Submitted: 10 April, 2023 Accepted: 12 May, 2023 Published: 03 November, 2023 DOI:10.22514/jocpd.2023.085

ORIGINAL RESEARCH

Journal of Clinical Pediatric Dentistry

Commercial artificial intelligence lateral cephalometric analysis: part 1—the possibility of replacing manual landmarking with artificial intelligence service

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

Artificial intelligence (AI) technology has recently been introduced to dentistry. AI-assisted cephalometric analysis is one of its applications, and several commercial AI services have already been launched. However, the performance of these commercial services is still unclear. This study aimed to determine whether commercially available AI cephalometric analysis can replace manual analysis by human examiners. Eighty-four pretreatment lateral cephalograms were traced and examined by two orthodontists and four commercial AIs, and 13 commonly used cephalometric variables were calculated. Then, the Bland-Altman analysis was conducted to evaluate systematic and random errors between examiners. The interchangeability of an AI was determined if the random errors of the AI were smaller than the clinically acceptable limits derived from the random errors between human examiners. Finally, the inter-examiner reliability index was calculated, and Cohen's kappa was determined to assess the actual classification reliability of each examiner. The systematic errors of the AIs were clinically insignificant in general. However, the random errors of the AIs were approximately twice those of human examiners, which did not satisfy the interchangeability condition. Furthermore, even though the reliability indices of the AIs were in the good-to-excellent range, their classification reliability was unacceptable. Commercial AI is still at a level that makes it challenging to replace manual landmarking by human experts. Thus, a human examiner's landmark position review is mandatory when using commercial AIs.

Keywords

Cephalometric; Artificial intelligence; Accuracy; Precision; Reliability

1. Introduction

Lateral cephalometrics is a useful diagnostic tool for assessing growth in children and adolescents. Specifically, if a maxillofacial deformity is present, lateral cephalometric evaluation is essential before orthopedic treatments. Despite this clinical importance, accurate landmarking has been constantly raised as difficult, and it has become a great entry barrier, especially for beginners in pediatric dentistry or orthodontic fields. Computer-aided landmarking has been consistently attempted; however, no significant results in accuracy and reliability have been achieved [1].

Recently, deep learning-based artificial intelligence (AI) has been introduced for dental imaging. In many studies published after 2017, the Euclidean distance (point-point distance) between landmarks detected by AI and the human examiner was <2 mm in more than 80% of cases [2–7]. Considering that the landmarking error between human examiners is approximately 1.5 mm [8], a dramatic improvement has undoubtedly been achieved in the accuracy of AI-assisted cephalometric landmark detection.

However, whether this technology is clinically applicable should be cautiously determined. For example, many studies have reported that the prediction performance of AI is degraded when evaluated with a test set whose source is different from that of the AI's learning set. This phenomenon, known as the lack of generalizability, is due to the difference in image qualities between learning and test data [2, 9–11]. Given that the source of cephalogram images can vary in actual clinical settings, this can be a significant issue. Furthermore, despite AI's improved performance, some landmarks showed different directional patterns in the scatter plot between AI and a human examiner [3, 5]. For instance, the landmark of the lower incisor edge was within a 2 mm error range for both human examiners and AI. However, while the human examiners showed an isotropic (circular) distribution, the AI showed an anisotropic (elliptic) pattern.

Considering the aforementioned issues, there is still a possibility that cephalometric variables measured by AI and human examiners may differ significantly. Nevertheless, various AI-assisted automatic cephalometric analysis services are commercially available [11]. Although some studies have

already evaluated the performance of commercial AI, these studies have shown limitations such as insufficient sample size or evaluation of only a limited number of AIs [10, 12, 13]. The accurate and objective performance evaluation of AI cephalometric services that are commercially available is still an unresolved issue for clinicians.

Thus, this study aimed to assess the performance of four commercial AI-assisted automatic cephalometric services using cephalometric variables. The agreement and reliability of AI were compared with those of human examiners. From this, we determined whether commercial automatic cephalometric services can replace manual analysis by human experts. To the best of our knowledge, this is the first study comparing the performance of four commercial AIs at once with a sufficient sample size.

2. Materials and methods

The study samples were lateral cephalograms, which were randomly collected from the diagnostic records of patients who visited the Department of Orthodontics and Pediatric Dentistry of Kyungpook National University Dental Hospital for the treatment of malocclusion between 2012 and 2021. This study was conducted on patients after mixed dentition with erupted permanent first molars and incisors. Patients with craniofacial malformations such as cleft lip and palate, evident facial asymmetry observed in posteroanterior cephalogram, or missing molars or incisors were excluded from the study. A total of 84 cephalogram images were obtained. All images that were taken using CX-90SP (an X-ray scanner, Asahi, Kyoto, Japan) had a resolution of 150 DPI and gray level of 24 in JPG format. The image size was calibrated using a marker ruler in each cephalogram. The characteristics of the samples are presented in Table 1.

TABLE 1. Sample characteristics.

Characteristics	N	Mean	SD
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
Age (yr)	84	11.13	3.52

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: $SN\text{-}MP \leq 27$.

Fifteen skeletal and dental landmarks were selected to calculate the 13 commonly used cephalometric variables (Tables 2 and 3, Fig. 1). To estimate the performance of human examiners, two human examiners, expert 1 (HKN) and expert 2 (SRB), who are board-certified orthodontists with more than 7 and 5 years of clinical experience, respectively, participated in this study. First, the examiners discussed and agreed upon the definition of landmarks using three cephalograms that were not included in the study samples. Subsequent analysis was performed independently, without any communication between the examiners. The first set of measurement data was obtained by manual landmark identification on a monitor screen using computer software (6.3 Sequential Tracing Mode, AudaxCeph, Ljubljana, Slovenia). Measurements were repeated for all samples after 1 month.

This study used four commercially available AI-supported automatic cephalometric analysis services, namely, CellmatIQ (GmbH, Hamburg, Germany), CephX (ORCA Dental AI, Herzliya, Israel), AudaxCeph Automatic Tracing Mode (6.3, Audax, Ljubljana, Slovenia) and WebCeph (1.0.0, Assemblecircle, Gyeonggi-do, Korea). Initially, after uploading the anonymized cephalogram images, each AI automatically performed landmark detection. Then, to assess the pure performance of the AI architectures, the results of the AI analysis were obtained without the landmark adjustments of the human examiners. Among measurement variables, the values of Wits, FMA, FH-U1, SN-OcP and FH-OcP could not be obtained using CellmatIQ because this service does not support these variables.

