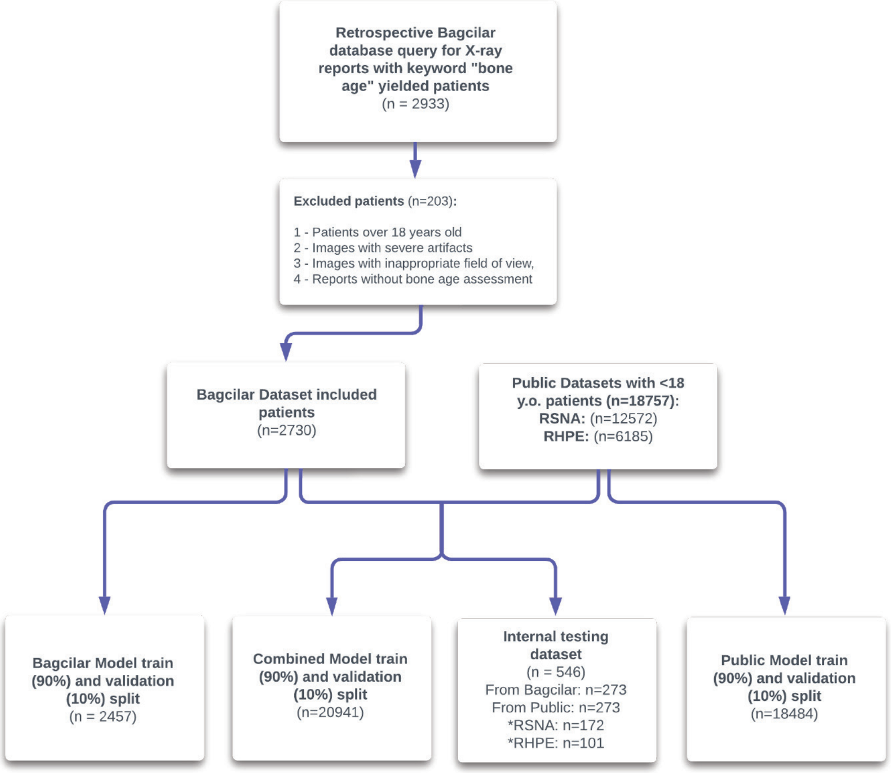
Retrospective Bagcilar database query for X-ray reports with keyword "bone age" yielded patients
(n = 2933)
Excluded patients (n=203):
1 - Patients over 18 years old
2 - Images with severe artifacts
3 - Images with inappropriate field of view,
4 - Reports without bone age assessment
Bagcilar Dataset included patients
(n=2730)
Public Datasets with <18 y.o. patients (n=18757):
RSNA: (n=12572)
RHPE: (n=6185)
Bagcilar Model train (90%) and validation (10%) split
(n = 2457)
Combined Model train (90%) and validation (10%) split
(n=20941)
Internal testing dataset
(n = 546)
From Bagcilar: n=273
From Public: n=273
*RSNA: n=172
*RHPE: n=101
Public Model train (90%) and validation (10%) split
(n=18484)

Figure 1. Flowchart illustrating the data collection process, inclusion and exclusion criteria, and dataset splitting methodology for the bone age prediction study. RSNA, Radiological Society of North America; RHPE, Radiological Hand Pose Estimation.

funding agreement with this company. The authors' models were built and trained "from scratch" using published architectures. As in the competition, a custom InceptionV3 model proved to be more suitable for this study. As a custom preprocessing step to improve model performance, a hand-detection model was also added using the YOLOv8m architecture. In the hospital's routine radiography acquisitions, some of the X-rays had a field of view large enough to include the elbow, whereas in others, phalanges were not included. The goal was to crop and adjust only the hand and wrist portion using YOLOv8m. Due to the heterogeneous nature of the hospital's dataset, encompassing images from diverse regions, a YOLOv8m model was initially trained for hand detection. Images were cropped with detected bounding boxes of the hand area before training the InceptionV3 model. All images were resized to 500 × 500 pixels, and an InceptionV3-based deep convolutional neural network (CNN) was constructed to process pixel information. Binary gender data (0 for female, 1 for male) were incorporated to account for gender effects via a densely connected layer with 32 neurons. Gender and pixel information were merged into a single network, followed by two densely connected layers with rectified linear unit activation, each containing 1,000 neurons, facilitating complex pattern learning. The output layer utilized mean absolute error (MAE) loss for regression simplification (Figure 2). A consistent model architecture was used throughout the study. It was trained and tested on three distinct datasets: the Bağcılar dataset, the public datasets (RSNA and RHPE), and a combined dataset consisting of both. For clarity, references to the "Bağcılar model," "Public model," and "Combined model" datasets pertain to the data used during model training and testing, not to distinct model architectures.

