Original Paper

# Identifying Data-Driven Clinical Subgroups for Cervical Cancer Prevention With Machine Learning: Population-Based, External, and Diagnostic Validation Study

Zhen Lu[1], MSc; Binhua Dong[2,3], MSc; Hongning Cai[4], MSc; Tian Tian[5], PhD; Junfeng Wang[6], PhD; Leiwen Fu[7], PhD; Bingyi Wang[8,9], PhD; Weijie Zhang[1], BD; Shaomei Lin[10], BD; Xunyuan Tuo[11], MSc; Juntao Wang[12], MSc; Tianjie Yang[13], MSc; Xinxin Huang[14], MSc; Zheng Zheng[13], MSc; Huifeng Xue[15], MSc; Shuxia Xu[16], BD; Siyang Liu[1], PhD; Pengming Sun[2,3,17], MD, PhD; Huachun Zou[18,19,20], PhD

[1]School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen, China
[2]Department of Gynecology, Laboratory of Gynecologic Oncology, Fujian Maternity and Child Health Hospital, College of Clinical Medicine for Obstetrics & Gynecology and Pediatrics, Fujian Medical University, Fuzhou, China
[3]Fujian Key Laboratory of Women and Children's Critical Diseases Research, Fuzhou, China
[4]Department of Gynecology, Maternal and Child Health Hospital of Hubei Province (Women and Children's Hospital of Hubei Province) Wuhan, Wuhan, China
[5]School of Public Health, Xinjiang Medical University, Urumqi, China
[6]Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands
[7]Beijing Chest Hospital, Capital Medical University, Beijing Tuberculosis and Thoracic Tumor Research Institute, Beijing, China
[8]Institute for HIV/AIDS Control and Prevention, Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China
[9]Department of HIV/AIDS Control and Prevention, Guangdong Provincial Academy of Preventive Medicine, Guangzhou, China
[10]Department of Gynecology, Shunde Women's and Children's Hospital of Guangdong Medical University, Foshan, China
[11]Department of Gynecology, Gansu Provincial Maternity and Child-care Hospital, Lanzhou, China
[12]Department of Gynecology, Guiyang Maternal and Child Health Care Hospital, Guiyang, China
[13]Department of Gynecology, Shenzhen Maternity & Child Healthcare Hospital, Shenzhen, China
[14]The Ministry of Health, Fujian Maternity and Child Health Hospital, College of Clinical Medicine for Obstetrics & Gynecology and Pediatrics, Fujian Medical University, Fuzhou, China
[15]Center for Cervical Disease Diagnosis and Treatment, Fujian Maternity and Child Health Hospital, College of Clinical Medicine for Obstetrics & Gynecology and Pediatrics, Fujian Medical University, Fuzhou, China
[16]Department of Pathology, Fujian Maternity and Child Health Hospital, College of Clinical Medicine for Obstetrics & Gynecology and Pediatrics, Fujian Medical University, Fuzhou, China
[17]School of Group Medicine and Public Health, Peking Union Medical College, Beijing, China
[18]School of Public Health, Fudan University, Shanghai, China
[19]Shenzhen Campus, Sun Yat-sen University, Shenzhen, China
[20]Fujian Maternity and Child Health Hospital, College of Clinical Medicine for Obstetrics and Gynecology and Pediatrics, Fujian Medical University, Fuzhou, China

**Corresponding Author:**

Huachun Zou, PhD
School of Public Health
Fudan University
Room 435, Bld #8, 130 Dongan Road, Xuhui District
Shanghai, 200032
China
Phone: 86 21 5423 7707
Email: zouhuachun@fudan.edu.cn

## Abstract

**Background:** Cervical cancer remains a major global health issue. Personalized, data-driven cervical cancer prevention (CCP) strategies tailored to phenotypic profiles may improve prevention and reduce disease burden.

**Objective:** This study aimed to identify subgroups with differential cervical precancer or cancer risks using machine learning, validate subgroup predictions across datasets, and propose a computational phenomapping strategy to enhance global CCP efforts.

**Methods:** We explored the data-driven CCP subgroups by applying unsupervised machine learning to a deeply phenotyped, population-based discovery cohort. We extracted CCP-specific risks of cervical intraepithelial neoplasia (CIN) and cervical cancer through weighted logistic regression analyses providing odds ratio (OR) estimates and 95% CIs. We trained a supervised machine learning model and developed pathways to classify individuals before evaluating its diagnostic validity and usability on an external cohort.

**Results:** This study included 551,934 women (median age, 49 years) in the discovery cohort and 47,130 women (median age, 37 years) in the external cohort. Phenotyping identified 5 CCP subgroups, with CCP4 showing the highest carcinoma prevalence. CCP2–4 had significantly higher risks of CIN2+ (CCP2: OR 2.07 [95% CI: 2.03-2.12], CCP3: 3.88 [3.78-3.97], and CCP4: 4.47 [4.33-4.63]) and CIN3+ (CCP2: 2.10 [2.05-2.14], CCP3: 3.92 [3.82-4.02], and CCP4: 4.45 [4.31-4.61]) compared to CCP1 ($P<.001$), consistent with the direction of results observed in the external cohort. The proposed triple strategy was validated as clinically relevant, prioritizing high-risk subgroups (CCP3-4) for colposcopies and scaling human papillomavirus screening for CCP1-2.

**Conclusions:** This study underscores the potential of leveraging machine learning algorithms and large-scale routine electronic health records to enhance CCP strategies. By identifying key determinants of CIN2+/CIN3+ risk and classifying 5 distinct subgroups, our study provides a robust, data-driven foundation for the proposed triple strategy. This approach prioritizes tailored prevention efforts for subgroups with varying risks, offering a novel and scalable tool to complement existing cervical cancer screening guidelines. Future work should focus on independent external and prospective validation to maximize the global impact of this strategy.

***JMIR Public Health Surveill 2025;11:e67840;*** doi: 10.2196/67840

**Keywords:** cervical cancer; human papillomavirus; screening; machine learning; cervical tumor; cancer; carcinoma; tumor; malignant; ML; phenomapping strategy; logistic regression; regression; population-based; validation study; cancer prevention; validity; usability; algorithm; surveillance; electronic health record; EHR

## Introduction

Cervical cancer is the fourth most common cancer among women, with an estimated 660,000 new cases and 350,000 deaths globally in 2022 [1,2]. It is the most common cancer in 25 countries and the leading cause of cancer death in 37 countries. Despite being largely preventable through human papillomavirus (HPV) vaccination [3], the high incidence and persistence of high-risk HPV (hrHPV) remain the primary risk factors for cervical cancer [4] and related diseases [5-7]. In November 2020, the World Health Organization (WHO) launched a global initiative [8] to eliminate cervical cancer as a public health problem, emphasizing a triple intervention strategy: vaccinating at least 90% of girls against HPV by age 15 years, screening 70% of women with a high-performance test by ages 35 and 45 years, and treating at least 90% of detected precancerous lesions and invasive cancers. Yet, globally, an estimated 1.6 billion (67%) of 2.3 billion women aged 20-70 years have never been screened for cervical cancer [9], and in China, 5-year screening coverage among women aged 35-49 years was only 33%. Establishing high-quality, sustainable, and acceptable cervical cancer prevention (CCP) with broad coverage in resource-limited regions remains a critical challenge [10-12]. While HPV-based screening has demonstrated significant benefits [13-18], leading the WHO to recommend it as the primary method for CCP globally [19], its adoption in low- and middle-income countries (LMICs) is limited due to health inequities, resource constraints, and limited access to affordable, clinically validated HPV tests [9,20]. Addressing these barriers requires personalized CCP strategies.

We hypothesize that data-driven CCP subgroups, identified using machine learning, can provide valuable insights into CCP, enabling personalized strategies to reduce the risk of cervical intraepithelial neoplasia (CIN) and cervical cancer. Personalized CCP strategies depend on the unique phenotypic profile of each individual. The population, with its varying phenotypic diversity, comprises heterogeneous subgroups that reflect multiple underlying behavior patterns and causes of disease. Clinical practice guidelines also recommend that cervical cancer screening should be tailored to an individual's risk profile of HPV infection [21,22]. Therefore, the discovery of CCP subgroups could inform and improve the development of new screening strategies, public health policies, and clinical decision-making, as well as contribute to trial design, despite the complex causal relationships among individual risk factors [4]. To date, no study has identified CCP subgroups. The incomplete understanding of these subgroups across large, diverse populations, coupled with insufficient validation through various methods, has hindered the implementation of the triple intervention strategy. Machine learning-based approaches allow for the phenotyping of entire populations based on individual characteristics, enabling the prediction of phenotypic clusters and disease onset [23,24]. Additionally, by defining a computational phenomap—a mathematical construct of individual phenotypes based on baseline measures [24]—computational methods can assess the heterogeneous risk effects of distinct subpopulations.

This approach accounts for the phenotypic diversity within populations and their subsequent histopathological diagnoses.

