Accuracy and reliability of 2 artificial intelligence platforms for cephalometric analysis compared with a semiautomatic computer program

Ian Raby,[a] Victor Rojas,[a] Andres Celis,[b] Catalina García-Duhalde,[a] and Macarena Martinac[a,b]
Santiago, Chile

Introduction: Web-based platforms offer cephalometric tracing using artificial intelligence (AI) with varying performance levels. This study assessed the accuracy, reliability, and time efficiency of cephalometric tracings performed with the AI Web-based platforms WebCeph (Assemble Circle, Seoul, South Korea) and CephX (ORCA Dental AI, Las Vegas, Nev) in both their automated and corrected forms. **Methods:** Fifty pretreatment lateral cephalograms of patients were randomly selected and traced using AI platforms WebCeph and CephX in both their automated and landmark-corrected forms, along with the Dolphin Imaging software (version 13.01; Dolphin Imaging and Management Solutions, Chatsworth, Calif) as the "gold standard." Twelve parameters involving sagittal, vertical, dental, and soft-tissue dimensions were selected. The time required for each analysis was measured using a stopwatch. Intersystem comparisons were performed using ordinary least squares linear regression models, with Dolphin Imaging software as the reference. The intraclass correlation coefficient was used to determine the agreement among systems. A significance level of $P < 0.05$ was applied, and 95% confidence intervals were calculated for all outcomes. Clinically relevant differences were defined as angular discrepancies greater than 2° or linear discrepancies exceeding 2 mm. **Results:** The AI systems in their corrected form showed similar results to those of Dolphin Imaging software. If a 14% error is accepted, they were accurate and reliable in 11 of 12 parameters. Moreover, it was possible to reduce the tracing time by 46% compared with Dolphin Imaging software. The automated systems demonstrated low reliability and accuracy for cephalometric analysis. CephX and WebCeph are still not suitable for assessing soft-tissue parameters. **Conclusions:** CephX and WebCeph platforms for cephalometric tracing are valuable diagnosis tools only when landmark correction is applied. (Am J Orthod Dentofacial Orthop 2025;168:505-14)

Cephalometrics have long dominated the orthodontic literature, providing orthodontics with critical diagnostic tools, such as superimposition techniques, to isolate skeletal changes and tooth movement and to confirm the clinical assessment of a patient's craniofacial complex.[1] Even with the shift from 2-dimensional (2D) images to 3-dimensional analysis, driven by the increasing use of cone-beam computed tomography, the routine analysis of 2D cephalometrics remains one of the most economical and practical tools in orthodontic treatment.[2] Cephalometric radiographic analysis is based on the identification of radiological landmarks to subsequently measure various angles, distances, and ratios for the interpretation of craniofacial structures.[3] Cephalometric analyses are typically performed using a semiautomatic computer-based software program, which enables direct landmark identification on screen-displayed digital images.[4] However, tracing landmarks remains a manual task that an orthodontic expert must perform, and the process may be time-consuming.[5,6]

Various cephalometric analysis programs are available for professional use on personal computers or local networks. However, many of these software programs have complicated installation procedures, expensive subscriptions, and update fees and require multiple

[a]Department of Orthodontics, Faculty of Dentistry, Universidad de los Andes, Santiago, Chile.
[b]Department of Public Health, Faculty of Dentistry, Universidad de los Andes, Santiago, Chile.
All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest, and none were reported.
Address correspondence to: Victor Rojas, Department of Orthodontics, Faculty of Dentistry, Universidad de Los Andes, Avenida Monseñor Álvaro del Portillo 12.455, Las Condes 7620086, Santiago, Chile; e-mail, vhrojas@miuandes.cl.
Submitted, February 2025; revised and accepted, April 2025.
0889-5406/$36.00

https://doi.org/10.1016/j.ajodo.2025.04.011

tutorials and significant practice to master the program.[7] In addition, these programs require high hardware specifications to run and digital storage to maintain records and data.[8]

Currently, technological advances such as artificial intelligence (AI) have had a significant impact on the field of dentistry, particularly in digital cephalometrics.[5] AI is defined as the capability of a machine to imitate intelligent human behavior, and machine learning is one of the core subfields of AI that enables a computer model to learn and make predictions by recognizing patterns.[9] The subset of machine learning algorithms that contain multiple hidden layers is known as deep learning (DL). One type of DL is the convolutional neural network, which is specifically designed for image analysis.[5] Studies suggest that AI using convolutional neural network methods can become a powerful decision-making tool because of its ability to perform accurate and fast automatic cephalometric landmark detection.[5,10,11] However, studies on the accuracy and reliability of AI systems for landmark detection and cephalometric tracing have reported conflicting results. Hwang et al[11] reported an error in landmark detection of 1.46-2.97 mm between an AI model and humans, whereas Kim et al[12] reported a precision of 84% for landmark detection in another AI model. Moreover, these results would fall outside the range of clinical acceptance, which considers discrepancies >2 mm.

In addition, some studies have concluded that specific AI systems can be considered experts in landmark detection, whereas others report that correcting the location of landmarks is still necessary.[4-6,11-13] A systematic review by Hung et al[9] reported that the existing models of AI could be used for the preliminary localization of the cephalometric landmarks. However, manual correction is still necessary before further cephalometric analyses.

Today, many Web-based platforms offering cephalometric tracing using AI are available, providing clinicians with access to fully or semiautomated cephalometrics through DL algorithms without requiring excessive hardware specifications.

