Frontiers in Oncology

TYPE Original Research
PUBLISHED 17 January 2024
DOI 10.3389/fonc.2023.1289030

OPEN ACCESS

EDITED BY
Manoj Menon,
Fred Hutchinson Cancer Center, United States

REVIEWED BY
Manuela Tamburro,
University of Molise, Italy
Matteo Pavone,
Agostino Gemelli University Polyclinic (IRCCS), Italy

*CORRESPONDENCE
Yali Sun
✉ syl197210@beihua.edu.cn

RECEIVED 05 September 2023
ACCEPTED 18 December 2023
PUBLISHED 17 January 2024

CITATION
Gong L, Tang Y, Xie H, Zhang L and Sun Y (2024) Predicting cervical intraepithelial neoplasia and determining the follow-up period in high-risk human papillomavirus patients. *Front. Oncol.* 13:1289030. doi: 10.3389/fonc.2023.1289030

Predicting cervical intraepithelial neoplasia and determining the follow-up period in high-risk human papillomavirus patients

Ling Gong[1], Yingxuan Tang[2], Hua Xie[3], Lu Zhang[3] and Yali Sun[1]*

[1]Department of Nursing, School of Nursing, Beihua University, Jilin, China, [2]Department of Computer Science and Technology, School of Computer Science, Northeast Electric Power University, Jilin, China, [3]Department of Gynecology, Jilin Central General Hospital, Jilin, China

Purpose: Despite strong efforts to promote human papillomavirus (HPV) vaccine and cervical cancer screening, cervical cancer remains a threat to women's reproductive health. Some high-risk HPV types play a crucial role in the progression of cervical cancer and precancerous lesions. Therefore, HPV screening has become an important means to prevent, diagnose, and triage cervical cancer. This study aims to leverage artificial intelligence to predict individual risks of cervical intraepithelial neoplasia (CIN) in women with high-risk HPV infection and to recommend the appropriate triage strategy and follow-up period according to the risk level.

Materials and methods: A total of 475 cases were collected in this study. The sources were from the Department of Gynecology and Obstetrics in a tertiary hospital, a case report on HPV from the PubMed website, and clinical data of cervical cancer patients from The Cancer Genome Atlas (TCGA) database. Through in-depth study of the interaction between high-risk HPV and its risk factors, the risk factor relationship diagram structure was constructed. A Classification of Lesion Stages (CLS) algorithm was designed to predict cervical lesion stages. The risk levels of patients were analyzed based on all risk factors, and follow-up periods were formulated for each risk level.

Results: Our proposed CLS algorithm predicted the probability of occurrence of CIN3—the precancerous lesion stage of cervical cancer. This prediction was based on patients' HPV-16 and -18 infection status, age, presence of persistent infection, and HPV type. Follow-up periods of 3–6 months, 6–12 months, and 3- to 5-year intervals were suggested for high-risk, medium-risk, and low-risk patients, respectively.

Conclusion: A lesion prediction model was constructed to determine the probabilities of occurrence of CIN by analyzing individual data, such as patient lifestyle, physical assessments, and patient complaints, in order to identify high-risk patients. Furthermore, the potential implications of the calculated features were mined to devise prevention strategies.

KEYWORDS

follow-up period, human papillomavirus, cervical intraepithelial neoplasia, prediction, genotype, cervical cancer

1 Introduction

Global statistics on cancer in women indicate that cervical cancer ranked fourth in both incidence and mortality rate in 2012 (1). In 2020, there were over 604,000 cervical cancer cases and 341,000 deaths worldwide (2). The efficacy of vaccination and screening in preventing cervical cancer has been established, leading to increased awareness and participation in prevention programs among women. However, globally, the incidence and mortality rates of cervical cancer remain substantially higher in low-income and middle-income countries than in high-income countries; this is attributed to the lack of vaccination coverage, high-quality screening, timely treatment, and follow-up care services. A priority for public health managers worldwide is to take proactive measures to address the need for continuous and improved prevention and monitoring of cervical cancer. This aligns with the targets of the World Health Organization elimination initiative launched in 2020 to reduce cervical cancer incidence to below four cases per 100,000 women-years in every country (3). Furthermore, advancements in effective disease prediction and diagnosis are crucial for accurately identifying the target population.

Persistent high-risk HPV infection is recognized as the primary cause of CIN and cervical cancer. The pathogenesis of cervical cancer involves a prolonged period of development of precancerous lesions, such as the CIN1, CIN2, and CIN3 stages. The risk of developing invasive cervical cancer associated with CIN 1, CIN2, and CIN3 is 4 times, 14.5 times, and 46.5 times, respectively, higher than that of non-CIN. While most CIN 1 lesions resolve naturally, CIN2 and CIN3 incur the risk of malignant transformation (4–6). Studies, including randomized clinical trials, have indicated that HPV-based screening—characterized by high sensitivity and long-term negative predictive value—plays a significant role in primary screening methods, along with cervical cytology, in identifying potential cervical cancer cases and triage (7–9). Additionally, electronic colposcopy of the cervix and cervical biopsy are employed to determine the cervical lesion stage based on primary screening results. However, it is not advisable for all patients to directly undergo biopsy due to its associated low detection rate, wastage of medical resources, and invasive nature of biopsies. Therefore, accurate prediction of the risk of cervical lesions holds crucial clinical implications for early diagnosis and prevention of cervical cancer.

There are still some challenges in predicting cervical cancer, such as missing data in medical records and transient HPV infection. Poor data quality affects the accuracy of prediction. The uncertainty of the prediction model and the deficiencies in the data would lead to poor performance of the model during prediction and affect the reliability of the prediction results. In recent years, artificial intelligence (AI) has been gradually applied in the field of clinical medicine, especially in disease diagnosis and detection, for greater ability of learning and strong potentials in data processing (10–12). The application of AI is conducive to reducing the rate of missed diagnoses, saving more time, and improving accuracy for clinicians. AI technology has greatly improved the diagnostic accuracy of lung cancer and breast cancer through training CT and ultrasound images (13, 14). AI liquid-based cytology has resulted in efficient referrals to colposcopy, with higher specificity than manual screening methods (15). The Colposcopic Artificial Intelligence Auxiliary Diagnostic System has been explored to classify colposcopic impressions and suggest biopsies (16). AI technology can not only overcome the limitations of doctors' subjective judgment and personal biases in diagnosis but also improve the accuracy of diagnosis and help to locate the lesion site (17–20). In the context of driving continuous progress in medical technology, there is an urgent need for an efficient and accurate method to determine the probabilities of occurrence of CIN through analysis of individual data such as information on lifestyle, physical assessments, and complaints so that high-risk patients can be identified and the potential implications of calculated features can be mined for further prevention strategies.

