Original Article

What amount of data is required to develop artificial intelligence that can accurately predict soft tissue changes after orthognathic surgery?

Jong-Hak Kim[a]; Naeun Kwon[a]; Ji-Ae Park[b]; Sung Bin Youn[c]; Byoung-Moo Seo[d]; Shin-Jae Lee[e]

ABSTRACT

Objectives: To suggest a sample size calculation method to develop artificial intelligence (AI) that can predict soft tissue changes after orthognathic surgery with clinically acceptable accuracy.

Materials and Methods: From data collected from 705 patients who had undergone combined surgical-orthodontic treatment, 10 subsets of the data were generated through random resampling procedures, specifically with reduced data sizes of 75, 100, 150, 200, 300, 400, 450, 500, 600, and 700. Resampling was repeated four times, and each subset was used to create a total of 40 AI models using a deep-learning algorithm. The prediction results for soft tissue change after orthognathic surgery were compared across all 40 AI models based on their sample sizes. Clinically acceptable accuracy was set as a 1.5-mm prediction error. The predictive performance of AI models was evaluated on the lower lip, which was selected as a primary outcome variable and a benchmark landmark. Linear regression analysis was conducted to estimate the relationship between sample size and prediction error.

Results: The prediction error decreased with increasing sample size. A sample size greater than 1700 datasets was estimated as being required for the development of an AI model with a prediction error $<$ 1.5 mm at the lower lip area.

Conclusions: A fairly large quantity of orthognathic surgery data seemed to be necessary to develop software programs for visualizing surgical treatment objectives with clinically acceptable accuracy. (*Angle Orthod*. 2025;95:467–473.)

KEY WORDS: Artificial intelligence; Sample size estimation; Surgical treatment objective; Orthognathic surgery

INTRODUCTION

When discussing the accuracy of orthognathic surgery, there are generally two categories in accuracy. The first involves comparing the planned osteotomy with the actual outcomes after surgery. For instance, the use of virtual surgical planning, along with a three-dimensional printed surgical guide, has significantly reduced discrepancies between the planned and actual results in orthognathic surgical procedures.[1–3] The second category addresses inconsistencies in surgical skeletal repositioning and the corresponding soft tissue response.[4–8] Although orthognathic surgeons primarily focused more on the first issue, orthodontic clinicians were more concerned with the second issue.[9,10] This might be due to the importance of establishing surgical treatment objectives (STO) right from the initial diagnostic and treatment planning stages, particularly for patients with skeletal malocclusion. Through combined surgical-orthodontic treatment, the facial soft tissue changes are more conspicuous than changes from orthodontic treatment alone. In this respect, providing treatment options to help patients choose an appropriate treatment plan has become essential in clinical orthodontic practice.

Today, the traditional use of STO and illustrations drawn on transparent sheets has been replaced by computer programs, as anticipated decades ago.[11] Automated cephalometric landmark detection, analysis,

The first two authors contributed equally to this work.

[a] Graduate Student (Ph.D), Department of Orthodontics, Seoul National University, Seoul, Korea.

[b] Clinical Lecturer, Department of Orthodontics, Seoul National University Dental Hospital, Seoul, Korea.

[c] Assistant Professor, Center for Orthognathic and Facial Contouring Surgery, Seoul National University Dental Hospital, Seoul, Korea.

[d] Professor, Department of Oral and Maxillofacial Surgery, Seoul National University School of Dentistry, Seoul, Korea.

[e] Professor, Department of Orthodontics and Dental Research Institute, Seoul National University School of Dentistry, Seoul, Korea.

Corresponding author: Shin-Jae Lee, 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: May 25, 2025. Submitted: January 1, 2025.
Published Online: June 18, 2025

Table 1. Summary of Surgery Prediction Errors for the Lower Lip Reported in Previous Publications

Research Group	Y	Subjects	Surgical Procedure	Prediction Error	Error Measurement	Prediction Method
Park et al.[4]	2024	705	Mixed	2.1 mm	Mean radial error[a]	Deep-learning
				2.2 mm	Mean radial error	Partial least squares
				1.9 mm	Mean absolute error[b]	Deep-learning
				2.1 mm	Mean absolute error	Partial least squares
				2.1 mm	Mean absolute error	Multiple linear regression
Suh et al.[5]	2019	318	Mixed	1.7 mm	Mean absolute error	Partial least squares
				1.8 mm	Mean absolute error	Sparse partial least squares
Lee et al.[6]	2014	204	Class III, 1-jaw	2.1 mm	Mean absolute error	Partial least squares
			Class III, 2-jaw	2.0 mm	Mean absolute error	Partial least squares
Lee et al.[7]	2014	80	Class II, 2-jaw	15.7 mm	Mean absolute error	Multiple linear regression
				3.9 mm	Mean absolute error	Partial least squares
Suh et al.[8]	2012	69	Class III, 1-jaw	4.1 mm	Mean absolute error	Partial least squares
				9.4 mm	Mean absolute error	Multiple linear regression

