

OPEN ACCESS

[DOI] 10.12016/j.issn.2096-1456.202660145

· 临床研究 ·

北京市5岁儿童新发龋预测模型

任雯, 赵梅, 张辉, 陈薇, 沈安琪, 刘敏

首都医科大学口腔医学院预防科, 北京(100070)

【摘要】目的 探讨北京市3岁无龋儿童5岁新发龋的影响因素,构建基于口腔健康行为数据的机器学习预测模型,为学龄前儿童龋病早期风险识别、分层防控及家-园-医协同干预提供科学依据与实用工具。**方法** 本研究已获得单位医学伦理委员会批准,并已获得研究参与者及其监护人的书面知情同意。选取参与北京市口腔公共卫生项目、具有完整3岁基线与5岁随访数据儿童为研究对象,依据世界卫生组织龋病诊断标准开展口腔检查与问卷调查。采用Logistic回归筛选危险因素,将总样本随机划分为测试集(20%)与验证集(80%),在验证集上采用5折交叉验证进行模型训练与调优,最终在测试集上评价模型效能。构建决策树、随机森林、极端梯度提升(XGBoost)、轻量级梯度提升机(LightGBM)四种预测模型,以曲线下面积(AUC)、准确率、特异度等评价模型效能,结合沙普利可加性解释方法(SHAP)方法进行模型解释。**结果** 纳入研究的北京市3372名5岁儿童新发龋发病率为50.96%;多因素Logistic回归显示,父母患龋、含奶入睡、夜奶频繁、餐后不清洁、高糖饮食、睡前刷牙不足、使用非含氟牙膏等为危险因素($P < 0.05$)。四种机器学习模型均表现良好,其中XGBoost模型预测效能最优($AUC = 0.806$)。特征重要性与SHAP分析提示,睡前刷牙频率、吃糖果频率、父母龋齿状况为核心预测因子。**结论** 基于3岁口腔健康行为数据构建的XGBoost模型可预测北京市5岁儿童新发龋风险,睡前刷牙、控糖、父母口腔健康管理是学龄前儿童龋病防控关键要点。

【关键词】 学龄前儿童; 龋病风险; 风险预测; 分层防控; 预测模型; 回归分析; 机器学习; 极端梯度提升; 口腔健康行为

【中图分类号】 R78 **【文献标志码】** A **【文章编号】** 2096-1456(2026)10-0982-11

【引用著录格式】 任雯,赵梅,张辉,等.北京市5岁儿童新发龋预测模型[J].口腔疾病防治,2026,34(10):982-992. doi:10.12016/j.issn.2096-1456.202660145.

The prediction model of incident dental caries in 5-year-old children in Beijing REN Wen, ZHAO Mei, ZHANG Hui, CHEN Wei, SHEN Anqi, LIU Min. Department of Preventive Dentistry, School of Stomatology, Capital Medical University, Beijing 100070, China

Corresponding author: LIU Min, Email: sdyd_2314@163.com

【Abstract】Objective To explore the influencing factors of incident caries at age 5 among caries-free 3-year-old children in Beijing, construct a machine learning prediction model based on oral health behavioral data, and provide a scientific basis and practical tool for early risk identification, stratified prevention and control, and home - kindergarten - medical collaborative intervention of incident dental caries in preschool children. **Methods** This study has been reviewed and approved by the Medical Ethics Committee, and written informed consent was obtained from the guardians of the study participants and their guardians. Children with complete baseline data at age 3 and follow-up data at age 5 from Beijing public oral health programs were included. Oral examinations were conducted according to World Health Organization diagnostic criteria, and questionnaires were completed by guardians. Risk factors were screened by Logistic regression. The total sample was randomly divided into a test set (20%) and a validation set (80%). Five-fold cross-validation was performed on the validation set for model training and tuning, and the final model performance was evalu-



微信公众号

【收稿日期】 2026-04-15; **【修回日期】** 2026-07-23

【基金项目】 中华口腔医学会口腔医生龋病防治能力提升计划项目(CSA-ICP2022-07)

【作者简介】 任雯,主治医师,博士研究生,Email: renwen90@foxmail.com

【通信作者】 刘敏,主任医师,博士研究生,Email: sdyd_2314@163.com

ated on the test set. Four machine learning models—decision tree, random forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM)—were constructed. Model performance was evaluated using the area under the curve (AUC), accuracy, and specificity. Shapley Additive Explanations (SHAP) was used for model interpretation. **Results** The incidence of incident caries for 3 372 5-year-old children in Beijing was 50.96%. Multivariate logistic regression showed that parental caries, sleeping with milk, frequent feeding at night, no postprandial cleaning, high sugar intake, insufficient bedtime tooth brushing, and non-fluoride toothpaste use were significant risk factors ($P < 0.05$). Although all four machine learning models performed well, the XGBoost model performed the best (AUC = 0.806). Feature importance and SHAP analysis indicated that the frequency of bedtime tooth brushing, the frequency of candy consumption, and parental caries status were the top three predictors. **Conclusion** The XGBoost model based on oral health behaviors at age 3 can accurately predict the risk of incident caries in 5-year-old children in Beijing. Bedtime tooth brushing, sugar control, and parental oral health management are key points for caries prevention in preschool children.

【Key words】 preschool children; caries risk; risk prediction; stratified prevention and control; prediction model; regression analysis; machine learning; extreme gradient boosting; oral health behavior

J Prev Treat Stomatol Dis, 2026, 34(10): 982-992.