Model training process

The study utilized Keras 3.02, TensorFlow 2.15, and Python 3.9 for training, using an Nvidia RTX 3090 24GB graphics card. Data augmentation techniques, including rotation (up to 20 degrees), horizontal/vertical shifting (up to 20%), zooming (up to 20%), and horizontal flipping, were applied across the entire dataset to encourage the learning of patient-specific features. The final model was trained using Adam optimization with a batch size of 32 for 500 epochs. Learning rate adjustments and early stopping mechanisms were implemented. Models were trained and validated (90% training, 10% validation) and tested on an initially separated internal testing dataset, which was composed of an equally distributed number of images from both local and public sources (Figure 1).

Statistical analysis

Normality analysis was conducted using the Kolmogorov–Smirnov test. For comparisons between variables showing normal distribution, t-tests were employed, while one-way analysis of variance (ANOVA) was utilized for multiple variable comparisons. The Mann–Whitney U test and Kruskal–Wallis analysis were employed for variables that were not normally distributed. Post-hoc analyses were conducted using Bonferroni-corrected Mann–Whitney U and Tukey tests. A significance threshold of $P < 0.05$ was applied. Python version 3.9 was utilized for statistical analyses and plot generation.

Results

A total of 21,487 patients were included in the study, with a mean bone age of 10.4 ± 3.5 years, and 51% were female. A total of 18,757 cases were from public datasets (RSNA and RHPE), with a mean bone age of 10.5 ± 3.4 years and 50% female representation (Figure 3). The Bağcılar dataset had a mean bone age of 9.8 ± 3.9 years, with 38% female patients. Table 1 shows demographic data and information regarding the referring departments and International Classification of Diseases-10 (ICD-10) diagnosis codes for the Bağcılar dataset. The primary referring departments were general pediatrics (48%) and pediatric endocrinology (47.5%). The majority of cases (81%) were referred with preliminary diagnoses under the ICD-10 main category "endocrine, nutritional, and metabolic diseases."

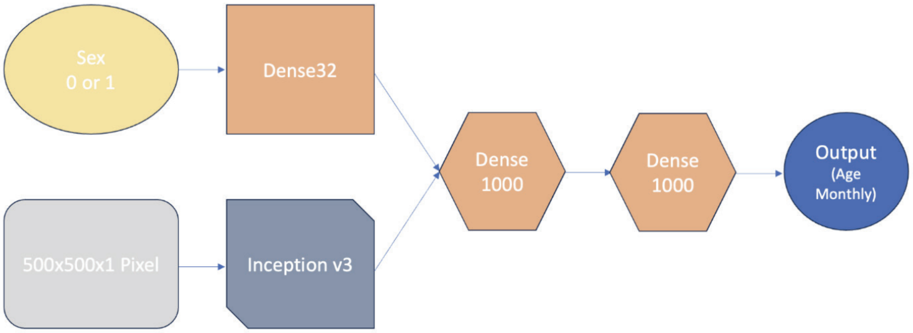

Figure 2. Architecture of the bone age prediction model: Combining sex input (encoded as 0 or 1) through a Dense32 layer and image input (500 × 500 × 1 pixels) processed via InceptionV3, followed by two dense layers (1,000 neurons each) to predict age in months.

The performance metrics for the models, evaluated in the internal testing dataset, showed that the Public model had an MAE of 11.3 months, with a mean squared error (MSE) of 302.1 and a root MSE (RMSE) of 17.4. The Bağcılar model (BM) showed a slightly higher MAE of 12.6 months but improved MSE and RMSE values of 260.3 and 16.1, respectively. The Combined model demonstrated the best overall performance, achieving the lowest MAE of 9.2 months, along with an MSE of 170.7 and an RMSE of 13.1, highlighting its superior accuracy compared with the other models.

Based on the internal testing dataset, the BM achieved bone age predictions within absolute differences of ≤6, ≤12, ≤18, and ≤24 months for 31%, 57%, 77%, and 88% of cases, respectively, with a Pearson correlation of 93%. The public dataset model (PM) achieved predictions within the same ranges for 45%, 69%, 81%, and 89% of cases, also with a Pearson correlation of 93%. The combined dataset model (CM) demonstrated the best performance, with predictions within ≤6, ≤12, ≤18, and ≤24 months for 44%, 73%, 87%, and 94% of cases, respectively, and a Pearson correlation of 96%, highlighting its superior accuracy and clinical utility.

Comparison of bone age predictions from the PM, BM, and CM models with the reference standard in the internal testing dataset revealed no statistically significant differences for any model, as determined by independent t-tests ($P > 0.05$).

The distribution of patients by age group and gender across the training, validation, and internal testing datasets is presented in Table 2. The mean and standard deviation of predicted bone ages alongside the reference standard for each age group and across models are shown in Table 3. Analyses of variance conducted for each age group between the three model assessments, and the reference standard revealed significant differences in the 0–3, 3–6, 6–9, 9–12, 12–15, and 15–18

Table 1. Clinical characteristics of patients in the Bağcılar dataset

	n	%
Gender		
Male	1,693	62
Female	1,037	38
Referring department		
General pediatrics	1,310	48
Pediatric endocrinology	1,296	47.5
Orthopedics	40	1.5
Other*	84	3
ICD-10 category**		
Endocrine, nutritional, and metabolic diseases	2,211	81
Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified	164	6
Factors influencing health status and contact with health services	104	3.8
Diseases of the blood and blood-forming organ and certain disorders involving the immune mechanism	65	2.4
Diseases of the respiratory system	46	1.7
Diseases of the genitourinary system	43	1.6
Diseases of the musculoskeletal system and connective tissue	33	1.2
Other	64	2.3

Average bone age was 9.8 years (standard deviation: 3.9). *Including mainly health board, family medicine, emergency department referrals; **ICD-10: International Classification of Diseases, tenth revision.

