Hong Kong J Radiol. 2024;27(4):e221-7. Published 28 November 2024 | https://doi.org/10.12809/hkjr2317736

ORIGINAL ARTICLE

Validation of Artificial Intelligence for Bone Age Assessment in Hong Kong Children

C Cheung, JPK Chan, CWK Ng, WT Lai, KKF Fung, EYL Kan

Department of Radiology, Hong Kong Children's Hospital, Hong Kong SAR, China

ABSTRACT

Introduction: *We sought to evaluate the accuracy of an artificial intelligence (AI)–automated bone age analysis software, BoneXpert 3.0, in determining bone age in children in Hong Kong.*

Methods: *All radiographs of the left hand and the wrist for bone age assessment at a tertiary referral centre in Hong Kong from January to December 2019 were included. We compared the bone ages from these radiographs assessed by two experienced paediatric radiologists with analysis by BoneXpert using the Greulich and Pyle method. Gender-based bone age comparisons were also performed. The assessment involved calculating the Spearman's correlation (r), the coefficient of determination (R^2), and accuracy (root mean square error). Agreement between manual and AI-generated assessments was evaluated by Bland-Altman analysis.*

Results: *A total of 99 bone age radiographs were analysed. The mean chronological age was 9.8 years (standard deviation [SD] = 3.9 years). Manual and AI analyses showed a strong correlation ($r = 0.98$, $R^2 = 0.97$; $p < 0.001$). Bland-Altman analysis showed a mean difference of -0.08 year (SD = 0.73 year) and limits of agreement between 1.35 and -1.51 years. The correlation between visual and AI-generated bone age assessment remained strong after stratification by sex ($r = 0.98$, $R^2 = 0.97$; $p < 0.001$). Accuracy of the AI bone age analysis was 0.74 year for all studies, 0.79 year for females, and 0.65 year for males.*

Conclusion: *BoneXpert is reliable and accurate in bone age assessment in the local paediatric population.*

Key Words: *Algorithms; Artificial intelligence; Bone and bones; Pediatrics; Radiologists*

Correspondence: *Dr C Cheung, Department of Radiology, Hong Kong Children's Hospital, Hong Kong SAR, China*
Email: cc755@ha.org.hk

Submitted: 16 June 2023; Accepted: 4 December 2023.

Contributors: All authors designed the study. CC, JPKC, WTL and KKFF acquired the data. CC, JPKC and KKFF analysed the data. CC and KKFF drafted the manuscript. CC, JPKC, CWKN and EYLK critically revised the manuscript for important intellectual content. All authors had full access to the data, contributed to the study, approved the final version for publication, and take responsibility for its accuracy and integrity.

Conflicts of Interest: All authors have disclosed no conflicts of interest.

Funding/Support: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability: All data generated or analysed during the present study are available from the corresponding author on reasonable request.

Ethics Approval: This research was approved by the Hospital Authority Central Institutional Review Board – Paediatrics Panel, Hong Kong (Ref No.: PAED-2023-049). The requirement for patient consent was waived by Board due to the retrospective nature of the research.

中文摘要

人工智能對香港兒童骨齡評估的驗證

張樂人、陳沛君、吳穎琦、黎永德、馮建勳、簡以靈

引言：我們評估人工智能自動骨齡分析軟體BoneXpert 3.0在確定香港兒童骨齡方面的準確性。

方法：本研究納入2019年1月至12月期間在香港一所三級轉診中心進行骨齡評估的所有左手／手腕 X 光片，比較了兩位經驗豐富的兒科放射科醫生評估這些X光片骨齡與BoneXpert使用Greulich和Pyle方法的分析結果，並比較了基於性別的骨齡。我們使用的比較方法包括Spearman相關性（r）、決定系數（R^2）和準確性（均方根誤差），並使用Bland-Altman分析評估人工評估和人工智能評估之間的一致性。

