Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Clinical Evaluation of an AI-Based System for Pediatric Growth Screening in Routine Practice: A Retrospective Cross-Sectional Study

Bhargavi Erravelli

kkfoundation2025@gmail.com

Rev & Dev Children Clinic

Lokesh Bandi

Mediciti Institute of Medical Sciences

Srujith CH

My Health Hospital

Prakash Raju Kodamala

Kodamalas Krushi Foundation

Kiran Kumar Ravulakollu

MNR University

Satyananda Siva Sagar Sambangi

MNR University

Devaraj Kuthadi

MNR Medical College

Research Article

Keywords: Pediatric growth screening, Artificial intelligence, Bone age assessment, Clinical imaging, Reliability, Screening tools, Deep learning, Greulich-Pyle method

Posted Date: April 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9190460/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Bone age assessment is fundamental to pediatric endocrinology and growth evaluation. Traditional manual methods using radiographic atlases suffer from inter-observer variability and time constraints. Artificial intelligence (AI) systems offer potential solutions for standardized, efficient bone age screening, though rigorous clinical validation in diverse populations remains essential.

Objective

To evaluate the clinical performance and accuracy of BoneAH, an AI-based bone age assessment system, as a screening tool in a cohort of Indian pediatric patients compared to reference determinations using the Greulich-Pyle method.

Methods

This retrospective cross-sectional observational study included 288 left hand-wrist radiographs from healthy pediatric patients aged 1 to 17 years. AI-predicted bone age was compared against consensus reference determinations by three blinded clinicians using the Greulich-Pyle atlas. Primary outcomes included mean absolute error (MAE), intraclass correlation coefficient (ICC), and Bland-Altman analysis. Secondary analyses examined performance across age groups and by gender.

Results

The AI system demonstrated high agreement with reference standards (ICC = 0.989; 95% CI: 0.986–0.991). Overall MAE was 0.58 years (95% CI: 0.53–0.63), with 83.3% of predictions within ± 1.0 year and 97.2% within ± 1.5 years of reference values. Pearson correlation was 0.993 ($p < 0.001$). A systematic positive bias of + 0.40 years was observed. Performance was comparable between males (MAE = 0.61 years) and females (MAE = 0.55 years; $p = 0.243$). Younger children (0–5 years) showed the lowest MAE (0.45 years).

Conclusions

BoneAH demonstrated high reliability and clinically acceptable accuracy for pediatric bone age screening in an Indian population. Its predictable nature supports potential calibration. The system shows promise as a first-level screening tool for growth assessment programs.

1. INTRODUCTION

Bone age assessment (BAA) is a necessary clinical tool in pediatric endocrinology, orthopedics and radiology, which serves as a critical indicator of skeletal maturity and biological development in children and adolescents. Unlike chronological age, which is determined solely by date of birth, bone age suggests the degree of skeletal maturation and provides valuable understandings related to growth potential, developmental disorders and overall physiological maturity [1, 2]. The assessment is particularly invaluable in the diagnostic evaluation of growth disorders, precocious or delayed puberty, endocrine abnormalities, prediction of adult height, and determining optimal timing for therapeutic interventions [3, 4].

The clinical applications of bone age assessment are diverse and clinically significant. In pediatric endocrinology, BAA plays a pivotal role in diagnosing and managing conditions such as growth hormone deficiency, constitutional delay of growth and puberty, idiopathic short stature, precocious puberty, congenital adrenal hyperplasia, and Turner syndrome [5, 6]. Advanced bone age is commonly observed in children with obesity, endocrine disorders causing excess hormone production, and precocious puberty, while delayed bone age is characteristic of growth hormone deficiency, hypothyroidism, constitutional growth delay, and chronic systemic illnesses [7, 8]. The accuracy of bone age determination is therefore paramount, as it directly influences clinical decision-making regarding treatment initiation, monitoring therapeutic response, and predicting final adult height [9, 10].

Traditionally, bone age assessment has relied on manual interpretation of left hand and wrist radiographs using standardized reference atlases. The Greulich-Pyle (GP) atlas method, published in 1959, remains the most widely utilized approach, employed by approximately 76% of pediatric endocrinologists and radiologists worldwide [11, 12]. This atlas-matching technique involves comparing a patient's hand-wrist radiograph with reference images representing specific chronological ages, offering simplicity and relative speed with assessment times averaging 1.4 minutes [13]. The Tanner-Whitehouse (TW2 and TW3) method represents an alternative approach based on detailed scoring of individual bone maturation stages across multiple skeletal elements, providing potentially greater objectivity at the cost of significantly increased assessment time, averaging 7.9 minutes per evaluation [13, 14].

Despite their widespread clinical adoption, conventional manual bone age assessment methods are fraught with inherent limitations that compromise their reliability and reproducibility. Inter-observer variability represents a significant challenge, with studies reporting average differences between radiologists of 0.69 years when using the GP method, with variations ranging from 0 to 1.95 years [15]. Intra-observer variability, though generally lower, remains substantial with 95% confidence intervals ranging from − 2.46 to 2.18 years for the GP method and − 1.41 to 1.43 years for TW2 [16]. This variability stems from multiple factors including differences in reader experience, subjective interpretation of atlas standards, difficulty in matching intermediate developmental stages, and potential bias introduced by knowledge of the patient's chronological age [15, 17]. Furthermore, the GP atlas, developed using data from Caucasian children in the 1930s-1950s, may not adequately represent contemporary populations or

diverse ethnic backgrounds, potentially introducing systematic bias when applied to non-Caucasian populations or modern cohorts with different growth patterns [18, 19].

The advent of artificial intelligence (AI) and deep learning technologies has ushered in a transformative era for automated bone age assessment, offering potential solutions to the limitations of traditional manual methods. Convolutional neural networks (CNNs) and other deep learning architectures have demonstrated remarkable capability in medical image analysis, achieving performance levels that rival or exceed human expert assessments [20, 21]. The Radiological Society of North America (RSNA) Pediatric Bone Age Machine Learning Challenge in 2017 demonstrated that top-performing AI algorithms could achieve mean absolute errors (MAE) as low as 4.3 months compared to reference standards, substantially outperforming the 7.3-month MAE typically observed with manual radiologist assessments [22, 23].

Several AI-based bone age assessment systems have been developed and commercialized, each demonstrating varying levels of clinical validation and adoption. BoneXpert, introduced in 2008, represents one of the first and most extensively validated automated systems [24]. Utilizing active appearance models and machine learning techniques, BoneXpert analyzes 13 bones and generates bone age estimates based on both GP and TW methodologies within approximately 15 seconds [24, 25]. Validation studies across diverse populations including Caucasian, Asian, Hispanic, and African children have reported MAE values ranging from 0.39 to 0.76 years, with strong correlations ($r > 0.98$) between automated and manual assessments [26–28]. The system has been particularly valuable in eliminating inter-observer variability and providing consistent, reproducible assessments across multiple clinical settings [29, 30].

VUNO Med-BoneAge, approved by the Korea Food and Drug Administration, represents another significant development in deep learning-based bone age assessment [31]. Trained on 18,940 hand radiographs analyzed using the GP method, this semi-automated system provides three ranked bone age predictions with associated probabilities and comparable reference images. The system demonstrates a first-rank accuracy of 69.5%, which increases to 93% when considering the top three predictions, with reported MAE values of approximately 4.9 months in validation studies [31, 32]. Other commercially available solutions including BoneView (Gleamer) and IB Lab PANDA have also emerged, each employing various deep learning architectures to automate the bone age assessment process [33].

Recent research has pushed the boundaries of AI-based bone age assessment even further. Advanced deep learning models utilizing state-of-the-art CNN architectures such as ResNet, InceptionV3, and VGG networks have achieved MAE values as low as 0.28 to 0.45 years on diverse datasets [34, 35]. Annotation-free pipelines that eliminate the need for manual bone region marking have been developed, incorporating attention mechanisms to automatically localize critical bone regions and integrate gender information as auxiliary inputs, thereby streamlining the clinical workflow while maintaining high accuracy [36]. Population-specific calibration approaches have addressed concerns about algorithmic

bias across different ethnic groups, demonstrating that locally calibrated AI models can achieve superior performance when tailored to specific demographic populations [37, 38].

Despite these technological advances, several critical challenges remain in the widespread clinical implementation of AI-based bone age assessment systems. External validation in real-world clinical settings beyond the controlled environments in which these algorithms were originally trained remains essential for establishing generalizability and reliability [39, 40]. Population-specific variations in skeletal maturation patterns necessitate careful calibration of AI models for diverse ethnic and geographic populations to avoid systematic bias [18, 37]. The "black box" nature of many deep learning models raises concerns about interpretability and clinical trust, although recent developments in explainable AI and attention visualization techniques are beginning to address these limitations [36, 41]. Furthermore, the integration of AI systems into existing clinical workflows, regulatory approval processes, and considerations of cost-effectiveness represent practical barriers to widespread adoption [42].

The imperative for rigorous validation of AI-based bone age assessment tools in diverse clinical populations cannot be overstated. While many algorithms demonstrate impressive performance on standardized benchmark datasets, their real-world clinical utility must be established through independent external validation studies across varied patient populations, imaging equipment, and clinical conditions. Such validation efforts are essential for building clinical confidence, identifying potential failure modes or edge cases, understanding performance variations across demographic subgroups, and establishing appropriate clinical use cases and limitations.