The study design is illustrated in Fig. 2.

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 reliabilities were evaluated between the first and second measurements of human examiners with intraclass correlation coefficients (ICCs) using two-way mixed-effects, single-rater and absolute agreement models [14]. Then, Dahlberg's formula was used to calculate the method errors between the first and second attempts.

The Bland-Altman analysis was performed to estimate the measurement errors between conventional manual analysis and the newly developed AI technique [15, 16]. In the Bland-Altman protocol, measurements by expert 1 were set as the reference, and those of expert 2 and AIs were the evaluation targets. First, the means and differences between the two methods were calculated. The normality of the difference was confirmed using the Shapiro-Wilk test. The Bland-Altman statistics were then obtained. Bias measures the mean difference between the methods (systematic error), whereas the limit of agreement (LoA) represents the upper and lower limits that contain 95% of measurement errors (systematic and random errors) [17]. The maximum random error (MRE), defined as the half-width of the upper and lower LoAs, is an index of the magnitude of pure random errors between the methods [15, 16]. Finally, the Bland-Altman plots were drawn to visually understand the agreement.

The interchangeability between measurement methods can be judged by comparing the random error magnitude with the acceptable clinical limit, a priori criterion based on existing

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. Classification criteria.

Variables	Type I	Type II	Type III
SNA	80~84	84≤	≤80
SNB	78~82	≤78	82≤
ANB	0~4	4≤	≤0
Wits	−1~1	1≤	≤−1
FMA	22~28	28≤	≤22
SN-MP	27~37	37≤	≤27
Björk-Jarabak Sum	390~402	402≤	≤390
SN-U1	96~108	108≤	≤96
FH-U1	111~121	121≤	≤111
IMPA	88~102	102≤	≤88
U1L1	124~136	≤124	136≤
SN-OcP	11~17	17≤	≤11
FH-OcP	8~12	12≤	≤8

literature or widely accepted by most experts rather than statistically determined [15–17]. In particular, the new method can be interchangeably used with the previous method if the random error of a newly developed method is below or equal to the acceptable clinical limit [15, 16]. However, determining the acceptable clinical limit was challenging; there was neither an evident empirical standard nor previous studies explicitly reporting the allowable cephalometric error range. Therefore, the clinical criteria that coincided with the purpose of this study must be established. The study intended to determine whether the performance level of the AI matched that of a regular human examiner. Thus, the MREs between experts 1 and 2 were selected as the acceptable clinical limit for evaluating AI performance. Then, the MRE of the AI was compared to this clinical limit to determine interchangeability. The number of variables per AI that met this condition was counted.

The ICCs of the two-way random-effects and single-rater models were used to assess the inter-examiner reliability between the human reference (expert 1) and other examiners (expert 2 and AIs). The absolute agreement and consistency ICCs were calculated to evaluate the inter-examiner reliability with and without systematic error, respectively.

The classification reliability of the AIs was evaluated to illustrate their performance from a clinical perspective. The classification criteria are presented in Table 3. The classification by expert 1 was set as the true reference, whereas classifications by other examiners were predictions. Then, Cohen's kappa coefficients were calculated.

3. Results

The mean intra-examiner reliability measures were 0.97 for expert 1 and 0.94 for expert 2, showing excellent reliability (Table 4). Dahlberg's formula also estimated method errors of approximately 1 mm and 1° for linear and angular measurements, respectively. Therefore, the average of the first and second measurements were taken for the data of each human examiner, and these mean values were used in the subsequent analyses.

The descriptive and Bland-Altman statistics of cephalometric variables measured by each examiner are presented in

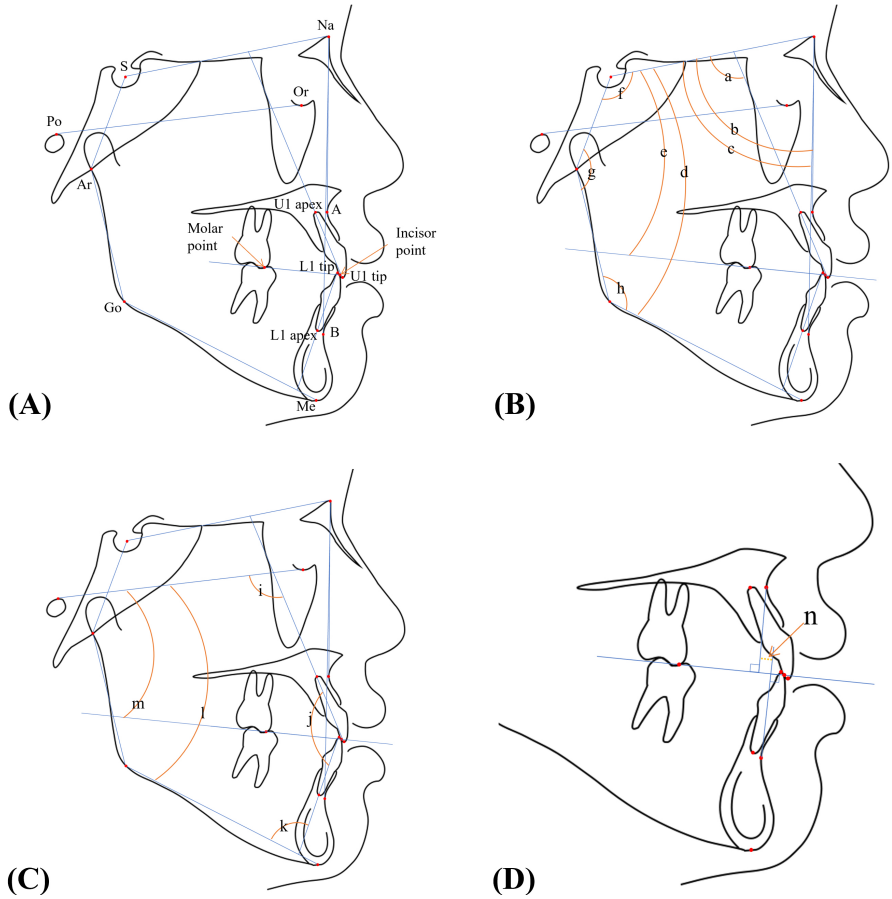

FIGURE 1. Graphical representation of landmarks and variables. (A) Cephalometric landmarks; (B–D) Cephalometric variables. a: SN-U1; b: SNA; c: SNB; d: SN-MP; e: SN-OcP; f–h: Björk-Jarabak Sum; i: FH-U1; j: U1L1; k: IMPA; l: FMA; m: FH-OcP; n: Wit's; S: sella; Na: nasion; Po: porion; Or: orbitale; Ar: articulare; Go: gonion; Me: menton; U1: upper central incisor; L1: lower central incisor.

