

Construction and Validation of a Risk Prediction Model for Malocclusion in Children: Based on Multifactorial Analysis

Xiaohua Zhang^{1,a}, Fan Zhang^{1,b}, Juncai Lin^{1,c}, Caiyin Qiu^{1,d}, Xiaochuan Zeng^{2,e,*}

¹Longyan First Affiliated Hospital of Fujian Medical University, Longyan 364000, Fujian Province, China

²Department of Stomatology, The Second Hospital of Longyan, Longyan 364000, Fujian Province, China

a.zxhfeiyang@163.com; b.zhangfantn@163.com; c.13645035528@163.com; d.18508459803@163.com; e.huadentist@163.com; *Corresponding author: Xiaochuan Zeng; Email: huadentist@163.com

CITATION

Xiaohua Z, Fan Z. et al. Construction and Validation of a Risk Prediction Model for Malocclusion in Children: Based on Multifactorial Analysis. *Psycho-Oncologie*.

ARTICLE INFO

Received:

Accepted:

Available online: September 29, 2025

COPYRIGHT



Copyright © 2025 by author(s).

Psycho-Oncologie is published by PsychoMed Publishing Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license.

<https://creativecommons.org/licenses/by/4.0/>

Abstract: Malocclusion, as one of the most common oral diseases in childhood, significantly affects children's oral health and quality of life. However, effective early screening and risk assessment tools are currently lacking. This study aimed to construct and validate a multifactorial analysis-based risk prediction model for childhood malocclusion. A cross-sectional study enrolled 1,140 children aged 6-12 years from five elementary schools between January 2022 and June 2024. Participants were randomly allocated to training (n=798) and validation (n=342) sets in a 7:3 ratio. Multidimensional information including demographic characteristics, oral bad habits, family history, early dental developmental abnormalities, and feeding history was collected through structured questionnaires and standardized oral examinations. Multivariable logistic regression analysis was used to identify independent risk factors and construct a nomogram prediction model. The study found an overall prevalence of malocclusion of 42.5%. Multivariable analysis identified seven independent predictive factors: current digit sucking habit (OR=3.42, 95%CI: 2.08-5.63), habitual mouth breathing (OR=2.86, 95%CI: 1.89-4.33), tongue thrusting (OR=2.24, 95%CI: 1.52-3.31), parental history of malocclusion in both parents (OR=2.68, 95%CI: 1.64-4.38), premature loss of primary teeth (OR=1.92, 95%CI: 1.36-2.71), breastfeeding duration <6 months (OR=1.78, 95%CI: 1.28-2.47), and bottle feeding >24 months (OR=1.65, 95%CI: 1.18-2.31). The prediction model achieved AUC values of 0.786 and 0.772 in training and validation sets. The Hosmer-Lemeshow test showed good calibration (P>0.05), and decision curve analysis confirmed positive net benefit (threshold probability 0.15-0.75). The developed nomogram provides practical instruments for early identification and risk stratification of childhood malocclusion.

Keywords: Childhood malocclusion; Risk prediction model; Nomogram; Oral bad habits; Multifactorial analysis

1. Introduction

Malocclusion is one of the most common oral diseases in childhood, severely affecting children's oral health and quality of life. Systematic reviews show

that the global prevalence of malocclusion varies significantly across different stages of dental development ^[1]. Recent meta-analyses further confirm that the incidence of primary dentition malocclusion in preschool children is on an upward trend, highlighting the urgency of early prevention and intervention ^[2]. In China, there are substantial differences in the prevalence of malocclusion among children from different regions and ethnic groups. A survey of 7-8 year-old children from the Zhuang ethnic group in southern China revealed that environmental factors and lifestyle habits have significant impacts on the occurrence of malocclusion ^[3].

Studies on the global distribution characteristics of malocclusion indicate that different types of malocclusion show heterogeneity in their distribution across continents and populations, reflecting the complex interactions between genetic background, environmental factors, and socioeconomic conditions ^[4]. Given the multiple impacts of malocclusion on children's maxillofacial development, masticatory function, language development, and psychological health, international expert consensus emphasizes the importance of orthodontic treatment during childhood and proposes staged, individualized treatment strategies ^[5]. However, large-scale epidemiological studies have found that a considerable proportion of malocclusion in healthy children and adolescents remains undiagnosed and untreated, which is closely related to the lack of effective early screening and risk assessment tools ^[6].

The relationship between oral bad habits and malocclusion has been a research hotspot. Survey studies of 5-7 year-old children have found significant associations between malocclusion and bad habits such as mouth breathing, thumb sucking, and lip biting, with the accumulation of multiple bad habits increasing the severity of malocclusion ^[7]. These findings provide an important epidemiological foundation for establishing risk factor-based prediction models.

With the development of precision medicine concepts, the application of clinical prediction models in disease risk assessment is becoming increasingly widespread. Decision curve analysis, as a novel model evaluation method, can comprehensively consider both the predictive accuracy and clinical utility of models, providing a more comprehensive assessment framework for clinical application ^[8]. In recent years, reporting standards for clinical prediction models have been continuously improved. The TRIPOD+AI statement provides standardized reporting guidelines for prediction models constructed using both traditional statistical methods and machine learning approaches, ensuring research transparency and reproducibility ^[9].

The application of artificial intelligence technology in orthodontics shows tremendous potential. Systematic reviews indicate that artificial intelligence has made significant progress in orthodontic diagnosis, treatment planning, and clinical decision support, particularly demonstrating superior performance to traditional methods in image analysis and treatment outcome prediction ^[10]. The advantages of machine learning algorithms in handling complex multidimensional

data, identifying nonlinear relationships, and discovering hidden patterns provide new technical pathways for constructing more accurate malocclusion risk prediction models^[11].

