Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Data-driven prediction model for periodontal disease based on correlational feature analysis and clinical validation

Woosun Beak [a,b,1], Jihun Park [c,1], Suk Ji [a,d,*]

[a] *Department of Dental Public Health, Ajou University Graduate School of Clinical Dentistry, Suwon, Republic of Korea*
[b] *Department of Dentistry, Gyeonggi Provincial Medical Center Suwon Hospital, Suwon, Republic of Korea*
[c] *Department of Materials Science and Engineering, University of Maryland, College Park, MD, USA*
[d] *Department of Periodontology, Institute of Oral Health Science, Ajou University School of Medicine, Suwon, Republic of Korea*

ARTICLE INFO

Keywords:
Periodontal disease
Periodontitis
Risk factor assessment
Data-driven prediction model
Clinical validation

ABSTRACT

Objectives: This study aimed to investigate the performance and reliability of data-driven models employing correlational feature analysis and clinical validation for predicting periodontal disease.

Methods: The 7th Korea National Health and Nutrition Examination Survey ($n = 10,654$) was used for correlation analysis to identify significant risk factors for periodontitis. Periodontal prediction models were developed with the selected factors and database, followed by internal validation with 5-fold cross-validation and 1000 bootstrap resampling. External validation was conducted with clinical data ($n = 120$) collected through self-reported questionnaires, clinical periodontal parameters, and radiographic image analysis. Predictive performance was assessed for logistics regression, support vector machine, random forest, XGBoost, and neural network algorithms using the area under the receiver operating characteristic curves (AUC) and other performance metrics.

Results: Correlation analysis identified 16 features from over 1000 potential risk factors for periodontitis. The best data-driven model (XGBoost) showed AUC values of 0.823 and 0.796 for internal and external validations, respectively. Modeling with clinical data revealed those same measures to be 0.836 and 0.649, respectively. In addition, the data-driven model could predict other clinical periodontal parameters including severe bone loss (AUC = 0.813), gingival bleeding (AUC = 0.694), and tooth loss (AUC = 0.734). A patient case study about prognostic predictions revealed that the probability of periodontitis can be reduced by 6.0 % (stop smoking) and 0.6 % (stop drinking) on average.

Conclusions: Data-driven models for predicting periodontitis and other periodontal parameters were developed from 16 risk factors, demonstrating enhanced prediction performance and reproducibility in internal–external validations.

Clinical significance: Clinicians can use data-driven prediction models for early detection and preventive care of periodontal disease. The models are expected to provide useful prognostic information about patient-specific periodontal risk factors.

[*] Corresponding author. Department of Periodontics, Ajou University Hospital, 164, World cup-ro, Yeongtong-gu, Suwon, Republic of Korea.
E-mail address: sukji@ajou.ac.kr (S. Ji).

[1] Equally contributed to this work.

https://doi.org/10.1016/j.heliyon.2024.e32496
Received 9 October 2023; Received in revised form 3 June 2024; Accepted 5 June 2024
Available online 5 June 2024

1. Introduction

Periodontal disease (PD) has become one of the most significant concerns in global oral healthcare, as demonstrated by its prevalence in 19 % of the adult population and one billion cases worldwide [1]. PD typically occurs progressively owing to several risk factors [2], starting with gingival bleeding or swelling in the early stage (gingivitis) to loss of teeth and bone in its more severe form (periodontitis). Due to this progressiveness with age as a critical risk factor, the disease is particularly prevalent and severe among older adults, with 70.1 % of people aged 65 years and older suffering from periodontitis in the United States [3]. However, such oral diseases are mostly preventable or delayable if diagnosed and treated at an early stage [1]. Therefore, early detection and preventive care are imperative for protecting community oral health from PD.

PD can be diagnosed through clinical methods such as periodontal probing depth, clinical attachment level, and radiographic bone loss analysis [4–11]. One of the standardized classifications for periodontal severity is the Community Periodontal Index (CPI) based on probing depth, as recommended by the World Health Organization [8]. While CPI can diagnose PD accurately, it can identify only the current state of the disease and has no predictive ability. This limitation makes it necessary to search for alternative methods to predict the disease for early detection and preventive measures.

Prediction models can be helpful for early-stage detection and prognostic analysis because they can identify high-risk patients before initiation or progression of the disease [9–20]. Such models can be built upon periodontal risk factors (predictor variables) from self-reported items, which are neither costly nor invasive to obtain and facilitate large-scale and epidemiologic studies [9–11]. In addition, this approach could efficiently assess patient-specific disease risk for individual prescriptions and preventive care [17–20]. Various assessment tools, guidelines, algorithms, and methods have been developed and evaluated for accurate clinical predictions [21–23].

Nevertheless, PD prediction models often encounter inherent reliability issues for the following reasons. First is the limited amount of data available for model training, which possibly causes random or systematic errors [21–25]. Prediction models are based on supervised learning and require costly and time-consuming clinical measurements for PD labeling, which is why most clinical prediction models utilize a small patient population (tens to a few hundred) [10–16]. Such a small sample size limits the models to overfitting, deteriorating model stability and reproducibility [23–27]. Also, the prevalence of PD in clinical patient groups may deviate from that in the national population, leading to biased or imbalanced data distribution. These problems increase the risk of bias for modeling and reduce model reliability, limiting the general use of prediction models [25,28]. Recent research has revealed, based on reliability assessment tools for prediction models (e.g., PROBAST [28] and TRIPOD [29]), that many studies on PD prediction have been exposed to a high risk of bias [23].

Modeling with a large-scale database (e.g., data from thousands to tens of thousands of patients) could reduce the risk of bias and enhance model reliability [25,27]. In addition, it provides a large pool of potential risk factors for feature selection [20]. Such databases include national health examination surveys [9,30–33] and electronic dental records [20,34,35] and allow reasonably accurate predictions. However, data-driven predictions for PD must be not merely accurate, but also reproducible and reliable enough to be applied in clinical practices. In this regard, clinical verification can be the most appropriate external validation of model performance

