Submitted: 10 April, 2023 Accepted: 12 May, 2023 Published: 03 November, 2023 DOI:10.22514/jocpd.2023.087

ORIGINAL RESEARCH

Journal of Clinical Pediatric Dentistry

Commercial artificial intelligence lateral cephalometric analysis: part 2—effects of human examiners on artificial intelligence performance, a pilot study

Jaesik Lee[1,†], Seong-Ryeol Bae[2,†], Hyung-Kyu Noh[2,*]

[1]Department of Pediatric Dentistry, School of Dentistry, Kyungpook National University, 41940 Daegu, Republic of Korea
[2]Department of Orthodontics, School of Dentistry, Kyungpook National University, 41940 Daegu, Republic of Korea

***Correspondence**
hknoh@knu.ac.kr
(Hyung-Kyu Noh)

† These authors contributed equally.

Abstract

At the current technology level, a human examiner's review must be accompanied to compensate for the insufficient commercial artificial intelligence (AI) performance. This study aimed to investigate the effects of the human examiner's expertise on the efficacy of AI analysis, including time-saving and error reduction. Eighty-four pretreatment cephalograms were randomly selected for this study. First, human examiners (one beginner and two regular examiners) manually detected 15 cephalometric landmarks and measured the required time. Subsequently, commercial AI services automatically identified these landmarks. Finally, the human examiners reviewed the AI landmark determination and adjusted them as needed while measuring the time required for the review process. Then, the elapsed time was compared statistically. Systematic and random errors among examiners (human examiners, AI and their combinations) were assessed using the Bland-Altman analysis. Intraclass correlation coefficients were used to estimate the inter-examiner reliability. No clinically significant time difference was observed regardless of AI use. AI measurement error decreased substantially after the review of the human examiner. From the standpoint of the human examiner, beginners could obtain better results than manual landmarking. However, the AI review outcomes of the regular examiner were not as good as those of manual analysis, possibly due to AI-dependent landmark decisions. The reliability of AI analysis could also be improved by employing the human examiner's review. Although the time-saving effect was not evident, commercial AI cephalometric services are currently recommendable for beginners.

Keywords

Cephalometric; Artificial intelligence; Efficacy; Accuracy; Precision; Reliability

1. Introduction

Automatic cephalometric landmark detection has attracted considerable research attention in dentistry [1]. With the introduction of artificial intelligence (AI) in this field in the past few years, the success rate of automated landmark identification is rapidly increasing [2–9]. Moreover, commercial AI-supported automatic cephalometric services have recently been launched. Accordingly, the accuracy and efficiency of cephalometric analysis are expected to improve considerably [1].

The primary clinical efficacy expected from automatic landmark identification is a decrease in the landmarking error and the required time by eliminating human examiner intervention [10, 11]. However, as indicated in our previous study (Part 1), which assessed the performance of commercial AI cephalometric services, an inspection of each landmark position by a human examiner was an indispensable requisite for commercial AI services. In other words, a human examiner could not be excluded from these commercial AI-supported cephalometric analyses for now.

The question is what benefits can be gained from AI-supported cephalometric analysis when accompanied by a human examiner's review. Whether time-saving and error-reduction effects can be expected even if less-experienced examiners, such as beginners in pediatric dentistry and orthodontic fields, review the AI results is unclear. Thus far, the mainstream of AI research has been to evaluate the performance of AI architectures [2–9]. The effect of a human examiner's expertise on the final results of the AI analysis has not yet been investigated. The effectiveness of these AI services under practical usage conditions is clinically relevant for clinicians because several commercial AI cephalometric services have already been released to the market with an increasing number of users.

This study aimed to evaluate the clinical efficacy of a commercial AI cephalometric analysis followed by a human examiner's inspection. The time-saving effect, agreement and

reliability of AI-assisted cephalometric analysis will be investigated among experienced and inexperienced human examiners, AI and combinations thereof.

2. Materials and methods

Eighty-four pretreatment cephalogram images were randomly selected from patients who visited the Department of Orthodontics and Pediatric Dentistry of Kyungpook National University Dental Hospital between 2012 and 2021 for the treatment of malocclusion. Patients with a history of orthodontic treatment or craniofacial malformations such as cleft lip and palate were excluded. All cephalogram images, taken using a CX-90SP (an X-ray scanner, Asahi, Kyoto, Japan), were in JPG format with a resolution of 150 DPI and a gray level of 24. Table 1 presents the characteristics of the sample.

TABLE 1. Sample characteristics.

Characteristics	N	Mean	SD
Age (yr)	84	11.13	3.52
Sex			
Male	46	-	-
Female	38	-	-
AP skeletal (ANB angle)			
Class I	35	1.76	1.11
Class II	24	5.83	1.38
Class III	25	−1.42	1.42
Vertical skeletal (SN-MP angle)			
Normal angle	53	33.18	2.55
High angle	27	40.42	2.94
Low angle	4	22.48	4.39

N, the number of samples; SD, standard deviation; Class I, $0 < ANB < 4$; Class II, $4 \leq ANB$; Class III, $ANB \leq 0$; Normal angle, $27 < SN\text{-}MP < 37$; High angle, $37 \leq SN\text{-}MP$; Low angle, $SN\text{-}MP \leq 27$.

Three human examiners participated in this study. Experts 1 (HKN) and 2 (SRB) are board-certified orthodontists with more than seven and five years of clinical experience, respectively, belonging to regular examiners. The beginner examiner (JSL) is a board-certified pediatric dentist with minimal expertise in tracing cephalograms, *i.e.*, less than 40 cases over the seven years of his career. Thirteen commonly used cephalometric variables were evaluated after identifying 15 dental and skeletal landmarks (Tables 2 and 3).

The manual landmarking process was as follows: First, human examiners discussed and agreed upon the landmark definition using three cephalogram images that were not included in the study samples. Then, the human examiners initiated manual landmark identification using computer software (6.3 Sequential Tracing Mode, AudaxCeph, Ljubljana, Slovenia). Detection was performed independently, without any communication between examiners. There was no time limit for landmarking, and re-examining landmark positions was always possible until the examiners were satisfied. Expert 2 and the beginner examiner measured the time required from the first landmark identification to the final approval of the overall landmark position under this condition. Repeated measurements were performed by expert 1 after one month.

TABLE 2. Cephalometric landmark definitions.