These AI clouds enable access from any device with an Internet connection and are compatible with all major operating systems.[14] In addition, orthodontists can obtain a cephalometric analysis within a few seconds, saving time.[7] WebCeph (Assemble Circle, Seoul, South Korea) and CephX (ORCA Dental AI, Las Vegas, Nev) are 2 of the most popular Web-based AI platforms for automated cephalometrics, with thousands of orthodontist subscribers worldwide. These AI platforms use DL algorithms; however, the specifics of their AI models are not publicly disclosed. Nevertheless, these platforms can automatically detect different cephalometric landmarks within seconds and allow for manual editing of the automatically calculated measurements.[15,16] Moreover, some studies have reported that when landmark correction is performed, the results can be compared with those obtained using programs such as Dolphin Imaging software (version 13.01; Dolphin Imaging and Management Solutions, Chatsworth, Calif) or FACAD software.[7,14] However, few studies have assessed the accuracy and reliability of cephalometric analysis on these platforms,[16,17] and tracing errors have been reported to lead to incorrect decision-making.[5] Yassir et al[16] concluded that these AI platforms should be used with caution when performing cephalometric analysis.

As more orthodontists adopt these AI cephalometric platforms, evaluating their performance has become increasingly essential. Therefore, this study aimed to assess the accuracy, reliability, and cephalometric tracing time of the AI platforms WebCeph and CephX in both their automatic and corrected forms and to compare them with the ground truth, represented by manual tracings performed by an experienced orthodontist using the Dolphin Imaging software.

MATERIAL AND METHODS

Pretreatment digital lateral cephalograms of 50 orthodontic patients who attended the Department of Orthodontics at the Faculty of Dentistry of the Universidad de los Andes between 2019 and 2023 were randomly selected using random-generated numbers from the orthodontic department database for this study. The sample was based on previous studies by Alqahtani,[7] Jeon and Lee,[6] and Mahto et al.[15] The study was approved by the Ethical Committee and Institutional Review Board of the Universidad de los Andes (CPI-ODO:05) and was conducted by the principles of the Declaration of Helsinki. Informed consent was obtained from all participants. All digital cephalograms were taken using the same radiographic unit (D-64625; Sirona Dental Systems GmbH, Bensheim, Germany), after the radiology protocol of the Universidad de los Andes (9.4 seconds, 77 kV, 14 mA), with a calibration ruler positioned to the side. No differentiation was made for age, gender, type of malocclusion, or skeletal pattern. Patient identifiers (name, age, gender, and date of examination) were removed from the original lateral cephalograms to maintain patient anonymity, and each cephalogram was assigned a number.

Inclusion criteria included (1) good quality image, (2) absence of artifacts, (3) a calibration ruler at the side, and (4) radiographs taken by the same radiographic unit. In contrast, the exclusion criteria included (1) radiographs that showed major asymmetry, such as

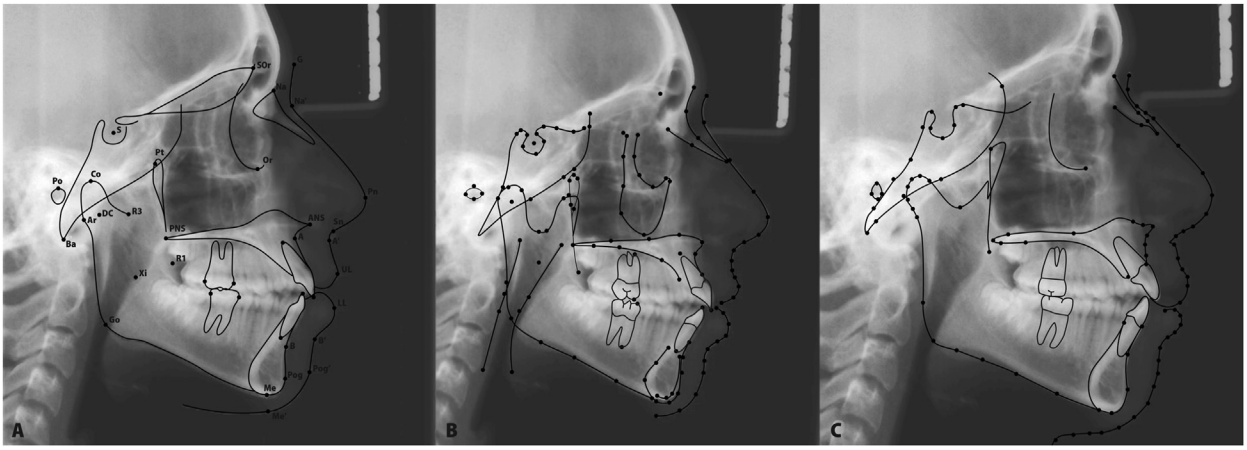

Fig 1. Lateral cephalometric radiograph traced with the fully automated landmark identification system: **A,** WebCeph AI platform; **B,** CephX AI platform; **C,** Dolphin computer program.

significant double borders of the mandible; (2) lack of resolution that could interfere with landmark location, and (3) craniofacial deformity.

An account was created on the CephX (https://cephx.com) and WebCeph (https://webceph.com) AI platforms using Google Chrome (Google LLC, Mountain View, Calif) as the standard Internet browser. Each participant was registered on each platform, and the digital radiographs were uploaded in JPG format. These AI Web sites locate cephalometric landmarks within seconds, enabling subsequent operator adjustments. In addition, it is possible to customize the cephalometric analysis and parameters according to the orthodontist's needs. The Dolphin Imaging software was used as the ground truth and compared with measurements obtained by WebCeph and CephX in the automatic form and corrected (with landmark correction). The digital films were calibrated by digitizing 2 points (10 mm) on the ruler within the digital cassette. Landmark identification was performed manually using a mouse-driven cursor for Dolphin Imaging software and the AI platforms with landmark correction. Figure 1, *A-C* shows an example of each system with the same cephalometric radiograph.

A total of 12 cephalometric parameters were selected for measurement and comparison: 4 sagittal parameters (ANB, SNA, SNB, and facial convexity), 3 vertical parameters (SNGoGn, facial axis, and PP-PM), 4 dental parameters (U1-APo, L1-Apo, U1-PP, and IMPA), and 1 soft-tissue parameter (NLA). The definition of the parameters is shown in Table I.

