

How reliable are AI-Assisted cephalometric programs in assessing measurements involving bilateral landmarks?

Hakan Gurcan Gurel*, Pinar Alkumru[†] and Pavel Sectakof[‡]

Department of Orthodontics, Faculty of Dentistry, Biruni University, Istanbul, Turkey*

Department of Orthodontics, Faculty of Dentistry, University of Toronto, Toronto, Canada[†]

Objective: Recent advancements in technology have promoted artificial intelligence (AI) to automatically detect landmarks on lateral cephalograms. This study aimed to evaluate the reliability of an AI-assisted cephalometric analysis program (Webceph, Gyeonggi-do, Republic of Korea) in the assessment of cephalometric measurements involving bilateral landmarks.

Materials and methods: Fifty-one high-quality cephalograms were used and inclusion/exclusion criteria were applied. Two researchers manually traced the cephalograms after which an AI-assisted cephalometric analysis program (Webceph) was applied. Both intra- and inter-operator reliability were tested. Independent Sample t-tests were used to compare the means of measurements.

Results: The inter- and intra-class correlation coefficients were 0.80 which indicated 'good' reliability. Statistically significant differences were found in the gonial angle and effective mandibular length measurements ($p < 0.05$), but not in the articular angle and FMA angle ($p > 0.05$). The results suggest that, while AI-assisted programs provide reliable measurements, differences in certain measurements may be attributed to inherent AI algorithm limitations. Clinicians should verify and, if needed, correct bilateral landmark locations after the initial AI digitisation.

Conclusion: AI-driven cephalometric analysis holds promise for improving diagnostic efficiency and precision. However, limitations and further AI advancements require consideration to ensure appropriate clinical use.

[Aust Orthod J 2025; 41: 123 - 129. DOI: 10.2478/aoj-2025-0017]

Received for publication: September, 2024

Accepted: March, 2025.

Hakan Gurcan Gurel: hgurel@biruni.edu.tr; Pinar Alkumru: pinar.alkumru@utoronto.ca; Pavel Sectakof: P.Sectakof@dentistry.utoronto.ca

Introduction

Since its introduction by Broadbent and Hofrath in 1931, cephalometry has contributed to the analysis of malocclusions and has become a 'gold standard' in orthodontic diagnosis.^{1,2} Initially, after manually tracing anatomical features and identifying landmarks on acetate paper, the linear and angular parameters were measured using rulers and protractors to generate a lateral cephalometric analysis. The entire process was laborious, time-consuming, prone to mistakes, and largely dependent on operator skill.^{3,4}

To address the issues, computer programmers have developed a number of cephalometric analysis programs. However, the programs still require a clinician's manual input for localising landmarks on the digitised cephalometric image.⁵ Artificial intelligence (AI) refers to machine intelligence rather than human intelligence. AI performs many tasks like speech recognition, computer vision, language translation, and other data mappings.⁶ Deep learning, which is a subset of AI, is a method inspired by the human brain that teaches computers to process data effectively. Deep learning models

excel at recognising intricate patterns of various data types, including images, text, and sounds thereby enabling accurate insights and predictions. Recently, deep learning algorithms have gained widespread use for automatically detecting landmarks on lateral cephalograms.⁶

However, there are two primary concerns regarding the algorithms noted as the number of landmarks which can be identified and the resulting accuracy. Accuracy is typically considered achieved when the distance between the gold standard position (manual localisation) and the automatic method's detected position is less than 2 mm, while a distance less than 4 mm is considered acceptable. Additionally, when bilateral anatomical structures do not overlap exactly, the resulting point should represent an average of the left and right landmark co-ordinates. Previous studies did not clearly describe this condition, which therefore could potentially impact the accuracy of landmark position and consequently affect the validity of cephalometric analysis.^{6,7}

The aim of the present study was to therefore compare the reliability of an AI-assisted cephalometric analysis program (Webceph) in the assessment of cephalometric measurements involving bilateral landmarks.

Materials and methods

The lateral cephalometric radiographs used in the study were selected from the radiology archive of the faculty of Dentistry, Biruni University. The archive housed the radiographs of previously treated patients who had provided consent for the use of their records. Two experienced orthodontists exclusively chose the cephalometric radiographs which exhibited deviation in the position of bilateral landmarks. All selected radiographs were obtained using the Orthophos Xg orthopantomograph (Sirona, Bensheim, Germany), which incorporated a metric scale within its cephalostat. The radiographs were captured with each subject's lips in the rest position, thereby ensuring that Frankfort horizontal plane was parallel to the floor and the jaws were in centric relation. Exposure settings were 73 kV, 15 mA and 9.4 s. Ethical approval for the study was granted by the Ethical Committee of Biruni University (Approval number 2023/78-25).