Landmarks	Definition
S	The center point of the sella turcica.
Na	The uppermost point of the frontonasal suture.
Po	The uppermost point of the external acoustic meatus.
Or	The lowermost point of the bony orbit.
Ar	The intersection of the cranial base and the posterior margin of the neck of condyles.
A-point	The most concave point of the curve between the anterior nasal spine and the most anterior-inferior point of the upper alveolar bone.
B-point	The most concave point of the curve between the most anterior-superior point of the lower alveolar bone and the most anterior point of the bony contour of the chin.
Go	The most posterior and inferior point of the angle of the mandible.
Me	The most inferior point of the bony contour of the chin.
Incisor point	The midpoint between U1 and L1 tips.
Molar point	The point where the upper and lower first molars occlude. The landmark was determined by the midpoint between the mesiobuccal cusp tips of the upper and lower first molars.
U1 tip	The incisal tip of the upper incisors.
U1 apex	The root apex point of the upper incisors.
L1 tip	The incisal tip of the lower incisors.
L1 apex	The root apex point of the lower incisors.

S, sella; Na, nasion; Po, porion; Or, orbitale; Ar, articulare; Go, gonion; Me, menton; U1, upper central incisor; L1, lower central incisor.

TABLE 3. Measures of intra-examiner reliability and method errors.

Variables	Expert 1		
	Coefficient	95% CI	Dahlberg
SNA	0.95	(0.93, 0.97)	0.79
SNB	0.97	(0.96, 0.98)	0.68
ANB	0.98	(0.97, 0.99)	0.46
Wits	0.98	(0.97, 0.99)	0.71
SN-MP	0.98	(0.97, 0.99)	0.77
FMA	0.97	(0.95, 0.98)	0.84
Bjork-Jarabak Sum	0.98	(0.97, 0.99)	0.77
SN-U1	0.98	(0.97, 0.99)	1.22
FH-U1	0.98	(0.97, 0.99)	1.22
IMPA	0.97	(0.96, 0.98)	1.27
U1L1	0.98	(0.97, 0.99)	1.68
SN-OcP	0.95	(0.93, 0.97)	0.92
FH-OcP	0.94	(0.91, 0.96)	0.98

ICC, intraclass correlation coefficient; CI, 95% confidence interval of ICC; Sig, significance; Dahlberg, method errors obtained by Dahlberg's formula ($\sqrt{\sum d^2/2n}$).

Commercial AI cephalometric service (WebCeph, 1.0.0, Assemblecircle, Gyeonggi-do, Korea) was used to obtain AI landmarking data. WebCeph automatically detected the landmark position and performed cephalometric analysis immediately after uploading the study samples. The landmark data from WebCeph were saved and delivered to expert 2 and the beginner examiner. Each examiner independently reviewed the positions of the 15 landmarks involved in this study. The review process was initiated 1 month after the previous manual detection. As in manual landmarking, each landmark position was examined and adjusted as necessary without any time limit. The examiners also measured the time required for the review process from the first landmark examination to the final approval.

The overall study design is shown in Fig. 1.

All statistical analyses were performed using the language R (4.3.1, R Foundation for Statistical Computing, Vienna, Austria) with a significance level of 0.05.

The intra-examiner reliability between the repeated measurements of expert 1 was evaluated with intraclass correlation coefficients (ICCs) using two-way mixed-effects, single rater and absolute agreement models [12]. The Dahlberg formula estimated method errors between repeated trials.

The time required for landmark identification was compared among expert 2, beginner, WebCeph and expert 2 and WebCeph and beginner. A one-way repeated-measures analysis of variance was used for comparison. Subsequent *post hoc* pairwise comparisons were performed using the Bonferroni correction.

The Bland-Altman analysis was used to evaluate measurement errors among examiners. Expert 1's data were set as the reference. Then, according to the Bland-Altman protocol, data from other examiners were analyzed relative to this reference [13]. Specifically, the mean and difference between expert 1 and other examiners were calculated. Data normality was confirmed using the Shapiro-Wilk test, and the Bland-Altman statistics was evaluated. The bias, which is the mean difference between examiners, measures the systematic error. On the contrary, the limits of agreement (LoA) are the upper and lower bounds containing 95% of the measurement errors between examiners [14]. As an index of pure random error size between examiners, the maximum random error (MRE) was defined as the half-width of the upper and lower LoA, eliminating the effect of bias included in LoA [13, 15]. To help visually understand the magnitudes of systematic and random errors, the Bland-Altman plots were drawn.

The number of variables that satisfied the agreement criteria was counted for each examiner. We applied the same criteria for agreement determination based on the rationale described in our previous study (Part 1).

The inter-examiner reliability between expert 1 and other examiners was estimated with ICCs using the two-way random-effects and single-rater models [12]. We calculated both absolute agreement and consistency ICCs to assess the impact of