These measurements were selected because ANB, SNA, SNB, SNGoGn, and IMPA are parameters that are commonly used by the American Board of Orthodontics when analyzing clinical patients. PP-PM, U1-APo, L1-Apo, U1-PP, and NLA are parameters that have not been previously reported but are used in the Universidad de los Andes cephalometric analysis. Moreover, the time required to obtain each cephalometric analysis was measured using a stopwatch for each system. Once the patient information was registered in the selected system and the cephalometric radiograph was uploaded, the time was ready to be assessed. The timer started when the landmark location began and stopped when the cephalometric tracing was completed, and calibration was performed.

The cephalometric measurements obtained for the 5 systems were downloaded in PDF and entered into the same Microsoft Excel spreadsheet.

To eliminate interobserver variability, all measurements were performed by a single examiner. To evaluate intraobserver reliability and reproducibility, a 2-stage training and calibration stage was performed. Fifteen radiographs after the inclusion and exclusion criteria but not included in the study sample were randomly selected using random-generated numbers from the orthodontic department database. These cephalograms were traced in a random order for each system. Measurements were obtained at 14-day intervals among stages, and the intraclass correlation coefficient (ICC) was used to assess the training phase. At the time the study was conducted, the examiner had 5 years of experience using the Dolphin Imaging software and 2 years of experience using CephX and WebCeph.

Statistical analysis

The Shapiro-Wilk test was used to assess the normality of the data distribution. Descriptive statistics,

Table I. Description of cephalometric parameters used in the study

Parameter	Definition and interpretation
Sagittal	
ANB, °	The angle formed among point A, nasion, and point B. Measures the skeletal class in the sagittal direction
SNA, °	The angle formed among sella, nasion, and tip A. Measures the sagittal position of the maxilla
SNB, °	The angle formed between sella, nasion, and tip B. Measures the sagittal position of the mandible
Facial convexity, mm	The distance that considers the facial plane (N-Pog) and point A. It measures the skeletal class in a sagittal direction
Vertical	
SN-GoGn, °	The angle formed between the gonion and gnation lines with the sella and nasion (mandibular plane to anterior skull base)
Facial axis, °	The angle formed by the intersection of the Ba-Na plane with the Pt-Gn line. The posterior angle is measured. It measures the mandibular divergence
PP-PM, °	Angle formed by the intersection of the palatal plane (ENA and ENP) and the mandibular plane (Go-Me). It indicates the vertical relationship between the jaws
Dental	
U1-Apo, mm	Distance from the maxillary incisor to the line generated by point A to Pogonion. Measures the position of the incisor
L1-Apo, mm	Distance from the mandibular incisor to the line generated by point A to Pogonion. Measures the position of the incisor
U1-PP, °	The angle formed by the longitudinal axis of the maxillary incisor and the palatal plane. Measures the inclination of the incisor
IMPA, °	The angle formed by the longitudinal axis of the mandibular incisor and the mandibular plane. It measures the inclination of the incisor
Soft tissue	
NLA, °	The angle formed by the columella-subnasal-upper lip. Measures the protrusion of the upper lip

including the minimum, maximum, mean, and standard deviation, were calculated to summarize the data. Intersystem comparisons were performed using ordinary least squares linear regression models with Dolphin Imaging software as the reference. The agreement among the measurements obtained from Dolphin Imaging software (reference), WebCeph (automated and corrected), and CephX (automated and corrected) was evaluated using the ICC. For this analysis, ICC threshold values were interpreted based on previous studies by Mahto et al[15] and Durão et al[18] as follows: <0.75 indicated poor-to-moderate agreement, 0.75-0.90 indicated good agreement, and >0.90 indicated excellent agreement. Clinically relevant differences were defined as angular discrepancies greater than 2° or linear discrepancies exceeding 2 mm. All statistical analyses were conducted using SAS software (version 9.4; SAS Institute Inc, Cary, NC). A significance level of $P < 0.05$ was applied, and 95% confidence interval values were calculated for all outcomes.

RESULTS

A total of 50 pretreatment digital lateral cephalograms were traced using 3 software programs and 5 systems: Dolphin Imaging software, CephX auto, CephX corrected, WebCeph auto, and WebCeph corrected, for 12 cephalometric parameters and time, resulting in a total of 3000 items for analysis.

The ICC for intraexaminer calibration was 0.994, obtaining high reliability and reproducibility. In addition, the distribution of the variables was normal ($P > 0.05$).

The minimum, maximum, mean, standard deviation, and time values for the measurements of the 5 systems were compared and are presented in Table II. The parameters ANB, SNA, SNB, facial convexity, facial axis, U1-APo, and L1-Apo were within the clinically accepted range based on their mean values across all systems with Dolphin Imaging software. The measurements for SNGoGn, PP-PM, U1-PP, IMPA, and NLA were at least 2 points beyond the clinically accepted range. Regarding the systems, CephX auto displayed 5 errors, WebCeph auto showed 3 errors, and CephX corrected, and WebCeph corrected each displayed only 2 errors. These 2 errors were related to the NLA and PP-PM parameters, which appeared in all 4 systems.

Intersystem comparisons were performed using ordinary least squares linear regression models, with Dolphin Imaging software as the reference. A significance level of $P < 0.05$ was applied, and 95% confidence interval values were calculated for all outcomes. The results for the 4 systems with Dolphin Imaging software as the reference are presented in Table III.

