AJO-DO

Sample size calculation for an artificial intelligence study

Jong-Hak Kim,[a] Naeun Kwon,[a] Nikolaos Pandis,[b,c] and Shin-Jae Lee[d]
Seoul, South Korea, and Bern, Switzerland, and Corfu, Greece

From a statistical and research design perspective, the importance of determining an adequate sample size before experimentation should not be underestimated. Studies with insufficient sample sizes often yield inconclusive results owing to low statistical power, whereas oversampling leads to unnecessary expenditure of time and resources. As such, one of the fundamental questions at the outset of the research design is, "How many participants are required for this study?"[1] To answer the question, various instructions, guidelines, and/or formulas exist to help researchers estimate the sample sizes required for statistical analyses.[1-4] Sample calculation methods for more sophisticated inferential tests can be assisted by various software programs available.[5]

However, unlike conventional statistical analyses, no sample size guidelines have been established for developing artificial intelligence (AI) models that have become popular in orthodontics. Thus, in the instance of AI studies, determining the optimal training sample size or learning datasets would require a heuristic or an empirical approach based on a simulation study using resampling and subsampling.[6] To restate, because there is no obvious formula or sample size determination software, researchers have no means except pilot studies to establish the research design.

Developing AI models requires not only significant computing resources but also a much larger sample size than conventional statistical analyses. For example, the development of an automatic landmark identification AI model was reported to require 2300 training data sets to detect 19 landmarks and 5400 data sets to detect 80 landmarks.[7] These findings clearly indicate that the necessary sample size for developing AI would be significantly larger than expected. Despite the recent surge in AI studies, many authors overlook the importance of appropriate sample sizes, which may need to be larger than currently used.[6,8]

This study describes a sample size estimation procedure for developing an AI model. The sequence for estimating the sample size is summarized as follows:

1. First, we must decide or define a clinically acceptable error or accuracy to be observed. In addition to defining the error, a threshold value for the prediction error or benchmark for accuracy should be established.
2. The next step was to build several subsets with smaller sample sizes using a random resampling method. Several repetitions were required. Then, train and create AI models using the resampled subsets (Fig 1).
3. The optimal data size can be estimated using a graph that visualizes the performance of the AI models according to different sample sizes (Fig 2).

To illustrate this procedure, we use an example of performing a sample size calculation to develop an AI model that can predict craniofacial growth using datasets consisting of longitudinal serial cephalometric radiographs. Example datasets are presented by Moon et al.[9,10]

For this example, we suggest the growth prediction error as the primary outcome variable, which is defined as the Euclidean distance or radial error (in millimeters) between the predicted and real growth changes of a cephalometric landmark. Allowing for lenient errors may reduce the required sample size; however, the resulting AI prediction model may lack clinical relevance because of a higher margin of error. Conversely, setting a very small acceptable error would significantly increase the required sample size, making it impractical.

[a]Department of Orthodontics, Seoul National University, Seoul, Korea.
[b]Department of Orthodontics and Dentofacial Orthopedics, Dental School, Medical Faculty, University of Bern, Bern, Switzerland.
[c]Private practice, Corfu, Greece.
[d]Department of Orthodontics and Dental Research Institute, Seoul National University School of Dentistry, Seoul, Korea.

All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest, and none were reported.

Address correspondence to: 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.

Am J Orthod Dentofacial Orthop 2025;167:616-20
0889-5406/$36.00

https://doi.org/10.1016/j.ajodo.2025.01.008

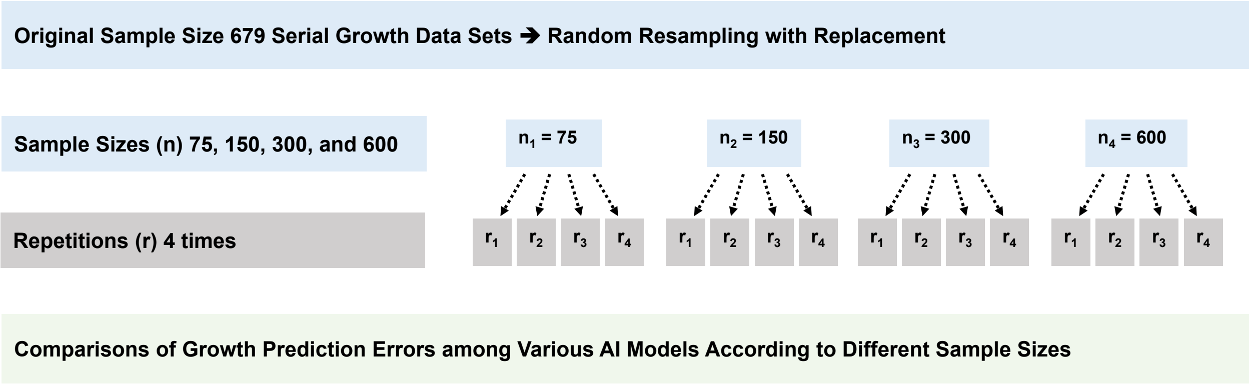
Original Sample Size 679 Serial Growth Data Sets ➔ Random Resampling with Replacement
Sample Sizes (n) 75, 150, 300, and 600
Repetitions (r) 4 times
n_1 = 75
n_2 = 150
n_3 = 300
n_4 = 600
r_1
r_2
r_3
r_4
Comparisons of Growth Prediction Errors among Various AI Models According to Different Sample Sizes