2 Materials and methods

2.1 Study design

The pathogenesis of cervical cancer usually involves a long period during which precancerous lesions (such as CIN3) form, mainly caused by persistent infection with high-risk HPV. The aim of this research is to achieve early detection of the predisposing factors for precancerous lesions, based on high-risk HPV infection, and implementation of preventive patient interventions. Data preprocessing—including dataset construction and mapping and mining of impact factors, along with the CLS algorithm proposed in our research—enabled prediction of cervical lesions, exploration of predictive indicators, and risk classification of CIN. The findings yield valuable suggestions for the formulation of guidelines for patient follow-up periods at all levels and for advance implementation of preventive interventions, to effectively enable precise prevention strategies and reduce the probability of occurrence of cervical cancer (Figure 1).

2.2 Data preprocessing

2.2.1 Datasets

The experimental environment is as follows: Python 3.7, Neo4j, and NetworkX 2.1 are configured under a Windows 10 operating system. Three Hadoop-distributed clusters of the CentOS 7 operating system were built, namely, HDFS, YARN, and Spark on YARN. A dataset constructing structure of the diagram for risk factors was collected by using crawler tools from the PubMed website, searching high-risk HPV, cervical cancer, HPV risk factor, and other similar terms, as literature retrieval words. The search yielded 2,221 pieces of medical literature.

The case data were collected mainly on the basis of cases with high-risk HPV infection and lesions, cases with high-risk HPV infection but no lesions, cases without high-risk HPV infection but lesions (i.e., cases that tested HPV-negative, but with lesions), and cases without high-risk HPV infection and no lesions.

A total of 475 cases were collected in this study. The sources were as follows: the Department of Gynecology and Obstetrics at

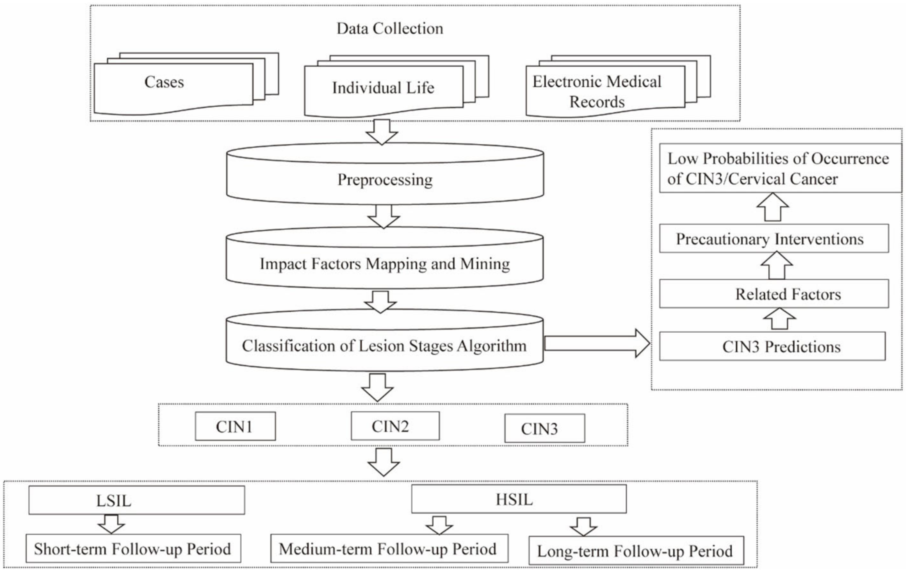
Data Collection
Cases
Individual Life
Electronic Medical Records
Preprocessing
Impact Factors Mapping and Mining
Classification of Lesion Stages Algorithm
Low Probabilities of Occurrence of CIN3/Cervical Cancer
Precautionary Interventions
Related Factors
CIN3 Predictions
CIN1
CIN2
CIN3
LSIL
HSIL
Short-term Follow-up Period
Medium-term Follow-up Period
Long-term Follow-up Period

FIGURE 1
Mapping and mining of impact factors.

Jilin Central General Hospital, case report articles about HPV on the PubMed website, and clinical data for cervical cancer in The Cancer Genome Atlas (TCGA) database.

2.2.2 Mapping of key risk factors

The electronic medical record text, which is different from ordinary text, usually has a relatively complete structure, including patients' personal information, main complaints, personal history, physical examination results, and auxiliary examination results, with little noise data. Examples of the style of entries include "the patient had vaginal bleeding one month ago," "denied history of drug allergy," and "denied familial inherited diseases". Therefore, the set of keywords for patient case data can be obtained by natural language processing methods. Key words representing textual information were directly extracted—e.g., "vaginal bleeding" for "the patient had vaginal bleeding one month ago"—and numerical information was extracted according to the rules shown in Table 1.

TABLE 1 Extraction Rules for Abnormality of Risk Factors.

Risk Factors	Extraction Rules for Abnormality
Age	> 30 years
Age at Menarche	> 14 years or <12 years
Age at First Sexual Intercourse	< 18 years
Number of Sexual Partners	>2
Age at First Full-term Pregnancy	<18 years
Number of Vaginal Births	> 2
Number of Pregnancies	> 2

2.3 Method

2.3.1 Classification of the lesion stage algorithm

The challenges involved in predicting lesion stage by machine learning methods involve determining what kind of data and what kind of features to analyze and calculate. The corresponding test values for patients are commonly used for training and analysis in machine learning methods, which poses great obstacles due to insufficient amount of data. The larger the amount of data and the more values available in machine learning, the more accurate the training is. However, there are many missing values and few positive samples when collecting data, which causes failures of application of many disease prediction models. Therefore, the mechanism of the disease should be fully considered when selecting features to enable more accurate prediction. The Classification of Lesion Stages (CLS) algorithm proposed in this study gives full consideration to the pathogenic mechanism and selects appropriate features for analysis, which has practical significance for the prediction of cervical lesions.

