Imaging Science in Dentistry 2020; 50: 237-43
https://doi.org/10.5624/isd.2020.50.3.237

Evaluation of the clinical efficacy of a TW3-based fully automated bone age assessment system using deep neural networks

Nan-Young Shin[1], Byoung-Dai Lee[2,3], Ju-Hee Kang[1], Hye-Rin Kim[1], Dong Hyo Oh[2], Byung Il Lee[2], Sung Hyun Kim[2], Mu Sook Lee[4], Min-Suk Heo[1,*]

[1]*Department of Oral and Maxillofacial Radiology and Dental Research Institute, School of Dentistry, Seoul National University, Seoul, Korea*
[2]*Center for Artificial Intelligence in Medicine and Imaging, HealthHub, Seoul, Korea*
[3]*Division of Computer Science and Engineering, Kyonggi University, Suwon, Korea*
[4]*Department of Radiology, Keimyung University, Dongsan Hospital, Daegu, Korea*

ABSTRACT

Purpose: The aim of this study was to evaluate the clinical efficacy of a Tanner-Whitehouse 3 (TW3)-based fully automated bone age assessment system on hand-wrist radiographs of Korean children and adolescents.

Materials and Methods: Hand-wrist radiographs of 80 subjects (40 boys and 40 girls, 7-15 years of age) were collected. The clinical efficacy was evaluated by comparing the bone ages that were determined using the system with those from the reference standard produced by 2 oral and maxillofacial radiologists. Comparisons were conducted using the paired *t*-test and simple regression analysis.

Results: The bone ages estimated with this bone age assessment system were not significantly different from those obtained with the reference standard ($P>0.05$) and satisfied the equivalence criterion of 0.6 years within the 95% confidence interval (−0.07 to 0.22), demonstrating excellent performance of the system. Similarly, in the comparisons of gender subgroups, no significant difference in bone age between the values produced by the system and the reference standard was observed ($P>0.05$ for both boys and girls). The determination coefficients obtained via regression analysis were 0.962, 0.945, and 0.952 for boys, girls, and overall, respectively ($P=0.000$); hence, the radiologist-determined bone ages and the system-determined bone ages were strongly correlated.

Conclusion: This TW3-based system can be effectively used for bone age assessment based on hand-wrist radiographs of Korean children and adolescents. ***(Imaging Sci Dent 2020; 50: 237-43)***

KEY WORDS: Age Determination by Skeleton; Radiography; Deep Learning; Artificial Intelligence

Introduction

In clinical trials aimed at evaluating the growth of children and adolescents, chronological age is not considered to be a reliable indicator due to individual variation of maturational patterns.[1] Of the various approaches for growth evaluation, bone age assessment (BAA) using hand-wrist radiographs is most commonly applied due to its simplicity and inexpensiveness, the availability of many ossification centers, and the low radiation exposure involved.[2]

The Greulich-Pyle (GP),[3] Tanner-Whitehouse 3 (TW3),[4] and Fishman[5] methods are generally used for BAA based on hand-wrist radiographs. Among them, the TW3 method calculates the bone age by classifying and scoring the developmental stages of regions in the radius, ulna, and short bones (RUS); summing these scores; and finding the corresponding age in a sum-score table.[4] Figure 1 shows the 13 short bones in the hand and wrist observed in the RUS scoring system of the TW3 method. Advantages of this method include relatively high accuracy and reproducibility.

Nan Young Shin and Byoung-Dai Lee contributed equally to this work as the first authors.
This work received financial support from HealthHub (No. 860-20190072).
Received March 25, 2020; Revised May 15, 2020; Accepted May 21, 2020
*Correspondence to : Prof. Min-Suk Heo
Department of Oral and Maxillofacial Radiology, School of Dentistry, Seoul National University, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea
Tel) 82-2-2072-3016, E-mail) hmslsh@snu.ac.kr

