International Journal of Women's Health

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Construction and Validation of a Model for Predicting Cervical Intraepithelial Neoplasia Grade II+: A Cross-Sectional Population Study via Machine Learning

Juan He[1,2,*], Kang-Jia Chen[1,3,*], Ya-Xing Fang[1,2], Yu-Feng He[1,2], Lan Xiang[1,2], Xue-Mei Wang[4], Ming-Li Zhou[4], Shu-Guang Zhou[1,3], Jing-Jing Hu[5]

[1]Department of Gynecology, Maternal and Child Medical Center of Anhui Medical University, Hefei, Anhui, 230032, People's Republic of China; [2]Department of Gynecology, Hefei Maternal and Child Health Hospital, Hefei, Anhui, 230001, People's Republic of China; [3]Department of Gynecology, Anhui Provincial Maternity and Child Healthcare Hospital, Hefei, Anhui, 230051, People's Republic of China; [4]Department of Gynecology, Linquan Maternity and Child Healthcare Hospital, Fuyang, 236400, People's Republic of China; [5]Department of Reproduction, The First Affiliated Hospital of Anhui Medical University, Hefei, Anhui, 230032, People's Republic of China

*These authors contributed equally to this work

Correspondence: Shu-Guang Zhou; Jing-Jing Hu, Email zhoushuguang@ahmu.edu.cn; hujingjing@ahmu.edu.cn

Background: Cervical cancer, as the leading malignant tumor among women globally, underscores the critical need for early screening; however, effective models for predicting cervical lesions remain lacking.

Objective: To construct a predictive model for cervical intraepithelial neoplasia II+(CINII+), and to compare the predictive performance of machine learning models integrating thinprep cytologic test (TCT) + human papillomavirus (HPV) testing with clinical data versus TCT combined with traditional clinical data for CIN II+.

Methods: Clinical data from women undergoing cervical cancer screening at Linquan Maternity and Child Healthcare Hospital (2020–2024) were collected, including TCT results, HPV status, cervical pathology, age, sexual history and other clinical data. Ten machine learning algorithms were applied to develop two predictive models: Model 1(TCT+HPV+clinical data) and Model 2(TCT +traditional clinical data). Model performance was evaluated using the area under the receiver operating characteristic curve (AUC), calibration curves and decision curve analysis (DCA).

Results: Multivariate logistic regression analysis showed that HPV positivity, TCT indicates High-Grade Squamous Intraepithelial Lesion(HSIL), colposcopy result indicates a high-grade lesion and the first age of pregnancy as predictors of CINII+. Model 1 (TCT +HPV+clinical data) demonstrated significantly higher predictive efficacy than Model 2(TCT+clinical data), the difference in AUC is statistically significant. (P=0.006 in training set; P=0.035 in testing set).

Conclusion: The TCT+HPV-integrated model outperformed the TCT-only model in predicting CIN II+, supporting the incorporation of HPV testing into routine screening to enhance early diagnostic accuracy.

Keywords: cervical lesions, thinprep cytologic test, TCT, human papillomavirus, HPV, predictive model, machine learning

Introduction

Cervical cancer remains one of the most common gynecologic malignancies threatening women's health, ranking second only to breast cancer in global incidence[1] and the leading cause of cancer-related mortality among females worldwide.[2] It is also the primary cause of cancer-related deaths in developing countries,[3,4] where it accounts for approximately 25% of all female cancers.[5] In recent decades, cervical cancer has emerged as the second most prevalent malignancy among Chinese women, with a trend toward younger age at diagnosis and increasing incidence rates,[6] This disease poses a severe threat to patient survival. Therefore, establishing predictive

Received: 17 July 2025
Accepted: 21 November 2025
Published: 5 December 2025

models for high-grade cervical lesions is critical, particularly in resource-limited settings in developing countries.[7,8] Similar to other malignancies, cervical cancer is a chronic, multifactorial disease driven by genetic and environmental interactions.[9] The pathological progression of cervical lesions follows a hierarchical sequence: Cervical Intraepithelial Neoplasia Grade I (CIN I), progressing to CIN II, then to CIN III, carcinoma in situ (CIS), early invasive carcinoma, and finally invasive carcinoma. The transition from CIN to CIS typically spans 3–8 years.[10,11] Notably, CIN I lesions often regress spontaneously, whereas CIN II+ (including CIN II, CIN III, and invasive cervical cancer) represents a high-risk category requiring intervention.[12] Early detection of precancerous lesions through screening is paramount, as cervical cancer remains the only gynecologic cancer amenable to early diagnosis and curative treatment.[13–16]

Due to advancements in cervical cancer screening technology, the incidence and mortality rates of cervical cancer have shown a downward trend in both developed and developing countries.[17,18] Early screening is key to reducing cervical cancer mortality. Thinprep cytologic test(TCT) using liquid-based methods has been widely applied for identifying abnormal cell morphology, but its sensitivity is limited and it is susceptible to sampling errors and subjective interpretation. In recent years, Human papillomavirus (HPV) nucleic acid testing has become an important supplementary method due to its direct detection of the pathogen, although its combined application with other clinical risk factors remains unclear. Current models for predicting cervical lesion risk primarily rely on single testing methods (such as TCT) or limited clinical variables (such as age and HPV infection history), leading to insufficient model accuracy. Some have proposed data-driven models,[18,19] but these are impractical in resource-poor environments. Therefore, it is crucial to determine whether integrating HPV testing with traditional clinical data can enhance model performance, which is significant for optimizing the screening process.

Existing predictive models for CIN II+ primarily rely solely on HPV as a single data source, exhibit suboptimal performance in specific populations, and lack integration with clinical factors. In contrast, our approach integrating TCT, HPV, and clinical data can enhance clinical discriminative performance, improve generalizability, and optimize clinical interpretability. Furthermore, the innovation of our machine learning framework, compared to traditional statistical models or prior machine learning studies, lies in its incorporation of ten machine learning algorithms, which optimizes algorithm selection.

In this study, we selected variables associated with cervical intraepithelial neoplasia grade II or higher (CIN II+) and performed logistic regression analysis to identify risk factors for CIN II+. Additionally, ten machine learning algorithms were employed to construct predictive models for CIN II+ risk, this can enhance their robustness and generalizability. The aim of this study was to construct a predictive model for CINII+, and to compare the predictive performance of machine learning models integrating TCT + HPV testing with clinical data versus TCT combined with traditional clinical data for CIN II+.

Materials and Methods

Research Design

Clinical data and screening results were collected from women aged 35 to 65 years old undergoing free cervical cancer screening at Linqian County Maternal and Child Health Hospital between January 2020 and December 2024, retrieved through the electronic medical record (EMR) system. Inclusion Criteria: 1. Underwent both liquid-based thin-layer cytology (TCT) and HPV testing, and cytology was interpreted using the Bethesda 2014 nomenclature, and HPV genotyping was performed via PCR assay to detect individual high-risk HPV genotypes; 2. Had complete pathological results; 3. Provided complete clinical data (eg, demographics, reproductive history). Exclusion Criteria: 1. Pregnant or breastfeeding women; 2. Incomplete clinical or pathological data; 3. History of endometrial cancer, ovarian cancer, fallopian tube cancer and primary vaginal cancer; 4. Lost-to-follow-up patients.

