



Modernization of bone age assessment: comparing the accuracy and reliability of an artificial intelligence algorithm and shorthand bone age to Greulich and Pyle

Mina Gerges¹ · Hayley Eng² · Harpreet Chhina^{1,3} · Anthony Cooper^{3,4}

Received: 18 November 2019 / Revised: 20 March 2020 / Accepted: 23 March 2020 / Published online: 23 April 2020
© ISS 2020

Abstract

Greulich and Pyle (GP) is one of the most common methods to determine bone age from hand radiographs. In recent years, new methods were developed to increase the efficiency in bone age analysis like the shorthand bone age (SBA) and automated artificial intelligence algorithms.

Objective

The aim of this study is to evaluate the accuracy and reliability of these two methods and examine if the reduction in analysis time compromises their efficacy.

Methods

Two hundred thirteen males and 213 females had their bone age determined by two separate raters using the SBA and GP methods. Three weeks later, the two raters repeated the analysis of the radiographs. The raters timed themselves using an online stopwatch. De-identified radiographs were securely uploaded to an automated algorithm developed by a group of radiologists in Toronto. The gold standard was determined to be the radiology report attached to each radiograph, written by experienced radiologists using GP.

Results

Intraclass correlation between each method and the gold standard fell within the range of 0.8–0.9, highlighting significant agreement. Most of the comparisons showed a statistically significant difference between the new methods and the gold standard; however, it may not be clinically significant as it ranges between 0.25 and 0.5 years. A bone age is considered clinically abnormal if it falls outside 2 standard deviations of the chronological age; standard deviations are calculated and provided in GP atlas.

Conclusion

The shorthand bone age method and the automated algorithm produced values that are in agreement with the gold standard while reducing analysis time.

Keywords Bone age · Greulich and Pyle · Shorthand bone age · 16 bit

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00256-020-03429-5>) contains supplementary material, which is available to authorized users.

✉ Anthony Cooper
externalfixators@cw.bc.ca

¹ Faculty of Medicine, University of British Columbia, Vancouver, Canada

² Faculty of Science, University of British Columbia, Vancouver, Canada

³ Department of Orthopaedics, BC Children's Hospital, 1D 64 4480 Oak Street, Vancouver, BC V6H 3V4, Canada

⁴ Department of Orthopaedics, Faculty of Medicine, University of British Columbia, Vancouver, Canada

Introduction

Bone age is a radiographical assessment used in pediatric medicine due to its relative objectivity compared to chronological age [1–3]. For example, it is used in patients with scoliosis to determine if surgical interventions are still viable [4]. It is also used in conjunction with other methods to estimate unknown chronological ages [5]. In orthopedics, bone age is essential for predicting outcomes with leg length discrepancy and determining the appropriate time for an epiphysiodesis [6].

The most common way to determine bone age is through an X-ray of the left hand. There are other models that rely on various skeletal regions, though less popular [6–14]. Greulich and Pyle (GP) is one of the most commonly used methods for analyzing hand radiographs [10, 15, 16]. It employs a collection of standardized images of the left wrist and hand at different developmental stages [4, 17]. A radiologist examines the patient's X-ray and compares it collectively to radiographs in an atlas; based on the closest image's assigned age, the patient is given a bone age [1, 4, 16, 18].

In recent years, new methods were developed to increase efficiency in bone age analysis. One method known as the shorthand bone age (SBA) was developed in 2013 at Boston's children's hospital. It is based on extracting high-yield landmark from GP where the rater looks for specific features on the radiograph such as epiphyseal fusion or sesamoid bone development (Fig. 1). This method is only valid for an age range of 12.5–16 for boys and 10–14 years for girls [1]. The developers of this method claim a significant reduction in

analysis time while maintaining a substantial agreement with GP [1].

More recently, an automated algorithm developed by a group radiologists in Toronto achieved the first place in the 2017 Radiological Society of North America (RSNA) machine learning challenge to predict pediatric bone age. This method employs a deep learning algorithm that was trained on over 14,000 images graded based on GP [19]. With this method, a physician can upload an X-ray and the algorithm will estimate the bone age, automating the entire process.

The aim of this study is to evaluate the accuracy and reliability of the two methods described above and examine if the reduction in analysis time compromises their efficacy. We hypothesize that the SBA method performed by inexperienced raters and the automated algorithm will produce values in agreement with GP performed by an experienced radiologist. In addition, we hypothesize high inter-rater and intra-rater reliability of the GP and SBA methods between two inexperienced raters over two trials.