In this context, the present study aims to evaluate the clinical performance and accuracy of BoneAH, an AI-powered bone age assessment tool, in a cohort of 288 pediatric patients from a South Indian population. By comparing BoneAH predictions against reference bone age determinations using the Greulich-Pyle method, this validation study seeks to assess the system's accuracy across different age groups, examine potential systematic biases, and evaluate its suitability for clinical implementation as a screening tool. The findings will contribute to the growing body of evidence regarding AI-assisted bone age assessment and provide insights into the practical application of automated systems in pediatric radiology and endocrinology practice.

2. MATERIALS AND METHODS

2.1 Study Design

This was a retrospective, cross-sectional observational study designed to evaluate the accuracy and reliability of an AI-based bone age assessment system compared to expert radiological interpretation using the Greulich-Pyle atlas method. The study was conducted at a Medical College, in India. The study protocol adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist for observational studies.

2.2 Study Population and Selection Criteria

The study population comprised pediatric patients aged 1 to 17 years who underwent left hand-wrist radiography at a Medical College between August 2025 and January 2026. The study included consecutive, healthy children without known chronic health issues who presented for routine bone age determination.

Inclusion Criteria

(1) Age between 1 and 17 years at the time of imaging; (2) Left hand posteroanterior (PA) radiograph available for analysis; (3) No known chronic systemic illnesses; (4) No documented endocrine disorders; (5) No current growth hormone therapy or other medications affecting growth; (6) Sufficient image quality for both manual and automated assessment.

Exclusion Criteria

(1) Known skeletal dysplasias; (2) Previous hand or wrist fractures; (3) Presence of bone tumors or lesions; (4) Poor image quality precluding reliable assessment (motion artifact, positioning errors, inadequate exposure); (5) Known genetic syndromes affecting growth; (6) Incomplete clinical documentation.

A total of 306 pediatric radiographic images were initially collected. Eighteen images (5.9%) were excluded due to inadequate image quality. The final analysis included 288 images from unique patients. This sample size exceeded recommendations from prior pediatric imaging and reliability studies, which indicate that a minimum of 100–200 cases is sufficient to achieve stable estimates of agreement metrics with acceptable precision.

2.3 Image Acquisition and Preprocessing

All radiographs were acquired using digital radiography (DR) equipment with standardized acquisition parameters: 50–60 kVp, 2 mAs, focal spot size 5-100 µm (0.005-0.1 mm). Images were obtained in the posteroanterior (PA) projection of the left hand following standard positioning protocols. Image acquisition time was under 10 seconds per study. All images were acquired at 300 dpi resolution and stored in PNG format for analysis. Quality assessment focused on appropriate positioning and adequate exposure for skeletal visualization.

2.4 AI System Description

BoneAH (version 2.6.2; Kodamalas Krushi Foundation, India) is a proprietary AI-powered bone age assessment system accessible via web interface (https://boneah.com/). The system accepts PNG image format inputs and performs automated region-of-interest (ROI) selection without manual intervention. Processing time per image was approximately 10 seconds. The system is currently undergoing regulatory approval through the Central Drugs Standard Control Organisation (CDSCO), India.

Technical specifications regarding the underlying algorithm architecture and training dataset are proprietary and not disclosed by the developer.

2.5 Reference Standard and Reader Methodology

The reference standard for bone age determination was established using the Greulich-Pyle atlas method (2nd edition, 1959) [1]. Three independent readers performed blinded assessments: (1) a radiologist from the Department of Radiology and Imaging Technology, a Medical College; (2) an orthopedic surgeon from another Medical College; and (3) a consultant physician and diabetologist. All readers were blinded to the AI predictions, patient chronological age, and other clinical information at the time of assessment.

Inter-reader agreement was evaluated prior to establishing consensus values. Disagreements between readers were minimal, and final reference bone age values were established through consensus discussion. The consensus bone age served as the ground truth for all subsequent comparative analyses.

2.6 Outcome Measures

Primary Outcomes

(1) Mean Absolute Error (MAE) with 95% confidence intervals (CI) representing the average magnitude of prediction errors; (2) Intraclass Correlation Coefficient (ICC, two-way random effects, absolute agreement, single measures; ICC(2,1)) to assess reliability between AI predictions and reference standard; (3) Proportion of predictions within clinically relevant error thresholds (± 0.5, ± 1.0, and ± 1.5 years).

Secondary Outcomes

(1) Pearson correlation coefficient (r) for linear association; (2) Spearman rank correlation coefficient (ρ) for monotonic relationship; (3) Root Mean Square Error (RMSE) to capture error magnitude including outlier sensitivity; (4) Bland-Altman analysis including mean bias, standard deviation of differences, and 95% limits of agreement (LoA); (5) Statistical testing for systematic bias using paired t-test; (6) Subgroup analyses by age group and gender.

2.7 Statistical Analysis

Descriptive statistics were presented as mean ± standard deviation (SD) for continuous variables and frequencies with percentages for categorical variables. The MAE was calculated as the mean of absolute differences between AI predictions and reference bone age values. The 95% CI for MAE was computed using bootstrap resampling (1000 iterations). ICC was calculated using a two-way random effects model for absolute agreement of single measures. Pearson and Spearman correlation coefficients assessed linear and monotonic relationships, respectively.

Bland-Altman analysis was performed by plotting the difference between AI prediction and reference standard against the mean of both measurements. Mean bias and 95% LoA (mean bias ± 1.96 × SD) were calculated. A one-sample t-test assessed whether the mean bias differed significantly from zero. Gender differences in prediction error were evaluated using independent samples t-test. Age group differences were assessed using Kruskal-Wallis test followed by post-hoc pairwise comparisons where appropriate. Linear regression analysis examined the relationship between prediction error and demographic variables (age, gender). Statistical significance was set at $\alpha = 0.05$ (two-tailed). All analyses were performed using Python (version 3.9) with SciPy, statsmodels, and pingouin packages.

2.8 Ethical Considerations

This study was approved by the Institutional Ethics Committee (Reference No: MNRU/IEC/2025/01). Given the retrospective nature of the study involving analysis of previously acquired clinical images, the requirement for informed consent was waived by the Ethics Committee, as the study posed minimal risk to participants and did not involve direct patient interaction. All patient data were fully de-identified prior to analysis; identifiers including patient names, hospital identification numbers, dates of birth, and acquisition metadata were removed or anonymized in compliance with data protection standards. Each image was assigned a unique study-specific code to ensure confidentiality. De-identified data were stored on secure, access-restricted systems with access limited to authorized study personnel only. All data handling and storage procedures complied with institutional data governance policies and applicable data protection regulations.

3. RESULTS

3.1 Dataset Characteristics

The final study cohort comprised 288 pediatric patients, including 162 males (56.3%) and 126 females (43.8%). All participants were of South Indian (Asian Indian) ethnicity from the Hyderabad region. The age distribution ranged from 1 to 17 years, with the following breakdown by age groups: 0–5 years (n = 54, 18.8%), 6–10 years (n = 57, 19.8%), 11–15 years (n = 110, 38.2%), and 16 + years (n = 67, 23.3%). Table 1 presents the detailed demographic characteristics of the study population.

Table 1
Demographic and Clinical Characteristics of the Study Population (N = 288)

Characteristic	Value
Total sample size	288
Age range (years)	1–17
Gender - Male, n (%)	162 (56.3%)
Gender - Female, n (%)	126 (43.8%)
Age group 0–5 years, n (%)	54 (18.8%)
Age group 6–10 years, n (%)	57 (19.8%)
Age group 11–15 years, n (%)	110 (38.2%)
Age group 16 + years, n (%)	67 (23.3%)
Ethnicity	South Indian (Asian Indian)
Geographic region	Hyderabad, Telangana, India

3.2 Primary Outcomes

The AI system demonstrated high agreement with the reference standard across all primary outcome measures. The overall MAE was 0.58 years (95% CI: 0.53–0.63 years), equivalent to approximately 7 months. The median absolute error was 0.50 years. The ICC (2,1) was 0.989 (95% CI: 0.986–0.991), indicating near-perfect reliability between AI predictions and reference bone age values. Table 2 presents the comprehensive performance metrics.

Clinical accuracy thresholds analysis revealed that 52.8% of predictions fell within ± 0.5 years, 83.3% within ± 1.0 year, and 97.2% within ± 1.5 years of the reference bone age values. Only 8 predictions (2.8%) demonstrated errors exceeding 1.5 years. Figure 1 illustrates the scatter plot of AI-predicted versus reference bone age values, demonstrating strong linear agreement across the entire age range with minimal deviation from the line of identity.

Table 2
Overall Performance Metrics of the AI System

Metric	Value (95% CI)
Mean Absolute Error (MAE), years	0.58 (0.53–0.63)
Median Absolute Error, years	0.50
Root Mean Square Error (RMSE), years	0.73
Intraclass Correlation Coefficient (ICC)	0.989 (0.986–0.991)
Pearson correlation coefficient (r)	0.993 ($p < 0.001$)
Spearman correlation coefficient (ρ)	0.985 ($p < 0.001$)
Predictions within ± 0.5 years, %	52.8%
Predictions within ± 1.0 year, %	83.3%
Predictions within ± 1.5 years, %	97.2%

3.3 Agreement and Bias Analysis

Bland-Altman analysis revealed a mean systematic bias of + 0.40 years (SD = 0.60 years), indicating that the AI system, on average, overestimated bone age relative to the reference standard. This bias was statistically significant (one-sample t-test, $p < 0.0001$). The 95% limits of agreement ranged from − 0.78 years to + 1.59 years (Fig. 2). The asymmetry in limits of agreement, with a wider range on the overestimation side, reflects the positive systematic bias (Table 3).

