Original Article

Factors influencing the predictive performance of artificial intelligence for craniofacial growth

Naeun Kwon[a]; Jong-Hak Kim[a]; Heeyeon Suh[b]; Heesoo Oh[c]; Shin-Jae Lee[d]

ABSTRACT

Objectives: To evaluate factors influencing the prediction error of artificial intelligence (AI) that predict craniofacial growth and to identify an optimal AI training condition to improve the predictive performance of the AI model.

Materials and Methods: Original growth data were collected from the Mathews longitudinal serial growth study. From the original data consisting of 1257 datasets from 33 growing children of northern European descent, 60 data subsets were generated using random resampling procedures to include 12, 18, and 24 subjects, with data sizes of 100, 200, 300, 400, and 500 datasets. The resampling procedures were repeated four times. Each subset was used to train and create a total of 60 AI models. The prediction accuracy of these models was evaluated using growth prediction errors at the lower lip landmark, labrale inferius, as a benchmark indicator. The prediction errors of the 60 AI models were analyzed according to the number of subjects and data sizes.

Results: Prediction error decreased as the data size increased. However, increasing the number of subjects within the growth data led to higher prediction errors. Notably, the increase in prediction error caused by adding more subjects was more substantial than the improvement achieved by increasing the data size.

Conclusions: The findings suggest that developing highly accurate AI-based craniofacial growth prediction models remains a significant challenge, even with extensive datasets. (*Angle Orthod*. 2025;00:000–000.)

KEY WORDS: Artificial intelligence; Craniofacial growth; Prediction error; Data quantity; Individual variability; Generalizability

INTRODUCTION

Recent advancements in artificial intelligence (AI) have led to superior accuracy in growth prediction methods compared to traditional approaches.[1-4] At present, however, the application of AI in orthodontics is still in its developmental stage and has room for improvement. In the past, one of the most challenging aspects in growth evaluation and prediction studies was collecting a sufficient amount of longitudinal serial growth data from children due to ethical concerns, including radiation exposure risk. However, with the American Association of Orthodontists Foundation (AAOF) Craniofacial Growth Legacy Collection, comprising a total sample size of 762 subjects and more than 20,000 digital images, data availability is no longer a major limitation.[5,6] Nevertheless, the optimal data training design to develop an AI-based growth prediction model remains unclear.[7]

In growth research design, the sample size or data size refers to the number of longitudinal serial datasets that contain paired data on growth before and after. Various factors such as data size, the number of unique subjects from whom the longitudinal serial datasets were collected, and population ethnicity differences may all influence the accuracy of the resultant growth prediction model. Yet, the exact effects of those factors are still

[a] Graduate Student (Ph.D.), Department of Orthodontics, Seoul National University, Seoul, Korea.
[b] Assistant Professor, Department of Orthodontics, Arthur A. Dugoni School of Dentistry, University of the Pacific, San Francisco, CA, USA.
[c] Professor and Chair, Department of Orthodontics, Arthur A. Dugoni School of Dentistry, University of the Pacific, San Francisco, CA, USA.
[d] Professor, Department of Orthodontics; and Dental Research Institute, Seoul National University School of Dentistry, Seoul, Korea.
Corresponding author: Dr Shin-Jae Lee, Professor, Department of Orthodontics and Dental Research Institute, Seoul National University School of Dentistry, 101 Daehakro, Jongro-Gu, Seoul 03080, Korea
(e-mail: nonext@snu.ac.kr)

Accepted: August 31, 2025. Submitted: March 10, 2025.
Published Online: September 29, 2025

debated since previous studies have reported conflicting findings.[2,7,8] In general, it has been recognized that the performance of AI could improve as the size of the training data increased,[9] a notion that was consistent with findings from other studies. For instance, research on sample size determination for AI suggested that more than 1600 datasets might be required to develop AI systems that could ensure clinically acceptable accuracy in predicting and visualizing treatment outcomes for orthodontic treatment and orthognathic surgical changes.[8,10,11]

However, in the craniofacial growth prediction problem, the situation was different. While one study suggested that larger sample sizes, greater than 1700 datasets, might achieve craniofacial growth prediction accuracy within a clinically acceptable level,[8] another study argued that increasing data size also introduces individual variability, thereby reducing accuracy.[7] Additionally, a recent article reported that AI models could predict the craniofacial growth of American children more accurately than Korean children, but the reasons for this remained unclear.[2] Possible explanations for the ethnicity difference in growth prediction accuracy might include the following: (1) a genuine ethnicity difference in the prediction accuracy between the two distinct populations, (2) differences in the number of unique subjects between the two databases, 33 American children in the Mathews growth collection and 410 Korean children, (3) disparity in the number of longitudinal serial records between the two groups, 1257 and 679 datasets, for American and Korean, respectively, or (4) a combination or interaction of these reasons. With the AAOF Craniofacial Growth Legacy Collection providing extensive longitudinal growth data, researchers now have the opportunity to refine AI models by optimizing data input strategies.