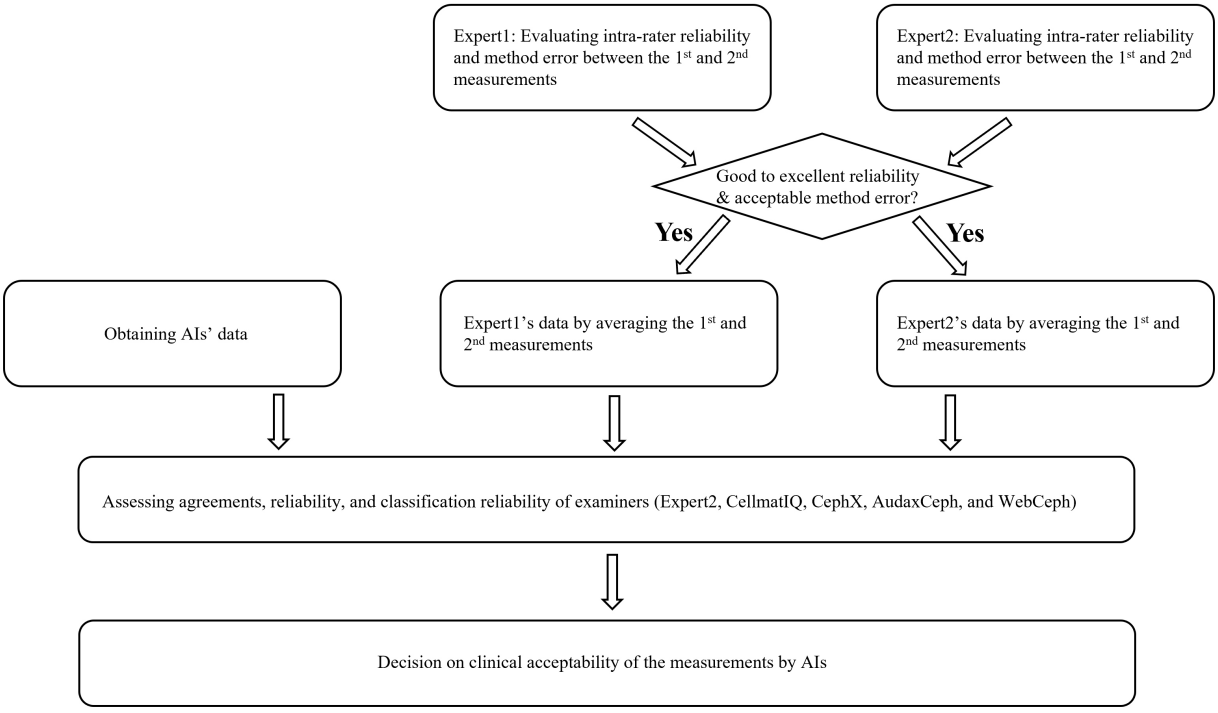

FIGURE 2. A flow chart summarizing the study design. AI: artificial intelligence.

TABLE 4. Measures of intra-examiner reliability and method errors.

Variables	Expert 1			Expert 2		
	ICC	95% CI	Dahlberg	ICC	95% CI	Dahlberg
SNA	0.95	(0.93, 0.97)	0.79	0.93	(0.90, 0.96)	0.90
SNB	0.97	(0.96, 0.98)	0.68	0.97	(0.96, 0.98)	0.67
ANB	0.98	(0.97, 0.99)	0.46	0.97	(0.95, 0.98)	0.58
Wits	0.98	(0.97, 0.99)	0.71	0.96	(0.94, 0.97)	1.03
SN-MP	0.98	(0.97, 0.99)	0.77	0.96	(0.94, 0.97)	1.01
FMA	0.97	(0.95, 0.98)	0.84	0.88	(0.82, 0.92)	1.67
Bjork-Jarabak Sum	0.98	(0.97, 0.99)	0.77	0.96	(0.94, 0.97)	1.01
SN-U1	0.98	(0.97, 0.99)	1.22	0.97	(0.96, 0.98)	1.52
FH-U1	0.98	(0.97, 0.99)	1.22	0.95	(0.92, 0.97)	1.99
IMPA	0.97	(0.96, 0.98)	1.27	0.94	(0.91, 0.96)	1.87
U1L1	0.98	(0.97, 0.99)	1.68	0.98	(0.96, 0.98)	2.08
SN-OcP	0.95	(0.93, 0.97)	0.92	0.92	(0.88, 0.95)	1.13
FH-OcP	0.94	(0.91, 0.96)	0.98	0.81	(0.73, 0.88)	1.75
Mean	0.97		0.71 mm (Linear) 0.97° (Angular)	0.94		1.03 mm (Linear) 1.35° (Angular)

ICC: intraclass correlation coefficient; CI: 95% confidence interval of ICC; Sig: significance; Dahlberg: method errors obtained by Dahlberg's formula ($\sqrt{\sum d^2/2n}$); All measurements showed statistical significance with p-values less than 0.001.

Tables 5 and 6. The bias existed not only in expert 2 but also in the AIs. However, the magnitude of the bias was small. Exceptionally, some variables, such as SN-MP, FMA, Björk-Jarabak sum and IMPA, showed an unexpectedly large bias when measured by CephX. Meanwhile, the MREs of AI measurements were more extensive than that of human examiners, indicating more prominent random errors in AI measurements (Table 6). The magnitudes of the MREs among examiners are shown in Fig. 3. The MREs of expert 2 were always smaller than those of the AIs. Consequently, the number of variables for each AI that satisfied the agreement criteria was zero (Table 7). As an illustrative example, the Bland-Altman plots of the SN-MP are shown in Fig. 4.

The mean inter-examiner reliability measures were 0.96 and 0.96 for absolute agreement and consistency between the human examiners (Table 8), respectively. However, the mean ICCs for AIs were between 0.83 and 0.91, smaller than those of the human examiners. The difference between the absolute agreement and consistency, when measured using CephX, was evident in SN-MP, FMA, Björk-Jarabak sum and IMPA.