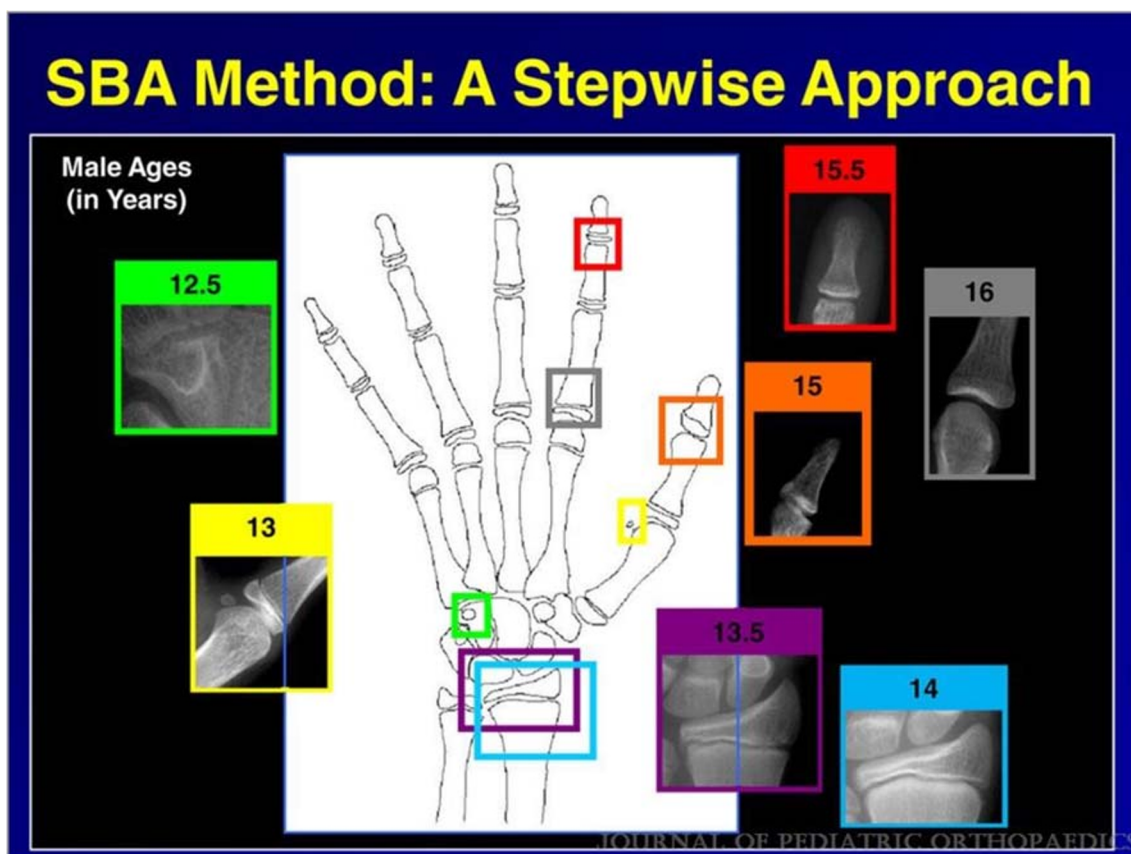


Figure 1 Sample of SBA process of determining bone age [1]. Analysis: Starting at site 1, if the landmark is present, the examiner moves on to landmark two and so on. The last landmark with a fulfilled criteria is the assigned bone age of the patient. Heyworth, Benton E.; Osei, Daniel A.; Fabricant, Peter D.; Schneider, Robert; Doyle, Shevaun M.; Green,

Daniel W.; Widmann, Roger F.; Lyman, Stephen; Burke, Stephen W.; Scher, David M. *Journal of Pediatric Orthopaedics* 33(5):569–574, July/August 2013. doi:<https://doi.org/10.1097/BPO.0b013e318293e5f2>. Image reproduced with permission from *Journal of Pediatric Orthopedics*.

Methods

Participants were selected from the BC Children’s Hospital radiology database, located in Vancouver, British Columbia, Canada. Ethics approval was obtained from the University of British Columbia research ethics board (H18-02756). Two hundred males and 200 females chronologically aged 12.5–16 and 10–14 years old, respectively, were selected. The age restrictions were put in place to match the allowed ranges by the SBA model [1]. Exclusion criteria included diagnosis or evidence of ipsilateral hand or wrist fracture within the past 2 years and diagnosis of endocrine disease that may affect stature such as growth hormone deficiency, congenital adrenal hyperplasia, elevated sex hormones, hypothyroidism, malnutrition, or chronic disease; skeletal dysplasia was also excluded.