Table 2. Age and gender distribution of training and validation datasets for each model and internal testing dataset

			Age groups (years)					
	Gender	Split	0–3	3–6	6–9	9–12	12–15	15–18
Bağcılar model	**Female**	**Train**	64	184	363	462	165	131
		Val	9	19	46	52	18	13
	Male	**Train**	97	139	143	227	168	68
		Val	8	18	9	26	20	8
Public model	**Female**	**Train**	273	716	2,242	3,344	1,477	201
		Val	33	94	249	347	168	17
	Male	**Train**	251	945	1,265	1,807	3,445	669
		Val	27	91	146	208	387	82
Combined model	**Female**	**Train**	338	923	2,624	3,785	1,661	321
		Val	41	90	276	420	167	41
	Male	**Train**	346	1,065	1,395	2,030	3,615	743
		Val	37	128	168	238	405	84
Internal testing set	**Female**	**Test**	9	30	76	121	46	17
	Male	**Test**	21	37	36	57	71	25

groups ($P < 0.001$). Subsequently, a Tukey post-hoc analysis was carried out to elucidate the differences between models and the reference standard for each respective age group where significance was observed in the ANOVA:

- 0–3 and 3–6 years: The PM differed significantly from both the reference standard and the other models ($P < 0.001$).
- 6–9 years: Significant differences were observed between the BM and both the reference standard ($P = 0.016$) and the CM ($P = 0.042$). The PM also differed significantly from the reference standard ($P = 0.0003$) and the CM ($P = 0.001$).
- 9–12 years: Significant differences were found between the reference standard and the BM ($P = 0.034$) as well as between the BM and PM ($P = 0.002$).
- 12–15 years: The BM differed significantly from the reference standard ($P = 0.005$) and the CM ($P = 0.002$).
- 15–18 years: The BM showed significant differences compared with the reference standard ($P = 0.011$) and the CM ($P = 0.003$).

These findings are also shown with boxplots in Figure 4 and indicate that while significant differences exist among certain models and age groups, the degree of discrepancy varies, emphasizing the variability in model performance across age groups. Notably, no significant difference was found between the reference standard and the CM across all age groups.

The MAEs of the models in the Public internal testing data were 6.2, 6.9, and 12.5 months for the PM, CM, and BM, respectively. The ANOVA conducted on the absolute error differences of the three model predictions in the public internal testing dataset revealed a statistically significant difference ($s = 60.01$, $P < 0.001$). Tukey post-hoc analysis of the model assessments in the Public internal testing dataset showed that the BM had a statistically significantly lower performance compared with the PM and CM, with MAE differences of 6.3 and 5.5 months, respectively ($P < 0.05$). There was no significant difference between the PM and CM ($P > 0.05$) (Table 4).

The MAEs of the models in the Bağcılar internal testing dataset were 16.5, 11.4, and 12.7 months for the PM, CM, and BM respectively. Analyses of variance among the absolute error differences of the three models in the Bağcılar internal testing dataset found a statistically significant difference ($s = 11.19$, $P < 0.001$). In the Tukey post-hoc analysis conducted among the model assessments in the Bağcılar test dataset, the PM showed a statistically significantly lower performance compared with the BM and CM, with MAE differences of 3.8 and 5 months, respectively ($P < 0.05$). However, no significant difference was observed between the BM and CM ($P > 0.05$) (Table 5).

Bland–Altman plots were generated to display the differences between the BM, PM, CM, and the reference standard in months,

Table 3. Mean and standard deviation of reference standard and predicted bone ages for different age groups in the internal testing dataset

(Months)	Public model predictions	Bağcılar model predictions	Combined model predictions	Reference standard
0–36	46.8 ± 27.3	30.6 ± 11.2	31.6 ± 10.8	29.9 ± 6.4
36–72	71.6 ± 17.6	64.2 ± 19.5	57.8 ± 14.1	60.6 ± 9.8
72–108	105.9 ± 19.9	103.3 ± 18.2	97.3 ± 18.1	96.6 ± 9.2
108–144	131.8 ± 16.1	126.4 ± 15.9	128.5 ± 14.7	130.5 ± 9.1
144–180	162.2 ± 16.5	155.5 ± 19.5	163.2 ± 17.9	162.6 ± 9.2
180–216	194.8 ± 14.4	189.0 ± 16.2	199.8 ± 15.2	198.6 ± 9.3