Fig 1. Pilot study design summary. From the original data source (N = 679), 4 subsamples with reduced sample sizes (75, 150, 300, and 600) were resampled. Each resampling was repeated 4 times.

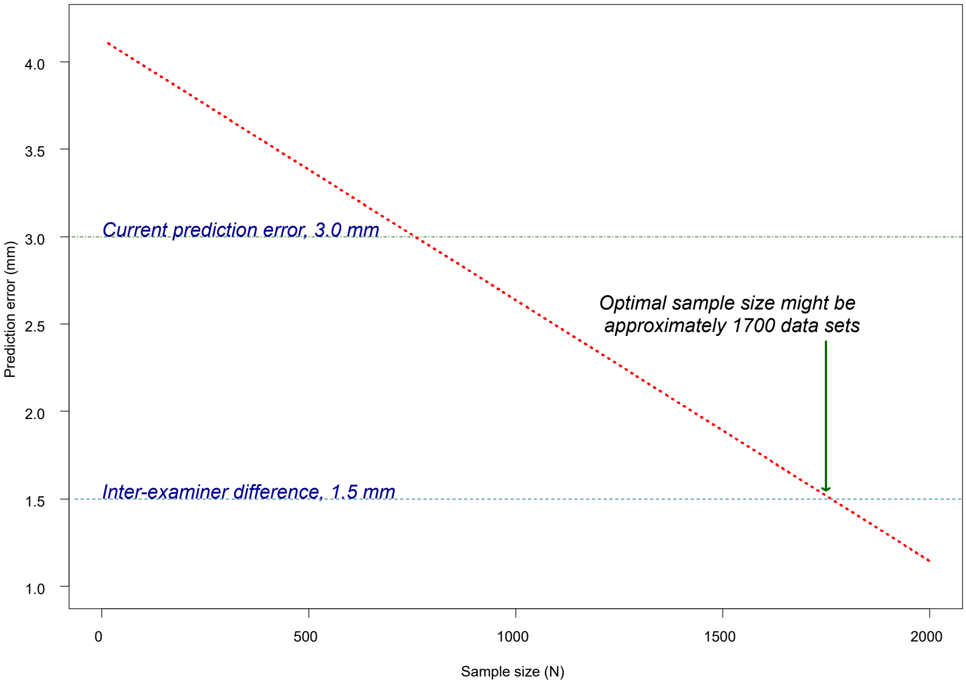

Fig 2. The hypothetical estimation line suggested that approximately 1700 data sets might be required to develop an AI model that can predict growth. The 1.5 mm threshold line was assumed to be the clinically acceptable accuracy, also known as the common interexaminer reliability among human examiners.

For clinically acceptable accuracy, let us consider the growth prediction error <1.5 mm. Selecting a threshold value of 1.5 mm might be reasonable because a 1.5 mm error has traditionally been recognized as an overall landmark identification error in cephalometrics.[11] In addition, studies have shown that the interexaminer difference in landmark identification among various human examiners is also 1.5 mm.[12-14]

Although the original datasets included growth changes across 78 cephalometric landmarks, for illustrative purposes, this article focuses only on a single landmark, the most prominent point of the lower lip, also called the labrale inferius. So far, the most accurate growth prediction error in the lower lip was known to be 3.0 mm.[10] This article supposes a problem formulation situation for developing a more accurate growth