The purpose of the present study was to analyze the factors influencing craniofacial growth prediction error systematically.

MATERIALS AND METHODS

The University of the Pacific Human Subjects Protection Office of Research and Sponsored Programs approved the research protocol, and the project received an exempt review (University of the Pacific IRB 2023-28).

Growth Data and Random Resampling Subsets

The longitudinal serial growth data were collected from the University of the Pacific Mathews Growth Study posted on the AAOF Craniofacial Growth Legacy Collection website, https://www.aaoflegacycollection.org/aaof_collection.html?id = UOPMathews, which is the only longitudinal serial cephalometric dataset from subjects with Björk-type implants available.[5] All 33 subjects (21 girls and 12 boys) included in the Mathews Growth Study were of northern European origin, for whom metal implants in the maxilla and mandible were placed by an open surgical method between the ages of 4 and 7 years.[5] The subjects had lateral cephalograms taken annually from two to 14 times, summing to a total of 1257 serial growth datasets.

In this data resampling study, the matrix-formed 1257 datasets were provided by Roseth et al. (2025), who compared the accuracy between AI and conventional craniofacial growth prediction methods.[1] The data matrix contained 159 columns of input, and 156 columns of output variables. The 159 input variables were designed to encompass nearly all skeletal and soft-tissue characteristics, including vertical and anteroposterior cephalometric patterns of an individual subject, along with demographic information such as age, sex, and growth observation intervals. The 156 output variables represented growth changes at 78 cephalometric landmarks, described in terms of Cartesian coordinate information.[1]

Random resampling procedures were utilized to generate subsets of longitudinal growth datasets from three different subject groups: 12 subjects (six girls, six boys), 18 subjects (nine girls, nine boys), and 24 subjects (12 girls, 12 boys), from whom the longitudinal serial growth data were collected. This process was conducted to generate five different dataset sizes: 100, 200, 300, 400, and 500. Each random resampling procedure was repeated four times, resulting in a total of 60 resampled-data subsets, which were used to create 60 different AI models (Figure 1).

Craniofacial Growth Prediction AI Model Building

To develop AI models for craniofacial growth prediction, the TabNet Deep Neural Network algorithm (Arik and Pfister, 2021, Stanford, California, USA),[12] was applied to train the 60 data subsets. This algorithm was selected because of its capability to handle multiple input and output variables required to reflect the craniofacial growth phenomenon that would be multifactorial in nature. Utilizing Python programming (Python Software Foundation, Wilmington, Delaware, USA), TabNet was tailored to include almost all cephalometric analysis variables as input variables to reflect an individual's soft-tissue characteristics, vertical and anteroposterior skeletal patterns, as well as age, sex, and growth observation period.

AI models incorporated 159 input variables, many of which were specifically designed to represent complex skeletal morphology quantitatively. These included linear and angular descriptors such as the gonial angle, antegonial curvature, and condylar-ramal height, allowing the model to encode structural patterns that are traditionally assessed subjectively by clinicians.

Although the generated samples spanned a broad range of ages, this heterogeneity did not compromise