Our study included a total of 26 variables. 577 cases were divided into a training set and a testing set in a 7:3 ratio. Therefore, the sample size for the training set was 405 cases and the sample size for the testing set was 172 cases.

Data Collection

Clinical data from participants were collected using a self-developed scale, including: HPV status, TCT status, colposcope results, vaccination status, history of cervical cancer screening, history of smoking, symptoms of contact bleeding, number of

sexual partners, history of CIN, marital status, gravidity, parity, history of oral contraceptives, work situation, educational level, history of drinking, menopausal status, human immunodeficiency virus status, residential area, number of abortions, history of cervical treatment; family history, age, age at first pregnancy, age of menarche, BMI=weight(kg)/[height(m)]^2.

Grouping Criteria

Based on the final pathological diagnosis, participants were categorized into two groups: CIN II- group (managed with follow-up, exhibiting regressive potential); CIN II+ group (requiring interventions due to high-risk progression to cervical cancer).

Variables and Model Development

Twenty-six risk factors were defined as independent variables. Univariate logistic regression analysis was performed with cervical lesion types (determined by pathological results) as the dependent variable. The association strength between each independent variable and the outcome was quantified. Variables with $P < 0.05$ were selected for multivariate logistic regression to explore their joint effects on cervical lesions.

Machine Learning Modeling

Significant variables identified through logistic regression were further modeled using ten machine learning algorithms: logistic regression (LR), support vector machine (SVM), gradient boosting machine(GBM), artificial neural network (ANN), random forest (RF), extreme gradient boosting (XGBoost), k-nearest neighbors (KNN), AdaBoost, LightGBM, and CatBoost.

Model Performance Evaluation

The performance of the model was evaluated around three aspects: discrimination, calibration and clinical utility. In this study, AUC was used to evaluate the model's discrimination ability. Calibration curves were used to determine the degree of agreement between predicted probabilities and observed outcomes. Decision curve analysis (DCA) was used to assess clinical validity. The same three methods are applied to validate the model using the testing set data. All analyses in this study were performed using R software (version 4.4.1). All tests were two-tailed, and the difference was statistically significant at $P < 0.05$.

Comparison of Cervical Lesion Prediction Models

The performance of paired sample models was evaluated using Receiver Operating Characteristic (ROC) curve analysis, with the Area Under the Curve (AUC), sensitivity, and specificity calculated. All statistical tests were two-tailed, and a P-value < 0.05 was considered statistically significant.

Statistical Methods

All analyses were performed using R software (version 4.4.1). Measurement data are presented as ($x \pm s$). For inter-group comparisons, the Wilcoxon rank-sum test is used. Count data are expressed as n (%), and inter-group comparisons are conducted using the chi-square test or Fisher's exact test. The differences between the two models were compared in the same sample set. The AUC values that obeyed the normal distribution were tested by paired sample *T*-test, and the AUC values that did not obey the normal distribution were tested by paired sample nonparametric test. All tests are two-tailed, and a P value of<0.05 is considered statistically significant.

Results

Univariate Analysis of Variable Factors for CINII+

405 cases were included in the training set, divided into CINII-group (n=299) and CINII+group (n=106) based on the degree of cervical lesions. A one-way analysis was conducted on twenty-six variables for both groups, revealing statistically significant differences in TCT, HPV, colposcopy, work status, and age at first pregnancy ($P<0.05$), while no other variables showed statistical significance ($P>0.05$), as detailed in Table 1.

Table 1 Univariate Analysis of Risk Factors for CIN II+

	CINII- Group (n=299)	CIN II+ Group (n=106)	P
HPV (%)			<0.001
No infection	115 (38.5)	13 (12.3)	
16/18 infection	31 (10.4)	48 (45.3)	
Other infections	153 (51.2)	45 (42.5)	
TCT (%)			<0.001
NILM/ASCU-S	237 (79.3)	42 (39.6)	
ASC-H/LSIL	59 (19.7)	34 (32.1)	
HSIL	3 (1.0)	30 (28.3)	
Colposcope (%)			<0.001
No lesion	40 (13.4)	2 (1.9)	
Low-grade lesion	238 (79.6)	49 (46.2)	
High-grade lesion	17 (5.7)	52 (49.1)	
Other lesion	4 (1.3)	3 (2.8)	
Vaccination = Yes (%)	12 (4.0)	2 (1.9)	0.471
History of cervical cancer screening = Yes (%)	139 (46.5)	45 (42.5)	0.546
History of smoking = Yes (%)	8 (2.7)	2 (1.9)	0.932
Symptoms of contact bleeding = Yes (%)	2 (0.7)	1 (0.9)	1.000
Number of sexual partners (%)			0.892
1	283 (94.6)	101 (95.3)	
2	14 (4.7)	4 (3.8)	
3	2 (0.7)	1 (0.9)	
History of CIN = Yes (%)	10 (3.3)	4 (3.8)	1.000
Marital status = singleton (%)	3 (1.0)	0 (0.0)	0.707
Gravidity (%)			0.944
2	48 (16.1)	15 (14.2)	
3	141 (47.2)	49 (46.2)	
4	96 (32.1)	37 (34.9)	
5	14 (4.7)	5 (4.7)	
Parity (%)			0.360
1	112 (37.5)	36 (34.0)	
2	166 (55.5)	58 (54.7)	
3	21 (7.0)	12 (11.3)	
History of oral contraceptives = Yes (%)	17 (5.7)	4 (3.8)	0.611
Work situation = Yes (%)	156 (52.2)	68 (64.2)	0.044
Educational level (%)			0.297
Middle school and below	129 (43.1)	52 (49.1)	
High schooler secondary school	144 (48.2)	42 (39.6)	
College or Undergraduate	26 (8.7)	12 (11.3)	
History of drinking = Yes (%)	157 (52.5)	52 (49.1)	0.619
Menopausal = Yes (%)	18 (6.0)	5 (4.7)	0.800
HIV status = Yes (%)	133 (44.5)	53 (50.0)	0.386
Residential area = city (%)	148 (49.5)	48 (45.3)	0.527
Number of abortions (%)			0.808
0	101 (33.8)	37 (34.9)	
1	89 (29.8)	34 (32.1)	
2	109 (36.5)	35 (33.0)	
History of cervical treatment = Yes (%)	146 (48.8)	59 (55.7)	0.273
Family history = Yes (%)	6 (2.0)	0 (0.0)	0.317
Age (mean (SD))	49.67 (8.05)	50.58 (7.81)	0.309
Age at first pregnancy (mean (SD))	24.85 (3.51)	25.73 (3.54)	0.029

(*Continued*)

Table 1 (Continued).

	CINII- Group (n=299)	CIN II+ Group (n=106)	P
Age of menarche (mean (SD))	12.53 (1.03)	12.76 (1.18)	0.056
BMI (mean (SD))	21.60 (2.83)	21.12 (2.80)	0.134

Notes: p<0.05 indicates that it is statistically significant.
Abbreviations: CIN, cervical intraepithelial neoplasia; HPV, human papillomavirus; TCT, thinPrep cytologic test; HIV, human immunodeficiency virus; BMI, body mass index; SD, standard deviation.