Despite important progress in understanding the epidemiological characteristics and risk factor identification of malocclusion, significant limitations remain. Existing studies are mostly cross-sectional surveys, lacking research that integrates multiple risk factors into practical prediction tools; most studies focus only on the role of single risk factors, neglecting interaction effects among multiple factors; there is a lack of rigorously validated risk prediction models with good clinical applicability; current risk assessment mainly relies on clinicians' subjective judgment, lacking objective, standardized assessment tools. These limitations restrict the early identification and preventive intervention effectiveness for malocclusion.

This study aims to construct and validate a multifactor-based risk prediction model for childhood malocclusion. Through systematic collection of multidimensional information including children's demographic characteristics, oral bad habits, and family history, the study will use multifactor logistic regression analysis to identify independent risk factors, construct a nomogram prediction model, and evaluate the model's predictive performance through internal and external validation. The study will also develop a web-based clinical application tool to provide convenient screening tools for primary healthcare institutions and schools. The establishment of this model will facilitate early identification and risk stratification of malocclusion, provide scientific evidence for developing individualized prevention strategies, ultimately reduce the incidence of childhood malocclusion, and improve children's oral health status.

2. Materials and Methods

2.1 Study Design and Participants

This study employed a cross-sectional design with internal and external validation to evaluate model performance, and was designed and implemented following the reporting standards of the TRIPOD statement^[12]. The research protocol was developed with full consideration of the methodological requirements for prediction model development, including key elements such as selection of candidate predictive variables, sample size calculation, model construction strategies, and validation methods^[13].

The study was conducted from January 2022 to June 2024 in five elementary schools in a city, using cluster sampling method to select study participants. Inclusion criteria included: enrolled students aged 6-12 years; obtaining consent from the children themselves and informed consent from parents; ability to cooperate with oral examination and questionnaire survey; absence of severe systemic diseases. Exclusion criteria included: previous orthodontic treatment; presence of

craniofacial developmental abnormalities such as cleft lip and palate; history of maxillofacial trauma affecting occlusal relationships; presence of systemic diseases affecting jaw development; incomplete clinical data.

According to the sample size calculation principles for logistic regression models, considering the expected incidence rate of malocclusion of approximately 40% and the planned inclusion of 10 candidate predictive variables, following the empirical rule of requiring at least 10 events per variable, the minimum sample size should be 250 cases^[14]. To ensure model stability and generalizability, and considering a 15% loss to follow-up rate, the study planned to enroll 1,200 children. After screening and exclusion, 1,140 eligible participants were finally enrolled and randomly allocated to training (n=798) and validation (n=342) sets in a 7:3 ratio. This sample size design meets the statistical requirements for multivariable prediction model development^[15]. The study flow is shown in **Figure 1**.

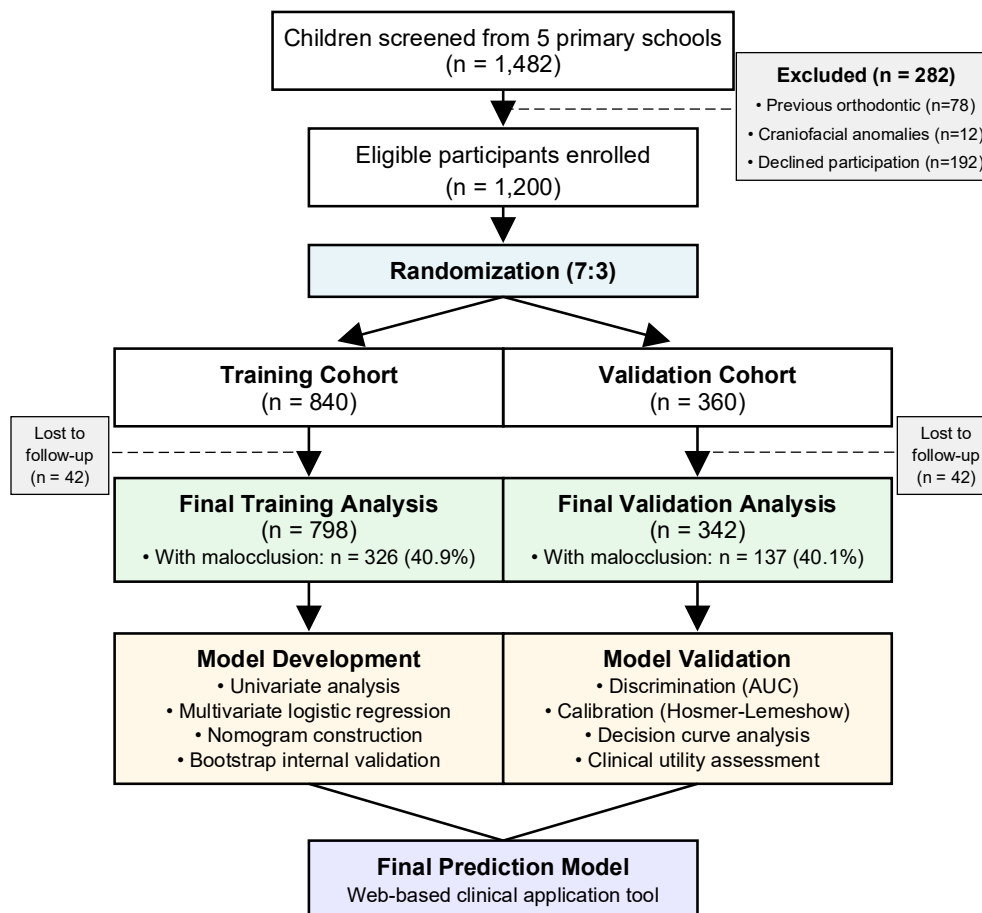


Figure 1. Study flowchart of participant enrollment and model development

2.2 Diagnostic Criteria and Data Collection

The diagnosis of malocclusion was based on the basic methods for oral health surveys recommended by the World Health Organization^[16], combined with the Angle classification system for evaluation^[17]. All oral examinations were performed by two orthodontic specialists who had undergone standardized

training. Consistency testing was conducted before examination, with Kappa values reaching above 0.85. Examinations were conducted under natural light using standardized tools including plane mirrors, probes, and millimeter rulers.