The overall classification reliability was estimated by Cohen's kappa coefficients (Table 9). The reliability of expert 2 was 0.76, where the reliability of the AI was generally inferior to that of human examiners, ranging from 0.53 to 0.68. In particular, the coefficients for SN-MP, FMA, Björk-Jarabak sum and IMPA using CephX decreased to 0.2.

4. Discussion

Bias estimates the magnitude of the systematic error between the two methods [17]. The overall systematic errors between a human examiner and the AIs were clinically insignificant (Table 6). Exceptionally, only the mandibular plane-associated variables, *i.e.*, SN-MP, FMA, Björk-Jarabak sum and IMPA, showed considerable bias when evaluated by CephX. Thus, bias correction may be necessary when evaluating these variables using CephX. Fortunately, bias correction is technically easy; *i.e.*, the results can be quickly revised by subtracting the bias from the original measurement value [15, 16].

However, a random error correction is not available. Thus, the magnitude of the random error is the key criterion in determining the interchangeability of two methods. Accordingly, the overall performance of commercial AIs was unsatisfactory. The magnitude of the random errors of the AIs was generally more prominent than that of expert 2 (Table 6, Fig. 3). Consequently, the number of cephalometric variables that met the interchangeability criteria for each AI was nearly zero (Table 7). Thus, these currently available commercial AI services could not replace manual analysis by regular examiners.

This finding contradicts the conclusions of previous studies that reported the acceptable quality of AI cephalometrics. For instance, Kunz *et al.* [13] reported that CellmatIQ could analyze at the same level as human experts. Similarly, Jeon and Lee [12] concluded that CephX offers clinically acceptable results. Although these studies analyzed the measurement errors between humans and AIs using the Bland-Altman analysis, their results were interpreted without considering the acceptable clinical limit. To the best of our knowledge, this is the first attempt to determine the interchangeability between regular orthodontists and AIs according to the Bland-Altman protocol using a clinical limit. By adopting the MREs between regular, not highly experienced orthodontists as the allowable clinical limit for the random error magnitude of AIs, we could draw a different conclusion on the interchangeability between a human examiner and AI.

It might be argued that employing regular orthodontists as

TABLE 5. Descriptive statistics.

	Expert 1	Expert 2	CellmatIQ	CephX	AudaxCeph	WebCeph
SNA	80.60 ± 3.62	80.22 ± 3.40	80.78 ± 4.29	80.49 ± 3.71	80.96 ± 3.49	81.92 ± 3.09
SNB	78.52 ± 4.16	78.30 ± 4.00	78.28 ± 4.54	78.23 ± 4.12	78.69 ± 4.15	78.76 ± 3.57
ANB	2.06 ± 3.06	1.92 ± 3.08	2.50 ± 3.19	2.27 ± 3.23	2.26 ± 3.05	3.16 ± 3.17
Wits	−2.28 ± 5.11	−2.69 ± 5.07	-	−1.52 ± 5.39	−2.45 ± 5.01	−0.88 ± 5.03
SN-MP	35.12 ± 5.24	35.48 ± 5.06	35.63 ± 5.33	40.21 ± 4.77	35.54 ± 5.23	34.29 ± 4.89
FMA	26.43 ± 4.77	25.18 ± 4.65	-	30.20 ± 4.38	28.14 ± 4.59	26.48 ± 4.67
Bjork-Jarabak Sum	395.12 ± 5.23	395.48 ± 5.06	395.63 ± 5.33	400.19 ± 4.77	395.54 ± 5.23	394.29 ± 4.89
SN-U1	106.86 ± 9.16	106.60 ± 8.92	106.90 ± 7.72	106.78 ± 7.46	105.21 ± 8.30	105.14 ± 8.48
FH-U1	115.55 ± 8.73	116.90 ± 8.79	-	115.50 ± 7.16	112.61 ± 8.31	112.95 ± 8.15
IMPA	92.98 ± 7.82	91.62 ± 7.73	92.08 ± 7.79	86.22 ± 6.64	91.17 ± 7.21	92.16 ± 6.58
U1L1	125.05 ± 13.01	126.31 ± 13.27	125.89 ± 10.91	126.78 ± 10.91	128.08 ± 12.42	128.41 ± 11.13
SN-OcP	17.81 ± 4.20	18.80 ± 3.88	-	17.29 ± 3.92	18.19 ± 4.04	16.99 ± 3.92
FH-OcP	9.12 ± 3.99	8.50 ± 3.86	-	8.58 ± 3.63	10.79 ± 3.65	9.18 ± 3.57

Values were mean ± standard deviation.

TABLE 6. Bland-Altman statistics.

	Expert 1-Expert 2		Expert 1-CellmatIQ		Expert 1-CephX		Expert 1-AudaxCeph		Expert 1-WebCeph	
	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI
SNA										
Bias	0.37	(0.17, 0.58)	−0.18	(−0.70, 0.33)	0.11	(−0.27, 0.48)	−0.36	(−0.88, 0.15)	−1.33	(−1.92, −0.73)
Upper LoA	2.22	(1.87, 2.57)	4.45	(3.57, 5.33)	3.48	(2.84, 4.13)	4.27	(3.39, 5.15)	4.06	(3.04, 5.08)
Lower LoA	−1.47	(−1.82, −1.12)	−4.82	(−5.70, −3.94)	−3.27	(−3.91, −2.63)	−5.00	(−5.88, −4.12)	−6.71	(−7.73, −5.69)
MRE	1.84		4.63		3.38		4.63		5.39	
SNB										
Bias	0.22	(0.03, 0.40)	0.24	(−0.19, 0.67)	0.29	(−0.07, 0.65)	−0.17	(−0.62, 0.28)	−0.24	(−0.71, 0.23)
Upper LoA	1.87	(1.56, 2.19)	4.15	(3.41, 4.89)	3.56	(2.94, 4.18)	3.90	(3.12, 4.67)	4.00	(3.20, 4.81)
Lower LoA	−1.44	(−1.75, −1.12)	−3.66	(−4.41, −2.92)	−2.98	(−3.60, −2.36)	−4.24	(−5.01, −3.47)	−4.48	(−5.28, −3.67)
MRE	1.66		3.91		3.27		4.07		4.24	
ANB										
Bias	0.15	(0.01, 0.28)	−0.44	(−0.70, −0.18)	−0.20	(−0.44, 0.04)	−0.20	(−0.39, −0.02)	−1.10	(−1.42, −0.78)
Upper LoA	1.36	(1.13, 1.59)	1.92	(1.47, 2.37)	1.95	(1.54, 2.36)	1.47	(1.15, 1.79)	1.81	(1.26, 2.36)
Lower LoA	−1.07	(−1.30, −0.84)	−2.80	(−3.25, −2.35)	−2.36	(−2.77, −1.95)	−1.87	(−2.19, −1.56)	−4.01	(−4.56, −3.46)
MRE	1.21		2.36		2.16		1.67		2.91	
Wits										
Bias	0.41	(0.16, 0.66)	-	-	−0.76	(−1.20, −0.33)	0.17	(−0.10, 0.43)	−1.40	(−1.87, −0.94)
Upper LoA	2.68	(2.25, 3.11)	-	-	3.16	(2.42, 3.91)	2.56	(2.11, 3.02)	2.78	(1.98, 3.57)
Lower LoA	−1.86	(−2.30, −1.43)	-	-	−4.68	(−5.43, −3.94)	−2.23	(−2.69, −1.77)	−5.58	(−6.37, −4.79)
MRE	2.27		-		3.92		2.40		4.18	