龋病是学龄前儿童最常见的慢性非传染性疾病,患病率高、危害广泛,可影响咀嚼、生长发育及心理健康,给家庭和社会带来较重负担^[1-3]。第四次全国口腔健康流行病学调查显示,3岁组儿童患龋率为50.8%,5岁年龄组龋病患病率为71.9%,且分布不均^[4-5]。北京市对于低龄儿童的公共卫生项目有两个:2011年开展的针对3~6岁学龄前儿童的氟化泡沫防龋项目;2021年启动的北京市0~3岁儿童口腔综合干预项目^[6],虽然实现学龄前儿童年龄全覆盖,但风险区分和分层防控仍是当前口腔公共卫生的重点工作。

国内外研究证实,早期含糖饮食、睡前刷牙、含奶瓶入睡、父母口腔健康、氟化物使用等因素与儿童龋病密切相关^[5, 7-13]。传统回归模型难以捕捉多变量间的复杂关系,而机器学习可利用高维行为数据构建高精度预测模型,已逐渐成为儿童龋病风险评估的重要方法^[14-16]。学龄前儿童龋病防控高度依赖家庭、幼儿园与医疗机构的协同联动,但幼儿园保健医与家长普遍缺乏简便、精准的龋病风险识别工具。鉴于此,本研究以3岁无龋儿童为研究对象,基于口腔健康行为数据,采用传统回归模型筛选关键危险因素,联合四种机器学习算法构建并验证5岁新发龋风险预测模型,为北京市学龄前儿童龋病早期识别、分层管理及精准干预提供科学依据,同时为家-园-医协同防控提供简便易用的风险识别工具。

1 资料和方法

1.1 研究对象及数据来源

本研究为回顾性队列研究,严格遵循《赫尔辛基宣言》伦理原则,并获得首都医科大学附属北京口腔医院伦理委员会批准(CMUSH-IRB-KJ-PJ-2023-41)。研究对象来自参与2021–2023年北京市氟化泡沫项目及0~3岁儿童口腔综合干预项目的儿童,并已获得研究参与者及其监护人的书面知情同意。本研究报告严格按照《加强流行病学中观察性研究报告质量(STROBE)声明:观察性研究报告规范》进行撰写。纳入标准:①同时参与两个项目,3岁基线口腔检查为无龋(龋失补指数(乳牙)=0),且具有完整的3岁基线问卷数据及5岁随访检查数据的儿童;②儿童无系统性疾病及严重颌面部发育异常。排除标准:存在明显逻辑错误、重复录入或无法匹配的无效数据所对应的儿童。

1.2 口腔检查及问卷调查

口腔检查中龋病诊断和记录标准参照世界卫生组织流行病学标准(第5版)^[17]。调查问卷为3岁基线时监护人在线填写的问卷,包括父母龋病情况、口腔健康行为指标、饮食指标等。

1.3 样本量估计

本研究预测变量数为26个,结局指标“是否有龋”为二分类变量,结局事件5岁患龋率参照2022年北京市哨点监测结果,为65.3%^[6],利用R语言pmsampsize包计算样本量^[18],具体参数为:type = b, csrsquare = 0.2, parameters = 26, prevalence =

0.653。计算结果所需样本量至少为1 035人。按上述纳入排除标准,最终3 372名儿童纳入研究。

1.4 偏倚控制

为减少选择偏倚,本研究采用北京口腔公共卫生项目连续纳入的儿童数据,避免了便利抽样造成的样本选择偏倚。信息偏倚方面,口腔检查由经过统一培训和标准一致性检验($\kappa > 0.8$)的医生执行,检查结果实行双录入核查,以降低测量偏倚和录入错误。问卷调查由监护人回顾性填写,可能存在回忆偏倚和社会期望偏倚,但问卷中多数行为问题采用频率等级(如“天天刷”“很少”等),具有较好可重复性。此外,数据分析时采用多因素分析和机器学习交叉验证,以减少混杂偏倚的影响。

1.5 统计学分析

采用Python 3.9和R 4.5.2进行统计分析。

Logistic 回归筛选标准:将所有预测变量逐个进行单因素Logistic 回归分析,以 $P < 0.1$ 为初步筛选标准,选出与5岁新发龋相关的候选变量。将单因素Logistic 回归筛选出的候选变量纳入多因素Logistic 回归模型,采用向前逐步回归,最终确定独立危险因素($P < 0.05$)。

数据集划分:将总样本($n = 3\,372$)划分为训练集(80%, $n = 2\,698$)与测试集(20%, $n = 674$),其中测试集仅用于最终模型性能评估。在训练集上,采用5折交叉验证进行模型训练与超参数调优,确定最优超参数后,使用全部训练集重新训练最终模型,并在测试集上评价其泛化性能。

机器学习模型输入变量与模型验证策略:将多因素Logistic 回归最终筛选的独立危险因素($P < 0.05$)纳入决策树、随机森林、极端梯度提升(ex-

treme gradient boosting, XGBoost)、轻量级梯度提升机(light gradient boosting machine, LightGBM)4种模型。以曲线下面积(area under curve, AUC)、准确率、精准度、召回率、特异度评价模型效能。采用DeLong 检验比较AUC 差异,并计算95%置信区间(confidence interval, CI);采用McNemar 检验比较准确率、精准度、召回率、特异度上的差异。为控制多重比较的假阳性率,对每个评价指标的两两比较采用Holm-Bonferroni 法进行校正。采用沙普利可加性解释方法(shapley additive explanations, SHAP)汇总图进行模型解释^[19]。校准度评估采用校准曲线,并计算校准斜率与截距。临床适用性评价采用决策曲线分析,计算不同风险阈值下的净获益,以评估模型的临床净获益。

2 结 果

2.1 研究对象基本情况

根据纳入排除标准,调取数据库中所有参与两个项目的儿童,删除有缺项及逻辑错误的条目,以儿童ID号作为唯一识别编码,最终筛选出3 372名5岁儿童进入最终研究,其中男性1 706名(50.59%),女性1 666名(49.41%);5岁新发龋儿童1 718名,发病率为50.96%。

2.2 新发龋相关危险因素分析

单因素Logistic 回归分析显示,父母龋齿情况、共用餐具行为、含奶入睡习惯、夜奶次数、餐后饮水/漱口、日常饮水类型、吃零食频率、吃糖果频率、睡前刷牙频率、是否使用牙膏、牙膏类型、是否使用牙线与5岁儿童新发龋相关($P < 0.05$),详见表1。

表1 北京市5岁儿童新发龋相关因素的单因素Logistic 回归分析

Table 1 Univariate logistic regression analysis of factors associated with incident dental caries in 5-year-old children in Beijing

Factors	Total/[n (%)]	Caries		χ^2	P
		Number	Rate (%)		
Gender				0.834	0.361
Male	1 706 (50.59)	856	50.18		
Female	1 666 (49.41)	862	51.74		
Parental caries status				17.584	< 0.001
Neither parent had caries	1 548 (45.91)	719	46.45		
Father or mother had caries	1 019 (30.22)	560	54.96		
Both parents had caries	805 (23.87)	439	54.53		
Sharing eating utensils				6.449	0.011
Almost never	1 415 (41.96)	686	48.48		
Rarely	1 680 (49.82)	877	52.20		