Predictive Model Development

For the twenty-six included variables, a stepwise logistic regression method was used to screen model variables. Univariate logistic regression showed that TCT, HPV, colposcopy, work situation and age at first pregnancy were risk factors for CINII+. Multivariate logistic regression of the univariate results identified HPV positivity, TCT indicates HSIL, colposcopy result indicates a high-grade lesion and the first age of pregnancy as independent risk factors for CINII+, as detailed in Table 2. Ten machine learning methods were used to build models using these four variables. Additionally, HPV was excluded, models were established for TCT, colposcopy and age at first pregnancy to evaluate the added value of HPV in cervical lesion screening.

Model Evaluation of TCT+HPV Combined Clinical Data

The training set was evaluated in terms of discrimination, calibration and clinical utility. The AUC value is used to evaluate the discrimination ability of the predictive model by examining the occurrence of CIN II +. The AUC values for ten machine learning models are as follows: 0.804, 95% CI: 0.752–0.855; 0.771, 95% CI: 0.712–0.829; 0.866, 95% CI: 0.826–0.906; 0.839, 95% CI: 0.793–0.844; 0.827, 95% CI: 0.781–0.872; 0.834, 95% CI: 0.788–0.880; 0.914, 95% CI: 0.884–0.944; 0.752, 95% CI: 0.702–0.802; 0.957, 95% CI: 0.937–976; 0.867, and 95% CI: 0.825–0.908. The sensitivities of the ROC curves for the ten machine learning models are respectively 0.547,0.613,0.830,0.821,0.660,0.830,0.840,0.557,0.840,0.764. And the specificities of the ROC curves for the ten machine learning models are respectively 0.936,0.890,0.729,0.696,0.993,0.692,0.826,0.933, 0.916,0.819. (see Table 3). The results show that the line graph has good distinguishing and predictive value, capable of accurately identifying individuals with or without CIN II +. The calibration plot of the model demonstrates excellent predictive accuracy between actual and predicted probabilities. The clinical utility of the model was evaluated using the DCA curve. The decision curve indicates that the net benefit of the predictive model is significantly higher than in both extreme scenarios, suggesting that the predictive model has a higher net benefit. The results are shown in Figure 1.

The testing set was evaluated in terms of discrimination, calibration and clinical utility. The AUC value is used to evaluate the discrimination ability of the predictive model by examining the occurrence of CIN II +. The AUC values for ten machine learning models are as follows: 0.807, 95% CI: 0.735–0.880; 0.768, 95% CI: 0.683–0.854; 0.817, 95% CI: 0.748–0.887; 0.835, 95% CI: 0.774–0.897; 0.680, 95% CI: 0.598–0.763; 0.823, 95% CI: 0.756–0.891; 0.835, 95% CI: 0.768–0.901; 0.696, 95% CI: 0.617–0.775; 0.771, 95% CI: 0.691–0.852; 0.834, and 95% CI: 0.768–0.899. The sensitivities of the ROC curves for the ten machine learning models are respectively 0.733,0.733,0.800,0.822,0.511,0.800,0.844,0.467,0.733,0.911. And the specificities of the ROC curves for the ten machine learning models are respectively 0.748,0.732,0.677,0.709,0.835,0.717,0.693,0.913, 0.732,0.598.(see Table 3). The results show that the line graph has good distinguishing and predictive value, capable of accurately identifying individuals with or without CIN II +. The calibration plot of the model demonstrates excellent predictive accuracy between actual and predicted probabilities. The clinical utility of the model was evaluated using the DCA curve. The decision curve indicates that the net benefit of the predictive model is significantly higher than in both extreme scenarios, suggesting that the predictive model has a higher net benefit. The results are shown in Figure 2.

Model Evaluation of TCT Combined Clinical Data

The training set was evaluated in terms of discrimination, calibration and clinical utility. The AUC value is used to evaluate the discrimination ability of the predictive model by examining the occurrence of CIN II +. The AUC values for ten machine learning models are as follows: 0.793, 95% CI: 0.739–0.847; 0.794, 95% CI: 0.740–0.848; 0.807, 95% CI: 0.753–0.861; 0.768, 95% CI: 0.709–0.826; 0.775, 95% CI: 0.727–0.823; 0.800, 95% CI: 0.745–0.856; 0.828, 95% CI:

https://doi.org/10.2147/IJWH.S554552

Table 2 Logistic Regression Analysis of Risk Factors for CIN II+

	Desc	CINII- Group (n=299)	CIN II+ Group (n=106)	OR (Univariable)	OR(Multivariable)
HPV	No infection	115 (38.5%)	13 (12.3%)		
	16/18 infection	31 (10.4%)	48 (45.3%)	13.70(6.60–28.42, p<0.001)	7.25 (3.04–17.30, p<0.001)
	Other infections	153 (51.2%)	45 (42.5%)	2.60 (1.34–5.05, p=0.005)	2.45 (1.13–5.32, p=0.024)
TCT	NILM/ASCU-S	237 (79.3%)	42 (39.6%)		
	ASCU-H/LSIL	59 (19.7%)	34 (32.1%)	3.25 (1.91–5.55, p<0.001)	1.59 (0.83–3.04, p=0.166)
	HSIL	3 (1%)	30 (28.3%)	56.43(16.48–193.27, p<0.001)	17.16(4.47–65.82, p<0.001)
Colposcope	No lesion	40 (13.4%)	2 (1.9%)		
	Low-grade lesion	238 (79.6%)	49 (46.2%)	4.12(0.96–17.61, p=0.056)	2.59 (0.59–11.48, p=0.209)
	High-grade lesion	17 (5.7%)	52 (49.1%)	61.18(13.35–280.26, p<0.001)	16.64(3.32–83.34, p<0.001)
	Other lesion	4 (1.3%)	3 (2.8%)	15.00(1.91–118.08, p=0.010)	3.10 (0.22–44.44, p=0.404)
Vaccination	No	287 (96%)	104 (98.1%)		
	Yes	12 (4%)	2 (1.9%)	0.46 (0.10–2.09, p=0.315)	
History of cervical cancer screening	No	160 (53.5%)	61 (57.5%)		
	Yes	139 (46.5%)	45 (42.5%)	0.85 (0.54–1.33, p=0.474)	
History of smoking	No	291 (97.3%)	104 (98.1%)		
	Yes	8 (2.7%)	2 (1.9%)	0.70 (0.15–3.35, p=0.655)	
Symptoms of contact bleeding	No	297 (99.3%)	105 (99.1%)		
	Yes	2 (0.7%)	1 (0.9%)	1.41(0.13–15.76, p=0.778)	
Number of sexual partners	1	283 (94.6%)	101 (95.3%)		
	2	14 (4.7%)	4 (3.8%)	0.80 (0.26–2.49, p=0.701)	
	3	2 (0.7%)	1 (0.9%)	1.40(0.13–15.62, p=0.784)	
History of CIN	No	289 (96.7%)	102 (96.2%)		
	Yes	10 (3.3%)	4 (3.8%)	1.13 (0.35–3.69, p=0.836)	
Marital status	Married	296 (99%)	106 (100%)		
	Singleton	3 (1%)	0 (0%)	0.00 (0.00-Inf, p=0.986)	
Gravidity	2	48 (16.1%)	15 (14.2%)		
	3	141 (47.2%)	49 (46.2%)	1.11 (0.57–2.16, p=0.754)	
	4	96 (32.1%)	37 (34.9%)	1.23 (0.62–2.47, p=0.553)	
	5	14 (4.7%)	5 (4.7%)	1.14 (0.35–3.70, p=0.824)	
Parity	1	112 (37.5%)	36 (34%)		
	2	166 (55.5%)	58 (54.7%)	1.09 (0.67–1.76, p=0.733)	
	3	21 (7%)	12 (11.3%)	1.78 (0.80–3.97, p=0.160)	
History of oral contraceptives	No	282 (94.3%)	102 (96.2%)		
	Yes	17 (5.7%)	4 (3.8%)	0.65 (0.21–1.98, p=0.449)	
Work situation	No	143 (47.8%)	38 (35.8%)		
	Yes	156 (52.2%)	68 (64.2%)	1.64 (1.04–2.59, p=0.034)	1.26 (0.71–2.26, p=0.432)
Educational level	Middle school and below	129 (43.1%)	52 (49.1%)		
	High schooler secondary school	144 (48.2%)	42 (39.6%)	0.72 (0.45–1.16, p=0.178)	
	College or Undergraduate	26 (8.7%)	12 (11.3%)	1.14 (0.54–2.44, p=0.726)	
History of drinking	No	142 (47.5%)	54 (50.9%)		
	Yes	157 (52.5%)	52 (49.1%)	0.87 (0.56–1.36, p=0.541)	