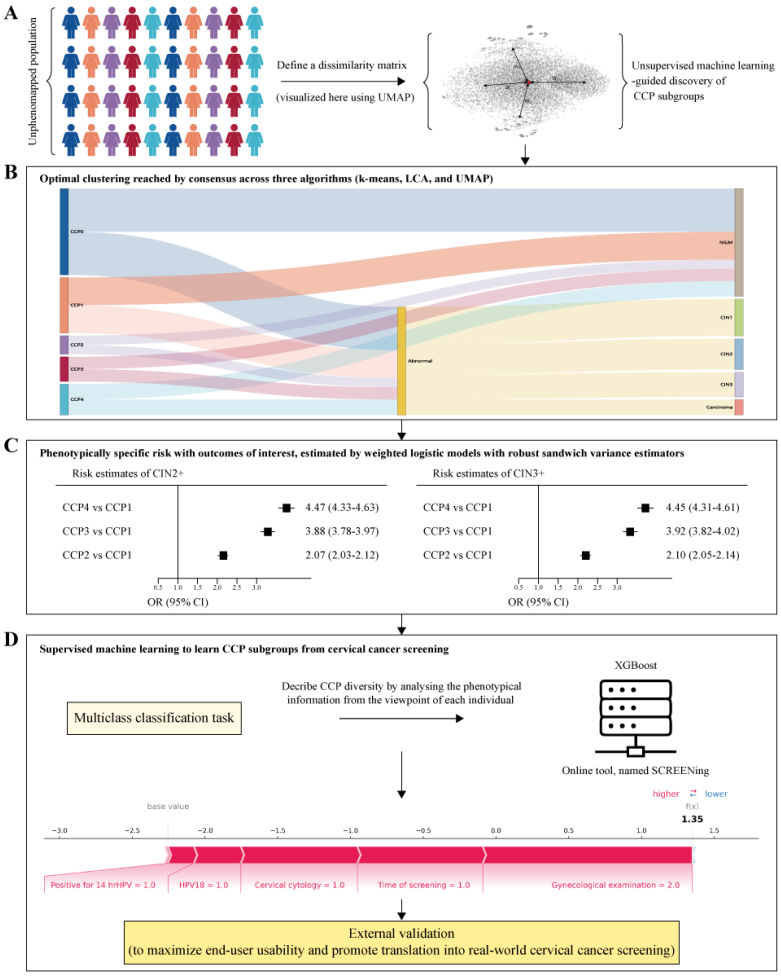
As such, in this study, we aimed to: (1) test the hypothesis that distinct CCP subgroups exhibit differential risks of cervical precancer or cancer based on their complex phenotypic profiles (development); (2) demonstrate internal validity (within a dataset and across methods), as well as external and diagnostic validity (across datasets; validation); and (3) propose a computational phenomapping strategy with clinical relevance and pathways to improve global access to CCP (impact).

# Methods

## Data Source

Deidentified data were extracted from electronic health records (EHRs) of the national cervical cancer screening program in China. In summary, our study included eligible women aged 25-65 years who participated in the cervical cancer screening. Data from Fujian Province (2014-2023) were used to establish a discovery cohort to train the models. Additionally, EHR data from 5 other regions—Shenzhen City, Foshan City, Hubei Province, Gansu Province, and Guizhou Province—were employed as an external cohort to validate the generalizability of the models across diverse populations. Details on the study design, as illustrated in Figure 1, are available in Appendix S1 in Multimedia Appendix 1.

**Figure 1.** Study design. (A) We explored the data-driven cervical cancer prevention (CCP) subgroups by applying unsupervised machine learning to a deeply phenotyped, population-based discovery cohort. In this population-based, external, and diagnostic validation study, we aimed to use different machine learning methods to test our hypothesis (B) that individuals within distinct CCP subgroups exhibit differential risks of cervical precancer or cancer based on their complex phenotypic profiles. After identifying CCP subgroups, we extracted CCP-specific risks of outcomes of interest (C) conditionally on all predefined and algorithmically selected features through weighted logistic regression analyses providing average odds ratio (OR) estimates. Additionally, we stratified individuals based on their key features to conduct subgroup analyses. Finally, we trained a supervised machine learning model and developed pathways (D) to classify individuals using features most consistently linked to CCP subgroups before evaluating its diagnostic validity and usability on an external cohort. Pathologically abnormal diagnoses consisted of CIN1, CIN2, CIN3, and carcinoma. UMAP: uniform manifold approximation and projection; CCP: cervical cancer prevention; LCA: latent class analysis; NILM: negative for intraepithelial lesion or malignancy; CIN1: cervical intraepithelial neoplasia grade 1; CIN2: cervical intraepithelial neoplasia grade 2; CIN3: cervical intraepithelial neoplasia grade 3; CIN2+: cervical intraepithelial neoplasia grade 2 or worse; CIN3+: cervical intraepithelial neoplasia grade 3 or worse; OR: odds ratio; XGBoost: eXtreme gradient boosting; SCREENing: clinical Subgroups for CeRvical cancEr prEvention using computational phenomaps and machine learning; hrHPV: high-risk human papillomavirus; HPV: human papillomavirus.




## Ethical Considerations

This study was approved by the Ethics Committee of Fujian Maternal and Child Health Hospital (2023KY141). No additional informed consent was required for this analysis. All participant data were deidentified prior to analysis. No images of identifiable individuals are included in the manuscript or supplementary materials.

## Data Preprocessing

In line with the consensus that detecting and treating CIN2 or CIN3 [25,26], key premalignant cervical conditions, can prevent progression to invasive cervical cancer, our primary and secondary outcomes were CIN2+ and CIN3+, respectively. These outcomes were confirmed by histopathology and align with those commonly used in previous studies [26-29]. We selected features based on prior research [22,30-34] and input from clinical, biostatistical, and epidemiological experts. These included demographic characteristics, cervical cancer screening history, HPV infection status, and medical examination results. Details on data preprocessing are provided in Appendix S1 in Multimedia Appendix 1.

## Unsupervised Machine Learning

We calculated phenotypic distances between individuals using Gower's distance [35], a dissimilarity metric suitable for mixed continuous and categorical data. To visualize phenotypic variation in the population, we employed uniform manifold approximation and projection (UMAP) [36] to construct a phenomap. This approach enhances interpretability by presenting the distributions of individuals within a multidimensional phenotypic space, capturing the full range of baseline phenotypes. To identify and categorize

the underlying CCP subgroups in which individuals exhibit phenotypic similarity, we employed and compared 3 methods to capture and validate CCP diversity: k-means [37], latent class analysis [38], and UMAP [39]. The final number of CCP subgroups was determined by reaching a consensus across all 3 approaches. Further details are provided in Appendix S1 in Multimedia Appendix 1.

### Risk Estimates and an Algorithm to Identify Subgroups

To further explore the association between CCP subgroups and outcomes of interest, we calculated CCP-specific risk estimates (odds ratios, ORs) [40] and 95% CIs using inverse probability weights [41]. Additionally, we trained an eXtreme gradient boosting (XGBoost) algorithm [42] to predict 5 CCP subgroups with differential risks of outcomes, as described in our previous work [23]. Briefly, the model was trained and calibrated using an isotonic regression algorithm, and internally validated in the discovery cohort. The SHapley Additive exPlanations (SHAP) method was employed to identify each feature's relative contribution [23,43] and enhance the model's explainability. Model performance was evaluated using OVR AUROC (one-versus-rest area under the receiver operating characteristics curve, extended for multiple classes), Brier score, and calibration curves as primary metrics. In external validation, the model's diagnostic validity was assessed by comparing CCP-specific ORs between datasets to evaluate cross-cluster and cross-dataset risk differences. Further details are provided in Appendix S1 in Multimedia Appendix 1.

### Developing Pathways to Improve Impact

We named our computational phenomapping strategy SCREENing (clinical Subgroups for CeRvical cancEr prEvention using computational pheNomaps and machine learnING). To assess real-world usability and effectiveness, we consulted 11 clinical experts and 3 epidemiologists on its clinical relevance, justification, result interpretability, and potential impact on screening strategies and public health policymaking.

### Statistical Analysis

Given the importance of menopause in women, we specifically examined CCP-specific risks across age using spline analyses, with interaction tests to assess whether age modified these risks. Subgroup analyses were also performed by stratifying women based on key features. Details are provided in Appendix S1 in Multimedia Appendix 1. We followed the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [44] and the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis statement [45]. Data management was performed using the lulab.utils R package [46]. All analyses [47] were conducted using Python (version 3.11.6), SAS Enterprise Guide (version 7.1), and R (version 4.3.2; R Foundation for Statistical Computing).

## Results

### Characteristics of Cohorts

The study included 551,934 women (median age 49 years [42,48]; 10% infected with hrHPV, 1.2% with HPV-16, and 0.6% with HPV-18) in the discovery cohort (Table 1) and 47,130 women (median age 37 years [32,45]; 16.6% with hrHPV, 3.9% with HPV-16, and 1.4% with HPV-18) in the external cohort (Table S3 in Multimedia Appendix 1). In the discovery cohort, 9932 were pathologically diagnosed with cervical abnormalities, including 533 with carcinoma. In contrast, the external cohort showed a higher prevalence of cervical abnormalities at 11.1% (5251).

**Table 1.** Characteristics by cervical cancer prevention (CCP) subgroups in the discovery cohort. Categorical features are summarized as numbers (percentages), and continuous features as median (Q1, Q3), as appropriate.

| Characteristics | CCP0[a] (n=542,002) | CCP1 (n=2242) | CCP2 (n=3770) | CCP3 (n=2278) | CCP4 (n=1642) | Total (n=551,934) |
|---|---|---|---|---|---|---|
| Age (years), median (Q1, Q3) | 48.00 (42.00, 54.00) | 48.00 (42.00, 54.00) | 48.00 (43.00, 55.00) | 51.00 (44.00, 57.00) | 48.00 (43.00, 54.00) | 49.00 (42.00, 54.00) |
| **Race/ethnicity, n (%)** | | | | | | |
| Han | 534,023 (98.53) | 2214 (98.75) | 3707 (98.33) | 2236 (98.16) | 1607 (97.87) | 543,787 (98.52) |
| Others | 4528 (0.84) | 15 (0.67) | 30 (0.80) | 29 (1.27) | 20 (1.22) | 4622 (0.84) |
| Missing | 3451 (0.64) | 13 (0.58) | 33 (0.88) | 13 (0.57) | 15 (0.91) | 3525 (0.64) |
| **History of cervical cancer screening, n (%)** | | | | | | |
| Missing | 234 (0.04) | 0 (0.00) | 2 (0.05) | 0 (0.00) | 0 (0.00) | 236 (0.04) |
| No | 411,029 (75.84) | 0 (0.00) | 3768 (99.95) | 2278 (100.00) | 1642 (100.00) | 418,717 (75.86) |
| Yes | 130,739 (24.12) | 2242 (100.00) | 0 (0.00) | 0 (0.00) | 0 (0.00) | 132,981 (24.09) |
| **Time of previous screening, n (%)** | | | | | | |
| Missing | 785 (0.14) | 13 (0.58) | 2 (0.05) | 0 (0.00) | 0 (0.00) | 800 (0.14) |
| No previous screening | 411,029 (75.84) | 0 (0.00) | 3768 (99.95) | 2278 (100.00) | 1642 (100.00) | 418,717 (75.86) |
| Within 3 years from now | 65,519 (12.09) | 1091 (48.66) | 0 (0.00) | 0 (0.00) | 0 (0.00) | 66,610 (12.07) |

