Original Article | Pediatric Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2023.0092
Korean J Radiol 2023;24(11):1151-1163

Korean Journal of Radiology
KJR

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

Pyeong Hwa Kim[1], Hee Mang Yoon[1], Jeong Rye Kim[2], Jae-Yeon Hwang[3], Jin-Ho Choi[4], Jisun Hwang[5], Jaewon Lee[6], Jinkyeong Sung[6], Kyu-Hwan Jung[6], Byeonguk Bae[6], Ah Young Jung[1], Young Ah Cho[1], Woo Hyun Shim[1,7], Boram Bak[8], Jin Seong Lee[1]

[1]Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[2]Department of Radiology, Dankook University Hospital, Dankook University College of Medicine, Cheonan, Republic of Korea
[3]Department of Radiology, Research Institute for Convergence of Biomedical Science and Technology, Pusan National University Yangsan Hospital, Pusan National University School of Medicine, Yangsan, Republic of Korea
[4]Department of Pediatrics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[5]Department of Radiology, Ajou University Hospital, Ajou University School of Medicine, Suwon, Republic of Korea
[6]VUNO Inc., Seoul, Republic of Korea
[7]Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[8]University of Ulsan Foundation for Industry Cooperation, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

Objective: To develop a deep-learning-based bone age prediction model optimized for Korean children and adolescents and evaluate its feasibility by comparing it with a Greulich-Pyle-based deep-learning model.

Materials and Methods: A convolutional neural network was trained to predict age according to the bone development shown on a hand radiograph (bone age) using 21036 hand radiographs of Korean children and adolescents without known bone development-affecting diseases/conditions obtained between 1998 and 2019 (median age [interquartile range {IQR}], 9 [7–12] years; male:female, 11794:9242) and their chronological ages as labels (Korean model). We constructed 2 separate external datasets consisting of Korean children and adolescents with healthy bone development (Institution 1: n = 343; median age [IQR], 10 [4–15] years; male: female, 183:160; Institution 2: n = 321; median age [IQR], 9 [5–14] years; male: female, 164:157) to test the model performance. The mean absolute error (MAE), root mean square error (RMSE), and proportions of bone age predictions within 6, 12, 18, and 24 months of the reference age (chronological age) were compared between the Korean model and a commercial model (VUNO Med-BoneAge version 1.1; VUNO) trained with Greulich-Pyle-based age as the label (GP-based model).

Results: Compared with the GP-based model, the Korean model showed a lower RMSE (11.2 vs. 13.8 months; $P = 0.004$) and MAE (8.2 vs. 10.5 months; $P = 0.002$), a higher proportion of bone age predictions within 18 months of chronological age (88.3% vs. 82.2%; $P = 0.031$) for Institution 1, and a lower MAE (9.5 vs. 11.0 months; $P = 0.022$) and higher proportion of bone age predictions within 6 months (44.5% vs. 36.4%; $P = 0.044$) for Institution 2.

Conclusion: The Korean model trained using the chronological ages of Korean children and adolescents without known bone development-affecting diseases/conditions as labels performed better in bone age assessment than the GP-based model in the Korean pediatric population. Further validation is required to confirm its accuracy.

Keywords: Pediatrics; Bone age; Deep-learning; Convolutional neural network

Received: July 8, 2022 **Revised:** August 1, 2023 **Accepted:** August 6, 2023

Corresponding author: Hee Mang Yoon, MD, PhD, Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea
• E-mail: espoirhm@gmail.com

See the invited Editorial by Cheon et al., in volume 24(11) on page 1059 to 1060, https://doi.org/10.3348/kjr.2023.0895.

INTRODUCTION

Bone age, generally evaluated using hand and wrist radiographs, is a representative index that reflects skeletal maturation in children and adolescents. Growth disorders should be considered when there is a considerable discrepancy between chronological and observed bone ages (discrepancy > 2 standard deviations [SDs]) [1,2]. Determining bone age is also useful for making surgical decisions in orthopedics [3] and forensics [4]. Among the methods available for bone age determination, the atlas-based Greulich-Pyle (GP) method is one of the most widely used methods [5]. However, there is an unresolved issue with the conventional GP method that needs to be addressed: no standardized protocol is available for determining how different bones should be weighted when assessing bone age, leading to unignorable inter-institutional and inter- and intraobserver variability [6-8]. Therefore, an automated GP-based bone age prediction system was introduced, which demonstrated high accuracy, reproducibility, and time efficiency [9-12].

However, whether this method applies to the current Korean pediatric population remains questionable, as the GP method was derived from a mostly white pediatric population at the upper socioeconomic level almost a century ago [5]. A meta-analysis demonstrated significant differences between GP-based bone and chronological ages in Asian boys [13]. Furthermore, Zhang et al. [14] showed advanced bone age in Asian boys (11–15 years) and girls (10–13 years) compared with Caucasians when bone age was based on the GP method. Ontell et al. [15] also reported delayed bone age in the preadolescent period and advanced bone age in the adolescent period in Asian boys. A previous study also affirmed that the contemporary Korean pediatric population showed different rates of skeletal maturation compared with GP-based bone age estimates [16]. Therefore, a standard bone age chart based on the Tanner–Whitehouse 2 (TW2)-20 score using 3407 radiographs of Korean children was introduced in 1996 [17]. However, the relatively small sample size and considerable time required for assessment based on the TW2 method [18] limits its wide application in clinical practice.

In this context, a deep-learning-based bone age prediction model focusing on the Korean pediatric population may offer a simple and reproducible bone age assessment optimized for the Korean pediatric population, reflecting relevant ethnic and environmental factors. Therefore, we developed a deep-learning-based bone age prediction model using hand and wrist radiographs obtained from healthy Korean children and adolescents with their chronological age as the label and subsequently evaluated its feasibility by comparing it with a GP-based deep-learning bone age assessment system.

MATERIALS AND METHODS

This retrospective study was conducted in accordance with the checklist for Artificial Intelligence in Medical Imaging [19]. Approval was obtained from the Institutional Review Board of each participating institution. The requirement for informed patient consent was waived by the institutional review boards because of the retrospective nature of the study.