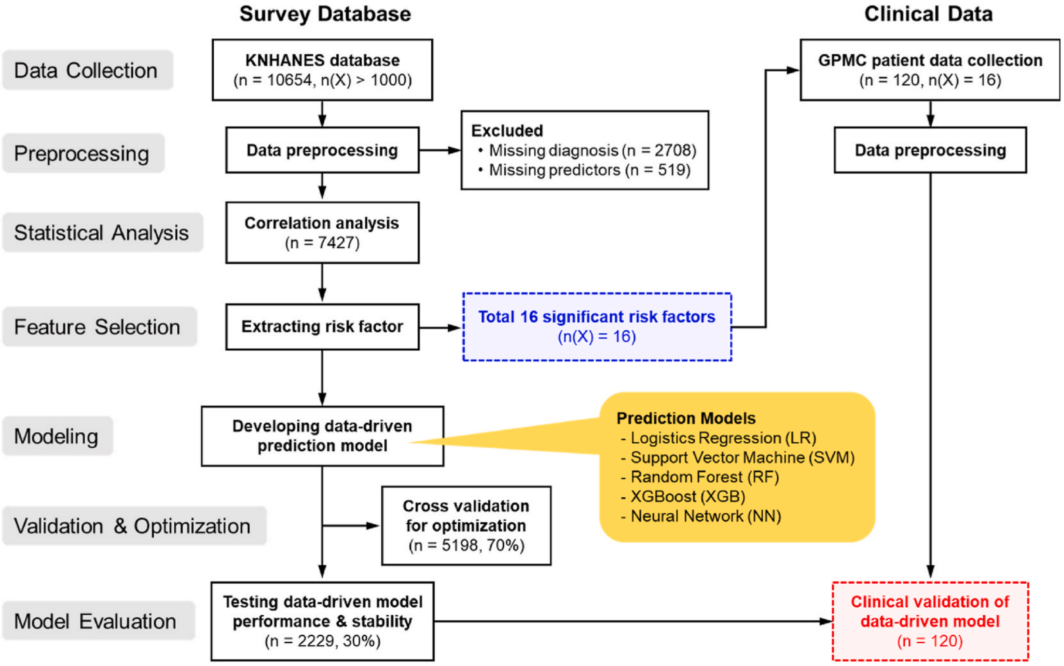
Survey Database
Clinical Data
Data Collection
Preprocessing
Statistical Analysis
Feature Selection
Modeling
Validation & Optimization
Model Evaluation
KNHANES database (n = 10654, n(X) > 1000)
Data preprocessing
Excluded
• Missing diagnosis (n = 2708)
• Missing predictors (n = 519)
Correlation analysis (n = 7427)
Extracting risk factor
Total 16 significant risk factors (n(X) = 16)
Developing data-driven prediction model
Prediction Models
- Logistics Regression (LR)
- Support Vector Machine (SVM)
- Random Forest (RF)
- XGBoost (XGB)
- Neural Network (NN)
Cross validation for optimization (n = 5198, 70%)
Testing data-driven model performance & stability (n = 2229, 30%)
GPMC patient data collection (n = 120, n(X) = 16)
Data preprocessing
Clinical validation of data-driven model (n = 120)

Fig. 1. Overall workflow chart.

and reproducibility. Procedures and protocols for data-driven approaches are also needed regarding feature selection, clinical data collection, modeling with large databases, optimization, and internal-external validations.

Therefore, this study aimed to develop data-driven prediction models for assessing PD risk and to evaluate their performance and reliability with clinical patient data for external validation.

2. Materials and methods

2.1. Study design and data preparation

Fig. 1 shows the overall flowchart of the present work. The 7th Korea National Health and Nutrition Examination Survey (KNHANES) conducted from 2016 to 2018 by the Korea Center for Disease Control and Prevention was employed to construct data-driven prediction models. This survey database involves a total data set of 10,654 subjects ($n = 10,654$) with more than 1000 potential risk factors (n(X)) as predictor variables. These potential risk factors can be broadly categorized as follows: data information, personal information, lifestyle, education level, economic activities, oral health conditions, disease and treatment, checkup and health screening, blood test, and nutritional intake. Of the 10,654 patients, we excluded 2708 and 519 due to missing clinical diagnosis (i.e., no CPI score) and missing predictor variables (i.e., missing any significant risk factor), respectively. Consequently, we used data from 7427 patients ($n = 7427$) for prediction modeling. The demographic and clinical characteristics of the patients in the KNHANES database are summarized in Table 1.

2.2. Statistical analysis and feature selection

We performed statistical analysis on the KNHANES database to select significant risk factors for PD, where the Pearson correlation was employed to examine pair-wise correlation [36]. First, we removed features with apparent irrelevance to PD (e.g., data collection date, patient serial number) or features with a risk of diminishing data volumes (more than 10 % of the total data volume) due to missing predictor variables. This step filtered out more than 700 potential risk factors, leaving 304 factors. An initial correlation analysis was conducted between the remaining potential risk factors and PD (CPI score = 3, 4; CPI3–4). We extracted significant risk factors from the initial correlation result for prediction modeling (Fig. S1). The extraction process also includes eliminating redundant or repetitive factors to ensure factor uniqueness. In this process, we considered the inter-correlations between risk factors as well as the correlations between a risk factor and PD in order to identify the effectiveness of features in PD prediction modeling. We have selected the optimal number of significant risk factors based on computational efficiency since a complicated model with various prediction variables (X) likely show a trade-off between computation costs (time and hardware output) and predictive performance depending on training data volume. Last, the final correlation analysis was performed for the selected features. All statistical analyses were performed using SPSS 26.0 (IBM, Chicago, IL, USA) and the Python Scikit-learn library.

2.3. Clinical data collection

According to the significant risk factors extracted, we collected clinical patient data from Gyeonggi Provincial Medical Center (GPMC) Suwon Hospital in South Korea through self-report questionnaires, clinical periodontal measurements, and radiographic

Table 1
Demographic and clinical characteristics of the KNHANES and GPMC patient data.

Characteristics	KNHANES database	GPMC clinical data
Number of patients, *n*	7427	120
Age		
10–29, *n* (%)	1702 (22.9)	11 (9.2)
30–49, *n* (%)	2834 (38.2)	26 (21.6)
50–69, *n* (%)	2331 (31.4)	48 (40.0)
≥ 70, *n* (%)	560 (7.5)	35 (29.2)
mean ± SD, years	43.8 ± 17.4	57.5 ± 17.0
Sex		
Male, *n* (%)	3258 (43.9)	70 (58.3)
Female, *n* (%)	4169 (56.1)	50 (41.7)
CPI Score		
0 (healthy), *n* (%)	2503 (33.7)	10 (8.3)
1 (bleeding), *n* (%)	472 (6.4)	6 (5.0)
2 (calculus), *n* (%)	2484 (33.4)	2 (1.7)
3 (PD = 4–5 mm), *n* (%)	1485 (20.0)	42 (35.0)
4 (PD ≥ 6 mm), *n* (%)	483 (6.5)	60 (50.0)
Mean ± SD	1.59 ± 1.31	3.13 ± 1.21

KNHANES, Korea National Health and Nutrition Examination Survey; GPMC, Gyeonggi Provincial Medical Center; SD, standard deviation; CPI, community periodontal index; PD, probing depth.

image analysis. The Institutional Review Board for Human Subjects of Ajou University Dental Hospital approved the clinical data collection (IRB No. AJOUIRB-SB-2023-008). All participants were recruited from January 2023 to May 2023, were older than 18 years, and voluntarily participated in this study with written informed consent. Table 1 includes the demographic and clinical characteristics of the GPMC patient data.

The number of required participants was estimated based on the power calculation with the KNHANES database. We estimated that 120 participants ($n = 120$) were required for this study after considering a p-value ≤ 0.05 with 90 % power, modeling data volume for train–test split, and a possible drop-out and missing data rate of 30 %.