续表1

continue table 1

Factors	Total/[n (%)]	Caries		χ^2	P
		Number	Rate (%)		
Frequently	277 (8.22)	155	55.96	28.337	< 0.001
Habit of sleeping with milk					
Bottle discontinued	2 006 (59.49)	954	47.56		
Rarely	1 233 (36.57)	678	54.99	16.261	< 0.001
Frequently	133 (3.94)	86	64.66		
Night feeding frequency					
0 times	3 056 (90.63)	1 523	49.84	30.715	< 0.001
1 - 2 times	285 (8.45)	171	60.00		
≥3 times	31 (0.92)	24	77.42		
Post-meal water drinking/rinsing				9.861	0.002
Frequently	1 772 (52.55)	811	45.77		
Rarely	1 317 (39.06)	751	57.02		
Almost never	283 (8.39)	156	55.12	25.112	< 0.001
Daily drinking water					
Warm boiled water	2 831 (83.96)	1 399	49.42		
Others	541 (16.04)	319	58.96	63.552	< 0.001
Snack frequency					
Never	124 (3.68)	49	39.52		
Rarely	1 368 (40.57)	623	45.54	68.881	< 0.001
Once daily	1 447 (42.91)	732	50.59		
≥2 times daily	433 (12.84)	264	60.97		
Candy frequency				4.347	0.037
Never	347 (10.29)	128	36.89		
Rarely	1 853 (54.95)	854	46.09		
Once daily	895 (26.54)	473	52.85	6.590	0.010
≥2 times daily	277 (8.22)	194	70.04		
Bedtime toothbrushing frequency					
Every day	2645 (78.44)	1 251	47.30	17.978	< 0.001
Rarely	685 (20.31)	445	64.96		
Never	42 (1.25)	22	52.38		
Using toothpaste ^a (n=3 330)				0.120	0.729
Yes	3 121 (93.72)	1 575	50.46		
No	209 (6.28)	121	57.89		
Toothpaste type ^b (n=3 121)				0.120	0.729
Fluoride toothpaste	2 504 (80.23)	1 222	48.80		
Non-fluoride toothpaste	406 (13.01)	218	53.69		
Unknown	211 (6.76)	135	63.98	17.978	< 0.001
Using dental floss					
Frequently	2 135 (63.32)	1 161	54.38		
Rarely	984 (29.18)	450	45.73	0.120	0.729
Never	253 (7.50)	107	42.29		
Visiting dentist					
Yes	2 270 (67.32)	1 161	51.15	0.120	0.729
No	1 102 (32.68)	557	50.54		

a: the number of people who use toothpaste needs to exclude those who never brush their teeth (n=42), so the total number of this item should be 3 330 children; b: the number of people using various types of toothpaste is only counted among those who have the habit of using toothpaste, so it is 3 121 children

以是否新发龋为因变量,多因素 Logistic 回归显示,父母口内存在龋齿、经常含奶入睡、夜奶 ≥ 3 次、餐后几乎不饮水/漱口、日常饮用非白开水、吃零食频率 ≥ 2 次/d、高频率吃糖果、睡前很少或不刷牙、使用非含氟牙膏、不使用牙线会显著提高新发龋风险($P < 0.05$),详见表2。

表2 北京市5岁儿童新发龋相关因素的多因素 Logistic 回归分析

Table 2 Multivariate logistic regression analysis of factors associated with incident dental caries in 5-year-old children in Beijing

Factors	<i>B</i>	<i>SE</i>	<i>Wald</i>	<i>P</i>	<i>OR</i> (95% <i>CI</i>)
Parental caries status (reference: neither parent had caries)					
Father or mother had caries	0.269	0.092	8.587	0.003	1.309(1.093-1.568)
Both parents had caries	0.307	0.085	13.066	< 0.001	1.359(1.151-1.605)
Sharing eating utensils (reference: almost never)					
Rarely	-0.011	0.077	0.022	0.882	0.989(0.850-1.149)
Frequently	-0.058	0.142	0.164	0.686	0.944(0.715-1.247)
Habit of sleeping with milk(reference: bottle discontinued)					
Rarely	0.396	0.207	3.654	0.056	1.486(0.990-2.230)
Frequently	0.180	0.078	5.386	0.020	1.197(1.028-1.393)
Night feeding frequency(reference: 0 times)					
1 - 2 times	0.114	0.139	0.675	0.411	1.121(0.854-1.470)
≥ 3 times	0.979	0.457	4.591	0.032	2.661(1.087-6.516)
Post-meal water drinking/rinsing (reference: frequently)					
Rarely	0.065	0.138	0.220	0.640	1.067(0.814-1.400)
Almost never	0.256	0.078	10.844	< 0.001	1.292(1.109-1.505)
Daily drinking water (reference: warm boiled water)					
Others	0.506	0.171	8.768	0.003	1.658(1.186-2.317)
Snack frequency (reference: never)					
Rarely	1.039	0.536	3.757	0.053	2.825(0.988-8.076)
Once daily	0.951	0.538	3.126	0.077	2.588(0.902-7.422)
≥ 2 times daily	1.249	0.539	5.375	0.020	3.485(1.213-10.015)
Candy frequency (reference: never)					
Rarely	0.546	0.126	18.825	< 0.001	1.726(1.350-2.209)
Once daily	0.834	0.173	23.266	< 0.001	2.303(1.641-3.233)
≥ 2 times daily	0.872	0.151	33.371	< 0.001	2.393(1.780-3.217)
Bedtime toothbrushing frequency (reference: every day)					
Rarely / never	0.484	0.096	25.250	< 0.001	1.623(1.344-1.960)
Using toothpaste (reference: yes)					
No	0.165	0.155	1.130	0.288	1.179(0.870-1.598)
Toothpaste type (reference: fluoride toothpaste)					
Non-fluoride toothpaste	0.460	0.155	8.746	0.003	1.583(1.168-2.147)
Unknown	0.092	0.112	0.685	0.408	1.097(0.881-1.364)
Using dental floss (reference: frequently)					
Rarely	0.013	0.147	0.008	0.043	1.513(1.352-1.756)
Never	0.275	0.141	3.824	0.021	1.817(1.734-1.999)