[a] Mean radial error (mean Euclidian distance error) = mean $\{\sqrt{[(\text{anteroposterior error})^2 + (\text{vertical error})^2]}\}$.
[b] Mean absolute error = $\sqrt{\{[\text{mean absolute (anteroposterior error)}]^2 + [\text{mean absolute (vertical error)}]^2\}}$.

and treatment planning have already become an integral part of the initial stage of orthodontic treatment.[12–15] In addition, advances in predicting and visualizing treatment changes have become relatively accurate.[4,16] However, the accuracy of STO has yet to be improved. For example, although surgery prediction errors have decreased over the past decade, the prediction error still remains slightly over 2 mm (Table 1).[4] Considering that a 1.5-mm error has conventionally been recognized as an overall landmark identification error in cephalometrics,[15] and 1.5 mm is known to be the interexaminer difference among human examiners,[14,17,18] if the errors in predicting surgical changes could be reduced to 1.5 mm, it would be helpful to develop a more practical STO software product.

To increase prediction accuracy, the present study focused primarily on the number of data samples since the size of the data sample has been known to be a crucial factor in developing artificial intelligence (AI).[18,19] However, unlike conventional statistical models, no sample size guidelines have been established for developing AI models that became popular in orthodontics. Since there is no clear answer as to how much data are necessary for developing an effective AI model, an empirical approach based on a simulation study seemed to be a reasonable method for estimating the optimal data size.[18–20]

The aim of this study was to estimate the sample size required for developing an AI model that could predict soft tissue changes after orthognathic surgery with clinically acceptable accuracy.

MATERIALS AND METHODS

The institutional review board of the Seoul National University School of Dentistry approved the research protocol (S-D20240021).

Problem Formulation

As the first step toward estimating the sample size, a primary outcome variable on which the sample size estimation should be based was selected.[20] The primary outcome variable was defined as the radial error of the prediction result. The radial error is equivalent to the Euclidian distance measure between the predicted and real soft tissue changes after orthognathic surgery. The clinically acceptable prediction accuracy was considered to be less than a 1.5-mm prediction error, as suggested by previous publications.[13,14,18,20]

Random Resampling Subsets

The original data was provided by Park et al. (2024),[4] who evaluated performance of an AI model in predicting orthognathic surgical outcomes compared to conventional prediction methods. The data included preoperative and post-treatment lateral cephalograms from 705 patients who had undergone combined surgical-orthodontic treatment. Among the patients, 23% had Class II malocclusion, whereas 72% had Class III malocclusion. In cases involving maxillary surgery, 83% underwent Le Fort I osteotomy, whereas only 1% had Le Fort II osteotomy. For mandibular surgery, 86% received bilateral sagittal split ramus osteotomy, and 9% underwent intraoral vertical ramus osteotomy. In addition, genioplasty was performed on 60% of the patients. The predictors included 254 input variables and the outcome variables were posttreatment changes in 32 soft tissue landmarks from the forehead (glabella) to the terminal point on the neck.[4]

From the original data, 10 subsamples were generated through random resampling procedures, specifically with reduced data sizes of 75, 100, 150, 200, 300, 400, 450, 500, 600, and 700. The resampling procedures were repeated four times, and each subset was used to create a total of 40 AI models (Figure 1).