| Characteristics | CCP0[a] (n=542,002) | CCP1 (n=2242) | CCP2 (n=3770) | CCP3 (n=2278) | CCP4 (n=1642) | Total (n=551,934) |
|---|---|---|---|---|---|---|
| More than 3 years ago | 64,669 (11.93) | 1138 (50.76) | 0 (0.00) | 0 (0.00) | 0 (0.00) | 65,807 (11.92) |
| **Gynecological examination, n (%)** | | | | | | |
| Missing | 2180 (0.40) | 1 (0.04) | 1 (0.03) | 0 (0.00) | 0 (0.00) | 2182 (0.40) |
| Normal | 417,921 (77.11) | 1508 (67.26) | 3763 (99.81) | 1568 (68.83) | 0 (0.00) | 424,760 (76.96) |
| Abnormal | 121,901 (22.49) | 733 (32.69) | 6 (0.16) | 710 (31.17) | 1642 (100.00) | 124,992 (22.65) |
| **Positive for high-risk HPV[b, c], n (%)** | | | | | | |
| No | 496,234 (91.56) | 99 (4.42) | 206 (5.46) | 27 (1.19) | 67 (4.08) | 496,633 (89.98) |
| Yes | 45,768 (8.44) | 2143 (95.58) | 3564 (94.54) | 2251 (98.81) | 1575 (95.92) | 55,301 (10.02) |
| **Positive for low-risk HPV[d], n (%)** | | | | | | |
| No | 534,313 (98.58) | 2035 (90.77) | 3664 (97.19) | 1813 (79.59) | 1602 (97.56) | 543,427 (98.46) |
| Yes | 7689 (1.42) | 207 (9.23) | 106 (2.81) | 465 (20.41) | 40 (2.44) | 8507 (1.54) |
| **Positive for possible high-risk HPV[e], n (%)** | | | | | | |
| No | 537,999 (99.26) | 2130 (95.00) | 3686 (97.77) | 1987 (87.23) | 1617 (98.48) | 547,419 (99.18) |
| Yes | 4003 (0.74) | 112 (5.00) | 84 (2.23) | 291 (12.77) | 25 (1.52) | 4515 (0.82) |
| Number of HPV infections | 0.00 (0.00, 0.00) | 1.00 (1.00, 2.00) | 1.00 (1.00, 1.00) | 2.00 (2.00, 3.00) | 1.00 (1.00, 1.00) | 0.00 (0.00, 0.00) |
| **Positive for HPV-16, n (%)** | | | | | | |
| No | 538,375 (99.33) | 1684 (75.11) | 2745 (72.81) | 1494 (65.58) | 1199 (73.02) | 545,497 (98.83) |
| Yes | 3627 (0.67) | 558 (24.89) | 1025 (27.19) | 784 (34.42) | 443 (26.98) | 6437 (1.17) |
| **Positive for HPV-18, n (%)** | | | | | | |
| No | 539,716 (99.58) | 2002 (89.30) | 3442 (91.30) | 1887 (82.84) | 1508 (91.84) | 548,555 (99.39) |
| Yes | 2286 (0.42) | 240 (10.70) | 328 (8.70) | 391 (17.16) | 134 (8.16) | 3379 (0.61) |
| **Positive for HPV-31, n (%)** | | | | | | |
| No | 540,143 (99.66) | 2138 (95.36) | 3662 (97.14) | 2063 (90.56) | 1588 (96.71) | 549,594 (99.58) |
| Yes | 1859 (0.34) | 104 (4.64) | 108 (2.86) | 215 (9.44) | 54 (3.29) | 2340 (0.42) |
| **Positive for HPV-33, n (%)** | | | | | | |
| No | 539,290 (99.50) | 2096 (93.49) | 3608 (95.70) | 1978 (86.83) | 1562 (95.13) | 548,534 (99.38) |
| Yes | 2712 (0.50) | 146 (6.51) | 162 (4.30) | 300 (13.17) | 80 (4.87) | 3400 (0.62) |
| **Positive for HPV-35, n (%)** | | | | | | |
| No | 540,820 (99.78) | 2197 (97.99) | 3708 (98.36) | 2160 (94.82) | 1622 (98.78) | 550,507 (99.74) |
| Yes | 1182 (0.22) | 45 (2.01) | 62 (1.64) | 118 (5.18) | 20 (1.22) | 1427 (0.26) |
| **Positive for HPV-39, n (%)** | | | | | | |
| No | 537,675 (99.20) | 2092 (93.31) | 3638 (96.50) | 2021 (88.72) | 1600 (97.44) | 547,026 (99.11) |
| Yes | 4327 (0.80) | 150 (6.69) | 132 (3.50) | 257 (11.28) | 42 (2.56) | 4908 (0.89) |
| **Positive for HPV-45, n (%)** | | | | | | |
| No | 540,998 (99.81) | 2205 (98.35) | 3741 (99.23) | 2204 (96.75) | 1632 (99.39) | 550,780 (99.79) |
| Yes | 1004 (0.19) | 37 (1.65) | 29 (0.77) | 74 (3.25) | 10 (0.61) | 1154 (0.21) |
| **Positive for HPV-51, n (%)** | | | | | | |
| No | 537,654 (99.20) | 2053 (91.57) | 3590 (95.23) | 1954 (85.78) | 1565 (95.31) | 546,816 (99.07) |
| Yes | 4348 (0.80) | 189 (8.43) | 180 (4.77) | 324 (14.22) | 77 (4.69) | 5118 (0.93) |
| **Positive for HPV-52, n (%)** | | | | | | |
| No | 525,904 (97.03) | 1622 (72.35) | 3019 (80.08) | 1396 (61.28) | 1276 (77.71) | 533,217 (96.61) |
| Yes | 16,098 (2.97) | 620 (27.65) | 751 (19.92) | 882 (38.72) | 366 (22.29) | 18,717 (3.39) |
| **Positive for HPV-56, n (%)** | | | | | | |
| No | 539,439 (99.53) | 2133 (95.14) | 3682 (97.67) | 2076 (91.13) | 1610 (98.05) | 548,940 (99.46) |
| Yes | 2563 (0.47) | 109 (4.86) | 88 (2.33) | 202 (8.87) | 32 (1.95) | 2994 (0.54) |
| **Positive for HPV-58, n (%)** | | | | | | |

| Characteristics | CCP0[a] (n=542,002) | CCP1 (n=2242) | CCP2 (n=3770) | CCP3 (n=2278) | CCP4 (n=1642) | Total (n=551,934) |
|---|---|---|---|---|---|---|
| No | 534,764 (98.66) | 1877 (83.72) | 3297 (87.45) | 1691 (74.23) | 1413 (86.05) | 543,042 (98.39) |
| Yes | 7238 (1.34) | 365 (16.28) | 473 (12.55) | 587 (25.77) | 229 (13.95) | 8892 (1.61) |
| **Positive for HPV-59, n (%)** | | | | | | |
| No | 540,034 (99.64) | 2180 (97.23) | 3710 (98.41) | 2124 (93.24) | 1620 (98.66) | 549,668 (99.59) |
| Yes | 1968 (0.36) | 62 (2.77) | 60 (1.59) | 154 (6.76) | 22 (1.34) | 2266 (0.41) |
| **Positive for HPV-66, n (%)** | | | | | | |
| No | 540,640 (99.75) | 2186 (97.50) | 3720 (98.67) | 2180 (95.70) | 1614 (98.29) | 550,340 (99.71) |
| Yes | 1362 (0.25) | 56 (2.50) | 50 (1.33) | 98 (4.30) | 28 (1.71) | 1594 (0.29) |
| **Positive for HPV-68, n (%)** | | | | | | |
| No | 538,250 (99.31) | 2124 (94.74) | 3654 (96.92) | 2018 (88.59) | 1604 (97.69) | 547,650 (99.22) |
| Yes | 3752 (0.69) | 118 (5.26) | 116 (3.08) | 260 (11.41) | 38 (2.31) | 4284 (0.78) |
| **Positive for HPV-11, n (%)** | | | | | | |
| No | 541,435 (99.90) | 2218 (98.93) | 3764 (99.84) | 2235 (98.11) | 1631 (99.33) | 551,283 (99.88) |
| Yes | 567 (0.10) | 24 (1.07) | 6 (0.16) | 43 (1.89) | 11 (0.67) | 651 (0.12) |
| **Positive for HPV-42, n (%)** | | | | | | |
| No | 540,587 (99.74) | 2199 (98.08) | 3745 (99.34) | 2175 (95.48) | 1639 (99.82) | 550,345 (99.71) |
| Yes | 1415 (0.26) | 43 (1.92) | 25 (0.66) | 103 (4.52) | 3 (0.18) | 1589 (0.29) |
| **Positive for HPV-43, n (%)** | | | | | | |
| No | 541,315 (99.87) | 2212 (98.66) | 3759 (99.71) | 2210 (97.01) | 1641 (99.94) | 551,137 (99.86) |
| Yes | 687 (0.13) | 30 (1.34) | 11 (0.29) | 68 (2.99) | 1 (0.06) | 797 (0.14) |
| **Positive for HPV-44, n (%)** | | | | | | |
| No | 540,505 (99.72) | 2212 (98.66) | 3753 (99.55) | 2201 (96.62) | 1636 (99.63) | 550,307 (99.71) |
| Yes | 1497 (0.28) | 30 (1.34) | 17 (0.45) | 77 (3.38) | 6 (0.37) | 1627 (0.29) |
| **Positive for HPV-6, n (%)** | | | | | | |
| No | 540,748 (99.77) | 2207 (98.44) | 3748 (99.42) | 2199 (96.53) | 1635 (99.57) | 550,537 (99.75) |
| Yes | 1254 (0.23) | 35 (1.56) | 22 (0.58) | 79 (3.47) | 7 (0.43) | 1397 (0.25) |
| **Positive for HPV-81, n (%)** | | | | | | |
| No | 539,354 (99.51) | 2178 (97.15) | 3747 (99.39) | 2136 (93.77) | 1632 (99.39) | 549,047 (99.48) |
| Yes | 2648 (0.49) | 64 (2.85) | 23 (0.61) | 142 (6.23) | 10 (0.61) | 2887 (0.52) |
| **Positive for HPV-53, n (%)** | | | | | | |
| No | 538,445 (99.34) | 2139 (95.41) | 3689 (97.85) | 2015 (88.45) | 1617 (98.48) | 547,905 (99.27) |
| Yes | 3557 (0.66) | 103 (4.59) | 81 (2.15) | 263 (11.55) | 25 (1.52) | 4029 (0.73) |
| **Cervical cytology examination, n (%)** | | | | | | |
| NILM[f] | 42,746 (7.89) | 489 (21.81) | 721 (19.12) | 438 (19.23) | 345 (21.01) | 44,739 (8.11) |
| No examination due to negative for high-risk HPV | 487,235 (89.90) | 3 (0.13) | 7 (0.19) | 0 (0.00) | 0 (0.00) | 487,245 (88.28) |
| ASC-US[g] | 7577 (1.40) | 808 (36.04) | 1185 (31.43) | 721 (31.65) | 531 (32.34) | 10822 (1.96) |
| LSIL[h] | 2245 (0.41) | 488 (21.77) | 772 (20.48) | 546 (23.97) | 286 (17.42) | 4337 (0.79) |
| AGC[i] | 99 (0.02) | 18 (0.80) | 29 (0.77) | 13 (0.57) | 30 (1.83) | 189 (0.03) |
| Missing but positive for high-risk HPV | 779 (0.14) | 99 (4.42) | 231 (6.13) | 159 (6.98) | 81 (4.93) | 1349 (0.24) |
| ASC-H[j] | 902 (0.17) | 147 (6.56) | 363 (9.63) | 163 (7.16) | 145 (8.83) | 1720 (0.31) |
| AGC-FN[k] | 7 (0.00) | 4 (0.18) | 9 (0.24) | 3 (0.13) | 6 (0.37) | 29 (0.01) |
| HSIL[l] | 411 (0.08) | 185 (8.25) | 443 (11.75) | 233 (10.23) | 215 (13.09) | 1487 (0.27) |
| Carcinoma | 1 (0.00) | 1 (0.04) | 10 (0.27) | 2 (0.09) | 3 (0.18) | 17 (0.00) |
| **Histopathological diagnosis, n (%)** | | | | | | |