Table 4. Post-hoc Tukey test for Public internal testing data. In the analysis of mean absolute error (MAE) for Public internal testing data, no significant difference was observed between the Combined model (CM) and the Public model (PM)

Post-hoc, Tukey test for Public test data						
Group 1	Group 2	MAE difference (months)	*P* value	Lower (months)	Upper (months)	Significant
BM MAE	CM MAE	−5.50	<0.001	−6.97	−4.03	Yes
BM MAE	PM MAE	−6.28	<0.001	−7.75	−4.81	Yes
CM MAE	PM MAE	−0.78	0.43	−2.24	0.68	No

Table 5. Post-hoc Tukey test for Bağcılar internal testing data. In the analysis of mean absolute error (MAE) for Bağcılar internal testing data, no significant difference was observed between the Bağcılar model (BM) and the Combined model (CM)

Post-hoc, Tukey test for Bağcılar internal testing data						
Group 1	Group 2	Mean difference (months)	*P* value	Lower (months)	Upper (months)	Significant
BM MAE	CM MAE	−1.24	0.496	−3.84	1.35	No
BM MAE	PM MAE	3.77	0.002	1.18	6.37	Yes
CM MAE	PM MAE	5.02	<0.001	2.43	7.62	Yes

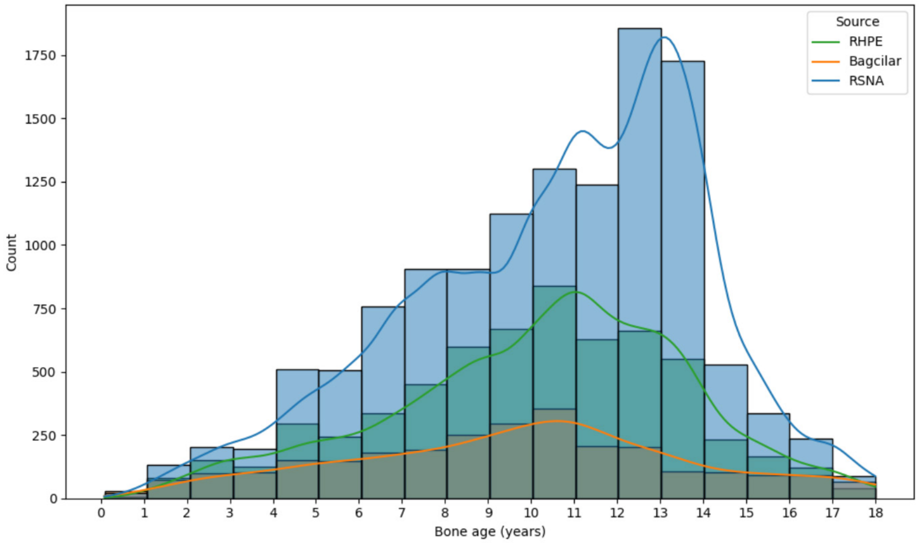

Figure 3. Bone age distribution across datasets: Histogram plots showing the distribution of bone ages (in years) for the Radiological Hand Pose Estimation, Bağcılar, and Radiological Society of North America datasets.

RHPE, Radiological Hand Pose Estimation; RSNA, Radiological Society of North America.

highlighting the variance between the model assessments and the reference standard within the internal testing dataset (Figure 5). Additionally, scatter plots with linear regression lines were created for each model to provide a clearer understanding of their performance across different internal testing datasets (Figure 6).

Discussion

The accuracy of BAA is largely dependent on the experience of the physician, as traditional evaluation is often a subjective estimation. Traditional assessment studies are typically conducted by experienced physicians through visual inspection and manual marking. This process requires significant time and effort, and different physicians may have varying standards when evaluating the same radiograph. Therefore, automated assessment approaches for BAA are increasingly gaining interest.

On average, an experienced radiologist spends approximately 1.4 minutes using the GP method and 7.9 minutes using the TW method to assess a patient.[7] Moreover, both methods are associated with high intra- and inter-observer variability. The reported range of BAA results averages 0.96 years (11.5 months) for GP and 0.74 years (8.9 months) for TW.[8] In some stages of child development, changes can be very subtle, especially after the age of 14 years, and the sensitivity perceivable by the human eye through radiological examination may be lacking.[9] The absence of significant differences between the predicted bone age using our proposed models and those obtained using the GP and TW methods enhances the reliability of our approach.

Our model utilized upper extremity radiographs containing hand and wrist regions with bone age reports sourced from Bağcılar. The images were taken with different presets and exposures, resulting in an inhomogeneous dataset. Some radiographs did not include joints prioritized in

BAA. Others were not captured at the correct position or angle. In some cases, bones were superimposed, and the parent's hand was often visible in infant radiographs. Considering this data heterogeneity, our model better reflects daily clinical practice compared to similar studies.

In this study, public datasets and data from Bağcılar Hospital were used as sources. Racial differences, imaging parameters, and image quality may have influenced the results. However, we believe that a model trained with these parameters could be more consistent than the inter- and intra-observer variability associated with the GP and TW methods.[8] Further prospective studies are needed to assess the added value of such models in daily clinical practice.

