DOI: 10.35191/medsoft_2025_1_37_kvak

RETROSPECTIVE VALIDATION OF AN AI ALGORITHM FOR AUTOMATED BONE AGE ASSESSMENT IN PAEDIATRIC HAND RADIOGRAPHS

Karolína Kvaková, Daniel Kvak, Zdeněk Straka

Abstract

Bone age is a radiological indicator of bone maturity that is routinely assessed in children and adolescents to evaluate growth and diagnose endocrine or chronic diseases. This retrospective study verifies the accuracy of an artificial intelligence algorithm (Carebot AI Bones, Bone Age function; Carebot s. r. o.) for automatically estimating bone maturity from dorsopalmary X-ray images. We analyzed 96 anonymized images (20–216 months; median 108) taken between January and June 2025. The reference standard was independently established by a radiologist and anthropologist according to the GP atlas, with consensus in case of disagreement. The index test was the algorithm's prediction in months. The primary endpoint was the mean absolute error (MAE) compared to a pre-specified non-inferiority limit of 12 months. Secondary measures included RMSE, bias, Pearson's r, Bland–Altman limits of agreement, and proportions within ±6/±12/±24 months. The algorithm showed a high correlation with the reference standard (r = 0.981; 95% CI 0.970–0.989). MAE was 5.97 months (95% CI 4.76–7.28), RMSE was 8.70, and bias was −0.27 with LoA −17.40 to +16.86. Predictions were within ±6/±12/±24 months in 66.7%/82.3%/96.9% of cases. Non-inferiority was met (t=−9.29; p<0.001). By gender, the MAE was 5.04 months for men (bias +2.79) and 6.82 months for women (bias −3.09). The lowest error was ≤60 months (MAE 3.40), with a slight underestimation occurring at 121–180 months (MAE 7.15; bias −3.47). The results show that the AI algorithm achieves an average error of less than 1 year across the entire pediatric age spectrum and meets the criteria for clinical acceptability, supporting its use as a tool to aid radiological decision-making.

Keywords

bone age, Greulich–Pyle, paediatric radiology, artificial intelligence, validation, Bland–Altman, concordance correlation

1 Introduction

Bone age (BA)—the radiographic estimate of skeletal maturity—remains a cornerstone of paediatric radiology and endocrinology because it contextualizes growth velocity, clarifies the aetiology of stature abnormalities, and informs both therapeutic timing and adult-height prediction [1]. In standard care, clinicians determine BA by visually matching a left-hand/wrist radiograph to reference standards. Greulich–Pyle (GP) atlas [2], derived from mid-twentieth century North American children, is favored in everyday practice because an experienced reader can complete the match in 1–2 minutes. Recently, computer-assisted BA estimation has advanced rapidly. Early rule-based tools such as BoneXpert segmented predefined features and showed consistent performance across several European cohorts [3] but still required expert oversight in atypical cases. The advent of deep learning transformed this landscape, with convolutional neural networks (CNNs) trained on tens of thousands of radiographs capture maturation cues beyond handcrafted features. In the Radiological Society of North America's 2017 machine-learning challenge, leading CNNs achieved mean absolute errors (MAE) of roughly 4–5 months, on par with senior paediatric radiologists [4]. Subsequent external validations reported comparable accuracy, substantial time savings, and improved inter-reader agreement [5,6]. These advances have enabled clinical translation. Several deep-learning BA systems now hold regulatory clearance, nevertheless, regulatory approval does not guarantee generalizable performance, with domain shift in acquisition parameters, demographics, or disease prevalence lowering accuracy. Independent, site-specific validation is therefore necessary to demonstrate real-world safety and effectiveness [7].

The present retrospective study addresses this need by quantifying agreement between developed AI algorithm and expert GP assessments in 96 routine hand radiographs from a tertiary hospital. By analyzing overall error metrics and age-stratified performance, we evaluate whether the algorithm maintains a clinically acceptable MAE (< 12 months) and supports integration into everyday paediatric practice.

2 Materials and Methods

2.1 Study Design

This retrospective, single-centre diagnostic-accuracy study compares the AI algorithm automated bone age estimates with a dual-expert Greulich–Pyle consensus reference on the same de-identified left-hand/wrist radiographs. Human readers were blinded to the AI output. For each case, the AI prediction and the consensus reference formed a paired observation for head-to-head analysis.