Imaging Science in Dentistry · pISSN 2233-7822 eISSN 2233-7830

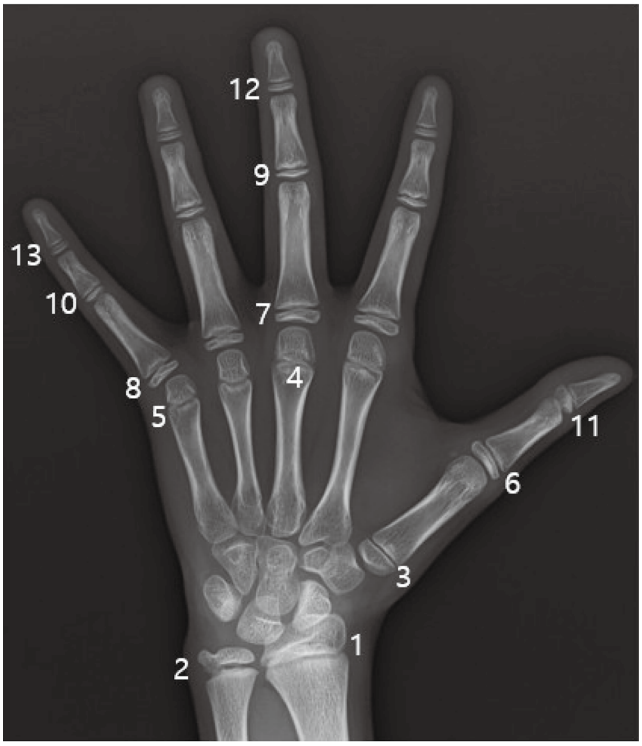

Fig. 1. Thirteen short bones in the hand and wrist observed in the scoring system based on the radius, ulna, and short bones.[4] 1. Radius. 2. Ulna. 3. First metacarpus. 4. Third metacarpus. 5. Fifth metacarpus. 6. First proximal phalanx 7. Third proximal phalanx. 8. Fifth proximal phalanx. 9. Third middle phalanx. 10. Fifth middle phalanx. 11. First distal phalanx. 12. Third distal phalanx. 13. Fifth distal phalanx.

The TW3 method consists of 2 main processes: the extraction of the 13 regions of interest (ROIs) on a hand-wrist radiograph and the evaluation of the skeletal maturity level of each extracted ROI. Complete automation of these processes and acceptable BAA accuracy require a high level of computing technique, which is hard to achieve. Previous automated BAA solutions based on the TW3 method have had limitations, including the need for radiologists to manually extract the ROI, the application of extra ROI features, and the usage of additional information such as bone age distribution data associated with ethnicities or countries. The differences between the bone ages estimated by these automated solutions and those obtained from the corresponding reference standards have ranged from 0.8 to 0.9 years.[6-8]

The recent rapid development of deep-learning technology based on artificial neural networks has led to the expansion of its applications, particularly in the context of medical imaging analysis. A level of performance comparable to that of radiologists has been achieved in some studies, such as the detection of diabetic retinopathy in photographs of the retinal fundus,[9] the classification of skin cancer,[10] lung cancer screening,[11] and breast cancer screening.[12] BAA is also considered an ideal target of object detection and classification using deep-learning technology. In particular, convolutional neural networks (CNNs) and their variants are being increasingly used to automate BAA, and they have shown promising results.[2,13-15]

However, the majority of existing deep learning-based BAA systems are based on the GP method and are potentially vulnerable to the low repeatability of measurements and the systematic errors that are inherent to the GP method.[16]

Therefore, a TW3-based BAA system using deep neural networks was developed to automate the entire process, from the localization of the 13 epiphysis-metaphysis growth regions to the output of the estimated bone age.[16] The software was trained to use the TW3 method to automatically analyze hand-wrist radiographs entered in the form of image files and to present bone ages in 0.1 years. It was aimed to utilize the BAA system to provide more efficient evaluation of skeletal maturity in clinical practice.

Accordingly, this study was performed to evaluate the clinical efficacy of this TW3-based BAA system by comparing bone ages measured with the system with those measured by 2 oral and maxillofacial radiologists.

Materials and Methods

This study complied with the management standards of medical device clinical trials and the fundamental principles of the Declaration of Helsinki in conducting the test and evaluating and recording its results. The institutional review board of Seoul National University Dental Hospital approved this retrospective study and waived the requirement to obtain informed consent.