The gold standard was determined to be the radiology report attached to each radiograph. These reports are written by experienced radiologists that routinely use GP in their practice delete rely on GPely on GP. In our study, each participant had their bone age determined by two separate raters using both the GP and SBA methods. The raters were a medical student and a science undergraduate student with no radiological or clinical training and no prior experience in bone age assessment. The GP method was performed first as it was more complex and required analysis of the entire radiograph, whereas the SBA only required examining 4–5 landmarks. This is also why the order of methods were not randomized, performing SBA before GP could have allowed the rater to only examine a few landmarks rather than examining the whole radiographs in GP. Patients whose bone age was outside the SBA age range were excluded from our statistical analysis. This reduced our sample size to 150 males and 134 females. Therefore, more radiographs were recruited using the same inclusion and exclusion criteria above until a final sample size of 213 males and 213 females was reached (Fig. 2). Two weeks later, the SBA method was used to analyze all radiographs. Three weeks later, the two raters repeated the analysis of the radiographs using the two methods in the same order. The raters timed themselves using an online stopwatch while analyzing the radiograph on a computer screen. The finally de-identified JPEG versions of the radiographs were securely uploaded to server for automated analysis.

Statistical analysis

Shapiro-Wilk and Kolmogorov-Smirnov normality tests showed that the distribution of bone ages in all methods and the times for completion were not normally distributed (supplementary table). Intraclass correlation between each method and the gold standard was performed. The median difference and interquartile ranges (IQR) between each method and the gold standard values were obtained in addition to a Wilcoxon

signed ranks test to assess statistical significance. For intra-rater variability, intraclass correlation analysis between trial 1 (T1) and trial 2 (T2) for each rater and method was performed. For inter-rater variability, intraclass correlation was performed between rater 1 (R1) and rater 2 (R2) for each method and trial. Finally, the median times for completion and IQRs for each method were obtained.

Results

Combined data

Correlation, median difference, and statistical significance

Intraclass correlation between each method and the gold standard are shown in Table 1. The automated algorithm showed the highest correlation ($\kappa = 0.880$), and R1 one using the GP method in trial 1 showed the lowest correlation ($\kappa = 0.825$) with the gold standard. The table also shows the median difference and IQR between each method and the gold standard. The methods that show a statistically significant difference are in bold in the table below ($P < 0.05$).

Time for completion

The median time for each method and the IQR are shown in Table 2. A box and whisker plot of time and variability for completion is shown in Fig. 2.

Female data

Intraclass correlation, median difference, IQR, and statistical significance

Intraclass correlation, median difference, and IQRs for females are shown in Table 3. R2 showed the highest correlation with GP in trial 2 ($\kappa = 0.703$), and R1 using GP in T1 showed the lowest correlation ($\kappa = 0.623$).

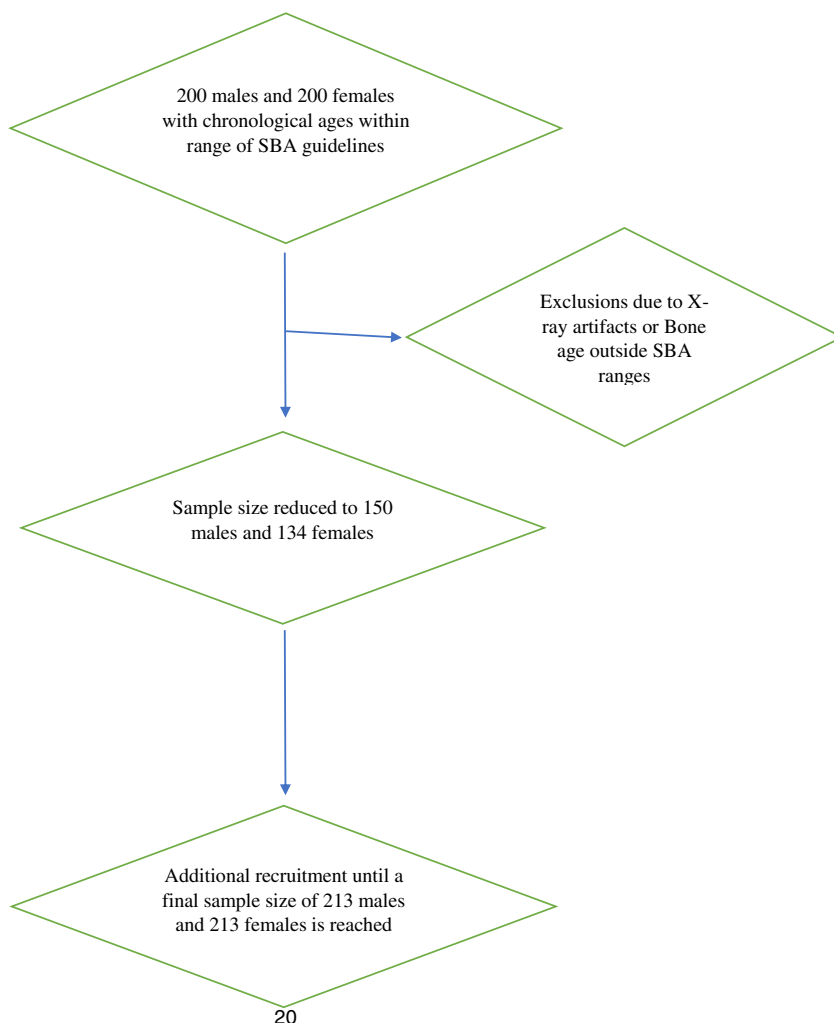
Time for completion

The median time for each method and the IQR for females are shown in Table 4.