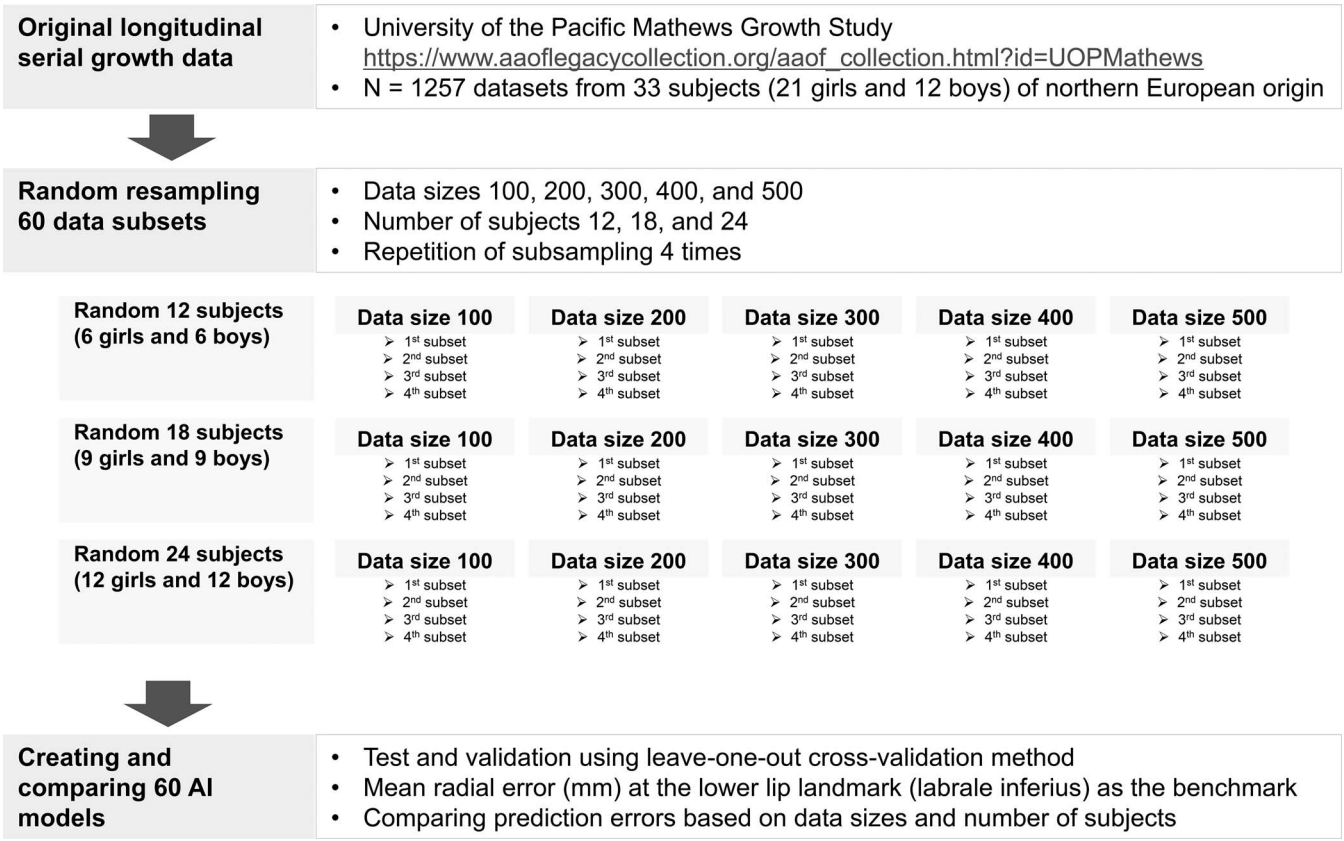

Figure 1. Experimental design summary.

model performance by incorporating chronological age and the interval of growth observation as input variables. This enabled the AI model to contextualize and predict craniofacial growth patterns relative to age, analogous to how a clinician considers both age and observation interval during longitudinal assessment.

The point Sella was set as the coordinate origin (0.0, 0.0). To establish orientation, the x-axis was defined as a horizontal reference plane constructed by rotating the Sella–Nasion (SN) line downward by 7 degrees. The y-axis was constructed as a perpendicular line to the x-axis at Sella. The outcome variables were predicted changes in 46 skeletal and 32 soft-tissue landmarks after a certain growth observation period.

All 60 AI models were trained using Linux desktop computers. To calculate test or validation error, one data point was omitted when building each AI model. For example, when developing an AI model using a subset of 100-sized data, a total of 100 AI models had to be created, each sequentially omitting a single data point. Each AI model was then used to calculate the test or validation error for that specific omitted data point. This type of test or validation method has often been called the leave-one-out cross-validation (LOOCV). Although LOOCV requires much more computation resources than other test/validation methods, LOOCV was frequently selected in clinical studies due to its unique capability for detecting individual variability, random aberrations, and outliers.[13,14]

Statistical Analysis

To evaluate prediction accuracy, the Euclidean distance was measured between the actual growth and the predicted change at the most anterior point of the lower lip landmark (labrale inferius). This distance, often referred to as the mean radial error, was expressed in millimeters.