Sample collection

Digital left hand-wrist radiographs were retrospectively and randomly selected from the picture archiving and communication system (PACS) at Seoul National University Dental Hospital. All of the radiographs were taken between 2012 and 2017 for the purpose of growth evaluation related to orthodontic treatment. The chronological age of the subjects ranged from 7 to 15 years old; this age was calculated by subtracting the birth date of the subject from the date on which the radiograph was taken. Considering that the maximum bone age interpretable with

the TW3 method is 15 years (for girls) and 16.5 years (for boys) and that the prediction could be unreliable in cases of complete fusion of the radius and ulna, the upper limit of the sample chronological age was set at 15 years.

The exclusion criteria were as follows: 1) systemic disease such as developmental or endocrinological disorders, 2) bony abnormalities of the hands and wrists due to trauma or disease, and 3) inappropriate radiographs (poor image quality, poor positioning, or patient movement).

A total of 80 radiographs (40 from boys and 40 from girls) were collected for this study. The sample size required to satisfy conditions set for the consistency test was calculated using PASS 2019 software (NCSS, LLC, Kaysville, UT, USA). Table 1 shows the sample distribution according to gender and age.

Acquisition and observation of hand-wrist radiographs

All of the hand-wrist radiographs were taken with a REX 650R device (Listem Co., Ltd., Wonju, Korea) under a protocol of 50 kV and 8 mAs. An FCR XG5000 apparatus (Fujifilm, Tokyo, Japan) was used for image acquisition. The images were obtained and visualized without patient information using Infinitt® PACS software (Infinitt Healthcare Co. Ltd., Seoul, South Korea) with tools such as window width/level adjustment and zoom. All radiographs were evaluated on a diagnostic display screen (Nio Color 2MP LED 21.3-inch monitor with 1200-1600 resolution; BARCO, Kortrijk, Belgium) in a quiet room under dim lighting conditions.

Reference standard

Two observers, oral and maxillofacial radiologists with 4 and 7 years of experience, assessed the bone ages from the 80 selected hand-wrist radiographs using the TW3 method. The estimation was performed twice by each observer, with estimates separated by a 3-week interval. The observers were unaware of each other's assessments and their first estimation during the second assessment, and they were similarly unaware of the measurements produced by the BAA system. In the event of a disagreement, the final reference standard was established by a consensus reached through discussion.

Table 1. Sample distribution according to age and gender

Age (years)	Male	Female	Overall
7	2	6	8
8	7	4	11
9	4	7	11
10	8	5	13
11	6	6	12
12	6	6	12
13	1	2	3
14	2	1	3
15	4	3	7
Total	40	40	80

Bone age assessment by the system

The TW3-based fully automated BAA system was developed based on 2 CNNs: Faster-R-CNN, which is the region-based CNN for the extraction of actual ROIs from bounding ROIs, and VGGNet-BA CNN, used for classification of the skeletal maturity level of an ROI. Hand-wrist radiographs of 3,027 Korean male and female children and adolescents under 18 years old, labeled by 2 radiologists based on the TW3 method, were used to train the system. The details of the system have been previously described.[16] After a hand-wrist radiograph (in the JPG file format) was entered into the BAA software and a rough area containing 13 ROIs was selected using a computer mouse, assessment was activated. Upon completion of the process after a few seconds, the predicted bone age was displayed along with skeletal maturity ratings of each of the 13 ROIs (Fig. 2).

Statistical analysis

Cohen kappa coefficients were calculated to evaluate the reliability of the reference standard. These coefficients were interpreted according to the definitions shown in Table 2. Using the paired *t*-test, the primary efficacy of the developed BAA system was evaluated via a comparison between the bone ages from the reference standard and those estimated with the BAA system ($P<0.05$). An upper and lower limit of ± 0.6 years in the 95% confidence interval was set as the equivalence criterion. The second-

Table 2. Cohen kappa (κ) coefficient definitions

κ	Meaning
$\kappa<0$	No agreement
$0\leq\kappa\leq0.2$	Slight agreement
$0.2<\kappa\leq0.4$	Fair agreement
$0.4<\kappa\leq0.6$	Moderate agreement
$0.6<\kappa\leq0.8$	Substantial agreement
$0.8<\kappa\leq1.0$	Almost perfect agreement

Fig. 2. Screenshot example of the bone age assessment (BAA) report presented by the TW3-based fully automated BAA system.

ary efficacy evaluation was performed by comparison between the bone ages from the reference standard and those estimated with the BAA system in the gender subgroups, also using the paired *t*-test ($P<0.05$). In addition, correlations between the bone ages from the reference standard and those estimated with the BAA system were evaluated using simple regression analysis. For statistical calculations, IBM SPSS Statistics version 23 (SPSS Corp., Armonk, NY, USA) was used.