結果：本研究總共分析了99張骨齡 X光片，平均實際年齡為9.8歲（標準差 = 3.9 歲）。人工和人工智能分析顯示出強相關（$r = 0.98$，$R^2 = 0.97$；$p < 0.001$）。Bland-Altman分析顯示平均差異為-0.08年（標準差 = 0.73年），一致限度為1.35至-1.51年。按性別分層後，人工骨齡評估和人工智能骨齡評估之間的相關性仍然強（$r = 0.98$，$R^2 = 0.97$；$p < 0.001$）。所有研究的人工智能骨齡分析準確度為0.74歲，女性為0.79歲，男性為0.65歲。

結論：BoneXpert 對本港兒科族群的骨齡評估可靠且準確。

INTRODUCTION

Bone age assessment is an integral part in the evaluation of paediatric growth and pubertal disorders. Accurate determination of bone age is important in assessing growth potential and timing of therapeutic interventions. For instance, in children with idiopathic short stature undergoing growth hormone treatment, continuous monitoring of bone age is vital to estimate potential height gain and adjust the treatment dosage.[1]

Conventional bone age assessment most frequently utilises the Greulich and Pyle (GP) or the Tanner and Whitehouse (TW) methods, both of which rely on visual comparison of radiographs of the left hand and the wrist of the patient against matching reference radiographs stratified by age and sex. However, these manual grading methods are subjective and prone to inter- and intra-rater variability.[1-3] Longitudinal assessment of multiple bone age radiographs for the same patient over time can yield inconsistent results when interpreted by different radiologists. Moreover, manual bone age assessment is time-consuming, particularly for inexperienced raters, with average reported rating time being 1.4 minutes for the GP method and 7.9 minutes for the TW method.[4] In addition, calculations of standard deviations of bone age using data in the atlas may introduce errors in the reports.

To address these challenges, artificial intelligence (AI)–based algorithms have been developed to reduce inconsistencies and eliminate inter-rater and intra-rater variability in bone age assessment in children. The evolution of AI-based bone age analysis has closely followed the advancements in machine learning through the decades.[5] BoneXpert (Visiana, Hørsholm, Denmark), launched in 2009, is the first AI-automated bone age assessment software that is commercially available and licensed for use in Europe.[6] The program utilises traditional machine-learning methodology and determines bone age based on shape, intensity, and texture scores. The algorithm segments the radius, ulna, metacarpals, and phalanges and determines an independent bone age value for each. A self-validation mechanism exists to reject bones for analysis if their morphologies lie out of the expected range of the bone-finding model or if their bone age values deviate by more than a predefined threshold from the mean bone age determined from all the tubular bones (Figure 1a). The algorithm also rejects the image if there are fewer than eight accepted bones to prevent erroneous bone age assessments.[6,7]

The final result is computed as the mean age of all the included bones.[6] The process is almost instantaneous and produces an annotated Digital Imaging and