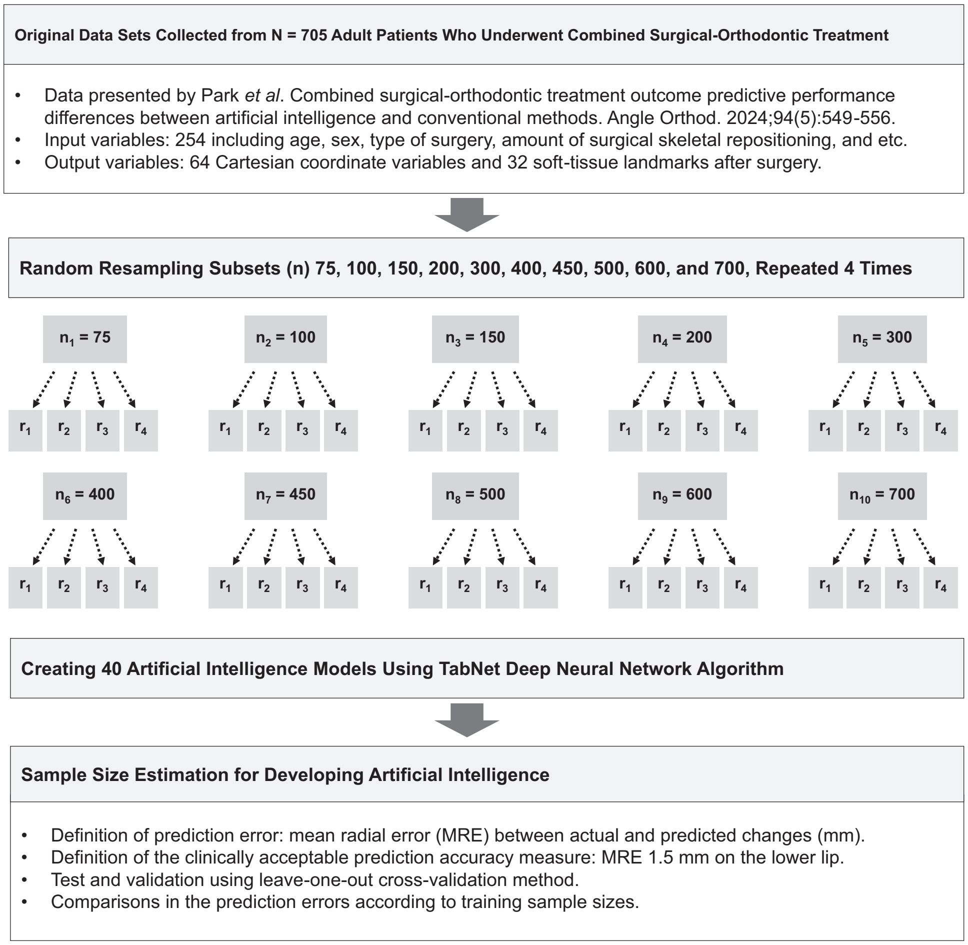

Figure 1. Experimental design summary.

To develop AI models, TabNet Deep Neural Network (Arik and Pfister, 2021), a type of convolutional neural network, was applied to the 40 data subsets. Among various deep-learning algorithms, convolutional neural networks are currently the most popular architecture for image analysis. This algorithm was selected because TabNet is applicable to table-shaped data that include numerous input and output variables relevant to surgical outcome prediction scenarios.[21] The 40 AI models were trained using ordinary desktop computers operated on a Linux environment.

Error Evaluation and Estimation Procedures for Optimal Sample Size

The prediction result for an individual subject was tested and validated utilizing the leave-one-out cross-validation method, as this validation method has been known to be particularly useful in clinical studies.[22]

The prediction results for soft tissue change after orthognathic surgery were compared across all 40 AI models. Among 32 soft tissue landmarks from the forehead to the neck,[4] the performance of AI models was evaluated specifically on the lower lip landmark (*labrale inferius*), the most anterior point of the lower lip. The lower lip landmark was selected as the benchmark because the lower lip is highly variable and, therefore, has often been used as a primary outcome variable and a benchmark in many other studies.[4,9,16,17,19,23]

The prediction error patterns resulting from the 40 AI models were evaluated using scatterplots with 95% confidence ellipses that could visualize error patterns, including the bias, variance, and reliability of the error in each model.[24]

The prediction errors resulting from the 40 AI models were analyzed using linear regression analysis. The regression line was depicted on a graph to examine the relationship between the error and sample sizes, which was used to estimate the optimal sample size. Statistical analyses were conducted using Language R (R Foundation for Statistical Computing, Vienna, Austria).[25]

RESULTS

The results of the analysis of variance on the 40 AI models did not demonstrate a statistically significant