TABLE 6. Continued.

	Expert 1-Expert 2		Expert 1-CellmatIQ		Expert 1-CephX		Expert 1-AudaxCeph		Expert 1-WebCeph	
	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI
Wits										
Bias	0.41	(0.16, 0.66)	-	-	−0.76	(−1.20, −0.33)	0.17	(−0.10, 0.43)	−1.40	(−1.87, −0.94)
Upper LoA	2.68	(2.25, 3.11)	-	-	3.16	(2.42, 3.91)	2.56	(2.11, 3.02)	2.78	(1.98, 3.57)
Lower LoA	−1.86	(−2.30, −1.43)	-	-	−4.68	(−5.43, −3.94)	−2.23	(−2.69, −1.77)	−5.58	(−6.37, −4.79)
MRE	2.27		-		3.92		2.40		4.18	
SN-MP										
Bias	−0.36	(−0.57, −0.15)	−0.51	(−0.98, −0.05)	−5.09	(−5.47, −4.71)	−0.42	(−0.90, 0.06)	0.83	(0.33, 1.32)
Upper LoA	1.53	(1.17, 1.89)	3.68	(2.88, 4.48)	−1.69	(−2.33, −1.04)	3.93	(3.10, 4.76)	5.32	(4.47, 6.17)
Lower LoA	−2.25	(−2.61, −1.89)	−4.71	(−5.51, −3.91)	−8.49	(−9.14, −7.85)	−4.77	(−5.60, −3.94)	−3.67	(−4.52, −2.81)
MRE	1.89		4.20		3.40		4.35		4.49	
FMA										
Bias	1.25	(0.98, 1.52)	-	-	−3.77	(−4.23, −3.32)	−1.71	(−1.99, −1.42)	−0.05	(−0.38, 0.29)
Upper LoA	3.70	(3.23, 4.16)	-	-	0.35	(−0.43, 1.13)	0.86	(0.37, 1.35)	2.99	(2.41, 3.57)
Lower LoA	−1.19	(−1.66, −0.73)	-	-	−7.90	(−8.68, −7.11)	−4.28	(−4.76, −3.79)	−3.08	(−3.66, −2.50)
MRE	2.45		-		4.12		2.57		3.03	
Sum										
Bias	−0.36	(−0.57, −0.15)	−0.51	(−0.97, −0.05)	−5.08	(−5.45, −4.70)	−0.42	(−0.90, 0.06)	0.83	(0.33, 1.33)
Upper LoA	1.54	(1.18, 1.90)	3.68	(2.88, 4.47)	−1.67	(−2.32, −1.03)	3.93	(3.11, 4.76)	5.32	(4.46, 6.17)
Lower LoA	−2.25	(−2.61, −1.89)	−4.70	(−5.49, −3.90)	−8.48	(−9.13, −7.83)	−4.77	(−5.60, −3.94)	−3.66	(−4.51, −2.81)
MRE	1.89		4.19		3.40		4.35		4.49	
SN-U1										
Bias	0.27	(−0.17, 0.70)	−0.04	(−1.06, 0.98)	0.09	(−0.84, 1.02)	1.65	(0.93, 2.37)	1.73	(0.69, 2.76)
Upper LoA	4.18	(3.44, 4.93)	9.15	(7.40, 10.89)	8.49	(6.89, 10.09)	8.18	(6.94, 9.42)	11.06	(9.29, 12.84)
Lower LoA	−3.65	(−4.40, −2.91)	−9.22	(−10.97, −7.48)	−8.32	(−9.91, −6.72)	−4.88	(−6.12, −3.64)	−7.61	(−9.38, −5.84)
MRE	3.92		9.19		8.40		6.53		9.34	
FH-U1										
Bias	−1.34	(−1.85, −0.84)	-	-	0.06	(−0.81, 0.93)	2.94	(2.32, 3.56)	2.60	(1.67, 3.53)
Upper LoA	3.26	(2.38, 4.13)	-	-	7.91	(6.42, 9.41)	8.56	(7.50, 9.63)	11.00	(9.40, 12.59)
Lower LoA	−5.94	(−6.82, −5.07)	-	-	−7.80	(−9.29, −6.31)	−2.68	(−3.75, −1.62)	−5.79	(−7.39, −4.20)
MRE	4.60		-		7.86		5.62		8.39	

TABLE 6. Continued.