2.2 Software

The investigated AI algorithm (Carebot AI Bones, Bone Age function; Carebot s. r. o.) uses a convolutional neural network with a ResNet-50 backbone that outputs bone age in months. Because skeletal maturation exhibits sex-specific patterns, we trained separate models for males and females. The final training dataset was collected retrospectively from multiple institutions to improve generalizability. All training images were standard dorsopalmar hand/wrist radiographs from patients < 18 years. Sex labels were extracted from study metadata and verified during curation. After quality filtering (exclusion of non-paediatric or incomplete views), the training dataset includes 1,893 male images and 1,442 female images, totaling 3,335 hand radiograph X-rays. Source DICOMs were converted for training, with intensity normalization applied. To mitigate domain shift and enhance robustness, we used data augmentation comprising random intensity jitter (brightness/contrast), and standard geometric transforms (e.g., random flips/rotations within small ranges). Normalization statistics were tuned to the training distribution. Models were trained with mini-batches (typical batch size = 16) using Adam-type optimization and a low learning rate (e.g., 5×10^{-5} for the reference male run). Model selection was pre-specified by lowest validation MAE on a held-out internal split from the multi-centre training pool. At AI algorithm's inference, the system reads the de-identified radiograph, applies the same normalization pipeline, routes by sex to the corresponding model, and returns a continuous bone-age estimate (months) together with per-study metadata for audit.

2.3 Data Collection

This single-centre, retrospective cohort was assembled at the Department of Radiology, University Hospital Olomouc from consecutive dorsopalmar left-hand/wrist X-rays acquired between 1 January and 30 June 2025. Case identification and data transfer were conducted under an institutional co-operation and data-sharing agreement between University

Figure 1 – Example of the standard dorsopalmar left-hand/wrist X-ray (left) and the output of the AI algorithm (Carebot AI Bones, Bone Age function; Carebot s.r.o.) for automated bone age assessment (right)

Hospital Olomouc and Carebot (agreement ref.: Agreement on cooperation in the development of software) from 14th November 2025. Before transfer, all studies were fully de-identified at source, with DICOM headers scrubbed of direct; no re-identification keys left the hospital network. Inclusion criteria included (i) patient age 0–18 years at imaging, (ii) native dorsopalmar left hand/wrist radiograph, and (iii) complete, artefact-free visualisation of hand and wrist. Exclusion criteria were (i) insufficient diagnostic quality (e.g., marked motion artefact, severe under/over-exposure, obscuration/cropping of key anatomy); (ii) presence of a cast, external fixation, or prominent postoperative hardware; (iii) gross traumatic deformity precluding reliable assessment; and (iv) duplicate examinations within the window, in which case the first complete study was retained. Applying these criteria yielded a final analytic dataset of n = 96 valid X-ray images.

2.4 Reference Standard

To establish the reference standard, each dorsopalmar left-hand/wrist X-ray was first assessed by a board-certified radiologist using the Greulich–Pyle (GP) atlas to assign an initial bone-age estimate. A physical anthropologist then performed an independent second reading. Discrepancies were resolved at an adjudication session, and the consensus value ("atlas age") was used as the reference standard. Human readers were blinded to the AI output for all cases.

The distribution in the analyzed cohort spanned 20–192 months (median 108 months). The cohort comprised 46 males (47.9%) and 50 females (52.1%). Distribution across predefined bone-age (BA) bands is shown below.

Bone-age band (months)	n (%)
≤ 60	24 (25.0)
61–120	35 (36.5)
121–180	35 (36.5)
> 180	2 (2.0)
Total	96

Table 1 – Distribution of patients across predefined bone-age bands in the analyzed cohort

2.5 Statistical Analysis

Accuracy of the AI algorithm was quantified against the reference atlas age on a per-case basis. For each X-ray, we paired the AI-predicted bone age with the atlas age and computed mean absolute error (MAE), root-mean-square error (RMSE), mean bias (AI – reference), Pearson's correlation (r), Bland–Altman bias and limits of agreement, and the proportions within ±6, ±12, and ±24 months of the reference. The prespecified primary endpoint was non-inferiority of MAE to a clinical margin δ = 12 months. Non-inferiority was concluded if the one-sided 97.5% upper confidence bound for MAE was < 12 months (equivalently, the upper bound of the two-sided 95% CI < 12). Confidence intervals for MAE, RMSE, bias, and Bland–Altman parameters were obtained by percentile bootstrap (10,000 resamples). As a sensitivity analysis, a one-sample, one-sided t-test versus δ was also reported. To relate estimates to biological maturation, we compared absolute deviations from chronological age: for each case we computed |AI – chronological| and |reference – chronological| and tested their paired difference using the Wilcoxon signed-rank test (two-tailed, α = 0.05), which avoids distributional assumptions for absolute-error data.