Menopausal	No	281 (94%)	101 (95.3%)		
	Yes	18 (6%)	5 (4.7%)	0.77 (0.28–2.14, p=0.619)	
HIV status	No	166 (55.5%)	53 (50%)		
	Yes	133 (44.5%)	53 (50%)	1.25 (0.80–1.95, p=0.328)	
Residential area	Countryside	151 (50.5%)	58 (54.7%)		
	City	148 (49.5%)	48 (45.3%)	0.84 (0.54–1.32, p=0.456)	
Number of abortions	0	101 (33.8%)	37 (34.9%)		
	1	89 (29.8%)	34 (32.1%)	1.04 (0.60–1.80, p=0.880)	
	2	109 (36.5%)	35 (33%)	0.88 (0.51–1.50, p=0.630)	
History of cervical treatment	No	153 (51.2%)	47 (44.3%)		
	Yes	146 (48.8%)	59 (55.7%)	1.32 (0.84–2.05, p=0.227)	
Family history	No	293 (98%)	106 (100%)		
	Yes	6 (2%)	0 (0%)	0.00 (0.00-Inf, p=0.981)	
Age	Mean ± SD	49.7 ± 8.0	50.6 ± 7.8	1.01 (0.99–1.04, p=0.308)	
Age at first pregnancy	Mean ± SD	24.9 ± 3.5	25.7 ± 3.5	1.07 (1.01–1.14, p=0.029)	1.09 (1.00–1.19, p=0.039)
Age of menarche	Mean ± SD	12.5 ± 1.0	12.8 ± 1.2	1.22 (0.99–1.49, p=0.057)	
BMI	Mean ± SD	21.6 ± 2.8	21.1 ± 2.8	0.94 (0.87–1.02, p=0.134)	

Notes: Unbolded values in parentheses are 95% CI; Bolded values in parentheses are percentage. p<0.05 indicates that it is statistically significant.

Abbreviations: CIN, cervical intraepithelial neoplasia; HPV, human papillomavirus; TCT, thinPrep cytologic test; HIV, human immunodeficiency virus; BMI, body mass index; SD, standard deviation.

https://doi.org/10.2147/IJWH.S554552

Table 3 Results of Training and Validation Analysis

Data	Methods	Accuracy	Sensitivity	Specificity	Precision
Train (TCT+HPV models)	Logistic	0.835	0.547	0.936	0.753
	SVM	0.817	0.613	0.890	0.663
	GBM	0.756	0.830	0.729	0.521
	NeuralNetwork	0.728	0.821	0.696	0.489
	RandomForest	0.906	0.660	0.993	0.972
	Xgboost	0.728	0.830	0.692	0.489
	KNN	0.830	0.840	0.826	0.631
	Adaboost	0.835	0.557	0.933	0.747
	LightGBM	0.896	0.840	0.916	0.781
	CatBoost	0.805	0.764	0.819	0.600
Test (TCT+HPV models)	Logistic	0.744	0.733	0.748	0.508
	SVM	0.733	0.733	0.732	0.493
	GBM	0.709	0.800	0.677	0.468
	NeuralNetwork	0.738	0.822	0.709	0.500
	RandomForest	0.750	0.511	0.835	0.523
	Xgboost	0.738	0.800	0.717	0.500
	KNN	0.733	0.844	0.693	0.494
	Adaboost	0.797	0.467	0.913	0.656
	LightGBM	0.733	0.733	0.732	0.493
	CatBoost	0.680	0.911	0.598	0.446
Train (TCT models)	Logistic	0.830	0.557	0.926	0.728
	SVM	0.827	0.585	0.913	0.705
	GBM	0.802	0.642	0.860	0.618
	NeuralNetwork	0.832	0.519	0.943	0.764
	RandomForest	0.874	0.566	0.983	0.923
	Xgboost	0.82	0.604	0.896	0.674
	KNN	0.854	0.632	0.933	0.770
	Adaboost	0.825	0.491	0.943	0.754
	LightGBM	0.852	0.689	0.910	0.730
	CatBoost	0.830	0.613	0.906	0.699
Test (TCT models)	Logistic	0.715	0.778	0.693	0.473
	SVM	0.715	0.778	0.693	0.473
	GBM	0.727	0.733	0.724	0.485
	NeuralNetwork	0.733	0.733	0.732	0.493
	RandomForest	0.767	0.489	0.866	0.564
	Xgboost	0.715	0.800	0.685	0.474
	KNN	0.733	0.600	0.780	0.491
	Adaboost	0.709	0.578	0.756	0.456
	LightGBM	0.773	0.578	0.843	0.565
	CatBoost	0.802	0.600	0.874	0.628

0.776–0.879; 0.761, 95% CI: 0.706–0.816; 0.885, 95% CI: 0.846–923; 0.809, 95% CI: 0.756–0.861. The sensitivities of the ROC curves for the ten machine learning models are respectively 0.557, 0.585, 0.642, 0.519, 0.566, 0.604,0.632, 0.491, 0.689, 0.613. And the specificities of the ROC curves for the ten machine learning models are respectively 0.926, 0.913, 0.860, 0.943, 0.983, 0.896, 0.933, 0.943, 0.910, 0.906 (see Table 3). The results show that the line graph has good distinguishing and predictive value, capable of accurately identifying individuals with or without CIN II +. The calibration plot of the model demonstrates excellent predictive accuracy between actual and predicted probabilities. The clinical utility of the model was evaluated using the DCA curve. The decision curve indicates that the net benefit of the predictive model is significantly higher than in both extreme scenarios, suggesting that the predictive model has a higher net benefit. The results are shown in Figure 3.