Multiple linear regression analysis was conducted to evaluate the influence of the number of subjects and data sizes on the growth prediction error. To determine which factor had a greater impact on the craniofacial growth prediction errors, a conditional inference tree structure was analyzed using the R package *partykit* (R Development Core Team, Vienna, Austria). The statistical significance level was set at $P < 0.05$.

RESULTS

The fastest AI model training time was 150 minutes for a subset of 100 data points from 12 children, whereas the longest was 37 hours and 30 minutes for a subset of 500 from 24 children.

Table 1. Growth Prediction Error Measured on the Lower Lip According to the Data Size (100, 200, 300, 400, and 500) and Number of Subjects (12, 18, and 24). Mean Radial Errors in Millimeters Were Calculated Between the Actual Growth and Predicted Growth Changes

	Mean ± Standard Deviation (mm)					
Data size	100	200	300	400	500	*P* Value
Number of subjects						
12 (six girls, six boys)	1.63 ± 1.20	1.36 ± 0.86	1.38 ± 0.85	1.27 ± 0.83	1.26 ± 0.84	<.0001
18 (nine girls, nine boys)	1.83 ± 1.20	1.75 ± 1.09	1.58 ± 0.98	1.46 ± 0.89	1.50 ± 0.91	<.0001
24 (12 girls, 12 boys)	2.10 ± 1.43	1.85 ± 1.17	1.66 ± 1.07	1.66 ± 1.07	1.63 ± 0.99	<.0001

The prediction error decreased as the data size increased. In contrast, increasing the number of subjects included in the AI training data led to a rise in the prediction error (Table 1). When error scatter plots were depicted, the AI prediction model trained on 100 datasets from 24 children exhibited the greatest error range and variability in elliptical shapes (Figure 2).