Recently, many deep learning methods have been developed for BAA, and RSNA even organized a competition for this purpose.[5] With the developed methods, the timing and pattern of ossification centers according to age can be extracted from images using deep learning techniques for BAA. Thus, this process, which is time-consuming and subjective, with differences between evaluators and even variations within the evaluator, can be carried out on more solid foundations.[10] In recent studies, we see models where various ensemble techniques are employed, combining multiple models into one.[11] Liu et al.[12] suggested that ranking learning may be a more suitable approach for the BAA task than classification and regression. In their study, they achieved accurate BAAs with an MAE of approximately 6 months using a proposed method based on a rank-monotonic enhanced ranking CNN.[12] Li et al.[13] developed a two-stage, fully automatic model

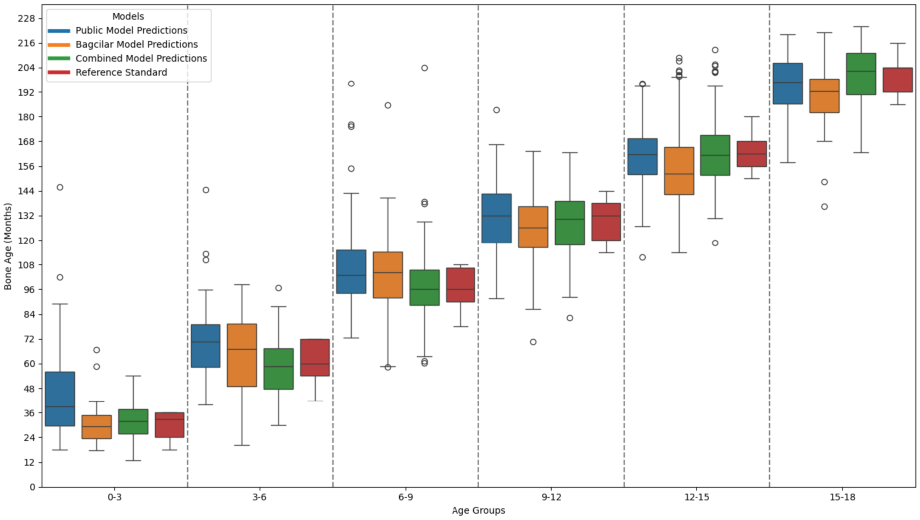

Figure 4. Boxplot of bone age predictions for each model across age groups.