All analyses were performed in Python 3.10 (pandas, NumPy, SciPy); key results were cross verified in R 4.3. Bootstrap procedures used a fixed random seed to ensure reproducibility.

3 Results

3.1 Overall Accuracy and Sex-Stratified Performance

The AI algorithm's predictions showed strong agreement with the atlas age (consensus Greulich–Pyle reference standard). The linear association was high (r = 0.981, p < 0.001). The mean absolute error (MAE) was 5.97 months (95% CI 4.76–7.25), with an RMSE of 8.70 months. Systematic error was negligible: the mean bias (AI – reference) was –0.27 months (95% CI –2.02 to +1.43), not different from zero. In terms of clinically relevant bands, 66.7% of predictions were within ±6 months, 82.3% within ±12 months, and 96.9% within ±24 months of the reference. Sex-stratified analyses demonstrated consistently high performance in both groups, with a small, opposing bias by sex that largely cancels in the pooled data. In males (n = 46), MAE was 5.03 months, RMSE 7.49 months, bias +2.79 months, and r = 0.990; 73.91% and 91.30% of predictions were within ±6 and ±12 months, respectively. In females (n = 50), MAE was 6.82 months, RMSE 9.68 months, bias –3.09 months, and r = 0.976; 60.00% and 74.00% were within ±6 and ±12 months, respectively. Virtually all cases of both sexes were within ±24 months (≥96%).

3.2 Age-Stratified Performance

We examined performance across the defined age groups (≤60, 61–120, 121–180, >180 months). In the youngest group (≤60 months, *n* = 24), the AI algorithm achieved the highest accuracy, with MAE 3.40 months and RMSE 4.00 months; 87.5 % of estimates were within ±6 months and all were within ±12 and ±24 months. The correlation in this infant/toddler group was very high (r = 0.951), and a slight positive bias was observed (AI on average overestimated by +2.66 months). In mid-childhood (61–120 months; *n* = 35), error increased to MAE 6.47 months, RMSE 9.33 months. About 62.9% of predictions in this group were within ±6 months. The correlation remained strong (r = 0.888), and bias was small (+1.3 months). In early adolescents (121–180 months, *n* = 35), error was slightly higher (MAE 7.15 months, RMSE 10.2), with 57.1 % within ±6 months and 77.1 % within ±12 months. For the oldest patient group (>180 months,

DOI: 10.35191/medsoft_2025_1_37_kvak

Sex	N	Pearson r	MAE (mo)	RMSE (mo)	Bias (mo)	± 6 mo (%)	± 12 mo (%)	± 24 mo (%)
Overall	**96**	**0.981**	**5.97**	**8.7**	**-0.27**	**66.67**	**82.29**	**96.88**
Male	46	0.990	5.03	7.54	+2.83	73.91	91.30	97.83
Female	50	0.976	6.78	9.67	–3.06	60.00	74.00	96.00

Table 2 – Overall and sex-stratified performance of the AI algorithm versus the Greulich–Pyle reference standard (atlas age)