Figure 3 displays the distribution of prediction errors, showing both signed and absolute error histograms. The signed error distribution demonstrates the positive shift corresponding to the systematic overestimation, while the absolute error distribution confirms the majority of errors cluster below 1.0 year.

Table 3
Bland-Altman Analysis Results

Parameter	Value
Mean bias (Predicted - Reference), years	+ 0.40
Standard deviation of differences, years	0.60
Upper limit of agreement (+ 1.96 SD), years	+ 1.59
Lower limit of agreement (-1.96 SD), years	-0.78
Bias significance (t-test p-value)	< 0.0001

The residual plot (Fig. 4) demonstrates that the prediction error pattern remains relatively consistent across the entire bone age range, with no substantial proportional bias. The smoothed trend line confirms the stable positive bias across all age ranges.

3.4 Subgroup Analysis by Age Group

Performance metrics varied across age groups (Table 4; Figs. 5 and 6). The youngest age group (0–5 years) demonstrated the lowest MAE at 0.45 years, while the 11–15 years group showed the highest MAE at 0.63 years. The Kruskal-Wallis test indicated a statistically significant difference in absolute error distribution across age groups ($p = 0.030$). Despite this variation, all age groups maintained MAE values below 0.7 years and achieved greater than 79% of predictions within ± 1.0 year.

Table 4
Performance Metrics by Age Group

Age Group	n	MAE (years)	RMSE (years)	Within ± 1 year (%)
0–5 years	54	0.45	0.61	91%
6–10 years	57	0.61	0.76	88%
11–15 years	110	0.63	0.80	80%
16 + years	67	0.58	0.75	79%
p-value (Kruskal-Wallis)	-	0.030	-	-

3.5 Subgroup Analysis by Gender

Gender-stratified analysis (Table 5; Fig. 6) revealed comparable performance between males and females. Males (n = 162) demonstrated a MAE of 0.61 years (SD = 0.46) compared to 0.55 years (SD = 0.40) in females (n = 126). This difference was not statistically significant (independent samples t-test, $p = 0.243$). RMSE values were 0.76 years for males and 0.68 years for females. The proportion of predictions within ± 1.0 year was 78% for males and 90% for females. Linear regression analysis confirmed that prediction error was not systematically associated with gender ($R^2 = 0.018$), indicating that the AI system's performance was not meaningfully influenced by patient sex.

Table 5
Performance Metrics by Gender

Metric	Male (n = 162)	Female (n = 126)	p-value
MAE (years)	0.61	0.55	0.243
SD (years)	0.46	0.40	-
RMSE (years)	0.76	0.68	-
Within ± 1 year (%)	78%	90%	-

3.6 Clinical Threshold Analysis

The cumulative distribution of absolute errors (Figs. 7 and 8) provides insight into the clinical utility of the AI system across different error tolerance thresholds. The percentage of predictions within clinical error thresholds varied across subgroups, with the youngest age group (0–5 years) achieving the highest proportion within ± 1.0 year (91%) and females showing higher accuracy (90%) compared to males (78%) at the same threshold. Figure 9 displays the agreement plot with error magnitude visualization, confirming high agreement across the measurement range with most errors in the acceptable range.

4. DISCUSSION

This study evaluated the clinical performance of BoneAH, an AI-based bone age assessment system, in a cohort of South Indian pediatric patients. Our findings demonstrate that the system achieves high reliability (ICC = 0.989) and clinically acceptable accuracy (MAE = 0.58 years) when compared to expert consensus using the Greulich-Pyle method. These results support the potential utility of BoneAH as a screening tool in pediatric growth assessment, while also identifying systematic bias that warrants consideration in clinical implementation.

4.1 Interpretation of Performance Metrics

The observed MAE of 0.58 years (approximately 7 months) positions BoneAH competitively among validated AI bone age assessment systems. Published validation studies of BoneXpert, perhaps the most extensively studied automated system, report MAE values ranging from 0.39 to 0.76 years across different populations [26–28], placing our results within this established performance range. The RSNA Pediatric Bone Age Challenge demonstrated that top-performing algorithms achieve MAE values of approximately 0.36 years (4.3 months), while average manual radiologist assessment yields MAE around 0.61 years (7.3 months) [22, 23]. Our findings suggest that BoneAH achieves accuracy comparable to or better than typical manual assessment, which is an important benchmark for clinical screening applications.

The exceptionally high ICC (0.989) indicates near-perfect reliability, a critical attribute for screening tools where consistency across assessments is paramount. This metric surpasses the commonly accepted threshold of 0.75 for excellent agreement and approaches the upper theoretical limit. The strong correlation coefficients (Pearson r = 0.993; Spearman ρ = 0.985) further confirm that the AI system accurately preserves the rank order and linear relationship of bone age across the pediatric age spectrum.

4.2 Clinical Relevance of Observed Error Margins

The clinical significance of prediction error must be interpreted within the context of pediatric growth assessment practice. In routine clinical scenarios, bone age differences of up to ± 1 year from chronological age are generally considered within normal variation, with discrepancies exceeding 2

standard deviations (approximately 2 years) typically triggering further investigation. Our finding that 83.3% of predictions fell within ± 1.0 year and 97.2% within ± 1.5 years suggests that the vast majority of AI assessments would support appropriate clinical decision-making in a screening context.

Importantly, the intended clinical application influences the acceptable error threshold. For first-level population screening, such as school health programs or community growth surveillance, higher error tolerance is acceptable compared to diagnostic contexts requiring precise bone age determination for treatment decisions. The performance profile of BoneAH, characterized by high reliability with occasional larger errors (as reflected by RMSE exceeding MAE), is well-suited for screening applications where false negatives (missed abnormalities) carry greater consequence than false positives, which can be addressed through secondary expert review.

4.3 Systematic Bias: Interpretation and Implications

The statistically significant positive bias (+ 0.40 years) observed in this study merits careful consideration. While the presence of systematic error may initially appear concerning, several factors contextualize this finding. First, the magnitude of bias remains within clinically acceptable limits for screening applications—a consistent 5-month overestimation, while detectable statistically, would rarely alter clinical categorization of skeletal maturity. Second, systematic bias, by definition, represents a predictable measurement shift rather than random error, making it amenable to post-hoc calibration if deemed necessary for specific clinical applications.

The origin of systematic bias in AI bone age systems often reflects population differences between training and validation cohorts. The Greulich-Pyle atlas was developed using Caucasian American children from the 1930s-1950s, and published evidence suggests that contemporary children and those from Asian populations may demonstrate accelerated skeletal maturation relative to these historical standards [18, 19]. If the AI system was trained primarily on data labeled using GP reference standards without population-specific adjustment, it would be expected to inherit any systematic offset present in the reference methodology. Notably, other validated AI systems have demonstrated similar patterns of systematic bias, which can be effectively addressed through population-specific calibration coefficients [37, 38].

4.4 Age-Stratified Performance and Clinical Implications

The observation of superior performance in younger children (MAE = 0.45 years in 0–5 years group) compared to older age groups (MAE = 0.63 years in 11–15 years group) aligns with biological and radiological expectations. Younger children demonstrate more discrete, easily distinguishable skeletal maturation stages, whereas pubertal-age children exhibit greater variability in timing and progression of secondary ossification center development. This pattern has been consistently reported across multiple AI bone age systems and reflects inherent challenges in assessing skeletal maturity during periods of rapid pubertal development.

From a public health perspective, the finding of lowest error in youngest children is clinically advantageous, as early childhood represents a critical window for detecting growth abnormalities amenable to intervention. Conditions such as growth hormone deficiency, hypothyroidism, and constitutional delay are optimally diagnosed and treated in early childhood, where intervention can maximize height potential and minimize psychosocial sequelae. The system's strongest performance in this age range supports its utility in early detection screening programs.

4.5 Comparison with Published Literature

Our results are consistent with published validation studies of AI bone age systems across diverse populations. A recent comprehensive validation of BoneXpert in Czech children (n = 3,398) reported MAE values of 0.45–0.47 years with ICC values exceeding 0.98 [28]. Studies of VUNO Med-BoneAge in Korean populations reported MAE of approximately 0.41 years (4.9 months) [31, 32]. A recent Portuguese validation study of another AI system reported MAE of 0.46 years with similar patterns of systematic bias [39]. Comparative studies between automated and manual methods consistently demonstrate that AI systems achieve accuracy comparable to or better than single-reader manual assessment while eliminating inter-observer variability [29, 30].

4.6 Implications for Large-Scale Screening Programs

The performance characteristics of BoneAH support its potential integration into population-level pediatric growth screening programs, particularly in resource-limited settings. The system's key advantages for screening applications include rapid processing time (approximately 10 seconds per image), elimination of inter-observer variability, scalability for large-volume assessment, and consistent performance across the pediatric age range. In settings where access to specialized pediatric radiologists may be limited, AI-assisted screening could facilitate earlier identification of children requiring specialist evaluation.