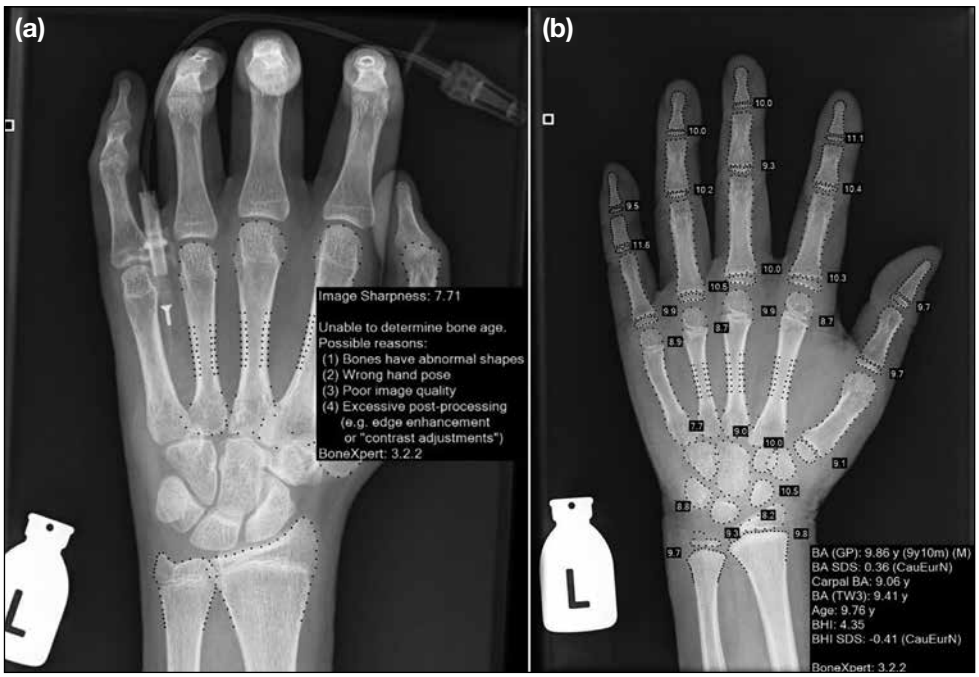

Figure 1. (a) An anteroposterior radiograph of the left hand in a 14-year-old boy rejected by the artificial intelligence (AI)–automated program BoneXpert for bone age assessment. The 2nd to 5th fingers were in flexion due to contractures and AI-automated software was unable to analyse the bone outline. The AI-generated algorithm also rejects the image if there are fewer than eight accepted bones to avoid assigning an erroneous bone age. (b) An annotated anteroposterior radiograph of the left hand and the wrist in a 9-year-and-10-month–old boy generated by bone age assessment using the same program. The algorithm segmented the bones and assigned a Greulich and Pyle (GP) bone age to each of them. The average bone age (BA) for the 21 tubular bones was reported as 'BA (GP): 9.86 y (M),' where 'M' indicates male sex. It also reported a bone age standard deviation score (SDS) of 0.36, which meant that the bone age was 0.36 standard deviation above the bone age expected at this chronological age. Chronological age was indicated below the bone age SDS at 9.76 years. The remaining reported numbers were: carpal BA = mean bone age in the visible carpals, BA (TW3) = bone age assessed by Tanner and Whitehouse version 3; BHI = bone health index; and its SDS relative to boys with the same bone age.

Communications in Medicine file (Figure 1b) with the software-calculated bone age data, which can be stored in a picture archiving and communication system as a permanent electronic medical record.

While the program has been validated in different Asian ethnicities,[8,9] including Chinese[10] and Japanese[11] children, it is crucial to validate its accuracy before implementing it in our local paediatric population in Hong Kong. The objective of our study is to evaluate the accuracy of this program in determining the bone age of children in Hong Kong, compared with visual bone age assessment by experienced paediatric radiologists.

METHODS

This study was performed as part of a quality assurance initiative. We retrospectively reviewed all bone age radiographs of the left hand and the wrist, as well as their radiology reports performed at Hong Kong Children's Hospital, a tertiary referral centre in Hong Kong, from January to December 2019. In accordance with our institutional practice, each radiograph was evaluated by two of seven experienced paediatric radiologists (with 6 to 7 years of experience in bone age assessment) using the GP method. The manual bone age of the patient was determined by consensus and recorded in the radiology report. Patient demographics, including sex, chronological age, diagnosis, and ethnicity (Chinese, South Asian, and Caucasian), were retrieved from the electronic patient record. All of the bone age radiographs were then analysed by BoneXpert 3.0 utilised in this study. The AI-generated bone age of the patient (the GP method) was determined by the aforementioned algorithm and was documented in an annotated image which was stored in a picture archiving and communication system. The interpreting radiologists were completely blinded to the AI-generated analysis results at the time of reporting. The Guidelines for Reporting Reliability and Agreement Studies were implemented.[12]

We compared the AI-generated bone age to the manual bone age for each patient. We also performed the comparison based on sex. We used the Spearman's correlation (r) and the coefficient of determination (R^2) when comparing AI-generated and manual bone age. Bland-Altman analysis was used to assess agreement between AI-generated and manual ratings. The accuracy of the AI-generated rating compared to manual rating using the GP method was defined as the root mean square error (RMSE) measured in years. Quantitative data are expressed in means ± standard deviations for comparing bone age as determined by the manual method versus the AI-automated method. Agreement was evaluated by Bland-Altman analysis. A p value < 0.05 was defined as statistically significant. All statistical analyses were performed with commercial software SPSS (Windows version 26.0; IBM Corp, Armonk [NY], United States).