The selection of candidate predictive variables was based on literature review and clinical experience, covering multiple dimensions including demographic characteristics, oral bad habits, family history, and early dental developmental abnormalities. Data collection combined structured questionnaires with clinical examinations. Questionnaires were completed through face-to-face interviews by trained investigators. For younger children, questionnaires were completed with parental assistance. **Table 1** provides detailed definitions and measurement methods for all candidate predictive variables.

Table 1. Definition and measurement of candidate predictive variables

Variable	Definition	Measurement Method	Categories/Units
Demographic Characteristics			
Age	Chronological age at examination	Date of birth to examination date	Years (continuous)
Sex	Biological sex	Parent-reported	Male/Female
BMI	Body mass index	Weight(kg)/Height(m) ²	kg/m ² (continuous)
Oral Habits			
Thumb sucking	Habitual thumb or finger sucking	Parent questionnaire + clinical observation	Never/Past (stopped >6 months)/Current
Mouth breathing	Predominant mouth breathing pattern	Clinical assessment + parent report	No/Occasional/Habitual
Tongue thrusting	Anterior tongue positioning during swallowing	Clinical examination during water swallow test	Absent/Present
Lip biting	Habitual lip biting or sucking	Clinical signs + parent report	Absent/Present
Nail biting	Chronic nail biting habit	Clinical examination of nails + parent report	Absent/Present
Family History			
Parental malocclusion	History of malocclusion in biological parents	Parent self-report + photographs	Neither/One parent/Both parents
Sibling malocclusion	Malocclusion in biological siblings	Parent report + clinical records if available	No/Yes/No siblings
Early Dental Development			
Premature tooth loss	Early loss of primary teeth (>6 months before normal)	Clinical examination + dental history	No/Yes (number of teeth)
Delayed tooth eruption	Eruption delay >6 months from normal range	Clinical examination + radiographs if indicated	No/Yes
Supernumerary teeth	Presence of extra teeth	Clinical + radiographic examination	Absent/Present
Feeding History			
Breastfeeding duration	Exclusive breastfeeding period	Parent recall	Months (continuous)
Bottle feeding >24 months	Extended bottle use beyond 2 years	Parent report	No/Yes
Pacifier use	History and duration of pacifier use	Parent report	Never/<12 months/>12 months

Data quality control measures included: investigator training and assessment; independent double data entry; logical and range checks; random re-examination of 10% of samples with consistency requirements reaching above 95%.

2.3 Statistical Analysis Methods

Statistical analysis was performed using R software version 4.3.0. Continu-

ous variables were described as mean $\pm$ standard deviation or median (interquartile range), and categorical variables were described as frequency (percentage). Between-group comparisons used t-tests, Mann-Whitney U tests, chi-square tests, or Fisher's exact tests.

Univariable analysis screened potential predictive factors associated with malocclusion, with $P < 0.1$ as the criterion for entering multivariable analysis. Multivariable logistic regression employed a backward variable selection strategy, retaining variables with $P < 0.05$. Model multicollinearity was assessed through variance inflation factors (VIF), with $VIF > 5$ indicating collinearity problems. Harrell's C-statistic was used to evaluate model discrimination ^[18].

2.4 Model Construction and Validation

A nomogram prediction model was constructed based on multivariable logistic regression analysis results, converting regression coefficients into a 0-100 point scoring system for convenient clinical application. Bootstrap resampling method (1000 iterations) was used for internal validation to assess the degree of overfitting and model stability ^[19].

Model performance evaluation followed the latest clinical prediction model evaluation guidelines ^[20]. Discrimination was assessed through receiver operating characteristic (ROC) curves and area under the curve (AUC), with 95% confidence intervals for AUC calculated using the DeLong method ^[21]. The optimal cutoff value based on ROC curves was determined using the Youden index, with concurrent reporting of sensitivity, specificity, positive predictive value, and negative predictive value at that cutoff point ^[22].

Calibration assessment employed the Hosmer-Lemeshow goodness-of-fit test and calibration plots, evaluating model calibration performance by comparing the consistency between predicted probabilities and actual observed probabilities. Calibration plots divided samples into 10 groups based on predicted probabilities, comparing predicted risk with actual risk for each group ^[23].

Clinical utility was assessed through decision curve analysis (DCA), calculating net benefit at different threshold probabilities and comparing with "treat all" and "treat none" strategies. DCA can comprehensively consider the clinical consequences of true positives and false positives, providing quantitative assessment of the model's clinical application value ^[24].

The model underwent external validation in an independent validation cohort using the same performance evaluation metrics. If the AUC difference between the validation and training sets was less than 0.05 and calibration was good (Hosmer-Lemeshow test $P > 0.05$), the model was considered to have good generalizability. All statistical tests were two-sided, with $P < 0.05$ considered statistically significant.

3. Results

3.1 Basic Characteristics of Study Population

The study screened 1,482 children and finally enrolled 1,140 eligible participants, including 798 in the training set and 342 in the validation set. In the training set, there were 423 boys (53.0%) and 375 girls (47.0%), with a mean age of 8.7 ± 1.8 years; in the validation set, there were 178 boys (52.0%) and 164 girls (48.0%), with a mean age of 8.6 ± 1.9 years. The prevalence of malocclusion in the training and validation sets was 42.9% (342/798) and 41.5% (142/342), respectively, with no statistically significant difference between the two groups ($P=0.668$).