SE: standard error; *OR*: odds ratio; *CI*: confidence interval

2.3 机器学习模型效能比较

将多因素 Logistic 回归有统计学意义的变量纳入机器学习算法模型中,4种机器学习模型均具有一定预测能力,其中 XGBoost 模型表现最优,AUC 为 0.806,准确率 0.747,精准度 0.819,特异度 0.852,

召回率 0.641。随机森林模型次之,AUC 为 0.786,LightGBM 与决策树模型预测效能相对较低(表3)。

进一步采用 DeLong 检验和 McNemar 检验对模型进行两两比较,并采用 Holm-Bonferroni 法校正多

表3 模型性能评价

Table 3 Model performance evaluation

Model	Accuracy (95%CI)	Precision (95%CI)	Recall (95%CI)	Specificity (95%CI)	AUC (95%CI)
Decision Tree	0.706 (0.684-0.728)	0.734 (0.700-0.768)	0.662 (0.629-0.695)	0.748 (0.717-0.779)	0.718 (0.688-0.748)
Random Forest	0.715 (0.693-0.737)	0.772 (0.739-0.805)	0.621 (0.588-0.654)	0.808 (0.779-0.837)	0.786 (0.758-0.814)
XGBoost	0.747 (0.717-0.777)	0.819 (0.787-0.851)	0.641 (0.608-0.674)	0.852 (0.824-0.880)	0.806 (0.778-0.834)
LightGBM	0.744 (0.722-0.766)	0.841 (0.811-0.871)	0.613 (0.580-0.646)	0.879 (0.852-0.906)	0.712 (0.682-0.742)

AUC: area under curve; CI: confidence interval

重比较(表4)。结果显示:XGBoost模型的AUC显著高于LightGBM和决策树(校正后 $P < 0.05$),但与随机森林的差异未达到校正后显著性水平(校正后 $P = 0.155$);各模型在准确率、精准度、召回率、

特异度上的差异均未达到校正后显著性阈值(校正后均 $P > 0.05$)。上述统计推断结果证实XGBoost模型在区分度上具有相对优势。

表4 四种模型性能指标的两两比较 Holm-Bonferroni 校正后P值

Table 4 Comparisons of four models with Holm-Bonferroni corrected P-values

Model Comparison	Accuracy [†]	Precision [†]	Recall [†]	Specificity [†]	AUC [‡]
XGBoost vs. LightGBM	1	1	1	1	< 0.001*
XGBoost vs. Decision Tree	0.096	0.114	0.126	0.054	< 0.001*
Random Forest vs. LightGBM	0.354	0.372	0.366	0.33	< 0.001*
Random Forest vs. Decision Tree	1	1	1	1	0.008*
XGBoost vs. Random Forest	0.162	0.174	0.186	0.132	0.155
LightGBM vs. Decision Tree	0.168	0.186	0.198	0.156	1

[†]Comparisons of accuracy, precision, recall, and specificity were all performed using McNemar's test. [‡]Area under curve (AUC) comparisons were performed using DeLong's test. * $P < 0.05$

2.4 特征重要性分析

基于XGBoost模型的特征重要性排序显示,吃糖果频率、父母龋齿情况、睡前刷牙频率为影响北京市5岁儿童新发龋的前三位核心预测因子,其余关键变量依次为牙膏类型、夜奶次数、日常饮水、吃零食频率、含奶入睡习惯、餐后饮水/漱口(图1)。

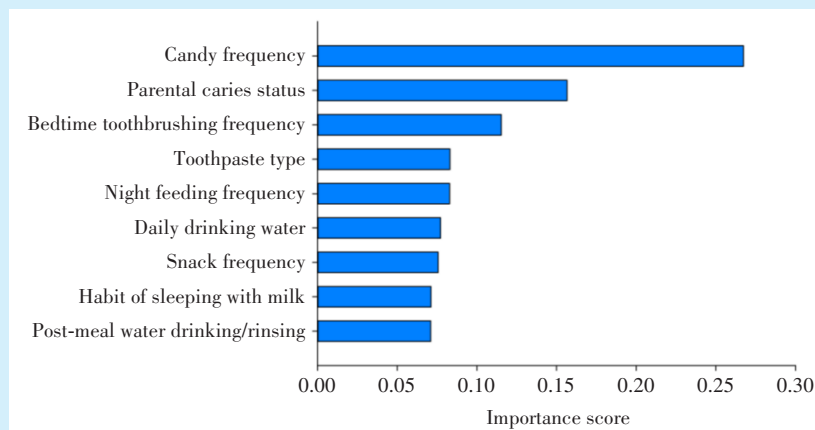
SHAP蜂群图直观呈现了各特征对模型输出的影响方向、效应强度及变量取值特征(图2)。为量化各特征的贡献程度,进一步计算了每个特征的平均绝对SHAP值、SHAP中位数及四分位间距(interquartile range, IQR)(表5)。结合图2与表5可知,睡前刷牙频率的平均绝对SHAP值最高(0.32),SHAP中位数为-0.30(IQR: -0.50, -0.10),且图2中该特征的点大部分分布于负半轴,是最强的保护因素;吃糖果频率和父母龋齿状况在正负两侧均有分布但正向贡献更明显,是主要的风险

升高因素。上述SHAP量化结果与多因素Logistic回归结论高度一致,进一步验证了模型的稳定性与临床可解释性。

2.5 模型校准度与临床效用评估

对性能最优的XGBoost模型进行校准度评估。校准曲线显示(图3),模型预测概率与实际观测概率沿45°对角线分布良好,校准斜率为0.96(95%CI: 0.89~1.03),校准截距为-0.02(95%CI: -0.09~0.06),提示模型预测概率具有较好的校准度。