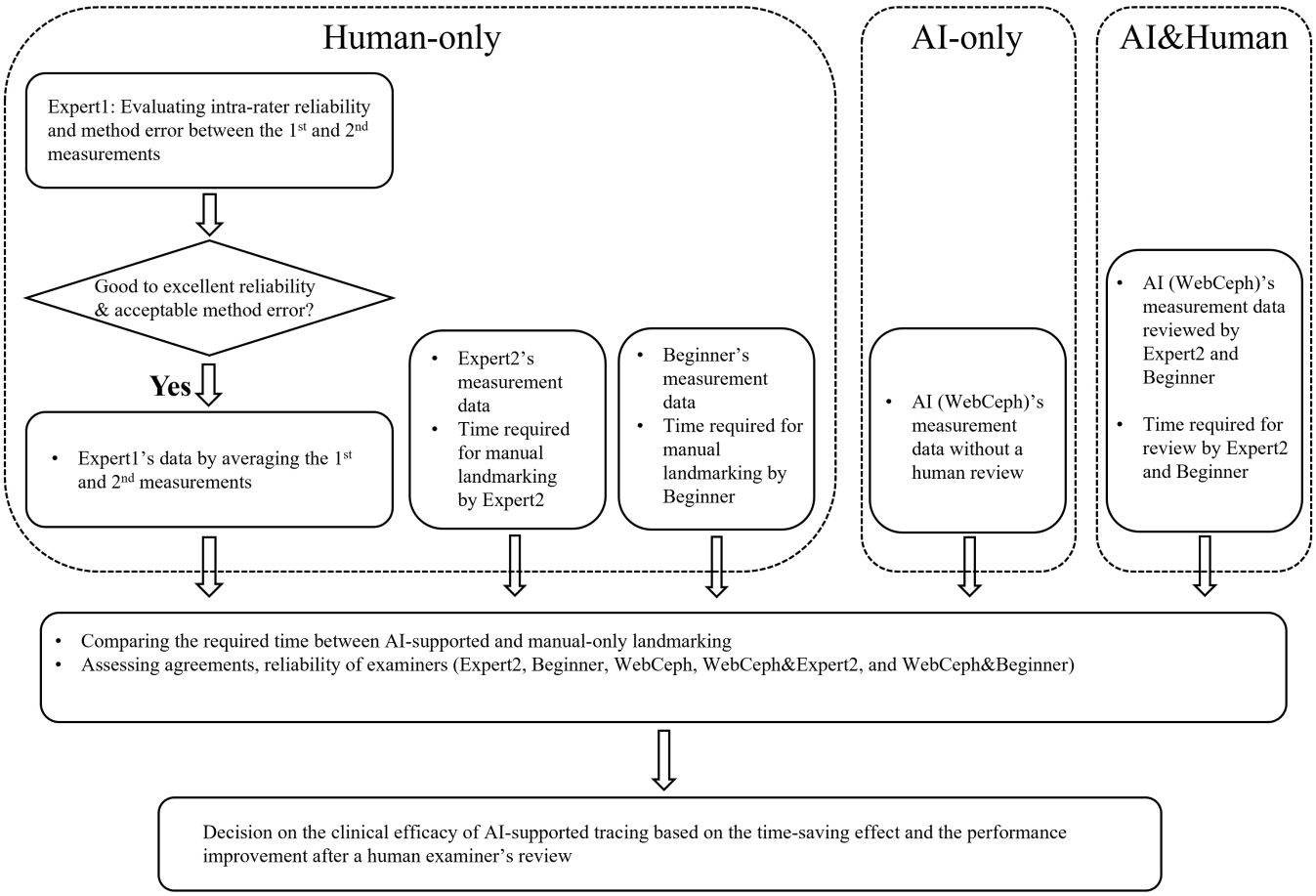

FIGURE 1. Study flow chart. AI, artificial intelligence.

systematic errors.

3. Results

The mean intra-examiner reliability of expert 1 was 0.97, and the mean method errors were 0.95° and 0.73 mm for angular and linear measurements, respectively (Table 3). Therefore, we took the mean of the first and second measurements and used these average values in the subsequent analysis.

Table 4 presents the time required for landmarking. The results of Mauchly's test confirmed that the sphericity assumption was violated ($p < 0.001$). Hence, we report the test results adjusted by Greenhouse and Geisser ($\epsilon = 0.736$). According to the analysis, at least one measurement time was significantly different: F (2.208, 218.592) = 229, $p < 0.001$. The subsequent *post-hoc* test showed significant differences in all comparison pairs ($p < 0.001$). Specifically, when comparing the required time between examiners according to the landmarking methods, expert 2 took significantly longer than the beginner for both manual landmarking and AI review. However, no consistency was observed when comparing each examiner's required time between the landmarking methods. Expert 2 took approximately 30 s more for AI review than manual landmarking, whereas the beginner showed approximately 10 s shorter time in the same case.

TABLE 4. The *post hoc* pairwise comparison of the required time.

	Expert 2	Beginner	Mean difference	*p*-value
Manual landmarking	75.27 ± 20.39	59.71 ± 20.62	15.56	<0.001
AI review	107.80 ± 27.25	47.54 ± 10.72	60.26	<0.001
Mean difference	−32.53	12.17	-	-
p-value	<0.001	<0.001	-	-

The values are mean ± standard deviation in the unit of seconds; p-values were adjusted with Bonferroni correction; AI, artificial intelligence.

The descriptive and Bland-Altman statistics of cephalometric variables measured by each examiner are presented in Tables 5 and 6. Statistically significant biases with confidence intervals not containing zero were common among examiners; however, their magnitudes were generally small. Exceptionally, the sella-nasion plane (SN plane)-associated variables (SNA, SNB, SN-MP, Björk-Jarabak sum, SN-U1 and SN-OcP) measured by the beginner showed relatively large systematic errors.

The magnitudes of random errors showed a clear pattern among examiners: expert 2 < WebCeph and expert 2 ≤ WebCeph and beginner < WebCeph < beginner (Fig. 2). The MRE of the beginner was 2.19 times larger than that of expert 2 on average. Similarly, WebCeph, WebCeph and beginner, and WebCeph and expert 2 showed 1.78, 1.35 and 1.18 times greater MREs than expert 2 (Table 6 and Fig. 2). As an illustrative example, the Bland-Altman plots of the SN-MP are shown in Fig. 3.

In this study, most of the examiners did not meet the agreement criteria (Table 7). Only WebCeph and expert 2 could satisfy the standard over three variables (Wits, IMPA and U1L1).

The mean inter-examiner reliability between expert 1 and expert 2 was 0.95 for absolute agreement and 0.96 for consistency (Table 8). Similar but slightly low ICC values were observed in WebCeph and expert 2 and WebCeph and beginner, *i.e.*, 0.91–0.94, corresponding to excellent ICCs. By contrast, Webceph showed relatively compromised ICCs between 0.84 and 0.86. Finally, the beginner revealed the lowest ICC values of 0.78–0.82, with a large difference between absolute agreement and consistency.

4. Discussion

Time-saving may be one of the major clinical benefits that clinicians expect from using AI-supported cephalometric analysis [10, 11]. However, contrary to usual expectations, the statistical comparison of the elapsed time showed no consistent trend among examiners, *i.e.*, that the beginner could save approximately 10 s on average, whereas expert 2 spent 30 s longer (Table 4). This is probably because the landmark position review process may negate the time saved by AI and can take more in some cases. In addition, regardless of whether it increased or decreased, the time difference between the measurement methods was <1 min, which was clinically insignificant. Therefore, contrary to the usual expectations, obtaining a clinically relevant time-saving effect from AI cephalometric analysis may be challenging as long as the examination by a human examiner is essential.

The overall systematic error was not clinically significant. However, the SN plane-associated variables measured by the beginner showed substantial systematic errors (Table 6). This is an example of a subjective error that has long been reported as a major error source for landmark detection [8, 10, 16]. Interestingly, this systematic error disappeared in WebCeph and beginner (Table 6, Fig. 3), supporting that subjective error reduction is achievable even after a human review.

In Part 1, we revealed that the random error size of WebCeph was clinically unacceptable. A human examiner's review has been suggested as a countermeasure against the unreliable performance of AI [17, 18]. The study results showed that the MRE of WebCeph decreased when reviewed by a human examiner, regardless of their expertise level (Table 6, Fig. 2). These results may justify AI review by a human examiner as a method compensating for AI's performance.