Implementation in school health programs or community screening initiatives would require consideration of several operational factors: availability of radiographic equipment and appropriate radiation safety protocols, defined referral pathways for children with abnormal bone age assessments, clinician training on interpretation of AI results, and quality assurance mechanisms to monitor ongoing system performance. The absence of statistically significant gender differences in our study ($p = 0.243$) supports equitable deployment across mixed-gender pediatric populations without requirement for sex-specific adjustment.

4.7 Ethical and Operational Considerations

Several ethical considerations merit attention in clinical deployment of AI bone age assessment. Transparency regarding system limitations, including the presence of systematic bias and age-related performance variation, is essential for informed clinical decision-making. Clear communication to patients and families that AI assessment serves a screening function requiring clinical interpretation, rather than definitive diagnosis, helps establish appropriate expectations. The proprietary nature of the

algorithm, while common among commercial AI medical devices, limits full assessment of potential failure modes or hidden biases—a consideration for regulatory evaluation and clinical governance [42].

5. LIMITATIONS

This study has several limitations that should be considered when interpreting the findings. First, the retrospective single-cohort design limits generalizability; while internal validity is supported by blinded assessment and rigorous methodology, external validation across different institutions, imaging equipment, and populations is required to establish broader applicability. Second, a systematic positive bias was observed, which, although within clinically acceptable limits for screening applications and amenable to calibration, represents a deviation from the reference standard that clinicians should acknowledge. Third, wider limits of agreement were noted on the overestimation side, reflecting occasional larger errors that, while infrequent, may have clinical significance in individual cases. Fourth, performance varied modestly across age groups, with superior accuracy in younger children; while clinically advantageous for early detection, this pattern warrants awareness when interpreting results in pubertal-age patients. Fifth, the absence of an independent external validation cohort means that the reported performance estimates may represent optimistic bounds that could diminish in truly independent datasets. Sixth, the proprietary nature of the AI algorithm limits mechanistic understanding of prediction behavior and potential failure modes. Finally, the single-ethnicity population (South Indian) may limit applicability to other ethnic groups, given known population differences in skeletal maturation patterns.

It should be explicitly noted that prospective clinical validation using newly acquired imaging data is planned as part of future work to address several of these limitations.

6. FUTURE DIRECTIONS

Several priorities for future research emerge from this validation study. Prospective clinical validation with temporally separated data collection is planned to confirm generalizability and assess real-world performance under routine clinical conditions. Multi-center studies incorporating diverse geographic populations across India and internationally would strengthen evidence for widespread deployment. Development and validation of population-specific calibration coefficients could potentially eliminate the observed systematic bias, thereby improving absolute accuracy while preserving the system's excellent reliability characteristics.

Integration studies examining workflow efficiency, clinician acceptance, and cost-effectiveness of AI-assisted bone age assessment in various clinical settings (tertiary hospitals, community clinics, school health programs) would inform optimal implementation strategies. Longitudinal studies tracking clinical outcomes in children screened with AI assistance compared to conventional assessment would provide evidence on downstream effects of early detection. Finally, investigation of system performance in

clinical populations with known endocrine or growth disorders, who were excluded from this healthy-cohort validation, would establish performance boundaries in diagnostically challenging cases.

7. CONCLUSION

This retrospective validation study demonstrates that BoneAH, an AI-based bone age assessment system, achieves high reliability (ICC = 0.989) and clinically acceptable accuracy (MAE = 0.58 years) when evaluated against expert consensus using the Greulich-Pyle method in a South Indian pediatric population. The system demonstrated consistent performance across both genders and maintained acceptable accuracy across the pediatric age range, with superior performance in younger children where early detection is most clinically impactful.

While systematic positive bias was identified, its magnitude remains within clinically acceptable limits for screening applications and can be addressed through calibration. The findings support the potential utility of BoneAH as a screening tool—rather than a diagnostic replacement—for pediatric growth assessment, particularly in settings where access to specialist interpretation may be limited. The system shows promise for integration into large-scale pediatric screening programs aimed at early detection of growth abnormalities.

Prospective multi-center validation is warranted before widespread clinical implementation to confirm generalizability across diverse populations and clinical settings. With appropriate validation and operational safeguards, AI-assisted bone age screening holds potential to enhance access to standardized growth assessment and support early intervention for pediatric growth disorders. The study shows promise as a first-level screening tool for growth assessment programs.

Declarations

8. ETHICS STATEMENT

This study was approved by the Institutional Ethics Committee of (Reference No: MNRU/IEC/2025/01). The study was conducted in accordance with the Declaration of Helsinki and applicable institutional guidelines. Given the retrospective nature of the study involving analysis of previously acquired, de-identified clinical images, the requirement for individual informed consent was waived by the Ethics Committee. All patient data were fully anonymized prior to analysis, with removal of all direct and indirect identifiers. De-identified data were stored on secure, access-restricted systems with access limited to authorized study personnel. The study posed minimal risk to participants and did not involve direct patient interaction or intervention.

9. CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

10. FUNDING

There is no funding support for the study.

References

1. Greulich WW, Pyle SI. Radiographic Atlas of Skeletal Development of the Hand and Wrist. 2nd ed. Stanford: Stanford University Press; 1959.
2. Satoh M. Bone age: assessment methods and clinical applications. Clin Pediatr Endocrinol. 2015;24(4):143-152.
3. Martin DD, Wit JM, Hochberg Z, et al. The use of bone age in clinical practice - part 1. Horm Res Paediatr. 2011;76(1):1-9.
4. Cavallo F, Mohn A, Chiarelli F, Giannini C. Evaluation of bone age in children: a mini-review. Front Pediatr. 2021;9:580314.
5. Balducci R, Toscano V. Bone age assessment in the workup of children with endocrine disorders. J Endocrinol Invest. 2010;33(3):168-173.
6. Cohen P, Rogol AD, Deal CL, et al. Consensus statement on the diagnosis and treatment of children with idiopathic short stature. J Clin Endocrinol Metab. 2008;93(11):4210-4217.
7. Sopher AB, Thornton JC, Silfen ME, et al. Bone age advancement in prepubertal children with obesity and premature adrenarche. Obesity. 2011;19(6):1259-1264.
8. Weise M, De-Levi S, Barnes KM, et al. Effects of estrogen on growth plate senescence and epiphyseal fusion. Proc Natl Acad Sci USA. 2001;98(12):6871-6876.
9. Deodati A, Cianfarani S. Impact of growth hormone therapy on adult height of children with idiopathic short stature: systematic review. BMJ. 2011;342:c7157.
10. Albanese A, Stanhope R. Predictive factors in the determination of final height in boys with constitutional delay of growth and puberty. J Pediatr. 1995;126(4):545-550.
11. Gaskin CM, Kahn SL, Bertozzi JC, Bunch PM. Skeletal Development of the Hand and Wrist: A Radiographic Atlas and Digital Bone Age Companion. Oxford: Oxford University Press; 2011.
12. Ontell FK, Ivanovic M, Ablin DS, Barlow TW. Bone age in children of diverse ethnicity. AJR Am J Roentgenol. 1996;167(6):1395-1398.
13. Bull RK, Edwards PD, Kemp PM, et al. Bone age assessment: a large scale comparison of the Greulich and Pyle and Tanner and Whitehouse methods. Arch Dis Child. 1999;81(2):172-173.
14. Tanner JM, Whitehouse RH, Cameron N, et al. Assessment of Skeletal Maturity and Prediction of Adult Height (TW2 Method). 2nd ed. London: Academic Press; 1983.
15. 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.

16. Berst MJ, Dolan L, Bogdanowicz MM, et al. Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards. AJR Am J Roentgenol. 2001;176(2):507-510.
17. Lynnerup N, Belard E, Buch-Olsen K, et al. Intra- and inter-observer error of the Greulich-Pyle method as used on a Danish forensic sample. Forensic Sci Int. 2008;179(2-3):242.e1-242.e6.
18. Mora S, Boechat MI, Pietka E, et al. Skeletal age determinations in children of European and African descent: applicability of the Greulich and Pyle standards. Pediatr Res. 2001;50(5):624-628.
19. 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(1):228-235.
20. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.
21. 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.
22. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology. 2019;290(2):498-503.
23. 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.
24. 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.
25. van Rijn RR, Thodberg HH. Bone age assessment: automated techniques coming of age? Acta Radiol. 2013;54(9):1024-1029.
26. Thodberg HH, Sävendahl L. Validation and reference values of automated bone age determination for four ethnicities. Acad Radiol. 2010;17(11):1425-1432.
27. Martin DD, Sato K, Sato M, Thodberg HH, Tanaka T. Validation of a new method for automated determination of bone age in Japanese children. Horm Res Paediatr. 2010;73(5):398-404.
28. Maratova K, Zapletalova J, Zemkova D, et al. A comprehensive validation study of the latest version of BoneXpert on a large cohort of Caucasian children and adolescents. Front Endocrinol. 2023;14:1130580.
29. Booz C, Wichmann JL, Boettger S, et al. Evaluation of a computer-aided diagnosis system for automated bone age assessment in comparison to the Greulich-Pyle atlas method. J Comput Assist Tomogr. 2019;43(1):39-45.
30. Larson N, Mahomed N, van Wyk N. Comparison of bone age assessment using manual Greulich and Pyle method versus automated BoneXpert method in South African children. S Afr J Radiol. 2024;28(1):2794.
31. Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging. 2017;30(4):427-441.