Study Design and Datasets

As we can assume that bone age is the same as chronological age in individuals whose bone development is normal, our strategy for training a model (Korean model) was to use hand and wrist radiographs from Korean pediatric individuals with normal bone development as input data and their chronological ages as labels. Therefore, we collected left hand and wrist radiographs of Korean children and adolescents who showed normal bone development without any genetic, endocrinologic, or other chronic diseases. A systematic computerized search of the database of Asan Medical Center was performed to identify all left hand and wrist radiographs of eligible pediatric patients (aged < 20 years) obtained between 1998 and 2019. Radiographs of patients meeting any of the following criteria were excluded: 1) confirmed precocious puberty (testis ≥ 4 mL before the age of 9 years in males; Tanner 2 or higher stage secondary sexual characteristics before the age of 8 years in females), 2) confirmed delayed puberty (absent secondary sexual characters in males aged ≥ 14 years or females aged ≥ 13 years), 3) abnormal growth rate (< 4–6 cm/year) in the prepubertal period [20], 4) abnormally short stature compared with the normal Korean pediatric population of the same age (height less than the third percentile for age and sex according to the Korean population-based reference) [21], 5) any confirmed congenital anomalies, 6) underlying chronic disease potentially affecting growth, 7) use of medications affecting bone growth or metabolism (recombinant human growth hormone therapy or corticosteroids), 8) evident soft tissue or bone tumors noted on the radiographs, 9) fractures with/without dislocations noted on the radiographs, 10) amputation or excision state, and 11) radiographs with poor image quality or wrong patient positioning. If the relevant information was unavailable (growth rate or height), the radiograph was considered normal because little medical attention was required in such cases. Images and clinical charts were initially screened by one researcher (Boram Bak with 2 years of clinical experience as a radiology technician). During screening, the inclusion/exclusion of complex cases was confirmed daily by an experienced pediatric radiologist (H.M.Y., with 9 years of experience in pediatric radiology). Additionally, randomly selected radiographs were double-checked by a pediatric radiologist (P.H.K., with 2 years of experience in pediatric radiology). The exclusion was confirmed by consensus between the two radiologists, considering the extracted clinical information and image quality.

Model Development: Preprocessing

Several preprocessing steps were performed to achieve consistency and reduce the complexity of the input data. These steps included the use of a background removal network and transformation network intended to locate the hand in a consistent position [22]. To train the preprocessing models, all labels were manually annotated using the Radiological Society of North America hand bone age dataset [23] by an experienced musculoskeletal radiologist (J.S., with 7 years of experience in musculoskeletal radiology). The images were first resized to a uniform size of 512 x 512 pixels, and background removal was performed on the downsampled image. Finally, a series of transformations (translation and rotation) was performed to make the scale and position of the hands constant across all the radiographic images. Both hand segmentation and transformation networks used a high-resolution network for the network architecture [24]. The preprocessing steps were conducted by an experienced computer programmer (J.L., with 3 years of experience in programming).

Model Development: Convolutional Neural Network

The ResNet-50 deep convolutional neural network model [25] was trained to estimate the chronological age at 1-month intervals (Fig. 1). After each convolutional layer, a rectified linear unit activation function was employed for nonlinearity, and batch normalization was performed to avoid overfitting. Finally, the flattened feature vectors from the global average pooling layer were fed into a final fully connected layer with 256 nodes. Each node of the last fully connected layer corresponded to a patient's chronological age at 1-month intervals.

To accelerate model convergence and achieve a better estimation performance, the model parameters were pretrained on ImageNet [26], except for the last fully connected layer, whose biases were set to zero. All weights were randomly initialized within a range of -0.5 to 0.5. Data augmentation was performed with random rotation, scaling, and horizontal flipping. The model was trained by minimizing the DLDLv2 objective function [27] using an Adam optimizer. The initial learning rate was $1e^{-3}$, which was dropped to $1e^{-5}$ over 300 training epochs. The model was trained using PyTorch 1.7 (Linux; https://pytorch.org) in Python 3.6. The performance of the developed Korean model was internally validated using four-fold cross-validation.

External Validation Dataset

For external validation, two separate datasets were obtained from the Pusan National University Yangsan Hospital (Institution 1; from January 2008 to November 2022) and Dankook University Hospital (Institution 2; from January 2005 to November 2022). These consisted of left hand and wrist radiographs obtained in the clinical setting of trauma in Korean children and adolescents without known underlying diseases or conditions affecting bone development. Similarly, if relevant information was unavailable (growth rate or height), the radiograph was considered normal because little medical attention was required in such cases. Radiographs were included only when two radiologists determined them to be satisfactory for bone age assessment (H.M.Y. and P.H.K., with 9 and 2 years of experience in pediatric radiology, respectively).

Statistical Analysis

The developed Korean model was externally validated using two separate datasets: one from Institution 1 and the other from Institution 2. The reference standard was the chronological age of the participants. For comparison, a GP-based automated bone age prediction model (VUNO Med-BoneAge version 1.1; VUNO) was applied to the test datasets. VUNO Med-BoneAge (VUNO) is a commercially available automated bone age prediction system trained using left-hand radiographs of a Korean pediatric population, with GP-based bone ages as labels. The performance of each system was graphically estimated using a scatter plot. Notably, for statistical analysis, the

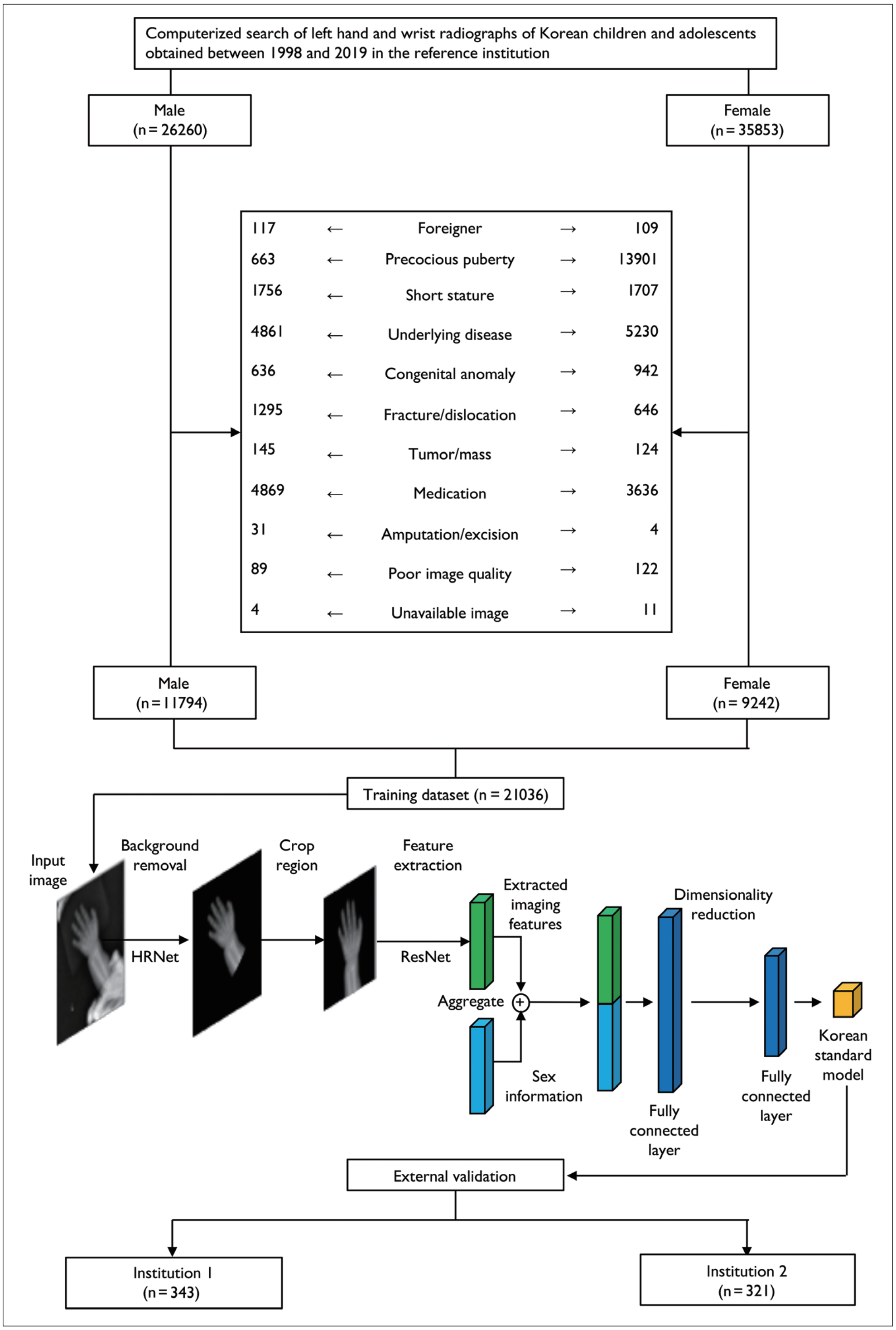
Computerized search of left hand and wrist radiographs of Korean children and adolescents obtained between 1998 and 2019 in the reference institution
Male (n = 26260)
Female (n = 35853)
117 ← Foreigner → 109
663 ← Precocious puberty → 13901
1756 ← Short stature → 1707
4861 ← Underlying disease → 5230
636 ← Congenital anomaly → 942
1295 ← Fracture/dislocation → 646
145 ← Tumor/mass → 124
4869 ← Medication → 3636
31 ← Amputation/excision → 4
89 ← Poor image quality → 122
4 ← Unavailable image → 11
Male (n = 11794)
Female (n = 9242)
Training dataset (n = 21036)
Input image
Background removal
HRNet
Crop region
Feature extraction
ResNet
Extracted imaging features
Aggregate
Sex information
Dimensionality reduction
Fully connected layer
Fully connected layer
Korean standard model
External validation
Institution 1 (n = 343)
Institution 2 (n = 321)