Results

The intra-observer reliability values were 0.846 (for observer 1) and 0.817 (for observer 2), with almost perfect agreement. Regarding inter-observer reliability, substantial agreement levels of 0.737, 0.763, and 0.750 were found for boys, girls, and overall, respectively. These values indicated that the reference standard was sufficiently reliable. The kappa coefficients for the 13 ROIs are shown in Table 3.

Table 4 shows the difference between the bone ages from the reference standard and those obtained with the BAA system. This difference was assessed using the paired *t*-test. No statistically significant difference was found between the bone ages from the reference standard and those obtained with the BAA system for the gender

Table 3. Intra- and inter-observer reliability (as indicated by kappa coefficients) by region of interest (ROI)

ROI	Intra-observer (observer 1)	Intra-observer (observer 2)	Inter-observer
Ulna	0.830	0.819	0.726
Radius	0.807	0.533	0.545
Third distal phalanx	0.814	0.931	0.766
Third middle phalanx	0.827	0.921	0.736
Third proximal phalanx	0.919	0.937	0.968
Third metacarpus	0.808	0.709	0.762
Fifth distal phalanx	0.792	0.902	0.730
Fifth middle phalanx	0.875	0.891	0.692
Fifth proximal phalanx	0.906	0.829	0.784
Fifth metacarpus	0.782	0.765	0.765
First distal phalanx	0.816	0.751	0.817
First proximal phalanx	0.935	0.919	0.854
First metacarpus	0.892	0.709	0.604
Overall	0.846	0.817	0.750

Table 4. Analysis of difference in bone ages (BAs) between the reference standard and the bone age assessment (BAA) system conducted via the paired *t*-test

Gender	Mean BA (years)		Mean difference[§] (years)	Significance probability	Upper and lower limits of difference in BA**
	Reference standard	BAA system			
Male	10.87 ± 3.12	11.06 ± 2.94	0.19 ± 0.62	0.060*	−0.01, 0.39
Female	11.29 ± 2.78	11.26 ± 2.74	−0.04 ± 0.65	0.737*	−0.24, 0.17
Overall	11.08 ± 2.94	11.16 ± 2.82	0.08 ± 0.64	0.285*	−0.07, 0.22

*: $P<0.05$, **: 95% confidence interval, §: Mean difference = (BA obtained by the system) − (BA from the reference standard)

subgroups or for the overall group. In addition, the upper and lower limits of the difference in bone age satisfied the equivalence criterion of 0.6 years in the 95% confidence interval.

For the simple regression analysis, in which the bone age obtained with the BAA system was the independent variable and the bone age from the reference standard was the dependent variable, the correlation coefficients were 0.981, 0.972, and 0.976 and the determination coefficients were 0.962, 0.945, and 0.952 for boys, girls, and overall, respectively. These values indicate that 96.2%, 94.5%, and 95.2% of the reference standard-based bone ages for boys, girls, and the overall cohort could be explained by the bone ages determined with the BAA system. Moreover, a statistically significant correlation was found between the bone ages from the BAA system and those from the reference standard ($P=0.000$). The linear equations obtained via regression analysis were $y=-0.629+1.040x$, $y=0.201+0.985x$, and $y=-0.253+1.016x$ for boys, girls, and overall, respectively. The correlation graphs for boys, girls, and overall are shown in Figure 3.