2.3.2 Types of high-risk HPV and classification of lesion stages

There are more than 100 types of high-risk HPV; 16 common types, namely, HPV-16, -18, -58, -52, -31, -51, -33, -35, -56, -26, -39, -53, -66, -67, -70, and -45, were analyzed and used for the calculations in this study, as nodes in the structure of the risk factor graph and connected with many other factors.

Cervical lesions are divided into three grades: CIN1, CIN2, and CIN3. The CIN3 stage has a high probability of transformation into cervical cancer. In 2014, the World Health Organization reclassified it into low-grade squamous intraepithelial lesion (LSIL) and high-grade squamous intraepithelial lesion (HSIL), further simplifying the original classification. LSIL refers to the original CIN1 stage, and HSIL includes the original CIN2 and CIN3 stages. In this study, both classification methods were adopted in the analysis and calculation stage, which was conducive to more detailed analysis. We described the differences in neighbor risk factor nodes and neighbor HPV genotypes between CIN1, CIN2, and CIN3. The factors with node relation value greater than 3 were selected as close factors, among which differences were compared and the degree of difference was calculated.

2.3.3 Prediction of lesion stages

Based on the set of key risk factors, we extracted the risk factors that were abnormal in case history and the HPV types with which the patient was infected by natural language processing. According to the principle of abnormal extraction of risk factors, we identified key risk factors for patients with abnormal p collection $AF_p = \{af_{1,} af_2, \ldots\ldots, af_n\}$, including patients' HPV types and their risk factors that were abnormal. The predictive value of a patient's classification relative to CIN1 was calculated by the following Equation 1.

$$CIN1_p = \sum_{m=1}^{n} W_{(AF_m, cin1)} \tag{1}$$

When the abnormal factors for a patient included those in *cin2Element* or *cin3Element*, it indicated that the patient had factors unique to CIN2 or CIN3. To describe this difference, the degree of difference was introduced to calculate the extent of difference of CIN2 or CIN3 relative to CIN1 in the current situation for each patient, using Equation 2. Abnormal factors as unique ones that appeared in CIN2 or CIN3 were remembered as $CIN2ELE_p = \{cin2Ele_{(p,1)}, cin2Ele_{(p,2)}, \ldots\ldots, cin2Ele_{(p,n)},\}$, $CIN3ELE_p = \{cin3Ele_{(p,1)}, cin3Ele_{(p,2)}, \ldots\ldots, cin3Ele_{(p,n)},\}$。

$$Diff_{(p,\ cin2)} = \frac{\sum_{m=1}^{n} W_{(cin2Ele_{(p,m)}, cin2)}}{n} \tag{2}$$

$W_{(cin2Ele_{(p,m)}, cin2)}$—connected edge weights of factor *of* $cin2Ele_{(p,m)}$ and CIN2 in risk factors—figure structure.

n— number of elements in $CIN2ELE_p Diff_{(p,cin2)}$—degree of difference in patients' p between CIN2 and CIN1.

The degree of difference in patients with p between CIN3 and CIN1 $Diff_{(p,cin2)}$ was calculated in the same way.

The degree of difference calculation should be introduced into the classification predicted value of CIN2 or CIN3, which is calculated by Equations 3, 4.

$$CIN2_p = Diff_{(p,\ cin2)} * \sum_{m=1}^{n} W_{(AF_m, cin2)} \tag{3}$$

$$CIN3_p = Diff_{(p,\ cin3)} * \sum_{m=1}^{n} W_{(AF_m, cin3)} \tag{4}$$

From the above calculation, the three classification predictive values of patient "p" can be obtained. In order to more accurately determine which category the patient belongs to, the risk level of the patient is introduced into the analysis. Patients at a low-risk level—which means that their risk of infection with high-risk HPV is very low—have low possibility of cervical lesion. Therefore, we predict that patients at a low-risk level will be disease-free (CIN−). In high-risk patients, i.e., those with a high risk of infection with high-risk HPV, the likelihood of lesions is also high. The prediction result with the largest predictive value of the three-stage classification is selected as the final prediction result. If the maximum value is the predicted value FOR CIN2 or CIN3, it is classified as HSIL; if the maximum value is the predicted value FOR CIN1, it is classified as LSIL. For intermediate-risk patients, this analysis is somewhat difficult, because for these patients, the risk level value is around 0.5, which represents an almost risk of occurrence of HSIL or LSIL. Patients in this category require more cautious management. In order to reduce the rate of missed diagnoses rate, degree of difference analysis is conducted in the present study. If the value of the degree of difference of patients with CIN2/CIN3 is greater than 2 at any stage, it is identified as a large difference and directly classified as HSIL because the possibility of CIN2/CIN3 stage is stronger. If the value of degree of difference is not greater than 2 at either stage of CIN2/CIN3, the patient does not have a high stage difference. In this scenario, identification as CIN1 and classification as LSIL is more likely.

3 Results

3.1 Follow-up period

After calculating risk levels for all patients, follow-up periods for patients at different risk levels were statistically analyzed (Figure 2). Each blue circle represents the suggested follow-up period for a patient at a high level of risk, green ones show follow-up periods for patients at medium risk, and beige ones represent suggested follow-up periods for patients at low risk. There was a clustering of data at different levels.

After summarizing the data for the above groups, the follow-up periods for patients at different risk levels were obtained. Respectively, for high-risk, medium-risk, and low-risk patients, follow-up at 3 to 6 months, 6 to 12 months, and 3 to 5 year intervals was suggested (Table 2).

3.2 Prediction of CIN

The 16 types of high-risk HPV, CIN1, CIN2, and CIN3 exist in the risk factor graph structure as nodes connected with many other factors, and the weight of the edge represents the closeness between them and the risk factors. The neighbor nodes in the graph structure were used to observe the relevant factors for different disease stages and the nodes' characteristics. The names and edge weights of key risk factors and high-risk HPV types that were directly related to CIN1, CIN2, and CIN3 were the output. CIN3 node's top-5 neighbor nodes and their relationship values are addressed as below (Table 3).

The risk classification of each patient warranted consideration. In addition, for lesions at different stages, their close risk factor neighbor nodes and the relationship value was different. Consideration of the difference of factors at different stages was conducive to better classification and prediction of patients' lesions.