Fig. 1. Flow diagram of patient selection and dataset organization with a schematic illustration of the development of the deep convolutional neural network model (Korean model). Among the 62113 radiographs identified through the computerized search, 21036 were used for model development. A convolutional neural network was trained to predict choronologic age (reference standard) at 1-month intervals. Two separate external datasets consisting of Korean children and adolescents with healthy bone development were used to test the model performance. HRNet = high-resolution net

bone age prediction of the GP-based model was calculated by summing all bone ages multiplied by their predicted probabilities (the VUNO score). The mean absolute error (MAE) and root mean square error (RMSE) were also calculated and compared between the Korean- and GP-based models using generalized estimating equations to account for patient clustering effects. The proportions of bone age predictions within 6, 12, 18, and 24 months of chronological age were calculated and compared between the Korean and GP-based models using chi-square tests. Additionally, Bland–Altman plot analysis was performed between the chronological age and ages predicted by the Korean and GP-based models to identify any systemic differences between the measurements. The presence of systemic trend differences between chronological and predicted ages, that is, age-dependent bias, was assessed using univariable linear regression analysis based on the Bland–Altman plot.

Because the most probable bone age and not the VUNO score (calculated bone age weighted by probabilities) is generally used in clinical practice, its accuracy should also be evaluated. However, directly comparing the most probable GP-based bone age with bone age predicted by the Korean model was inappropriate because the Korean model was presented in months. Therefore, to ensure the comparability between the two models, we also obtained the GP-based bone age that was most similar to bone age predicted by the Korean model. We subsequently compared the most probable bone age by the GP model with the GP-based bone age that was most similar to the bone age predicted by the Korean model.

Given that the GP-based model has shown suboptimal predictive performance for ages < 2 years [28] and growth reaches a plateau around the age of 14 years in females and 16 years in males [21], we performed a subgroup analysis restricted to ages of 2–16 years for males and 2–14 years for females. MAE, RMSE, and Bland–Altman plot analyses were similarly performed for this subgroup. To compare the magnitude of age-dependent bias between the two models, the coefficients of linear regression model in the Bland–Altman plot were compared by testing the interaction effect of the bone age prediction model.

Generalized estimating equations were performed using IBM SPSS Statistics for Windows (version 23.0; IBM Corp.), whereas other statistical analyses were performed using R software (version 3.6.3.; R Foundation for Statistical Computing). Differences were considered statistically significant at $P < 0.05$.

RESULTS

Dataset Characteristics

Among the 62113 radiographs identified through the computerized search, 21036 (median age [interquartile range {IQR}], 9 [7–12] years; male: female, 11794:9242) were used for model development (Fig. 1). The reasons for the radiograph acquisitions were as follows: growth evaluation (n = 16984); trauma (n = 3056); congenital anomaly work-up (n = 604); pain/swelling (n = 330); and soft tissue, bone mass, or tumor evaluation (n = 62).

For external validation, 343 radiographs from Institution 1 (median age [IQR], 10 [4–15] years; male: female, 183:160) and 321 radiographs from Institution 2 (median age [IQR], 9 [5–14] years; male: female, 164:157) were used. The age distributions of the datasets are shown in Figure 2.

Internal Validation of the Korean Model

The RMSE and MAE of the predicted bone ages (in comparison with chronological ages) were 8.4 and 6.1 months, respectively. The proportions of participants with absolute differences ≤ 6, 12, 18, and 24 months were 60.7% (12770 of 21036), 87.5% (18408 of 21036), 96.3% (20266 of 21036), and 98.8% (20774 of 21036), respectively.

External Validation

The results of the concordance analysis between the chronological and bone ages predicted by the GP-based and Korean models are summarized in Table 1 and Figure 3. For the data from Institution 1, both the RMSE and MAE were significantly lower in the Korean model than in the GP-based model (RMSE, 11.2 vs. 13.8 months, $P = 0.004$; MAE, 8.2 vs. 10.5 months, $P = 0.002$). Additionally, the proportion of participants with an absolute difference ≤ 18 months was also higher in the Korean model (88.3% vs. 82.2%; $P = 0.031$). For Institution 2, MAE was significantly lower in the Korean model than in the GP-based model (9.5 vs. 11.0 months; $P = 0.022$), and the proportion of participants with an absolute difference ≤ 6 months was also higher in the Korean model (44.5% vs. 36.4%; $P = 0.044$).

The Bland–Altman results are summarized in Table 1 and Figure 3. The Korean model showed a trend of underestimating age as chronological age increased (Institution 1: slope, -0.067; $P < 0.001$; Institution 2: slope, -0.056; $P < 0.001$).