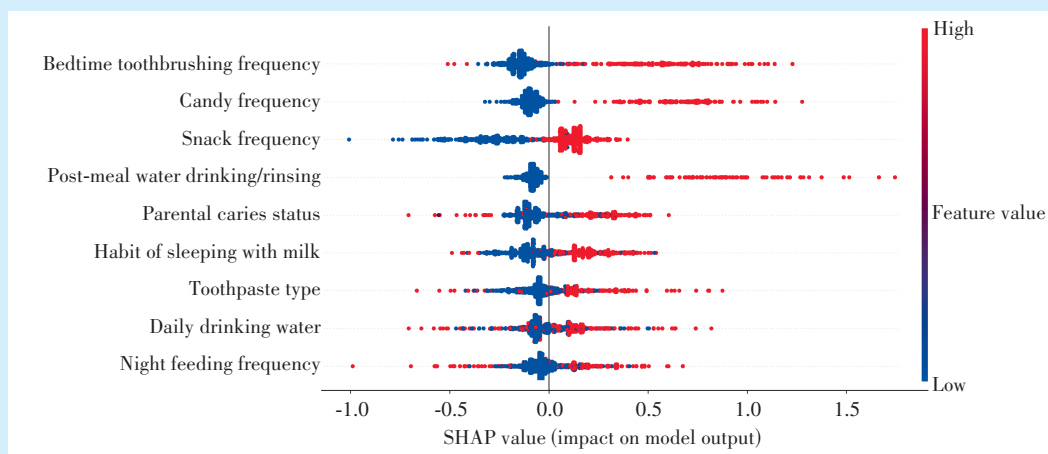
决策曲线分析结果显示(图4),在风险阈值概率为0.15~0.40,XGBoost模型的净获益高于随机森林、LightGBM和决策树模型,且优于“全部干预”及“全部不干预”策略。当阈值概率低于0.15时,“全部干预”策略净获益最高,反映该阈值下干预成本极低;当阈值概率高于0.45时,各模型净获益趋近于0,提示在高阈值下进行干预的临床价值有限。



Feature importance scores were calculated based on the XGBoost algorithm using risk factors screened by multivariate Logistic regression. Candy consumption frequency, parental caries status and bedtime toothbrushing frequency were identified as the top three predictive features for childhood dental caries

Figure 1 Ranking of importance features in the XGBoost model

图1 XGBoost模型中重要特征排名



SHAP (Shapley additive explanations) values quantify the contribution of each variable to individual prediction results. Points distributed on the negative axis represent protective factors against dental caries, while points on the positive axis represent risk factors. The color gradient reflects the variable value, and point density indicates the sample distribution

Figure 2 SHAP beeswarm plot of feature importance for the preschool children's caries risk prediction model

图2 学龄前儿童新发龋风险预测模型的SHAP特征重要性蜂群图

总体而言, XGBoost模型在较宽的风险阈值区间内具有正向净获益, 表明其具备良好的临床适用性。

3 讨论

在2025年12月国家卫生健康委员会印发的《儿童青少年“五健”促进行动计划(2026—2030年)》中明确指出, 要强化龋齿防治健康促进以及做好儿童青少年口腔健康服务; 到2030年5岁儿童患龋得到有效干预。北京市作为全国口腔卫生服务体系与公共卫生管理水平领先的地区, 在

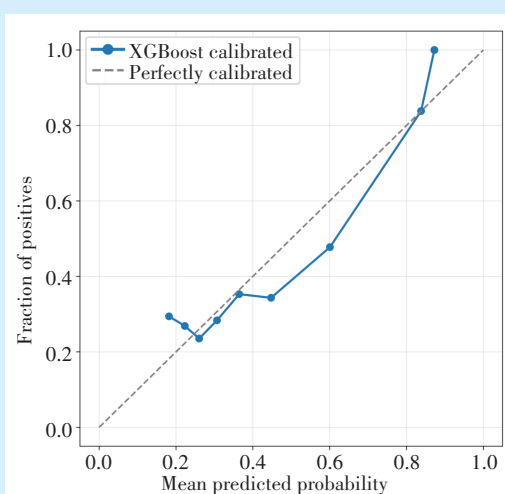
2011年开展了针对3~6岁学龄前儿童氟化泡沫防龋项目, 2021年正式启动0~3岁儿童口腔综合干预项目。即便在资源投入充足、项目覆盖广泛的前提下, 北京市学龄前儿童龋病防控仍面临较大压力, 传统的普适性预防亟须向风险分层、精准预防转型。本研究基于3372名3岁无龋儿童的口腔健康行为数据, 采用独立测试集+5折交叉验证的严格策略, 构建了四种机器学习模型, 其中XGBoost表现最佳, 在独立测试集上AUC达0.806(95%CI: 0.778~0.834), 准确率0.747, 特异度0.852, 提示使

表5 各特征SHAP值分布

Table 5 Distribution of SHAP values for each factor

Factors	Mean absolute SHAP value	SHAP median (<i>IQR</i>)	Risk direction
Bedtime toothbrushing frequency	0.32	-0.30 (-0.50, -0.10)	↓
Candy frequency	0.28	0.05 (-0.15, 0.20)	↑
Snack frequency	0.22	0.02 (-0.12, 0.12)	↑
Post-meal water drinking/rinsing	0.18	-0.04 (-0.18, 0.04)	↓
Parental caries status	0.20	0.05 (-0.08, 0.20)	↑
Habit of sleeping with milk	0.16	0.08 (-0.02, 0.18)	↑
Toothpaste type	0.12	0.08 (0, 0.14)	↑
Daily drinking water	0.14	0.12 (0.04, 0.20)	↑
Night feeding frequency	0.18	0.15 (0.06, 0.24)	↑

SHAP: Shapley additive explanations. *IQR*: interquartile range



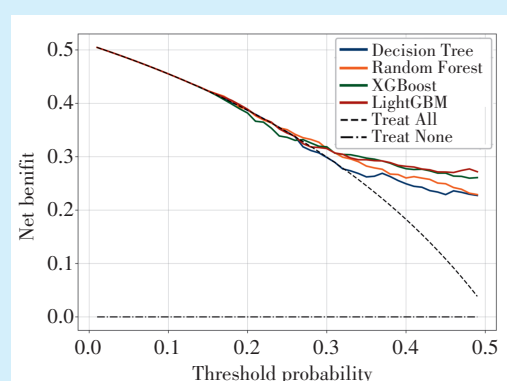
The horizontal axis represents the predicted caries risk probability of the model, and the vertical axis represents the actual observed caries incidence. The 45° diagonal line indicates perfect calibration. The XGBoost model showed favorable calibration performance with a calibration slope of 0.96 (95% *CI*: 0.89-1.03) and an intercept of -0.02 (95% *CI*: -0.09-0.06)