Male data

Intraclass correlation coefficients between gold standard different methods used in this study are shown in Table 5. Median differences, IQRs, and statistical significance between each method and gold standard are also shown in Table 5. Median time for completion and IQRs are shown in Table 6.

Fig. 2 Recruitment process



Intra-rater variability

Intraclass correlation between T1 and T2 for the same method and same rater are shown in Table 7. There is an excellent correlation between T1 and T2 in all methods indicating high reliability of the two manual methods (SBA and GP).

Inter-rater variability

Correlation between the two raters in each method and trial are shown in Table 7. There is excellent correlation between both raters in each method and trial.

Table 1 Intraclass correlation, median difference, and IQR between the various methods and raters and the gold standard

Rater/method/trial	Intraclass correlation coefficient (κ)	Median (IQR) difference in years	Z score (two tailed) Wilcoxon signed ranks test
R1M1T1	0.825 ± 0.028	0.5000 (1.00)	0.224
R1M2T1	0.855 ± 0.024	0.5000 (1.00)	0.000
R2M1T1	0.871 ± 0.022	0.5000 (1.00)	0.016
R2M2T1	0.862 ± 0.023	0.5000 (1.00)	0.000
R1M1T2	0.839 ± 0.026	0.5000 (1.00)	0.001
R1M2T2	0.851 ± 0.024	0.5000 (1.00)	0.006
R2M1T2	0.876 ± 0.020	0.2500 (1.00)	0.000
R2M2T2	0.860 ± 0.023	0.5000 (1.00)	0.000
16 Bit. AI.	0.880 ± 0.020	0.3300 (0.50)	0.032

Statistically significant differences are in bold. R1, rater 1; R2, rater 2; method 1(M1), GP; M2, SBA; T, 1st or 2nd round of radiograph analysis

Table 2 The median times and IQRs for each method

Rater/method/trial	Median time (IQR) in seconds
R1M1T1	21.8300 (15.3550)
R1M2T1	7.0000 (6.1950)
R2M1T1	9.3000 (18.8500)
R2M2T1	5.0000 (2.7000)
R1M1T2	15.280000 (11.7650)
R1M2T2	5.53000 (4.0900)
R2M1T2	3.40000 (1.3000)
R2M2T2	3.8000 (0.9000)

R, rater 1 or 2; M1, GP; M2, SBA; T, 1st or 2nd round of radiograph analysis

Discussion

The first objective of this study was to examine the accuracy of the SBA method compared to the gold standard. The intra-class correlation analysis showed high correlation with $\kappa = 0.855$ for R1 and $\kappa = 0.862$ for R2 [20]. This is a stronger correlation than that reported by the developers of SBA [1]. On the other hand, the Wilcoxon signed rank test showed a statistically significant difference between the SBA values and the gold standard. The median difference between SBA and the gold standard was 0.5 years. Clinically, however, this difference may be insignificant. The GP and SBA methods provide bone ages in 0.5–1 year increments. For example, the options for females between the age of 10 and 14 are 10, 11, 12, 13, 13.5, and 14; therefore, a bone age that is 0.5 years higher or lower than the correct value may not be clinically significant. Furthermore, a bone age is considered clinically abnormal if it falls outside 2 standard deviations of the chronological age; standard deviations are calculated and provided in GP atlas [4, 17, 21]. For a 10-year-old female, 2 standard deviations constitute 21.6 months which far outweighs the difference reported here between SBA and the gold standard.