The results comparing the most probable bone age by the

GP model and the GP-based bone age most similar to the bone age predicted by the Korean model are presented in Table 2. However, both the RMSE and MAE in Institution 1 and the MAE in Institution 2 were significantly lower than those in the Korean model. Additionally, the proportions of participants with an absolute difference ≤ 12 and 18 months

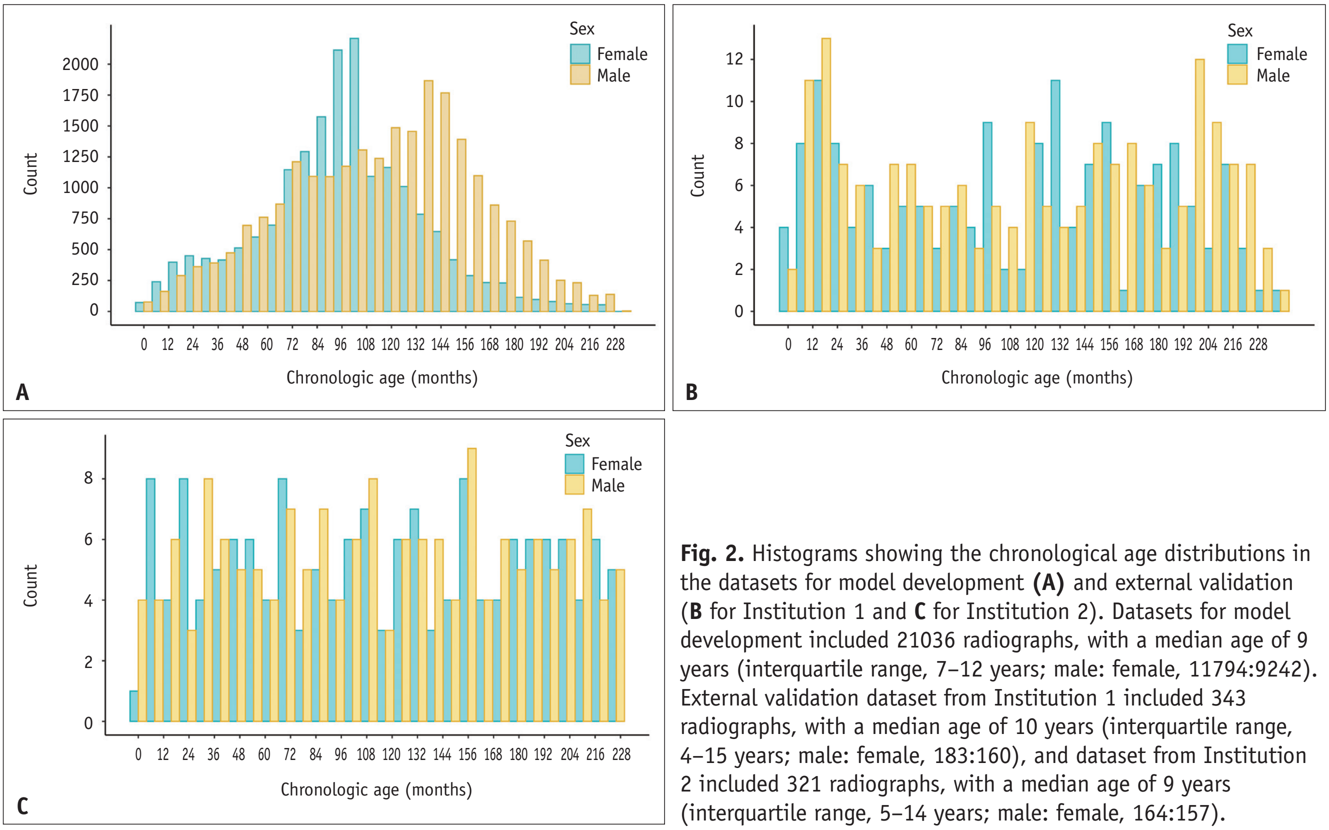

Fig. 2. Histograms showing the chronological age distributions in the datasets for model development **(A)** and external validation (**B** for Institution 1 and **C** for Institution 2). Datasets for model development included 21036 radiographs, with a median age of 9 years (interquartile range, 7–12 years; male: female, 11794:9242). External validation dataset from Institution 1 included 343 radiographs, with a median age of 10 years (interquartile range, 4–15 years; male: female, 183:160), and dataset from Institution 2 included 321 radiographs, with a median age of 9 years (interquartile range, 5–14 years; male: female, 164:157).

Table 1. Concordance and Bland-Altman analysis between chronological age and bone age predicted by the Greulich-Pyle-based model and Korean model

Parameters	Institution 1			Institution 2		
	GP	Korean	*P*	GP	Korean	*P*
RMSE, month	13.8	11.2	0.004	14.3	13.1	0.250
MAE, month	10.5	8.2	0.002	11.0	9.5	0.022
Percentage of subjects with absolute difference‡						
≤ 6 months	43.1 (148/343)	47.2 (162/343)	0.319	36.4 (117/321)	44.5 (143/321)	0.044
≤ 12 months	67.6 (232/343)	74.1 (254/343)	0.078	66.0 (212/321)	71.7 (230/321)	0.147
≤ 18 months	82.2 (282/343)	88.3 (303/343)	0.031	81.3 (261/321)	85.7 (275/321)	0.167
≤ 24 months	92.1 (316/343)	93.6 (321/343)	0.553	90.3 (290/321)	91.3 (293/321)	0.785
Bland-Altman parameters*						
Slope	0.016	-0.067	-	0.012	-0.056	-
Intercept, month	-1.8	7.0	-	-1.3	6.1	-
Bias, month	-0.02	-0.47	-	0.14	-0.24	-
Standard deviation, month	14.0	11.8	-	14.2	12.5	-
95% limits of agreement, month	-27.5 to 27.5	-23.5 to 22.6	-	-27.7 to 28.0	-24.7 to 24.2	-
P†	0.140	< 0.001	-	0.126	< 0.001	-

*Bland-Altman plot analysis was performed between chronologic age as the reference and the ages predicted by the GP-based (VUNO score) and Korean models, †*P*-value was calculated using univariable linear regression analysis based on the Bland-Altman plot, with the independent variable being the mean value of the chronologic age and predicted bone age, and the dependent variable being the difference between the chronologic age and predicted bone age, ‡Numbers in parentheses indicate numerators and denominators.
GP = Greulich-Pyle, RMSE = root mean square error, MAE = mean absolute error