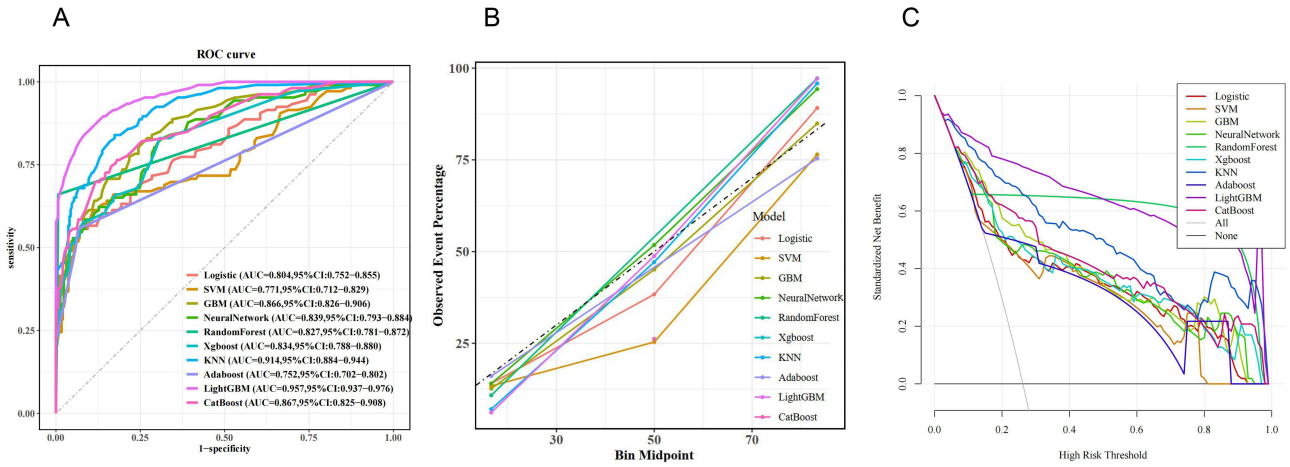

Figure 1 Training Sets Performance Evaluation for HPV+TCT models.(**A**) ROC curve (**B**) calibration curve (**C**) DCA curve.

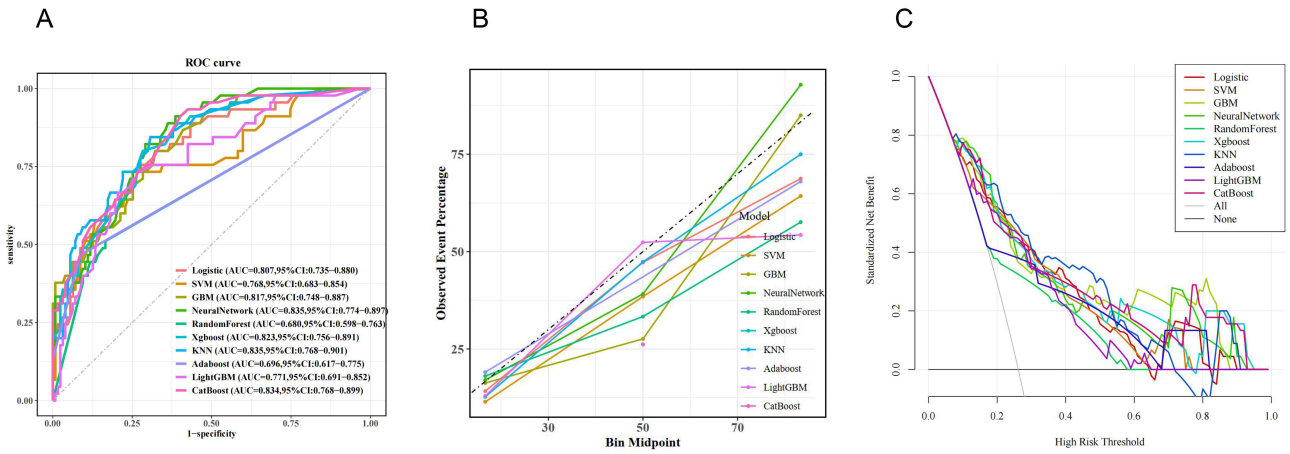

Figure 2 Testing Sets Performance Evaluation for HPV+TCT models.(**A**) ROC curve (**B**) calibration curve (**C**) DCA curve.

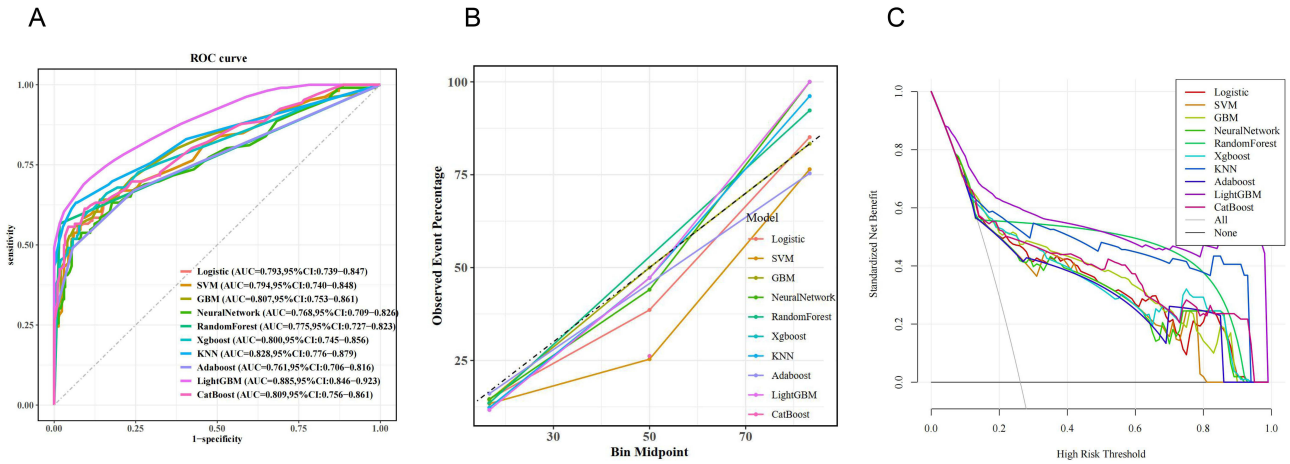

Figure 3 Training Sets Performance Evaluation for TCT models.(**A**) ROC curve (**B**) calibration curve (**C**) DCA curve.