Table 4 Median time and IQR for each method in seconds for females

Rater/method/trial	Median time (IQR) in seconds
R1M1T1	18.34 (12.90)
R1M2T1	6.60 (5.11)
R2M1T1	5.6 (9.05)
R2M2T1	4.60 (1.75)
R1M1T2	13.18 (10.70)
R1M2T2	6.33 (4.57)
R2M1T2	3.30 (1.20)
R2M2T2	3.80 (1.10)

R, rater 1 or 2; M1, GP; M2, SBA; T, 1st or 2nd round of radiograph analysis

The same principle applies for male bone ages. Consequently, even though the SBA method performed by inexperienced researchers produced bone ages that are statistically different from an experienced radiologist using the GP method, these values still lie within the normal clinical ranges for bone age and therefore deeming the method efficacious. In addition, the SBA method does not require the use of an atlas making it easier to access in circumstances/areas where the GP method is not accessible [1]. On the other hand, SBA only addresses females between the ages of 10–14 and males between 12 and 16; thus, if a practitioner works with various ages, the SBA may not be most appropriate method. The SBA was designed to be an efficient, easy to use method; we found that to be true for the most part. However, the time required to use GP was not significantly longer. For example, R1 required a median of 21.83 s for GP vs. 7.00 s for SBA. R2 required 9.30 s for GP vs. 5.00 s for SBA. These time requirements were reduced as the raters gain more experience, leading to a median of 1 min or less per radiograph regardless of the method. It is important to highlight however that an average radiologist will not assess 200 consecutive bone age radiographs within a constricted age range consecutively; therefore, the numbers reported

Table 3 Intraclass correlation coefficient, median difference, and IQRs between gold standard and various analysis methods utilized in this study for females

Rater/method/trial	Intraclass correlation coefficient (κ)	Median (IQR) difference in years	Z (2 tailed) Wilcoxon signed ranks test
R1M1T1	0.623 ± 0.076	1.000 (1.00)	0.149
R1M2T1	0.650 ± 0.071	0.5000 (1.00)	0.000
R2M1T1	0.680 ± 0.066	0.5000 (1.00)	0.362
R2M2T1	0.677 ± 0.067	0.5000 (1.00)	0.133
R1M1T2	0.644 ± 0.052	1.000 (1.10)	0.000
R1M2T2	0.636 ± 0.073	0.5000 (1.00)	0.006
R2M1T2	0.703 ± 0.062	0.5000 (1.00)	0.114
R2M2T2	0.682 ± 0.066	0.5000 (1.00)	0.038
16 Bit. AI.	0.699 ± 0.063	0.4167 (0.67)	0.973

Statistically significant differences are in bold. R, rater 1 or 2; M1, GP; M2, SBA; T, 1st or 2nd round of radiograph analysis

Table 5 Intraclass correlation coefficient, median difference, and IQRs between gold standard and various analysis methods utilized in this study for males

Rater/method/trial	Intraclass correlation coefficient (κ)	Median (IQR) difference in years	Z (2 tailed) Wilcoxon signed ranks test
R1M1T1	0.861 ± 0.031	0.5000 (0.5)	0.382
R1M2T1	0.842 ± 0.035	0.5000 (0.50)	0.000
R2M1T1	0.885 ± 0.026	0.3750 (0.50)	0.006
R2M2T1	0.844 ± 0.035	0.5000 (1.00)	0.000
R1M1T2	0.877 ± 0.028	0.5000 (0.50)	0.875
R1M2T2	0.856 ± 0.033	0.5000 (0.50)	0.235
R2M1T2	0.868 ± 0.029	0.5000 (0.50)	0.001
R2M2T2	0.842 ± 0.035	0.5000 (0.50)	0.000
16 Bit. AI.	0.893 ± 0.025	0.3300 (0.41)	0.001

Statistically significant differences are in bold. R, rater 1 or 2; M1, GP; M2, SBA; T, 1st or 2nd round of radiograph analysis

above likely underestimate the time it takes to determine a single bone age. This is supported by the work of Kim et al. (2017) where it required 188 min and 22 s for a radiologist to assess 200 radiographs using GP (~0.94 min per X-ray). It required a second radiologist approximately 0.9 min per X-ray. In this study, 100 radiographs were assessed by a radiologist, then a 1-week washout period, and then the second 100 were assessed, whereas in our study, 200 radiographs for one gender were performed in one reading session leading to improved pattern recognition by the rater [22]. This is likely the cause for the discrepancy in reading time between the two studies and the overall shortened time required to assess bone age. In clinical practice, a radiologist may assess 1–2 bone age radiographs, perform other more frequently ordered tasks, then return to a few more bone age scans, etc. In this context, it is likely to require even longer time to assess a single radiograph. Overall, SBA maybe sufficient for physicians who specialize in populations within the allowed age range (i.e., adolescent medicine), but it does not provide a total replacement to GP.

An automated algorithm was chosen in this study due to its potential in diagnostic radiology and bone age in particular.