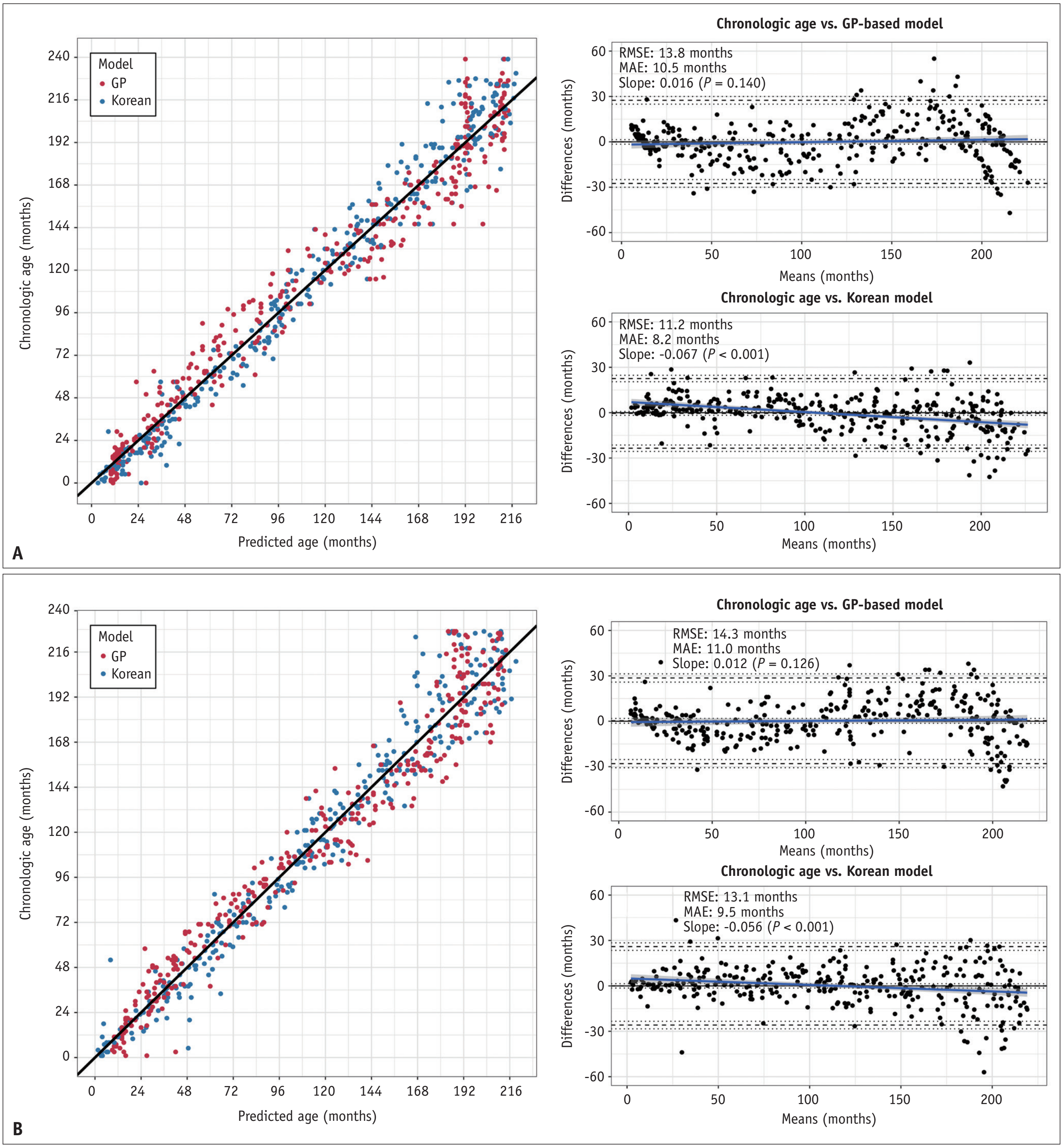

Fig. 3. Bivariate scatterplots showing associations between reference (chronological age) and bone ages predicted by the Greulich-Pyle (GP) (red dots) and Korean (blue dots) models, and Bland–Altman plots showing the difference between chronological and predicted bone ages in datasets from Institution 1 **(A)** and Institution 2 **(B)**. In the bivariate scatter plot, perfect concordance is represented by a 45° line (black line). In the Bland–Altman plot, the top and bottom dashed lines denote 1.96 standard deviations above and below the mean difference, respectively. The dotted lines represent 95% confidence intervals for these three values. The black line at 0 is the reference representing the situation with no bias (mean or slope) existing. The blue line represents the estimated bias from 0 with respect to age, with 95% confidence intervals (gray shaded area). Note that the root mean square error (RMSE) of the Korean model is significantly lower for Institution 1 ($P = 0.004$), and the mean absolute error (MAE) of the Korean model is significantly lower for both institutions (Institution 1, $P = 0.002$; Institution 2, $P = 0.022$).

https://doi.org/10.3348/kjr.2023.0092

in Institution 1 and those with an absolute difference ≤ 6 and 12 months in institution 2 were significantly higher than those in the Korean model.

Subgroup Analysis

The results of subgroup analyses restricted to ages 2–16 years for males and 2–14 years for females are summarized in Table 3 and Figure 4. The GP-based model tended to underestimate the bone age before the age of 8 years and overestimate the bone age after the age of 8 years (Institution 1: slope, 0.15; *P* < 0.001; Institution 2: slope, 0.15; *P* < 0.001). Contrastingly, the Korean model showed no significant age-dependent bias for Institution 1 (*P* = 0.266), and the slope in the Bland–Altman plot was significantly lower in the Korean model (slope; 0.15 [GP-based model] vs. –0.017 [Korean model]; *P* < 0.001). Although the Korean model showed a significant age-dependent bias for Institution 2 (*P* = 0.048), the slope in the Bland–Altman plot was significantly lower in the Korean model (slope; 0.15 [GP-based model] vs. 0.03 [Korean model]; *P* < 0.001). Additionally, in the Korean model, RMSE and MAE were lower, and the proportions of participants with absolute differences ≤ 6, 12, and 18 months were higher than those in the GP-based model.

DISCUSSION

In this study, we developed a deep-learning model for predicting bone age (termed the “Korean model”) in a contemporary healthy Korean pediatric population using chronological age of the participants as the label. Compared with the GP-based model, the Korean model showed better prediction performance. Superior accuracy was also observed even when comparing the most probable bone age and in the subgroup analysis restricted to ages of 2–16 years (boys) and 2–14 years (girls). Furthermore, the magnitude of age-dependent bias observed in the GP-based model was significantly reduced in the Korean model. Therefore, our newly developed Korean model appears to be a feasible method for assessing normal skeletal development in the Korean pediatric population.