Regarding the systems individually, only 4 parameters showed significant differences: SNGoGn, PP-PM, U1-PP, and NLA. SNGoGn presented significant differences between Dolphin Imaging software and CephX

Table II. Comparison of skeletal, dental, and soft-tissue measurements and time for 4 different cephalometric analysis methods

	Dolphin Imaging software			*CephX auto*			*CephX corrected*			*Webceph auto*			*Webceph corrected*		
Measurements	*Mean ± SD*	*Min*	*Max*	*Mean ± SD*	*Min*	*Max*	*Mean ± SD*	*Min*	*Max*	*Mean ± SD*	*Min*	*Max*	*Mean ± SD*	*Min*	*Max*
ANB, °	4.30 ± 2.21	−1.11	7.94	3.88 ± 2.33	−2.63	8.16	4.25 ± 2.23	−1.51	7.92	4.02 ± 2.28	−1.33	9.71	4.01 ± 2.15	−1.56	7.74
SNA, °	82.46 ± 3.15	76.98	89.42	82.52 ± 3.19	76.64	89.16	82.73 ± 3.11	77.99	89.46	83.32 ± 3.31	77.62	91.17	82.47 ± 3.17	76.94	89.46
SNB, °	78.18 ± 3.94	70.65	87.77	78.69 ± 3.95	71.55	89.21	78.46 ± 4.01	70.45	89.64	79.29 ± 3.45	72.41	89.33	78.44 ± 3.94	70.45	89.64
Convexity, mm	3.91 ± 2.43	−1.94	7.63	3.66 ± 2.33	0.04	7.68	3.98 ± 2.36	0.13	7.91	3.36 ± 2.41	−2.82	8.21	3.69 ± 2.33	−1.06	7.63
SNGoGn, °	33.49 ± 5.63	19.23	47.81	38.02 ± 5.21	23.23	50.39	34.15 ± 5.64	20.46	48.93	31.99 ± 5.13	17.12	45.92	33.43 ± 5.34	19.85	48.54
Facial axis, °	87.82 ± 4.32	80.81	97.22	86.73 ± 4.16	79.76	97.66	87.79 ± 4.39	81.35	98.05	86.82 ± 4.12	80.88	97.76	87.71 ± 4.31	81.05	98.82
PP-PM, °	32.79 ± 40.58	12.53	38.54	28.65 ± 4.65	15.20	39.36	27.24 ± 4.93	12.21	40.13	24.56 ± 4.75	12.31	35.55	26.76 ± 4.91	11.08	39.45
U1-Apo, mm	6.77 ± 2.63	2.93	12.59	7.45 ± 2.51	2.92	12.65	6.82 ± 2.56	2.99	12.24	6.55 ± 2.71	1.58	11.84	6.61 ± 2.54	2.82	12.78
L1-Apo, mm	2.46 ± 2.35	−1.82	9.71	2.58 ± 2.16	−1.17	7.66	2.36 ± 2.34	−2.01	8.76	2.57 ± 2.38	−1.95	7.97	2.37 ± 2.32	−1.96	8.97
U1-PP, °	112.96 ± 5.36	100.08	123.72	115.21 ± 4.43	105.71	124.60	113.61 ± 5.49	101.10	125.12	115.57 ± 5.67	107.26	128.61	113.65 ± 5.18	102.12	125.22
IMPA, °	92.60 ± 6.32	83.47	111.45	90.40 ± 5.88	80.62	103.69	92.61 ± 5.99	82.63	109.68	92.58 ± 5.32	81.48	104.17	93.06 ± 5.74	83.33	109.45
NLA, °	101.69 ± 9.77	87.54	138.21	109.78 ± 9.35	92.48	131.59	110.09 ± 10.02	92.48	138.21	98.7 ± 14.98	73.60	138.21	97.33 ± 14.96	68.72	142.12
Time, s	127.81 ± 12.85	100.96	155.23	16.73 ± 0.77	15.60	20.08	76.35 ± 6.98	63.86	93.17	6.48 ± 0.65	5.59	8.60	61.37 ± 11.34	39.94	100.37

Min, minimum; *Max*, maximum; *SD*, standard deviation.

auto. PP-PM presented significant differences between Dolphin Imaging software and WebCeph auto. U1-PP presented significant differences between Dolphin Imaging software and both CephX auto and WebCeph auto. NLA presented significant differences with all systems.

On the basis of these results, when considering significant differences for the parameters SNGoGn, PP-PM, U1-PP, and NLA, WebCeph corrected, and CephX corrected exhibited the least variation, with 1 error each, followed by the automated versions of both programs, which each exhibited 3 errors.

The ICC values for repeated cephalometric measurements are reported in Table IV. The ICC was >0.9 for CephX corrected and WebCeph corrected in 10 of the 12 parameters, indicating excellent agreement. Two measurements were <0.75, indicating poor agreement. CephX auto had 8 of 12 parameters with an ICC >0.9, 2 measurements between 0.75-0.90, and 2 parameters with an ICC <0.75. WebCeph auto had 7 of 12 parameters with an ICC >0.9, 3 measurements between 0.75-0.90, and 2 parameters <0.75. PP-PM and NLA for all systems were the parameters with poor agreement. Figure 2 shows dispersion graphs for all systems, illustrating the range of clinical acceptance when there is a variation of 2° or 2 mm.

For the automatic AI systems, both software programs had 4 of the 12 parameters within the accepted range: CephX corrected had 10 of 12 parameters within the range, whereas WebCeph corrected had 9 of 12 parameters. If a 14% measurement error is accepted, both CephX corrected and WebCeph corrected presented 11 of 12 parameters within the acceptable range.

Finally, the time dimension was assessed for each system, obtaining a mean value of 6.4 seconds for WebCeph auto, 16.73 seconds for CephX auto, 61.37 seconds for WebCeph corrected, and 76.35 seconds for CephX corrected. The mean tracing time for Dolphin Image software was 127.81 seconds. As expected, the AI systems were faster than manual tracing. In addition, the time taken was significant when comparing all systems individually with Dolphin Image software.