However, these advantages of AI review were not equally applicable to examiners with different expertise levels. Compared with manual landmarking, the average MRE decreased by 2.63 for beginners with AI aid, whereas expert 2 conversely increased by 0.57, albeit slightly (Table 6, Fig. 2). The number of variables meeting the agreement condition also illustrated the same points. The performance of expert 2 became worse with the help of Webceph (Table 7).

TABLE 5. Descriptive statistics.

	Expert 1	Expert 2	Beginner	WebCeph	WebCeph_Beginner	WebCeph_Expert2
SNA	80.60 ± 3.62	80.08 ± 3.45	82.85 ± 5.10	81.92 ± 3.09	81.33 ± 3.38	80.42 ± 3.42
SNB	78.52 ± 4.16	78.29 ± 4.05	80.88 ± 5.48	78.76 ± 3.57	78.54 ± 3.90	77.78 ± 3.99
ANB	2.06 ± 3.06	1.79 ± 3.13	1.97 ± 3.50	3.16 ± 3.17	2.79 ± 3.09	2.64 ± 3.10
Wits	−2.28 ± 5.11	−2.99 ± 5.06	−3.28 ± 5.31	−0.88 ± 5.03	−1.30 ± 5.56	−1.40 ± 5.26
SN-MP	35.12 ± 5.24	35.64 ± 5.13	32.86 ± 5.70	34.29 ± 4.89	35.46 ± 5.09	36.03 ± 5.11
FMA	26.43 ± 4.77	25.38 ± 4.56	27.60 ± 4.91	26.48 ± 4.67	26.48 ± 4.87	25.89 ± 4.76
Bjork-Jarabak Sum	395.12 ± 5.23	395.64 ± 5.13	392.86 ± 5.70	394.29 ± 4.89	395.46 ± 5.09	396.03 ± 5.11
SN-U1	106.86 ± 9.16	106.38 ± 8.86	111.47 ± 11.41	105.14 ± 8.48	105.65 ± 8.71	105.34 ± 9.22
FH-U1	115.55 ± 8.73	116.64 ± 8.76	116.69 ± 10.31	112.95 ± 8.15	114.63 ± 8.47	115.47 ± 9.05
IMPA	92.98 ± 7.82	92.31 ± 7.80	94.33 ± 8.18	92.16 ± 6.58	90.40 ± 6.95	90.67 ± 7.66
U1L1	125.05 ± 13.01	125.68 ± 13.21	121.38 ± 13.71	128.41 ± 11.13	128.50 ± 12.31	127.96 ± 13.36
SN-OcP	17.81 ± 4.20	19.00 ± 4.01	16.38 ± 5.53	16.99 ± 3.92	17.18 ± 4.22	18.41 ± 3.99
FH-OcP	9.12 ± 3.99	8.74 ± 3.74	11.16 ± 4.97	9.18 ± 3.57	8.20 ± 4.10	8.28 ± 3.81

The values are mean ± standard deviation.

TABLE 6. Bland-Altman statistics.

	Expert 1-Expert 2		Expert 1-Beginner		Expert 1-WebCeph		Expert 1-WebCeph_Expert2		Expert 1-WebCeph_Beginner	
	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI
SNA										
Bias	0.52	(0.27, 0.76)	−2.26	(−3.06, −1.45)	−1.33	(−1.92, −0.73)	0.18	(−0.16, 0.51)	−0.73	(−1.05, −0.41)
Upper LoA	2.74	(2.32, 3.16)	5.02	(3.64, 6.40)	4.06	(3.04, 5.08)	3.18	(2.61, 3.75)	2.19	(1.64, 2.75)
Lower LoA	−1.71	(−2.13, −1.28)	−9.54	(−10.92, −8.15)	−6.71	(−7.73, −5.69)	−2.83	(−3.40, −2.26)	−3.65	(−4.21, −3.10)
MRE	2.22		7.28		5.39		3.00		2.92	
SNB										
Bias	0.23	(0.04, 0.43)	−2.36	(−3.12, −1.61)	−0.24	(−0.71, 0.23)	0.74	(0.48, 1.01)	−0.02	(−0.29, −0.25)
Upper LoA	2.00	(1.67, 2.34)	4.47	(3.17, 5.77)	4.00	(3.20, 4.81)	3.13	(2.68, 3.59)	2.38	(1.93, 2.84)
Lower LoA	−1.54	(−1.87, −1.20)	−9.20	(−10.49, −7.90)	−4.48	(−5.28, −3.67)	−1.65	(−2.10, −1.19)	−2.42	(−2.88, −1.96)
MRE	1.77		6.83		4.24		2.39		2.40	
ANB										
Bias	0.27	(0.13, 0.42)	0.09	(−0.24, 0.42)	−1.10	(−1.42, −0.78)	−0.58	(−0.78, −0.38)	−0.72	(−0.94, 0.50)
Upper LoA	1.59	(1.34, 1.84)	3.08	(2.51, 3.64)	1.81	(1.26, 2.36)	1.21	(0.87, 1.55)	1.25	(0.87, 1.62)
Lower LoA	−1.05	(−1.30, −0.80)	−2.89	(−3.45, −2.32)	−4.01	(−4.56, −3.46)	−2.37	(−2.71, −2.03)	−2.69	(−3.07, −2.32)
MRE	1.32		2.98		2.91		1.79		1.97	
Wits										
Bias	0.71	(0.40, 1.02)	1.00	(0.58, 1.42)	−1.40	(−1.87, −0.94)	−0.89	(−1.20, −0.57)	−0.99	(−1.42, 0.56)
Upper LoA	3.52	(2.98, 4.05)	4.77	(4.05, 5.49)	2.78	(1.98, 3.57)	1.96	(1.42, 2.50)	2.90	(2.16, 3.64)
Lower LoA	−2.10	(−2.63, −1.57)	−2.77	(−3.49, −2.06)	−5.58	(−6.37, −4.79)	−3.73	(−4.27, −3.19)	−4.87	(−5.61, −4.14)
MRE	2.81		3.77		4.18		2.85		3.89	
SN-MP										
Bias	−0.52	(−0.73, −0.31)	2.26	(1.49, 3.03)	0.83	(0.33, 1.32)	−0.91	(−1.22, −0.60)	−0.34	(−0.72, 0.04)
Upper LoA	1.38	(1.02, 1.74)	9.20	(7.88, 10.52)	5.32	(4.47, 6.17)	1.88	(1.35, 2.41)	3.11	(2.45, 3.76)
Lower LoA	−2.42	(−2.78, −2.06)	−4.69	(−6.01, −3.37)	−3.67	(−4.52, −2.81)	−3.70	(−4.23, −3.17)	−3.79	(−4.44, −3.13)
MRE	1.90		6.95		4.49		2.79		3.45	

TABLE 6. Continued.