| Characteristics | CCP0[a] (n=542,002) | CCP1 (n=2242) | CCP2 (n=3770) | CCP3 (n=2278) | CCP4 (n=1642) | Total (n=551,934) |
|---|---|---|---|---|---|---|
| NILM | 542,002 (100.00) | 0 (0.00) | 0 (0.00) | 0 (0.00) | 0 (0.00) | 542,002 (98.20) |
| CIN1[m] | 0 (0.00) | 1542 (68.78) | 2456 (65.15) | 1445 (63.43) | 948 (57.73) | 6391 (1.16) |
| CIN2[n] | 0 (0.00) | 22 (0.98) | 29 (0.77) | 15 (0.66) | 16 (0.97) | 82 (0.01) |
| CIN2/3[o] | 0 (0.00) | 551 (24.58) | 1113 (29.52) | 716 (31.43) | 485 (29.54) | 2865 (0.52) |
| CIN3[p] | 0 (0.00) | 16 (0.71) | 19 (0.50) | 13 (0.57) | 13 (0.79) | 61 (0.01) |
| Carcinoma | 0 (0.00) | 111 (4.95) | 153 (4.06) | 89 (3.91) | 180 (10.96) | 533 (0.10) |

[a]CCP: cervical cancer prevention.
[b]HPV: human papillomavirus.
[c]high-risk HPV: HPV-16/18/31/33/35/39/45/51/52/56/58/59/66/68.
[d]low-risk HPV: HPV-11/40/42/43/44/6/61/72/81.
[e]possible high-risk HPV: HPV-53/70/73/82/83.
[f]NILM: negative for intraepithelial lesion or malignancy.
[g]ASC-US: atypical squamous cells of undetermined significance.
[h]LSIL: low-grade squamous intraepithelial lesion.
[i]AGC: atypical glandular cells.
[j]ASC-H: atypical squamous cells, cannot exclude high-grade squamous intraepithelial lesion.
[k]AGC-FN: AGC-favor neoplastic.
[l]HSIL: high-grade squamous intraepithelial lesion.
[m]CIN1: cervical intraepithelial neoplasia grade 1.
[n]CIN2: cervical intraepithelial neoplasia grade 2.
[o]CIN2/3 is reflective of CIN2 or CIN3, i.e., HSIL; carcinoma consists of AIS (adenocarcinoma in situ) and cancer.
[p]CIN3: cervical intraepithelial neoplasia grade 3.

## Data-Driven Subgroups

The discovery cohort was extensively phenotyped based on 31 features (Figure 2). Visual assessment of the risk phenomaps revealed that nearly all 6 features were heterogeneously distributed in the phenomic space to varying degrees, indicating distinct phenotypic neighborhoods. In the discovery cohort, the CCP subgroups identified across 3 algorithms and all metrics were stable (Figure 3A and Table S4 in Multimedia Appendix 1), with the optimal number of clusters being 4 in the subpopulation of 9932 women with cervical abnormalities. The distribution of cervical abnormalities across subgroups is presented in Figure 3B. Additionally, women with normal screening results were treated as another CCP subgroup (Table 1). Following a detailed analysis of each subgroup's features, we labeled the 5 identified subgroups as follows: (0) healthy, (1) early onset, (2) screening-targeted, (3) late onset, and (4) carcinoma-specific. CCP1 had the highest prevalence of CIN1, while CCP4 had the highest prevalence of carcinoma (Figure 3C). Most features were able to discriminate well between the subgroups. In the external cohort, the subgroups were consistent (Table S3 in Multimedia Appendix 1), though CCP1 was not identified due to missing information on previous screenings. The distribution of subgroups was similar across the discovery and external cohorts, with CCP0 being the most common subgroup and CCP4 having the highest prevalence of carcinoma.

**Figure 2.** Manifold representations of the phenotypic architecture of 9932 individuals pathologically diagnosed with cervical abnormality from the discovery cohort (N=551,934). A total of 9932 individuals are embedded in the phenotypic space based on dissimilarity metrics (Gower's distance) derived from 31 included phenotypic features; thus, phenotypically similar women tend to be topologically closer. In the subfigures, each dot represents an individual, with coloring based on the value of features. Since the dimensionality reduction is nonlinear, axes have been omitted, and only the comparisons between distances are meaningful. Pathologically abnormal diagnoses consisted of CIN1, CIN2, CIN3, and carcinoma. 14 high-risk HPV: HPV-16/18/31/33/35/39/45/51/52/56/58/59/66/68; low-risk HPV: HPV-11/40/42/43/44/6/61/72/81; possible high-risk HPV: HPV-53/70/73/82/83.