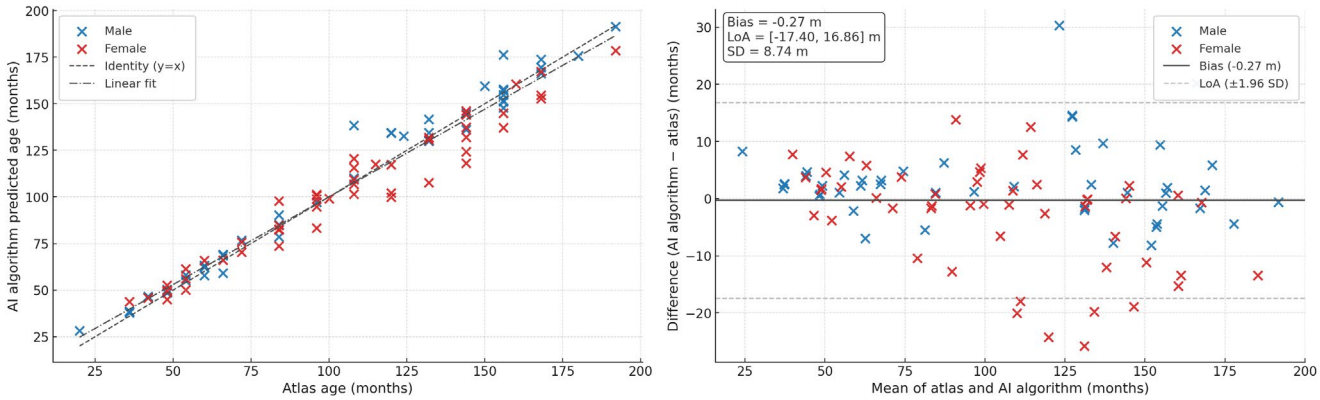

Figure 2 – Scatter (left) and Bland–Altman (right) plots illustrating agreement between AI-predicted and atlas bone age. Strong correlation and narrow limits of agreement indicate high accuracy with minimal bias.

n = 2), results are difficult to generalize due to the very small sample. In those two cases, the MAE was ~7.0 months with one prediction within 6 months of the reference and one large error (~13 months underestimation). Correlation is not meaningful for 2 cases (essentially N/A), but no systematic conclusion can be drawn from such a limited data subset. Notably, bias flipped sign across development: the AI algorithm overestimated in early ages (≤ 60: +2.66 m; 61–120: +1.30 m) and underestimated during puberty (121–180: –3.47 m). Across all subgroups, ≥ 94 % of predictions were within ±24 months.

3.3 Non-Inferiority Test and Power Analysis

To relate estimates to biological maturation, we compared absolute deviations from chronological age for the AI and for the reference method. Median absolute error from true age was 12.35 months for AI algorithm and 11.02 months for the reference atlas age; the paired Wilcoxon signed-rank test showed no significant difference (p = 0.11), indicating that AI-derived bone ages diverge from true age to a similar extent as expert GP readings.

Age Group (mo)	N	Pearson r	MAE (mo)	RMSE (mo)	Bias (mo)	±6 mo (%)	±12 mo (%)	±24 mo (%)
≤ 60	24	0.951	3.40	4.00	+2.66	87.50	100.00	100.00
61 – 120	35	0.888	6.47	9.33	+1.30	62.86	77.14	97.14
121 – 180	35	0.813	7.15	10.22	–3.47	57.14	77.14	94.29
> 180	2	N/A	7.04	9.52	–7.04	50.00	50.00	100.00

Table 3 – Performance by atlas-derived bone-age group

Figure 3 – Examples of the output of the AI algorithm (Carebot AI Bones, Bone Age function; Carebot s.r.o.) for automated bone age assessment. For each case, the AI algorithm displays the chronological age (white) and the AI-estimated bone age (blue), while the column on the right shows nearest atlas exemplars (ranked candidate ages).

For the prespecified non-inferiority assessment (margin δ = 12 months), the observed MAE was 5.97 months (SD 6.41). A one-sided one-sample t-test rejected H_0: MAE ≥ 12 (t = –9.29, df = 95, one-sided p < 0.001), confirming that accuracy of the AI algorithm is non-inferior, substantially better than the 12-month threshold. As a sensitivity analysis, the one-sided 97.5% upper confidence bound for MAE lay below δ, likewise supporting non-inferiority.

A post-hoc power calculation based on the observed mean (5.97 months), SD (6.41), n = 96, and α = 0.025 (one-tailed) indicated ≈100% power to demonstrate non-inferiority at δ = 12 months. Given the large margin between the point estimate and δ, the probability of a Type II error was effectively negligible. These findings show that the AI algorithm meets the predefined clinical acceptability criterion across the cohort and remain consistent within age- and sex-stratified analyses.