	Expert 1-Expert 2		Expert 1-Beginner		Expert 1-WebCeph		Expert 1-WebCeph_Expert2		Expert 1-WebCeph_Beginner	
	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI
FMA										
Bias	1.05	(0.81, 1.30)	−1.17	(−1.73, −0.60)	−0.05	(−0.38, 0.29)	0.54	(0.20, 0.88)	−0.05	(−0.41, −0.31)
Upper LoA	3.27	(2.84, 3.69)	3.96	(2.98, 4.93)	2.99	(2.41, 3.57)	3.60	(3.01, 4.18)	3.22	(2.60, 3.84)
Lower LoA	−1.16	(−1.58, −0.74)	−6.29	(−7.26, −5.31)	−3.08	(−3.66, −2.50)	−2.52	(−3.10, −1.94)	−3.31	(−3.93, −2.69)
MRE	2.21		5.12		3.03		3.06		3.26	
Sum										
Bias	−0.52	(−0.73, −0.31)	2.26	(1.49, 3.03)	0.83	(0.33, 1.33)	−0.91	(−1.22, −0.60)	−0.34	(−0.72, 0.04)
Upper LoA	1.38	(1.02, 1.74)	9.21	(7.89, 10.53)	5.32	(4.46, 6.17)	1.88	(1.35, 2.41)	3.11	(2.46, 3.76)
Lower LoA	−2.42	(−2.78, −2.06)	−4.69	(−6.01, −3.37)	−3.66	(−4.51, −2.81)	−3.70	(−4.23, −3.17)	−3.78	(−4.44, −3.13)
MRE	1.90		6.95		4.49		2.79		3.45	
SN-U1										
Bias	0.48	(0.01, 0.95)	−4.61	(−5.81, −3.40)	1.73	(0.69, 2.76)	1.52	(0.86, 2.19)	1.22	(0.49, −1.94)
Upper LoA	4.73	(3.92, 5.54)	6.26	(4.20, 8.32)	11.06	(9.29, 12.84)	7.54	(6.40, 8.69)	7.76	(6.52, 9.01)
Lower LoA	−3.76	(−4.57, −2.96)	−15.47	(−17.53, −13.41)	−7.61	(−9.38, −5.84)	−4.50	(−5.64, −3.35)	−5.33	(−6.58, −4.09)
MRE	4.25		10.87		9.34		6.02		6.55	
FH-U1										
Bias	−1.09	(−1.64, −0.54)	−1.14	(−2.19, −0.10)	2.60	(1.67, 3.53)	0.08	(−0.60, 0.76)	0.93	(0.23, −1.62)
Upper LoA	3.90	(2.95, 4.85)	8.31	(6.52, 10.11)	11.00	(9.40, 12.59)	6.23	(5.06, 7.40)	7.20	(6.01, 8.39)
Lower LoA	−6.08	(−7.02, −5.13)	−10.59	(−12.39, −8.80)	−5.79	(−7.39, −4.20)	−6.07	(−7.24, −4.90)	−5.34	(−6.54, −4.15)
MRE	4.99		9.45		8.39		6.15		6.27	

TABLE 6. Continued.

	Expert 1-Expert 2		Expert 1-Beginner		Expert 1-WebCeph		Expert 1-WebCeph_Expert2		Expert 1-WebCeph_Beginner	
	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI
IMPA										
Bias	0.67	(0.17, 1.18)	−1.35	(−2.15, −0.56)	0.81	(−0.02, 1.64)	2.31	(1.80, 2.81)	2.58	(1.92, −3.24)
Upper LoA	5.23	(4.37, 6.10)	5.81	(4.45, 7.18)	8.30	(6.88, 9.72)	6.85	(5.99, 7.71)	8.57	(7.43, 9.71)
Lower LoA	−3.89	(−4.76, −3.03)	−8.52	(−9.88, −7.16)	−6.68	(−8.10, −5.26)	−2.24	(−3.10, −1.38)	−3.41	(−4.55, −2.27)
MRE	4.56		7.17		7.49		4.54		5.99	
U1L1										
Bias	−0.62	(−1.29, 0.04)	3.67	(2.63, 4.72)	−3.36	(−4.37, −2.34)	−2.91	(−3.54, −2.29)	−3.45	(−4.32, 2.58)
Upper LoA	5.41	(4.26, 6.55)	13.11	(11.32, 14.90)	5.79	(4.06, 7.53)	2.73	(1.66, 3.80)	4.42	(2.93, 5.92)
Lower LoA	−6.66	(−7.80, −5.51)	−5.77	(−7.56, −3.97)	−12.51	(−14.25, −10.77)	−8.55	(−9.63, −7.48)	−11.32	(−12.82, −9.82)
MRE	6.03		9.44		9.15		5.64		7.87	
SN-OcP										
Bias	−1.19	(−1.57, −0.80)	1.43	(0.65, 2.21)	0.82	(0.18, 1.46)	−0.60	(−1.01, −0.19)	0.64	(0.21, 1.06)
Upper LoA	2.27	(1.62, 2.93)	8.45	(7.11, 9.78)	6.60	(5.50, 7.69)	3.13	(2.42, 3.83)	4.44	(3.72, 5.17)
Lower LoA	−4.64	(−5.30, −3.99)	−5.59	(−6.92, −4.25)	−4.95	(−6.05, −3.85)	−4.33	(−5.03, −3.62)	−3.17	(−3.89, −2.45)
MRE	3.46		7.02		5.77		3.73		3.81	
FH-OcP										
Bias	0.38	(−0.04, 0.80)	−2.03	(−2.70, −1.37)	−0.05	(−0.55, 0.44)	0.84	(0.41, 1.28)	0.92	(0.50, −1.35)
Upper LoA	4.17	(3.45, 4.89)	3.96	(2.82, 5.09)	4.43	(3.58, 5.28)	4.76	(4.01, 5.50)	4.79	(4.05, 5.52)
Lower LoA	−3.40	(−4.12, −2.68)	−8.02	(−9.16, −6.89)	−4.54	(−5.39, −3.69)	−3.07	(−3.81, −2.33)	−2.94	(−3.67, −2.20)
MRE	3.78		5.99		4.48		3.91		3.86	

Upper CI, upper limit of 95% confidence interval; lower CI, lower limit of 95% confidence interval; LoA, limit of agreement; MRE, maximum random error calculated by (Upper LoA − Lower LoA)/2.