The baseline characteristics and univariable analysis results of study participants are shown in **Table 2**. Univariable analysis showed that factors such as mouth breathing, digit sucking habit, tongue thrusting, lip biting, parental history of malocclusion, premature loss of primary teeth, breastfeeding duration, bottle feeding >24 months, and pacifier use were significantly associated with the occurrence of malocclusion ($P<0.05$).

Table 2. Baseline characteristics and univariate analysis of risk factors for malocclusion

Variable	Training Cohort (n=798)			Validation Cohort (n=342)	
	No Malocclusion (n=456)	Malocclusion (n=342)	P value	No Malocclusion (n=200)	Malocclusion (n=142)
Demographic Characteristics					
Age (years), mean $\pm$ SD	8.6 $\pm$ 1.7	8.8 $\pm$ 1.9	0.124	8.5 $\pm$ 1.8	8.7 $\pm$ 2.0
Sex, n (%)			0.386		
Male	248 (54.4)	175 (51.2)		108 (54.0)	70 (49.3)
Female	208 (45.6)	167 (48.8)		92 (46.0)	72 (50.7)
BMI (kg/m ²), mean $\pm$ SD	17.2 $\pm$ 2.8	17.5 $\pm$ 3.1	0.156	17.1 $\pm$ 2.9	17.4 $\pm$ 3.0
Oral Habits, n (%)					
Thumb sucking			<0.001		
Never	378 (82.9)	218 (63.7)		168 (84.0)	92 (64.8)
Past	52 (11.4)	68 (19.9)		22 (11.0)	28 (19.7)
Current	26 (5.7)	56 (16.4)		10 (5.0)	22 (15.5)
Mouth breathing			<0.001		
No	356 (78.1)	186 (54.4)		158 (79.0)	78 (54.9)
Occasional	72 (15.8)	98 (28.7)		32 (16.0)	42 (29.6)
Habitual	28 (6.1)	58 (17.0)		10 (5.0)	22 (15.5)
Tongue thrusting	42 (9.2)	86 (25.1)	<0.001	18 (9.0)	35 (24.6)
Lip biting	68 (14.9)	94 (27.5)	<0.001	28 (14.0)	38 (26.8)
Nail biting	98 (21.5)	82 (24.0)	0.408	44 (22.0)	36 (25.4)
Family History, n (%)					
Parental malocclusion			<0.001		
Neither	326 (71.5)	168 (49.1)		144 (72.0)	72 (50.7)
One parent	102 (22.4)	126 (36.8)		44 (22.0)	52 (36.6)
Both parents	28 (6.1)	48 (14.0)		12 (6.0)	18 (12.7)
Sibling malocclusion	86 (18.9)	108 (31.6)	<0.001	38 (19.0)	44 (31.0)
Early Dental Development, n (%)					
Premature tooth loss	72 (15.8)	98 (28.7)	<0.001	30 (15.0)	40 (28.2)
Delayed tooth eruption	38 (8.3)	42 (12.3)	0.068	16 (8.0)	18 (12.7)
Supernumerary teeth	12 (2.6)	14 (4.1)	0.252	6 (3.0)	6 (4.2)
Feeding History					

Breastfeeding (months), median (IQR)	8 (6-12)	6 (3-9)	<0.001	8 (6-12)	6 (3-9)
Bottle feeding >24 months, n (%)	96 (21.1)	118 (34.5)	<0.001	42 (21.0)	48 (33.8)
Pacifier use, n (%)			0.002		
Never	298 (65.4)	186 (54.4)		132 (66.0)	78 (54.9)
<12 months	108 (23.7)	92 (26.9)		46 (23.0)	38 (26.8)
>12 months	50 (11.0)	64 (18.7)		22 (11.0)	26 (18.3)

3.2 Risk Factor Analysis for Malocclusion

Variables with $P < 0.1$ in univariable analysis were included in multivariable logistic regression analysis. The results showed that current digit sucking habit (OR=3.42, 95%CI: 2.08-5.63), habitual mouth breathing (OR=2.86, 95%CI: 1.89-4.33), tongue thrusting (OR=2.24, 95%CI: 1.52-3.31), parental history of malocclusion in both parents (OR=2.68, 95%CI: 1.64-4.38), premature loss of primary teeth (OR=1.92, 95%CI: 1.36-2.71), breastfeeding duration <6 months (OR=1.78, 95%CI: 1.28-2.47), and bottle feeding >24 months (OR=1.65, 95%CI: 1.18-2.31) were independent risk factors for childhood malocclusion (**Table 3**).

Table 3. Multivariate logistic regression analysis and model performance metrics

Variable	β coefficient	SE	OR (95% CI)	P value
Multivariate Analysis				
Thumb sucking (ref: Never)				
Past	0.52	0.22	1.68 (1.09-2.59)	0.018
Current	1.23	0.25	3.42 (2.08-5.63)	<0.001
Mouth breathing (ref: No)				
Occasional	0.68	0.18	1.97 (1.38-2.81)	<0.001
Habitual	1.05	0.21	2.86 (1.89-4.33)	<0.001
Tongue thrusting	0.81	0.20	2.24 (1.52-3.31)	<0.001
Parental malocclusion (ref: Neither)				
One parent	0.58	0.16	1.79 (1.31-2.44)	<0.001
Both parents	0.99	0.25	2.68 (1.64-4.38)	<0.001
Premature tooth loss	0.65	0.18	1.92 (1.36-2.71)	<0.001
Breastfeeding <6 months	0.58	0.17	1.78 (1.28-2.47)	0.001
Bottle feeding >24 months	0.50	0.17	1.65 (1.18-2.31)	0.003
Constant	-2.18	0.28	-	<0.001
Model Performance Metrics				
	Training Cohort		Validation Cohort	
AUC (95% CI)	0.786 (0.754-0.818)		0.772 (0.721-0.823)	
Sensitivity	0.743		0.732	
Specificity	0.716		0.705	
PPV	0.652		0.638	
NPV	0.798		0.791	
Accuracy	0.727		0.716	
Brier Score	0.198		0.206	
Hosmer-Lemeshow χ^2 (P value)	8.42 (0.394)		7.86 (0.447)	