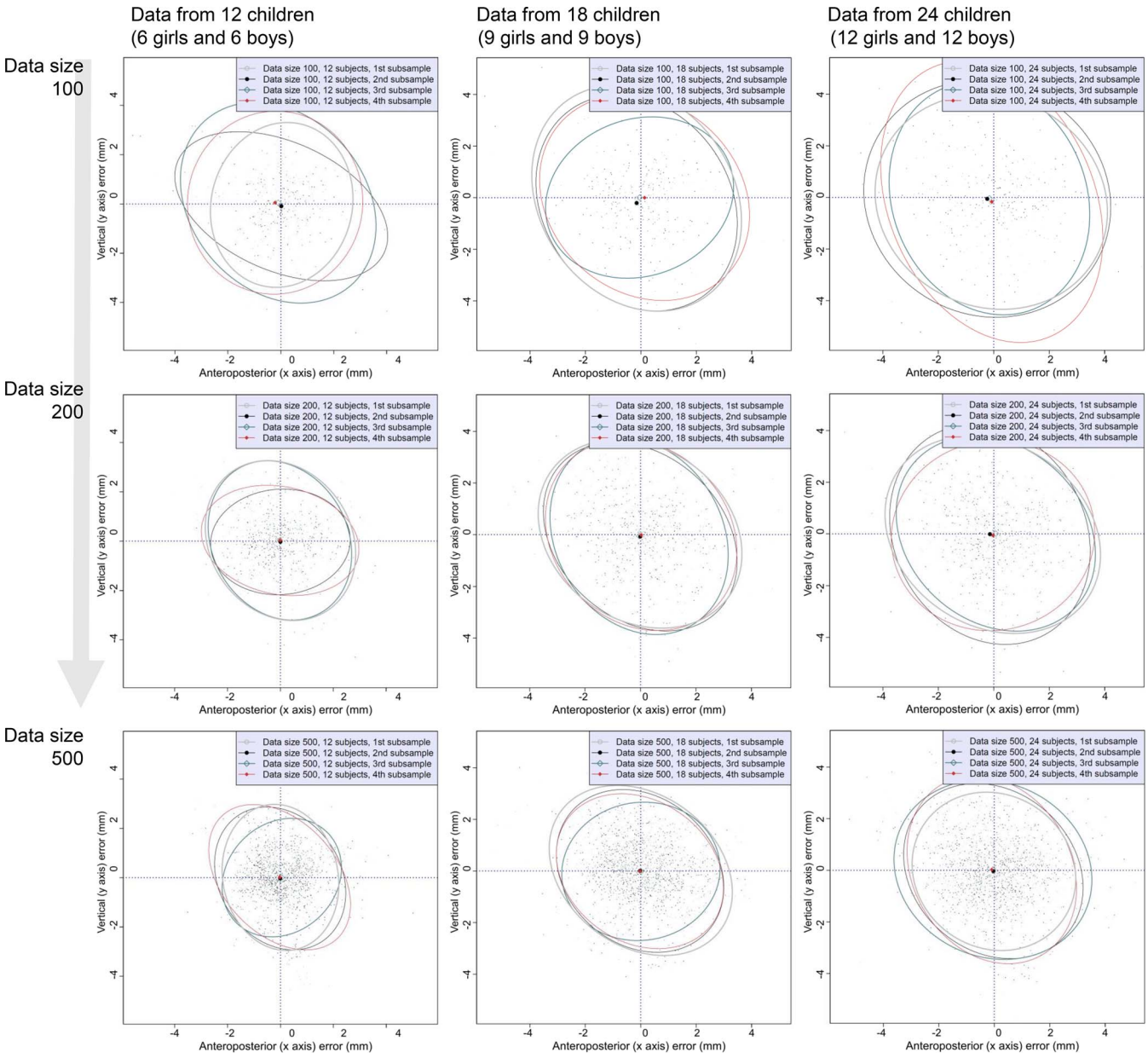

Figure 2. Scatterplots with 95% confidence boundaries of prediction errors according to the data sizes and number of subjects from whom the longitudinal serial growth data were collected.

Table 2. Multiple Linear Regression Analysis of Factors Influencing Growth Prediction Error

	β (mm)	SE (β)	*P* Value
(intercept)	1.2450	0.0352	<.0001
Number of subjects	0.0317	0.0015	<.0001
Data size (unit 100)	−0.0778	0.0059	<.0001

β indicates regression coefficient estimates; SE, standard error.

Results of the multiple linear regression analysis indicated that the number of subjects and data size had a statistically significant impact on the prediction errors ($P < .0001$), whereas the interaction term between them was not statistically significant. From the regression coefficients, an additional subject resulted in an increase of 0.03 mm in the prediction errors, while including every additional 100 data points led to a decrease of 0.08 mm in the prediction errors (Table 2).

When the growth prediction errors were illustrated based on data sizes and the number of subjects, the error decreased with larger data sizes (Figure 3, *right*) but increased with more subjects (Figure 3, *left*).

Since these two factors significantly influenced the prediction error in opposing ways, a conditional inference tree structure was devised to determine whether the error would converge or diverge in the end. Figure 4 illustrates a conditional inference tree that identifies which factor is more prominent than the others. The number of subjects from whom longitudinal serial growth data were collected emerged as the most significant factor, substantially contributing to the prediction error in craniofacial growth.

DISCUSSION

This study was inspired by recent advancements in AI-based data simulation studies and the abundant growth data available on the AAOF Craniofacial Growth Legacy Collection website.[5] Initially, it was expected that accuracy could be further improved if sufficient training data were utilized, and that an optimal data input strategy could be implemented by analyzing the factors influencing the accuracy of the growth prediction model. However, contrary to these initial expectations, results of the present study implied that developing an accurate growth prediction model might be fundamentally unachievable, even with an extensive amount of growth data.

Although increasing dataset size led to a reduction in prediction errors to some extent, the declining pattern did not seem drastic, and the error values did not fall below a certain threshold either. Additionally, as is inherent in longitudinal data collection, increasing dataset size inevitably involved including more subjects from whom the longitudinal growth data were collected. In other words, the number of subjects is inherently associated with the amount of longitudinal serial data. When developing AI models to predict craniofacial growth, incorporating as much growth data as possible from a considerable number of subjects may enhance the generalizability of the prediction method.[1-5,7,8] However, this also introduces greater variability in the prediction results. Consequently, the errors in predicting craniofacial growth are unlikely to converge below a certain threshold but may, instead, continue to increase as more subjects are included.