32. 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.
33. Báez-Suárez A, Martín-González JM, García-Hernández C, Palacios-Navarro G. Artificial intelligence-based models for automated bone age assessment from posteroanterior wrist X-rays: a systematic review. Appl Sci. 2025;15(11):5978.
34. Kasani PH, Kasani S, Kim JY, Jang R, Oh SL. Bone age assessment from hand radiographs using divide-and-conquer based lightweight CNN architecture. Comput Biol Med. 2023;157:106734.
35. Ren X, Li T, Yang X, et al. Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. IEEE J Biomed Health Inform. 2019;23(5):2030-2038.
36. 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.
37. Gordeladze M, Bjørk MH, Grønnesby M, et al. Population-specific calibration and validation of an open-source bone age AI. Sci Rep. 2025;15(1):1234.
38. Özmen E, Özen Atalay H, Uzer E, Veznikli M. A comparison of two artificial intelligence-based methods for assessing bone age in Turkish children. Diagn Interv Radiol. 2024;30(4):242-248.
39. Simões AM, Meneses JP, Oliveira PG, et al. Clinical validation of an Artificial Intelligence software for bone age assessment based on Greulich and Pyle method in a Portuguese paediatric cohort. Eur J Radiol Artif Intell. 2025;6:100027.
40. Shin NY, Lee YS, Choi JW, et al. External validation of deep learning-based bone-age software: a preliminary study with real world data. Sci Rep. 2022;12(1):1401.
41. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128(2):336-359.
42. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Proc ACM Conf Health Inference Learn. 2020;2020:151-159.

Figures

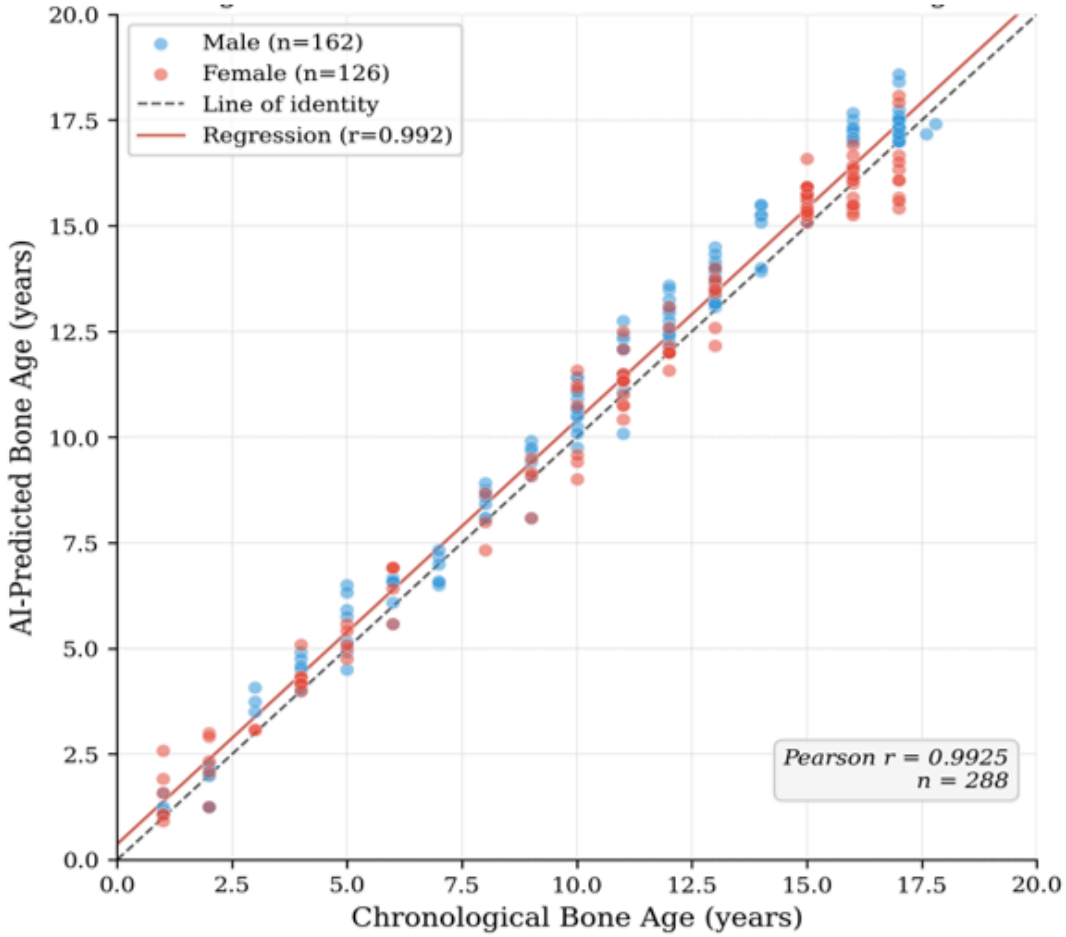


Figure 1

Scatter plot of AI-predicted bone age versus ground truth (reference) bone age. Blue circles represent male subjects (n = 162) and red circles represent female subjects (n = 126). The dashed black line indicates the line of identity (perfect agreement), and the solid red line represents the linear regression fit. Pearson correlation coefficient r = 0.993 (p < 0.001).

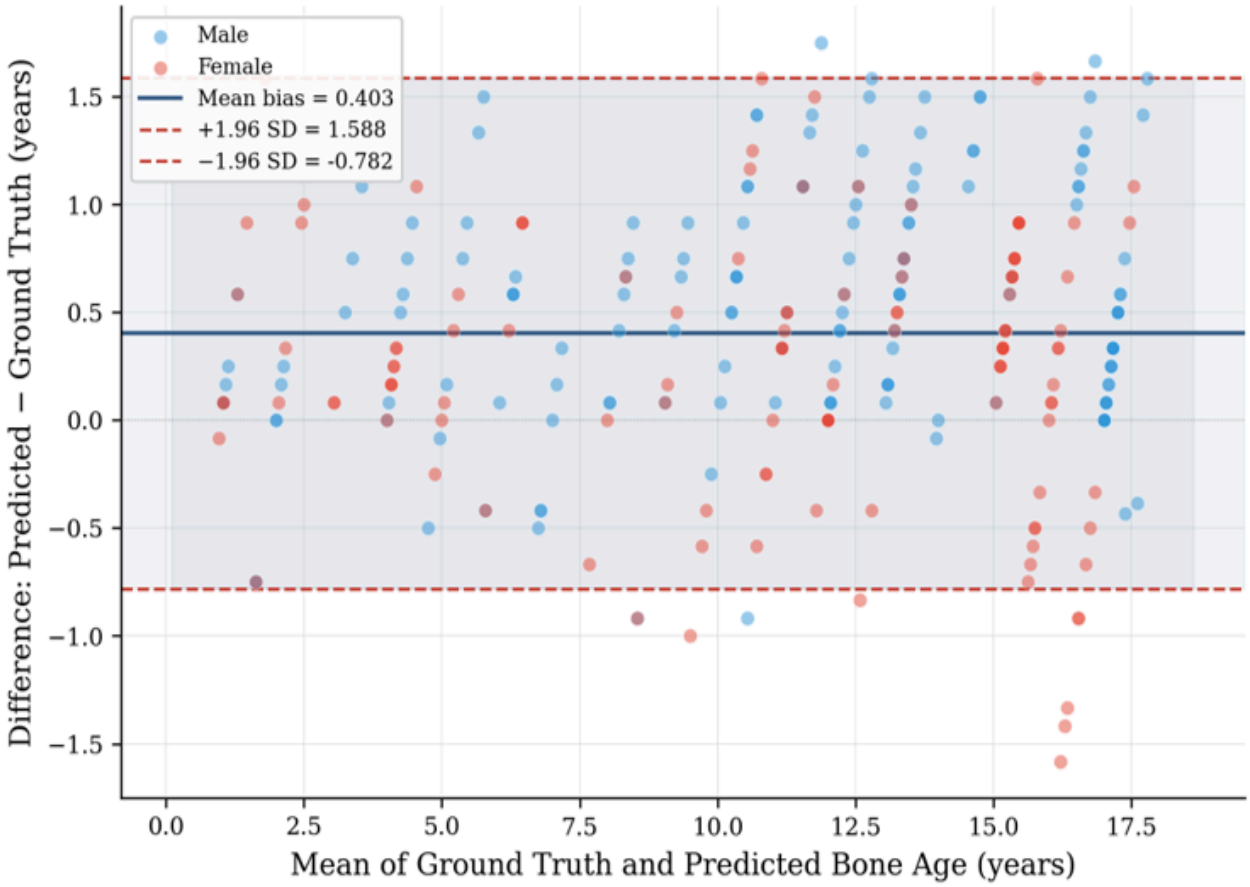


Figure 2

Bland-Altman plot showing the difference between AI-predicted and reference bone age plotted against the mean of both measurements. The solid horizontal line indicates the mean bias (+0.403 years), and dashed lines indicate the 95% limits of agreement (−0.782 to +1.588 years). Points are color-coded by gender.