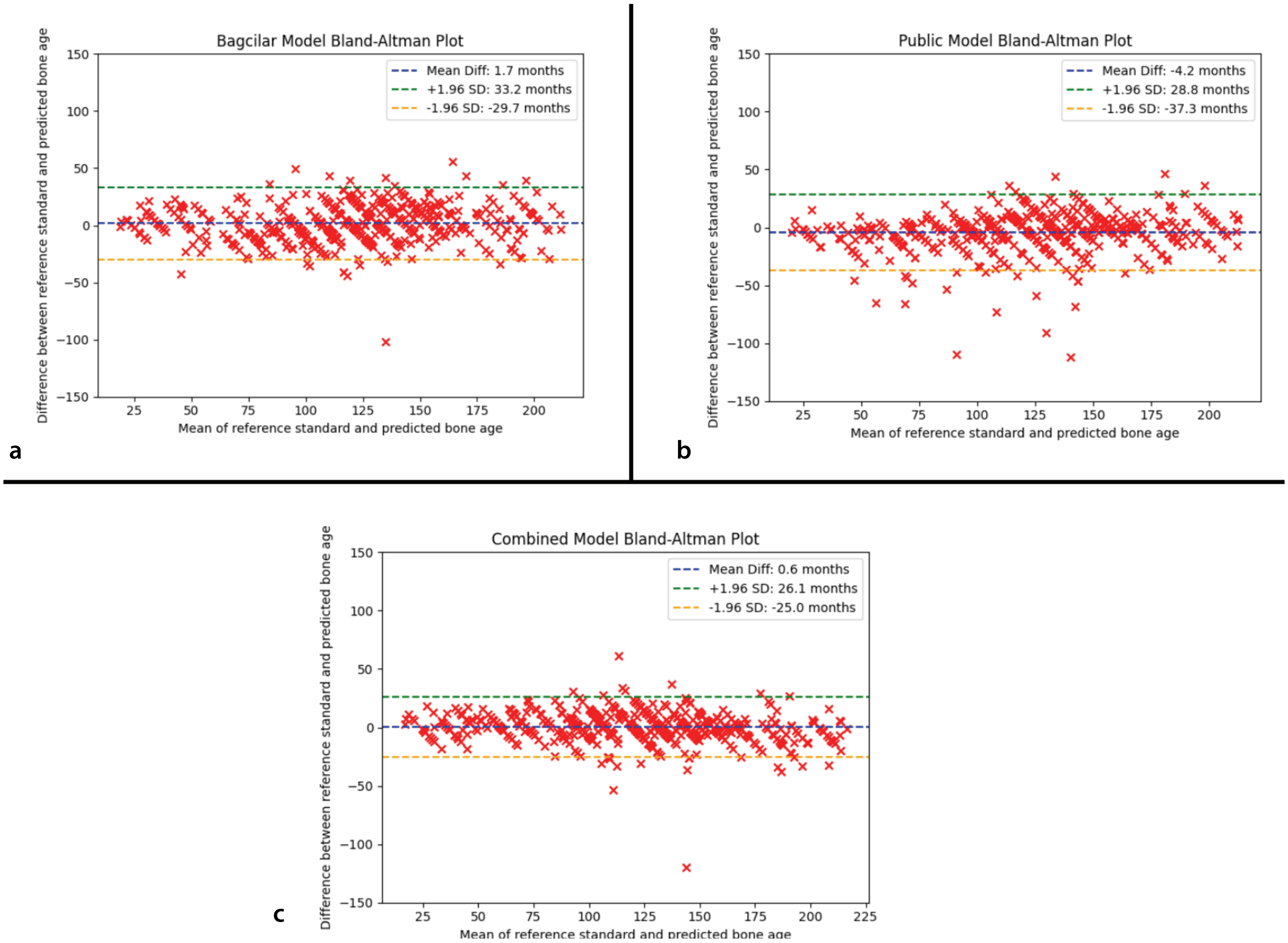

Figure 5. Bland–Altman plots showing the difference between the Bağcılar model, Public model, Combined model, and the reference standard in months, illustrating the variance between the reference standard and the model assessment in the internal testing dataset.

that does not require manual annotation. They demonstrated MAEs of 5.45 months on the RSNA dataset and 3.34 months on a specific dataset.[13] Similarly, our model does not involve annotation. It is an end-to-end model where the bone age is directly assessed by using the cropped hand portion of the X-ray alongside gender information as input. Since our main goal is to show the effect of population differences on model performances, we preferred a validated method that produced the best performance in the RSNA 2017 bone age prediction challenge.

Kim et al.[14] developed a model based on a completely Korean, healthy population, assuming chronological age as the real bone age, such as an atlas study. The developed deep learning model followed a rigorous preprocessing process for estimating chronological age from hand radiographic images. Background removal and transformation networks were applied using manual annotations from an experienced musculoskeletal radiologist. ResNet-50 was used as the basic architecture for age estimation. Compared with their GP-based model, the Korean model showed a lower MAE (8.2 vs. 10.5 months; *P* = 0.002). Additionally, the rate of BAAs within 6 months of chronological age was higher (44.5% vs. 36.4%; *P* = 0.04) with the Korean model. Similarly, our study is also a population-specific model study. In their model, many radiographs were not used as it was based on a non-patient population, such as an atlas. Consequently, there were 21,036 training sets left, and separate test datasets were obtained from two institutions, consisting of 343 and 321 data sets, respectively. Manual annotations were used in creating the model, which is generally time-consuming and cumbersome. Our developed model demonstrated performance comparable with existing models. Utilizing heterogeneous datasets plays a critical role in enhancing model generalizability by exposing the algorithm to a wider range of population and imaging variations. This diversity allows the model to better identify under-represented patterns and reduces the risk of overfitting to specific subsets. The improved performance of the Combined model compared with the locally trained BM underscores the importance of incorporating data from heterogeneous sources in achieving better generalization. Furthermore, increasing the diversity of the included population and imaging modalities can further enhance these models by enabling them to capture relevant information from under-represent-