3.3 Establishment of Prediction Model

Based on the multivariable analysis results, a nomogram prediction model

incorporating 7 independent predictive factors was constructed (**Figure 2**). This model assigns corresponding scores to each risk factor, and individual malocclusion risk can be rapidly assessed by summing the total score. For example, a child with habitual mouth breathing (28 points), premature loss of primary teeth (18 points), and 3 months of breastfeeding (22 points) would have a total score of 68 points, corresponding to approximately 65% risk of malocclusion.

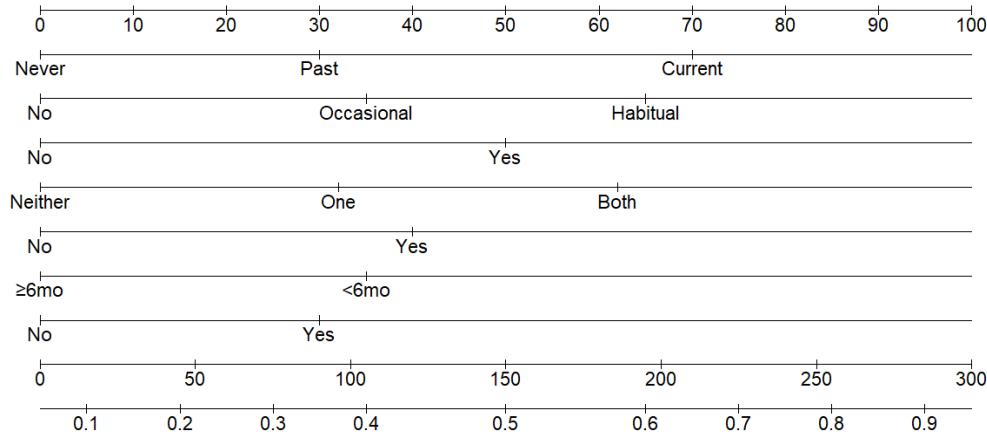


Figure 2. Nomogram for predicting the risk of malocclusion in children

3.4 Model Performance Evaluation and Validation

Model performance metrics are shown in the lower half of **Table 3**. The model achieved an AUC of 0.786 (95%CI: 0.754-0.818) in the training set and 0.772 (95%CI: 0.721-0.823) in the validation set, demonstrating good discrimination. At the optimal cutoff value of 0.42, the training set showed sensitivity of 74.3% and specificity of 71.6%. In the validation set, the sensitivity was 73.2% and specificity was 70.5% at the same cutoff value. The Hosmer-Lemeshow test indicated good calibration of the model (training set $P=0.394$, validation set $P=0.447$).

The comprehensive performance evaluation of the model is shown in **Figure 3**. ROC curve analysis showed that the curves for the training and validation sets were essentially overlapping, suggesting good model stability. The calibration plot demonstrated high consistency between predicted probabilities and actual observed probabilities, with deviations between predicted and observed values being less than 5% for all risk groups. Decision curve analysis indicated that within the threshold probability range of 0.15-0.75, the net benefit of using this prediction model was significantly higher than both “treat all” and “treat none” strategies.