Discussion

Bone age estimation from hand-wrist radiographs is currently in the most general use among various age assessment methods. The GP method involves the evaluation of bone age in direct comparison with bone age-labeled reference images in an atlas. This method is widely used in many cases (including fully automated BAA systems that have been previously reported), since it is relatively uncomplicated in that the whole hand is simply observed based on atlas images; however, it can be inefficient and lack reproducibility for the same reason. In contrast, the TW3 method outperforms the GP method

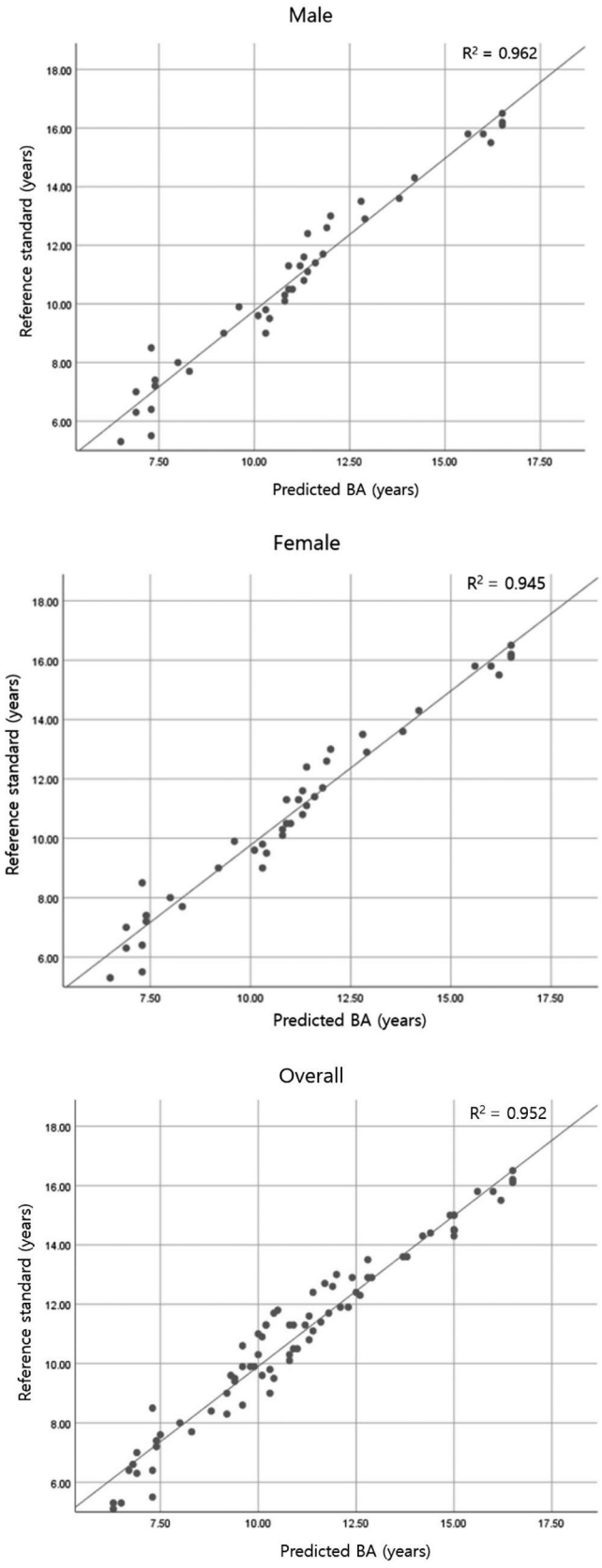

Fig. 3. Correlations between bone ages (BAs) estimated by the bone age assessment system and those from the reference standard.

in accuracy and reproducibility, since it is more complex and elaborative to evaluate the skeletal maturity levels of specific bones from the hand and wrist area, compute the score based on maturity levels, and convert the score to a bone age using a correlation matrix between maturity scores and bone ages. The BAA system in this study is the first TW3-based fully automated system using deep CNNs for ROI extraction and skeletal maturity evaluation.

In this study, we found no significant difference between the bone ages determined with the BAA system and those obtained with the reference standard in the overall group or the gender subgroups. The system demonstrated excellent performance by satisfying the equivalence criterion of 0.6 years in the 95% confidence interval: −0.07 to 0.22 years in overall, −0.01 to 0.39 years in boys, and −0.24 to 0.17 years in girls. Given that existing TW3-based BAA systems that were not fully automated have produced estimation errors of 0.8-0.9 years[6], and that 0.42 years is the lowest value that has been reported among GP-based fully automated BAA systems,[17] the system in this study can be concluded to be reliable and have excellent accuracy for BAA. The regression analysis also showed excellent determination coefficients and linear regression equations with significant probability, demonstrating a statistically significant and very high correlation between bone ages obtained with the system and those from the reference standard.