Periodontal risk factors were obtained from a self-report questionnaire and oral examination of 120 participants. Table S1 shows the normalization of these predictor variables, which were attained based on data synchronization with the KNHANES database and model simplification for effective clinical validation.

Clinical measurements were performed by one qualified examiner to secure consistent clinical periodontal parameters from the 120 participants. The descriptions and classification criteria for all the clinical parameters used in this study are presented in Table S2. The CPI score was obtained based on probing depth, blood on probing, and visual inspection. In addition, the gingival index was measured from all participants. The average probing depth of total sections and individual teeth from each sextant was also investigated, corresponding to Probing Depth (Mean) and Probing Depth (Individual). Loss of supporting bone tissue, root length, tooth length, and number of lost teeth were estimated for all teeth from each participant based on radiographic image analysis. We defined Bone Loss Ratio (PD) as a ratio of supporting bone tissue loss to root length greater than one-third and Bone Loss Ratio (PD+) as a ratio larger than one-third in >30 % of total sites, as previously defined [6].

2.4. Development and optimization of data-driven models

The detailed workflow of building and optimizing data-driven predictive models is depicted in Fig. S2(a). We randomly split the KNHANES data set ($n = 7427$) into the training set (70 %, $n = 5198$) and the test set (30 %, $n = 2229$) for model development and evaluation, respectively. The extracted periodontal risk factors were used as predictor variables, and CPI 3–4 was used as the outcome variable for model training. We examined five model algorithms that have been reported to successfully predict PD and other dental diseases: logistics regression (LR) [10–14,30–32], support vector machine (SVM) [15,23,37], random forest (RF) [23,36,37], extreme gradient boosting (XGB) [20,36], and neural network (NN) [23,33,37]. A grid search with a stratified five-fold cross-validation technique was utilized for optimizing model-specific control parameters (hyperparameters), followed by final model retraining with the entire training set. The stratification process allows us to obtain five data manifolds with the least biased and uniform data distribution from the training set ($n = 5198$), where one manifold was used as a validation set ($n = 1040$) and the others were used as a sub-train set ($n = 4158$). As shown in Fig. S2(b), the 5-fold cross-validation process of these sub-datasets provides average prediction accuracy for different hyperparameter combinations, from which the best model hyperparameters can be obtained. This stratified 5-fold cross-validation approach enables reduced standard deviations of prediction performance for validation sets by more than a factor of two, compared to non-stratified 5-fold cross-validation, thus providing a better choice of model hyperparameters. After final retraining, the relative feature importance of the prediction models for the RF and XGB algorithms was calculated as the mean decrease Gini index and average of the tree values, respectively [36].

2.5. Model evaluation

Evaluation of data-driven models involves two validation processes defined as follows.

(1) Internal validation: testing the performance of prediction models using the test set ($n = 2229$) after model completion through cross-validation and model optimization.
(2) External validation: verifying the performance of prediction models with the clinical data set ($n = 120$). This process also includes predicting clinical parameters other than CPI3–4 (Table S2).

These model evaluations provide four prediction outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP and TN denote the numbers of correct predictions for the positive and negative cases, respectively. FP and FN represent the numbers of false prediction cases. Each prediction model's sensitivity, specificity, precision, and accuracy are defined based on these values. The probability of each prediction yields receiver operating characteristic (ROC) curves, with the area under the curve (AUC) as one of the most comprehensive indicators for prediction performance. Such values, called performance metrics, indicate the performance of the prediction models. The following equations show the definitions of the performance metrics used in this study [25,36]:

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

To examine the dependence of prediction performance on random sampling states in the test-train set splitting, the prediction stability of the trained model was evaluated using bootstrap resampling (1000 times) as an advanced internal validation technique, where a narrower distribution in the performance metrics indicated a stabler model [10,23]. Prediction reproducibility was estimated from the difference in performance metrics (e.g., ROC AUC) between internal and external validations. We also performed a comparative analysis between prediction modeling with a small dataset and our data-driven models. Different random sampling (five times) and hyperparameter conditions (e.g., test–train set split ratio) were tested to examine prediction reproducibility and stability in internal-external validations. The small-dataset model used clinical data ($n = 120$) for internal model training/testing and KNHANES data ($n = 7427$) for external validation.

The model development and evaluation processes were performed using Python (version 3.8.8) with Scikit-learn (version 1.3.0), Numpy (version 1.22.0), Matplotlib (version 3.3.4), and Seaborn (version 0.11.1) as the main packages.

3. Results

3.1. Correlation analysis

A correlation analysis of the selected KNHANES database ($n = 7427$) was conducted after extracting significant risk factors related to periodontitis (CPI3–4). Table 2 shows the Pearson correlation coefficient of the 16 selected risk factors ($n(X) = 16$). These factors can be broadly categorized as age, sex, smoking, drinking, disease, obesity, and oral health conditions with slightly different sub-categorical descriptions. The most substantial risk factors were age ($R = 0.298$, $p < 0.001$), smoking duration ($R = 0.257$, $p < 0.001$), and self-perceived oral health ($R = 0.209$, $p < 0.001$), where R refers to the Pearson correlation coefficient for each variable. Periodontitis also exhibited a significant correlation with sex (male: 1, female: 2) with a negative correlation coefficient ($R = -0.121$, $p < 0.001$). On the other hand, both the amount of alcohol consumption ($R = 0.032$, $p = 0.006$) and the number of walking days per week ($R = -0.022$, $p = 0.055$) have significant but relatively lower Pearson correlations with PD.

3.2. Model performance

We developed data-driven models for predicting periodontitis (CPI3–4) with different machine-learning algorithms using the risk factors obtained from the correlation analysis. Table 3 presents the performance metrics for five models for the KNHANES test set ($n = 2229$): LR, SVM, RF, XGB, and NN. While the prediction models exhibited similar performance, the SVM prediction algorithm demonstrated the best performance, with an AUC of 0.828, along with sensitivity (0.394), specificity (0.924), precision (0.646), and accuracy (0.786). The ROC curves of the data-driven models are shown in Fig. S3.

The prediction stability of the data-driven model was examined with 1000 random resamplings via bootstrapping. The bootstrap resampling result for the SVM algorithm is illustrated in Fig. S4. The histogram illustrates the distribution of each metric value, and the red dashed lines correspond to the upper and lower confidence intervals (CI) of 0.95 and the median level. The data-driven model, in particular, exhibited reasonably narrow distributions for specificity (0.911, 0.924, 0.937), accuracy (0.769, 0.787, 0.804), and AUC (0.811, 0.828, 0.845), which correspond to lower CI and median and upper CI values, respectively.

The feature importance of the prediction models was also investigated. Fig. 2(a) and (b) show the feature importance of the XGB and RF models, respectively. Both models assign age as the most prominent feature and smoking period as the second most important

Table 2
Significant periodontal risk factors extracted from the correlation analysis of the KNHANES database.