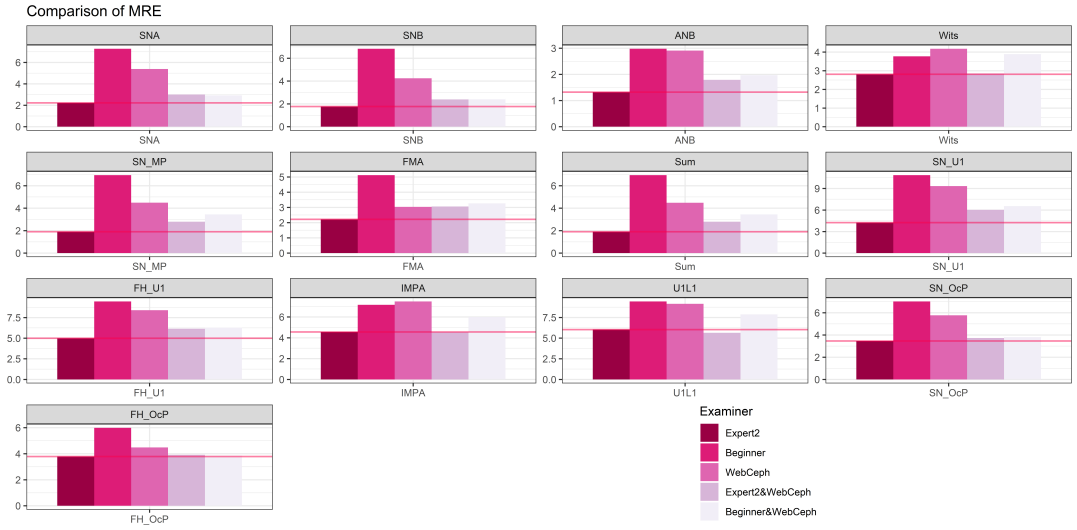

FIGURE 2. Bar graphs of the magnitude of MREs. The horizontal solid line represents the acceptable clinical limit for random error.

TABLE 7. The number of variables per each examiner meeting the interchangeability criterion.

Examiner	N	Variables
Beginner	0	-
WebCeph	0	-
WebCeph & Expert 2	3	Wits, IMPA, U1L1
WebCeph & Beginner	0	-

The ICC values could confirm the aforementioned study findings again. The inter-examiner reliability was calculated to evaluate the effect of the reduced measurement error on discriminating individuals in the study group (Table 8) [12]. As in the previous study, both absolute agreement and consistency ICC were measured. No considerable difference was found between the two ICCs, which further supported the previous decision that systematic error was not clinically relevant. Moreover, the beginner's reliability could be improved to a level close to that of expert 2 with the help of AI. However, expert 2 showed a slight decrease in reliability when assisted by AI. These conflicting reliability results between the beginner and expert 2 may imply that AI-aided cephalometric analysis cannot be practical for all human examiners.

Overall, adding a human review to AI could significantly reduce measurement errors, which led to an apparent improvement in agreement and reliability compared with using AI alone in cephalometric analysis. However, in terms of superiority to manual landmarking, conflicting results were noted according to the examiner's expertise. When using AI together, less-experienced examiners could obtain better results than manual alone, whereas regular examiners obtained slightly worse results than manual landmarking. Considering that the time-saving effect is not evident when using AI, AI-supported commercial cephalometric analysis may be more recommendable for less-experienced examiners such as general practitioners or a beginner in the pediatric dentistry or orthodontic field.

The reason for these conflicting results is unclear. As one hypothesis, expert 2 and the beginner might have made AI-dependent decisions during the landmark review. WebCeph's performance is approximately halfway between the beginner and expert 2 (Fig. 2). Thus, WebCeph's landmarking may have suggested a valuable guide for the beginner, whereas it may have hindered expert 2. For accurate verification, analysis at the landmark coordinate level may be required; however, this goes beyond the scope of this study. Further research may be needed.

As another limitation, the number of examiners participating was small due to its pilot nature. Consequently, various examiner factors may have affected the study results. If a sufficient number of qualified examiners had participated, these examiner-origin biases would have been controlled. The reference positions of the landmarks could have also set closer to the ground truth. Further studies with more examiners may be needed.

5. Conclusions

In this study, the clinical effectiveness of commercially available AI-supported cephalometric services was evaluated under the assumption of a mandatory human examiner review. Within the limitations of this study, the following conclusions were drawn:

1. Contrary to expectations, the time-saving effect was not evident when using AI services together with an examiner's review.

2. AI measurement errors generally decreased when reviewed by a human examiner. However, the actual benefits were limited to beginners. The regular examiner's error slightly increased more than manual landmarking when using AI.

3. Overall, commercial AI-supported cephalometric services are recommendable as an aid for less-experienced examiners or clinicians.