The values of the relationship with the CIN3 weight of all the top-5 neighbor nodes were between 3 and 3.005, which means that

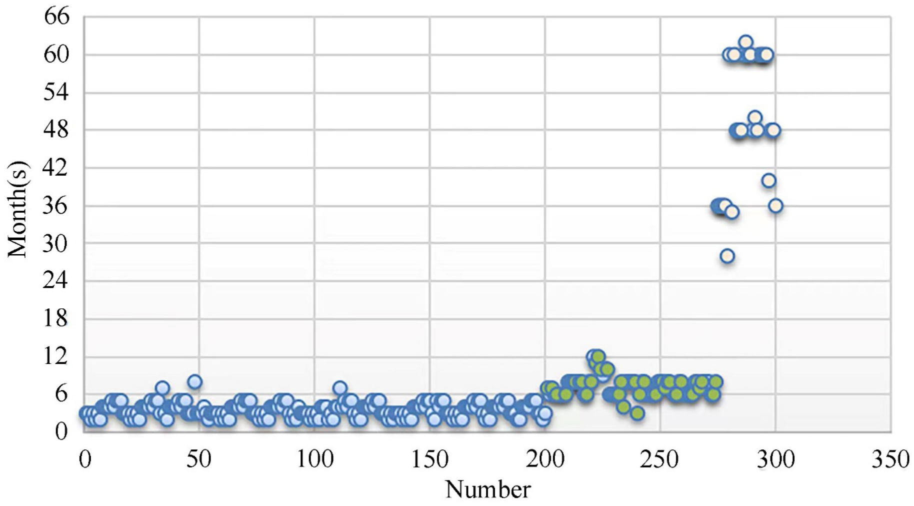

FIGURE 2
Follow-up periods for patients at the three risk levels (Blue: Low-risk Level, Green: Medium-risk Level, Beige: High-risk Level).

the factors were reliable predictive parameters for CIN3; in order, they were HPV-16, HPV-18, age, persistent HPV infection, and HPV type. HPV type and infection represent four of the five closest neighbor nodes of CIN3. Obviously, factors closely related to HPV made a large contribution to precancerous progression. The top two factors were the two high-risk genotypes 16 and 18, in line with current studies that consider them the predominant causes of precancer or cervical squamous cell carcinoma. In recent years, extant works have yielded similar results as our study: HPV subtypes in different age groups and different regions have different characteristics, according to epidemiological statistical data. Moreover, the differences also reflect the different levels of cervical lesions (21).

When analyzing the CLS and identifying related risk factors and high-risk HPV types, we found that different lesion stages had different correlations with high-risk HPV types. Three genotypes, mentioned in Table 4, describe CIN1-related high-risk HPV and relation value with CIN1. In descending order of risk, they are HPV-18, -16, and -45. The top two values are above 3.0, which is remarkably higher than the value for HPV 45. It is suggested that HPV 18 has the closest relationship with CIN1 and HPV 45 takes the third place with a relatively low value.

As shown in Table 5, CIN2-related high-risk HPV genotypes include those found in CIN1 as well as HPV 31. HPV 16 is the primary type. The relation values with CIN2 for HPV-16 and -18 are greater than 3.0, although the values for HPV-31 and -45 are just over 1.0. Evidently, HPV-16 and -18 are predominant factors leading to CIN2 among high-risk HPV genotypes. Despite the values for the other genotypes not being as high, HPV-31 and -45 emerge, among many other genotypes, as CIN2-relevant high-risk HPV genotypes.

Calculations implicate 14 genotypes as causes of CIN3 from the perspective of high-risk HPV (Table 6). They can be divided into three echelons according to relation value with CIN3 ≥3.0, ≥2.0 and<3.0, and ≥1.0 and<2.0. In the first echelon, HPV-16 and -18 display the most intimate relationship with CIN3. HPV-58 and -31 appear in the second echelon and HPV-52, -56, -66, -51, -39, -35, -33, -45, and -26 emerge in the third echelon, in descending order.

CIN3 was correlated with multiple high-risk HPV types; in other words, when these high-risk HPV types occur, there is a greater probability of development of CIN3. At the same time, we found that HPV-16 and -18 have a strong impact on each of the three stages. A number of studies over the years have also shown that these two HPV genotypes are associated with the highest risk of occurrence of lesions and even cervical cancer, and the three common types of cervical cancer vaccines inevitably cover these two genotypes. Compared with the CIN1 stage, it was found that the HPV31 genotype was a unique high-risk type for the CIN2 stage, indicating that upon infection with HPV31, the likelihood of development into the CIN2 stage is higher. High-risk HPV

TABLE 2 Follow-up Periods for Patients of Different Risk Levels.

Risk Level	Follow-up Period
Low-Risk	3-5 years
Medium-Risk	6-12 months
High-Risk	3-6 months

TABLE 3 CIN3 Node's Top 5 Neighbor Nodes.

Neighbor Nodes	Value of the Relationship with CIN3 Weight
HPV 16	3.004129552
HPV 18	3.00267266
Age	3.001884209
Persistent HPV Infection	3.001449165
HPV Type	3.000758473

TABLE 4 CIN1 Relevant High-risk HPV and Value.

Type of High-risk HPV	Relation Value with CIN1
HPV 18	3.001518086
HPV 16	3.000897878
HPV 45	1.000011528

infection warrants more attention. It also indicates that multiple genotypes of infection leads to greater likelihood of high-grade lesions.

3.3 Experimental analysis

We introduced an experimental evaluation index and conducted evidence-based analysis based on the diagram structure of risk factors. Finally, classification to predict the cervical lesion stage of patients and experimental verification through a total of 125 collected case data, excluding the data for cases that have developed into cervical cancer, was carried out. Comparative experimental analysis between the CLS algorithm proposed in this study and SMOTE-LSTM (22) was conducted.

At a statistical level, the results of disease diagnosis are described in terms of sensitivity and specificity. Sensitivity refers to the ability of diagnostic tests to detect disease when people are sick, as shown in the calculation Equation 5. Specificity refers to the ability of diagnostic tests to exclude disease when people are not sick, as shown in the calculation Equation 6.