	Pearson	
Risk factors ($n = 7427$)	Coefficient	p-value
Sex	−0.121	<0.001
Age	0.298	<0.001
Self-perceived oral health	0.209	<0.001
Maxillary prosthetic condition	0.169	<0.001
Mandibular prosthetic condition	0.118	<0.001
Duration of hypertension	0.134	<0.001
Duration of dyslipidemia	0.073	<0.001
Duration of diabetes	0.094	<0.001
Drinking frequency per year	0.057	<0.001
Amount of alcohol consumption	0.032	0.006
Heavy drinking frequency	0.061	<0.001
Smoking period	0.257	<0.001
Average smoking per day	0.167	<0.001
Number of walking days per week	−0.022	0.055
Waist	0.181	<0.001
Body mass index	0.117	<0.001

Table 3
Performance metrics of the data-driven prediction models evaluated using the KNHANES database with 95 % confidence intervals.

Performance metric (n = 2229)	Prediction models				
	LR	SVM	RF	XGB	NN
Sensitivity	0.399 (0.360–0.439)	0.394 (0.359–0.435)	0.342 (0.302–0.382)	0.354 (0.314–0.395)	0.399 (0.362–0.440)
Specificity	0.907 (0.893–0.920)	0.924 (0.911–0.937)	0.932 (0.920–0.943)	0.919 (0.907–0.931)	0.910 (0.897–0.923)
Precision	0.600 (0.553–0.646)	0.646 (0.599–0.696)	0.637 (0.583–0.690)	0.607 (0.553–0.655)	0.609 (0.561–0.656)
Accuracy	0.775 (0.757–0.793)	0.786 (0.769–0.804)	0.778 (0.761–0.796)	0.773 (0.755–0.789)	0.777 (0.760–0.794)
ROC AUC	0.822 (0.804–0.840)	0.828 (0.811–0.845)	0.824 (0.806–0.841)	0.823 (0.804–0.842)	0.823 (0.805–0.840)

Abbreviations: ROC, response operation characteristic; AUC, area under the curve; LR, logistics regression; SVM, support vector machine; RF, random forest; XGB, extreme gradient boosting; NN, neural network.

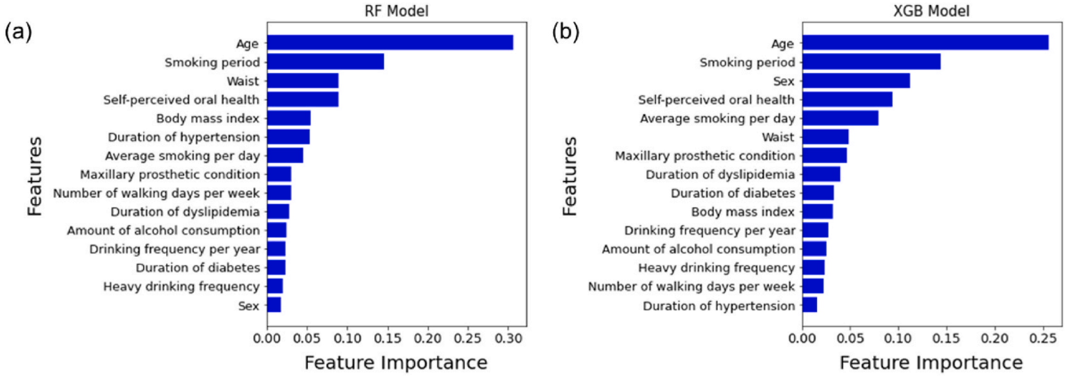

Fig. 2. Feature importance of the data-driven prediction models. (a) RF and (b) XGB models.

in predicting periodontitis (CPI3–4). Overall feature importance ranking revealed a similar trend as the correlation analysis result. Remarkably, the XGB algorithm identified sex as having feature importance (0.11) in predicting periodontitis, in contrast to the RF model (0.02).

3.3. Clinical validation

External validation with clinical data was carried out using the prediction models developed and validated with the KNHANES database. The average (± standard deviation) CPI score and gingival index of clinical data ($n = 120$) were 3.13 ± 1.21 and 1.85 ± 0.98, respectively. The ROC curves of five data-driven models for predicting periodontitis (CPI3–4) are presented in Fig. 3. The AUC values for these models were 0.780 (LR), 0.786 (SVM), 0.794 (RF), 0.796 (XGB), and 0.795 (NN). The XGB algorithm demonstrated the best prediction performance, although the differences were not large, similar to the evaluation results with the KNHANES data. Table S3 shows the performance metrics of the data-driven models for clinical evaluation. Similar to the interval validation results, the data-

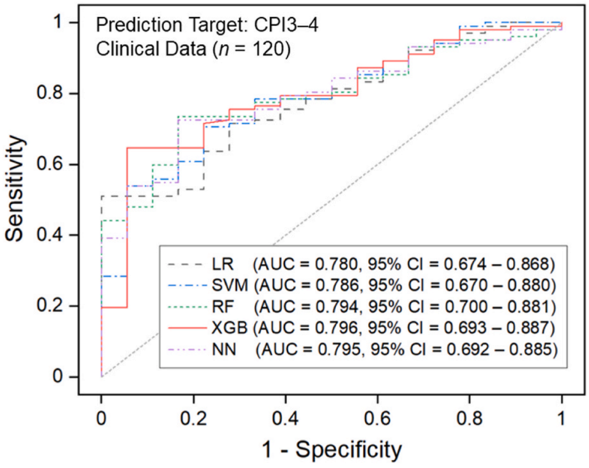

Fig. 3. ROC curves of the data-driven prediction models evaluated using GPMC clinical data.

driven models show high specificity and precision (>0.9) in clinical validations, whereas prediction sensitivity was observed to be relatively low. The data-driven prediction models with different algorithms show comparable ROC AUC values and the corresponding 95 % confidence intervals regardless of differences in other performance metrics.

3.4. *Prediction for other clinical factors*

The data-driven models were also used to predict clinical periodontal parameters other than CPI3–4. These parameters are CPI4 (CPI score = 4), Probing Depth (Mean, Individual), Bone Loss Ratio (PD, PD+), Gingival Bleeding, and Tooth Loss. Details of the descriptions and classification criteria for these parameters are provided in Table S2. These PD-related clinical parameters were predicted with the data-driven model (XGB) and the GPMC clinical patient data ($n = 120$). We also estimated 95 % lower and upper confidence intervals to identify statistical distributions in each of the prediction performance metrics.

Fig. 4 shows the ROC curves of the XGB model for predicting the eight clinical parameters. As shown in Fig. 4, the AUC values of the prediction parameters are 0.796 (a: CPI3–4), 0.736 (b: CPI4), 0.743 (c: Probing Depth (Mean)), 0.781 (d: Probing Depth (Individual)), 0.792 (e: Bone Loss Ratio (PD)), 0.813 (f: Bone Loss Ratio (PD+)), 0.694 (g: Gingival Bleeding), and 0.734 (h: Tooth Loss). Bone Loss Ratio (PD+) was predicted most accurately by the data-driven model, with the highest AUC value (0.813). These results indicate that the XGB model developed in this work allows us to predict other PD-related clinical parameters at significant levels. Fig. 4 also includes other performance metrics (sensitivity, specificity, precision, and accuracy) for the prediction of the PD-related clinical parameters.