	Expert 1-Expert 2		Expert 1-CellmatIQ		Expert 1-CephX		Expert 1-AudaxCeph		Expert 1-WebCeph	
	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI
IMPA										
Bias	1.36	(0.96, 1.76)	0.89	(−0.32, 2.11)	6.75	(6.08, 7.42)	1.80	(1.37, 2.23)	0.81	(−0.02, 1.64)
Upper LoA	4.98	(4.29, 5.67)	11.87	(9.79, 13.96)	12.78	(11.64, 13.93)	5.67	(4.94, 6.41)	8.30	(6.88, 9.72)
Lower LoA	−2.26	(−2.95, −1.57)	−10.09	(−12.17, −8.00)	0.72	(−0.42, 1.87)	−2.06	(−2.80, −1.33)	−6.68	(−8.10, −5.26)
MRE	3.62		10.98		6.03		3.87		7.49	
U1L1										
Bias	−1.26	(−1.82, −0.70)	−0.83	(−1.90, 0.23)	−1.73	(−2.75, −0.71)	−3.03	(−3.71, −2.34)	−3.36	(−4.37, −2.34)
Upper LoA	3.80	(2.84, 4.76)	8.76	(6.94, 10.58)	7.47	(5.73, 9.22)	3.13	(1.96, 4.30)	5.79	(4.06, 7.53)
Lower LoA	−6.31	(−7.27, −5.35)	−10.43	(−12.25, −8.60)	−10.93	(−12.68, −9.18)	−9.19	(−10.36, −8.02)	−12.51	(−14.25, −10.77)
MRE	5.06		9.59		9.20		6.16		9.15	
SN-OcP										
Bias	−0.99	(−1.29, −0.69)	-	-	0.52	(−0.01, 1.05)	−0.37	(−0.93, 0.18)	0.82	(0.18, 1.46)
Upper LoA	1.71	(1.20, 2.22)	-	-	5.30	(4.39, 6.21)	4.65	(3.69, 5.60)	6.60	(5.50, 7.69)
Lower LoA	−3.68	(−4.20, −3.17)	-	-	−4.26	(−5.17, −3.35)	−5.39	(−6.35, −4.44)	−4.95	(−6.05, −3.85)
MRE	2.70		-		4.78		5.02		5.77	
FH-OcP										
Bias	0.62	(0.24, 1.00)	-	-	0.55	(0.05, 1.05)	−1.66	(−2.04, −1.29)	−0.05	(−0.55, 0.44)
Upper LoA	4.08	(3.42, 4.74)	-	-	5.06	(4.20, 5.91)	1.73	(1.09, 2.37)	4.43	(3.58, 5.28)
Lower LoA	−2.84	(−3.49, −2.18)	-	-	−3.96	(−4.82, −3.10)	−5.06	(−5.70, −4.41)	−4.54	(−5.39, −3.69)
MRE	3.46		-		4.51		3.39		4.48	

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.

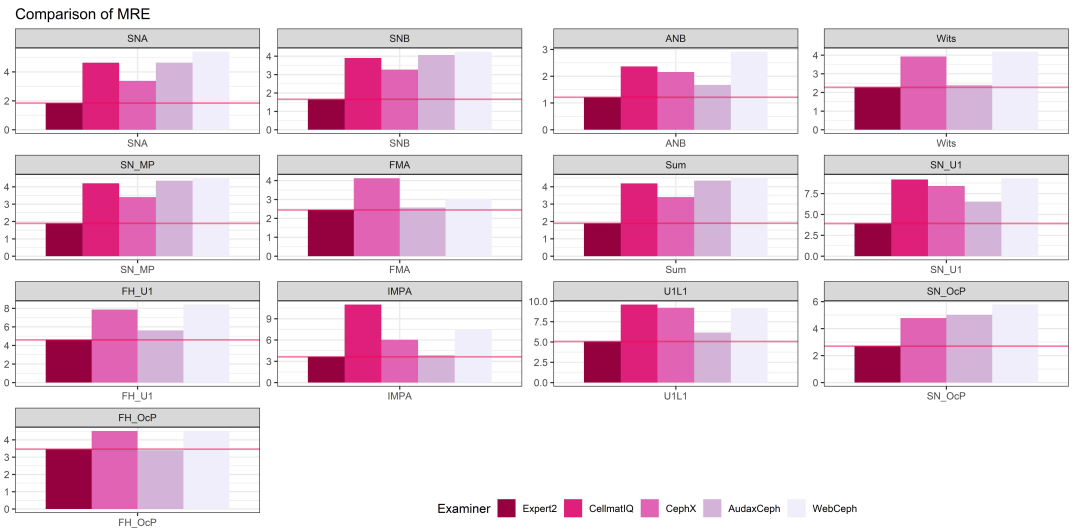

FIGURE 3. Bar graphs of the magnitude of MREs. The horizontal solid line represents the acceptable clinical limit for random error.

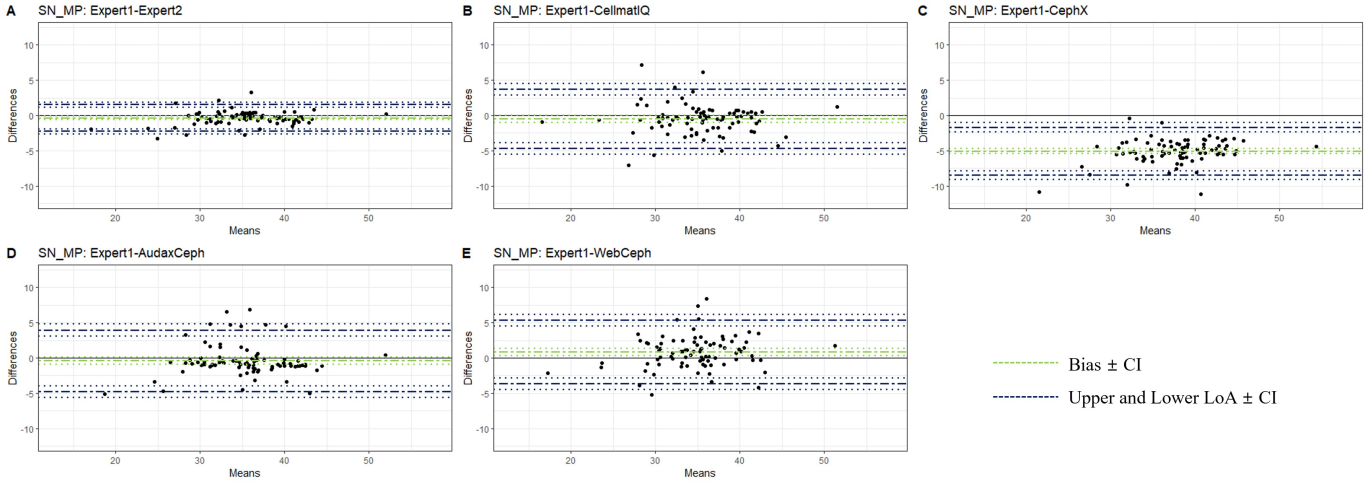

FIGURE 4. Bland-Altman plots for SN-MP. In each graph, the x-axis denotes the average of two examiners' measurement outcomes, while the y-axis is the difference between the examiners. Green dotted lines represent the bias with a 95% confidence interval. Blue dotted lines stand for the upper and lower limit of agreement with a 95% confidence interval. (A) between expert 1 and expert 2; (B) between expert 1 and CellmatIQ; (C) between expert 1 and CephX; (D) between expert 1 and AudaxCeph; (E) between expert 1 and WebCeph. CI: 95% confidence interval; LoA: limit of agreement.