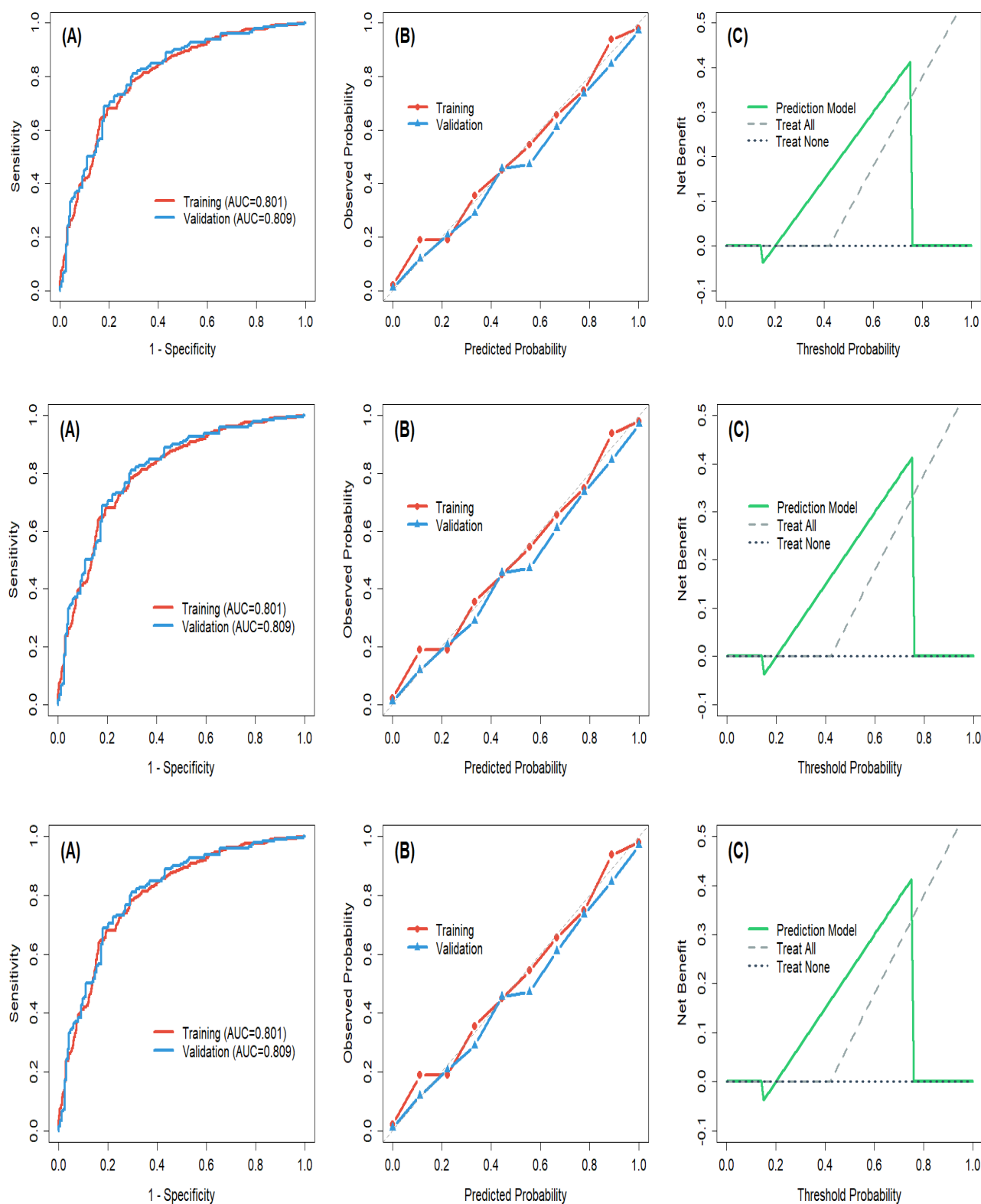


Figure 3. Model validation and performance evaluation

Bootstrap internal validation showed that the mean AUC from 1000 resampling iterations was 0.779 (95%CI: 0.768-0.790), differing by less than 0.01 from the original model's AUC, indicating good internal stability of the model. All performance metrics in the external validation cohort were essentially consistent with the training set, confirming the model's generalizability and clinical applicability.

3.5 Development of Clinical Application Tool

To facilitate clinical application, the study developed a web-based risk assessment tool (Figure 4). This tool features a clean and user-friendly interface where healthcare professionals need only input basic information about the child and relevant risk factors, and the system automatically calculates the risk score and provides individualized risk level assessment. The tool also provides risk stratification management recommendations: low risk (<30%) suggests regular observation; moderate risk (30-60%) recommends strengthening oral health education and correction of bad habits; high risk (>60%) suggests timely referral to orthodontic specialists for early intervention.

Pediatric Malocclusion Risk Calculator

Patient Information

Risk Assessment

Total Score: 68 points

Risk Level: HIGH (65%)

0% — 65% — 100%

Clinical Recommendations:

- Immediate orthodontic consultation
- Habit correction therapy indicated
- Myofunctional therapy beneficial
- Follow-up evaluation in 3 months

Calculate

Reset

Export PDF

Figure 4. Web-based clinical application tool interface

4. Discussion

This study constructed and validated a multifactorial analysis-based risk prediction model for childhood malocclusion, which integrates multidimensional information including oral bad habits, family history, early dental developmental

abnormalities, and feeding history, demonstrating good predictive performance and clinical applicability. The study found that 42.5% of children had malocclusion, a result similar to recent research conducted in Vietnam, which reported significant associations between different oral habits and malocclusion^[25].

Oral bad habits as important risk factors for malocclusion were thoroughly validated in this study. The research found that current digit sucking habit was the strongest independent predictive factor (OR=3.42), which is consistent with results from Japanese scholars studying preschool children aged 3-6 years, who found that incomplete lip closure and nail-biting habits were important risk factors for malocclusion^[26]. The impact of oral bad habits on childhood malocclusion has been widely recognized, with recent studies emphasizing the etiological significance of these habits and the importance of early preventive intervention^[27]. A study on parental perception of oral habits revealed the high incidence of these habits in school-age children and their predictive factors, providing important epidemiological support for the risk assessment in this study^[28].

Habitual mouth breathing showed strong predictive value in this study (OR=2.86), a finding supported by systematic review evidence. Lin et al.'s review comprehensively described the adverse effects of mouth breathing on craniofacial development, including changes such as high palatal vault, mandibular retrusion, and dental arch narrowing^[29]. Systematic reviews and meta-analyses have confirmed the close association between mouth breathing and abnormal swallowing patterns, with this superimposed effect of functional abnormalities potentially exacerbating the severity of malocclusion^[30]. Wang et al.'s research found correlations among malocclusion, poor oral habits, and caries rates, suggesting that these factors may affect oral health through common pathophysiological mechanisms^[31].

The prediction model constructed in this study achieved AUC values of 0.786 and 0.772 in the training and validation sets, respectively, showing good discriminative ability. This performance is comparable to recently published orthodontic treatment decision prediction studies, where Huang et al. used four machine learning methods to predict orthodontic extraction decisions, with model prediction accuracy similar to this study^[32]. Volovic et al.'s orthodontic treatment duration prediction model also demonstrated the application potential of machine learning in orthodontics^[33]. Mohammad-Rahimi et al.'s scoping review systematically summarized the current status and future opportunities of machine learning applications in orthodontics, providing theoretical basis for the methodological choices in this study^[34].