4. Discussion

This work identified 16 significant factors from over 1000 potential risk factors through correlational feature analysis, and these features were used to collect clinical data and develop data-driven models. The study also provided detailed pipelines and workflow for building prediction models with various algorithms and evaluating them with a large-scale database (the KNHANES data) and clinical patient data (the GPMC data) for internal-external validations. Prediction stability and reproducibility of the data-driven models were studied with consideration of data distribution and collection protocols. Furthermore, our data-driven models were explored for predicting other PD-related clinical parameters and their prognostic performances.

The ROC AUC values revealed a prominent performance of the data-driven models compared with previous results. In our study, the prediction models with the five algorithms exhibited decent predictive performance for periodontitis (CPI3–4) in internal validation (AUC = 0.822 to 0.828) and external validation (AUC = 0.780 to 0.796). The literature includes several AUC values for internal validation of data-driven models: 0.60–0.86 [10], 0.69–0.72 [20], 0.702–0.712 [31], and 0.770–0.878 [33]. Our prediction models show comparable or higher AUC values than those in previous results. This can be attributed to (1) well-performing algorithms for periodontal risk assessment [20,23,37], (2) well-designed model optimization processes based on grid search and cross-validation (Fig. S2) [23], and (3) effective feature selection through correlation analysis [20]. In contrast, a data-driven model with a decision tree algorithm revealed less accurate prediction in internal validation (AUC = 0.796), as shown in Fig. S5.

The data-driven models also demonstrated remarkable prediction reproducibility in internal–external validations through a

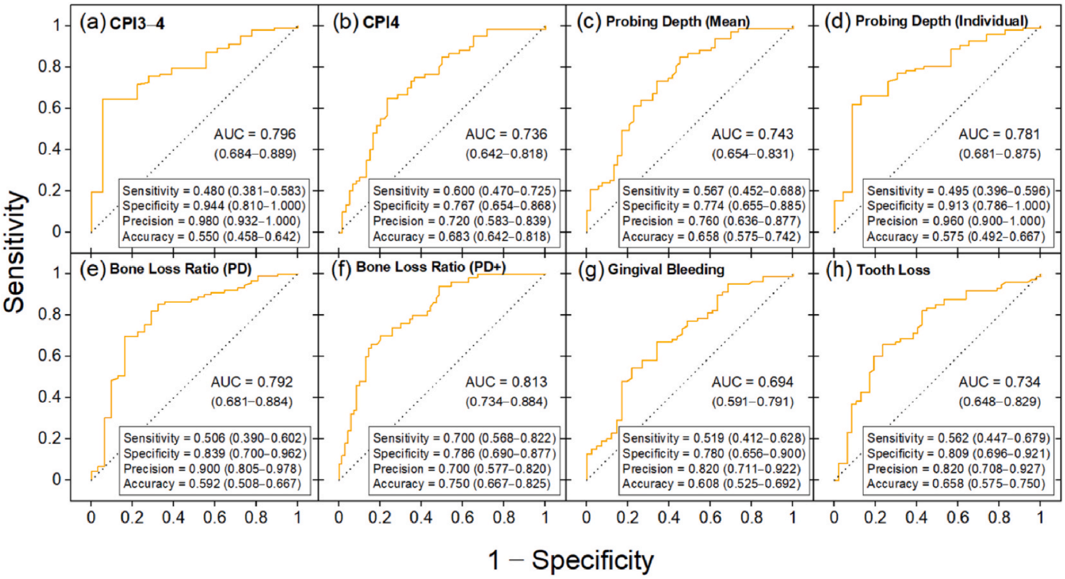

Fig. 4. Performance of the data-driven model in predicting other clinical factors (XGB model). (a) CPI3 –4, (b) CPI4, (c) probing depth (mean), (d) probing depth (individual), (e) bone loss ratio (PD), (f) bone loss ratio (PD+), (g) gingival bleeding, and (h) tooth loss. Values in the parentheses correspond to 95 % lower and upper confidence intervals.

comparative analysis. The best (or most reproducible) data-driven model (XGB) showed accurate prediction for periodontitis (CPI3–4) in internal validation (AUC = 0.823) and clinical validation (AUC = 0.796), with a difference (ΔAUC) of 0.027. For comparison, we developed a prediction model with the small clinical data set (n = 120): model training with 70 % of the clinical data (n = 84) and model evaluation with the remaining 30 % (n = 36). The KNHANES database *(n* = 7427) was used for external validation. Fig. S6 shows the ROC curves of the prediction model developed with the small clinical data set, with AUC values of internal and external validations of 0.836 and 0.649, respectively. Therefore, the data-driven prediction model demonstrated enhanced reproducibility, showing a difference (ΔAUC = 0.027) approximately seven times smaller than that of the prediction model using the small clinical data (ΔAUC = 0.187).

Prediction reproducibility and stability for these two approaches were further investigated. The performance metrics (accuracy and AUC) for internal and external validations with varying test-training set split ratios are presented in Fig. S7. These performance metrics were obtained from five random samplings for internal model training and testing, with the error bars corresponding to the standard deviations. The data-driven prediction model showed AUC and accuracy values with (1) smaller difference between internal and external validation values, (2) lower deviations within the same conditions (i.e., error bar), and (3) less difference among the test/train set ratios. These results indicate that the data-driven approach based on the KNHANES database can improve reproducibility and stability in predicting PD (CPI3–4).

The balanced and unbiased distribution of training data for the data-driven prediction models can explain the enhanced reliability. The KNHANES data were collected with stratified multistage probability sampling based on the geographical area, sex, and age of the non-institutionalized civilian population in South Korea [38]. The detailed sampling methods and guidelines are described in the KNHANES report [39]. The KNHANES data set represents the entire Korean population and is characterized by minimum selection bias and balanced data distribution [40], providing a lower risk of modeling bias [25,28]. On the other hand, there seems to be significant imbalance and deviation in the patient data despite the random data collection at the GPMC Suwon Hospital. For example, the proportion of periodontitis (CPI3–4) patients was larger in the GPMC data set (85 %) compared to the KNHANES database (26.5 %). In addition, the clinical data revealed an older age distribution of patients, with 69.2 % of the sample group older than 50 years of age, while 38.9 % of patients in the database were over 50 years. Therefore, enhanced reliability and reproducibility of the data-driven prediction model can be attributed to the balanced and less-biased data distributions by a well-designed sampling protocol in the KNHANES database.