Table 2. Results of the Analysis of Variance Among Resampling Subsets

Resampling Sample Size (n)	Lower Lip Prediction Error (Mean Radial Error ± Standard Deviation, mm)				*P* Values
	Repetition 1	Repetition 2	Repetition 3	Repetition 4	
75	3.0 ± 3.2	2.8 ± 2.1	3.0 ± 3.9	3.1 ± 2.1	.706
100	2.9 ± 2.7	2.8 ± 2.5	2.8 ± 2.1	2.8 ± 2.5	.729
150	2.4 ± 1.5	2.7 ± 2.9	2.5 ± 1.6	2.7 ± 1.9	.450
200	2.5 ± 1.8	2.5 ± 1.7	2.6 ± 1.8	2.8 ± 2.5	.120
300	2.4 ± 1.5	2.5 ± 2.1	2.4 ± 1.5	2.4 ± 1.7	.875
400	2.4 ± 1.5	2.4 ± 1.5	2.4 ± 1.4	2.4 ± 1.5	.937
450	2.3 ± 1.4	2.4 ± 1.5	2.5 ± 1.6	2.4 ± 1.5	.127
500	2.4 ± 1.8	2.4 ± 1.5	2.4 ± 1.5	2.4 ± 1.6	.996
600	2.3 ± 1.3	2.3 ± 1.4	2.3 ± 1.4	2.3 ± 1.4	.503
700	2.2 ± 1.3	2.3 ± 1.4	2.2 ± 1.4	2.3 ± 1.4	.678

difference among the four repetition subsets of the same sample size (Table 2).

The AI development time for subsets with a sample size of 75 was the shortest, taking 102 minutes, whereas the subset with a sample size of 700 took the longest, at 3447 minutes when computed by an ordinary desktop computer at the authors' lab.

The magnitude of the bias, range, and variance of the prediction errors, expressed as the 95% confidence boundary ellipses, decreased as the sample sizes increased (Figure 2). Although the scatterplots for sample sizes of 100 and 300, shown in Figure 2, might appear to show discrepancies in error ranges, there was no statistically significant difference among the subsets that were repeated four times at the same sample size (Table 2).

The result of the linear regression analysis indicated that the prediction error could decrease by 0.7 mm with every increase of 1000 in sample size. When the linear regression line was plotted, a sample size of approximately 1700 datasets was estimated to be the optimal sample size (Figure 3).

DISCUSSION

The ultimate goal of this study was to estimate the necessary size of longitudinal serial data collection from patients who have undergone combined surgical-orthodontic treatment. The result demonstrated increased prediction accuracy with increasing sample sizes. Although the study by Lee et al. suggested that increasing the sample size for growth predictions led to higher prediction errors,[19] the results of the present study were contrary to this finding. In fact, the current results were in greater alignment with the common belief that a larger sample size enhances the performance of a developed prediction model.[26,27]

The result of the present study also suggested that collecting data from approximately 1700 patients might be necessary to develop an AI model with an error $<$ 1.5 mm at the lower lip area. When developing an AI model, one of the challenging obstacles is collecting sufficient data. In general, a larger dataset is more beneficial, which may play an essential role in developing AI systems for accurately predicting the outcomes of orthognathic surgery.[20] If a more accurate visual treatment objective for use in clinical orthodontic practice can be developed, it will function as an efficient consulting tool to enhance communication between patients and clinicians.

Currently, many commercial STO software programs are available to visualize changes after orthodontic and surgical treatments. However, these programs typically rely on a fixed value of 1-to-1 correspondence ratio between skeletal repositioning and specific soft tissue landmarks. This approach may oversimplify the complex nature of soft tissue response after surgery, potentially leading to prediction errors. For instance, the lower lip was previously considered as one of the most unpredictable soft tissue regions due to its variability to postural changes. Even minor adjustments in head or lip posture can result in significant variability of lower lip position. In addition, in patients with severe malocclusions, the lips are often strained or flaccid, amplifying the complexity in prediction. However, recent advancements in AI technology have significantly improved the accuracy of lower lip predictions.[4,9,10] Given this progress, if an AI model can reliably predict changes of lower lip, which is one of the most challenging areas to predict, it is likely to perform well for other soft tissue regions as well. Due to its predictive complexity, the lower lip has often been used as a benchmark in various studies to evaluate the predictive performance of AI models.[12,16,17,19,20,23,28]

Determining an adequate sample size prior to experimentation has been emphasized as an essential first step of research. Although studies with small samples tend to be less convincing and inconclusive due to the low statistical power, collecting more samples than required wastes resources. Accordingly, there are various instructions to calculate the optimal sample size. For example,