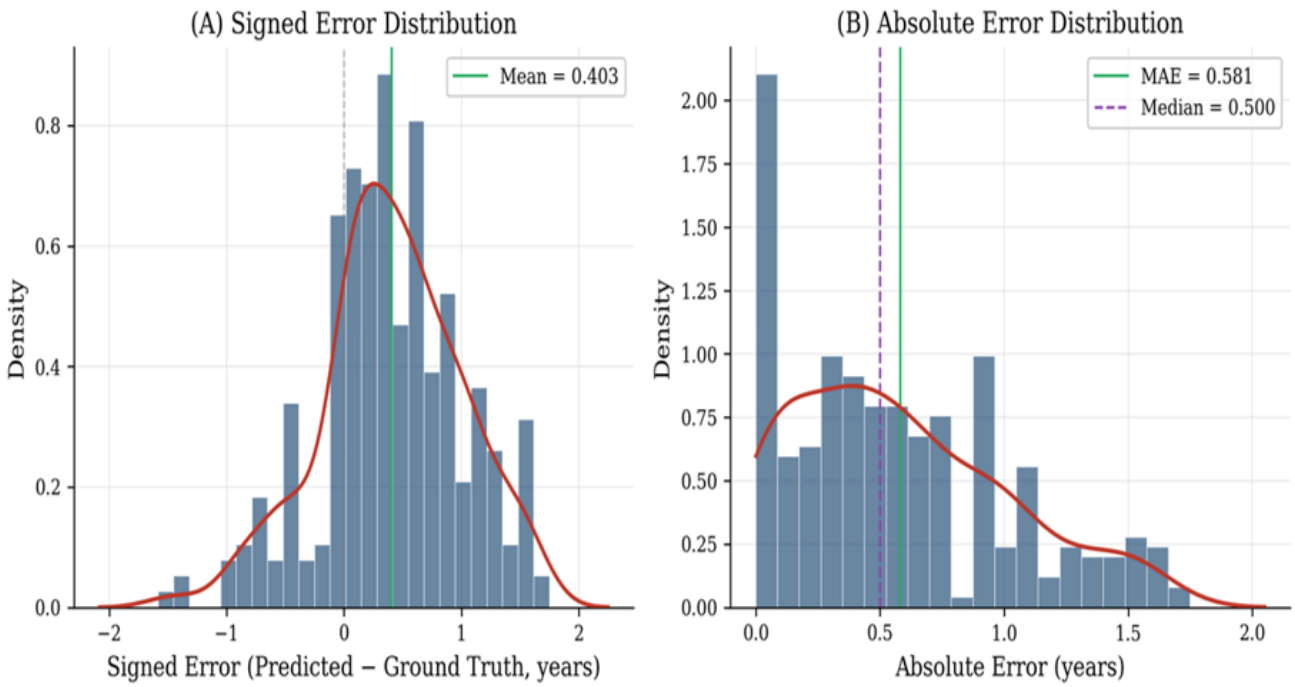


Figure 3

Distribution of prediction errors. (A) Signed error distribution showing the positive shift corresponding to systematic overestimation (mean = 0.403 years). (B) Absolute error distribution demonstrating that the majority of errors cluster below 1.0 year (MAE = 0.581 years; median = 0.500 years).

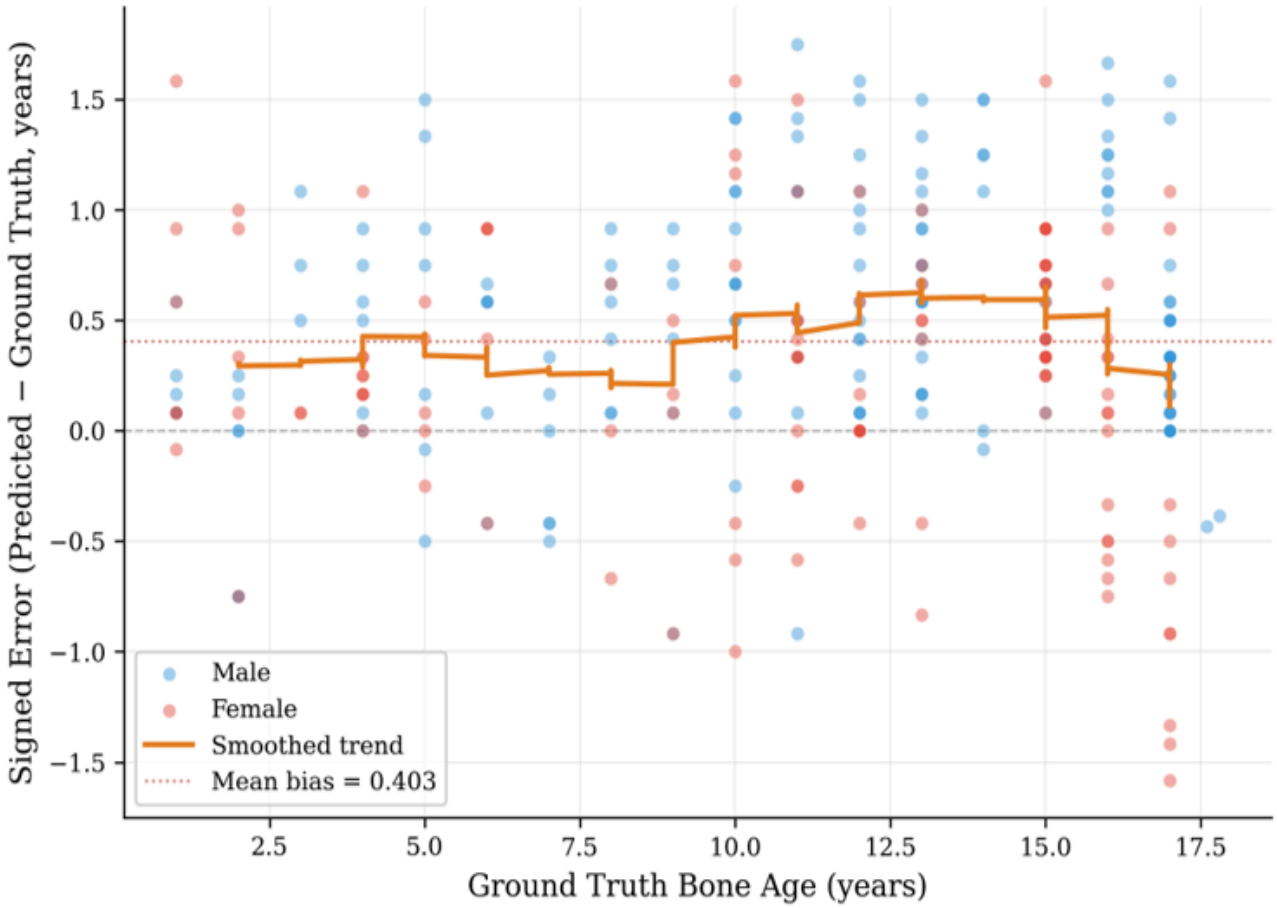


Figure 4

Residual plot showing prediction error (signed difference) versus ground truth bone age. The orange line represents the smoothed trend, and the horizontal dashed line indicates the mean bias. Points are color-coded by gender.

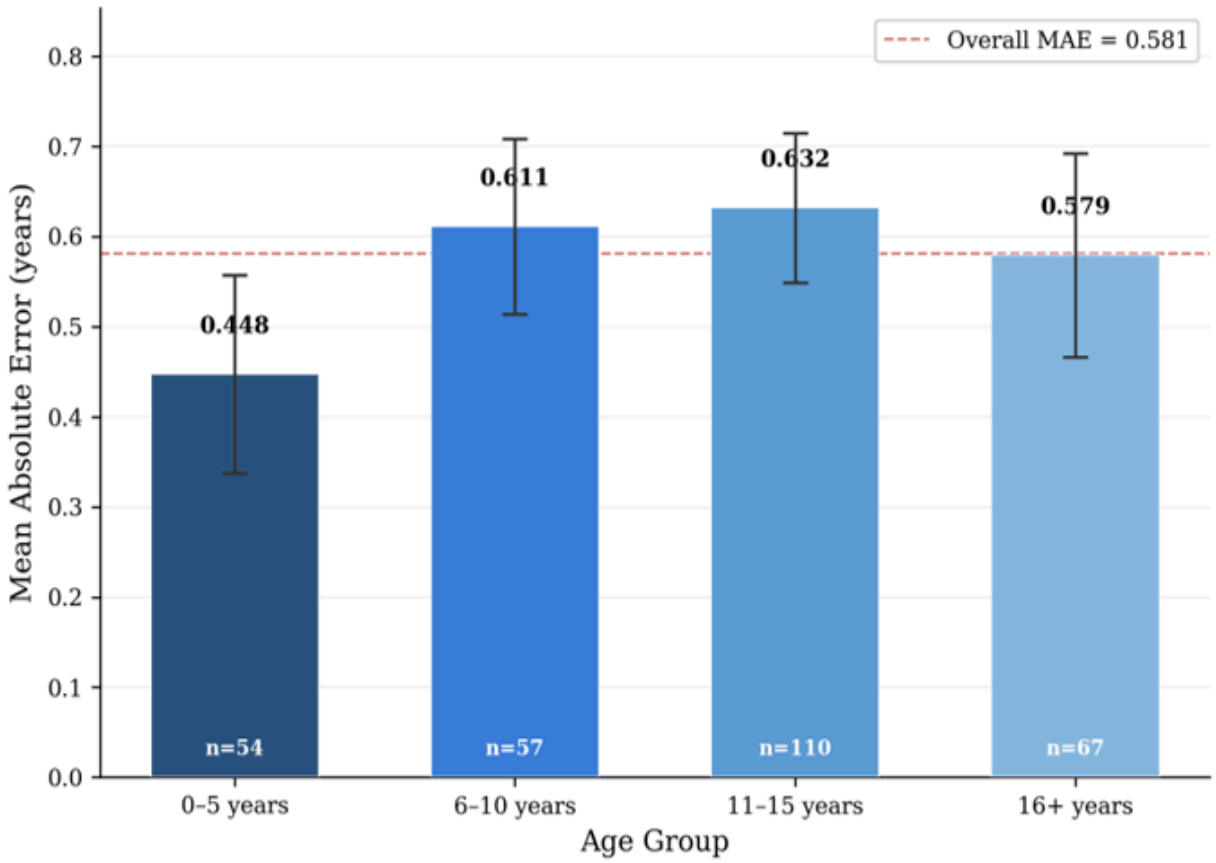


Figure 5

Mean absolute error (MAE) by age group with 95% confidence intervals. The dashed horizontal line indicates the overall MAE (0.581 years). Sample sizes: 0-5 years (n = 54), 6-10 years (n = 57), 11-15 years (n = 110), 16+ years (n = 67).