All 60 growth prediction models predicted growth changes in 78 landmarks. However, only the lower lip landmark was selected as the benchmark indicator of the craniofacial growth prediction accuracy. The lower lip was selected because it is known to be one of the most variable landmarks. Its position is highly sensitive to transient changes like lip posture and facial muscle tension, often resulting in substantial individual

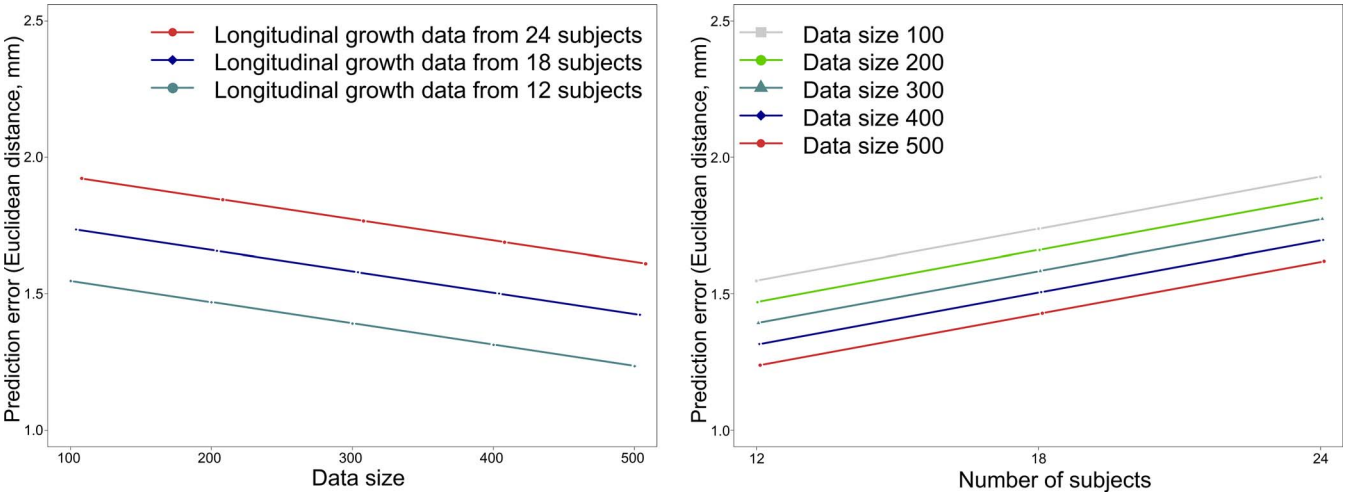

Figure 3. Estimated growth prediction errors based on the data sizes (*right*) and the number of subjects (*left*).