https://doi.org/10.2147/IJWH.S554552

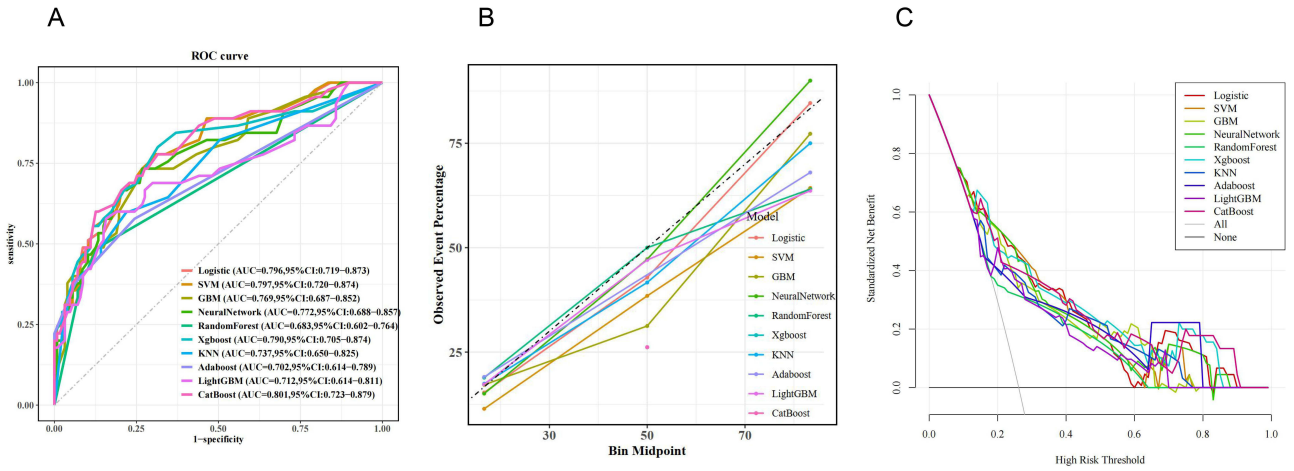

Figure 4 Testing Sets Performance Evaluation for TCT models.(**A**) ROC curve (**B**) calibration curve (**C**) DCA curve.

The testing set was evaluated in terms of discrimination, calibration and clinical utility. The AUC value is used to evaluate the discrimination ability of the predictive model by examining the occurrence of CIN II +. The AUC values for ten machine learning models are as follows: 0.796, 95% CI: 0.719–0.873; 0.797, 95% CI: 0.720–0.874; 0.769, 95% CI: 0.687–0.852; 0.772, 95% CI: 0.688–0.857; 0.683, 95% CI: 0.602–0.764; 0.790, 95% CI: 0.705–0.874; 0.737, 95% CI: 0.650–0.825; 0.702, 95% CI: 0.614–0.789; 0.712, 95% CI: 0.614–811; 0.801, and 95% CI: 0.723–0.879. The sensitivities of the ROC curves for the ten machine learning models are respectively 0.778, 0.778, 0.733, 0.733, 0.489, 0.800, 0.600, 0.578, 0.578, 0.600. And the specificities of the ROC curves for the ten machine learning models are respectively 0.693, 0.693, 0.724, 0.732, 0.866, 0.685, 0.780, 0.756, 0.843, 0.874 (see Table 3). The results show that the line graph has good distinguishing and predictive value, capable of accurately identifying individuals with or without CIN II +. The calibration plot of the model demonstrates excellent predictive accuracy between actual and predicted probabilities. The clinical utility of the model was evaluated using the DCA curve. The decision curve indicates that the net benefit of the predictive model is significantly higher than in both extreme scenarios, suggesting that the predictive model has a higher net benefit. The results are shown in Figure 4.

Model Performance Comparison

Model performance was evaluated by comparing the area under the receiver operating characteristic curve (AUC) using paired samples. Training Set: The AUC of the TCT+HPV+clinical data model (0.84 ± 0.06) was significantly higher than that of the TCT+clinical data model (0.80 ± 0.04, P = 0.006). Testing Set: The AUC of the TCT+HPV+clinical data model[Median (IQR): 0.81 (0.75–0.83)] was significantly higher than that of the TCT+clinical data model [Median (IQR): 0.77 (0.71–0.80), P = 0.035], as shown in Table 4 and Figure 5.

Discussion

Cervical cancer is a globally recognized public health issue. This type of cancer can lead to loss of life, but it is preventable if detected and treated early.[20] Previous studies have shown that sexually transmitted infections, hormonal influences, genetics, and participant factors (risky behaviors, drug abuse and alcohol consumption) are risk factors for

Table 4 Comparison of AUC Values Between Models

	TCT+HPV Models	TCT Models	P
Training set AUC (Mean±SD)	0.84±0.06	0.80±0.04	0.006
Testing set AUC (Median, Q1~Q3)	0.81,0.75~0.83	0.77,0.71~0.80	0.035

Notes: p<0.05 indicates that it is statistically significant.
Abbreviations: AUC, area under the curve; SD, standard deviation.

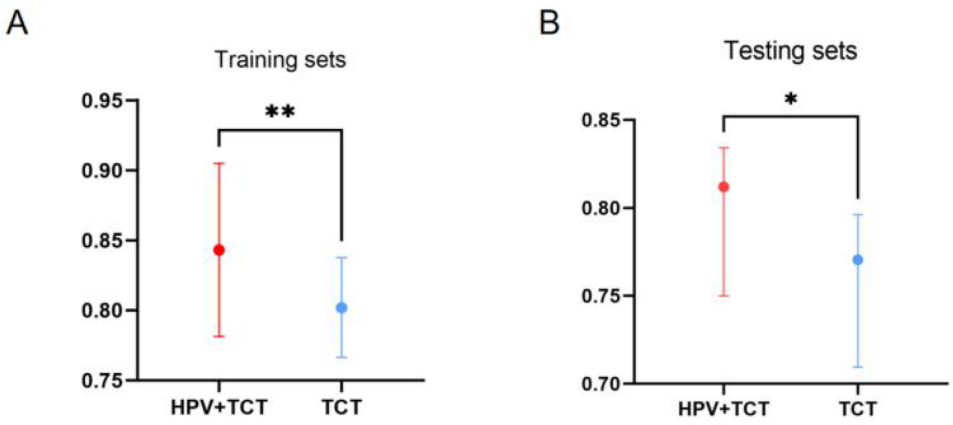

Figure 5 Comparison of AUC values between models.**indicates P<0.01;*indicates P<0.05. (**A**) Comparison of training sets. (**B**) Comparison of testing sets.

cervical cancer in developing countries. To date, predicting cervical lesions remains a challenge. Previous research has primarily focused on exploring risk factors for cervical lesions, with fewer studies on predictive models for cervical lesions. Binyue Sheng et al[21] developed a predictive model for high-grade cervical lesions that includes TCT results, HPV results, acetowhite epithelium, abnormal blood vessels and mosaic. The predictive indicators of this model are not routine examination items, which makes its application very difficult. Our study provides a predictive method to address this issue. Our model integrates routine clinical data, which can be easily obtained from electronic medical records. In addition, we investigated whether HPV adds value to the prediction of high-grade cervical lesions. Our study identified TCT results, HPV results, colposcopy results, and age at first pregnancy as variables for establishing the predictive model. The AUC values of the training sets of the ten machine learning models we developed ranged from 0.752 to 0.957, indicating high accuracy and specificity of our models. ThinPrep cytology test (TCT) is a common cytological screening method for cervical cancer.[22,23] HPV infection is a necessary condition for cervical lesions. HPV strains can be classified as low-risk or high-risk based on their propensity to cause cervical cancer.[24,25] A substantial amount of evidence shows that the rates of HPV infection and positive TCT results increase with the severity of cervical lesions. Previous studies have shown that 96% of epithelial lesions or malignancies (NILM) are negative. Abnormal results include atypical squamous cells of undetermined significance (ASC-US), low-grade squamous intraepithelial lesion (LSIL), high-grade squamous intraepithelial lesion (HSIL), atypical squamous cells—cannot exclude high-grade squamous intraepithelial lesion (ASC-H), and squamous cell carcinoma (SCC).[26] Our study found that colposcopy results are an independent risk factor for CIN II+, which is consistent with previous research. Priscila Thais et al[27] showed that the size of the lesion detected by colposcopy is associated with the severity of the lesion. Colposcopy involves examining the cervix under magnification to detect suspicious areas of cervical intraepithelial neoplasia (CIN) and invasive cancer in real-time.[28] It can also measure the size of cervical lesions and is useful for predicting the likelihood of a negative biopsy result in the future.[29] Mantoani et al[30] also found that in high-grade cervical lesions, the size of the lesion detected by colposcopy is associated with systemic inflammatory and immune responses.