Our application of the GP-based bone age prediction to approximately 20000 radiographs from a healthy Korean pediatric population confirmed the presence of a systemic bias in this population. In other words, skeletal maturation in contemporary Korean children and adolescents starts later and ends earlier than that in Caucasians. Zhang et al. [14] reported that bone age estimated using the GP method was more advanced in Asian boys (11–15 years) and girls

Table 2. Concordance and Bland-Altman analysis between chronologic age and the most probable/similar bone age by the Greulich-Pyle-based model and Korean model

Parameters	Institution 1			Institution 2		
	GP	Korean	*P*	GP	Korean	*P*
RMSE, month	14.0	11.7	0.014	14.4	13.5	0.270
MAE, month	10.6	8.6	0.005	11.1	9.8	0.017
Percentage of subjects with absolute difference[‡]						
≤ 6 months	41.4 (142/343)	47.5 (163/343)	0.124	35.8 (115/321)	48.3 (155/321)	0.002
≤ 12 months	65.3 (224/343)	74.3 (255/343)	0.013	62.3 (200/321)	72.3 (232/321)	0.009
≤ 18 months	80.2 (275/343)	87.5 (300/343)	0.013	81.9 (263/321)	85.4 (274/321)	0.286
≤ 24 months	91.8 (315/343)	95.0 (326/343)	0.123	90.3 (290/321)	91.9 (295/321)	0.579
Bland-Altman parameters*						
Slope	0.065	-0.075	-	-0.0007	-0.018	-
Intercept, month	-5.5	7.4	-	-0.0062	2.9	-
Bias, month	1.6	-1.4	-	-0.09	0.932	-
Standard deviation, month	13.7	12.8	-	13.6	14.6	-
95% limits of agreement, month	-25.3 to 28.4	-26.5 to 23.7	-	-26.8 to 26.7	-27.7 to 29.5	-
P[†]	< 0.001	< 0.001	-	0.960	0.138	-

*Bland-Altman plot analysis was performed between chronologic age as the reference and the ages predicted by the GP-based (the most probable GP-based bone age) and Korean models (GP-based bone age most similar to the predicted bone age by Korean model), [†]*P*-value was calculated using univariable linear regression analysis based on the Bland-Altman plot, with the independent variable being the mean value of the chronologic age and predicted bone age, and the dependent variable being the difference between the chronologic age and predicted bone age, [‡]Numbers in parentheses indicate numerators and denominators.
GP = Greulich-Pyle, RMSE = root mean square error, MAE = mean absolute error

Table 3. Results of subgroup analysis restricted to ages 2–16 years for males and 2–14 years for females

Parameters	Institution 1			Institution 2		
	GP	Korean	*P*	GP	Korean	*P*
RMSE, month	12.2	9.8	< 0.001	10.7	10.1	< 0.001
MAE, month	10.0	7.4	< 0.001	8.1	7.8	< 0.001
Percentage of subjects with absolute difference‡						
≤ 6 months	36.1 (75/208)	51.0 (106/208)	0.003	32.0 (70/219)	48.4 (106/219)	0.001
≤ 12 months	62.5 (130/208)	79.3 (165/208)	< 0.001	65.8 (144/219)	76.7 (168/219)	0.015
≤ 18 months	77.9 (162/208)	92.8 (193/208)	< 0.001	81.7 (179/219)	90.4 (198/219)	0.013
≤ 24 months	90.9 (189/208)	97.1 (202/208)	0.013	91.8 (201/219)	95.9 (210/219)	0.112
Bland-Altman parameters*						
Slope	0.15	-0.017	-	0.15	0.03	-
Intercept, month	-15	3	-	-14	-0.77	-
Bias, month	0.856	1.248	-	2.164	2.344	-
Standard deviation, month	15.137	9.804	-	13.849	10.477	-
95% limits of agreement, month	-28.8 to 30.5	-18.0 to 20.5	-	-25.0 to 29.3	-18.2 to 22.9	-
P†	< 0.001	0.266	-	< 0.001	0.048	-

*Bland-Altman plot analysis was performed between chronologic age as the reference and the ages predicted by the GP-based (VUNO score) and Korean models, †*P*-value was calculated using univariable linear regression analysis based on the Bland-Altman plot, with the independent variable being the mean value of the chronologic age and predicted bone age, and the dependent variable being the difference between the chronologic age and predicted bone age, ‡Numbers in parentheses indicate numerators and denominators. GP = Greulich-Pyle, RMSE = root mean square error, MAE = mean absolute error

(10–13 years) than that in Caucasians. Ontell et al. [15] also reported delayed bone age in the preadolescent period and advanced bone age in the adolescent period in Asian boys. This systemic bias was reproduced in our data (underestimation of bone age before the age of 8 years and overestimation of bone age after the age of 8 years). Considering the accuracy of deep-learning-based bone age assessment systems [9,29], this trend is probably not derived from the performance of the GP-based model but rather from different genetic factors, diet, and/or nutritional intake between the general Korean pediatric population and white pediatric population of an upper socioeconomic level (as was used to develop the GP method).

It should be emphasized that this systemic bias may affect treatment decisions. In general, delayed or advanced bone age is defined as the difference between bone age and chronological age > 2 SD of the mean [30]. This can be roughly interpreted as a difference between bone and chronological ages of greater than approximately 12 months between 2 and 4 years of chronological age, greater than 18 months between 4 and 12 years, and greater than 24 months after the age of 12 years [31]. Indeed, the proportions of participants with absolute differences ≤ 18 months in Institution 1 and those with absolute differences ≤ 6 months in Institution 2 were significantly higher in the Korean model than in the GP-based model. Furthermore, in the subgroup analysis restricted to participants aged 2–16 years for males and 2–14 years for females, the proportions of participants with absolute differences ≤ 6, 12, and 18 months were higher in the Korean model in both institutions. Considering the systemic bias in age prediction for the GP-based model, our deep-learning-based Korean model demonstrated the potential to minimize this age-dependent bias and might reduce inappropriate diagnosis and treatment. Although a statistically significant trend difference was noted in the Korean model when analyzing the whole study population (slope; Institution 1, -0.067; Institution 2, -0.056; both *P* < 0.001), this might be due to inevitable discrepancies between chronological and predicted bone ages after completion of bone development in the older age group; this would be supported by the fact that this systemic bias was minimized after exclusion of older adolescents.

This study had several limitations. First, the training dataset derived from a single institution might not reflect the general Korean pediatric population. Second, potentially eligible participants and radiographs were initially reviewed by a non-medical doctor researcher in a retrospective manner. Third, because the unavailable relevant information was considered normal (puberty, height, and growth rate), some included participants might have had undescribed endocrine problems. Fourth, the number of infants,