Figure 3 Calibration curve of the XGBoost model

图3 XGBoost模型的校准曲线

用简单、易获取的行为问卷变量可实现中等偏上的风险区分能力,可为基层大规模筛查提供可行工具。

尽管本研究显示 XGBoost 在儿童新发龋预测中综合表现最佳,但不同研究对机器学习模型性能的比较结果不尽相同。陈灵芝等^[15]在苏州 9 岁儿童第一恒磨牙龋病研究中的 AUC 从高到低分别为 XGBoost、随机森林、LightGBM、Logistic 回归,与本研究结果基本一致;Dey 等^[20]同样发现 XGBoost 是表现最佳的算法。严鑫森等^[16]在中国四川 12 岁儿童龋病预测中报告随机森林优于 XGBoost;



The net clinical benefit of each model was calculated across different risk threshold probabilities. Within the threshold range of 0.15 to 0.40, the XGBoost model yielded higher net benefit than the decision tree, Random Forest, LightGBM, and the two extreme strategies (intervention for all / no intervention for all), indicating satisfactory clinical applicability

Figure 4 Decision curve analysis of the four machine learning models

图4 四种机器学习模型的决策曲线分析

Sadegh-Zadeh 等^[21]也发现随机森林模型表现最优。而 Bahammam 等^[22]报告 LightGBM 总体表现最佳。Park 等^[23]基于韩国全国调查数据比较 XGBoost、随机森林、LightGBM 和逻辑回归四种模型,发现四个模型的 AUROC 值介于 0.774 至 0.785 之间,且差异无统计学意义,表明传统逻辑回归与机器学习方法的预测性能可以同样令人满意。模型性能的比较高度依赖于数据源、研究人群、样本量、预测变量选择和评价指标等多重因素,这解释了不同研究和不同算法间呈现性能差异的客观原因。仅凭 AUC 不足以评价模型的临床实用性,而校准曲线和决策曲线可分别评估预测概率的准确性与临床净获益,二者共同构成模型临床转化评价的核心

框架。目前,很多国内外针对龋病预测模型的研究也纳入这两项内容^[24-26](研究临床试验注册号分别为: ChiCTR2300076663^[24]、ChiCTR2500099297^[26], <https://www.chictr.org.cn/index.html>)。本研究在进一步评估模型的数据显示,在风险阈值0.15~0.40范围内, XGBoost的净获益持续高于“全部干预”和其他模型。该阈值区间对应临床“低成本、高收益”的干预策略,如涂氟、口腔卫生指导等,提示本模型具备有效识别真正需要干预的高风险儿童的能力,同时避免大量低风险儿童接受不必要的措施,对优化基层口腔公共卫生资源分配具有实际意义。同时 XGBoost的校准斜率为0.96(95%CI: 0.89~1.03)、校准截距为-0.02(95%CI: -0.09~0.06),证明 XGBoost不仅具有良好区分度,还具有令人满意的临床净获益。

除此之外,本研究也印证了龋病的保护因素和危险因素。本研究发现睡前刷牙是降低龋病风险最强的保护因素。国内外指南均推荐每日两次刷牙、睡前刷牙作为基础预防措施^[27-28]。多项机器学习研究显示,刷牙频率与含氟牙膏使用是最重要的保护变量^[21, 29]。此外,本研究发现使用含氟牙膏、餐后漱口、使用牙线可降低风险,而使用非含氟牙膏、口腔清洁不足显著升高风险,与先前研究结果一致^[30-31]。Toledo等^[32]基于行为学变量构建的预测模型同样支持口腔清洁行为的核心保护作用。多项研究一致证实,游离糖摄入频率是儿童龋病最重要的可干预危险因素^[33-35]。游离糖摄入量与龋病之间存在正向关联,每额外摄入5 g糖就与患龋的概率增加有关^[36]。其他机器学习研究表明,甜食与甜饮料摄入是龋病强预测变量,与本研究结果一致^[30, 37-38]。因此,减少游离糖摄入有助于降低患龋风险。母亲或看护人龋病状况可显著增加儿童患龋风险,其机制主要包括致龋微生物的垂直传播、家庭共享的饮食与口腔卫生习惯、口腔微生态环境的相似性以及遗传易感性的共同作用,本研究也发现父母患龋是儿童新发龋的重要因素^[14, 39-40]。上述结果共同表明,儿童龋病高发并非单一因素所致,而是饮食行为、口腔清洁、家庭成员口腔健康等多因素叠加作用的结果。

本研究存在若干局限性。第一,数据仅来源于参与北京市口腔公共卫生项目的儿童,虽然样本量充足,但单中心设计可能限制结论的推广。北京市口腔保健资源处于全国领先水平,模型在资源匮乏地区的预测效能可能下降。第二,行为

问卷由监护人回顾性填写,难以完全避免回忆倚倚和社会期望倚倚,可能低估某些危险因素与龋病的关联强度。第三,本研究仅在儿童3岁和5岁时进行两次口腔检查,缺少中间时间点的随访数据,无法精确确定龋病发生的具体时间。第四,未纳入社会经济地位、父母教育水平、分娩方式等潜在混杂因素,也未对龋坏牙位、牙面进行细分,可能导致部分效应估计值存在倚倚,但以适度简化换取更高的适用性与可及性,弥补了基层非口腔专业人员难以开展龋病风险评估的短板,更符合基层“早筛查、早分层、早干预”的防控需求。

3岁是乳牙列萌出完整、口腔行为模式快速定型、适宜开展家园共育的关键期,而5岁是国际上评价乳牙龋病流行状况的公认节点^[17]。以3岁行为预测5岁患龋情况,既符合儿童生长发育规律,也与幼儿园健康管理高度契合。本模型可为北京市学龄前儿童龋病家-园-医协同防控提供实用工具。未来将采用多源、多中心数据对模型进行外部验证并进一步提升预测能力。

【Author contributions】 Ren W designed the study, collected and analyzed data, and wrote the original manuscript. Zhao M analyzed the data. Zhang H, Shen AQ and Chen W collected data and conducted statistical analyses. Liu M conceived the study and revised the manuscript. All authors read and approved the final manuscript as submitted.

【Ethics statement】 The study adhered to the tenets of the Declaration of Helsinki and was approved by the Ethics Committee of Beijing Stomatological Hospital, Capital Medical University (No. CMUSH-IRB-KJ-PJ-2023-41), and written informed consent was obtained from all participants.