4 Discussion

The proposed AI algorithm (Carebot AI Bones, Bone Age function; Carebot s. r. o.) achieved sub-year error against a dual-expert Greulich–Pyle (GP) atlas reference (MAE 5.97 months; RMSE 8.70 months; bias −0.27 months; r ≈ 0.98), meeting the prespecified 12-month acceptability margin. In clinical terms, a ~half-year deviation from the atlas standard is negligible in paediatric endocrine practice and aligns with prior reports of deep-learning models showing sub-year errors and strong concordance with expert assessment [1]. Agreement within clinically relevant bands was high (±6 months: 66.7%, ±12 months: 82.3%, ±24 months: 96.9%). Given that human readers can differ by over a year in a substantial minority of cases, these findings are consistent with expected expert-level variability [8]. Subgroup analyses were consistent across demographics. Accuracy was highest in early childhood (≤ 60 months; MAE 3.40), with a modest underestimation around puberty (121–180 months; MAE 7.15, bias −3.47). Sex-stratified results showed small, opposing biases—+2.79 months in boys and −3.09 months in girls—that largely cancel in aggregate; both sexes remained well within the 12-month margin, in line with prior observations of no meaningful sex-related disparity. A few > ±24-month outliers clustered near early puberty, mirroring known variability during rapid growth.

Strengths include a consensus dual-reader reference, a prespecified clinical margin and analysis plan, and comprehensive agreement reporting. Limitations are the retrospective, single-centre design, modest size with under-representation at older adolescence, and lack of workflow endpoints; generalizability to endocrine pathologies and late adolescence warrants prospective, multi-centre confirmation.

In practice, an automated estimate available within seconds can support paediatric radiology and endocrinology by standardizing assessments and reducing manual workload. Notably, prior work has shown that automated bone-age tools can reduce reading times by up to 87% [9]. Consistency may also mitigate inter-observer variability, important given that differences ≥1 year are not uncommon between human interpreters [10]. The intention is augmentation, not replacement; as a second reader, the AI can assist trainees and provide confirmation for experienced radiologists. Observed performance (MAE < 12 months) is compatible with regulatory expectations for decision-support software [11].

5 Conclusion

The AI algorithm achieved MAE 5.97 months, r = 0.98, and ≥ 82% of predictions within ±12 months, with sub-year accuracy across all age and sex subgroups and negligible bias versus the consensus Greulich–Pyle reference. These results meet the prespecified ±12-month non-inferiority threshold, supporting use as a second-reader decision-support tool in paediatric radiology and endocrinology. Prospective, multi-centre studies—including older adolescents and children with growth disorders—should confirm generalizability and quantify workflow impact.

RETROSPEKTIVNÍ VALIDACE ALGORITMU UMĚLÉ INTELIGENCE PRO AUTOMATICKÉ STANOVENÍ KOSTNÍHO VĚKU NA PEDIATRICKÝCH RENTGENOVÝCH SNÍMCÍCH RUKY

Abstrakt

Kostní věk je radiologický ukazatel kostní zralosti, který se u dětí a adolescentů rutinně hodnotí k posouzení růstu a k diagnostice endokrinních či chronických onemocnění. Tato retrospektivní studie ověřuje přesnost algoritmu umělé inteligence (Carebot AI Bones, funkce Bone Age; Carebot s. r. o.) pro automatický odhad kostní zralosti z dorzopalmárních rentgenových snímků. Analyzovali jsme 96 anonymizovaných snímků (20–216 měsíců; medián 108) pořízených mezi lednem a červnem 2025. Referenční standard stanovili nezávisle radiolog a antropolog podle atlasu GP s konsenzem při neshodě. Indexovým testem byla predikce algoritmu v měsících. Primárním koncovým bodem byla průměrná absolutní chyba (MAE) ve srovnání s předem stanovenou mezí neinferiority 12 měsíců. Sekundární ukazatele zahrnovaly RMSE, zkreslení, Pearsonův r, Bland–Altmanovy meze shody a podíly v rozmezí ±6/±12/±24 měsíců. Algoritmus vykázal vysokou korelaci s referenčním standardem (r = 0,981; 95% CI 0,970–0,989). MAE činila 5,97 měsíce (95% CI 4,76–7,28), RMSE 8,70 a zkreslení −0,27 s LoA −17,40 až +16,86. Predikce byly v rozmezí ±6/±12/±24 měsíců v 66,7 %/82,3 %/96,9 % případů. Neinferiorita byla splněna (t=−9,29; p<0,001). Podle pohlaví byla MAE 5,04 měsíce u mužů (bias +2,79) a 6,82 měsíce u žen (bias −3,09). Nejnižší chyba byla u ≤60 měsíců (MAE 3,40), mírné podhodnocení se objevilo u 121–180 měsíců (MAE 7,15; bias −3,47). Výsledky ukazují, že algoritmus AI dosahuje průměrné chyby pod 1 rok v celém spektru pediatrického věku a splňuje kritéria klinické přijatelnosti, což podporuje jeho použití jako nástroje pro podporu radiologického rozhodování.