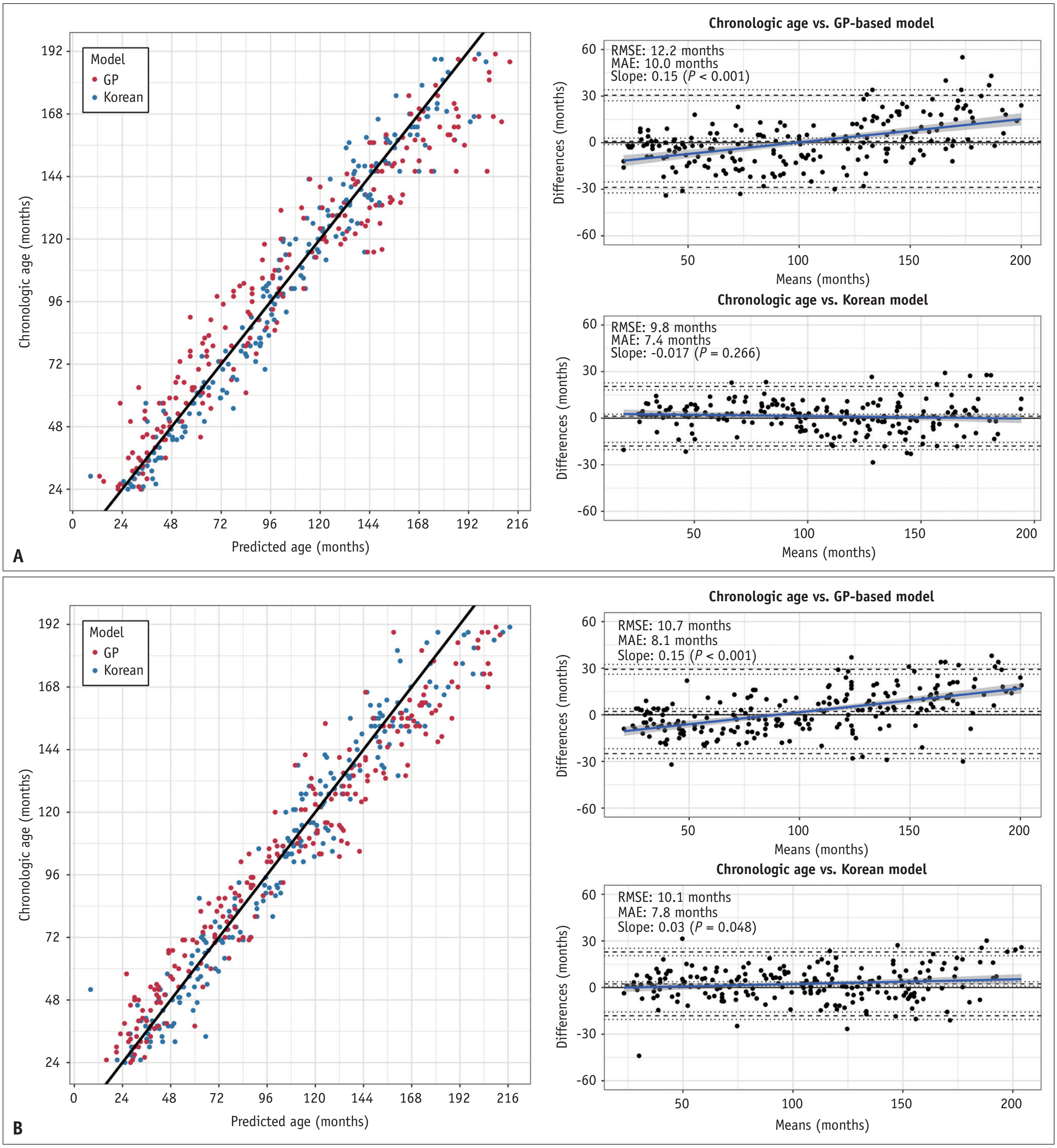

Fig. 4. Bivariate scatterplots showing associations between reference (chronological age) and bone ages predicted by the Greulich-Pyle (GP)-based (red dots) and Korean (blue dots) models for males aged 2–16 years and females aged 2–14 years. Bland–Altman plots show the difference between chronological and predicted bone ages in data from Institution 1 **(A)** and Institution 2 **(B)**. In the bivariate scatter plot, perfect concordance is represented by a 45° line (black line). In the Bland–Altman plot, the top and bottom dashed lines denote 1.96 standard deviations above and below the mean difference, respectively. The dotted lines represent 95% confidence intervals for these three values. The black line at 0 is the reference representing the situation with no bias (mean or slope) existing. The blue line represents the estimated bias from 0 with respect to age, with 95% confidence intervals (gray shaded area). Note that the magnitude of age-dependent bias (underestimation of bone age before the age of 8 years and overestimation of it after the age of 8 years) is reduced with the Korean model than that with the GP model for Institution 1 (slope, -0.017 [Korean] vs. 0.15 [GP model]; *P* < 0.001) and Institution 2 (slope, 0.03 [Korean] vs. 0.15 [GP model]; *P* < 0.001). RMSE = root mean square error, MAE = mean absolute error

toddlers, and older adolescents in the training set was relatively small compared with the number of participants aged approximately 10 years. Therefore, the model might not be optimized for these age groups. Indeed, for both the internal and external validation sets, the bone age predicted by the Korean model showed a wide dispersion from the chronological age. Therefore, it is necessary to collect additional training datasets containing more infants, toddlers, and older adolescents for future model modifications. Fifth, although chronological age was used as the label, skeletal maturation can vary within the same chronological age [32]. Sixth, the RSME of our model was larger than 6 months, which is generally considered the acceptable range for an accurate bone age assessment tool. This should be improved by further modifications. Lastly, we did not develop a bone age atlas based on the newly developed Korean model, which could simply be used as a reference in clinical practice.

In conclusion, a newly developed deep-learning-based Korean bone age assessment model trained using the chronological ages of Korean children and adolescents without known bone development-affecting diseases/conditions as labels showed better performance in bone age assessment than a GP-based model in the Korean pediatric population. Further validation is required to confirm the accuracy of this method.

Availability of Data and Material

The datasets generated or analyzed during the study are available from the corresponding author on reasonable request.

Conflicts of Interest

Jaewon Lee, Jinkyeong Sung, and Byeonguk Bae are employee of VUNO, and Kyu-Hwan Jung is shareholder of VUNO Inc., however this do not affect to publish this manuscript. All remaining authors have declared no conflicts of interest.

Author Contributions

Conceptualization: Hee Mang Yoon, Jin Seong Lee. Data curation: Pyeong Hwa Kim, Hee Mang Yoon, Jeong Rye Kim, Jae-Yeon Hwang, Boram Bak. Formal analysis: Pyeong Hwa Kim, Jaewon Lee. Funding acquisition: Hee Mang Yoon. Investigation: Pyeong Hwa Kim, Hee Mang Yoon. Methodology: Pyeong Hwa Kim, Hee Mang Yoon, Jin Seong Lee. Project administration: Hee Mang Yoon, Jin Seong Lee. Resources: Jaewon Lee, Jinkyeong Sung, Kyu-Hwan Jung, Byeonguk Bae, Ah Young Jung, Young Ah Cho, Woo Hyun Shim. Software: Jaewon Lee, Jinkyeong Sung, Kyu-Hwan Jung, Byeonguk Bae, Woo Hyun Shim. Supervision: Hee Mang Yoon, Jin Seong Lee, Jin-Ho Choi. Validation: Pyeong Hwa Kim, Hee Mang Yoon, Jisun Hwang, Jin Seong Lee. Visualization: Pyeong Hwa Kim, Jaewon Lee. Writing—original draft: Pyeong Hwa Kim. Writing—review & editing: all authors.