The relationship between correlation analysis and feature importance (RF and XGB) was also scrutinized. While most features identified in this study are consistent in their significance with previous studies [18,20,33,41], several exhibited differences in factor ranking (Table S4). Such differences are likely attributable to the mechanisms by which the significance values are calculated [36]. In detail, Pearson coefficients only consider the correlation between a risk factor (e.g., sex) and an outcome (e.g., CPI3–4). However, correlations between risk factors (e.g., sex and smoking period) can be involved. For example, the correlation between sex and smoking duration was significant (Pearson correlation coefficient = −0.55), as shown in Fig. S8. This implies that the significance of a relatively weak risk factor (sex) can be more readily varied than that of a stronger risk factor (smoking duration), depending on how the model ensembles were constructed [36,39]. Therefore, highly correlated factors can induce variability in feature importance, showing the need to verify factor-to-factor correlation in feature selection processes [36].

The data-driven models evidenced the predictability of other clinical factors in external validation (n = 120). In particular, the Bone Loss Ratio (PD+) was the best predictive parameter in clinical evaluation, with an AUC of 0.813 (Fig. 4), indicating that our approach can be useful in clinical practice (i.e., ROC AUC >0.8) [10]. However, model training only considered the 16 risk factors as predictors and CPI3–4 as an outcome, excluding all information about other clinical parameters (e.g., Bone Loss Ratio, Gingival Bleeding, and Tooth Loss). To address this, we investigated the Pearson correlation between the outcome variables (Fig. S9), which were significant

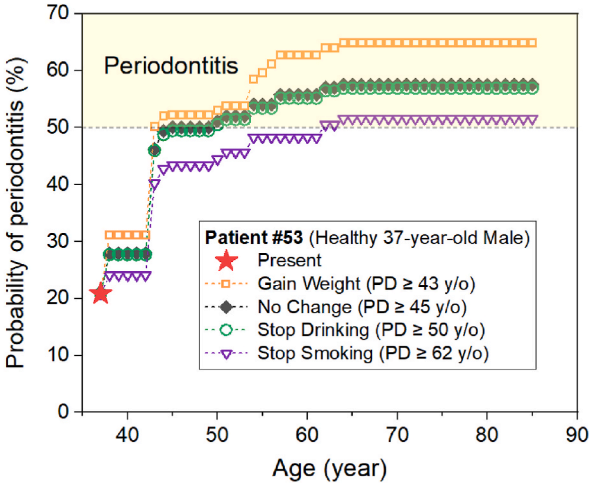

Fig. 5. Prediction of age-dependent periodontal disease probability using the data-driven model (XGB).

regardless of the types of clinical measurements. This is reasonable considering PD is one of the major causes of tooth loss, gingival inflammation, and gingival recession [42]. The Pearson correlation coefficients of tooth loss or gingival bleeding become larger in line with periodontal severity.

Our results demonstrate the utility of the XGB model in predicting the risk of PD with pronounced versatility and efficacy across different validation metrics. The XGB algorithm has been explored and successfully adopted in other biomedical fields for clinical predictions. For instance, L. Xing et al. demonstrated that the XGB model outperforms six different machine-learning algorithms in predicting lip prominence based on hard-tissue measurements and the demographic characteristics of Asian people [43]. Also, N. Hou et al. reported the best performance of the XGB model to predict the 30-day mortality for MIMIC-III patients with sepsis-3 over traditional prediction modeling approaches [44]. These results imply that the XGB algorithm would be useful in making prediction models for clinical practice.

Last, the data-driven model was explored for prognostic prediction of periodontitis (CPI3–4) to provide preventive care in clinical practice. A 37-year-old healthy participant (Participant No. 53), the youngest among the smoking and drinking participants without periodontitis, was chosen for this analysis. The data-driven prediction model provided the prognostic probability of periodontitis (CPI3–4) for this participant as a function of age (37–85 years), with different conditions for risk factors considered (Fig. 5). According to the prognostic analysis, the probability of periodontitis was 20.8 % at current and increased to greater than 50 % at 45 years. This indicates the participant would likely suffer from periodontitis after ≥8 years without preventive measures. The prediction model also revealed that weight gain (i.e., increase in BMI and waist by approximately 18 % and 16 cm, respectively) can increase the risk of periodontitis by 13.4 % on average. The probability of periodontitis can be reduced by 6.0 % (stop smoking) or 0.6 % (stop drinking). In particular, the possibility of periodontitis (probability >50 %) can be delayed by 17 years if the patient stops smoking. This prognostic information can be helpful for patient-specific prescription and treatment. Therefore, prognostic prediction of data-driven models can be a useful clinical application for individual patient care for PD.

One possible application of the PD prediction model to clinical practice is screening high-risk PD patients without restrictions on time and location via mobile or online websites. Such prediction models enable one to readily evaluate the risk level of PD by filling out survey forms, thus alerting them to get urgent or periodic dental check-ups if necessary. High-specificity screening is useful and cost-effective in early-stage PD detection for the general public because most people with low PD risks will be diagnosed negatively and thus not required to get unnecessary clinical treatments [45,46]. Therefore, our data-driven PD prediction model with high specificity and decent ROC AUC is desirable for this purpose.

Yet, there are several potential challenges to be tackled for practical applications. Firstly, the sensitivity of the prediction model needs to be enhanced. This can be addressed by improving the data-driven prediction model employing refined databases. Several risk factors were excluded in this work due to the risk of diminishing data volume (missing predictor variables), factor uniqueness, modeling complexity, and computational costs despite non-negligible correlations with PD. These factors include household income, educational level, nutritional intake (e.g., calcium and vitamin), stress, and other diseases (e.g., rheumatoid arthritis, osteoporosis, and allergy), which are known to be prospective risk factors for PD [18,41,47]. Better prediction performance and higher sensitivity are expected if the modeling database is thorough enough to include such variables. More accurate PD prediction would also be possible via more appropriate verification of risk factors if up-to-date databases (e.g., electronic dental records) are used for model training [20]. In addition, the present study demonstrates the predictive performance of the data-driven model based on clinical validation of 120 patients, limiting in confirming the general utility and applicability of the model. Thus, further validation studies with larger populations are necessary. Lastly, it is essential to validate the data-driven model based on diverse and comprehensive populations with different racial, cultural, regional, and sociodemographic environments. Future studies based on data-driven modeling approaches are anticipated with improved experimental designs and high-quality databases, not only for PD research but also for predictions of general biomedical diseases.

5. Conclusion

The present work demonstrated the data-driven approach to predict PD and related factors. The statistical correlation analysis was performed using the KNHANES database to identify significant risk factors for PD, based on which clinical data were obtained. The data-driven method exhibited better performance and reproducibility compared to the modeling with a small dataset, as demonstrated by internal and external validations. In addition, the data-driven approach could predict other clinical parameters such as severe bone loss, gingival bleeding, and tooth loss. The prognosis prediction of periodontitis was further investigated using the prediction model to provide patient-specific preventive care. These results imply that the data-driven prediction approach can be useful in clinical practice.

Ethics statement

The study protocol including clinical data collection was approved by the Institutional Review Board for Human Subjects of Ajou University Dental Hospital (IRB No. AJOUIRB-SB-2023-008). The written informed consent was obtained from all participants individually. All processes were conducted in accordance with the principles of the Declaration of Helsinki.