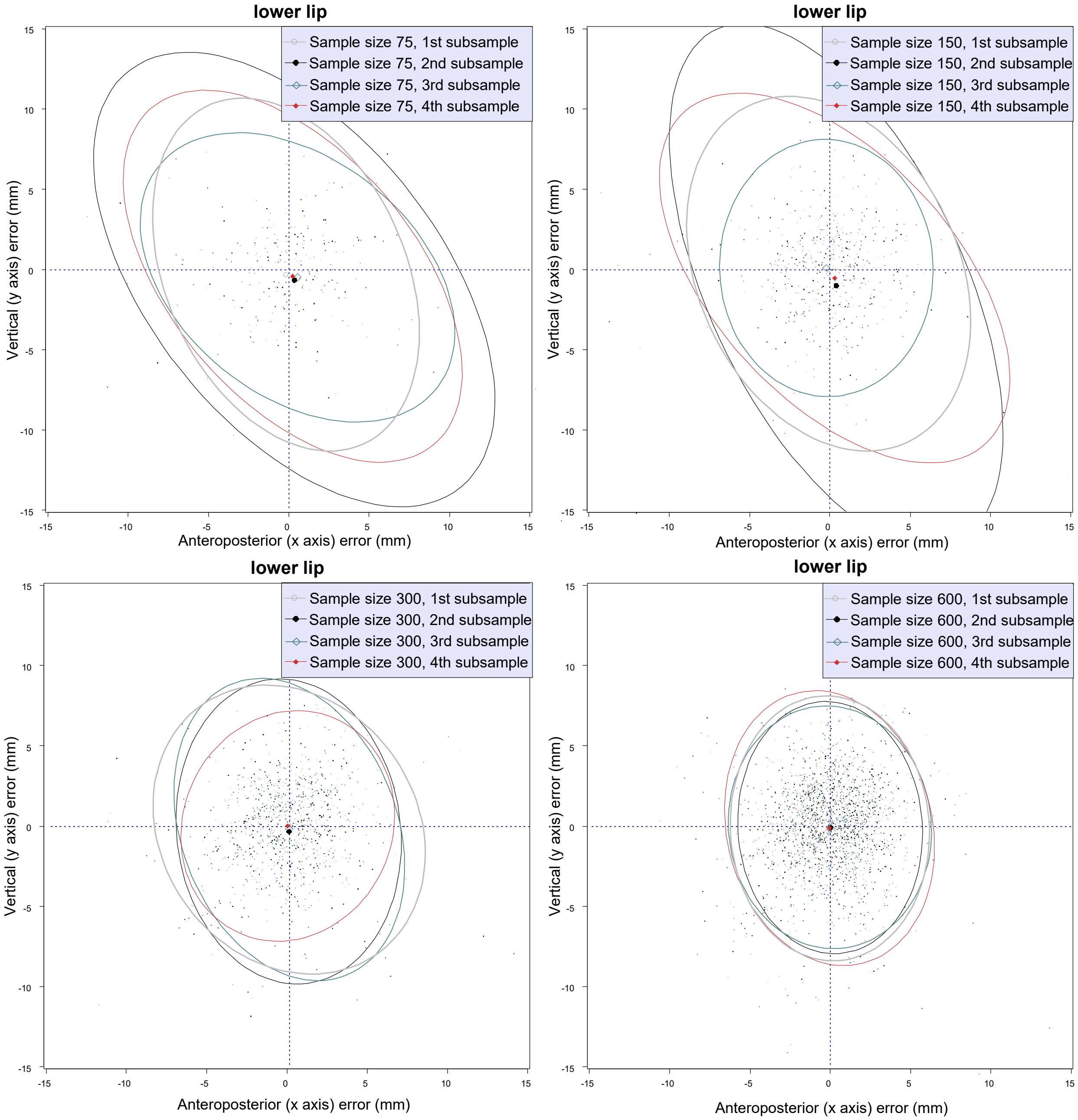

Fig 3. Scatterplots of the prediction errors of several AI models show that prediction errors decrease as sample sizes increase. As the sample size increases, the dispersion of prediction errors among the resampling subsets also decreases.

prediction AI model that will demonstrate a <1.5 mm prediction error in the lower lip.

Random resampling can be conducted using various software programs such as Microsoft Excel (Redmond, Wash). At least 3 sample sizes and 2 repetitions are recommended when selecting the number of reduced sample sizes. The next step was to train various AI models using the resampling subsets.

In this growth prediction example, four subsets and four repetitions were applied. From the original data source, which included 679 longitudinal serial growth datasets, 4 subsamples with smaller sample sizes, consisting of 75 (n_1), 150 (n_2), 300 (n_3), and 600 (n_4) of these datasets, were selected by random resampling with replacement procedures. Each was repeated 4 times (r_1, r_2, r_3, r_4), resulting in 16 subsets (Fig 1). To develop the AI models, the TabNet deep neural network (Arik and Pfister, 2021, Stanford, Calif), a type of convolutional neural network,[15] was applied to 16 data subsets.

The next step is specifying the relationship between AI models' predictive performance and training sample sizes. The expected predictive performance was determined through the aforementioned pilot experiments, using resampled subsets with reduced data sizes. Here, the performance is the resultant prediction error of the AI model. Figure 3 demonstrates that the growth prediction error decreased with increasing sample size.