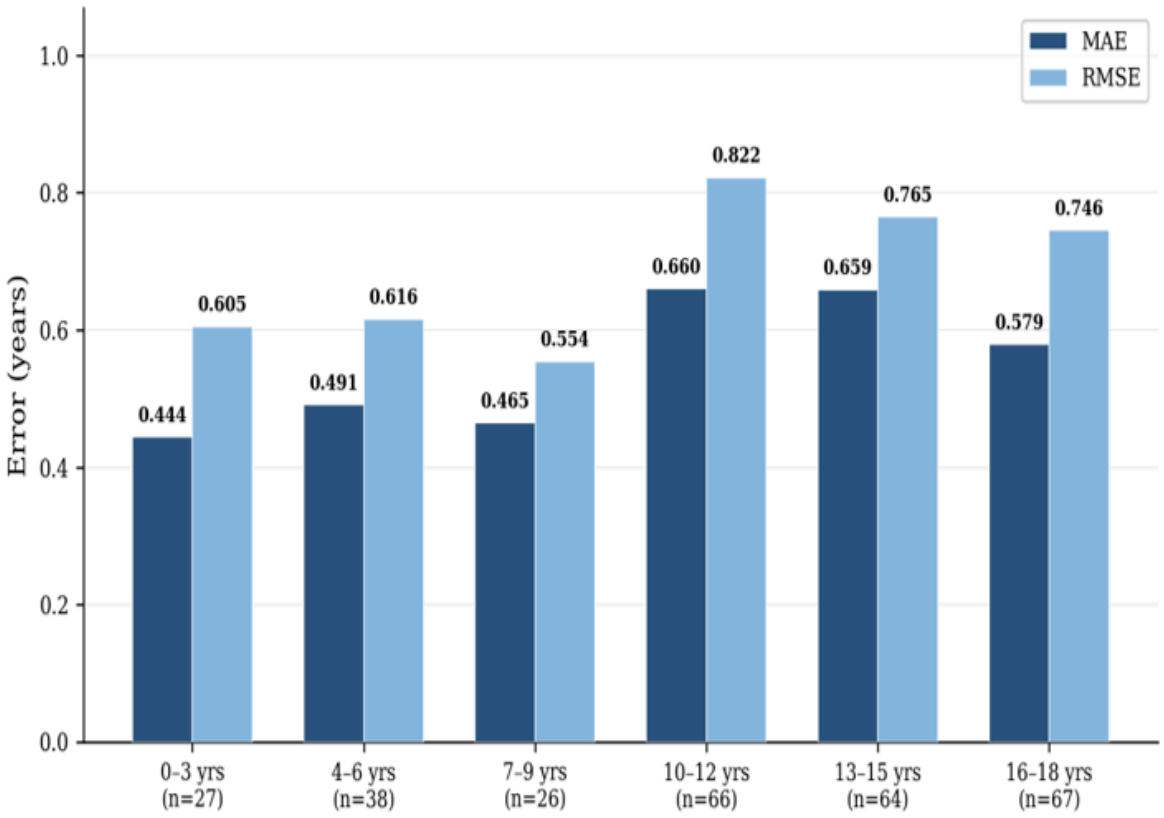


Figure 6

Mean absolute error (MAE) by age group with 95% confidence intervals. The dashed horizontal line indicates the overall MAE (0.581 years). Sample sizes: 0-5 years (n = 54), 6-10 years (n = 57), 11-15 years (n = 110), 16+ years (n = 67).

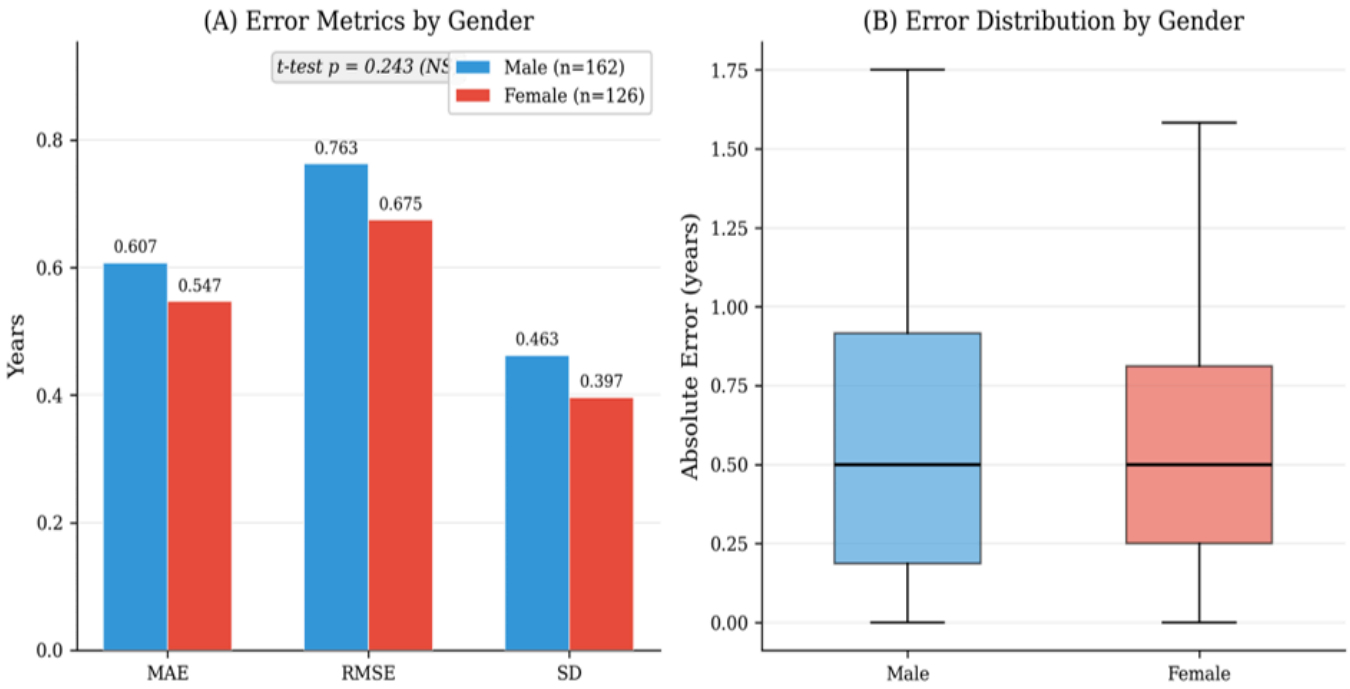


Figure 7

Figure 6. Performance comparison by gender. (A) Bar chart showing MAE, RMSE, and standard deviation for males (n = 162) and females (n = 126). No statistically significant difference was observed (t-test p = 0.243). (B) Box plots showing the distribution of absolute errors by gender.