Table. Patient demographics and study indications.

	No.	Spearman's correlation (r)	Coefficient of determination (R^2)	Accuracy (RMSE), y
Overall	99	0.98	0.97	0.74
Female	38	0.98	0.97	0.79
Male	61	0.98	0.97	0.65
Ethnicity				
Chinese	94	0.98	0.97	0.74
Others*	5	0.82	0.95	0.84
Study indication				
Pubertal disorders	23	0.92	0.95	0.75
Growth disorders	51	0.98	0.97	0.68
Miscellaneous†	25	0.98	0.97	0.83

Abbreviation: RMSE = root mean square error.
* Including South Asian (n = 4) and Caucasian (n = 1).
† Including bone marrow transplant workup (n = 21), adrenal disorders (n = 2), and orthopaedic assessment (n = 2).

RESULTS

Patients and Studies

A total of 99 bone age radiographs from January to December 2019 were analysed, 38 of which were from female patients and 61 were from male patients. The mean chronological age of the cohort was 9.8 ± 3.9 years (range, 1.5-17.8). The majority of patients (n = 94, 94.9%) were Chinese, with the rest being South Asian (n = 4, 4.0%) and Caucasian (n = 1, 1.0%). Regarding the indications for bone age assessment, 23 were evaluated for pubertal disorders, 51 for growth disorders, 21 for bone marrow transplant workup, two for adrenal disease, and two for orthopaedic assessment (Table).

For manual bone age, an exact bone age was determined in 93 radiographs while a bone age-range (e.g., between 3 years and 3 years and 6 months) was provided for six radiographs. For these six radiographs, the midpoint of bone age range was calculated as the manually rated bone age. For AI-generated bone age, the AI software was able to determine an exact bone age for all 99 radiographs in the sample. None of the radiographs was rejected by the software.

Comparison Between Artificial Intelligence–Generated and Manual Bone Age Analysis

A strong correlation was demonstrated between AI-generated and manual bone age, with r of 0.98 and R^2 of 0.97 (p < 0.001) [Figure 2a]. The Bland-Altman analysis