【Funding】 This study was supported by the grants from Program for Improving Dental Caries Prevention and Treatment Capabilities of Dental Practitioners, Chinese Stomatological Association (No. CSA-ICP2022-07). The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

【Competing interests】 All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: support from the Program for Improving Dental Caries Prevention and Treatment Capabilities of Dental Practitioners, Chinese Stomatological Association (No. CSA-ICP2022-07) for the submitted work; no financial relationships with any organisation that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

【Data availability statement】 Part of the data used in this study comes from the Beijing Center for Oral Public Health Information (<https://www.bjkqjk.cn/bjkq/login.jsp>). Further inquiries can be directed to the corresponding author upon reasonable request.

【Generative AI statement】 The authors declared that generative AI was not used in the creation of this manuscript.

参考文献

- [1] Huang C, Zhu K, Feng Y, et al. Related risk factors of early childhood caries in Zhejiang Province, China during the COVID-19 pandemic[J]. *Front Public Health*, 2022, 10: 879955. doi: 10.3389/fpubh.2022.879955.
- [2] Cao Z, Yu R, Ma J, et al. Association between early childhood caries and anthropometric growth and nutritional status in preschool children: a cross-sectional study[J]. *Front Pediatr*, 2025, 13: 1623666. doi: 10.3389/fped.2025.1623666.
- [3] Chen J, Chen W, Lin L, et al. The prevalence of dental caries and its associated factors among preschool children in Huizhou, China: a cross-sectional study[J]. *Front Oral Health*, 2024, 5: 1461959. doi: 10.3389/froh.2024.1461959.
- [4] Du MQ, Li Z, Jiang H, et al. Dental caries status and its associated factors among 3- to 5-year-old children in China: a national survey [J]. *Chin J Dent Res*, 2018, 21(3): 167-179. doi: 10.3290/j.cjdr.a41076.
- [5] 秦秀荣, 韩靖, 张瑜, 等. 维生素D水平与低龄儿童龋风险的相关性分析[J]. *口腔疾病防治*, 2022, 30(2): 111-116. doi: 10.12016/j.issn.2096-1456.2022.02.006.
Qin XR, Han J, Zhang Y, et al. Correlation between vitamin D level and caries risk in early childhood caries[J]. *J Prev Treat Stomatol Dis*, 2022, 30(2): 111-116. doi: 10.12016/j.issn.2096-1456.2022.02.006.
- [6] 孙鹏, 王宇, 刘敏, 等. 北京市2011—2022年5岁儿童乳牙龋纵向变化分析[J]. *北京口腔医学*, 2024, 32(3): 209-214. doi: 10.20049/j.bjkqyx.1006-673X.2024.03.011.
Sun P, Wang Y, Liu M, et al. Longitudinal changes of deciduous teeth caries in 5-year-old children in Beijing from 2011 to 2022[J]. *Beijing J Stomatol*, 2024, 32(3): 209-214. doi: 10.20049/j.bjkqyx.1006-673X.2024.03.011.
- [7] Boustedt K, Roswall J, Twetman S. Free sugars and early childhood caries development: a prospective cohort study[J]. *Eur Arch Paediatr Dent*, 2022, 23(5): 829-833. doi: 10.1007/s40368-022-00745-3.
- [8] Echeverria MS, Schuch HS, Cenci MS, et al. Trajectories of sugar consumption and dental caries in early childhood[J]. *J Dent Res*, 2022, 101(6): 724-730. doi: 10.1177/00220345211068743.
- [9] Sobiech P, Olczak-Kowalczyk D, Hosey MT, et al. Vitamin D supplementation, characteristics of mastication, and parent-supervised toothbrushing as crucial factors in the prevention of caries in 12- to 36-month-old children[J]. *Nutrients*, 2022, 14(20): 4358. doi: 10.3390/nu14204358.
- [10] Ganesh A, Muthu MS, Padmanabhan R, et al. Association of sleep-time feeding practices with early childhood caries: a case-control study[J]. *Caries Res*, 2022, 56(4): 399-406. doi: 10.1159/000527035.
- [11] Zhu H, Lian L, Zhu K, et al. Guardian factors affecting high prevalence of dental caries in preschool children[J]. *Oral Health Prev Dent*, 2022, 20: 157-164. doi: 10.3290/j.ohpd.b2960227.
- [12] Cagetti MG, Cocco F, Wierichs RJ, et al. Efficacy of HAF tooth-pastes in primary and permanent dentitions. A 2-years triple-blind RCT[J]. *J Dent*, 2022, 121: 104049. doi: 10.1016/j.jdent.2022.104049.
- [13] 马婷, 孟凡琦, 韩婧, 等. 乌鲁木齐市3~5岁儿童龋病相关因素的主成分分析[J]. *口腔疾病防治*, 2023, 31(3): 179-185. doi: 10.12016/j.issn.2096-1456.2023.03.004.
Ma T, Meng FQ, Han J, et al. Principal component analysis of dental caries related factors among 3- to 5-year-old children in Urumqi city[J]. *J Prev Treat Stomatol Dis*, 2023, 31(3): 179-185. doi: 10.12016/j.issn.2096-1456.2023.03.004.
- [14] Fontana M, Eckert GJ, Katz BP, et al. Predicting dental caries in young children in primary health care settings[J]. *J Dent Res*, 2023, 102(9): 988-998. doi: 10.1177/00220345231173585.
- [15] 陈灵芝, 王霞琴, 朱凯飞, 等. 基于机器学习的苏州地区9岁儿童第一恒磨牙龋病预测模型研究[J]. *华西口腔医学杂志*, 2025, 43(6): 871-880. doi: 10.7518/hxkq.2025.2025135.
Chen LZ, Wang XQ, Zhu KF, et al. Machine learning-based prediction model for caries in the first molars of 9-year-old children in Suzhou[J]. *West Chin J Stomatol*, 2025, 43(6): 871-880. doi: 10.7518/hxkq.2025.2025135.
- [16] 严鑫森, 孙桃兰, 卢雨航, 等. 基于机器学习的四川省12岁儿童龋齿预测模型[J]. *华西口腔医学杂志*, 2023, 41(6): 686-693. doi: 10.7518/hxkq.2023.2023124.
Yan XM, Sun TL, Lu YH, et al. Prediction model of dental caries in 12-year-old children in Sichuan Province based on machine learning[J]. *West Chin J Stomatol*, 2023, 41(6): 686-693. doi: 10.7518/hxkq.2023.2023124.
- [17] World Health Organization. Oral health surveys: basic methods [M]. 5th ed. Geneva: World Health Organization, 2013.
- [18] 卜志军, 丁申, 马文欣, 等. 临床预测模型构建中的常用样本量估算方法介绍及实例分析[J]. *现代中医临床*, 2024, 31(6): 32-37. doi: 10.3969/j.issn.2095-6606.2024.06.008.
Bu ZJ, Ding S, Ma WX, et al. Introduction and example analysis of common sample size estimation methods in clinical prediction model construction[J]. *Mod Chin Clin Med*, 2024, 31(6): 32-37. doi: 10.3969/j.issn.2095-6606.2024.06.008.
- [19] Chen H, Lundberg SM, Lee SI. Explaining a series of models by propagating Shapley values[J]. *Nat Commun*, 2022, 13: 4512. doi: 10.1038/s41467-022-31384-3.
- [20] Dey P, Ogwo C, Tellez M. Comparison of traditional regression modeling vs. AI modeling for the prediction of dental caries: a secondary data analysis[J]. *Front Oral Health*, 2024, 5: 1322733. doi: 10.3389/froh.2024.1322733.
- [21] Sadegh-Zadeh SA, Bagheri M, Saadat M. Decoding children dental health risks: a machine learning approach to identifying key influencing factors[J]. *Front Artif Intell*, 2024, 7: 1392597. doi: 10.3389/frai.2024.1392597.
- [22] Bahammam SA. Prediction of dental caries in children through machine learning[J]. *J Clin Pediatr Dent*, 2025, 49(5): 158-167. doi: 10.22514/jocpd.2025.110.
- [23] Park YH, Kim SH, Choi YY. Prediction models of early childhood caries based on machine learning algorithms[J]. *Int J Environ Res Public Health*, 2021, 18(16): 8613. doi: 10.3390/ijerph18168613.
- [24] Zhang Z, Dou Z, Zhang X, et al. Development and validation of a novel public prediction platform for deciduous caries in preschool