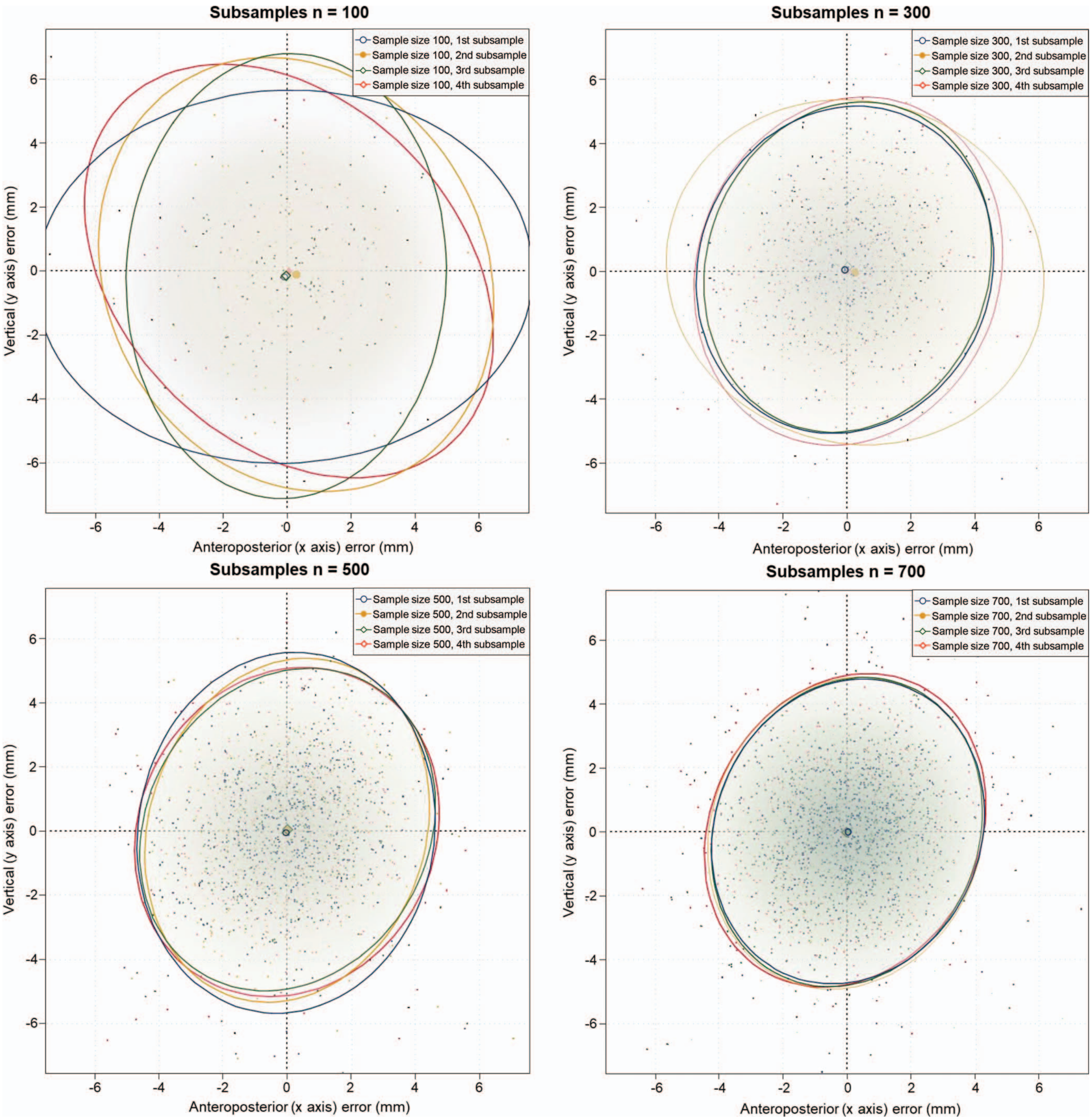

Figure. 2. Scatterplots with 95% confidence ellipses representing the patterns of prediction errors (mm) for the lower lip. Each scatterplot was generated for every resampling subset, but only four subsample sizes were selected to succinctly demonstrate that the variance of prediction errors decreases as the sample size increases.