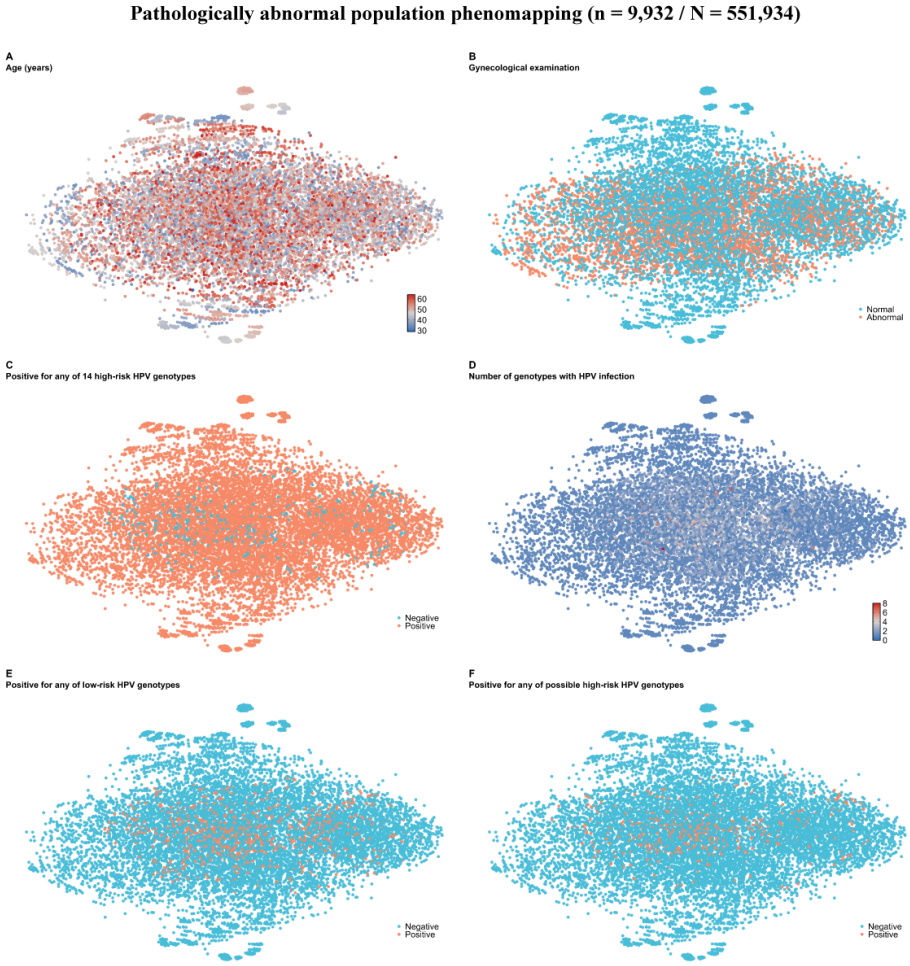
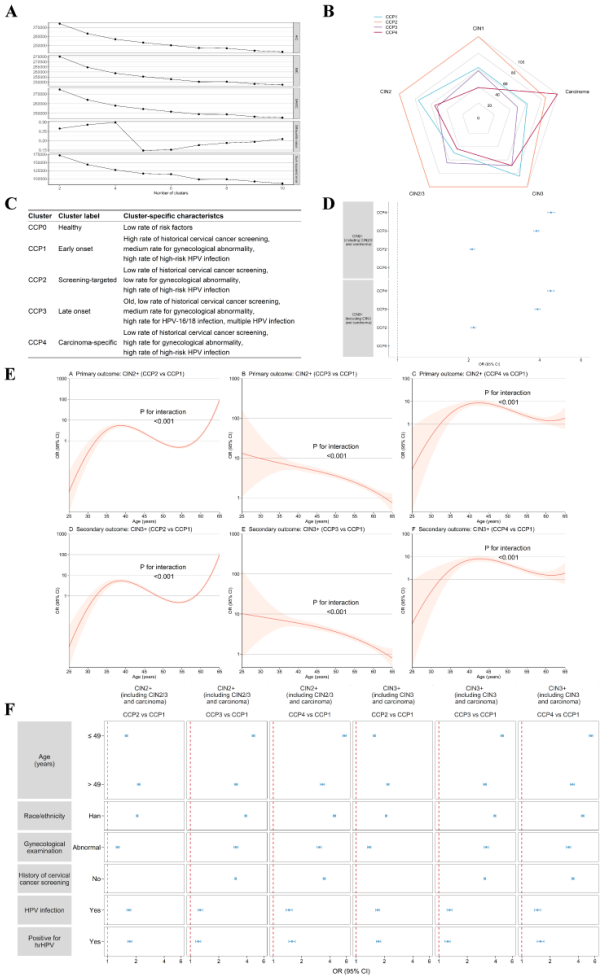


**Figure 3.** Determination of the 5 CCP subgroups with differential risks of CIN2+/CIN3+. (A) Elbow plot to determine the optimal number of clusters. Across 3 algorithms (k-means, LCA, and UMAP), identified CCP subgroups were stable, with the optimal number of clusters being 4 in 9932 women pathologically diagnosed with cervical abnormality. Women with normal screening test findings were considered as one independent CCP subgroup free of cervical cancer (Table 1). (B) Relative prevalence of pathologically abnormal diagnoses across CCP subgroups. For each pathologically abnormal diagnosis, the highest prevalence was designated as 100%, and the prevalence in each of the other CCP subgroups was relative to that prevalence (0-100). (C) Characteristics of CCP subgroups. (D) CCP-specific risk estimates of CIN2+/CIN3+ compared with the CCP1 subgroup. Adjusted ORs (dots) and 95% CIs (error bars) are presented. The dashed line marks an OR of 1.00; lower limits of 95% CIs with values greater than 1.00 indicate significantly increased risk. (E) The CCP-specific risks of CIN2+/CIN3+ across age compared with the CCP1 subgroup. Age was transformed into a restricted cubic spline function for the analyses. *P* value was based on 2-sided Chi-squared test on the interaction between age and CCP subgroups. A *P* value of <.05 suggests that age modifies the association between CCP subgroups and CIN2+/CIN3+. Light-colored bands represent 95% CIs. (F) Subgroup analyses of CCP-specific risk estimates of CIN2+/CIN3+ compared with the CCP1 subgroup. Adjusted ORs (dots) and 95% CIs (error bars) are presented. The dashed line marks an OR of 1.00; lower limits of 95% CIs with values greater than 1.00 indicate significantly increased risk. Empty subfigures suggest insufficient samples for the analysis. AIC: Akaike information criterion; BIC: Bayesian information criterion; SABIC: sample size-adjusted BIC; CCP: cervical cancer prevention; CIN1: cervical intraepithelial neoplasia grade 1; CIN2: cervical intraepithelial neoplasia grade 2; CIN3: cervical intraepithelial neoplasia grade 3; CIN2+: cervical intraepithelial neoplasia grade 2 or worse; CIN3+: cervical intraepithelial neoplasia grade 3 or worse; OR: odds ratio; hrHPV: high-risk human papillomavirus.



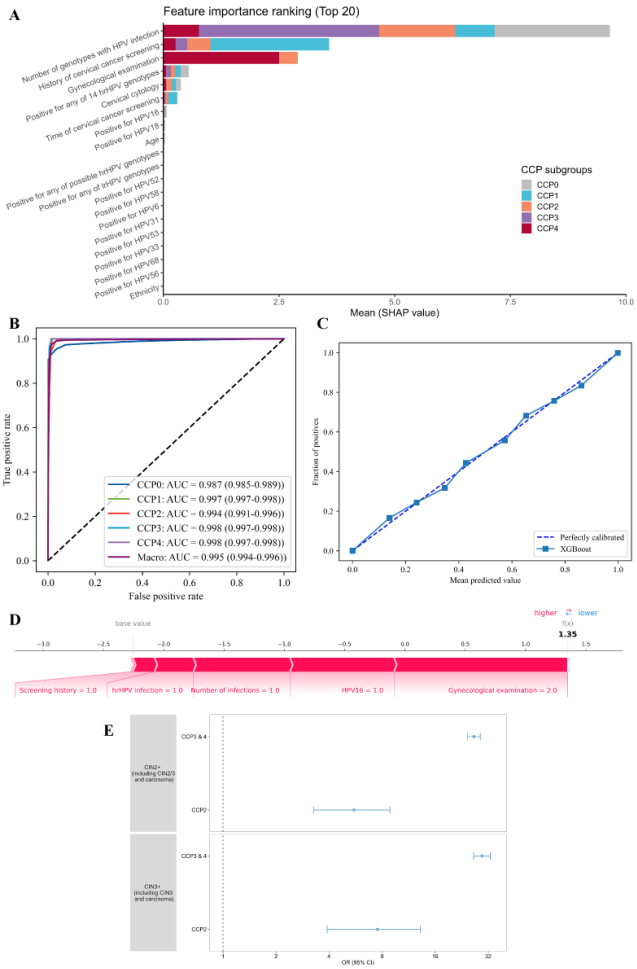

## Risk Estimates and Diagnostic Validity

In the diagnostic validity analysis of the discovery cohort (Figure 3D), women in CCP2-4 exhibited a significantly increased risk of both CIN2+ (CCP2: OR 2.07, 95% CI [2.03-2.12]; CCP3: 3.88 [3.78-3.97]; CCP4: 4.47 [4.33-4.63]) and CIN3+ (2.10 [2.05-2.14]; 3.92 [3.82-4.02]; 4.45 [4.31-4.61]) compared to CCP1. Risk analysis across age groups showed that the risks for CIN2+/CIN3+ were evident in the age ranges of 34-47 and 60-65 in CCP2, 26-61 in CCP3, and 35-60 in CCP4, respectively (Figure 3E). Interaction analyses between age and subgroups revealed that the risks of CIN2+/CIN3+ increased with age in the 34-39 and 60-65 age ranges in CCP2, in the 35-43/44 age range in CCP4. Conversely, the risks decreased with age in CCP2 within the 40-47 age range and in CCP4 within the 45-60 age range (all *P* for interaction <.001). Subgroup analyses indicated that the risks for CIN2+/CIN3+ were present across nearly all subgroups, categorized by age, race, gynecological examination, screening history, hrHPV infection, and number of infections (Figure 3F and Table S5 in Multimedia Appendix 1).

SHAP analysis identified the top 10 key features for prediction, including the number of infections, screening history, gynecological examination, hrHPV infection, cervical cytology, time since previous screening, HPV-16/18 infection, age, and possible hrHPV infection (Figure 4A). The final XGBoost model demonstrated excellent discrimination

(OVR AUROC 0.995 [0.994-0.996]), with CCP0 showing the lowest AUROC of 0.987 [0.985-0.989], and strong calibration (Brier score 0.021 [0.020-0.022]) (Figure 4B–C and Table S6 in Multimedia Appendix 1). For the external diagnostic validity of the identified subgroups, women in CCP2-4 showed differential and increased risk for both CIN2+ (CCP2: 5.54 [3.27-8.86]; CCP3&4: 26.56 [24.44-28.88]) and CIN3+ (7.53 [3.90-13.18]; 29.47 [26.46-32.86]) compared to CCP0. The cross-cluster risk differences were consistent across the discovery and external cohorts. To illustrate the model's explainability, Figure 4D presents the SHAP plot for a woman in the CCP4 subgroup, who had no previous screening and tested positive for HPV-16 and CIN1 based on cervical cytology. Given this information, our model predicted an elevated risk of CCP4. Additionally, a screenshot of our browser-accessible tool based on the model developed in this study, to clarify how the model can be easily accessed via a browser in clinical settings, is shown in Figure S1 in Multimedia Appendix 1.