The parameters within the range of clinical acceptance were ANB, U1-APo, L1-Apo, and facial convexity. The parameters that were reliable only with landmark correction were SNA, SNB, and facial axis. The parameters considered reliable only for the corrected systems when accounting for a 14% error, were SNGoGn, IMPA, U1-PP, and PP-PM. The NLA parameter was not reliable for any system.

Considering the final results in terms of accuracy, reliability, and tracing time for cephalometric measurements, CephX corrected was the best system. WebCeph

Table III. Comparison of the intersystem least squares means for effect software

		Least squares mean for effect software (t for H0:Pr)			
Parameters	*P* value	CephX auto	CephX corrected	Webceph auto	Webceph corrected
ANB	0.87	0.35	0.91	0.54	0.51
SNA	0.62	0.92	0.66	0.17	0.98
SNB	0.66	0.51	0.72	0.15	0.73
Convexity	0.71	0.62	0.86	0.25	0.66
SNGoGn	<0.0001*	<0.001*	0.54	0.16	0.95
Facial axis	0.51	0.21	0.99	0.24	0.91
PP-PM	0.02*	0.26	0.13	0.02*	0.10
U1-Apo	0.41	0.14	0.80	0.77	0.86
L1-Apo	0.98	0.80	0.83	0.81	0.84
U1-PP	0.04*	0.03*	0.53	0.01*	0.51
IMPA	0.17	0.06	0.99	0.98	0.73
NLA	<0.0001*	0.0009*	0.0006*	0.02*	0.03*
Time	<0.0001*	<0.0001*	<0.0001*	<0.0001*	<0.0001*

*Values with significant differences ($P < 0.05$).

Table IV. ICC for reproducibility of each cephalometric analysis method with Dolphin Imaging software as ground truth

Parameters	*CephX AU*	*CephX CO*	*Webceph AU*	*Webceph CO*	*ICC, 95% confidence interval*
ANB	0.95	0.98	0.83	0.97	0.87-0.93
SNA	0.93	0.92	0.81	0.93	0.83-0.91
SNB	0.96	0.95	0.91	0.96	0.87-0.93
Convexity	0.87	0.97	0.91	0.97	0.87-0.93
SNGoGn	0.95	0.98	0.92	0.97	0.73-0.85
Facial axis	0.94	0.97	0.95	0.97	0.92-0.96
PP-PM	0.25	0.32	0.19	0.29	−0.01 to 0.18
U1-APO	0.94	0.98	0.93	0.97	0.91-0.95
L1-APO	0.92	0.97	0.91	0.97	0.92-0.96
U1-PP	0.86	0.97	0.81	0.98	0.78-0.88
IMPA	0.91	0.98	0.83	0.98	0.84-0.91
NLA	0.63	0.71	0.57	0.65	0.38-0.59

AU, automated; *CO*, corrected.

corrected had a shorter tracing time but lower levels of accuracy and reliability, which were deemed insignificant. Moreover, both systems yielded similar results, and if we accept a 14% error margin, tracing time can be reduced by 46% compared with Dolphin Image software. Automatic systems, although they have very low tracing times, exhibit lower reliability and accuracy in cephalometric analysis.

DISCUSSION

Cephalometric measurements were used instead of landmark identification in this study, as measurements are the end products of the cephalometric tracing process and provide data for the treatment plan.[15] In addition, errors in landmark detection, when used in combination to obtain measurements, may either cancel each other out or increase the discrepancy.[19,20] The parameters chosen in this study included sagittal, vertical, dental, and soft-tissue dimensions for a more reliable comparison. As interexaminer errors are more frequent than intraexaminer errors,[16] all measurements were taken by a previously calibrated single operator (ICC = 0.994). A clinically relevant difference was determined when the difference in the angular and linear measurements was greater than 2° or 2 mm.[9]

This study provided a detailed analytical assessment of the accuracy and reliability of linear and angular cephalometric measurements obtained using WebCeph and CephX. Overall, AI platforms with landmark correction performed well when using the Dolphin Imaging program as a reference; 11 of 12 parameters were reliable, accepting a 14% error margin. However, automatic tracing by these platforms was not trustworthy, as it presented only 4 of 12 parameters within the accepted range. This finding aligns with those of other studies, including those by Jeon and Lee,[6] Nishimoto et al[21,9],