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

AI	N	Variables
CellmatIQ	0	-
CephX	0	-
AudaxCeph	1	FH-OcP
WebCeph	0	-

the standard for AI is an overly strict criterion. However, fully automated cephalometric analysis users may range from beginners to highly experienced experts. Therefore, the practical performance of AI may need to be superior to at least that of a beginner, which is why regular orthodontists were set as the evaluation criteria in this study.

While the random error sizes of the AIs were approximately twice those of Expert 2, the corresponding reliability index (ICCs) was reduced by approximately 0.1, showing values between 0.8–0.9 (Table 8). This relatively small reduction of ICCs can be understood by regarding the definition of the reliability index. The reliability index estimates the effect of the measurement error in differentiating the difference between samples and is calculated as the variance between subjects/total variance, where total variance = variance between subjects + measurement error [14, 15]. For instance, an ICC of 0.8 means that 80% of the total variance in the measurement outcomes is due to the sample's unique variance, and 20% is owing to measurement error. The ICC formula implies that the magnitude of reliability can be less affected by the measurement error size if the variance between subjects is sufficiently large in the corresponding study group [15]. In fact, as shown in Table 1, the study samples include almost all anatomical features both anteroposteriorly and vertically, showing large variance. In summary, under a typical clinical situation that deals with various malocclusions, regular orthodontists can identify about 96% of group characteristics, while commercial AI can confirm about 86%.

In other words, if the same AI is applied to measure a group with a small variance, such as a class III or II high-angle group, the influence of the measurement error may increase, resulting in a smaller ICC value. This situation can arise especially when AI services are used for research. Bulatova *et al.* [10] suggested that AI-driven cephalometric analysis may facilitate research progress dealing with large samples. However, considering the relatively large measurement error and the dependence of reliability on the sample variance, caution may be needed to apply commercial AI in research.

According to the empirical standard, an ICC of 0.8–0.9 falls into the good to excellent range. However, it may be necessary to verify whether such measurement results are clinically acceptable. The dental/skeletal classification is an example of the application of cephalometric variables in actual clinical practice. Accordingly, classification reliability was evaluated by calculating Cohen's kappa coefficients, an indicator of the classification agreement between two examiners (Table 9). Expert 2 showed the highest mean reliability index of 0.76. On the contrary, the mean reliability indices of AIs were 0.53–0.68. Depending on individual variables, there were even cases where the reliability index was <0.1 (Björk-Jarabak sum by CephX). Caution must be exercised when claiming that these relatively low-reliability indices are clinically acceptable. The classification results using commercial AI should be carefully interpreted.

We could estimate the effect of bias by comparing absolute agreement and consistency ICCs (Table 8). The difference between these two ICC types was generally between 0.01 and 0.07, supporting our previous interpretation that bias was clinically insignificant in most variables. Exceptionally, the mandibular plane-associated variables evaluated by CephX showed substantial changes from 0.6 (absolute agreement) to 0.9 (consistency). This observation indicates the importance

TABLE 8. Inter-examiner reliability.

Examiner	Variables	Absolute Agreement		Consistency	
		Coefficient	95% CI	Coefficient	95% CI
Expert 1-Expert 2					
	SNA	0.96	(0.93, 0.98)	0.96	(0.95, 0.98)
	SNB	0.98	(0.96, 0.99)	0.98	(0.97, 0.99)
	ANB	0.98	(0.97, 0.99)	0.98	(0.97, 0.99)
	Wits	0.97	(0.95, 0.98)	0.97	(0.96, 0.98)
	SN-MP	0.98	(0.97, 0.99)	0.98	(0.97, 0.99)
	FMA	0.93	(0.59, 0.98)	0.96	(0.95, 0.98)
	Bjork-Jarabak Sum	0.98	(0.97, 0.99)	0.98	(0.97, 0.99)
	SN-U1	0.98	(0.96, 0.98)	0.98	(0.96, 0.98)
	FH-U1	0.95	(0.89, 0.98)	0.96	(0.95, 0.98)
	IMPA	0.96	(0.85, 0.98)	0.97	(0.96, 0.98)
	U1L1	0.98	(0.95, 0.99)	0.98	(0.97, 0.99)
	SN-OcP	0.92	(0.74, 0.96)	0.94	(0.91, 0.96)
	FH-OcP	0.89	(0.82, 0.93)	0.90	(0.85, 0.93)
	Mean	0.96		0.96	
Expert1-Cellmatiq					
	SNA	0.82	(0.74, 0.88)	0.82	(0.74, 0.88)
	SNB	0.90	(0.84, 0.93)	0.90	(0.84, 0.93)
	ANB	0.92	(0.86, 0.95)	0.93	(0.89, 0.95)
	Wits	-	-	-	-
	SN-MP	0.91	(0.87, 0.94)	0.92	(0.88, 0.95)
	FMA	-	-	-	-
	Bjork-Jarabak Sum	0.91	(0.87, 0.94)	0.92	(0.88, 0.95)
	SN-U1	0.85	(0.78, 0.90)	0.85	(0.78, 0.90)
	FH-U1	-	-	-	-
	IMPA	0.74	(0.63, 0.82)	0.74	(0.63, 0.83)
	U1L1	0.92	(0.87, 0.94)	0.92	(0.87, 0.95)
	SN-OcP	-	-	-	-
	FH-OcP	-	-	-	-
	Mean	0.87		0.88	
Expert1-CephX					
	SNA	0.89	(0.84, 0.93)	0.89	(0.84, 0.93)
	SNB	0.92	(0.88, 0.95)	0.92	(0.88, 0.95)
	ANB	0.94	(0.90, 0.96)	0.94	(0.91, 0.96)
	Wits	0.92	(0.86, 0.95)	0.93	(0.89, 0.95)
	SN-MP	0.62	(0.05, 0.89)	0.94	(0.91, 0.96)
	FMA	0.67	(0.08, 0.89)	0.89	(0.84, 0.93)
	Bjork-Jarabak Sum	0.62	(0.05, 0.89)	0.94	(0.91, 0.96)
	SN-U1	0.87	(0.81, 0.91)	0.87	(0.81, 0.91)
	FH-U1	0.88	(0.81, 0.92)	0.88	(0.81, 0.92)
	IMPA	0.94	(0.07, 0.88)	0.91	(0.86, 0.94)
	U1L1	0.91	(0.86, 0.95)	0.92	(0.88, 0.95)
	SN-OcP	0.81	(0.73, 0.88)	0.82	(0.73, 0.88)
	FH-OcP	0.81	(0.72, 0.87)	0.82	(0.73, 0.88)
	Mean	0.83		0.90	

TABLE 8. Continued.