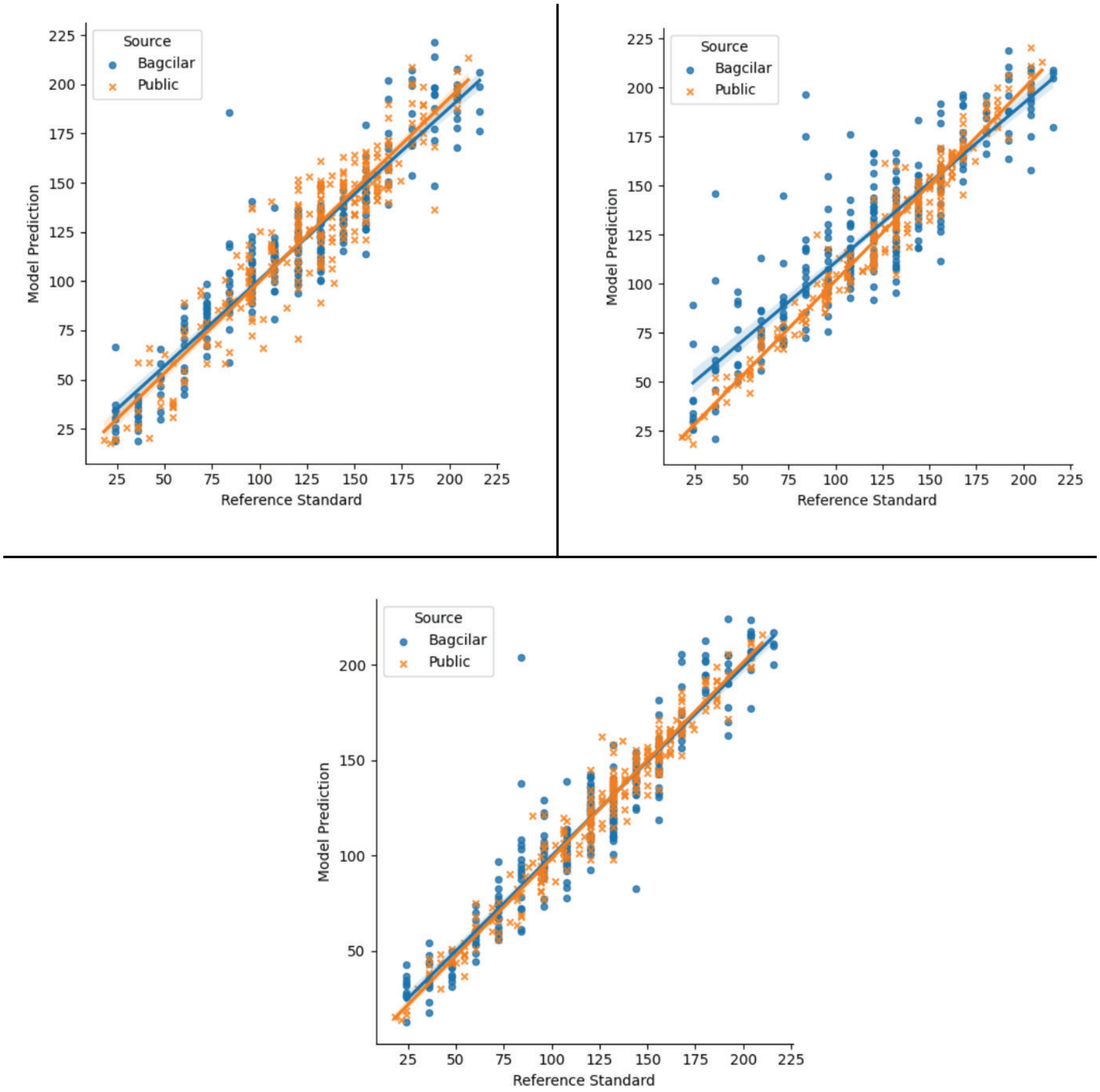

Figure 6. Bone age assessments of the **(a)** Bağcılar model, **(b)** Public model, and **(c)** Combined model on the internal testing dataset in months. Translucent bands around the regression line represent confidence intervals.

ed portions of the data. Greater diversity ultimately strengthens model robustness and improves its capacity to extract meaningful insights. In our model, X-rays requested for BAA and previously reported by radiologists were used. Even though the demographic and diagnostic information was not extensively available in public datasets, models developed using these sources performed worse on local data, indicating important population differences alongside data curation-related information loss.

Spampinato et al.[15] achieved an MAE of 9.6 months using Bonet and the RSNA dataset. Larson et al.[16] achieved an MAE of 6.24 months on the RSNA dataset with a deep residual network structure based on the GP mapping method using ResNet50. Pan et al.[17] used a U-Net model to segment hand mask images from raw X-ray images, employing a deep active learning technique that reduces annotation burden, achieving an MAE of 8.59 months on the RSNA dataset. In our developed Combined model, the MAE value for all data was 9.2 months, 6.9 months for the public dataset, and 11.4 months for the Bağcılar dataset. The described methods, similar to our model, do not involve annotation. Annotation-based methods involve using processed images and adding manual bounding box annotations to these images. These strategies can extract features from specific regions based on prior knowledge and then generate age estimates. Annotation-based methods, which involve additional manual annotations, generally exhibit better performance and higher accuracy compared with annotation-free methods. However, manual annotation is time-consuming and has made it difficult for experimental methods to transition to clinical applications.

Unlike many previous studies that rely on homogeneous datasets, our model was trained and validated using a heterogeneous dataset that includes radiographs from both Bağcılar and public datasets (RSNA and RHPE). This dataset reflects a wide range of imaging conditions, patient demographics, and ethnic backgrounds, thereby increasing the model's robustness and generalizability to real-world clinical settings. The inclusion of such diverse data sources is crucial, as it enables the model to handle a broader spectrum of clinical scenarios, making it more applicable across different populations. The results of our study are promising and highlight the potential of automated BAA models. The Combined model, which integrated data from both Bağcılar and public datasets, demonstrated a high Pearson correlation of 96% with the reference standard, indicating strong predictive accuracy. Specifically, the Public model achieved an MAE of 11.3 months when tested across all test data, while the BM had a higher MAE of 12.6 months. However, when data from both sources were combined, the MAE improved to 9.2 months, highlighting the advantage of integrating diverse datasets to enhance model performance. This improvement could be attributed to the increase in the number of data and the model's increased focus on significant areas due to heterogeneity, enabling the model to account for these differences more effectively, resulting in more accurate and reliable assessments.