Our study found that CINII+ group had a larger age of first pregnancy than CINII-group, which may be because that advanced maternal age at the time of first pregnancy may be associated with prolonged exposure to estrogen priming, which accelerates cervical epithelial maturation. Mature epithelial cells exhibit decreased regenerative capacity, affecting immune clearance of HPV infection.[31]

In addition, the core finding of this study is that the predictive performance of the combined TCT + HPV model significantly outperformed the TCT + clinical data model (training set comparison: $p = 0.006$; testing set comparison: $p = 0.035$). This confirms that HPV testing is a core variable in cervical lesion prediction models, and its inclusion significantly enhances the model's ability to identify high-risk populations. Previous studies have shown that cytology test results are generally considered to reflect the current risk of CIN, while HPV status reflects both the current and future risks of CIN.[32,33] The AUC values of the traditional TCT + clinical data model's training set ranged from 0.761 to

https://doi.org/10.2147/IJWH.S554552

0.885, and the testing set's AUC values ranged from 0.683 to 0.801. This indicates that even without incorporating HPV data, integrating clinical variables still holds some value. All in all, earlier detection and treatment of smaller, less advanced lesions may lead to improved surgical outcomes and reduced recurrence rates, while the decreasing prevalence of high-risk HPV types in vaccinated populations which may alter the overall burden of CIN.

Our model was constructed using machine learning methods with R, which offers significant advantages over traditional logistic regression models, especially in handling complex data, capturing non-linear relationships, and enhancing predictive performance. The predictive model we developed achieves predictive performance comparable to or even better than that of previous prospective cohorts. This study used a 7:3 ratio to divide the data into training and validation sets, with the testing set used for internal validation. The results showed that the model's AUC values in internal validation ranged from 0.680 to 0.835, and the AUC values of the TCT-only + clinical data model ranged from 0.683 to 0.801. This indicates that the model still has good discrimination and calibration correction ability, and also shows that the model remains well-calibrated. DCA is a model evaluation method based on continuous potential risk thresholds (x-axis), while the net benefit (y-axis) of using the model for patient risk stratification demonstrates the model's clinical utility. This study shows high net benefit of the model in clinical application. In resource-limited settings for HPV testing or regions where HPV detection methods remain unavailable, a cervical lesion prediction model tailored to such scenarios can be developed using an integrated modeling strategy that combines the thinprep cytologic test (TCT) with patient clinical characteristics. Such models leverage TCT results as the core data foundation, integrate synchronous clinical information, and enable quantitative risk assessment and prospective prediction of cervical lesions through multi-dimensional information fusion analysis. Ultimately, this approach provides a evidence-based scientific framework for cervical cancer screening and early intervention in contexts where HPV testing is inaccessible. In clinical practice, we can provide lifestyle interventions and targeted health education to patients. Additionally, for patients with a history of CIN II+, the model can integrate dynamic data such as HPV clearance status and follow-up TCT results to predict the risk of recurrence, which is a direction for future research.

This study has certain limitations. First, a major limitation of our research is that it was conducted at a single center, which may affect the generalizability and validity of the model application. Second, external validation of the model is needed to assess its value in clinical practice. Third, our study requires a large amount of prospective data to improve the accuracy of the model's predictions. To address these, we will conduct multi-center external validation with partner institutions, evaluating the model's performance in diverse populations to enhance real-world applicability.

Conclusion

In this study, it was found that HPV positivity, TCT indicates HSIL, colposcopy result indicates a high-grade lesion and the first age of pregnancy were risk factors for CIN II +. The TCT+HPV-integrated model outperformed the TCT-only model in predicting CIN II+, supporting the incorporation of HPV testing into routine screening to enhance early diagnostic accuracy. According to the findings, we constructed a prediction model based on the information of routine clinical data that could accurately predict the risk of CIN II + patients, which may provide a reference for clinicians to identify high-risk groups early.

Abbreviations

CIN, cervical intraepithelial neoplasias; HPV, human papilloma virus; TCT, ThinPrep cytological test; NILM, negative for intraepithelial lesion or malignancy; ASC-US, atypical cytology of undetermined significance; LSIL, low-grade squamous intraepithelial lesion; HSIL, high-grade squamous intraepithelial lesion; ASC-H, atypical squamous cells cannot exclude high-grade; ROC, receiver operating characteristic; AUC, area under the curve; DCA, decision curve analysis.

Data Sharing Statement

Data will be available upon request from the corresponding author, Professor Zhou.

Ethics Approval and Consent to Participate

This was a cross-sectional study of routinely collected clinical data, and exemption from informed consent was approved by the Ethics Review Committee of Linquan County Maternal and Child Health Hospital (approval number PJ-KY20250523-1). All experimental protocols were approved by the the Ethics Review Committee of Linquan County Maternal and Child Health Hospital and all methods were carried out in accordance with the Declarations of Helsinki. All patient data used in this study were anonymized prior to analysis, with direct identifiers removed and replaced by unique codes. Data access was restricted to authorized research personnel.

Acknowledgments

The authors wish to thank all of the staff members at the Department of Obstetrics and Gynaecology, Linquan County Women and Children's Hospital. The authors wish to thank all the staff members at the Department of Obstetrics and Gynaecology, Linquan County Women and Children's Hospital for their strong support of this study.

Funding

This work was supported by Fuyang Health Research Projects (FYZC2024 −038; FYZC2024-031), the Natural Science Foundation of Higher Education Institutions of Anhui Province (grant number KJ2021A0352), the Research Fund Project of Anhui Medical University (2020xkj236), the Applied Medicine Research Project of Hefei Health Commission (HWKJ2019-172-14; Hwk2023zd001), and Anhui Provincial Health Research Project(AHWJ2023BAa10009).

Disclosure

The authors declare no conflicts of interest in this work.