Table 6 Median time and IQR for each method in seconds for males

Rater/method/trial	Median time (IQR) in seconds
R1M1T1	26.0 (17)
R1M2T1	7.00 (7.44)
R2M1T1	19.05 (36.45)
R2M2T1	5.45 (3.65)
R1M1T2	17.04 (11.43)
R1M2T2	4.97 (3.30)
R2M1T2	3.7 (1.40)
R2M2T2	3.80 (0.80)

R, rater 1 or 2; M1, GP; M2, SBA; T, 1st or 2nd round of radiograph analysis

The algorithm provided by 16 bit relied on convolutional neural networks (CNN) and used both the pixel and sex information in the same network. For more technical details of the 16-bit algorithm, see Halabi et al. (2019) [19]. There were multiple algorithms available to conduct this study, and we chose to adopt this algorithm specifically because it achieved the best performance measured by mean absolute error (MAE) in the 2017 RSNA Machine Learning Challenge for Pediatric Bone Age. The MAE achieved was 4.265 months from the gold standard which was established by averaging the reads from 3 pediatric radiologists to the final report (for a total of 4 reads). The training and test data used in the RSNA challenge was the same used by Larson et al. (2018); their algorithm achieved a MAE of 6 months and was published prior to the challenge [23]. The training set consisted of 12,612 images from 2 US pediatric hospitals, and the test set consisted of 200 images.

Work by Lee et al. (2017) tested a variety of CNN architectures. A data set of ~8325 radiographs for training and 15% of those for testing, patients age 0–4 were excluded. They reported a female bone age within 1 year 90.39% and

Table 7 Intra-rater (first half table) and inter-rater (second half table) variability

Rater/method/trial	Intraclass correlation coefficient (κ)
R1M1T1 vs R1M1T2	0.942 ± 0.011
R1M2T1 vs R1M2T2	0.958 ± 0.008
R2M1T1 vs R2M1T2	0.955 ± 0.004
R2M2T1 vs R2M2T2	0.972 ± 0.005
R1M1T1 vs R2M1T1	0.914 ± 0.015
R1M2T1 vs R2M2T1	0.945 ± 0.009
R1M1T2 vs R2M1T2	0.919 ± 0.017
R1M2T2 vs R2M2T2	0.953 ± 0.009

R, rater 1 or 2; M1, GP; M2, SBA; T, 1st or 2nd round of radiograph analysis

within 2 years 98.11% of the time and a male bone age within 1 year 94.18% and within 2 years 99.00% of the time [24]. Similarly, Kim et al. (2017) used a training set of 18,940 images and a test set of 200 images; they achieved a root mean square error (RMSE) of 0.60 years whereas 16 bit achieved RMSE of 0.47 years [22]. Finally, the work by Mutasa et al. (2018) showed a MAE of 0.561 years which is higher than that achieved by 16 bit [25].

In this study, the automated algorithm showed good correlation ($\kappa = 0.880$) and a median difference of 0.33 years (~4 months) from the gold standard which coincides with the results on the RSNA challenge test set [21]. The difference between this method and the gold standard was statistically significant. Like SBA, this difference is not clinically significant for the same reasons explained above. Considering the good correlation between this method and the gold standard and the 4 months median difference, this algorithm can be considered efficacious, at least for age range examined in this study. There are additional benefits to this process that are not available for SBA and GP. It does not have age restrictions and provides an opportunity to automate the entire process. A technician can simply upload the image to the server and obtain the bone age within seconds. It is unlikely that this method can eliminate the role of the radiologist completely; however, it can streamline the patient's journey. Currently, a patient may see a physician who refers them for an X-ray. There, an image of the hand is taken and sent to the radiologist who analyzes it and issues a report. With the use of AI, the physician can get the bone age once the patient gets their X-ray taken which can speed up the diagnostic process and improve the patient's quality of care. One additional benefit to using the 16-bit algorithm compared to other AI algorithms is that it produces an output in months, allowing for bone age values outside the ages set by GP, i.e., 11.5 years, 15.8 years, etc. This increases the accuracy of assessment and is especially useful when the bone age does not match a specific standard.

Upon examining the male and female data separately, the male values correlate with the gold standard similarly to the combined data; however, the female values show lower correlation coefficients. The SBA method yielded values of $\kappa = 0.650$ for rater 1 and $\kappa = 0.677$ for rater 2 in females. Similarly, method 3 produced a correlation coefficient of 0.699. In addition, the median difference of the SBA from the gold standard in females was 0.68 (~8 months) for rater 1, 0.6 (~7 months) for rater 2, and 0.65 years (~8 months) for method 3, which are higher than the values provided by the combined and male data. One explanation is that there are fewer options of ages for females compared to males in the age ranges allowed in this study. For males between the ages of 12.5 and 16, there are 7 options to choose from, whereas for females between the ages of 10 and 14, there are 6 options. In that sense, there is one less bone age a rater can use to analyze

a radiograph. We examined the correlation between method 3 and all the other methods used in this study (except the gold standard) for females, and it was within the range of 0.8–0.9. This could indicate that the gold standard itself has its own limitations and variability [1]. In recent years, there has been evidence of changes in puberty patterns in females over the last century [26, 27]. Future studies should explore the efficacy of GP on current populations and whether it needs to be adjusted.