Bone age determination plays essential roles in various fields, including growth evaluation in paediatrics or orthodontics, identification in forensic medicine, and legal proceedings. For instance, significant growth deviation may indicate endocrine or genetic disorders as well as psychosocial problems. Meanwhile, the optimal timing and device for orthodontic treatment can be determined with reference to the bone age rather than the chronological age. In this regard, this BAA system could be very efficiently utilized for more rapid and accurate age prediction and skeletal maturity estimation.

The reliability of the reference standard is one of the important factors involved in evaluating the clinical efficacy of a system. Cohen kappa coefficients in this study showed good intra- and inter-observer reliability. The kappa values for the ROIs were at or above the level of substantial agreement except for the radius, which was associated with the lowest kappa value (less than 0.6). it is necessary to perform a very detailed and careful calibration of the observers prior to the evaluation. In addition, more clear and unambiguous definition of each stage of

the ROIs may be helpful.

This study had some limitations. First, this study was retrospective, involving 80 radiographs from a single institution. Since the conditions (such as hand positioning) under which the radiographs were taken were not strictly controlled, it was sometimes difficult to observe the developmental status of specific regions, such as the middle phalanges of the fingers, due to overlapping or superimposition. However, the image quality was generally acceptable. Additionally, since the test subjects consisted of individuals of a single race only, future research should be conducted in multiple institutions and include a multi-racial sample of subjects. Second, since X-ray images in infants aged 0-6 years are not commonly obtained in non-emergency situations due to concerns about radiation exposure, further studies will be needed to expand the range of automated bone age prediction.

In conclusion, this study demonstrated that this BAA system can be effectively used for TW3-based BAA from hand-wrist radiographs of Korean children and adolescents aged 7-15 years.

Conflicts of Interest: Two of the authors (Byoung-Dai Lee and Byoung-Il Lee) developed the TW3-based BAA system at HealthHub (Seoul, Korea) and are still working for the company.

References

1. Benjavongkulchai S, Pittayapat P. Age estimation methods using hand and wrist radiographs in a group of contemporary Thais. Forensic Sci Int 2018; 287: 218.e1-8.
2. 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.
3. Gilsanz V, Ratib O. Hand bone age: A digital atlas of skeletal maturity. Berlin: Springer; 2005.
4. Tanner JM, Healy MJR, Cameron N, Goldstein H. Assessment of skeletal maturity and prediction of adult height (TW3 method). Philadelphia: W. B. Saunders; 2001.
5. Fishman LS. Radiographic evaluation of skeletal maturation. A clinically oriented method based on hand-wrist films. Angle Orthod 1982; 52: 88-112.
6. Tristan-Vega A, Arribas JI. A radius and ulna TW3 bone age assessment. IEEE Trans Biomed Eng 2008; 55: 1463-76.
7. Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging 2009; 28: 52-66.
8. Liu J, Qi J, Liu Z, Ning Q, Luo X. Automatic bone age assessment based on intelligent algorithms and comparison with TW3 method. Comput Med Imaging Graph 2008; 32: 678-84.
9. Gulshan, V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316: 2402-10.
10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115-8.
11. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019; 25: 954-61.
12. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020; 577: 89-94.
13. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017; 30: 427-41.
14. Kim JR, Shim WH, Yoon HM, Hong SH, Lee JS, Cho YA, et al. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. AJR Am J Roentgenol 2017; 209: 1374-80.
15. Ren X, Li T, Yang X, Wang S, Ahmad S, Xiang L, et al. Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. IEEE J Biomed Health Inform 2019; 23: 2030-8.
16. Son SJ, Song Y, Kim N, Do Y, Kwak N, Lee MS, et al. TW3-based fully automated bone age assessment system using deep neural networks. IEEE Access 2019; 7: 33346-58.
17. Iglovikov VI, Rakhlin A, Kalinin AA, Shvets AA. Paediatric bone age assessment using deep convolutional neural networks. In: Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T, Martel A, Maier-Hein L, et al. Deep learning in medical image analysis and multimodal learning for clinical decision support. Cham: Springer; 2018. p. 300-8.