References

1. Lei J, Ploner A, Elfström KM, et al. HPV vaccination and the risk of invasive cervical cancer. New Engl J Med. 2020;383(14):1340–1348. doi:10.1056/nejmoa1917338
2. Jin XW, Cash J, Kennedy AW, Kennedy AW. Human papillomavirus typing and the reduction of cervical cancer risk. Cleve Clin J Med. 1999;66 (9):533–539. doi:10.3949/ccjm.66.9.533
3. Bedell SL, Goldstein LS, Goldstein AR, Goldstein AT. Cervical cancer screening: past, present, and future. Sexual Med Rev. 2020;8(1):28–37. doi:10.1016/j.sxmr.2019.09.005
4. Shrestha AD, Neupane D, Vedsted P, Kallestrup P. Cervical cancer prevalence, incidence and mortality in low and middle income countries: a systematic review. Asian Pac J Cancer Prev. 2018;19(2):319–324. doi:10.22034/APJCP.2018.19.2.319
5. Harro CD, Pang YY, Roden RB, et al. Safety and immunogenicity trial in adult volunteers of a human papillomavirus 16 L1 virus-like particle vaccine. J National Cancer Inst. 2001;93(4):284–292. doi:10.1093/jnci/93.4.284
6. Maver PJ, Poljak M. Primary HPV-based cervical cancer screening in Europe: implementation status, challenges, and future plans. Clin Microbiol Infect. 2020;26(5):579–583. doi:10.1016/j.cmi.2019.09.006
7. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA. 2011;61(2):69–90. doi:10.3322/caac.20107
8. Yang S, Wu Y, Wang S, et al. HPV-related methylation-based reclassification and risk stratification of cervical cancer. Mol oncol. 2020;14 (9):2124–2141. doi:10.1002/1878-0261.12709
9. Hu Z, Ma D. The precision prevention and therapy of HPV-related cervical cancer: new concepts and clinical implications. Cancer Med. 2018;7 (10):5217–5236. doi:10.1002/cam4.1501
10. Miao Z, Shen J, Zhang FQ, et al. The relationship between HPV integration and prognosis of cervical cancer. Zhonghua zhong liu za zhi. 2020;42 (12):1014–1019. doi:10.3760/cma.j.cn112152-20191031-00705
11. Khan AA, Abuderman AA, Ashraf MT, Khan Z. Protein-protein interactions of HPV-Chlamydia trachomatis-human and their potential in cervical cancer. Future Microbiol. 2020;15:509–520. doi:10.2217/fmb-2019-0242
12. Wang M, Hou B, Wang X, et al. Diagnostic value of high-risk human papillomavirus viral load on cervical lesion assessment and ASCUS triage. Cancer Med. 2021;10(7):2482–2488. doi:10.1002/cam4.3653
13. Leite KRM, Pimenta R, Canavez J, et al. HPV genotype prevalence and success of vaccination to prevent cervical cancer. Acta cytologica. 2020;64 (5):420–424. doi:10.1159/000506725
14. Kelly-Hanku A, Newland J, Aggleton P, et al. HPV vaccination in Papua New Guinea to prevent cervical cancer in women: gender, sexual morality, outsiders and the de-feminization of the HPV vaccine. Papillomavirus Res. 2019;8:100171. doi:10.1016/j.pvr.2019.100171
15. Molina R, Barak V, van Dalen A, et al. Tumor markers in breast cancer- European Group on Tumor Markers recommendations. Tumour Biol. 2005;26(6):281–293. doi:10.1159/000089260
16. Shao Y, Zhu F, Zhu S, Bai L. HDAC6 suppresses microRNA-199a transcription and augments HPV-positive cervical cancer progression through Wnt5a upregulation. Int J Biochem Cell Biol. 2021;136:106000. doi:10.1016/j.biocel.2021.106000
17. Pimple SA, Mishra GA. Optimizing high risk HPV-based primary screening for cervical cancer in low- and middle-income countries: opportunities and challenges. Minerva ginecologica. 2019;71(5):365–371. doi:10.23736/s0026-4784.19.04468-x

18. Langberg G, Nygård JF, Gogineni VC, Nygård M, Grasmair M, Naumova V. Towards a data-driven system for personalized cervical cancer risk stratification. Sci Rep. 2022;12(1):12083. doi:10.1038/s41598-022-16361-6
19. Kruczkowski M, Drabik-Kruczkowska A, Marciniak A, Tarczewska M, Kosowska M, Szczerska M. Predictions of cervical cancer identification by photonic method combined with machine learning. Sci Rep. 2022;12(1):3762. doi:10.1038/s41598-022-07723-1
20. Abera GB, Yebyo HG, Hailekiros H, et al. Epidemiology of pre-cancerous cervical lesion and risk factors among adult women in Tigray, Ethiopia. PLoS One. 2023;18(1):e0280191. doi:10.1371/journal.pone.0280191
21. Sheng B, Yao D, Du X, Chen D, Zhou L. Establishment and validation of a risk prediction model for high-grade cervical lesions. Eur J Obstet Gynecol Reprod Biol. 2023;281:1–6. doi:10.1016/j.ejogrb.2022.12.005
22. Chen H, Shu HM, Chang ZL, et al. Efficacy of Pap test in combination with ThinPrep cytological test in screening for cervical cancer. Asian Pac J Cancer Prev. 2012;13(4):1651–1655. doi:10.7314/apjcp.2012.13.4.1651
23. Liu Y, Zhang L, Zhao G, Che L, Zhang H, Fang J. The clinical research of Thinprep Cytology Test (TCT) combined with HPV-DNA detection in screening cervical cancer. Cell Mol Biol. 2017;63(2):92–95. doi:10.14715/cmb/2017.63.2.14
24. Egawa N, Doorbar J. The low-risk papillomaviruses. Virus Res. 2017;231:119–127. doi:10.1016/j.virusres.2016.12.017
25. Bonde JH, Sandri MT, Gary DS, Andrews JC. Clinical utility of human papillomavirus genotyping in cervical cancer screening: a systematic review. J Low Genit Tract Dis. 2020;24(1):1–13. doi:10.1097/lgt.0000000000000494
26. Katki HA, Schiffman M, Castle PE, et al. Benchmarking CIN 3+ risk as the basis for incorporating HPV and Pap cotesting into cervical screening and management guidelines. J Low Genit Tract Dis. 2013;17(5 Suppl 1):S28–35. doi:10.1097/lgt.0b013e318285423c
27. Mantoani PTS, Jammal MP, Caixeta JM, et al. Association of lesion area measured by colposcopy and cervical neoplasia. J Obstetrics Gynaecol. 2022;42(2):306–309. doi:10.1080/01443615.2021.1904218
28. O'Neill E, Reeves MF, Creinin MD. Baseline colposcopic findings in women entering studies on female vaginal products. Contraception. 2008;78 (2):162–166. doi:10.1016/j.contraception.2008.04.002
29. Jarmulowicz MR, Jenkins D, Barton SE, Goodall AL, Hollingworth A, Singer A. Cytological status and lesion size: a further dimension in cervical intraepithelial neoplasia. Br J Obstet Gynaecol. 1989;96(9):1061–1066. doi:10.1111/j.1471-0528.1989.tb03381.x
30. Mantoani PTS, Micheli DC, Jammal MP, et al. High-grade cervical intraepithelial neoplasia: impact of colposcopic lesion area on systemic immune responses. Int J Women's Health. 2025;17:345–353. doi:10.2147/ijwh.s503028
31. Athanasiou A, Bowden S, Paraskevaidi M, et al. HPV vaccination and cancer prevention. Best Pract Res Clin Obstet Gynaecol. 2020;65:109–124. doi:10.1016/j.bpobgyn.2020.02.009
32. Perkins RB, Guido RS, Castle PE, et al. 2019 ASCCP risk-based management consensus guidelines for abnormal cervical cancer screening tests and cancer precursors. J Low Genit Tract Dis. 2020;24(2):102–131. doi:10.1097/lgt.0000000000000525
33. Insinga RP, Perez G, Wheeler CM, et al. Incident cervical HPV infections in young women: transition probabilities for CIN and infection clearance. Cancer Epidemiol Biomarkers Prev. 2011;20(2):287–296. doi:10.1158/1055-9965.epi-10-0791