TP (true positive): The prediction corresponds to the number of people diagnosed with a certain stage of the disease.

FP (false positive): The prediction does not correspond to the number of people diagnosed with a certain stage of the disease.

FN (false negative): The prediction does not correspond to the number of people free from disease.

TN (true negative): The prediction corresponds to the number of people free from disease.

$$sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$specificity = \frac{TN}{TN + FP} \tag{6}$$

TABLE 5 CIN2 Relevant High-risk HPV and Value.

Type of High-risk HPV	Relation Value with CIN2
HPV 16	3.003927177
HPV 18	3.003502006
HPV 31	1.00005263
HPV 45	1.000007755

Sensitivity and specificity are often used to evaluate the authenticity of outpatient results. In order to evaluate the classification results of disease prediction more accurately, the definition of true positive in this study has been modified. In general, true positive indicates the condition of finding disease and predicting disease; that is, patients with the disease are correctly predicted to be patients with the disease. However, in the study, true positive is to predict not disease but accurate disease stage. These changes were made to improve the accuracy of CLS.

The CLS algorithm put forward in this research and the SMOTE-LSTM algorithm were compared based on the two aspects of specificity and sensitivity. Sensitivity represents the ability to identify patients, and specificity represents the ability to identify non-patients, i.e., the ability to be assessed as disease-free.

The experimental results of lesion prediction are shown in Figure 3. It can be clearly seen that the sensitivity and specificity of the CLS algorithm proposed in this study are higher than those of the comparison algorithm. This is because we have fully considered the principle of disease application, that is, the relationship between high-risk HPV infection and cervical lesions. A comprehensive analysis of the infection risk level of the patients themselves was carried out, so as to avoid missed diagnoses of those patients who have not tested positive for high-risk HPV but do have lesions. Degree of difference analysis was introduced to analyze the differences between related risk factors at different disease stages, so as to classify and predict the disease stages of patients better. In addition, the specificity of the CLS algorithm proposed in this study reached 92.7%, which indicates that our algorithm shows good ability to distinguish non-patients from patients. The CLS algorithm is therefore a tool for medically assisted decision-making that can effectively reduce the occurrence of overexamination.

TABLE 6 CIN3 Relevant High-risk HPV and Value.

Type of High-risk HPV	Relation Value with CIN3
HPV16	3.004129552
HPV 18	3.00267266
HPV 58	2.000134558
HPV 52	1.000043481
HPV 31	2.000129077
HPV 51	1.000015006
HPV 33	1.000014283
HPV 35	1.000014325
HPV 56	1.000015488
HPV 26	1.000007551
HPV 39	1.000014413
HPV 66	1.00001536
HPV 45	1.000007829

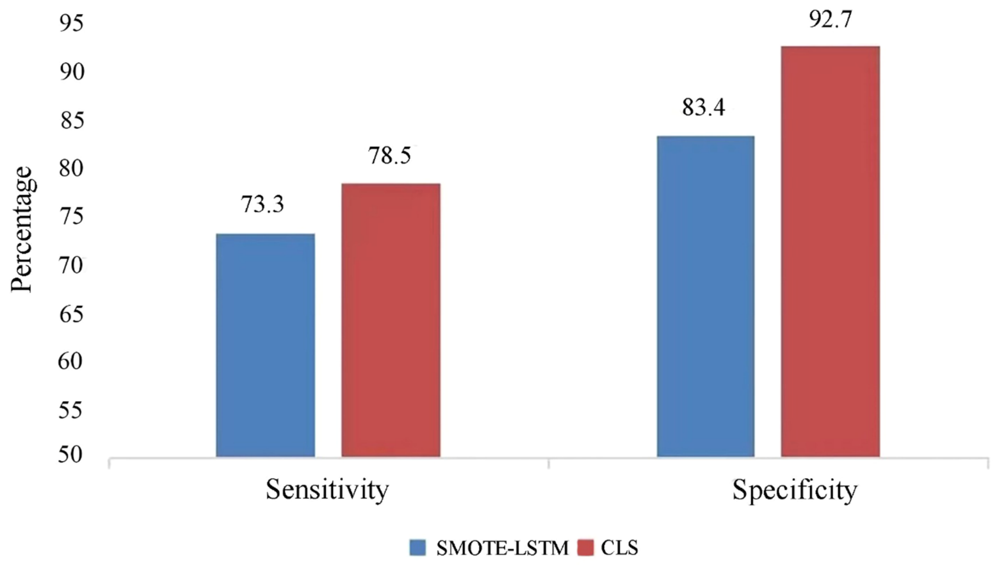

FIGURE 3
Comparison of sensitivity and specificity of prediction methods for cervical lesions.

4 Discussion

Accurate decision-making regarding the appropriate follow-up period for a target population with high-risk HPV infection can be time-consuming and challenging for clinicians, given the multitude of factors to consider. Prolonging the follow-up period increases the risk of missing the occurrence of cervical lesions, potentially leading to missed diagnostic opportunities before lesions develop. If the follow-up period is set too short, it may result in excessive examination, wasting medical resources and posing harm to patients' health. In 2020, the American Cancer Society updated its guidance to extend HPV screening intervals to 5 years based on accumulated evidence (23). However, disparities exist in recommendations from different academic organizations. According to the ATHENA trial, colposcopy is recommended if the patient tests positive for either HPV 16 or 18. Unfortunately, HPV testing can detect viral subtypes rather than persistent infection, which is an important factor in carcinogenesis. Girls and women tested positive for HPV subtypes -35, -39, -51, -56, -59, -66, or -68 are advised to undergo rescreening in 12 months (24). HPV infection genotype is important in detecting cancer and should be considered in triage management (25). To avoid excessive examinations and reduce the burden on patients, a more personalized diagnosis is recommended based on individual conditions. Physicians often manage patients according to their practical experience, acquired knowledge, guidance from predecessors, or research reports in journals, mainly relying on their subjective judgment. Evidence-based medicine not only focuses on doctors' clinical experience but also emphasizes the use of scientific evidence to guide clinical practice. Computers, as tools for data mining and knowledge discovery, can extract scientific evidence, providing an auxiliary support to doctors in clinical diagnosis. Therefore, evidence-based analysis of medical evidence can not only validate the accuracy of research but also play a relevant and conclusive role in clinical decision-making.