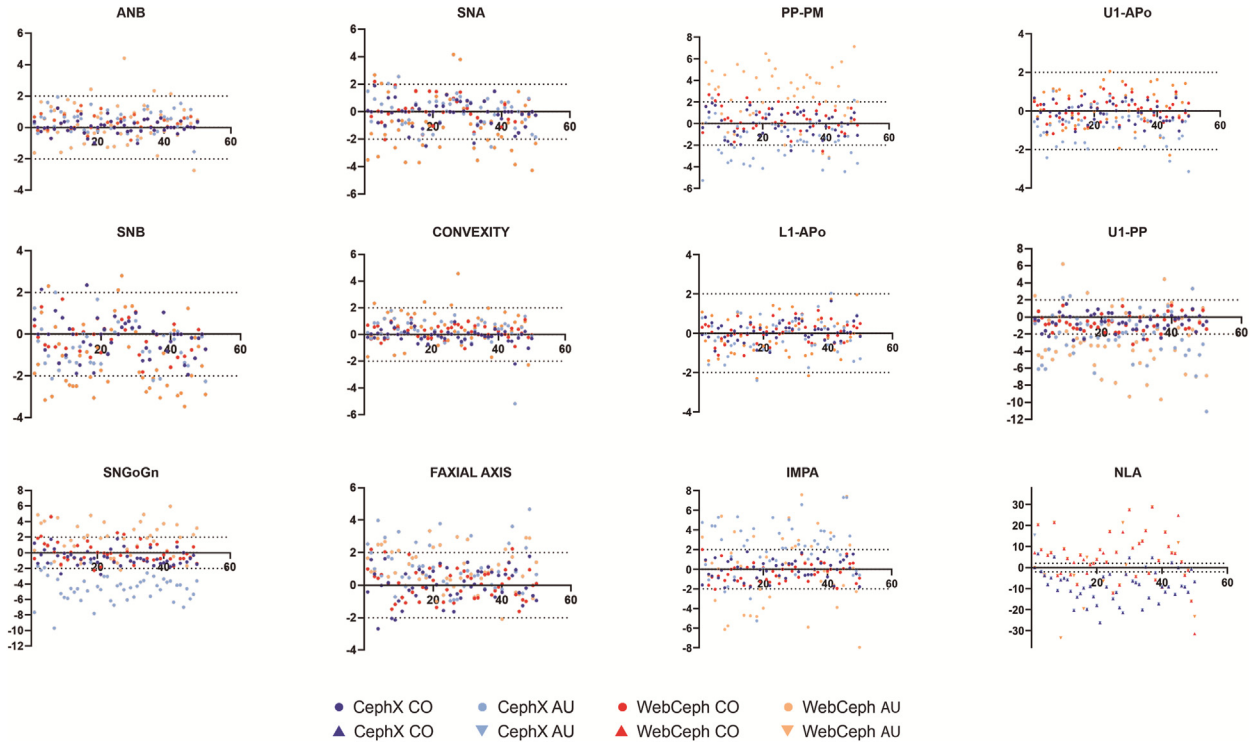

Fig 2. Dispersion graphs for all systems in which value 0 is for Dolphin as reference standard. Plots of the measurements obtained with CephX CO, CephX AU, WebCeph CO, and WebCeph AU. *Dashes* indicate the cutoff for clinically relevant differences of 2° or 2 mm. *CO*, corrected; *AU*, automated.

Hung et al,[21,9] and Meriç and Naoumova,[14] who recommended using landmark correction when employing automated traces, as AI algorithms are still not sufficiently developed to be reliable for all parameters. Meriç and Naoumova[14] reported similar results between Dolphin Imaging software and CephX corrected. Alqathtani[7] reported reliable measurements between the FACAD computer program and CephX corrected. However, Yassir et al[16] found poor landmark identification and inconsistent measurements in the automatic WebCeph; they also reported that WebCeph corrected could overcome some of the limitations of the automatic form. Mahto et al[15] reported that orthodontic diagnosis and treatment planning based on cephalometric measurements obtained from automated WebCeph could be misleading and that manual landmark correction may improve the accuracy of these measurements. This finding was also consistent with that of El-Dawlatly et al,[22] who mentioned that WebCeph corrected delivered similar results to traditional methods of tracing.

ANB, U1-APo, L1-Apo, and facial convexity were parameters that presented accuracy and reliability for all 4 systems with Dolphin Imaging software. These parameters have been reported as reliable in other studies assessing AI performance.[4,6,7] Regarding other parameters, difficulty in localizing landmarks such as the nasion, gnathion, gonion, and mandibular incisor apex has been reported, even by orthodontic experts.[3,5-7,9] Jeon and Lee,[6] and Meriç and Naoumova,[14] have reported that SNGoGn and IMPA showed significant variation in automated cephalometrics when assessing accuracy. In addition, a systematic review by Hung et al[9] confirmed that these parameters were difficult to localize for automated AI systems and that the algorithm has not yet been sufficiently developed to detect soft-tissue landmarks, which helps explain the differences found between the systems and Dolphin Imaging software. In our results, NLA was the only parameter that was not reliable for any system when compared with Dolphin Imaging software, which is consistent with the findings of Hung et al.[9] Neither AI platform is precise enough to detect soft-tissue landmarks within the range of clinical acceptance. Perhaps, in future studies, when the algorithm is more developed, this parameter should be reassessed.

ICCs showed good agreement for all systems in 10 of the 12 parameters, with ICC values >0.75. CephX corrected and WebCeph corrected showed 10 of 12

parameters with an ICC value >0.90, indicating excellent agreement. For automated CephX, 8 of 12 parameters and for automated WebCeph, 7 of 12 parameters had ICC values >0.90, but not necessarily an ICC >0.75, which is considered good agreement and is within the range of clinical acceptance. When considering accuracy and reliability, the minimum acceptable value should be an ICC >0.90 for all parameters. As reported by Kazimierczak et al,[17] accuracy sharply decreased when the threshold was <2 mm. A systematic review by de Queiroz et al,[19] showed that the agreement between AI and manual detection ranged 79%-90%, depending on the margin of error, with a mean divergence of 2.05 compared with manual landmarking. Reasons for lower ICC values may include the software program's faulty identification of landmarks.[15] In addition, the AI systems do not allow for a fine layout for landmark detection as with Dolphin Imaging software.

In the dispersion graphs, we sought to measure differences in the range of clinical acceptance within ± 2° or mm among the 4 systems with Dolphin Imaging software. These differences were mapped and identified (Fig 2) to assess which parameters were more accurate and reliable, considering a 10% error threshold for data that may have been affected by patient variation, such as the double contour of bony structures or anatomic discrepancies. Analyzing the data dispersion, we observed a positive trend for WebCeph and a negative trend for CephX. The parameters with the most significant variation were IS-PP, SNGoGn, and IMPA for automated AI and NLA for all systems. When assessing the mean value of IMPA for automated WebCeph, the dispersion of data presented a negative and a positive tendency that compensated for the mean value of this parameter. This suggests that assessing only mean values for cephalometric tracing using AI should be used with caution.