Inclusion criteria

- (1) High-quality patient cephalograms without artifacts, the presence of which could hinder the accurate identification of anatomical points.
- (2) Cephalograms must be accompanied by a calibration ruler to enable the determination of the magnification.

Exclusion criteria

- (1) Cephalograms displaying excessive soft tissue that may impede the precise localisation of anatomical points.
- (2) Cephalograms of patients with a cleft lip and palate, a congenital craniofacial deformity or a marked asymmetry.
- (3) Cephalograms of patients exhibiting positional errors, as indicated by ear rod markers.
- (4) Cephalograms wherein landmarks cannot be identified due to motion, significant differences in resolution, or lack of contrast.

The comparison of two independent groups was planned in the main hypothesis of the study. Similar studies were examined, and the sample size calculation that gave the highest number according to the applied statistical methods in the main hypothesis was considered. Therefore, the sample size was calculated at a 95% confidence level using the 'G. Power-3.1.9.2' program (Heinrich Heine Universität, Düsseldorf, Germany).⁸ As a result, with an $\alpha=0.05$ and a standardised effect size of 0.915 obtained from a previous study,⁹ a minimum sample of 33 radiographs was calculated to achieve a theoretical power of 0.95. Considering the possibility of a lack of a normal distribution, the sample size was increased by 50% which generated a final sample size of 51 radiographs (30 females aged 16.93 ± 3.2 and 21 males aged 17.7 ± 2.1).

The lateral cephalometric radiographs were originally in JPEG format but were converted into high-quality hard copies for manual tracing. The images were printed on A4-sized glossy photo paper using a high-resolution inkjet printer. To maintain measurement accuracy, the digital images were adjusted to 1:1 magnification before printing and by using the embedded metric ruler in the cephalograms as a reference. Subsequently, the radiographs were



Figure 1. Landmarks used in the study; Ar (Articulare): The point at the junction of the posterior border of the ramus and the inferior border of the posterior cranial base. Co (Condylion): The most superior point of the mandibular condyle. Gn (Gnathion): The most anterior and inferior point on the mandibular symphysis. Go (Gonion): The most posterior and inferior point on the angle of the mandible. Menton (Me): The most inferior point on the mandibular symphysis. Or (Orbitale): The most inferior point on the inferior margin of the orbit. Po (Porion): The superior-most point of the external auditory meatus. S (Sella): The centre of the sella turcica. (For the bilateral landmarks Ar, Co, Go, Or and Po, midpoints of the left and right images were used). Measurements used in the study; Frankfort-mandibular plane angle (FMA): The angle between Frankfort horizontal plane (Po-Or) and mandibular plane (Go-Me). Articular angle: The angle formed by joining the points S, Ar and Go. Gonial angle: The angle formed by joining the points Ar, Go and Me. Effective mandibular length: The distance between the points Co and Gn.

manually traced using a 0.5 mm mechanical pencil and ultra-thin (0.003" thick) acetate overlay sheets, following established cephalometric tracing protocols.¹⁰ All tracings were performed independently by two experienced orthodontists to ensure reliability and reproducibility. Following manual tracing, the same radiographs were assessed using Webceph AI-assisted cephalometric analysis software (AssembleCircle Corp., Gyeonggi-do, Republic of Korea). All of the cephalometric images were calibrated according to the instructions provided by Webceph and only measurements involving bilateral landmarks were included. This selection was based on several previous studies which assessed the overall reliability of AI-assisted cephalometric tracing programs.^{11,12} The cephalometric landmarks

and the description of measurements are presented in Figure 1 and the tracing of a study subject using Webceph is shown in Figure 2.

To assess intra-operator reliability, the first experienced investigator manually traced 25 lateral cephalometric images twice to evaluate consistency. For inter-operator reliability, a second experienced investigator independently performed manual tracings of the same images. The 25 tracings from each investigator were then compared to assess agreement and a kappa coefficient of 0.8 was chosen as the threshold to indicate satisfactory calibration and reliability between the two investigators.