ORCID IDs

Pyeong Hwa Kim
https://orcid.org/0000-0003-4276-8803
Hee Mang Yoon
https://orcid.org/0000-0001-6491-5734
Jeong Rye Kim
https://orcid.org/0000-0001-6762-4174
Jae-Yeon Hwang
https://orcid.org/0000-0003-2777-3444
Jin-Ho Choi
https://orcid.org/0000-0003-1196-7826
Jisun Hwang
https://orcid.org/0000-0002-7593-2246
Jaewon Lee
https://orcid.org/0000-0002-8039-6222
Jinkyeong Sung
https://orcid.org/0000-0003-3546-6081
Kyu-Hwan Jung
https://orcid.org/0000-0002-6626-6800
Byeonguk Bae
https://orcid.org/0000-0003-2309-8517
Ah Young Jung
https://orcid.org/0000-0002-7427-6240
Young Ah Cho
https://orcid.org/0000-0001-6722-121X
Woo Hyun Shim
https://orcid.org/0000-0002-7251-2916
Boram Bak
https://orcid.org/0000-0003-0409-6292
Jin Seong Lee
https://orcid.org/0000-0002-8470-4595

Funding Statement

This study was supported by a grant (2020IE0016-1) from Asan Medical Center Children’s Hospital (Heart Institute), Seoul, Korea. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the

Korea government (MSIT) (2020R1G1A1004591). This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI22C1723). This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0026).

REFERENCES

1. Oh MS, Kim S, Lee J, Lee MS, Kim YJ, Kang KS. Factors associated with advanced bone age in overweight and obese children. *Pediatr Gastroenterol Hepatol Nutr* 2020;23:89-97
2. Kim D, Cho SY, Maeng SH, Yi ES, Jung YJ, Park SW, et al. Diagnosis and constitutional and laboratory features of Korean girls referred for precocious puberty. *Korean J Pediatr* 2012;55:481-486
3. Kelly PM, Diméglio A. Lower-limb growth: how predictable are predictions? *J Child Orthop* 2008;2:407-415
4. Creo AL, Schwenk WF 2nd. Bone age: a handy tool for pediatric providers. *Pediatrics* 2017;140:e20171486
5. Greulich WW, Pyle SI. *Radiographic atlas of skeletal development of the hand and wrist*. 2nd ed. Redwood City: Stanford University Press, 1999
6. Lee BD, Lee MS. Automated bone age assessment using artificial intelligence: the future of bone age assessment. *Korean J Radiol* 2021;22:792-800
7. Roche AF, Rohmann CG, French NY, Dávila GH. Effect of training on replicability of assessments of skeletal maturity (Greulich-Pyle). *Am J Roentgenol Radium Ther Nucl Med* 1970;108:511-515
8. Bull RK, Edwards PD, Kemp PM, Fry S, Hughes IA. Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods. *Arch Dis Child* 1999;81:172-173
9. Kim JR, Shim WH, Yoon HM, Hong SH, Lee JS, Cho YA, et al. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. *AJR Am J Roentgenol* 2017;209:1374-1380
10. 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;301:692-699
11. Booz C, Yel I, Wichmann JL, Boettger S, Al Kamali A, Albrecht MH, et al. Artificial intelligence in bone age assessment: accuracy and efficiency of a novel fully automated algorithm compared to the Greulich-Pyle method. *Eur Radiol Exp* 2020;4:6
12. Lee KC, Lee KH, Kang CH, Ahn KS, Chung LY, Lee JJ, et al. Clinical validation of a deep learning-based hybrid (Greulich-Pyle and modified Tanner-Whitehouse) method for bone age assessment. *Korean J Radiol* 2021;22:2017-2025
13. Alshamrani K, Messina F, Offiah AC. Is the Greulich and Pyle atlas applicable to all ethnicities? A systematic review and meta-analysis. *Eur Radiol* 2019;29:2910-2923
14. Zhang A, Sayre JW, Vachon L, Liu BJ, Huang HK. Racial differences in growth patterns of children assessed on the basis of bone age. *Radiology* 2009;250:228-235
15. Ontell FK, Ivanovic M, Ablin DS, Barlow TW. Bone age in children of diverse ethnicity. *AJR Am J Roentgenol* 1996;167:1395-1398
16. Hwang J, Yoon HM, Hwang JY, Kim PH, Bak B, Bae BU, et al. Re-assessment of applicability of Greulich and Pyle-based bone age to Korean children using manual and deep learning-based automated method. *Yonsei Med J* 2022;63:683-691
17. Yeon KM. Standard bone-age of infants and children in Korea. *J Korean Med Sci* 1997;12:9-16
18. King DG, Steventon DM, O'Sullivan MP, Cook AM, Hornsby VP, Jefferson IG, 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;67:848-851
19. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell* 2020;2:e200029
20. Nejedly N. Normal and abnormal growth in the pediatric patient. *Curr Probl Pediatr Adolesc Health Care* 2020;50:100771
21. Kim JH, Yun S, Hwang SS, Shim JO, Chae HW, Lee YJ, et al. The 2017 Korean National Growth Charts for children and adolescents: development, improvement, and prospects. *Korean J Pediatr* 2018;61:135-149
22. Iglovikov VI, Rakhlin A, Kalinin AA, Shvets AA. *Paediatric bone age assessment using deep convolutional neural networks*. In: Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T, Martel A, Maier-Hein L, et al., eds. *Deep learning in medical image analysis and multimodal learning for clinical decision support. DLMIA ML-CDS 2018. Lecture notes in computer science, vol 11045*. Cham: Springer, 2018:300-308
23. 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
24. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 2020;43:3349-3364
25. He K, Zhang X, Ren S, Sun J. *Deep residual learning for image recognition*. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, USA: IEEE; 2016:770-778
26. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. *ImageNet: a large-scale hierarchical image database*. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20-25; Miami, USA: IEEE; 2009:248-255

27. Gao BB, Zhou HY, Wu J, Geng X. *Age estimation using expectation of label distribution learning*. Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18); 2018 Jul 13-19; Stockholm, Sweden: IJCAI; 2018:712-718
28. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology* 2018;287:313-322
29. Lea WW, 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
30. Martin DD, Wit JM, Hochberg Z, Sävendahl L, van Rijn RR, Fricke O, et al. The use of bone age in clinical practice - part 1. *Horm Res Paediatr* 2011;76:1-9
31. Richmond EJ, Rogol AD. Diagnostic approach to children and adolescents with short stature [accessed on April 12, 2022]. Available at: https://www.uptodate.com/contents/diagnostic-approach-to-children-and-adolescents-with-short-stature
32. Gilsanz V, Ratib O. *Hand bone age: a digital atlas of skeletal maturity*. 1st ed. Berlin, Heidelberg: Springer-Verlag, 2005