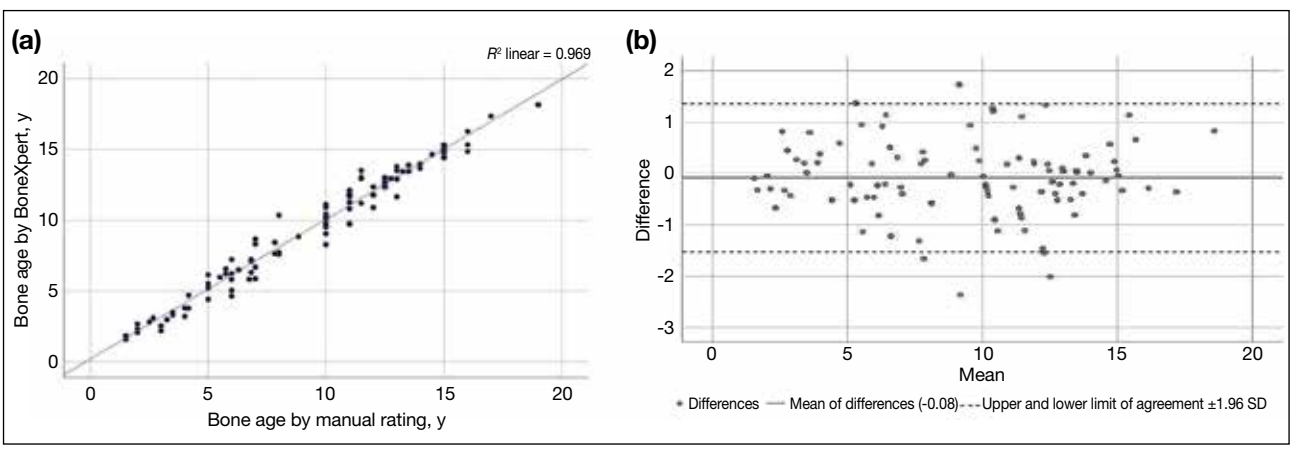

Figure 2. (a) Artificial intelligence (AI)–generated versus manual bone age correlation for the overall study population. Scatterplot showing that readings were strongly correlated (R^2 = 0.97, r = 0.98; p < 0.001). (b) Bland-Altman plot shows the difference between AI-generated and manual bone age on the vertical axis against the mean of the AI-generated and manual bone age on the horizontal axis for all patients. The solid line represents bias while the dotted line represents the upper and lower 95% limits of agreement. The mean differences in bone age between the AI-automated assessment and the manual method were -0.08 ± 0.73 years.
Abbreviation: SD = standard deviation.

also showed good agreement between manual rating and AI-generated bone age. The mean of differences was -0.08 ± 0.73 year and limits of agreement was between 1.35 and -1.51 years (Figure 2b). When stratified based on sex, the correlation between manual and AI-generated bone age assessment remained strong, with *r* of both male and female subgroups being 0.98 and R^2 being 0.97 ($p < 0.001$) [Figure 3]. The Bland-Altman bias was 0.16 ± 0.78 years in males and -0.03 ± 0.66 years in females (Figure 4). RMSE of the AI-generated bone age analysis was 0.74 year for all studies, 0.79 years for females, and 0.65 years for males.

When comparing bone ages for different study indications, a strong correlation remained between manual and AI-generated bone age. The *r* for growth disorders and miscellaneous conditions were 0.98 while that for pubertal disorders was 0.92. RMSE was best for growth disorders (0.68 year) and worst for miscellaneous conditions (0.83 year) [Figure 5].

DISCUSSION

Good agreement between the manual and AI-generated bone age rating was demonstrated in our local paediatric population in this study, with correlation remaining strong after stratification by sex. Minimal bias was detected in the Bland-Altman analysis. The small discrepancies amongst the ratings may be attributed to inclusion of the carpal bones or the presence of a sesamoid bone during manual bone age assessment, neither of which is included