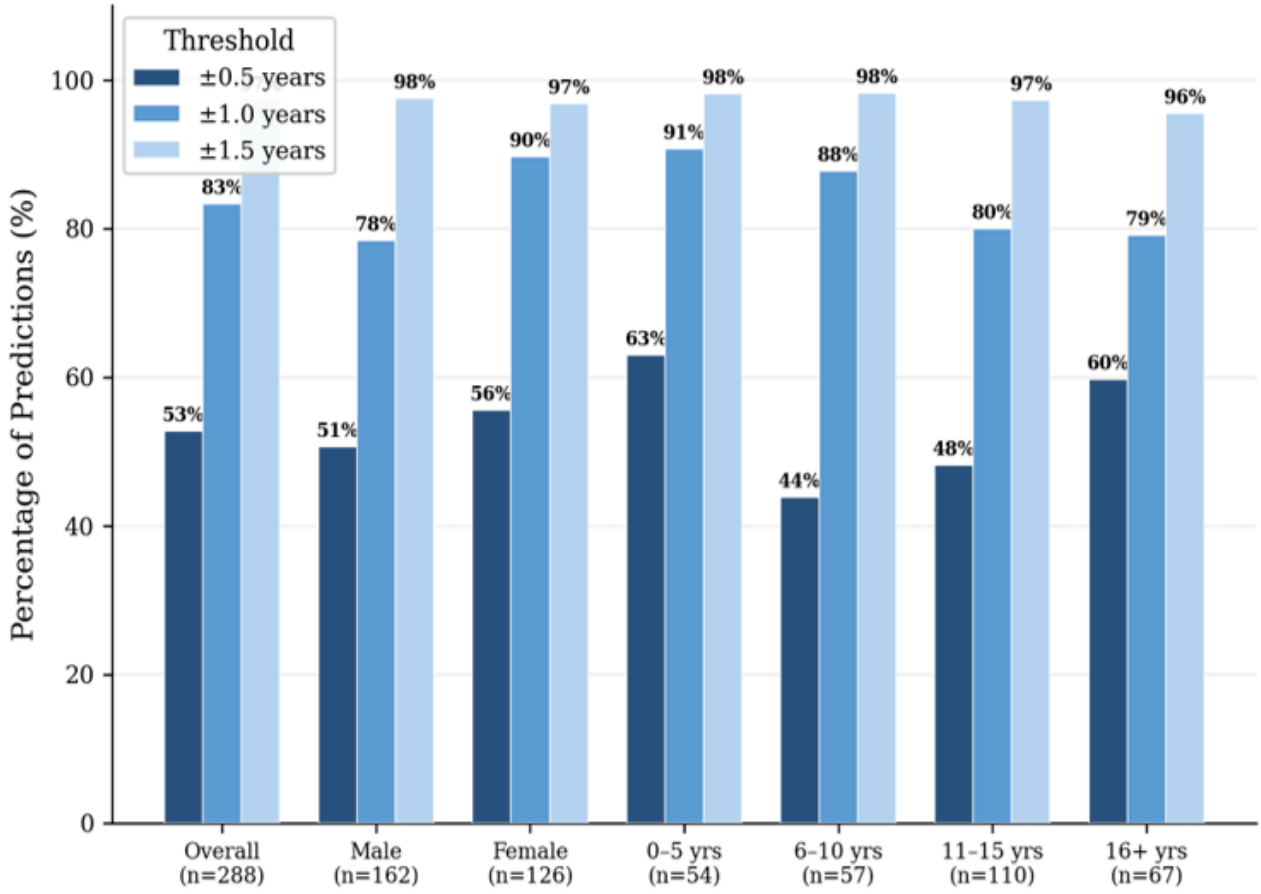


Figure 8

Figure 7. Percentage of predictions within clinical error thresholds (±0.5, ±1.0, and ±1.5 years), stratified by overall sample, gender, and age group.

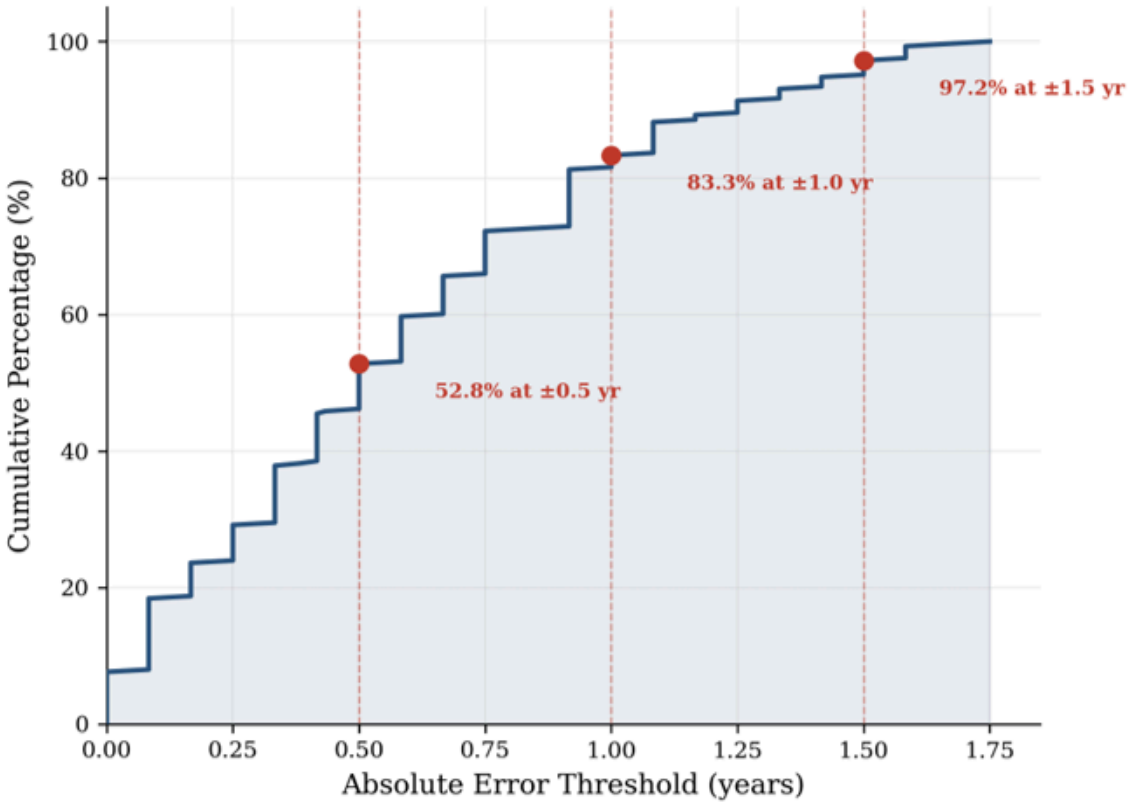


Figure 9

Figure 8. Cumulative distribution of absolute errors showing the percentage of predictions within increasing error thresholds. Key thresholds: 52.8% at ±0.5 years, 83.3% at ±1.0 year, and 97.2% at ±1.5 years.

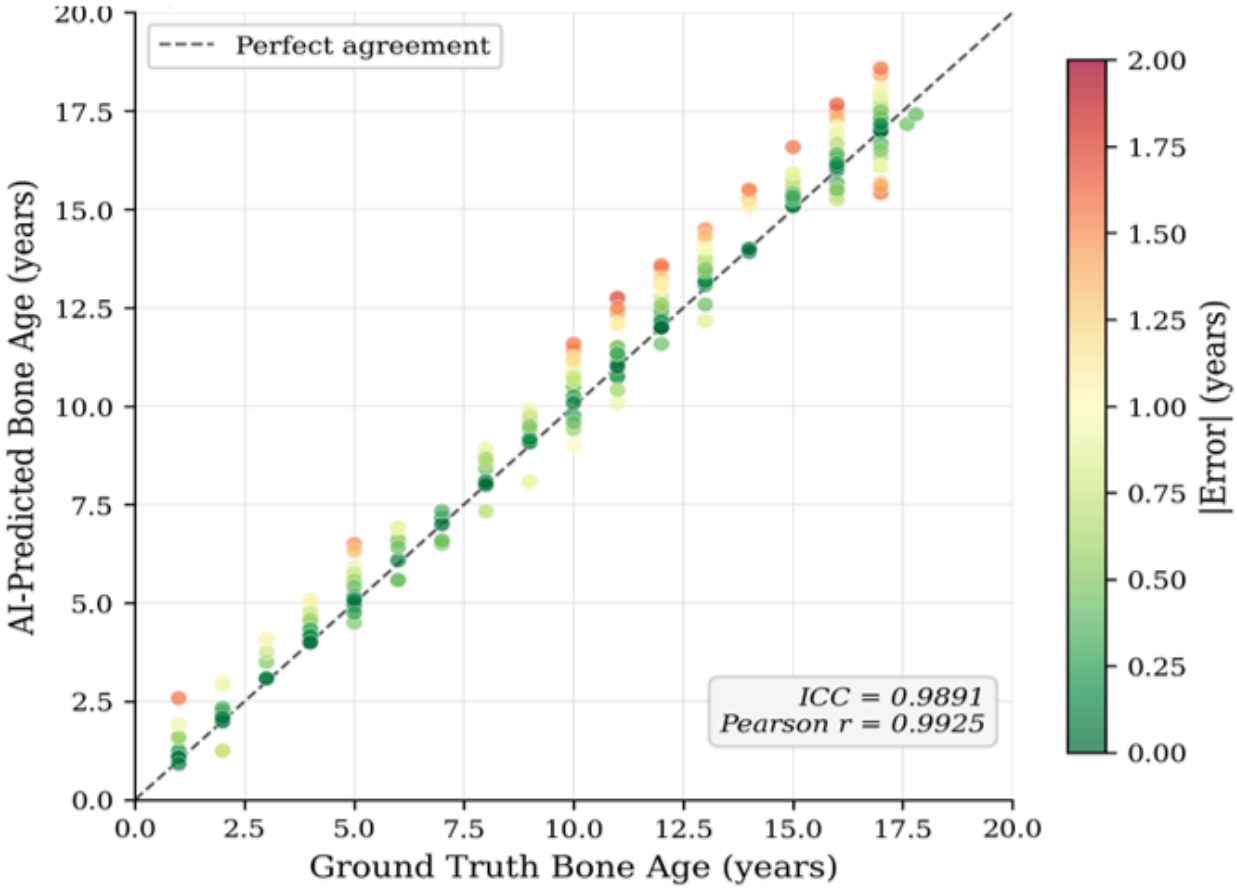


Figure 10

Figure 9. Agreement plot with error magnitude visualization. Points are color-coded by absolute error magnitude (green = low error, red = high error). The dashed line represents perfect agreement. ICC = 0.989; Pearson r = 0.993.