The application of artificial intelligence and machine learning technologies in orthodontic diagnosis and treatment is increasingly widespread. Bichu et al.'s scoping review detailed the applications of these technologies in diagnosis, treatment planning, and prognosis assessment, confirming the value of prediction models in improving clinical decision-making^[35]. This study employed traditional logistic regression methods to construct the model, which, although relatively

simple, has good interpretability and clinical practicality. Similar nomogram models have been successfully applied in oral disease prediction, such as Duan et al.'s nomogram for screening severe caries in 12-year-old children, providing a convenient tool for school oral health screening^[36].

Nomograms, as intuitive risk assessment tools, have shown unique advantages in pediatric disease prediction. Jiang et al.'s nomogram model for predicting recurrence risk of complicated urinary tract infections in children achieved individualized risk assessment by integrating clinical features^[37]. In the field of oral medicine, Ravidà et al.'s nomogram for predicting tooth loss in periodontal disease successfully predicted long-term prognosis using staging and grading systems^[38]. The nomogram model in this study enables clinicians to rapidly assess children's malocclusion risk through a simple scoring system, providing quantitative evidence for formulating individualized prevention strategies.

Decision curve analysis showed that this model had positive net benefit within the threshold probability range of 0.15-0.75, indicating that the model has practical value in clinical practice. Wang et al.'s systematic review and meta-analysis of caries risk prediction models in children and adolescents emphasized the important role of risk prediction models in early identification and prevention of oral diseases^[39].

The sample size calculation in this study followed statistical principles for multivariable prediction model development. de Jong et al.'s research provided important guidance for sample size determination in multinomial logistic prediction models, emphasizing the impact of adequate sample size on model performance^[40]. Pate et al. further refined the minimum sample size calculation methods for multivariable prediction models, providing methodological support for the sample size design in this study^[41]. Notably, Collins et al.'s systematic evaluation of prostate cancer prediction model studies found that many studies lacked sufficient justification for sample size and were often inadequate, reminding researchers to emphasize scientific sample size calculation^[42].

Several limitations should be acknowledged in interpreting our findings. While we validated our model using an independent dataset, the cross-sectional nature of our data collection limits our ability to establish definitive causal relationships. Our study population was drawn from schools within a single geographic region, raising questions about how well the model might perform in different settings or populations. Additionally, we relied on parental recall for certain variables like feeding history, which inevitably introduces some degree of recall bias. We also recognize that incorporating objective measures—such as genetic markers or imaging parameters—could potentially enhance the model's predictive capability, though these were not included in the current study.

Nevertheless, our prediction model represents a valuable step forward in addressing childhood malocclusion. The integration of multiple risk factors allows for personalized risk assessment, making it easier to identify children who would benefit most from early intervention. The accompanying web-based appli-

cation transforms this research into a practical screening tool that primary care providers and schools can readily use. This accessibility is particularly important for implementing population-level screening programs and developing targeted management strategies. Moving forward, it would be worthwhile to incorporate more objective biomarkers, expand validation efforts across different populations and regions, and investigate whether machine learning approaches could further refine predictive accuracy. Such advances could strengthen our ability to prevent and manage malocclusion in pediatric populations.

5. Conclusion

This study successfully constructed and validated a multifactorial analysis-based risk prediction model for childhood malocclusion, which integrates seven independent predictive factors including oral bad habits, family history, early dental developmental abnormalities, and feeding history, achieving AUC values of 0.786 and 0.772 in the training and validation sets, respectively, demonstrating good predictive performance. The study clearly identified current digit sucking habit, habitual mouth breathing, tongue thrusting, parental history of malocclusion in both parents, premature loss of primary teeth, insufficient breastfeeding duration, and prolonged bottle use as major risk factors for childhood malocclusion. The prediction model presented in nomogram form transforms complex statistical models into simple and practical clinical tools, with decision curve analysis confirming that the model has positive net benefit within a wide range of threshold probabilities. The developed web-based risk assessment tool further enhances the accessibility of the model, providing technical support for large-scale screening in primary healthcare institutions and schools. The establishment of this prediction model helps achieve early identification and risk stratification management of malocclusion, enabling high-risk children to receive timely preventive intervention. The study results provide scientific evidence for formulating individualized prevention strategies based on risk stratification, which is of great significance for optimizing medical resource allocation and improving the efficiency of children's oral health management, and is expected to play a positive role in improving children's oral health levels.

FUNDING

This work was supported by Missing repair implants digital clinical application promotion in the whole process of the research (2023ZD062).

References

1. Stomatologic, S., Worldwide prevalence of malocclusion in the different stages of dentition: A systematic review and meta-analysis. *Eur J Paediatr Dent*, 2020. 21(2): p. 115-22.
2. Chen, H., et al. Prevalence of Malocclusion Traits in Primary Dentition, 2010–2024: A Systematic Review. in

Healthcare. 2024. MDPI.