Examiner	Variables	Absolute Agreement		Consistency	
		Coefficient	95% CI	Coefficient	95% CI
Expert1-Audax					
	SNA	0.78	(0.68, 0.85)	0.78	(0.68, 0.85)
	SNB	0.88	(0.82, 0.92)	0.88	(0.81, 0.92)
	ANB	0.96	(0.94, 0.97)	0.96	(0.94, 0.97)
	Wits	0.97	(0.96, 0.98)	0.97	(0.96, 0.98)
	SN-MP	0.91	(0.86, 0.94)	0.91	(0.86, 0.94)
	FMA	0.90	(0.23, 0.97)	0.96	(0.94, 0.97)
	Bjork-Jarabak Sum	0.91	(0.86, 0.94)	0.91	(0.86, 0.94)
	SN-U1	0.91	(0.83, 0.95)	0.93	(0.89, 0.95)
	FH-U1	0.89	(0.43, 0.96)	0.94	(0.91, 0.96)
	IMPA	0.83	(0.60, 0.97)	0.96	(0.94, 0.97)
	U1L1	0.94	(0.67, 0.98)	0.97	(0.95, 0.98)
	SN-OcP	0.81	(0.72, 0.87)	0.81	(0.72, 0.87)
	FH-OcP	0.82	(0.32, 0.93)	0.90	(0.85, 0.93)
	Mean	0.89		0.91	
Expert1-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	

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

TABLE 9. Classification reliability index (Cohen's kappa).

Variables	Expert 1-Expert 2	Expert 1-Cellmatiq	Expert 1-CephX	Expert 1-Audax	Expert 1-Webceph
SNA	0.65	0.48	0.62	0.50	0.30
SNB	0.93	0.61	0.79	0.78	0.53
ANB	0.76	0.69	0.76	0.83	0.44
Wits	0.77	-	0.64	0.82	0.55
FMA	0.70	-	0.25	0.66	0.76
SN-MP	0.86	0.70	0.22	0.70	0.62
Bjork-Jarabak Sum	0.82	0.66	0.09	0.60	0.44
SN-U1	0.84	0.65	0.67	0.68	0.52
FH-U1	0.76	-	0.63	0.73	0.63
IMPA	0.81	0.58	0.24	0.76	0.62
U1L1	0.76	0.69	0.75	0.83	0.68
SN-OcP	0.67	-	0.64	0.58	0.47
FH-OcP	0.56	-	0.52	0.37	0.48
Mean	0.76	0.63	0.53	0.68	0.54

All variables showed statistical significance with p-values less than 0.001.

of bias correction for these variables when using CephX.

Going through a review by a human examiner has been suggested as a countermeasure to compensate for the relatively low AI performance [12, 18]. However, in this case, the advantages of AI analysis may be limited. Reducing the tedious workload while eliminating subjective errors caused by human examiners may be the biggest advantage of AI cephalometric analysis [19, 20]. However, if a human examiner has to review and correct landmark positions individually, it is questionable whether time-saving and subjective error reduction can be obtained as expected. Some researchers argued that checking the landmarking by AI would be much easier than manual landmark detection; however, no concrete evidence supports this argument [18]. If the human examiner's calibration is inevitable in applying AI technology, the human factor may affect the clinical effectiveness of AI cephalometrics in some way. We will address this topic in a subsequent study (Part 2).

This study has several limitations. The soft tissue measurement was not included in the study variable. In addition, the AI comparison was limited to four commercial services. Furthermore, only two human examiners participated in setting the ground truth of the variables and evaluating inter-examiner errors. These limitations are related to the pilot nature of this study. The clinical significance of this study may be to provide relevant information to clinicians by quickly verifying the actual performance of commercially available AI cephalometric service in line with rapidly developing AI technology. Through this study, clinicians may get a brief clue about the current level of AI technology, but caution is needed when generalizing the results. Future studies are needed to supplement these issues.

5. Conclusions

In this study, the performance of commercially available AI cephalometric services was evaluated in comparison with the inter-examiner error between two regular human examiners. Within the limitations of this study, the following conclusions were drawn:

1. The systematic error in commercial AIs was clinically insignificant. However, the results of the mandibular plane-associated cephalometric variables of CephX must be interpreted cautiously.

2. The random error in commercial AI was significantly larger than that of human experts.

3. The reliability index of commercial AI was between 0.8 and 0.9, which corresponds to good to excellent levels according to empirical standard. However, caution may be needed when applied to a group with a small between-subject variance.

4. Cohen's kappa coefficients, which measure the actual classification reliability of commercial AIs, were only 0.5–0.6, which were clinically unacceptable.

Therefore, commercial AIs still cannot replace manual landmarking by human experts. A human examiner's landmark position review is mandatory when using commercial AIs.

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 and SRB—collected experimental data; HKN, SRB and JSL—analyzed the 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] 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.

[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] 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.

[9] Tanikawa C, Oka A, Lim J, Lee C, Yamashiro T. Clinical applicability of automated cephalometric landmark identification: Part II—number of images needed to re-learn various quality of images. Orthodontics & Craniofacial Research. 2021; 24: 53–58.

[10] Bulatova G, Kusnoto B, Grace V, Tsay TP, Avenetti DM, Sanchez FJC. Assessment of automatic cephalometric landmark identification using artificial intelligence. Orthodontics & Craniofacial Research. 2021; 24: 37–42.

[11] Schwendicke F, Chaurasia A, Arsiwala L, Lee J, 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.

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

[13] Kunz F, Stellzig-Eisenhauer A, Zeman F, Boldt J. Artificial intelligence in orthodontics: evaluation of a fully automated cephalometric analysis using a customized convolutional neural network. Journal of Orofacial Orthopedics. 2020; 81: 52–68.

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

[15] 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.

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

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

[18] 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.

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

[20] 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.

How to cite this article: Jaesik Lee, Seong-Ryeol Bae, Hyung-Kyu Noh. Commercial artificial intelligence lateral cephalometric analysis: part 1—the possibility of replacing manual landmarking with artificial intelligence service. Journal of Clinical Pediatric Dentistry. 2023; 47(6): 106-118. doi: 10.22514/jocpd.2023.085.