Our final objective was to assess the time required for cephalometric tracing across systems and how it relates to accuracy and reliability. Tracing time was fast for both automated and corrected systems. When comparing traditional tracing in Dolphin Imaging software with automated AI systems, the time difference was considerable. The mean tracing times for WebCeph auto, CephX auto, WebCeph corrected, and CephX corrected were 6.4, 16.73, 61.37, and 76.35 seconds, respectively. The mean tracing time for Dolphin Imaging software was 127.81 seconds, similar to the time reported by Meriç and Naoumova,[14] of 129.4 seconds. Jeon and Lee[6] reported a 6-minute mean for conventional tracing with the OrthoCeph program. In Meriç and Naoumova's study,[14] CephX corrected had a mean time of 58.7 seconds, and El-Dawlatly et al,[22] reported a mean time for WebCeph corrected of 31.07 seconds. Although these times were faster than ours, the differences may be attributed to factors such as Internet connection, the number of users connected at the time of tracing, or variations in how time was measured. Nevertheless, even though time was a significant factor between Dolphin Imaging software and the automated AI systems, the tracing algorithm for automated WebCeph and CephX requires further development to improve accuracy and reliability. AI systems with landmark correction presented results similar to those of Dolphin Imaging software in considerably less time, which is consistent with the findings of Jeon and Lee,[6] Livas et al[4,14], Meriç and Naoumova,[4,14] and El-Dawlatly et al.[22]

Considering the final results in terms of accuracy, reliability, and time for cephalometric traces, CephX corrected was the best system. WebCeph corrected presented a shorter tracing time but lower levels of accuracy and reliability, which were considered of no significance. Moreover, both systems yielded similar results if we accept a 14% error margin, and it is possible to reduce tracing time by 46% compared with Dolphin Imaging software. Although presenting 91% less tracing time, automatic systems are less accurate and reliable than expected.

As mentioned by El-Dawlatly et al,[22] the primary purpose of developing fully automated tracing software is to reduce the time required by clinicians to locate landmarks. Testing the accuracy of these technologies is crucial to confirm their reliability and to provide clinicians with a more straightforward approach to cephalometric analysis. Moreover, A clinically relevant difference threshold of 2 points is essential when considering supplementary information for diagnosis. A difference between an ANB of 0° and an ANB of 2° may influence decision-making with borderline extraction, especially for inexperienced orthodontists. As reported by Devereux et al,[23] a significant change in the extraction decision occurred when a lateral cephalometric radiograph was provided to a group of orthodontists for a single patient. If cephalometric analyses are inaccurate, it may lead to an incorrect diagnosis. Durão et al,[24] also reported that a significant number of orthodontists consider cephalometric analysis necessary when producing a treatment plan. In addition, a systematic review by Durão et al[25] highlighted that one of the significant benefits of 2D cephalometry is that it can lead to changes in the treatment plans compared with when only clinical evaluation is performed.

The workflow of these AI platforms is straightforward. They offer several advantages, including easy access, competitive subscriptions, and a significant

improvement in the efficiency of orthodontists when conducting cephalometric analysis in both routine clinical practice and research.[15] Nevertheless, time savings do not justify the lack of oversight when accuracy is crucial, especially in analyzing borderline extraction patients, in which the cephalometric analysis may significantly impact decision-making. In addition, because a large group of clinicians uses these platforms, it is essential to inform them that the accuracy of the analysis could influence their final diagnosis.

As of today, it remains imperative that landmarks and tracings obtained through fully automated software programs be supervised by an experienced orthodontist. It should also be mandatory for clinicians to perform landmark corrections to achieve accurate final readings, regardless of the AI platform used.

The limitations of our study include the use of a single examiner, the specific patient population, and the limited parameters chosen for analysis. Regarding the use of a single examiner, we aimed to replicate the experience of an orthodontist when choosing one program over another, as well as the different results that may be obtained by selecting a particular system. Although the examiner was an expert in all systems involved and rest periods were considered, a potential risk of bias remains. A single examiner may tend to reproduce the same error in locating a specific landmark. However, this error may be replicated across other systems, and there may be compensation for this discrepancy. Moreover, a single examiner may be more comfortable with a specific system over others, potentially leading to errors in how the cephalometric trace is performed. Future studies would benefit from analyzing measurements across a panel of experts, automated AI, and landmark correction, similar to the study by Hwang et al,[11] in which a panel of experts was calibrated and evaluated an automated AI model for cephalometric analysis. However, seasoned orthodontists tend to prefer traditional, time-proven software such as Dolphin Imaging software over AI platforms for cephalometrics.

The specific patient population, small sample size, and limited parameter sets may affect the generalizability of our results. The orthodontic database used is primarily composed of Caucasian and Hispanic populations, and the results may not be extrapolated to other ethnic groups. In addition, only 12 cephalometric parameters were selected to avoid exceeding 3000 data points for analysis, as this was an exploratory study on AI performance. We suggest that future studies consider involving a more diverse range of ethnic groups, a larger sample size, and additional parameters covering all dimensions, along with improvements to the AI models' algorithms to generate more comprehensive results.

Although cephalometric measurements carried out using Dolphin Imaging software are considered highly reliable, manual tracing is still regarded as the gold standard in cephalometrics. Future studies are needed to investigate the accuracy and reliability of the systems involved, considering the ground truth provided by manual tracings performed by a panel of orthodontic experts.

CONCLUSIONS

AI Web platforms for cephalometric tracing in their automated form require further development of their algorithms to be comparable to Dolphin Imaging software. When landmark correction is performed, it delivers similar results in a shorter time. CephX and WebCeph are still not capable of assessing soft-tissue parameters.

Given the accessibility of complex programs such as Dolphin Imaging software to orthodontists, the CephX and WebCeph platforms for cephalometric tracing serve as valuable complementary diagnostic tools, but only when landmark correction is applied. From a critical clinical perspective, these AI platforms are not yet capable of replacing trained orthodontists in cephalometric tracing or specialized programs such as Dolphin Imaging software.

As a practical recommendation for clinicians, these platforms can serve as a guide for an initial, rapid cephalometric analysis; however, manual verification is essential, and they should not be relied on as standalone tools. If a more detailed analysis is required, we recommend using time-proven software such as Dolphin Imaging software.