Intra-rater variability and inter-rater variability

A secondary objective of this study was to examine the intra-rater variability in each method. The correlation between the two trials in each method and rater (i.e., R1M1T1 vs. R1M1T2) was excellent, between the range of 0.9 and 1. This confirms the reliability of these methods. Similarly, with inter-rater variability, we compared each method between the two raters, and we obtained excellent correlation within the range of 0.9–1. This highlights the limited variability between raters who may use these methods.

Limitations

This study relied on GP performed by experienced licensed pediatric radiologists to be the gold standard but did not account for variability between radiologists, although all were experienced radiologists working in an academic pediatric hospital. Due to time, resources, and logistical limitations, this was not possible to obtain a large enough sample size of radiographs assessed by only one radiologist. As a result, we opted to rely on all radiologists as a gold standard. In addition, there is evidence of increased variability in GP with ethnically diverse populations or in children with musculoskeletal conditions [1, 15, 28–30]. The efficacy of the automated algorithm was tested within a constricted age range defined by SBA. Examining a wider range was the original goal of this study as it would have allowed to assess the full capability of the 16-bit algorithm. However, it would have required 2 comparison parameters, making the study overly complex. The logical expansion for this research group is to examine the efficacy of 16-bit algorithm across all ages compared to the gold standard. The raters timing themselves could have also created a bias which could have been minimized by using a software for timing; however, this was not possible due to resource limitation.

Only 2 methods were utilized in this study, and compared to GP, there are multiple models (manual and automated) that have been developed in recent years that are worth examining [1, 31–34]. Tanner-Whitehouse III (TW) is another method that is widely used to assess hand bone age. Originally, TW was included in this study; however, the original text describing this method is out of print and inaccessible; therefore, it

was removed [35]. Finally, in this study, 2 raters and 213 males and females were recruited. A higher sample size can be beneficial in establishing better evidence for these methods, and more raters can be useful in understanding the inter-rater variability within each method.

Conclusion

The shorthand bone age method and the automated algorithm produced values that are in agreement with the gold standard Greulich and Pyle while reducing analysis time and maintaining a high inter-rater and intra-rater reliability.

Compliance with ethical standards

Conflict of interest Dr. Mark Cicero from 16 Bit Inc. The algorithm adopted in this study is the intellectual property of 16 Bit Inc. Members of the company assisted with the use of the algorithm for the purpose of the study; however, none of the authors are affiliated with the company, nor there are any financial association with the study.

Ethical approval Ethics approval was obtained from the University of British Columbia's research ethics board (H18-02756)