**Figure 4.** Feature importance, discrimination, calibration, explainability, and external diagnostic validity. (A) The top 20 features for prediction of CCP subgroups with differential risks of CIN2+/CIN3+ are shown. The y-axis represents the features included in the model development (in descending order of importance), and the x-axis indicates the mean of SHAP values. The 10 most important features were selected to train the final parsimonious model. (B) ROC curves in the internal validation from the discovery cohort. The dark purple line represents the macro-average of all 5 CCP subgroups. (C) Calibration curve of the alignment between predicted and observed CCP subgroups for the final model. The solid line corresponds to the calibration curve, with the dashed line corresponding to a reference for perfect calibration (ie, perfect alignment between the predicted and observed probabilities). (D) Explanation of the SCREENing tool (ie, the inference process of the final model with a woman in the CCP4 subgroup). (E) In external validation, to assess cross-cluster and cross-dataset risk differences, the model was measured and compared by diagnostic validity (CCP-specific ORs) between datasets. CCP: cervical cancer prevention; hrHPV: high-risk human papillomavirus; lrHPV: low-risk human papillomavirus; SHAP: SHapley Additive exPlanations; AUC: area under receiver operating characteristics curve; XGBoost: eXtreme gradient boosting.




## Developed Pathways

Based on our analyses, we proposed a triple SCREENing strategy to prioritize CCP subgroups with varying CIN2+/CIN3+ risks: (1) top priority for colposcopy referrals to women in CCP3-4 in resource-constrained settings, (2) higher priority given for scaling up organized, population-based HPV screening programs (with adequate follow-up) for women in CCP1-2, and (3) lowest priority for CCP0 women, with large-scale screening limited to resource-rich settings. To assess its impact, we evaluated the real-world usability and effectiveness of SCREENing. Sample clinicians reported that the identified subgroups and included features were clinically relevant for risk-based cervical cancer management, particularly for colposcopy referrals in CCP3-4. They also

found the strategy transparent, interpretable, and generalizable across clinical settings, with the browser-accessible model feasible and effective for indicating risk probabilities during consultations. Sample epidemiologists highlighted the strategy's potential to advance screening practices, public health policies, and trial design by enabling phenomapping, estimating subgroup-specific risk profiles, and establishing high-performance, cost-effective CCP for resource-limited regions. They noted its capability for testing effectiveness and cost-effectiveness in well-designed, prospective studies.

## Discussion

This study represents one of the largest EHR analyses to date, employing diverse machine learning methods and robust validation approaches to classify subgroups and predict CIN2+/CIN3+ risk. By leveraging comprehensive and interconnected phenotypic features, our study identified 5 CCP subgroups with varying risks. CCP2–4 had significantly higher risks of CIN2+ (CCP2: OR 2.07 [95% CI: 2.03-2.12], CCP3: 3.88 [3.78-3.97], CCP4: 4.47 [4.33-4.63]) and CIN3+ (CCP2: 2.10 [2.05-2.14], CCP3: 3.92 [3.82-4.02], CCP4: 4.45 [4.31-4.61]) compared to CCP1 ($P$<.001), consistent with the direction of results observed in the external cohort. Our findings offer a robust foundation for the proposed triple SCREENing strategy. This tailored approach prioritizes high-risk subgroups, providing actionable insights for cervical cancer prevention, particularly in LMICs.

To this end, various approaches, including single-cell transcriptomic analysis [49], sequence and phylogenetic analysis [50,51], and cluster analysis [48,52-54], have been employed to define genetic and cellular heterogeneity [55] in cervical cancer. These studies suggest that distinct subpopulations exhibit heterogeneous risk effects, linking individual features to varying absolute cervical cancer risk. Our 5 CCP subgroups align with findings from 2 major population-based clustering studies [56,57]. One study used a Poisson regression–based CEM clustering algorithm [56] to identify clusters of Indian states with similar cervical cancer incidence patterns. However, due to its focus on approximating missing data on sexual behavior, HPV prevalence, or cervical cancer incidence [56], the clusters were not ideal for screening purposes. That study did not provide sufficient details to estimate cluster-specific risks, and its reliance on features not typically available in routine screening limited real-world applicability. Additionally, strict model assumptions and the inclusion of relatively few features further constrained its generalizability. Another study performed hierarchical clustering of HPV-related methylation sites to identify subgroups of patients with cervical cancer [57]. While informative for prognosis and clinical management, the 3 clusters were not designed for cervical cancer screening. Instead, they are more suitable for guiding prognosis assessment, refining risk stratification, and optimizing treatment strategies in clinical practice.

In contrast, our study, specifically designed for cervical cancer screening, identified complex interactions among factors such as the number of infections, screening history, gynecological examination, hrHPV and HPV-16/18 infections, cervical cytology, time since previous screening, age, and possible hrHPV infection as key determinants of the 5 CCP subgroups with varying CIN2+/CIN3+ risks. Compared to previous studies [56,57], our design and identified subgroups are more representative of real-world cervical cancer screening. Leveraging the comprehensive and interconnected phenotypic features from EHRs, our algorithm extracted CCP-specific risk estimates, providing a robust foundation for the proposed triple SCREENing strategy. This strategy prioritizes subgroups with differential CIN2+/CIN3+ risks, offering a tailored approach to CCP. We recommend implementing SCREENing as a supplemental tool to existing guidelines [58-60], while accounting for the unique priorities and constraints of LMICs. SCREENing enables clinicians to perform effective, risk-based screening, followed by adequate diagnosis, surveillance, and management, while empowering policy makers to optimize public health policies and resource allocation. This approach has the potential to mitigate resource shortages in LMICs, reduce delays in diagnosis and treatment, and enhance screening efficiency by focusing efforts on high-risk populations, ultimately maximizing population-level benefits. Additionally, our analysis highlights the modifying effect of age on CCP-specific CIN2+/CIN3+ risks, emphasizing the need for greater attention to menopause and age [61] in cervical cancer screening strategies.

Methodologically, our study advances the external validation of machine learning algorithms for identifying subgroups and predicting cervical precancer or cancer risk using large-scale routine EHRs—a rarity in previous studies, which have often been limited to small samples. Our robust, structured framework of internal, external, and diagnostic validation enhances the acceptability and generalizability of unsupervised and supervised machine learning approaches in routine CCP and is adaptable to other diseases. The identified subgroups demonstrated good accuracy and diagnostic validity for CIN2+/CIN3+ both within and across datasets, though performance was lower in scenarios with missing key features. Differences in diagnostic validity across datasets may reflect variations in HPV infection patterns and previous screening histories, which influence the distribution of risk factors. Our findings of 5 subgroups and the proposed strategy are novel, offering a framework for assessing validity in screening, follow-up surveillance, and treatment for cervical cancer. This study signals the potential for more effective and targeted approaches to CCP.

This study represents one of the largest EHR analyses to date, employing multiple machine learning methods, datasets, and validation approaches to classify subgroups and predict CIN2+/CIN3+ risk. However, there are several limitations. First, a major limitation of this study is that our models were not externally validated beyond China. HPV infection patterns and the epidemiology of cervical abnormalities differ across countries and regions, which may impact the generalizability of our models. Given this variability, and although we used data from 5 distinct regions within China, we strongly recommend further validation of our models' performance

before applying them in settings not included in this study. We acknowledge that externally validating the model in additional multicenter studies worldwide is crucial for assessing its transferability and applicability across different clinical settings. Additionally, the tool may face challenges in fully adapting to various real-world scenarios when implemented outside the controlled validation environment of this study. To address these concerns, we plan to conduct independent external and prospective validations, as well as pilot implementation across diverse clinical settings in future studies. These efforts will evaluate the model's effectiveness using a wider range of data, particularly in the context of real-world cervical cancer prevention across regions beyond those included in this study. We also encourage independent researchers to validate our model in their own settings, where feasible. While this study marks a significant step in the development of the SCREENing tool, we acknowledge that further research and validation are necessary to establish its effectiveness in real-world applications.

Second, we acknowledge that the use of retrospective EHR data may introduce biases and may not fully capture the real-time challenges of clinical practice. These challenges include issues such as missing data, data errors, and the heterogeneity of EHR systems across different settings. Although we made efforts to control for the quality of the EHR data, it remains inherently difficult to fully address these concerns without significant infrastructure changes, as well as ongoing monitoring and data validation. Therefore, we explicitly caution that the findings of this study should be interpreted with caution. In light of this limitation, future studies should aim to bridge the gap between retrospective analyses and the practical challenges of data collection in clinical settings. Third, while this study focused on routine EHR data, we acknowledge that incorporating multi-omics data—such as genomic, proteomic, and imaging data—could further enhance the model's robustness and performance, providing valuable new insights into cervical abnormalities. We propose this as an area for future research to better understand the progression of cervical cancer and to offer novel perspectives on controlling and eliminating cervical cancer. Fourth, due to the limitations of routinely collected EHR content, we were unable to include several key sexual behavior features, such as the number of sexual partners and oral contraceptive use [62], in our analyses. These factors are known to significantly influence HPV infection and cervical cancer risk. While the model in this study demonstrated high performance, the absence of such data constrained our ability to further analyze the impact of sexual behaviors on findings, such as subgroup characterization, which could provide valuable insights. We recommend that future studies incorporate these data to validate our findings and offer a more comprehensive understanding of this field.

In conclusion, this study underscores the potential of leveraging machine learning algorithms and large-scale routine EHRs to enhance cervical cancer prevention strategies. By identifying key determinants of CIN2+/CIN3+ risk and classifying 5 distinct subgroups, our study provides a robust, data-driven foundation for the proposed triple SCREENing strategy. This approach prioritizes tailored prevention efforts for subgroups with varying risks, offering a novel and scalable tool to complement existing cervical cancer screening guidelines. Future work should focus on independent external and prospective validation to maximize the global impact of this strategy.

## Acknowledgments

We thank all the contributors for their efforts in this study. We also thank all funding sources for their funding. This research was funded by Fujian Province's Third Batch of Flexible Introduction of High-Level Medical Talent Teams (TD202307), Merck Investigator Studies Program [100073, 59484], National Natural Science Foundation of China (82271658), Major Scientific Research Program for Young and Middle-aged Health Professionals of Fujian Province, China (2021ZQNZD011), Fujian Province Central Government-Guided Local Science and Technology Development Project (2023L3019), Fujian Provincial Health Technology Project (2024GGA061), and Fujian Provincial Science and Technology Innovation Joint Fund (2021Y9171).