Data availability statement

The Korea National Health and Nutrition Examination Survey data are available at http://knhanes.kdca.go.kr/.

CRediT authorship contribution statement

Woosun Beak: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jihun Park:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Suk Ji:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C2004811) and a grant of 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 number: HI23C0409).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e32496.

References

[1] World Health Organization, Global Oral Health Status Report: towards Universal Health Coverage for Oral Health by 2030, World Health Organization, Geneva, 2022 (ISBN: 978-92-4-006148-4).

[2] L.J.A. Heitz-Mayfield, Disease progression: identification of high-risk groups and individuals for periodontitis, J. Clin. Periodontol. 32 (Suppl. 6) (2005) 196–209, https://doi.org/10.1111/j.1600-051X.2005.00803x.

[3] P.I. Eke, B.A. Dye, L. Wei, G.O. Thornton-Evans, R.J. Genco, Prevalence of periodontitis in adults in the United States: 2009 and 2010, J. Dent. Res. 91 (2012) 914–920, https://doi.org/10.1177/0022034512457373.

[4] G.C. Armitage, Periodontal diagnoses and classification of periodontal diseases, Periodontol 34 (2000) 9–21, https://doi.org/10.1046/j.0906-6713.2002.003421x, 2004.

[5] G.C. Armitage, Diagnosis of periodontal diseases, J. Periodontol. 74 (8) (2003) 1237–1247, https://doi.org/10.1902/jop.2003.74.8.1237.

[6] N. Rathnayake, S. Akerman, N. Lundegren, H. Jansson, Y. Rryselius, Salivary biomarkers of oral health – a cross-sectional study, J. Clin. Periodontol. 40 (2013) 140–147, https://doi.org/10.1111/jcpe.12038.

[7] D.F. Kinane, P.G. Stathopoulou, P.N. Papapanou, Periodontal diseases, Nat. Rev. Dis. Prim. 3 (2017) 17038, https://doi.org/10.1038/nrdp.2017.38.

[8] World Health Organization, Oral Health Surveys: Basic Methods, fourth ed., World Health Organization, Geneva, 1997 (ISBN: 92-4-154493-7).

[9] P.I. Eke, B.A. Dye, L. Wei, G.D. Slade, G.O. Thornton-Evans, J.D. Beck, G.W. Taylor, W.S. Borgnakke, R.C. Page, R.J. Genco, Self-reported measures for surveillance of periodontitis, J. Dent. Res. 92 (11) (2013) 1041–1047, https://doi.org/10.1177/0022034513505621.

[10] T. Dietrich, U. Stosch, D. Dietrich, W. Kaiser, J.-P. Bernimoulin, K. Joshipura, Prediction of periodontal disease from multiple self-reported items in a German practice-based sample, J. Periodontol. 78 (75) (2007) 1421–1428, https://doi.org/10.1902/jop.2007.060212.

[11] Y.-J. Maeng, B.-R. Kim, H.-I. Jung, U.-W. Jung, H.E. Kim, B.-I. Kim, Diagnostic accuracy of a combination of salivary hemoglobin levels, self-report questionnaires, and age in periodontitis screening, J. Periodontal Implant Sci. 46 (1) (2016) 10–21, https://doi.org/10.5051/jpis.2016.46.1.10.

[12] G.S. Chatzopoulos, L. Tsalikis, A. Konstantinidis, G.A. Kotsakis, A two-domain self-report measure of periodontal disease has good accuracy for periodontitis screening in dental school outpatients, J. Periodontol. 87 (10) (2016) 1165–1173, https://doi.org/10.1902/jop.2016.160043.

[13] M. Kuboniwa, A. Sakanaka, E. Hashino, T. Bamba, E. Fukusaki, A. Amano, Prediction of periodontal inflammation via metabolic profiling of saliva, J. Dent. Res. 95 (12) (2016) 1381–1386, https://doi.org/10.1177/0022034516661142.

[14] F.R.M. Leite, K.G. Peres, L.G. Do, F.F. Demarco, M.A.A. Peres, Prediction of periodontitis occurrence: influence of classification and sociodemographic and general health information, J. Periodontol. 88 (8) (2017) 731–743, https://doi.org/10.1902/jop.2017.160607.

[15] M. Feres, Y. Louzoun, S. Haber, M. Faveri, L.C. Figueiredo, L. Levin, Support vector machine-based differentiation between aggressive and chronic periodontitis using microbial profiles, Int. Dent. J. 68 (2018) 39–46, https://doi.org/10.1111/idj.12326.

[16] S.-H. Nam, H.-I. Jung, S.-M. Kang, D. Inaba, H.-K. Kwon, B.-I. Kim, Validity of screening methods for periodontitis using salivary hemoglobin level and self-report questionnaires in people with disabilities, J. Periodontol. 86 (4) (2015) 536–545, https://doi.org/10.1902/jop.2015.140457.

[17] N. Shimpi, S. McRoy, H. Zhao, M. Wu, A. Acharya, Development of a periodontitis risk assessment model for primary care providers in an interdisciplinary setting, Technol. Health Care 28 (2019) 143–154, https://doi.org/10.3233/THC-191642.

[18] R.J. Genco, W.S. Borgnakke, Risk factors for periodontal disease, Periodontol 62 (2013) 59–94, https://doi.org/10.1111/j.1600-0757.2012.00457x, 2000.

[19] R.I. Garcia, R. Compton, T. Dietrich, Risk assessment and periodontal prevention in primary care, Periodontol 200071 (2016) 10–21, https://doi.org/10.1111/prd.12124.

[20] J.S. Patel, C. Su, M. Tellez, J.M. Albandar, R. Rao, V. Iyer, E. Shi, H. Wu, Developing and testing a prediction model for periodontal disease using machine learning and big electronic dental record data, Front. Artif. Intell. 5 (2022) 979525, https://doi.org/10.3389/frai.2022.979525.

[21] F. Schwendicke, T. Singh, J.-H. Lee, R. Gaudin, A. Chaurasia, T. Wiegand, S. Uribe, J. Krois, The IADR e-oral health network, the ITU WHO focus group AI for Health, Artificial intelligence in dental research: checklist for authors, reviewers, readers, J. Dent. 107 (2021) 103610, https://doi.org/10.1016/j.jdent.2021.103610.

[22] G. Battineni, G.G. Sagaro, N. Chinatalapudi, F. Amenta, Applications of machine learning predictive models in the chronic disease diagnosis, J. Personalized Med. 10 (2) (2020) 21, https://doi.org/10.3390/jpm10020021.

[23] M. Du, D. Haag, Y. Song, J. Lynch, M. Mittinty, Examining bias and reporting in oral health prediction modeling studies, J. Dent. Res. 99 (4) (2020) 374–387, https://doi.org/10.1177/0022034520903725.

[24] K.J. Rothman, S. Greenland, T.L. Lash, Modern Epidemiology, third ed., Lippincott Wilkins & Williams, Philadelphia (PA), 2008 (ISBN-10: 0-7817-5564-6).