The descriptive statistics of the data are shown in Table I. As the first step of the statistical analysis, the normal distribution assumption was assessed using the Shapiro–Wilk test, and variance homogeneity was tested using the Levene test. When the normal distribution assumption was met, the agreement between measurements was evaluated using Pearson's moment product correlation. Independent samples *t* test was applied to compare the means of the two independent groups showing normal distribution. The analyses were carried out using IBM SPSS

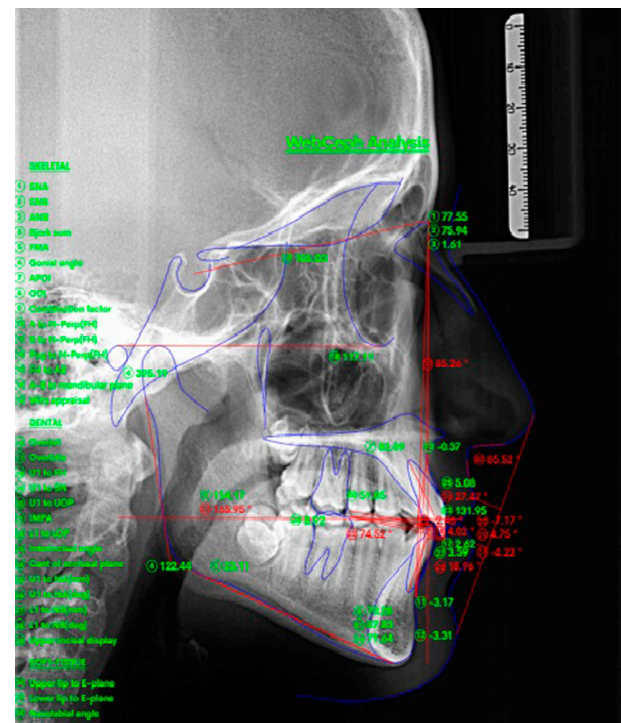


Figure 2. Cephalometric tracing of a study subject with analysis results on Webceph.

Table I. The descriptive statistics of the data used in the study

Methods		n	Minimum	Maximum	Mean	Standard deviation	Median
Webceph	Articular Angle (degree)	51	137.50	161.20	148.45	5.44	147.90
	Gonial Angle (degree)	51	108.00	133.10	122.72	5.73	123.40
	Effective Mandibular Length (mm)	51	98.50	122.60	108.95	6.36	110.00
	FMA (degree)	51	13.40	43.00	25.66	5.96	25.50
Manual Tracing	Articular Angle (degree)	51	134	163	149.39	6.84	150
	Gonial Angle (degree)	51	100	137	119.81	7.16	120
	Effective Mandibular Length (mm)	51	97	142	113.29	8.72	113
	FMA (degree)	51	12	37	25.96	5.56	26.5

Statistics for Windows, Version 25.0 (IBM Corp., Armonk, New York).

The test-retest analysis was applied to investigate whether the same variables were similar between different observers, different methods, or values measured at different times (Table II). The normality assumption was verified using the Shapiro-Wilk test to select the appropriate inter-class correlation test. The variability between measurements was evaluated using Pearson's moment product test. The correlation coefficients between different measurement methods for the articular angle and effective mandibular length parameters were below the minimum value of 0.70 required for similarity.

Therefore, it was determined that the articular angle and effective mandibular length measurements were not sufficiently similar between the methods. The correlation coefficients between different methods for gonial angle and FMA parameters were calculated above the minimum value of 0.70 required for similarity. It was determined that the gonial angle and FMA measurements were sufficiently similar and stable according to the measurement methods. When all *p*-values were examined, statistically significant relationships were obtained.

The assumptions were checked and Independent Sample *t* tests were applied to compare the means of the clinical measurements according to the methods applied in the study.

Table II. Test-retest and intraclass correlation coefficients for clinical measurements

			Manual Tracing
Webceph	Articular angle	rho	0.658
		<i>p</i>	0.000*
	Gonial angle	rho	0.753
		<i>p</i>	0.000*
	Effective Mandibular Length	rho	0.416
		<i>p</i>	0.002*
	FMA	rho	0.869
		<i>p</i>	0.000*

**p*<0.05.

Results

The inter- and intra-class correlation coefficients were 0.80, which indicated 'good' reliability.

Statistically significant differences were found in the means of the gonial angle and effective mandibular length measurements (*p*<0.05). The Webceph Gonial angle mean value was found to be higher than the manual tracing Gonial angle mean value. Manual tracing Effective Mandibular Length mean value was found to be greater than Webceph Effective Mandibular Length mean value.

No statistically significant differences were found between the mean values of the Articular angle and FMA angle (*p* > 0.05). The results are shown in Table III.