AUTHOR CREDIT STATEMENT

Ian Raby contributed to conceptualization, methodology, investigation, validation, visualization, data curation, original draft preparation, manuscript review and editing, supervision, and project administration; Victor Rojas contributed to investigation, manuscript review and editing, supervision, project administration, and methodology; Andres Celis contributed to formal analysis, methodology, and manuscript review and editing; Catalina García-Duhalde contributed to manuscript review and editing; and Macarena Martinac contributed to manuscript review and editing.

ACKNOWLEDGMENTS

The authors thank Nicole Feliu and Dr Manuel Carrasco for their invaluable support and guidance, as well as Enid Rosenstiel, MA (Columbus, Ohio), for her professional editing of this article in English.

REFERENCES

1. Hans MG, Palomo JM, Valiathan M. History of imaging in orthodontics from Broadbent to cone-beam computed tomography. Am J Orthod Dentofacial Orthop 2015;148:914-21.
2. Hassan MM, Alfaifi WH, Qaysi AM, Alfaifi AA, AlGhafli ZM, Mattoo KA, et al. Comparative evaluation of digital cephalometric tracing applications on mobile devices and manual tracing. Med Sci Monit 2024;30:e944628.
3. Arık SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging (Bellingham) 2017;4:014501.
4. Livas C, Delli K, Spijkervet FKL, Vissink A, Dijkstra PU. Concurrent validity and reliability of cephalometric analysis using smartphone apps and computer software. Angle Orthod 2019;89:889-96.
5. 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. J Orofac Orthop 2020;81:52-68.
6. Jeon S, Lee KC. Comparison of cephalometric measurements between conventional and automatic cephalometric analysis using convolutional neural network. Prog Orthod 2021;22:14.
7. Alqahtani H. Evaluation of an online website-based platform for cephalometric analysis. J Stomatol Oral Maxillofac Surg 2020;121:53-7.
8. Kumar M, Kumari S, Chandna A, Konark SA, Singh A, Kumar H, et al. Comparative evaluation of CephNinja for android and NemoCeph for computer for cephalometric analysis: a study to evaluate the diagnostic performance of CephNinja for cephalometric analysis. J Int Soc Prev Community Dent 2020;10:286-91.
9. Hung K, Montalvao C, Tanaka R, Kawai T, Bornstein MM. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: a systematic review. Dentomaxillofac Radiol 2020;49:20190107.
10. Mohammad-Rahimi H, Nadimi M, Rohban MH, Shamsoddin E, Lee VY, Motamedian SR. Machine learning and orthodontics, current trends and the future opportunities: a scoping review. Am J Orthod Dentofacial Orthop 2021;160:170-92.e4.
11. 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? Angle Orthod 2020;90:69-76.
12. Kim H, Shim E, Park J, Kim YJ, Lee U, Kim Y. Web-based fully automated cephalometric analysis by deep learning. Comput Methods Programs Biomed 2020;194:105513.
13. 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. Angle Orthod 2019;89:903-9.
14. Meriç P, Naoumova J. Web-based fully automated cephalometric analysis: comparisons between app-aided, computerized, and manual tracings. Turk J Orthod 2020;33:142-9.
15. Mahto RK, Kafle D, Giri A, Luintel S, Karki A. Evaluation of fully automated cephalometric measurements obtained from web-based artificial intelligence driven platform. BMC Oral Health 2022;22:132.
16. Yassir YA, Salman AR, Nabbat SA. The accuracy and reliability of WebCeph for cephalometric analysis. J Taibah Univ Med Sci 2022;17:57-66.
17. Kazimierczak W, Gawin G, Janiszewska-Olszowska J, Dyszkiewicz-Konwińska M, Nowicki P, Kazimierczak N, et al. Comparison of three commercially available, AI-driven cephalometric analysis tools in orthodontics. J Clin Med 2024;13:3733.
18. Durão APR, Morosolli A, Pittayapat P, Bolstad N, Ferreira AP, Jacobs R. Cephalometric landmark variability among orthodontists and dentomaxillofacial radiologists: a comparative study. Imaging Sci Dent 2015;45:213-20.
19. de Queiroz Tavares Borges Mesquita G, Vieira WA, Vidigal MTC, Travençolo BAN, Beaini TL, Spin-Neto R, et al. Artificial intelligence for detecting cephalometric landmarks: a systematic review and meta-analysis. J Digit Imaging 2023;36:1158-79.
20. Ongkosuwito EM, Katsaros C, van 't Hof MA, Bodegom JC, Kuijpers-Jagtman AM. The reproducibility of cephalometric measurements: a comparison of analogue and digital methods. Eur J Orthod 2002;24:655-65.
21. Nishimoto S, Sotsuka Y, Kawai K, Ishise H, Kakibuchi M. Personal computer-based cephalometric landmark detection with deep learning, using cephalograms on the Internet. J Craniofac Surg 2019;30:91-5.
22. El-Dawlatly M, Attia KH, Abdelghaffar AY, Mostafa YA, Abd El-Ghafour M. Preciseness of artificial intelligence for lateral cephalometric measurements. J Orofac Orthop 2024;85:27-33.
23. Devereux L, Moles D, Cunningham SJ, McKnight M. How important are lateral cephalometric radiographs in orthodontic treatment planning? Am J Orthod Dentofacial Orthop 2011;139:e175-81.
24. Durão AR, Alqerban A, Ferreira AP, Jacobs R. Influence of lateral cephalometric radiography in orthodontic diagnosis and treatment planning. Angle Orthod 2015;85:206-10.
25. Durão AR, Pittayapat P, Rockenbach MIB, Olszewski R, Ng S, Ferreira AP, et al. Validity of 2D lateral cephalometry in orthodontics: a systematic review. Prog Orthod 2013;14:31.