- children: an observational study from Northwest China[J]. BMC Pediatr, 2025, 25(1): 600. doi: 10.1186/s12887-025-05763-w.
- [25] Xu Y, Yang L, Wu J, et al. Development and validation of a predictive model for severe caries in 12-year-olds from resource-limited regions: a cross-sectional study in China[J]. BMC Oral Health, 2026, 26(1): 550. doi: 10.1186/s12903-026-07867-w.
- [26] Huang LR, Zhang CZ, Gong ML, et al. Development of a nomogram for root caries risk assessment in a Chinese elderly population[J]. J Dent, 2025, 156: 105624. doi: 10.1016/j.jdent.2025.105624.
- [27] Glenny AM, Walsh T, Iwasaki M, et al. Development of tooth brushing recommendations through professional consensus[J]. Int Dent J, 2024, 74(3): 526-535. doi: 10.1016/j.identj.2023.10.018.
- [28] Zou J, Du Q, Ge L, et al. Expert consensus on early childhood caries management[J]. Int J Oral Sci, 2022, 14(1): 35. doi: 10.1038/s41368-022-00186-0.
- [29] Hasan F, Tantawi ME, Haque F, et al. Early childhood caries risk prediction using machine learning approaches in Bangladesh[J]. BMC Oral Health, 2025, 25(1): 49. doi: 10.1186/s12903-025-05419-2.
- [30] Bai X. Effects of graded prevention intervention combined with minimally invasive restoration on the oral health trajectory of caries-susceptible children[J]. Front Public Health, 2026, 14: 1862313. doi: 10.3389/fpubh.2026.1862313.
- [31] Masaebi F, Salehi M, Ghorbani Z, et al. Exploring dental caries and associated factors in 3-year-old Iranian children: an application of random forest for zero-inflated poisson process[J]. Med J Islam Repub Iran, 2025, 39: 56. doi: 10.47176/mjiri.39.56.
- [32] Toledo Reyes L, Knorst JK, Ortiz FR, et al. Early childhood predictors for dental caries: a machine learning approach[J]. J Dent Res, 2023, 102(9): 999-1006. doi: 10.1177/00220345231170535.
- [33] Ha DH, Nguyen HV, Bell LK, et al. Trajectories of child free sugars intake and dental caries - a population-based birth cohort study [J]. J Dent, 2023, 134: 104559. doi: 10.1016/j.jdent.2023.104559.
- [34] Mututanthri SA, Thoradeniya T, Harris R. Quantitative relationship between free sugars intake and early childhood caries among preschoolers in Sri Lanka[J]. J Dent, 2025, 152: 105441. doi: 10.1016/j.jdent.2024.105441.
- [35] Nakai Y, Mori-Suzuki Y. Impact of dietary patterns on plaque acidogenicity and dental caries in early childhood: a retrospective analysis in Japan[J]. Int J Environ Res Public Health, 2022, 19(12): 7245. doi: 10.3390/ijerph19127245.
- [36] Szpunar SM, Eklund SA, Burt BA. Sugar consumption and caries risk in schoolchildren with low caries experience[J]. Community Dent Oral Epidemiol, 1995, 23(3): 142-146. doi: 10.1111/j.1600-0528.1995.tb00218.x.
- [37] Sun HB, Zhang W, Zhou XB. Risk factors associated with early childhood caries[J]. Chin J Dent Res, 2017, 20(2): 97-104. doi: 10.3290/j.cjdr.a38274.
- [38] Ogwo C, Brown G, Warren J, et al. Predicting longitudinal caries trajectories from childhood to early adulthood[J]. J Dent Res, 2026. doi: 10.1177/00220345261432531.
- [39] Hariyani N, Do LG, Spencer AJ, et al. Maternal caries experience influences offspring's early childhood caries - a birth cohort study [J]. Community Dent Oral Epidemiol, 2020, 48(6): 561-569. doi: 10.1111/cdoe.12568.
- [40] Duan Z, Du C, Tong M, et al. Association between first 2 years' feeding practices and early childhood caries: a birth cohort study in Shanghai[J]. BMC Oral Health, 2025, 25(1): 1531. doi: 10.1186/s12903-025-06794-6.

(编辑 周春华, 杨紫荆)



Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License.

Copyright © 2026 by Editorial Department of Journal of Prevention and Treatment for Stomatological Diseases



官网