Table III. Comparison of the means of clinical measurements based on the programs used in the study

	Methods	Mean	Standard deviation	Test statistics	<i>p</i>
Articular Angle (degree)	Webceph	148.45	5.44	-0.77	0.444
	Manual Tracing	149.39	6.84		
Gonial Angle (degree)	Webceph	122.72	5.73	2.26	0.026 *
	Manual Tracing	119.81	7.16		
Effective Mandibular Length (mm)	Webceph	108.95	6.36	-2.87	0.005 *
	Manual Tracing	113.29	8.73		
FMA (degree)	Webceph	25.66	5.96	-0.261	0.794
	Manual Tracing	25.96	5.56		

**p*<0.05.

Discussion

Cephalometric analysis plays a crucial role in orthodontic diagnosis, treatment planning, and the evaluation of treatment results. With the advancement of artificial intelligence (AI) technology, several studies have previously explored the accuracy and reliability of AI-driven cephalometric analysis programs.^{11–13} In the present study, it was aimed to compare the reliability of an AI-assisted cephalometric analysis program (Webceph) in the assessment of cephalometric measurements involving bilateral landmarks. In contrast to prior reports, the present study concentrated on exclusively evaluating the reliability of the AI-assisted software using meticulously chosen lateral cephalometric radiographs which displayed deviations between pairs of bilateral landmarks. This focused approach aimed to assess the AI algorithm's capability to accurately localise and subsequently average the bilateral landmarks, thereby evaluating its performance under challenging conditions that tested the program's operational limits.

The findings of the present study align with previous research that evaluated the accuracy of AI systems in the detection of cephalometric landmarks. An earlier study compared the accuracy of the AudaxCeph® software (Audax d.o.o., Ljubljana, Slovenia) with human tracers and found that the software was reliable for diagnosis and treatment planning. However, differences in landmark identification were observed between the software and the human tracers, particularly for specific bilateral landmarks, noted as Porion and Orbitale.¹³ For most measurements, the differences, although statistically

significant, were deemed clinically non-significant. In the current study, there was no statistically significant difference in the measurements involving Porion and Orbitale produced by Webceph and manual tracing. Similar results were obtained in a prior study that used two deep neural networks for cephalometric landmark identification. The AI system demonstrated comparable accuracy to previous systems and showed clinically non-significant errors, even in challenging cases such as cleft subgroups.¹⁴

Past studies evaluating the accuracy of AI systems in cephalometric landmark detection have highlighted the potential of these systems to assist novices and improve their performance. In a study focusing on AI-based assistance for beginners, it was found that the system aided in improving landmark detection, although the improvements were considered insignificant in accurately identifying landmark position.¹⁵ Similarly, a study proposing a fully automatic deep learning method for the identification of landmarks including Porion, Orbitale and Articulare reported lower errors compared to human experts, thereby reducing human-induced variability.¹⁶

The use of AI-driven cephalometric analysis platforms, such as CephX®, CEFBOT, and WebCeph™, has shown promise in achieving efficient and timely cephalometric analyses. A previous study assessed the accuracy of WebCeph™ and examined cephalometric measurements some of which involved the bilateral landmarks of Porion, Orbitale and Gonion. The results demonstrated high reliability and accuracy for most parameters, as evidenced by high intraclass correlation coefficients. The present

findings align with previous studies that have evaluated AI-driven software, in which statistically significant relationships and reasonable accuracy have been observed compared to manual tracing.¹⁷

Despite the promising results, it is important to acknowledge the limitations and challenges associated with AI-driven cephalometric analysis. Inaccuracy can arise from factors related to non-uniform image quality, an improper image angle, and the inability of AI models to recognise non-skeletal objects and markings.¹⁸ Additionally, while AI systems have shown improved accuracy and precision, specific landmarks, such as Condylion, are prone to incorrect identification. The condyle is a three-dimensional structure, and its position can vary depending on head posture and the angle of the X-ray beam. Secondly, the condyle can be partially obscured by other anatomical structures, identified as the temporal bone, the zygomatic arch, and the soft tissues. These factors may explain the difference in the effective mandibular length measurement between the two current methods. The difference in Gonial angle measurements, however, might be explained by the possible inherent AI algorithm limitations of the AI-assisted tracing program. Future developments in AI technology should focus on addressing these limitations to further enhance cephalometric accuracy and reliability.

With advancements in deep learning methodologies, the efficiency of AI-assisted cephalometric programs can be significantly improved. AI models can be trained with large, high-quality datasets specifically designed to enhance the detection of bilateral landmarks. By leveraging deep learning algorithms, particularly convolutional neural networks (CNNs), AI systems can improve the ability to correctly identify and average bilateral landmark positions thereby reducing discrepancies caused by anatomical asymmetry. These developments suggest that future iterations of AI-assisted cephalometry will likely achieve greater precision and reliability, making them more useful for both routine clinical practice and complex cases.