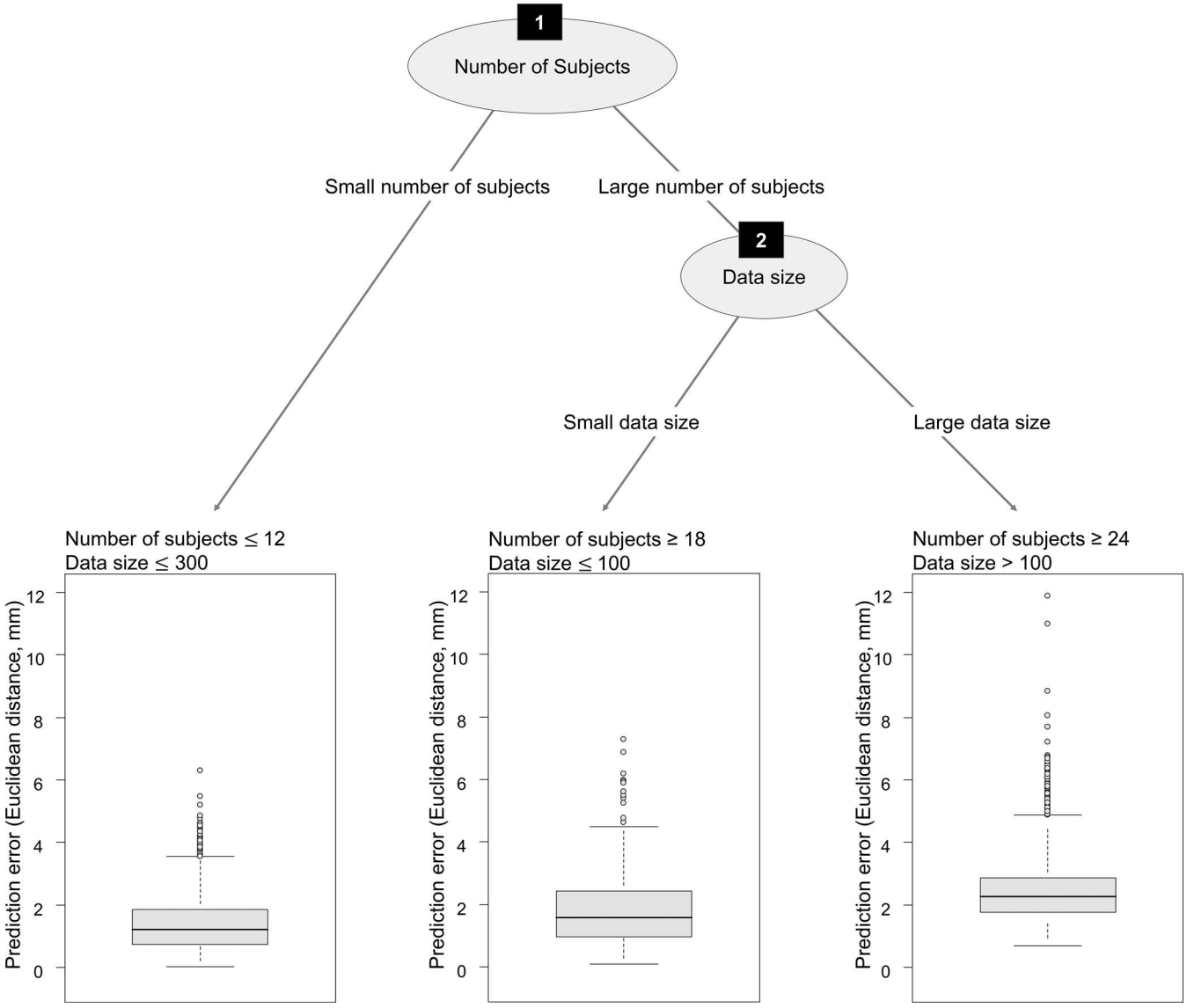

Figure 4. Conditional inference tree. The most significant factor contributing to the craniofacial growth prediction error was the number of subjects from whom longitudinal serial growth data were collected.

variation. In this respect, the prediction accuracy of the lower lip was indicative of more favorable results than in other regions.[10,15–17] Consequently, the lower lip has often been regarded as a representative landmark for evaluating the performance of AI models in orthodontics.[7,15,18,19]

To graphically visualize the form and shape of the growth prediction errors from the 60 AI models, scatterplots were depicted for selected data subsets to present the results concisely, as shown in Figure 2. Since the craniofacial growth prediction error using the cephalometric image was two-dimensional in nature, the degree and pattern of errors were expressed as scatterplots with 95% confidence boundary ellipses. While a confidence interval is a one-dimensional measure, a confidence ellipse is a two-dimensional extension based on a chi-square distribution with 2 degrees of freedom.[14]

The results did not demonstrate any significant improvement in prediction accuracy with an increased number of subjects. This contrasts with other AI applications, such as automatic landmark identification and predicting orthodontic and/or surgical treatment outcomes, where larger dataset sizes generally led to improved accuracy.[8–11] Presumably, craniofacial growth prediction appears to involve a different mechanism. Inter-individual variability in craniofacial development is substantially greater than the intra-individual variability observed across time within a single subject. As a result, adding more data from the same individual allows the model to learn smooth and structured growth trajectories. The model improves its ability to interpolate

between known time points. Conversely, increasing the number of subjects introduces greater heterogeneity in growth patterns. Differences in direction, magnitude, and timing may degrade the predictive precision of the model. This observation was supported by the conditional inference tree in Figure 4, in which the number of subjects contributed more significantly to residual deviance than data size. These findings suggest that biological diversity across individuals poses a fundamental challenge to the development of universally accurate AI models for craniofacial growth prediction.

Despite these limitations, it is important to consider the clinical implications of the reported prediction errors. A mean error of approximately 2 mm is comparable to the interexaminer variability observed in manual cephalometric tracing, which typically ranges between 1.5–2 mm.[7] Also, the purpose of AI is not to replace clinical judgment of humans, but to assist it. This is particularly relevant in the treatment planning of growing patients, where timely decisions often need to be made with incomplete information. In such scenarios, an AI-generated growth prediction may serve as a quick preliminary reference for clinicians to evaluate and refine individualized treatment strategies. In addition, clinical decisions in growth-modification treatment are generally based on jaw-level relationships rather than on precise single-tooth positioning. Thus, the observed prediction error is likely within a clinically acceptable range for decisions such as the timing for orthopedic intervention.

In summary, the development of a craniofacial growth prediction AI model exhibited distinctive characteristics compared to other AI applications. The conflicting effects of the training dataset size and subject variability imply that achieving highly accurate growth predictions may remain an unattainable goal, despite efforts to increase dataset size or subject variability. It was conjectured that this outcome, though significant, might not be entirely unexpected by readers.