[25] F. Pethani, Promises and perils of artificial intelligence in dentistry, Aust. Dent. J. 66 (2021) 124–135, https://doi.org/10.1111/adj.12812.

[26] D. Rajput, W.-J. Wang, C.-C. Chen, Evaluation of a decided sample size in machine learning applications, BMC Bioinf. 24 (2023) 48, https://doi.org/10.1186/s12859-023-05156-9.

[27] R.D. Riley, G.S. Collins, Stability of clinical prediction models developed using statistical or machine learning methods, Biom. J. 00 (2023) e2200302, https://doi.org/10.1002/bimj.202200302.

[28] R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett, PROBAST Group, PROBAST: a tool to assess the risk of bias and applicability of prediction model studies, Ann. Intern. Med. 170 (1) (2019) 51–58, https://doi.org/10.7326/M18-1376.

[29] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G.M. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, BMC Med. 13 (2015) 1–10, https://doi.org/10.1186/s12916-014-0241-z.

[30] P.I. Eke, X. Zhang, H. Lu, L. Wei, G. Thornton-Evans, K.J. Greenlund, J.B. Holt, J.B. Croft, Predicting periodontitis at state and local levels in the United States, J. Dent. Res. 95 (5) (2016) 515–522, https://doi.org/10.1177/0022034516629112.

[31] H. Lai, C.-W. Su, A.M.-F. Yen, S.Y.-H. Chiu, J.C.-Y. Fann, W.Y.-Y. Wu, S.-L. Chuang, H.-C. Liu, H.-H. Chen, L.-S. Chen, A prediction model for periodontal disease: modelling and validation from a National Survey of 4061 Taiwanese adults, J. Clin. Periodontol. 42 (2015) 413–421, https://doi.org/10.1111/jcpe.12389.

[32] Y.-C. Wu, L. Ning, Y.-K. Tu, C.-P. Huang, N.-T. Huang, Y.-F. Chen, P.-C. Chang, Salivary biomarker combination prediction model for the diagnosis of periodontitis in a Taiwanese population, J. Formos. Med. Assoc. 117 (9) (2018) 841–848, https://doi.org/10.1016/j.jfma.2017.10.004.

[33] J.-H. Lee, S.-N. Jeong, S.-H. Choi, Predictive data mining for diagnosing periodontal disease: the Korea national health and nutrition examination surveys (KNHANES V and VI) from 2010 to 2015, J. Publ. Health Dent. 79 (2019) 44–52, https://doi.org/10.1111/jphd.12293.

[34] V.P. Kearney, A.-I.M. Yansane, R.G. Brandon, R. Vaderhobli, G.-H. Lin, H. Hekmatian, W. Deng, N. Joshi, H. Bhandari, A.S. Sadat, J.M. White, A generative adversarial inpainting network to enhance prediction of periodontal clinical attachment level, J. Dent. 123 (2022) 104211, https://doi.org/10.1016/j.jdent.2022.104211.

[35] J.S. Patel, R. Brandon, M. Tellez, J.M. Albandar, R. Rao, J. Krois, H. Wu, Developing automated computer algorithms to phenotype periodontal disease diagnoses in electronic dental records, Methods Inf. Med. 61 (2022) e125–e133, https://doi.org/10.1055/s-0042-1757880.

[36] Y. Qu, Z. Lin, Z. Yang, H. Lin, X. Huang, L. Gu, Machine learning models for prognosis prediction in endodontic microsurgery, J. Dent. 118 (2022) 103947, https://doi.org/10.1016/j.jdent.2022.103947.

[37] G. Troiano, L. Nibali, H. Petsos, P. Eickholz, M.H.A. Saleh, P. Santamaria, J. Jian, S. Shi, H. Meng, K. Zhurakivska, H.-L. Wang, A. Ravidà, Development and international validation of logistic regression and machine-learning models for the prediction of 10-year molar loss, J. Clin. Periodontol. 50 (2023) 348–357, https://doi.org/10.1111/jcpe.13739.

[38] J.-B. Lee, H.-Y. Yi, K.-H. Bae, The association between periodontitis and dyslipidemia based on the fourth Korea national health and nutrition examination survey, J. Clin. Periodontol. 40 (2013) 437–442, https://doi.org/10.1111/jcpe.12095.

[39] Korea Disease Control and Prevention Agency, Korean national health and nutrition examination survey: the 7th survey. https://knhanes.kdca.go.kr/, 2018. (Accessed 21 September 2022).

[40] J.-R. Jeong, Y.-R. Choe, Health-promoting behaviors among middle-aged breast cancer survivors compared with matched non-cancer controls: a KNHANES VI-VII (2013–2018) study, Medicine (Baltim.) 102 (26) (2023) e34065, https://doi.org/10.1097/MD.0000000000034065.

[41] M. Du, T. Bo, K. Kapellas, M. A Peres, Prediction models for the incidence and progression of periodontitis: a systematic review, J. Clin. Periodontol. 45 (2018) 1408–1420, https://doi.org/10.1111/jcpe.13037.

[42] P. Meisel, B. Holtfreter, H. Völzke, T. Kocher, Self-reported oral health predicts tooth loss after five and ten years in a population-based study, J. Clin. Periodontol. 45 (10) (2018) 1164–1172, https://doi.org/10.1111/jcpe.12997.

[43] L. Xing, X. Zhang, Y. Guo, D. Bai, H. Xu, XGBoost-aided prediction of lip prominence based on hard-tissue measurements and demographic characteristics in an Asian population, AJODO 164 (3) (2023) 357–367, https://doi.org/10.1016/j.ajodo.2023.01.017.

[44] N. Hou, M. Li, L. He, B. Xie, L. Wang, R. Zhang, Y. Yu, X. Sun, Z. Pan, K. Wang, Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost, J. Transl. Med. 18 (2020) 462, https://doi.org/10.1186/s12967-020-02620-5.

[45] S.H.R. Cheong, Y.J.X. Ng, Y. Lau, S.T. Lau, Wearable technology for early detection of COVID-19: a systematic scoping review, Prev. Med. 162 (2022) 107170, https://doi.org/10.1016/j.ypmed.2022.107170.

[46] N. Castellanos-Ryan, M. O'Leary-Barrett, L. Sully, P. Conrod, Sensitivity and specificity of a brief personality screening instrument in predicting future substance use, emotional, and behavioral problems: 18-month predictive validity of the substance use risk profile scale, Alcohol Clin. Exp. Res. 37 (S1) (2013) E281–E290, https://doi.org/10.1111/j.1530-0277.2012.01931x.

[47] D. Bourgeois, C. Inquimbert, L. Ottolenghi, F. Carrouel, Periodontal pathogens as risk factors of cardiovascular diseases, diabetes, rheumatoid arthritis, cancer, and chronic obstructive pulmonary disease—is there cause for consideration? Microorganisms 7 (10) (2019) 424, https://doi.org/10.3390/microorganisms7100424.