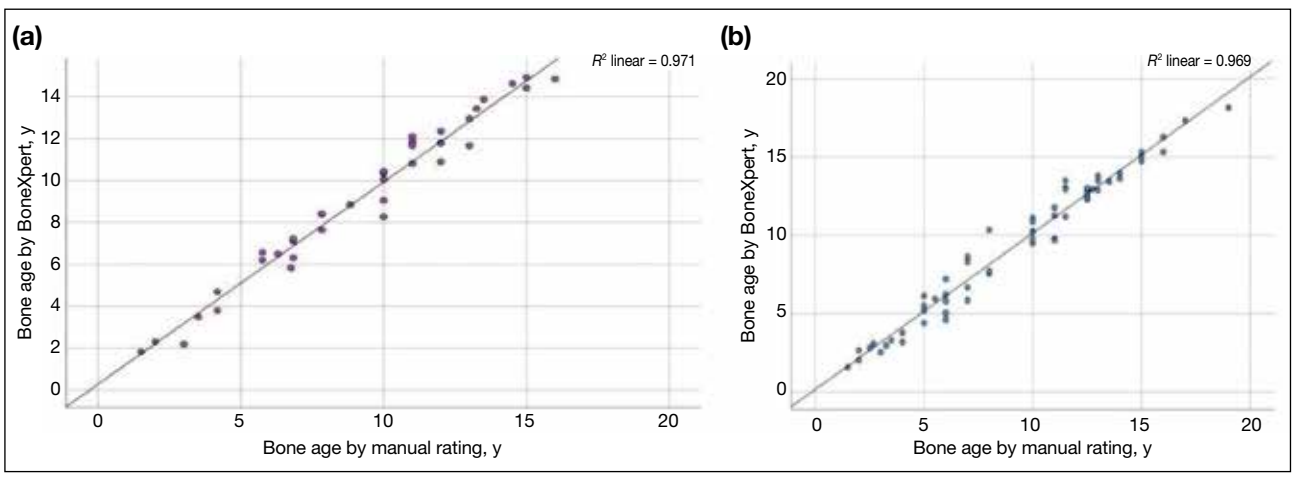

Figure 3. Scatterplots for bone age readings stratified by sex. Artificial intelligence–automated assessment and manual bone age scatterplots for females (a) and males (b) show strong correlations ($r = 0.98$ for both females and males, $R^2 = 0.97$; $p < 0.001$).

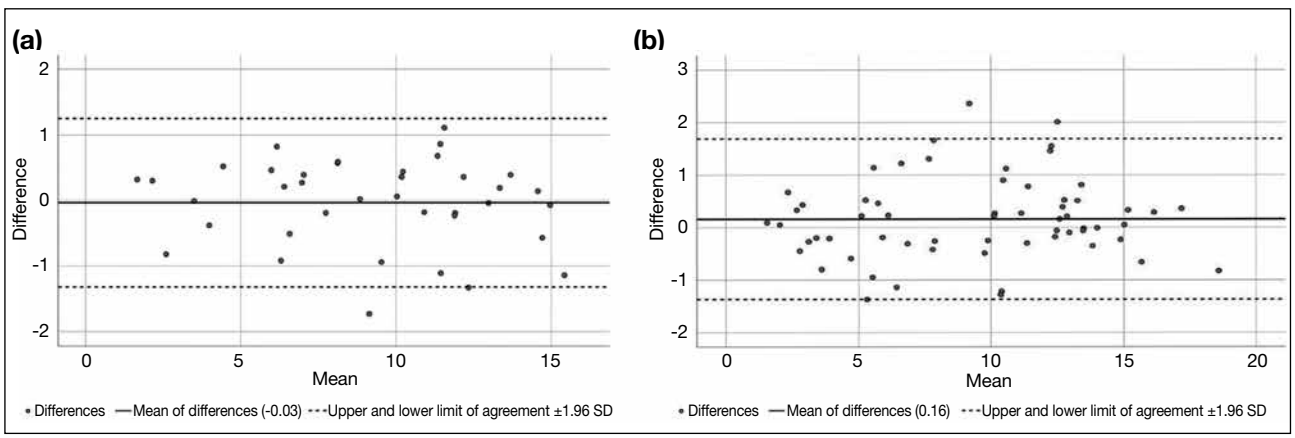

Figure 4. Bland-Altman plot of the female (a) and male (b) subgroups. The solid line represents bias while the dotted line represents the upper and lower 95% limits of agreement. The mean differences in bone age between the artificial intelligence–automated assessment and the manual method were -0.03 ± 0.66 years in females and 0.16 ± 0.78 years in males.
Abbreviation: SD = standard deviation.