CONCLUSIONS

- When developing AI models for predicting craniofacial growth, incorporating growth data from a considerable number of subjects might enhance the generalizability of the prediction method. However, this also led to greater prediction errors, which suggested that developing highly accurate AI-based craniofacial growth prediction models might remain a significant challenge, even with extensive growth datasets.

ACKNOWLEDGMENTS

This study was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant no. HI22C1518), and grant No. 02-2025-0507 from the SNUDH Research Fund.

DISCLOSURE

All authors of this study declare that they have no conflict of interest.

REFERENCES

1. Roseth J, Kim JH, Moon JH, et al. Comparison of individualized facial growth prediction models using artificial intelligence and partial least squares based on the Mathews growth collection. *Angle Orthod.* 2025;95:249–258.
2. Kim JH, Moon JH, Roseth J, Suh H, Oh H, Lee SJ. Craniofacial growth prediction models based on cephalometric landmarks in Korean and American children. *Angle Orthod.* 2025;95:219–226.
3. Moon JH, Shin HK, Lee JM, et al. Comparison of individualized facial growth prediction models based on the partial least squares and artificial intelligence. *Angle Orthod.* 2024;94:207–215.
4. Moon JH, Kim MG, Hwang HW, Cho SJ, Donatelli RE, Lee SJ. Evaluation of an individualized facial growth prediction model based on the multivariate partial least squares method. *Angle Orthod.* 2022;92:705–713.
5. AAOF craniofacial growth legacy collection. 2025. https://www.aaoflegacycollection.org/aaof_home.html. Accessed March, 2025.
6. Ferrillo M, Pandis N, Fleming PS. The effect of vertical skeletal proportions on overbite changes in untreated adolescents: a longitudinal evaluation. *Angle Orthod.* 2024;94:25–30.
7. Lee JM, Moon JH, Park JA, Kim JH, Lee SJ. Factors influencing the development of artificial intelligence in orthodontics. *Orthod Craniofac Res.* 2024;27 Suppl 2:6–12.
8. Kim JH, Kwon N, Pandis N, Lee SJ. Sample size calculation for an artificial intelligence study. *Am J Orthod Dentofacial Orthop*. 2025;167:616–620.
9. Moon JH, Hwang HW, Yu Y, Kim MG, Donatelli RE, Lee SJ. How much deep learning is enough for automatic identification to be reliable? A cephalometric example. *Angle Orthod.* 2020;90:823–830.
10. Kim JH, Kwon N, Lee SJ. Sample size estimation for developing artificial intelligence to predict orthodontic treatment outcomes. *J Korean Dent Sci.* 2025;18:12–19.
11. Kim JH, Kwon N, Park JA, Youn SB, Seo BM, Lee SJ. What amount of data is required to develop artificial intelligence that can accurately predict soft-tissue changes after orthognathic surgery? *Angle Orthod.* 2025;95:467–472.
12. Arik SÖ, Pfister T. TabNet: attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021;35:6679–6687.
13. Donatelli RE, Lee SJ. How to test validity in orthodontic research: a mixed dentition analysis example. *Am J Orthod Dentofacial Orthop.* 2015;147:272–279.
14. Moon JH, Lee JM, Park JA, Suh H, Lee SJ. Reliability statistics every orthodontist should know. *Semin Orthod.* 2024;30:45–49.
15. Han SH, Park YS. Growth patterns and overbite depth indicators of long and short faces in Korean adolescents:

revisited through mixed-effects analysis. *Orthod Craniofac Res*. 2019;22:38–45.
16. Yu W, Lee SJ, Cho H. Partial least squares regression trees for multivariate response data with multicollinear predictors. *IEEE Access.* 2024;12:36636–36644.
17. Kim K, Lee SJ, Eo SH, Cho SJ, Lee JW. Modified partial least squares method implementing mixed-effect model. *Commun Stat Appl Methods.* 2023;30:65–73.
18. Park JA, Moon JH, Lee JM, et al. Does artificial intelligence predict orthognathic surgical outcomes better than conventional linear regression methods? *Angle Orthod.* 2024;94:549–556.
19. Cho SJ, Moon JH, Ko DY, et al. Orthodontic treatment outcome predictive performance differences between artificial intelligence and conventional methods. *Angle Orthod.* 2024;94:557–565.