According to our study, patient risk was divided into three levels based on calculation of their total risk through the CLS algorithm. Then, every patient was assigned a serial number and a follow-up period. Follow-up periods were based on different risk levels. Individuals comprehensively understood to provide suggestions for the follow-up period compared with considering only single or several aspects. A reasonable and sufficient recommendation was expected.

Although studies published in 2006 and later show that fewer cases progressed to CIN3+, on average, in high-risk HPV-positive women compared with studies before 2006 (26), it remains critical to identify high-risk individuals to minimize the risk of their developing high-grade precancerous lesions. CIN3+ has been shown to be predominantly attributed to persistent HPV-16 and -18 infection, in line with the present study. However, it is difficult to identify the variations in the trends of the distribution of HPV genotypes in the target population due to the effects of vaccination and other factors such as patient age.

In extant studies of HPV, patient age has not received sufficient attention. Actually, age is non-negligible as one of other factors in present and potential CIN3 cases regardless of including or excluding HPV genotype. In our study, age was found to play a key role in CIN3 risk, ranking in significance only after HPV-16 and -18 infection status. Because of the limitations of the study, we did not analyze how or why age affects the infecting HPV genotypes and CIN risk. Extant studies show that the characteristics of distribution of high-risk HPV types differ with increasing age in patients with CIN2+. For instance, HPV-16 and -18 types cause CIN more often in younger women than in older women who are affected by genotypes other than those associated with non-high-risk HPV (27, 28). The reason for the atypical age-related distribution of HPV genotypes in older women is immunological status (29). Changes in the immunological status of older women weaken their immune systems, resulting in less effective immune clearance of uncommon HPV genotypes. Persistent infection may occur for the same reason and lead to high-grade cervical lesions. In the meanwhile, the incidence of CIN2 and CIN3 in the 20–29-year age group has doubled relative to the >60-year age group (30). Approximately 50% of cervical cancer cases in older women result in non-high-risk-HPV (31). Age and immunological status ought to be fully considered when investigating the distribution of HPV genotype.

The emergence and development of HPV vaccines, from bivalent to nonavalent, as the primary prevention method, has effectively protected more and more women of the appropriate age from HPV infection with certain genotypes, with well-established safety (32). The bivalent vaccine covers the HPV genotypes 16 and 18; the quadrivalent vaccine covers the low-risk genotypes 6 and 11 that contribute to most cases of genital warts (33, 34) and the two high-risk genotypes mentioned above. In addition to all these abovementioned genotypes, high-risk HPV genotypes 31, 33, 45, 52, and 58 are the other genotypes covered by the nonavalent vaccine (35).

Extant studies describe high efficacy (>90%) of the HPV vaccine against high-grade CIN-related genotypes and persistent high-risk HPV infection, and an efficacy of 64.6% against cross-protective types (HPV-31, -33, and -45). Additionally, the HPV vaccine shows robust and long-acting clinical efficacy in terms of protection and prevention (36, 37). Due to the effects of the uptake of the HPV vaccine, changes of prevalent HPV genotypes in women of different age groups have appeared globally. However, the unequal uptake of HPV vaccination program step by step in the world has led to variations in HPV genotype among countries and regions at a given time. Studies indicate the role of the HPV vaccine in preventing the occurrence of CIN2 and CIN3 in some countries (36, 37). Although the proportion of CIN3 due to genotypes covered by the nonavalent vaccine is high in the age group of 45 years and above, it seems that older women have significantly higher risk of high-grade CIN associated with the genotypes of HPV that are not covered by the nonavalent vaccine as well as non-high-risk HPV precancerous lesions.

In the present study, we find that HPV-16, -18, and -45 are the common types leading to all stages of CIN; this is consistent with a study that considers HPV-16 and -18 as the main types of cervical squamous carcinoma and HPV-18 and -45 as the primary types of cervical adenocarcinoma (38).The HPV types at the secondary level leading to CIN did not appear to be common features, likely because of the differences in race, region, and vaccination status.

In the future, the rates of HPV-16 and -18 infection are expected to gradually decrease, especially in young women, as a result of the effectiveness of bivalent and quadrivalent HPV vaccination programs. The influence of nonavalent vaccine on other prevalent high-risk genotypes is deemed to come out in a long term for relatively late implementation and stipulated younger age group between 9 and 26. Therefore, traditional high-risk HPV types may not be predictive of CIN or lesions. Conversely, the specific HPV genotypes excluded in the nonavalent vaccination, such as -56, -66, -51, -39, -35, and -26, are expected to be predictive of CIN3 and CIN3+.

Therefore, HPV genotyping test is a valuable screening method to predict risk value and guide individual management. Clinical decision-makers should regard age as a factor, together with HPV genotype, when managing CIN3 patients (39). Overall, these findings highlight the importance of regular cervical cancer screening throughout a woman's lifetime and tailoring management strategies based on individual risk profiles. This would allow unnecessary interventions to be minimized while ensuring early detection and treatment of precancerous lesions before they progress to invasive disease.

The HPV vaccine—an effective preventive strategy against HPV infection, related genital warts, and cervical cancer (40–42)—combined with HPV testing is expected to reduce cervical cancer rates. HPV testing has gradually become the main screening method due to its good sensitivity, and HPV self-sampling programs will be an available supplement to improve screening coverage. Although HPV self-sampling projects have been carried out only in a small number of countries, because of its advantages as a safe, simple, and private method, HPV self-sampling may have more widespread application in the future in additional countries (43). In addition, the vaccination status of girls and women should be taken into account during triage and to determine the frequency of HPV screening; these considerations should be explored in future studies (42).

5 Conclusion

Through in-depth study of the interactions between the risk factors for high-risk HPV, a risk factor relationship diagram structure was constructed. The risk level of patients was analyzed based on all risk factors, and a follow-up period for each risk level was formulated. According to the correlation between high-risk HPV genotypes and CIN, a lesion prediction model was constructed to predict the stage of cervical lesions within a reasonable follow-up period, provide a basis for pathological diagnosis, effectively reduce the risk of lesions, and even cancerization, and achieve primary prevention of cervical cancer.