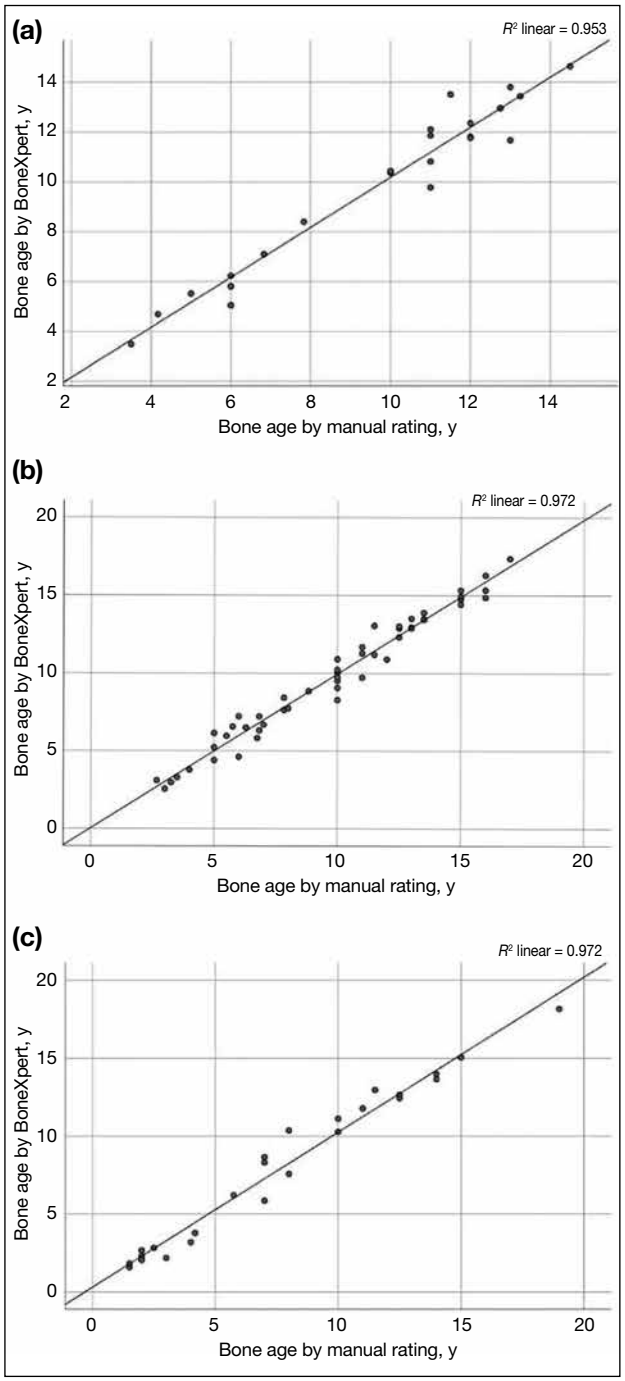

Figure 5. Scatterplots for bone age readings and study indications. Artificial intelligence–generated and manual bone age scatterplots for pubertal disorders (a), growth disorders (b), and miscellaneous conditions (c), showing strong correlation overall.

in the AI-generated assessment. These findings are similar to previous studies performed in healthy children of different races from different countries.[3,8,9,13,14]

One of the strengths of our study is that it demonstrates the high accuracy of AI-automated assessment when applied in real-life clinical practice. Many of the previous validation studies for AI-automated assessment included healthy children as subjects,[3,9,10] while the radiographs included in our study were performed in patients with pathologies clinically indicated for bone age assessment. These radiographs reflected actual clinical scenarios and the rating radiologists assessed the radiographs as part of their routine clinical practice. AI-automated assessment maintained high accuracy in our local paediatric population in Hong Kong (0.75 year) and this level of accuracy is comparable to validation studies previously published in healthy children in the Dutch population (0.71 year),[3] the Northern American population of four races (Caucasian, African-American, Hispanic and Asian) [0.74 year],[9] and in healthy Chinese (0.64 year)[10] and Japanese (0.71 year)[11] children.

The image rejection rate by the AI-automated analysis was 0% in our study. Other studies have reported an image rejection rate from 1.3% to 2%.[14,15] This very low rejection rate was likely a result of the absence of skeletal dysplasia cases in our study. As our service expands with wider clinical indications for bone age assessment, it is anticipated that there will be an increased number of rejected cases in clinical application of the software. Other studies have shown that AI-automated assessment was able to reject bone age radiographs with abnormal bone morphology and alert the reporting radiologist that an underlying metabolic or genetic bone disorder was possible, indicating the need for manual assessment.[14,15] Monitoring the rejection rate and the reasons for rejection of radiographs as part of a continuous quality improvement process would be helpful to monitor the performance of AI-automated assessment.