in the context of a *t*-test to compare two means, an obvious formula exists to calculate sample sizes, and the sample size calculation depends upon the statistical power (also called 1—beta, type II error rate, or false-negative), probability value (also called alpha, type I error rate, or false-positive), previously known means, and standard deviations.[29] Several well-known inferential tests, such as correlation statistics, also have formulae to calculate sample sizes.[26] However, for developing AI, since no such formula exists, pilot studies and an empirical approach using resampling and subsampling might be the only options.[19,20]

As the first step in sample-size calculation for a *t*-test is deciding what is an expected between-group difference to be pursued by the researcher, the first step in developing an AI prediction model may be deciding

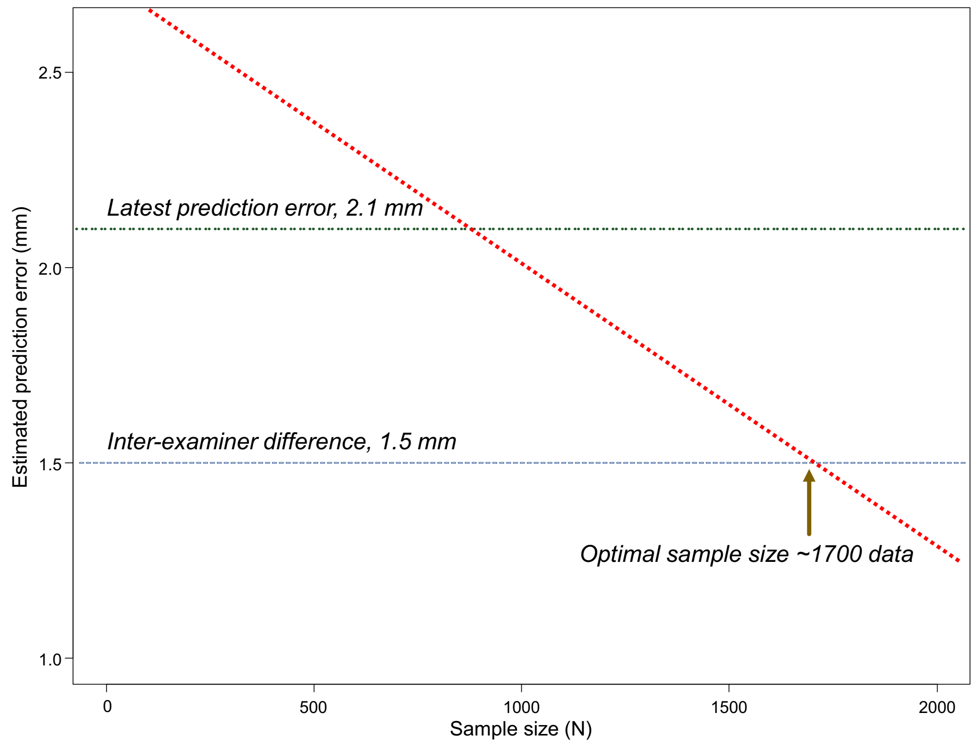

Figure. 3. The prediction error decreased as sample sizes increased. The regression line indicated that a sample size greater than 1700 would result in an error of less than 1.5 mm.

on an acceptable level of error, or a threshold value. As the threshold for clinically acceptable STO error, 1.5 mm was selected for the following reasons: (1) a 1.5-mm error has traditionally been recognized as an overall standard landmark identification error in cephalometrics[24]; (2) studies have shown that the interexaminer difference in landmark identification among various human examiners is 1.5 mm[17]; (3) although a 2.0-mm criterion is commonly used in AI performance contests and conferences organized by the International Symposium on Biomedical Imaging, 1.5 mm could be considered to be a stricter and more conservative standard.[15] As a result, the 1.5-mm threshold seemed to have been referenced in numerous previous publications.[13,14,18–20] Please note that radial error was used in this study as the prediction error instead of absolute error. In the past, reporting absolute error values was more common. However, radial error is now more widely used in fields such as computer science and statistics, making it a more popular choice for reporting errors currently, as shown in Table 1.[9,10,21]

The present study had a notable limitation in that it exclusively focused on statistical and AI study design aspects of sample size matters. For example, one of the hyperparameters that could account for prediction accuracy, the greatest number of training epochs (also called the early stopping condition), was fixed at 1000 epochs so that the computation procedures and pilot studies could be completed within a couple of months. This was because pilot studies applying 10,000 training epochs did not demonstrate a significant increase in prediction accuracy, but extended computation times considerably. In addition, since predictive performance was assessed only for the lower lip, the prediction accuracy might not be generalized to other soft tissue landmarks in the mid-face and chin. Also, as the subjects were of Korean ethnicity, the AI model might not be applicable to other populations.