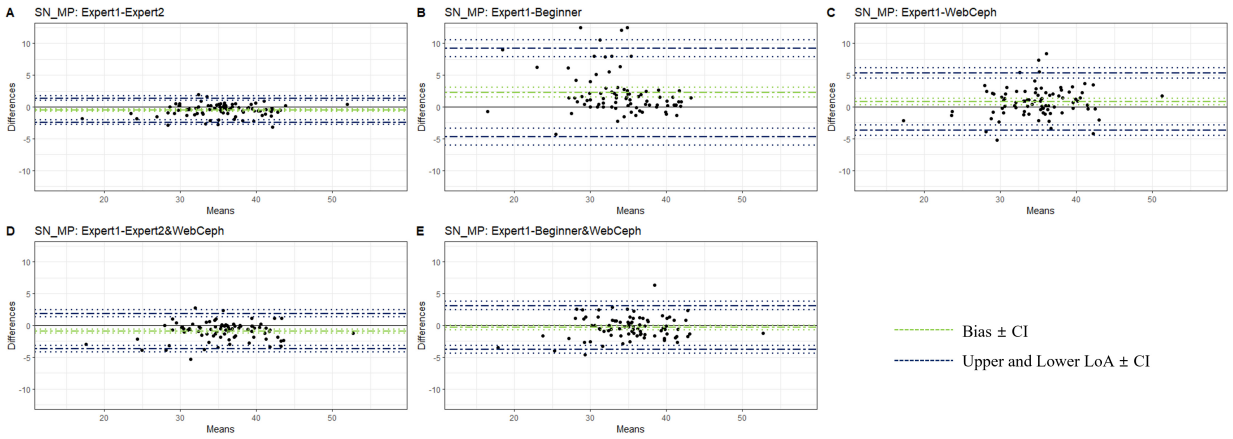

FIGURE 3. Bland-Altman plots for SN-MP. In each graph, the x-axis denotes the average of two examiners' measurement outcomes, while the y-axis is the difference between the examiners. Green dotted lines represent the bias with a 95% confidence interval. Blue dotted lines stand for the upper and lower limit of agreement with a 95% confidence interval. (A) between expert 1 and expert 2; (B) between expert 1 and beginner; (C) between expert 1 and Webceph; (D) between expert 1 and Webceph which was revised by expert2; (E) between expert 1 and WebCeph which was revised by beginner. CI, 95% confidence interval; LoA, limit of agreement.

TABLE 8. Inter-examiner reliability.

Examiner	Variables	Absolute Agreement		Consistency	
		Coefficient	95% CI	Coefficient	95% CI
Expert 1-Expert 2					
	SNA	0.94	(0.88, 0.97)	0.95	(0.92, 0.97)
	SNB	0.97	(0.96, 0.98)	0.98	(0.96, 0.98)
	ANB	0.97	(0.95, 0.98)	0.98	(0.96, 0.98)
	Wits	0.95	(0.90, 0.97)	0.96	(0.94, 0.97)
	SN-MP	0.98	(0.95, 0.99)	0.98	(0.97, 0.99)
	FMA	0.95	(0.71, 0.98)	0.97	(0.96, 0.98)
	Bjork-Jarabak Sum	0.98	(0.95, 0.99)	0.98	(0.97, 0.99)
	SN-U1	0.97	(0.95, 0.98)	0.97	(0.96, 0.98)
	FH-U1	0.95	(0.91, 0.97)	0.96	(0.94, 0.97)
	IMPA	0.95	(0.92, 0.97)	0.96	(0.93, 0.97)
	U1L1	0.97	(0.96, 0.98)	0.97	(0.96, 0.98)
	SN-OcP	0.87	(0.67, 0.94)	0.91	(0.86, 0.94)
	FH-OcP	0.87	(0.81, 0.92)	0.88	(0.81, 0.92)
	Mean	0.95		0.96	
Expert 1-Beginner					
	SNA	0.57	(0.28, 0.74)	0.65	(0.50, 0.76)
	SNB	0.67	(0.34, 0.82)	0.74	(0.63, 0.83)
	ANB	0.89	(0.84, 0.93)	0.89	(0.84, 0.93)
	Wits	0.92	(0.83, 0.95)	0.93	(0.90, 0.96)
	SN-MP	0.73	(0.45, 0.85)	0.79	(0.69, 0.86)
	FMA	0.83	(0.71, 0.90)	0.85	(0.78, 0.90)
	Bjork-Jarabak Sum	0.73	(0.45, 0.85)	0.79	(0.69, 0.86)
	SN-U1	0.78	(0.37, 0.90)	0.86	(0.79, 0.90)
	FH-U1	0.87	(0.80, 0.91)	0.87	(0.81, 0.92)
	IMPA	0.88	(0.81, 0.93)	0.90	(0.84, 0.93)
	U1L1	0.90	(0.68, 0.96)	0.94	(0.90, 0.96)
	SN-OcP	0.71	(0.55, 0.81)	0.73	(0.62, 0.82)
	FH-OcP	0.70	(0.39, 0.84)	0.77	(0.67, 0.84)
	Mean	0.78		0.82	

TABLE 8. Continued.

Examiner	Variables	Absolute Agreement		Consistency	
		Coefficient	95% CI	Coefficient	95% CI
Expert 1-Webceph					
	SNA	0.62	(0.41, 0.76)	0.67	(0.53, 0.77)
	SNB	0.84	(0.77, 0.90)	0.84	(0.77, 0.90)
	ANB	0.84	(0.54, 0.92)	0.89	(0.83, 0.92)
	Wits	0.88	(0.69, 0.94)	0.91	(0.87, 0.94)
	SN-MP	0.89	(0.82, 0.93)	0.90	(0.85, 0.93)
	FMA	0.95	(0.92, 0.97)	0.95	(0.92, 0.97)
	Bjork-Jarabak Sum	0.89	(0.82, 0.93)	0.90	(0.85, 0.93)
	SN-U1	0.84	(0.74, 0.90)	0.85	(0.78, 0.90)
	FH-U1	0.83	(0.63, 0.91)	0.87	(0.81, 0.91)
	IMPA	0.86	(0.79, 0.90)	0.86	(0.79, 0.91)
	U1L1	0.89	(0.68, 0.95)	0.93	(0.89, 0.95)
	SN-OcP	0.72	(0.60, 0.81)	0.74	(0.62, 0.82)
	FH-OcP	0.82	(0.73, 0.88)	0.82	(0.73, 0.88)
	Mean	0.84		0.86	
Expert 1-Expert 2_Webceph					
	SNA	0.91	(0.86, 0.94)	0.91	(0.86, 0.94)
	SNB	0.94	(0.85, 0.97)	0.96	(0.93, 0.97)
	ANB	0.95	(0.84, 0.97)	0.96	(0.93, 0.97)
	Wits	0.95	(0.86, 0.97)	0.96	(0.94, 0.97)
	SN-MP	0.95	(0.86, 0.98)	0.96	(0.94, 0.98)
	FMA	0.94	(0.90, 0.96)	0.95	(0.92, 0.96)
	Bjork-Jarabak Sum	0.95	(0.86, 0.98)	0.96	(0.94, 0.98)
	SN-U1	0.93	(0.86, 0.96)	0.94	(0.92, 0.96)
	FH-U1	0.94	(0.91, 0.96)	0.94	(0.91, 0.96)
	IMPA	0.92	(0.53, 0.97)	0.96	(0.93, 0.97)
	U1L1	0.95	(0.68, 0.98)	0.98	(0.96, 0.98)
	SN-OcP	0.88	(0.82, 0.93)	0.89	(0.84, 0.93)
	FH-OcP	0.85	(0.74, 0.91)	0.87	(0.81, 0.91)
	Mean	0.93		0.94	
Expert 1-Beginner_Webceph					
	SNA	0.89	(0.79, 0.94)	0.91	(0.86, 0.94)
	SNB	0.95	(0.93, 0.97)	0.95	(0.93, 0.97)
	ANB	0.92	(0.75, 0.96)	0.95	(0.92, 0.96)
	Wits	0.92	(0.83, 0.95)	0.93	(0.90, 0.95)
	SN-MP	0.94	(0.91, 0.96)	0.94	(0.91, 0.96)
	FMA	0.94	(0.91, 0.96)	0.94	(0.91, 0.96)
	Bjork-Jarabak Sum	0.94	(0.91, 0.96)	0.94	(0.91, 0.96)
	SN-U1	0.92	(0.87, 0.95)	0.93	(0.89, 0.95)
	FH-U1	0.93	(0.88, 0.95)	0.93	(0.90, 0.95)
	IMPA	0.86	(0.52, 0.94)	0.91	(0.87, 0.94)
	U1L1	0.92	(0.65, 0.97)	0.95	(0.92, 0.97)
	SN-OcP	0.88	(0.82, 0.93)	0.89	(0.84, 0.93)
	FH-OcP	0.86	(0.75, 0.92)	0.88	(0.82, 0.92)
	Mean	0.91		0.93	