In this study, we mined potential key risk factors, identified high-risk HPV patients, formulated follow-up periods for each risk level, and predicted cervical lesions, providing a new technological basis and ideas for related fields. Through evidence-based analysis, we demonstrated that the construction of a cervical cancer knowledge base and the structure of the risk factor relationship graph in this study play a key role in the evidence-based analysis of diseases and provide convenience and a scientific basis for evidence-based medicine. Furthermore, the findings offer time savings to doctors by enabling to assess and conduct decision making more efficiently. At the same time, the potential risk factors mined based on the structure of the risk factor map are also of significance for guiding clinical diagnosis and disease prevention. Altogether, the findings of this study can help the medical community to identify high-risk HPV patients more accurately, arrange follow-up more effectively, and improve the accuracy of cervical lesion prediction, thus providing more effective strategies for the prevention and treatment of cervical cancer.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Ethics statement

The studies involving humans were approved by Ethics Committee of Jilin Central General Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin because the local legislation and institutional requirements.

Author contributions

LG: Project administration, Writing – original draft, Conceptualization. YT: Data curation, Methodology, Writing – original draft. HX: Conceptualization, Investigation, Writing – original draft. LZ: Investigation, Resources, Writing – review & editing. YS: Formal analysis, Writing – review & editing, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Education Department Foundation of Jilin Province (Grant No. JJKH20210066KJ, Grant No. 2018ZCY129).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global cancer in women: burden and trends. *Cancer Epidemiol Biomarkers Prev* (2017) 26(4):444–57. doi: 10.1158/1055-9965.EPI-16-0858

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2021) 71:209–249. doi: 10.3322/caac.21660

3. *Global strategy to accelerate the elimination of cervical cancer as a public health problem*. Geneva: World Health Organization (2020).

4. Kremer WW, Steenbergen R, Heideman D, Kenter GG, Meijer C. The use of host cell DNA methylation analysis in the detection and management of women with advanced cervical intraepithelial neoplasia: a review. *BJOG* (2021) 128(3):504–14. doi: 10.1111/1471-0528.16395

5. Sammarco ML, Ucciferri C, Tamburro M, Falasca K, Ripabelli G, Vecchiet J. High prevalence of human papillomavirus type 58 in HIV infected men who have sex with men: A preliminary report in Central Italy. *J Med Virol* (2016) 88(5):911–4. doi: 10.1002/jmv.24406

6. Fonseca BO, Possati-Resende JC, Salcedo MP, Schmeler KM, Accorsi GS, Fregnani JHTG, et al. Topical imiquimod for the treatment of high-grade squamous intraepithelial lesions of the cervix: A randomized controlled trial. *Obstet Gynecol* (2021) 137(6):1043–53. doi: 10.1097/AOG.0000000000004384

7. Ogilvie GS, van Niekerk D, Krajden M, Smith LW, Cook D, Gondara L, et al. Effect of Screening with Primary Cervical HPV Testing vs Cytology Testing on High-grade Cervical Intraepithelial Neoplasia at 48 Months: The HPV FOCAL Randomized Clinical Trial. *JAMA* (2018) 320(1):43–52. doi: 10.1001/jama.2018.7464

8. Castle PE, Kinney WK, Xue X, Cheung LC, Gage JC, Zhao FH, et al. Effect of several negative rounds of human papillomavirus and cytology co-testing on safety against cervical cancer: an observational cohort study. *Ann Intern Med* (2018) 168 (1):20–9. doi: 10.7326/M17-1609

9. Frega A, Pavone M, Sesti F, Leone C, Bianchi P, Cozza G, et al. Sensitivity and specificity values of high-risk HPV DNA, p16/ki-67 and HPV mRNA in young women with atypical squamous cells of undetermined significance (ASCUS) or low-grade squamous intraepithelial lesion (LSIL). *Eur Rev Med Pharmacol Sci* (2019) 23 (24):10672–7. doi: 10.26355/eurrev_201912_19765

10. Kann BH, Hosny A, Aerts HJWL. Artificial intelligence for clinical oncology. *Cancer Cell* (2021) 39(7):916–27. doi: 10.1016/j.ccell.2021.04.002

11. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* (2019) 103(2):167–75. doi: 10.1136/bjophthalmol-2018-313173

12. Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* (2022) 40 (10):1095–110. doi: 10.1016/j.ccell.2022.09.012

13. Yanagawa M, Niioka H, Kusumoto M, Awai K, Tsubamoto M, Satoh Y, et al. Diagnostic performance for pulmonary adenocarcinoma on CT: comparison of radiologists with and without three-dimensional convolutional neural network. *Eur Radiol* (2021) 31(4):1978–86. doi: 10.1007/s00330-020-07339-x

14. Jiang M, Zhang D, Tang SC, Luo XM, Chuan ZR, Lv WZ, et al. Deep learning with convolutional neural network in the assessment of breast cancer molecular subtypes based on US images: a multicenter retrospective study. *Eur Radiol* (2021) 31(6):3673–82. doi: 10.1007/s00330-020-07544-8

15. Xue P, Xu H-M, Tang H-P, Wu W-Q, Seery S, Han X, et al. Assessing artificial intelligence enabled liquid-based cytology for triaging HPV-positive women: a population-based cross-sectional study. *Acta Obstet Gynecol Scand* (2023) 102:1026–33. doi: 10.1111/aogs.14611

16. Xue P, Tang C, Li Q, Li Y, Shen Y, Zhao Y, et al. Development and validation of an artificial intelligence system for grading colposcopic impressions and guiding biopsies. *BMC Med* (2020) 18(1):406. doi: 10.1186/s12916-020-01860-y

17. Sato M, Horie K, Hara A, Miyamoto Y, Kurihara K, Tomio K, et al. Application of deep learning to the classification of images from colposcopy. *Oncol Lett* (2018) 15 (3):3518–23. doi: 10.3892/ol.2018.7762

18. Dhombres F, Bonnard J, Bailly K, Maurice P, Papageorghiou AT, Jouannic JM. Contributions of artificial intelligence reported in obstetrics and gynecology journals: systematic review. *J Med Internet Res* (2022) 24(4):e35465. doi: 10.2196/35465

19. Peng G, Dong H, Liang T, Li L, Liu J. Diagnosis of cervical precancerous lesions based on multimodal feature changes. *Comput Biol Med* (2021) 130:104209. doi: 10.1016/j.compbiomed.2021.104209