The final step is estimating the sample size on a graph. At this stage, depicting a graph such as Figure 2 is essential. Different study formulations would reveal not only a linear relationship but also other types of associations between AI performance and sample sizes. For example, in instances of developing automated landmark detection AI models, the detection error has demonstrated a linear relationship between the detection accuracy and training sample sizes on a logarithmic scale.[7] By depicting a graph, the logarithmic, exponential, or polynomial relationships can be revealed and visualized.

This relationship can be expressed using a straightforward equation derived from a linear regression analysis. However, it is recommended that a graph be depicted rather than a regression equation. The sample size estimation for an AI study is an approximation of the empirical procedure. Sometimes, this could be a wild guess because of the many unexplainable factors that might have affected the outcome.

For example, as shown in Figure 2, we can estimate the required sample size on a graph by examining the linear relationship between the prediction error and the sample size. By extrapolating the expected error that is approximately close to the acceptable error value defined from the research formulation stage, Figure 2 indicates that the required sample size to gain less than 1.5 mm prediction error would be close to 1700 datasets. In other words, if we assume that the prediction error is directly proportional to the sample size used to train the AI models, then the optimal sample size may be 1700.

Here, we said "optimal" because a smaller sample size would never guarantee sufficient prediction accuracy, and a greater sample size would be unnecessary, not considering the innate and inevitable nature of growth prediction errors. In addition, it is conventional to acknowledge a total of 1.5 mm error in cephalometric procedures, from the image-taking stage to the landmark identification stage. Therefore, a growth prediction accuracy measure of <1.5 mm may be clinically irrelevant.

The sample size estimation result of this example, as shown in Figure 2, can be as follows: "Based on the linear relationship between the prediction accuracy over the training sample sizes, an optimal sample size was calculated to be approximately 1700."

In practical terms, when planning an AI study, pilot studies are mandatory not only for sample size estimation purposes but also for estimating the computation time required to develop the AI model. Without considering these 2 issues, the result may not guarantee sufficient predictive performance, or the training process may take too long (several months or years) to obtain the result, even with high-speed computers.[6]

The estimated sample size changes if any of the deep-learning algorithm's tuning hyperparameters change. For example, the early stopping condition or the number of training epochs significantly affects an AI model's predictive performance. Therefore, we recommend repeating the aforementioned sample size calculation procedure for several hyperparameter conditions.

CREDIT AUTHOR STATEMENT

Jong-Hak Kim: Software; Writing – original draft. **Naeun Kwon:** Data curation; Project administration. **Nikolaos Pandis:** Writing – review & editing. **Shin-Jae Lee:** Conceptualization; Formal analysis.

REFERENCES

1. Pandis N. Sample calculations for comparison of 2 means. Am J Orthod Dentofacial Orthop 2012;141:519-21.
2. Pandis N. Sample calculations for comparing proportions. Am J Orthod Dentofacial Orthop 2012;141:666-7.
3. Pandis N, Machin D. Sample calculations for comparing rates. Am J Orthod Dentofacial Orthop 2012;142:565-7.
4. Pandis N. The effect size. Am J Orthod Dentofacial Orthop 2012;142:739-40.
5. Guo Y, Pandis N. Sample-size calculation for repeated-measures and longitudinal studies. Am J Orthod Dentofacial Orthop 2015;147:146-9.
6. Lee JM, Moon JH, Park JA, Kim JH, Lee SJ. Factors influencing the development of artificial intelligence in orthodontics. Orthod Craniofac Res 2024;27:6-12.
7. Moon JH, Hwang HW, Yu Y, Kim MG, Donatelli RE, Lee SJ. How much deep learning is enough for automatic identification to be reliable? Angle Orthod 2020;90:823-30.

8. Rousseau AJ, Geubbelmans M, Burzykowski T, Valkenborg D. Deep learning. Am J Orthod Dentofacial Orthop 2024;165:369-71.
9. 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;92:705-13.
10. Moon JH, Shin HK, Lee JM, Cho SJ, Park JA, Donatelli RE, et al. Comparison of individualized facial growth prediction models based on the partial least squares and artificial intelligence. Angle Orthod 2024;94:207-15.
11. Moon J-H, Lee J-M, Park J-A, Suh H, Lee S-J. Reliability statistics every orthodontist should know. Semin Orthod 2024;30:45-9.
12. 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;95:219-26.
13. Moon JH, Kim MG, Cho SJ, Ko DY, Hwang HW, Park JA, et al. Evaluation of automated photograph-cephalogram image integration using artificial intelligence models. Angle Orthod 2024;94:595-601.
14. 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.
15. Arik SÖ, Pfister T. TabNet: attentive interpretable tabular learning. Proceedings of the AAAIconference on artificial intelligence 2021 2021;35:6679-87.