ICC, intraclass correlation coefficient; CI, 95% confidence interval of ICC; All variables showed statistical significance with p-values less than 0.001.

ABBREVIATIONS

AI, artificial intelligence; ICC, intraclass correlation coefficient; LoA, limit of agreement; MRE, the maximum random error.

AVAILABILITY OF DATA AND MATERIALS

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

HKN, SRB and JSL—collected and analyzed experimental data and drafted the manuscript. HKN—designed the research and critically supervised the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the Institutional Review Board of Kyungpook National University Dental Hospital (KNUDH-2022-05-02-00). Patient informed consent was waived due to the study's retrospective nature.

ACKNOWLEDGMENT

Thanks to Kyungpook National University Dental Hospital Dental Research Institute for support (2022).

FUNDING

This work was supported by Kyungpook National University Dental Hospital Institute for Dental Research (2022).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Schwendicke F, Chaurasia A, Arsiwala L, Lee JH, Elhennawy K, Jost-Brinkmann PG, *et al*. Deep learning for cephalometric landmark detection: systematic review and meta-analysis. Clinical Oral Investigations. 2021; 25: 4299–4309.

[2] Kim H, Shim E, Park J, Kim YJ, Lee U, Kim Y. Web-based fully automated cephalometric analysis by deep learning. Computer Methods and Programs in Biomedicine. 2020; 194: 105513.

[3] Hwang HW, Moon JH, Kim MG, Donatelli RE, Lee SJ. Evaluation of automated cephalometric analysis based on the latest deep learning method. The Angle Orthodontist. 2021; 91: 329–335.

[4] Gil SM, Kim I, Cho JH, Hong M, Kim M, Kim SJ, *et al*. Accuracy of auto-identification of the posteroanterior cephalometric landmarks using cascade convolution neural network algorithm and cephalometric images of different quality from nationwide multiple centers. American Journal of Orthodontics and Dentofacial Orthopedics. 2022; 161: e361–e371.

[5] Yoon HJ, Kim DR, Gwon E, Kim N, Baek SH, Ahn HW, *et al*. Fully automated identification of cephalometric landmarks for upper airway assessment using cascaded convolutional neural networks. European Journal of Orthodontics. 2022; 44: 66–77.

[6] Park JH, Hwang HW, Moon JH, Yu Y, Kim H, Her SB, *et al*. Automated identification of cephalometric landmarks: part 1—comparisons between the latest deep-learning methods YOLOV3 and SSD. The Angle Orthodontist. 2019; 89: 903–909.

[7] Hwang HW, Park JH, Moon JH, Yu Y, Kim H, Her SB, *et al*. Automated identification of cephalometric landmarks: part 2—might it be better than human? The Angle Orthodontist. 2020; 90: 69–76.

[8] Choi YJ, Lee K. Possibilities of artificial intelligence use in orthodontic diagnosis and treatment planning: image recognition and three-dimensional VTO. Seminars in Orthodontics. 2021; 27: 121–129.

[9] Leonardi R, Giordano D, Maiorana F, Spampinato C. Automatic cephalometric analysis: a systematic review. The Angle Orthodontist. 2008; 78: 145–151.

[10] Rudolph DJ, Sinclair PM, Coggins JM. Automatic computerized radiographic identification of cephalometric landmarks. American Journal of Orthodontics and Dentofacial Orthopedics. 1998; 113: 173–179.

[11] Kazandjian S, Kiliaridis S, Mavropoulos A. Validity and reliability of a new edge-based computerized method for identification of cephalometric landmarks. The Angle Orthodontist. 2006; 76: 619–624.

[12] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine. 2016; 15: 155–163.

[13] Haghayegh S, Kang HA, Khoshnevis S, Smolensky MH, Diller KR. A comprehensive guideline for Bland-Altman and intra class correlation calculations to properly compare two methods of measurement and interpret findings. Physiological Measurement. 2020; 41: 055012.

[14] van Stralen KJ, Jager KJ, Zoccali C, Dekker FW. Agreement between methods. Kidney International. 2008; 74: 1116–1120.

[15] Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research. 1999; 8: 135–160.

[16] Lagravère MO, Low C, Flores-Mir C, Chung R, Carey JP, Heo G, *et al*. Intraexaminer and interexaminer reliabilities of landmark identification on digitized lateral cephalograms and formatted 3-dimensional cone-beam computerized tomography images. American Journal of Orthodontics and Dentofacial Orthopedics. 2010; 137: 598–604.

[17] Tanikawa C, Lee C, Lim J, Oka A, Yamashiro T. Clinical applicability of automated cephalometric landmark identification: Part I—patient-related identification errors. Orthodontics & Craniofacial Research. 2021; 24: 43–52.

[18] Jeon S, Lee KC. Comparison of cephalometric measurements between conventional and automatic cephalometric analysis using convolutional neural network. Progress in Orthodontics. 2021; 22: 14.

How to cite this article: Jaesik Lee, Seong-Ryeol Bae, Hyung-Kyu Noh. Commercial artificial intelligence lateral cephalometric analysis: part 2—effects of human examiners on artificial intelligence performance, a pilot study. Journal of Clinical Pediatric Dentistry. 2023; 47(6): 130-141. doi: 10.22514/jocpd.2023.087.