Conclusion

The current study, along with the findings from previous research, has demonstrated the potential of AI-driven cephalometric analysis programs in providing accurate and reliable results. The

systems show promise in improving the efficiency and precision of cephalometric analysis, especially for less experienced clinicians. However, it is important to carefully consider the limitations and challenges associated with the systems to ensure their appropriate application in clinical practice. It is also recommended that the clinician should check and, if necessary, correct the location of bilateral landmarks after the initial digitisation by the AI-assisted cephalometric tracing program. Continued advancements in AI technology hold great potential for further enhancing cephalometric analysis and its clinical applicability.

Conflict of interest

The authors declare that there is no conflict of interest.

Corresponding author

Hakan Gurcan Gurel, DDS, PhD, MSc, Biruni University Istanbul, Istanbul, Turkey. E-mail: hgurel@biruni.edu.tr.

References

1. Broadbent BH. A new X-ray technique and its application to orthodontia. *Angle Orthodontist*. 1931; 1:45–66.
2. Hofrath H. Die bedeutung der roentgenfern der kiefer anomalien. *Fortschr Orthodontic*. 1931; 1:232–48.
3. Baumrind S, Frantz RC. The reliability of head film measurements 1 Landmark identification. *Am J Orthod*. 1971; 60:111–27.
4. Baumrind S, Frantz RC. The reliability of head film measurements 2. Conventional angular and linear measures. *Am J Orthod*. 1971; 60:505–17.
5. Wang C-W, Huang C-T, Hsieh M-C, Li C-H, Chang S-W, Li W-C, et al. Evaluation and comparison of anatomical landmark detection methods for cephalometric X-Ray Images. *IEEE Trans Med Imaging*. 2015; 34:1890–900.
6. Ding H, Wu J, Zhao W, Matinlinna JP, Burrow MF, Tsoi JKH. Artificial intelligence in dentistry—a review. *Front Dental Med*. 2023; 4:1–13.
7. Yao J, Zeng W, He T, Zhou S, Zhang Y, Guo J, et al. Automatic localization of cephalometric landmarks based on convolutional neural network. *Am J Orthod Dentofacial Orthop*. 2022; 161: 250–9.
8. Cohen J. *Statistical power analysis for the behavioral sciences*. (2nd ed.). Hillsdale, NJ: Erlbaum; 1988.
9. Duran GS, Gökmen Ş, Topsakal KG, Görgülü S. Evaluation of the accuracy of fully automatic cephalometric analysis software with artificial intelligence algorithm. *Orthod Craniofacial Res*. 2023; 26:481–90.
10. Hassan MM, Hadi Alfaifi WH, Qaysi AM, Alfaifi AA, AlGhaffi ZM, Mattoo KA, et al. Comparative evaluation of digital cephalometric tracing applications on mobile devices and manual tracing. *Med Sci Monitor*. 2024; 30:e944628.

11. Alessandri-Bonetti A, Sangalli L, Salerno M, Gallenzi P. Reliability of artificial intelligence-assisted cephalometric analysis. A pilot study. *BioMedInformatics*. 2023; 3:44–53.
12. Mercier J-P, Rossi C, Sanchez IN, Renovales ID, Sahagún PM-P, Templier L. Reliability and accuracy of Artificial intelligence-based software for cephalometric analysis. *BMC Oral Health*. 2024; 24:1309.
13. Ristau B, Coreil M, Chapple A, Armbruster P, Ballard R. Comparison of AudaxCeph*'s fully automated cephalometric tracing technology to a semi-automated approach by human examiners. *Int Orthod*. 2022; 20:100691.
14. Tanikawa C, Lee C, Lim J, Oka A, Yamashiro T. Clinical applicability of automated cephalometric landmark identification: Part I-Patient-related identification errors. *Orthod Craniofac Res*. 2021; 24(Suppl 2):43–52.
15. Le VNT, Kang J, Oh I-S, Kim J-G, Yang Y-M, Lee D-W. Effectiveness of human-artificial intelligence collaboration in cephalometric landmark detection. *J Personalized Med*. 2022; 12:387.
16. Kim YH, Lee C, Ha E-G, Choi YJ, Han S-S. A fully deep learning model for the automatic identification of cephalometric landmarks. *Imaging Sci Dentistry*. 2021; 51:299–306.
17. Mahto RK, Kaffle 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.
18. Bulatova G, Kusnoto B, Grace V, Tsay TP, Avenetti DM, Sanchez FJC. Assessment of automatic cephalometric landmark identification using artificial intelligence. *Orthod Craniofac Res*. 2021; 24(Suppl 2):37–42.