References

- Heyworth BE, Osei DA, Fabricant PD, et al. The shorthand bone age assessment: a simpler alternative to current methods. *J Pediatr Orthop*. 2013;33(5):569–74. <https://doi.org/10.1097/BPO.0b013e318293e5f2>.
- Bass S, Pearce G, Bradney M, et al. Exercise before puberty may confer residual benefits in bone density in adulthood: studies in active prepubertal and retired female gymnasts. *J Bone Miner Res*. 1998;13:500–7.
- 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. <https://doi.org/10.1159/000329372>.
- Satoh M. Bone age: assessment methods and clinical applications. 2015. *Clin Pediatr Endocrinol*. 2015;24(4):143–52. Published online 2015 Oct 24. <https://doi.org/10.1297/cpe.24.143>.
- Martin DD, Wit JM, Hochberg Z, et al. The use of bone age in clinical practice - part 2. *Horm Res Paediatr*. 2011;76(1):10–6. <https://doi.org/10.1159/000329374>.
- Makarov MR, Jackson TJ, Smith CM, Jo CH, Birch JG. Timing of epiphysiodesis to correct leg-length discrepancy: a comparison of prediction methods. *J Bone Joint Surg Am*. 2018;100(14):1217–22. <https://doi.org/10.2106/JBJS.17.01380>.
- Diméglio A, Charles YP, Daures JP, de Rosa V, Kaboré B. Accuracy of the Sauvegrain method in determining skeletal age during puberty. *J Bone Joint Surg Am*. 2005;87(8):1689–96.
- Bitan FD, Veliskakis KP, Campbell BC. Differences in the Risser grading systems in the United States and France. *Clin Orthop Relat Res*. 2005;436:190–5. <https://doi.org/10.1097/01.blo.0000160819.10767.88>.
- Wittschieber D, Vieth V, Domnick C, Pfeiffer H, Schmeling A. The iliac crest in forensic age diagnostics: evaluation of the apophyseal ossification in conventional radiography. *Int J Legal Med*. 2013;127(2):473–9. <https://doi.org/10.1007/s00414-012-0763-x>.
- Schmidt S, Schmeling A, Zwiesigk P, Pfeiffer H, Schulz R. Sonographic evaluation of apophyseal ossification of the iliac crest in forensic age diagnostics in living individuals. *Int J Legal Med*. 2011;125(2):271–6. <https://doi.org/10.1007/s00414-011-0554-9>.
- Mughal AM, Hassan N, Ahmed A. Bone age assessment methods: a critical review. *Pak J Med Sci*. 2014;30(1):211–5. <https://doi.org/10.12669/pjms.301.4295>.
- Su P, Zhang L, Peng Y, Liang A, Du K, Huang D. A histological and ultrastructural study of femoral head cartilage in a new type II collagenopathy. *Int Orthop*. 2010;34(8):1333–9. <https://doi.org/10.1007/s00264-010-0985-9>.
- Kaur G, Khandelwal N, Jasuja OP. Computed tomographic studies on ossification status of medial epiphysis of clavicle: effect of slice thickness and dose distribution. *J Indian Acad Forensic Med*. 32(4).
- Schmidt S, Mühler M, Schmeling A, Reisinger W, Schulz R. Magnetic resonance imaging of the clavicular ossification. *Int J Legal Med*. 2007;121(4):321–4.
- Hoerr NL. Radiographic atlas of skeletal development of the knee. Springfield: Charles C. Thomas; 1955.
- Zafar AM, Nadeem N, Husen Y, Ahmad MN. An appraisal of Greulich-Pyle atlas for skeletal age assessment in Pakistan. *J Pak Med Assoc*. 2010;60(7):552–5.
- Gaskin CM, Kahn SL, Bertozzi JC, Bunch PM. Skeletal development of the hand and wrist: a radiographic atlas and digital bone age companion: a radiographic atlas and digital bone age companion. Oxford: Oxford University Press; 2011.
- Greulich WW, Pyle SI. Radiograph atlas of skeletal development of the hand and wrist. 2nd ed. Palo Alto: Stanford University Press; 1959.
- Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA pediatric bone age machine learning challenge. *Radiology*. 2019;290(2):498–503. <https://doi.org/10.1148/radiol.2018180736>.
- Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. 2012;24(3):69–71.
- Nwosu BU, Lee MM. Evaluation of short and tall stature in children. *Am Fam Physician*. 2008;78(5):597–604.
- 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):1374–80.
- 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):313–22.
- Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. *J Digit Imaging*. 2017;30(4):427–41.
- Mutasa S, Chang PD, Ruzal-Shapiro C, Ayyala R. MABAL: a novel deep-learning architecture for machine-assisted bone age labeling. *J Digit Imaging*. 2018;31(4):513–9.
- Kaplowitz PB, Slora EJ, Wasserman RC, Pedlow SE, Herman-Giddens ME. Earlier onset of puberty in girls: relation to increased body mass index and race. *Pediatrics*. 2001;108(2):347–53.
- Herman-Giddens ME, Steffes J, Harris D, et al. Secondary sexual characteristics in boys: data from the pediatric research in office settings network. *Pediatrics*. 2012;130(5):e1058–68. <https://doi.org/10.1542/peds.2011-3291>.
- Ontell FK, Ivanovic M, Ablin DS, Barlow TW. Bone age in children of diverse ethnicity. *Am J Roentgenol*. 1996;167:1395.
- Loder RT, Estle DT, Morrison K, et al. Applicability of the Greulich and Pyle skeletal age standards to black and white children of today. *Am J Dis Child*. 1993;147:1329–33.
- Zhang A, Sayre JW, Vachon L, et al. Racial differences in growth patterns of children assessed on the basis of bone age. *Radiology*. 2009;250:228–35.
- Martin DD, Neuhof J, Jenni OG, et al. Automatic determination of left- and right-hand bone age in the first Zurich longitudinal study. *Horm Res Paediatr*. 2010;74:50–5.

32. Thodberg HH. Clinical review: an automated method for determination of bone age. *J Clin Endocrinol Metab.* 2009;94:2239–44.
33. Thodberg HH, Jenni OG, Caflisch J, et al. Prediction of adult height based on automated determination of bone age. *J Clin Endocrinol Metab.* 2009;94:4868–74.
34. Thodberg HH, Kreiborg S, Juul A, et al. The BoneXpert method for automated determination of skeletal maturity. *IEEE Trans Med Imaging.* 2009;28:52–66.
35. Tanner JM. Assessment of skeletal maturity and prediction of adult height (TW3 method). 3rd ed. London: W.B. Saunders; 2001.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.