The sample size estimation method of the present study was inspired by the method introduced by Kim et al. that emphasized the use of pilot studies based on resampling subsets with reduced sample sizes, repetitions, and preliminary creation of AI models.[20] The method suggested in the present study may help research design for developing AI models for use in clinical orthodontic practice.

CONCLUSIONS

- The present study described a method of estimating the necessary sample sizes required to develop an AI model prior to experimentation.
- From the statistical and research design point of view, it appears that a substantial amount of training data may be essential to develop more accurate surgical treatment objectives (STOs).

ACKNOWLEDGMENTS

This study was partly 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 partly supported by grant 05-2025-0507 from the SNUDH Research Fund.

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

REFERENCES

1. Youn SB, Oh HJ, Son IS, Lee SJ, Sohn HB, Seo BM. Does the sequence of bimaxillary orthognathic surgery affect accuracy in Skeletal Class Iii patients? *J Oral Maxillofac Surg.* 2024;82:1402–1415.
2. Oh HJ, Son IS, Lee SJ, Sohn HB, Seo BM. Effect of maxillary impaction on mandibular surgical accuracy in virtually-planned orthognathic surgery: a retrospective study. *J Craniomaxillofac Surg.* 2023;51:387–392.
3. Oh HJ, Moon JH, Ha H, et al. Virtually-planned orthognathic surgery achieves an accurate condylar position. *J Oral Maxillofac Surg.* 2021;79:1146.e1–1146.e25.
4. 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.
5. Suh HY, Lee HJ, Lee YS, Eo SH, Donatelli RE, Lee SJ. Predicting soft tissue changes after orthognathic surgery: the sparse partial least squares method. *Angle Orthod*. 2019;89:910–916.
6. Lee YS, Suh HY, Lee SJ, Donatelli RE. A more accurate soft-tissue prediction model for Class III 2-jaw surgeries. *Am J Orthod Dentofacial Orthop.* 2014;146:724–733.
7. Lee HJ, Suh HY, Lee YS, et al. A better statistical method of predicting postsurgery soft tissue response in Class II patients. *Angle Orthod.* 2014;84:322–328.
8. Suh HY, Lee SJ, Lee YS, et al. A more accurate method of predicting soft tissue changes after mandibular setback surgery. *J Oral Maxillofac Surg.* 2012;70:e553–562.
9. Yu W, Lee SJ, Cho H. Partial least squares regression trees for multivariate response data with multicollinear predictors. *IEEE Access.* 2024;12:36636–36644.
10. 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.
11. Haskell BS, Segal ES. Ethnic and ethical challenges in treatment planning: dealing with diversity in the 21st century. *Angle Orthod.* 2014;84:380–382.
12. Moon JH, Kim MG, Cho SJ, et al. Evaluation of automated photograph-cephalogram image integration using artificial intelligence models. *Angle Orthod.* 2024;94:595–601.
13. Hwang HW, Moon JH, Kim MG, Donatelli RE, Lee SJ. Evaluation of automated cephalometric analysis based on the latest deep learning method. *Angle Orthod.* 2021;91:329–335.
14. Hwang HW, Park JH, Moon JH, et al. Automated identification of cephalometric landmarks: Part 2-Might it be better than human? *Angle Orthod.* 2020;90:69–76.
15. Park JH, Hwang HW, Moon JH, et al. Automated identification of cephalometric landmarks: Part 1-Comparisons between the latest deep-learning methods YOLOV3 and SSD. *Angle Orthod.* 2019;89:903–909.
16. 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.
17. 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
18. 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.
19. 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.
20. 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.
21. Arik SÖ, Pfister T. TabNet: attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021;35:6679–6687.
22. 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.
23. 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.
24. Moon JH, Lee JM, Park JA, Suh H, Lee SJ. Reliability statistics every orthodontist should know. *Semin Orthod*. 2024;30:45–49.
25. R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2025. https://www.R-project.org/
26. Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*. St. Louis, Missouri: Mosby Year Book; 1994.
27. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol 2. 2nd ed. New York, NY: Springer; 2009.
28. 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.
29. Pandis N. Sample calculations for comparison of 2 means. *Am J Orthod Dentofacial Orthop.* 2012;141:519–521.