3. Mai, W., et al., Prevalence and contributing factors of malocclusion in Zhuang children aged 7–8 years in southern China. *Frontiers in Pediatrics*, 2024. 12: p. 1308039.
4. Alhammadi, M.S., et al., Global distribution of malocclusion traits: A systematic review. *Dental press journal of orthodontics*, 2018. 23(06): p. 40. e1-40. e10.
5. Zhou, C., et al., Expert consensus on pediatric orthodontic therapies of malocclusions in children. *International Journal of Oral Science*, 2024. 16(1): p. 32.
6. De Ridder, L., et al., Prevalence of orthodontic malocclusions in healthy children and adolescents: a systematic review. *International journal of environmental research and public health*, 2022. 19(12): p. 7446.
7. Kasparaviciene, K., et al., The prevalence of malocclusion and oral habits among 5–7-year-old children. *Medical science monitor: international medical journal of experimental and clinical research*, 2014. 20: p. 2036.
8. Vickers, A.J. and E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 2006. 26(6): p. 565-574.
9. Collins, G.S., et al., TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 2024. 385.
10. Khanagar, S.B., et al., Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making-a systematic review. *Journal of dental sciences*, 2021. 16(1): p. 482-492.
11. Liu, J., C. Zhang, and Z. Shan. Application of artificial intelligence in orthodontics: current state and future perspectives. in *Healthcare*. 2023. MDPI.
12. Collins, G.S., et al., Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Journal of British Surgery*, 2015. 102(3): p. 148-158.
13. Moons, K.G., et al., Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*, 2015. 162(1): p. W1-W73.
14. Bujang, M.A., et al., Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. *The Malaysian journal of medical sciences: MJMS*, 2018. 25(4): p. 122.
15. Riley, R.D., et al., Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. *Statistics in medicine*, 2019. 38(7): p. 1276-1296.
16. Organization, W.H., Oral health surveys: basic methods. 2013: World Health Organization.
17. Angle, E.H., Classification of malocclusion. 1899.
18. Harrell, F.E. and D.G. Levy, Regression modeling strategies. R package version, 2022: p. 6.3-0.
19. Fernandez-Felix, B., et al., Bootstrap internal validation command for predictive logistic regression models. *The Stata Journal*, 2021. 21(2): p. 498-509.
20. Collins, G.S., et al., Evaluation of clinical prediction models (part 1): from development to external validation. *Bmj*, 2024. 384.
21. Sadatsafavi, M., P. Saha-Chaudhuri, and J. Petkau, Model-based ROC curve: examining the effect of case mix and model calibration on the ROC plot. *Medical Decision Making*, 2022. 42(4): p. 487-499.
22. Verbakel, J.Y., et al., ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *Journal of Clinical Epidemiology*, 2020. 126: p. 207-216.
23. Van Calster, B., et al., Calibration: the Achilles heel of predictive analytics. *BMC medicine*, 2019. 17(1): p. 230.
24. Vickers, A.J., B. van Calster, and E.W. Steyerberg, A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and prognostic research*, 2019. 3(1): p. 18.
25. Vu, D.A., et al., Malocclusion among children in Vietnam: Prevalence and associations with different habits.

Journal of oral biology and craniofacial research, 2024. 14(2): p. 112-115.

26. Otsugu, M., et al., Incompetent lip seal and nail biting as risk factors for malocclusion in Japanese preschool children aged 3–6 years. *BMC pediatrics*, 2023. 23(1): p. 532.

27. Katib, H.S., et al., Influence of Oral Habits on Pediatric Malocclusion: Etiology and Preventive Approaches. *Cureus*, 2024. 16(11).

28. Khattab, N.M.A., M.A.A. Abd-Elsabour, and O.M. Omar, Parent-perceived oral habits among a group of school children: prevalence and predictors. *BDJ open*, 2024. 10(1): p. 77.

29. Lin, L., et al., The impact of mouth breathing on dentofacial development: A concise review. *Frontiers in public health*, 2022. 10: p. 929165.

30. Gómez-González, C., et al., Mouth Breathing and its impact on atypical swallowing: a systematic review and Meta-analysis. *Dentistry Journal*, 2024. 12(2): p. 21.

31. Wang, Z., et al., Analysis of the correlation between malocclusion, bad oral habits, and the caries rate in adolescents. *Translational Pediatrics*, 2021. 10(12): p. 3291.

32. Huang, J., et al., Evaluation of four machine learning methods in predicting orthodontic extraction decision from clinical examination data and analysis of feature contribution. *Frontiers in Bioengineering and Biotechnology*, 2024. 12: p. 1483230.

33. Volovic, J., et al., A novel machine learning model for predicting orthodontic treatment duration. *Diagnostics*, 2023. 13(17): p. 2740.

34. Mohammad-Rahimi, H., et al., Machine learning and orthodontics, current trends and the future opportunities: A scoping review. *American Journal of Orthodontics and Dentofacial Orthopedics*, 2021. 160(2): p. 170-192. e4.

35. Bichu, Y.M., et al., Applications of artificial intelligence and machine learning in orthodontics: a scoping review. *Progress in orthodontics*, 2021. 22(1): p. 18.

36. Duan, S., et al., A predictive nomogram: a cross-sectional study on a simple-to-use model for screening 12-year-old children for severe caries in middle schools. *BMC Oral Health*, 2021. 21(1): p. 457.

37. Jiang, J., X.-Y. Chen, and H. Guo, Clinical characteristics and nomogram model for predicting the risk of recurrence of complicated urinary tract infection in pediatric patients. *Scientific Reports*, 2024. 14(1): p. 25393.

38. Ravidà, A., et al., Development of a nomogram for the prediction of periodontal tooth loss using the staging and grading system: A long-term cohort study. *Journal of Clinical Periodontology*, 2020. 47(11): p. 1362-1370.

39. Wang, X., et al., Risk prediction models for dental caries in children and adolescents: a systematic review and meta-analysis. *BMJ open*, 2025. 15(3): p. e088253.

40. de Jong, V.M., et al., Sample size considerations and predictive performance of multinomial logistic prediction models. *Statistics in medicine*, 2019. 38(9): p. 1601-1619.

41. Pate, A., et al., Minimum sample size for developing a multivariable prediction model using multinomial logistic regression. *Statistical methods in medical research*, 2023. 32(3): p. 555-571.

42. Collins, S.D., et al., Sample sizes of prediction model studies in prostate cancer were rarely justified and often insufficient. *Journal of Clinical Epidemiology*, 2021. 133: p. 53-60.