The importance of data diversity is further emphasized when examining the model's performance across different age groups. The Combined model showed consistent accuracy across various age ranges, particularly during the critical growth periods of 9–12 and 12–15 years. In contrast, the BM alone exhibited significant deviations from the reference standard in these age groups. This consistency across age groups is crucial for clinical application, as it ensures that the model can be reliably used across a broad patient demographic, minimizing the risk of misclassification and improving overall patient care.

The study has several limitations. Primarily, the limited data quantity has been a key factor, particularly with a small number of radiographs for children under 3 years and a considerably low amount of high-quality data. Another limitation is the absence of a study demonstrating inter-observer differences in our Bağcılar dataset. However, there are many studies in the literature addressing this issue. Additionally, the bone ages in our data were determined using manual methods, such as GP and TW, which, despite having their own limitations, are commonly used in daily practice. Nevertheless, there was no statistically significant difference found between the bone ages obtained with our Combined model and those obtained with clinical methods. Furthermore, the model's performance in older adolescents (aged 15–18 years) showed higher MAEs compared with younger age groups. This could be due to the increased complexity of bone maturation patterns in these age ranges, where small differences in ossification can lead to significant variations in BAA. Addressing this issue may require the development of more specialized models or the inclusion of additional features, such as hormonal markers or elbow and shoulder X-rays, which could provide further insights into bone development in these populations.

In conclusion, this study presents the development of an automatic BAA model using data from Bağcılar, RSNA, and RHPE, making it one of the few studies to incorporate a Turkish population in deep learning-based BAA research. Our model is particularly notable for its ability to integrate heterogeneous data, demonstrating that the inclusion of diverse datasets can significantly enhance model performance. The proposed models offer the advantage of automated analysis without any need for annotation.

Despite the challenges posed by population-level differences, heterogeneous data, and image quality issues, these models can be effectively adopted in various clinical environments, and accuracies can be increased over time with prospectively cumulating data. By enabling more accurate and efficient BAAs, our approach offers valuable support to healthcare professionals, particularly in settings where traditional methods are time-consuming and subject to variability.

Future research should aim to expand the dataset, particularly for younger and older age groups, to improve the model's accuracy and generalizability. Additionally, exploring the incorporation of other clinical parameters, such as hormonal levels, could provide a more comprehensive assessment of bone age, particularly in complex cases. Finally, further validation studies, including prospective trials and cross-institutional collaborations, will be crucial for ensuring the widespread adoption and clinical utility of automated BAA models.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. *Am J Med Sci*. 1959;238(3). [Crossref]

2. Tanner JM. Assessement of skeletal maturity and predicting of adult height (TW2 method). Prediction of adult height. Published online 1983:22-37. [Crossref]

3. 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;209(6):1374-1380. [Crossref]

4. Nadeem MW, Goh HG, Ali A, Hussain M, Khan MA, Ponnusamy VA. Bone age assessment empowered with deep learning: a survey, open research challenges and future directions. *Diagnostics (Basel)*. 2020;10(10):781. [Crossref]

5. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA pediatric bone age machine learning challenge. *Radiology*. 2019;290(2):498-503. [Crossref]

6. Escobar M, González C, Torres F, Daza L, Triana G, Arbeláez P. Hand pose estimation for pediatric bone age assessment. In: Shen D, ed. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science). *Springer, Cham*; 2019. [Crossref]

7. Christoforidis A, Badouraki M, Katzos G, Athanassiou-Metaxa M. Bone age estimation and prediction of final height in patients with beta-thalassaemia major: a comparison between the two most common methods. *Pediatr Radiol*. 2007;37(12):1241-1246. [Crossref]

8. 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):848-851. [Crossref]

9. Yekeler E. Kemik Yaşı Atlası. First edition; 2021.

10. Lee BD, Lee MS. Automated bone age assessment using artificial intelligence: the future of bone age assessment. *Korean J Radiol*. 2021;22(5):792-800. [Crossref]

11. Pan I, Thodberg HH, Halabi SS, Kalpathy-Cramer J, Larson DB. Improving automated pediatric bone age estimation using ensembles of models from the 2017 RSNA machine learning challenge. *Radiol Artif Intell*. 2019;1(6):e190053. [Crossref]

12. Liu B, Zhang Y, Chu M, Bai X, Zhou F. Bone age assessment based on rank-monotonicity enhanced ranking CNN. *IEEE Access*. 2019;7:120976-120983. [Crossref]

13. Li Z, Chen W, Ju Y, et al. Bone age assessment based on deep neural networks with annotation-free cascaded critical bone region extraction. *Front Artif Intell*. 2023;6:1142895. [Crossref]

14. Kim PH, Yoon HM, Kim JR, 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(11):1151-1163. [Crossref]

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

16. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*. 2018;287(1):313-322. [Crossref]

17. Pan X, Zhao Y, Chen H, Wei D, Zhao C, Wei Z. Fully automated bone age assessment on large-scale hand X-ray dataset. *Int J Biomed Imaging*. 2020;2020:8460493. [Crossref]