Limitations

There are a few limitations to our study. Firstly, the longitudinal bone age assessment of the same patient was not evaluated due to the relatively short study period. With the inherent nature of AI-automated assessment, the risk of intra- and inter-rater variability is eliminated and previous studies have proven that repeated bone age assessment by AI-automated assessment has good agreement with manual rating in terms of bone age maturation.[14] The time required for manual bone age assessment was not formally documented in our study, limiting our ability to assess how AI-automated assessment can shorten reporting time. From our experience, manual bone age determination using the GP method commonly requires around 5 to 10 minutes to determine bone age from a radiograph of the left hand

and the wrist. Compared with the almost instantaneous process of bone age determination by AI-automated assessment, this significantly shortens reporting time and improves the efficiency of radiologists. Another limitation is that our study did not compare the agreement and accuracy of AI-automated assessment against manual rating using the TW method, which is utilised in a small number of centres in our locality.

CONCLUSION

BoneXpert, an AI-automated bone age analysis algorithm, was reliable and accurate in a real-life clinical setting in our local paediatric population in Hong Kong.

REFERENCES

1. Lepe GP, Villacrés F, Silva Fuente-Alba C, Guiloff S. Correlation in radiological bone age determination using the Greulich and Pyle method versus automated evaluation using BoneXpert software [in Spanish]. Rev Chil Pediatr. 2018;89:606-11.
2. Tajmir SH, Lee H, Shailam R, Gale HI, Nguyen JC, Westra SJ, et al. Artificial intelligence–assisted interpretation of bone age radiographs improves accuracy and decreases variability. Skeletal Radiol. 2018;48:275-83.
3. van Rijn RR, Lequin MH, Thodberg HH. Automatic determination of Greulich and Pyle Bone age in healthy Dutch children. Pediatric Radiol. 2009;39:591-7.
4. Satoh M. Bone age: assessment methods and clinical applications. Clin Pediatr Endocrinol. 2015;24:143-52.
5. Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J. Bone age assessment with various machine learning techniques: a systematic literature review and meta-analysis. PLoS One. 2019;14:e0220242.
6. Thodberg HH, Thodberg B, Ahlkvist J, Offiah AC. Autonomous artificial intelligence in pediatric radiology: the use and perception of BoneXpert for bone age assessment. Pediatr Radiol. 2022;52:1338-46.
7. Martin DD, Calder AD, Ranke MB, Binder G, Thodberg HH. Accuracy and self-validation of automated bone age determination. Sci Rep. 2022;12:6388.
8. Prokop-Piotrkowska M, Marszałek-Dziuba K, Moszczyńska E, Szalecki M, Jurkiewicz E. Traditional and new methods of bone age assessment—an overview. J Clin Res Pediatr Endocrinol. 2021;13:251-62.
9. Thodberg HH, Sävendahl L. Validation and reference values of automated bone age determination for four ethnicities. Acad Radiol. 2010;17:1425-32.
10. Zhang SY, Liu G, Ma CG, Han YS, Shen XZ, Xu RL, et al. Automated determination of bone age in a modern Chinese population. ISRN Radiol. 2013;874570.
11. 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;73:398-404.
12. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64:96-106.
13. Artioli TO, Alvares MA, Carvalho Macedo VS, Silva TS, Avritchir R, Kochi C, et al. Bone age determination in eutrophic, overweight and obese Brazilian children and adolescents: a comparison between computerized BoneXpert and Greulich-Pyle methods. Pediatr Radiol. 2019;49:1185-91.
14. Bowden JJ, Bowden SA, Ruess L, Adler BH, Hu H, Krishnamurthy R, et al. Validation of automated bone age analysis from hand radiographs in a North American pediatric population. Pediatr Radiol. 2022;52:1347-55.
15. Offiah AC. Current and emerging artificial intelligence applications for pediatric musculoskeletal radiology. Pediatric Radiol. 2022;52:2149-58.