Klíčová slova

kostní věk, Greulich–Pyle, pediatrická radiologie, umělá inteligence, validace, Bland–Altman, korelace shody

References

[1.] Cavallo, F., Mohn, A., Chiarelli, F., & Giannini, C. (2021). Evaluation of Bone Age in Children: A Mini-Review. Frontiers in pediatrics, 9, 580314. https://doi.org/10.3389/fped.2021.580314

[2.] Greulich, W. W., & Pyle, S. I. (1959). Radiographic atlas of skeletal development of the hand and wrist. The American Journal of the Medical Sciences, 238(3), 393.

[3.] Thodberg, H. H., Kreiborg, S., Juul, A., & Pedersen, K. D. (2008). The BoneXpert method for automated determination of skeletal maturity. IEEE transactions on medical imaging, 28(1), 52-66.

[4.] Halabi, S. S., Prevedello, L. M., Kalpathy-Cramer, J., Mamonov, A. B., Bilbily, A., Cicero, M., ... & Flanders, A. E. (2019). The RSNA pediatric bone age machine learning challenge. Radiology, 290(2), 498-503.

[5.] Lee, K. C., Lee, K. H., Kang, C. H., Ahn, K. S., Chung, L. Y., Lee, J. J., ... & Shim, E. (2021). Clinical validation of a deep learning-based hybrid (Greulich-Pyle and modified Tanner-Whitehouse) method for bone age assessment. Korean Journal of Radiology, 22(12), 2017.

[6.] Lea, W. W. I., Hong, S. J., Nam, H. K., Kang, W. Y., Yang, Z. P., & Noh, E. J. (2022). External validation of deep learning-based bone-age software: a preliminary study with real world data. Scientific reports, 12(1), 1232.

[7.] Kim, J. R., Shim, W. H., Yoon, H. M., Hong, S. H., Lee, J. S., Cho, Y. A., & Kim, S. (2017). Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. American Journal of Roentgenology, 209(6), 1374-1380.

[8.] Breen, A. B., Steen, H., Pripp, A., Gunderson, R., Mentzoni, H. K. S., Merckoll, E., ... & Horn, J. (2022). A comparison of 3 different methods for assessment of skeletal age when treating leg-length discrepancies: an inter-and intra-observer study. Acta orthopaedica, 222-228.

[9.] Booz, C., Yel, I., Wichmann, J. L., Boettger, S., Al Kamali, A., Albrecht, M. H., ... & Bodelle, B. (2020). Artificial intelligence in bone age assessment: accuracy and efficiency of a novel fully automated algorithm compared to the Greulich-Pyle method. European radiology experimental, 4(1), 6.

[10.] Breen, A. B., Steen, H., Pripp, A., Gunderson, R., Mentzoni, H. K. S., Merckoll, E., ... & Horn, J. (2022). A comparison of 3 different methods for assessment of skeletal age when treating leg-length discrepancies: an inter-and intra-observer study. Acta orthopaedica, 222-228.

[11.] Zulkifley, M. A., Abdani, S. R., & Zulkifley, N. H. (2020). Automated bone age assessment with image registration using hand X-ray images. Applied Sciences, 10(20), 7233.

Kontakt

Mgr. Daniel Kvak
Carebot s.r.o.
Rašínovo nábřeží 71/10
128 00 Praha 2
+420 739 174 316
daniel.kvak@carebot.com
https://www.carebot.com/

Bc. Karolína Kvaková
Carebot s.r.o.
karolina.kvakova@carebot.com

Ing. Zdeněk Straka Ph.D.
Carebot s.r.o.
zdenek.straka@carebot.com