## Data Availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

## Authors' Contributions

ZL conceived and designed the study in consultation with HZ. ZL analyzed the data and interpreted the results. ZL drafted the manuscript with all authors (BD, HC, TT, JW, LF, BW, WZ, TY, XT, JW, SL, XH, ZZ, HX, SX, SL, PS, and HZ) critically reviewing the manuscript. HZ and SL provided administrative and material support, as well as supervision and mentorship.

Each author contributed important intellectual content during drafting or revision of the manuscript and accepts accountability for the overall work by ensuring that questions pertaining to the accuracy or integrity of any portion of the work are appropriately investigated and resolved.

All authors had access to all the data in this study and approved the final version of the manuscript. The corresponding authors attests that all the listed authors meet the authorship criteria and that no others meeting the criteria have been omitted. We confirm that no generative AI tools, such as ChatGPT or other language models, were used in the writing or preparation of this manuscript.

S Liu, PS, and HZ are co-corresponding authors.

The complete address for the three co-corresponding authors are as follows:

S Liu: West Building 1 Medical Garden 4th floor, Gongchang Road 66, Sun Yat-sen University (Shenzhen Campus) Shenzhen, Guangming District, 518107, China; liusy99@mail.sysu.edu.cn (SL); +86 135 4329 6396.
PS: 18 Daoshan Road, Fuzhou 350001, Fujian, P.R. China (Phone: +86-591-87558732; Fax: +86-591-87551247; email: fmsun1975@fjmu.edu.cn).
HZ: Room 435, Bld #8, 130 Dongan Road, Xuhui District, Shanghai (200032), China; zouhuachun@fudan.edu.cn (HZ); +86 21 5423 7707.
ZL, BD, HC, TT, and Junfeng Wang are co-first authors.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Data Supplement.
[DOCX File (Microsoft Word File), 380 KB-Multimedia Appendix 1]

## References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Chen YC, Chen YY, Su SY, et al. Projected time for the elimination of cervical cancer under various intervention scenarios: age-period-cohort macrosimulation study. JMIR Public Health Surveill. Apr 18, 2024;10:e46360. [doi: 10.2196/46360] [Medline: 38635315]
3. Lama Y, Hu D, Jamison A, Quinn SC, Broniatowski DA. Characterizing trends in human papillomavirus vaccine discourse on Reddit (2007-2015): an observational study. JMIR Public Health Surveill. Mar 27, 2019;5(1):e12480. [doi: 10.2196/12480] [Medline: 30916662]
4. Walboomers JMM, Jacobs MV, Manos MM, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J Pathol. Sep 1999;189(1):12-19. [doi: 10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F]
5. Lu Z, Zhou X, Fu L, et al. Prognostic prediction models for oropharyngeal squamous cell carcinoma (OPSCC): a protocol for systematic review, critical appraisal and meta-analysis. BMJ Open. Oct 12, 2023;13(10):e073375. [doi: 10.1136/bmjopen-2023-073375] [Medline: 37827742]
6. Yang SP, Lin XY, Hu M, Cai CF. The prognostic and predictive effects of human papillomavirus status in hypopharyngeal carcinoma: population-based study. JMIR Public Health Surveill. Dec 16, 2022;8(12):e40185. [doi: 10.2196/40185] [Medline: 36525304]
7. Yang L, Lu Z, Bian J, Li F, Zou H. Association between chronic viral infection-related hospitalization and risk of cardiovascular disease: a population-based cohort study. J Med Virol. Jan 2024;96(1):e29350. [doi: 10.1002/jmv.29350] [Medline: 38180233]
8. WHO. Global strategy to accelerate the elimination of cervical cancer as a public health problem. 2020. URL: https://www.who.int/publications/i/item/9789240014107
9. Bruni L, Serrano B, Roura E, et al. Cervical cancer screening programmes and age-specific coverage estimates for 202 countries and territories worldwide: a review and synthetic analysis. Lancet Glob Health. Aug 2022;10(8):e1115-e1127. [doi: 10.1016/S2214-109X(22)00241-8] [Medline: 35839811]
10. Gravitt PE, Silver MI, Hussey HM, et al. Achieving equity in cervical cancer screening in low- and middle-income countries (LMICs): strengthening health systems using a systems thinking approach. Prev Med. Mar 2021;144:106322. [doi: 10.1016/j.ypmed.2020.106322]
11. Arbyn M, Simon M, Peeters E, et al. 2020 list of human papillomavirus assays suitable for primary cervical cancer screening. Clin Microbiol Infect. Aug 2021;27(8):1083-1095. [doi: 10.1016/j.cmi.2021.04.031]
12. Poljak M, Oštrbenk Valenčak A, Gimpelj Domjanič G, Xu L, Arbyn M. Commercially available molecular tests for human papillomaviruses: a global overview. Clin Microbiol Infect. Sep 2020;26(9):1144-1150. [doi: 10.1016/j.cmi.2020.03.033] [Medline: 32247892]
13. Arbyn M, Verdoodt F, Snijders PJF, et al. Accuracy of human papillomavirus testing on self-collected versus clinician-collected samples: a meta-analysis. Lancet Oncol. Feb 2014;15(2):172-183. [doi: 10.1016/S1470-2045(13)70570-9] [Medline: 24433684]
14. Mezei AK, Armstrong HL, Pedersen HN, et al. Cost-effectiveness of cervical cancer screening methods in low- and middle-income countries: a systematic review. Int J Cancer. Aug 1, 2017;141(3):437-446. [doi: 10.1002/ijc.30695] [Medline: 28297074]

15. Camara H, Nosi S, Munnull G, et al. Women's acceptability of a self-collect HPV same-day screen-and-treat program in a high burden setting in the Pacific. BMC Health Serv Res. Dec 12, 2022;22(1):1514. [doi: 10.1186/s12913-022-08842-1] [Medline: 36510192]
16. Ronco G, Dillner J, Elfström KM, et al. Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials. The Lancet. Feb 2014;383(9916):524-532. [doi: 10.1016/S0140-6736(13)62218-7]
17. Arbyn M, Ronco G, Anttila A, et al. Evidence regarding human papillomavirus testing in secondary prevention of cervical cancer. Vaccine (Auckl). Nov 20, 2012;30 Suppl 5:F88-99. [doi: 10.1016/j.vaccine.2012.06.095] [Medline: 23199969]
18. Bouvard V, Wentzensen N, Mackie A, et al. The IARC perspective on cervical cancer screening. N Engl J Med. Nov 11, 2021;385(20):1908-1918. [doi: 10.1056/NEJMsr2030640] [Medline: 34758259]
19. WHO. Guidelines for screening and treatment of precancerous lesions for cervical cancer prevention. 2021. URL: https://www.who.int/publications/i/item/9789240030824
20. Mungo C, Guliam A, Chinula L, et al. Comparison of the ScreenFire and Xpert HPV assays for the detection of human papillomavirus and cervical precancer among women living with HIV in Malawi. medRxiv. Feb 23, 2024:2024.02.21.24303142. [doi: 10.1101/2024.02.21.24303142] [Medline: 38617305]
21. WHO. Comprehensive cervical cancer control: a guide to essential practice. 2014.
22. Cheung LC, Egemen D, Chen X, et al. 2019 ASCCP risk-based management consensus guidelines: methods for risk estimation, recommended management, and validation. J Low Genit Tract Dis. Apr 2020;24(2):90-101. [doi: 10.1097/LGT.0000000000000528] [Medline: 32243306]
23. Zou H, Lu Z, Weng W, et al. Diagnosis of neurosyphilis in HIV-negative patients with syphilis: development, validation, and clinical utility of a suite of machine learning models. EClinicalMedicine. Aug 2023;62:102080. [doi: 10.1016/j.eclinm.2023.102080] [Medline: 37533423]
24. Oikonomou EK, Spatz ES, Suchard MA, Khera R. Individualising intensive systolic blood pressure reduction in hypertension using computational trial phenomaps and machine learning: a post-hoc analysis of randomised clinical trials. Lancet Digit Health. Nov 2022;4(11):e796-e805. [doi: 10.1016/S2589-7500(22)00170-4] [Medline: 36307193]
25. Melnikow J, Henderson JT, Burda BU, Senger CA, Durbin S, Weyrich MS. Screening for cervical cancer with high-risk human papillomavirus testing: updated evidence report and systematic review for the US Preventive Services Task Force. J Am Med Assoc. Aug 21, 2018;320(7):687-705. [doi: 10.1001/jama.2018.10400] [Medline: 30140883]
26. Xiao T, Wang C, Yang M, et al. Use of virus genotypes in machine learning diagnostic prediction models for cervical cancer in women with high-risk human papillomavirus infection. JAMA Netw Open. Aug 1, 2023;6(8):e2326890. [doi: 10.1001/jamanetworkopen.2023.26890] [Medline: 37531108]
27. van der Waal D, Bekkers RLM, Dick S, et al. Risk prediction of cervical abnormalities: the value of sociodemographic and lifestyle factors in addition to HPV status. Prev Med. Jan 2020;130:105927. [doi: 10.1016/j.ypmed.2019.105927]
28. Rothberg MB, Hu B, Lipold L, et al. A risk prediction model to allow personalized screening for cervical cancer. Cancer Causes Control. Mar 2018;29(3):297-304. [doi: 10.1007/s10552-018-1013-4]
29. Lee CH, Peng CY, Li RN, et al. Risk evaluation for the development of cervical intraepithelial neoplasia: development and validation of risk-scoring schemes. Int J Cancer. Jan 15, 2015;136(2):340-349. [doi: 10.1002/ijc.28982] [Medline: 24841989]
30. Bouvard V, Baan R, Straif K, et al. A review of human carcinogens--Part B: biological agents. Lancet Oncol. Apr 2009;10(4):321-322. [doi: 10.1016/s1470-2045(09)70096-8] [Medline: 19350698]
31. Schreiberhuber L, Barrett JE, Wang J, et al. Cervical cancer screening using DNA methylation triage in a real-world population. Nat Med. Aug 2024;30(8):2251-2257. [doi: 10.1038/s41591-024-03014-6] [Medline: 38834848]
32. Marcus JZ, Cason P, Downs LS, Einstein MH, Flowers L. The ASCCP Cervical Cancer Screening Task Force Endorsement and Opinion on the American Cancer Society updated cervical cancer screening guidelines. J Low Genit Tract Dis. Jul 1, 2021;25(3):187-191. [doi: 10.1097/LGT.0000000000000614] [Medline: 34138787]
33. Moy LM, Zhao FH, Li LY, et al. Human papillomavirus testing and cervical cytology in primary screening for cervical cancer among women in rural China: comparison of sensitivity, specificity, and frequency of referral. Int J Cancer. Aug 1, 2010;127(3):646-656. [doi: 10.1002/ijc.25071] [Medline: 19960441]
34. Saslow D, Solomon D, Lawson HW, et al. American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. Am J Clin Pathol. Apr 2012;137(4):516-542. [doi: 10.1309/AJCPTGD94EVRSJCG] [Medline: 22431528]
35. Gower JC. A general coefficient of similarity and some of its properties. Biometrics. Dec 1971;27(4):857. [doi: 10.2307/2528823]

36. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. arXiv. Preprint posted online on 2018. arXiv:180203426
37. Reese JT, Blau H, Casiraghi E, et al. Generalisable long COVID subtypes: findings from the NIH N3C and RECOVER programmes. eBioMedicine. Jan 2023;87:104413. [doi: 10.1016/j.ebiom.2022.104413] [Medline: 36563487]
38. Zhou W, He MM, Wang F, Xu RH, Wang F, Zhao Q. Latent class analysis-derived classification improves the cancer-specific death stratification of molecular subtyping in colorectal cancer. NPJ Precis Oncol. Jun 23, 2023;7(1):60. [doi: 10.1038/s41698-023-00412-w] [Medline: 37353681]
39. Faghri F, Brunn F, Dadu A, et al. Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study. Lancet Digit Health. May 2022;4(5):e359-e369. [doi: 10.1016/S2589-7500(21)00274-0] [Medline: 35341712]
40. Lu Z, Fu L, Yang L, et al. Hesitancy to undergo SARS-CoV-2 rapid antigen testing in China: nationwide cross-sectional study. JMIR Public Health Surveill. May 1, 2023;9:e43555. [doi: 10.2196/43555] [Medline: 36888911]
41. Xu E, Xie Y, Al-Aly Z. Long-term neurologic outcomes of COVID-19. Nat Med. Nov 2022;28(11):2406-2415. [doi: 10.1038/s41591-022-02001-z]
42. Chen T, Guestrin C, editors. XGBoost: a scalable tree boosting system. Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.
43. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. Jan 2020;2(1):56-67. [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
44. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. Dec 16, 2016;18(12):e323. [doi: 10.2196/jmir.5870] [Medline: 27986644]
45. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD statement. Ann Intern Med. Jan 6, 2015;162(1):55-63. [doi: 10.7326/M14-0697]
46. Lu Z. lulab.utils: Supporting Functions Maintained by Zhen Lu. 2024. URL: https://CRAN.R-project.org/package=lulab.utils
47. Lu Z. SCREENing. URL: https://github.com/Leslie-Lu/SCREENing
48. Rosty C, Sheffer M, Tsafrir D, et al. Identification of a proliferation gene cluster associated with HPV E6/E7 expression level and viral DNA load in invasive cervical carcinoma. Oncogene. Oct 27, 2005;24(47):7094-7104. [doi: 10.1038/sj.onc.1208854] [Medline: 16007141]
49. Li Y, Wang C, Ma A, et al. Identification of HPV oncogene and host cell differentiation associated cellular heterogeneity in cervical cancer via single-cell transcriptomic analysis. J Med Virol. Aug 2023;95(8). [doi: 10.1002/jmv.29060]
50. Sait K, Turki R, Abuzenadah AM, Jiffiri OH, Bohmaidah A, Sohrab SS. Genetic diversity and phylogenetic analysis of HPV 16 & 18 variants isolated from cervical specimens of women in Saudi Arabia. Saudi J Biol Sci. Feb 2019;26(2):317-324. [doi: 10.1016/j.sjbs.2018.05.005] [Medline: 31485171]
51. Balanda M, Fernández J, Vergara N, et al. Genetic variability of human papillomavirus type 66 L1 gene among women presenting for cervical cancer screening in Chile. Med Microbiol Immunol. Dec 2019;208(6):757-771. [doi: 10.1007/s00430-019-00621-w] [Medline: 31165237]
52. Wan F, Miao X, Quraishi I, Kennedy V, Creek KE, Pirisi L. Gene expression changes during HPV-mediated carcinogenesis: a comparison between an in vitro cell model and cervical cancer. Int J Cancer. Jul 1, 2008;123(1):32-40. [doi: 10.1002/ijc.23463] [Medline: 18398830]
53. Vázquez-Ortíz G, Ciudad CJ, Piña P, et al. Gene identification by cDNA arrays in HPV-positive cervical cancer. Arch Med Res. 2005;36(5):448-458. [doi: 10.1016/j.arcmed.2005.04.016] [Medline: 16099320]
54. Widschwendter A, Gattringer C, Ivarsson L, et al. Analysis of aberrant DNA methylation and human papillomavirus DNA in cervicovaginal specimens to detect invasive cervical cancer and its precursors. Clin Cancer Res. May 15, 2004;10(10):3396-3400. [doi: 10.1158/1078-0432.CCR-03-0143] [Medline: 15161694]
55. Shiri Aghbash P, Shirvaliloo M, Khalo Abass Kasho A, Alinezhad F, Nauwynck H, Bannazadeh Baghi H. Cluster of differentiation frequency on antigen presenting-cells: the next step to cervical cancer prognosis? Int Immunopharmacol. Jul 2022;108:108896. [doi: 10.1016/j.intimp.2022.108896] [Medline: 35640377]
56. Man I, Georges D, Bonjour M, Baussano I. Approximating missing epidemiological data for cervical cancer through footprinting: a case study in India. Elife. May 25, 2023;12:e81752. [doi: 10.7554/eLife.81752] [Medline: 37227260]
57. Yang S, Wu Y, Wang S, et al. HPV-related methylation-based reclassification and risk stratification of cervical cancer. Mol Oncol. Sep 2020;14(9):2124-2141. [doi: 10.1002/1878-0261.12709] [Medline: 32408396]
58. Li M, Wei L, Sui L, et al. Guidelines for cervical cancer screening in China. Gynecol Obstet Clin Med. Dec 2023;3(4):189-194. [doi: 10.1016/j.gocm.2023.10.005]

59. Sawaya GF, Smith-McCune K, Kuppermann M. Cervical cancer screening: more choices in 2019. J Am Med Assoc. May 28, 2019;321(20):2018-2019. [doi: 10.1001/jama.2019.4595] [Medline: 31135834]
60. Wang W, Arcà E, Sinha A, Hartl K, Houwing N, Kothari S. Cervical cancer screening guidelines and screening practices in 11 countries: a systematic literature review. Prev Med Rep. Aug 2022;28:101813. [doi: 10.1016/j.pmedr.2022.101813] [Medline: 35637896]
61. Hickey M, Basu P, Sassarini J, et al. Managing menopause after cancer. The Lancet. Mar 2024;403(10430):984-996. [doi: 10.1016/S0140-6736(23)02802-7]
62. Gierisch JM, Coeytaux RR, Urrutia RP, et al. Oral contraceptive use and risk of breast, cervical, colorectal, and endometrial cancers: a systematic review. Cancer Epidemiol Biomarkers Prev. Nov 1, 2013;22(11):1931-1943. [doi: 10.1158/1055-9965.EPI-13-0298]

## Abbreviations

**CCP:** cervical cancer prevention
**CIN:** cervical intraepithelial neoplasia
**EHR:** electronic health record
**HPV:** human papillomavirus
**hrHPV:** high-risk HPV
**LMIC:** low- and middle-income country
**OR:** odds ratio
**OVR AUROC:** one-versus-rest area under the receiver operating characteristics curve
**SCREENing:** clinical Subgroups for CeRvical cancEr prEvention using computational pheNomaps and machine learnING
**SHAP:** SHapley Additive exPlanations
**UMAP:** uniform manifold approximation and projection
**WHO:** World Health Organization
**XGBoost:** eXtreme gradient boosting

*Edited by Amaryllis Mavragani; peer-reviewed by Gabriel Guerrero-Contreras, Zhi-Ping Liu; submitted 22.10.2024; final revised version received 29.01.2025; accepted 29.01.2025; published 19.03.2025*

*Please cite as:*
*Lu Z, Dong B, Cai H, Tian T, Wang J, Fu L, Wang B, Zhang W, Lin S, Tuo X, Wang J, Yang T, Huang X, Zheng Z, Xue H, Xu S, Liu S, Sun P, Zou H*
*Identifying Data-Driven Clinical Subgroups for Cervical Cancer Prevention With Machine Learning: Population-Based, External, and Diagnostic Validation Study*
*JMIR Public Health Surveill 2025;11:e67840*
*URL: https://publichealth.jmir.org/2025/1/e67840*
*doi: 10.2196/67840*

© Zhen Lu, Binhua Dong, Hongning Cai, Tian Tian, Junfeng Wang, Leiwen Fu, Bingyi Wang, Weijie Zhang, Shaomei Lin, Xunyuan Tuo, Juntao Wang, Tianjie Yang, Xinxin Huang, Zheng Zheng, Huifeng Xue, Shuxia Xu, Siyang Liu, Pengming Sun, Huachun Zou. Originally published in JMIR Public Health and Surveillance (https://publichealth.jmir.org), 19.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Public Health and Surveillance, is properly cited. The complete bibliographic information, a link to the original publication on https://publichealth.jmir.org, as well as this copyright and license information must be included.