20. Crowell EF, Bazin C, Saunier F, Brixtel R, Caillot Y, Lesner B, et al. CytoProcessorTM: A new cervical cancer screening system for remote diagnosis. *Acta Cytol* (2019) 63(3):215–23. doi: 10.1159/000497111

21. Li K, Li Q, Song L, Wang D, Yin R. The distribution and prevalence of human papillomavirus in women in mainland China. *Cancer* (2019) 125(7):1030–7. doi: 10.1002/cncr.32003

22. Alex SA, Jhanjhi N, Humayun M, Ibrahim AO, Abulfaraj AW. Deep LSTM model for diabetes prediction with class balancing by SMOTE. *Electronics* (2022) 11 (17):2737. doi: 10.3390/electronics11172737

23. Fontham ETH, Wolf AMD, Church TR, Etzioni R, Flowers CR, Herzig A, et al. Cervical cancer screening for individuals at average risk: 2020 guideline update from the American Cancer Society. *CA Cancer J Clin* (2020) 70:321–46. doi: 10.3322/caac.21628

24. Athanasiou A, Bowden S, Paraskevaidi M, Fotopoulou C, Martin-Hirsch P, Paraskevaidis E, et al. HPV vaccination and cancer prevention. *Best Pract Res Clin Obstet Gynaecol* (2020) 65:109–24. doi: 10.1016/j.bpobgyn.2020.02.009

25. Tota JE, Bentley J, Blake J, Coutlée F, Duggan MA, Ferenczy A, et al. Approaches for triaging women who test positive for human papillomavirus in cervical cancer screening. *Prev Med* (2017) 98:15–20. doi: 10.1016/j.ypmed.2016.11.030

26. Malagón T, Volesky KD, Bouten S, Laprise C, El-Zein M, Franco EL. Cumulative risk of cervical intraepithelial neoplasia for women with normal cytology but positive for human papillomavirus: Systematic review and meta-analysis. *Int J Canc* (2020) 147 (10):2695–707. doi: 10.1002/ijc.33035

27. Sakai M, Ohara T, Suzuki H, Kadomoto T, Inayama Y, Shitanaka S, et al. Clinical impact of age-specific distribution of combination patterns of cytology and high-risk HPV status on cervical intraepithelial neoplasia grade 2 or more. *Oncol Lett* (2023) 26 (3):384. doi: 10.3892/ol.2023.13970

28. Giannella L, Delli Carpini G, Di Giuseppe J, Prandi S, Tsiroglou D, Ciavattini A. Age-related changes in the fraction of cervical intraepithelial neoplasia grade 3 related to HPV genotypes included in the nonavalent vaccine. *J Oncol* (2019) 11(6):7137891. doi: 10.1155/2019/7137891

29. González P, Hildesheim A, Rodríguez AC, Schiffman M, Porras C, Wacholder S, et al. Behavioral/lifestyle and immunologic factors associated with HPV infection among women older than 45 years. *Cancer Epidemiol Biomarkers Prev* (2010) 19 (12):3044–54. doi: 10.1158/1055-9965

30. Onuki M, Matsumoto K, Satoh T, Oki A, Okada S, Minaguchi T, et al. Human papillomavirus infections among Japanese women: age-related prevalence and type-specific risk for cervical cancer. *Cancer Sci* (2009) 100(7):1312–6. doi: 10.1111/j.1349-7006.2009.01161.x

31. Guardado-Estrada M, Juárez-Torres E, Román-Bassaure E, Medina-Martinez I, Alfaro A, Benuto RE, et al. The distribution of high-risk human papillomaviruses is different in young and old patients with cervical cancer. *PloS One* (2014) 9(10):e109406. doi: 10.1371/journal.pone.0109406

32. Centers for Disease Control and Prevention. *HPV vaccine information for clinicians-fact sheet*. Available at: https://www.cdc.gov/std/HPV.

33. Paz-Zulueta M, Álvarez-Paredes L, Rodríguez Díaz JC, Parás-Bravo P, Andrada Becerra ME, Rodríguez Ingelmo JM, et al. Prevalence of high-risk HPV genotypes, categorised by their quadrivalent and nine-valent HPV vaccination coverage, and the genotype association with high-grade lesions. *BMC Canc* (2018) 18(1):112. doi: 10.1186/s12885-018-4033-2

34. Food and Drug Administration. *FDA approves gardasil 9 for prevention of certain cancers caused by five additional types of HPV*. Available at: https://www.fda.gov/BiologicsBloodVaccines/Vaccines/ApprovedProducts/ucm426445.htm.

35. Arbyn M, Xu L. Efficacy and safety of prophylactic HPV vaccines. A Cochrane review of randomized trials. *Expert Rev Vaccines* (2018) 17(12):1085–91. doi: 10.1080/14760584.2018.1548282

36. Harper DM, DeMars LR. HPV vaccines - A review of the first decade. *Gynecol Oncol* (2017) 146(1):196–204. doi: 10.1016/j.ygyno.2017.04.004

37. Hoes J, King AJ, Berkhof J, de Melker HE. High vaccine effectiveness persists for ten years after HPV16/18 vaccination among young Dutch women. *Vaccine.* (2023) 41 (2):285–9. doi: 10.1016/j.vaccine.2022.11.057

38. Bhatla N, Singhal S. Primary HPV screening for cervical cancer. *Best Pract Res Clin Obstet Gynaecol* (2020) 65:98–108. doi: 10.1016/j.bpobgyn.2020.02.008

39. Giannella L, Giorgi Rossi P, Delli Carpini G, Di Giuseppe J, Bogani G, Gardella B, et al. Age-related distribution of uncommon HPV genotypes in cervical intraepithelial neoplasia grade 3. *Gynecol Oncol* (2021) 161(3):741–7. doi: 10.1016/j.ygyno.2021.03.025

40. Drolet M, Bénard É, Pérez N. Brisson M.HPV Vaccination Impact Study Group. Population-level impact and herd effects following the introduction of human papillomavirus vaccination programmes: updated systematic review and meta-analysis. *Lancet* (2019) 394(10197):497–509. doi: 10.1016/S0140-6736(19)30298-3

41. Palmer T, Wallace L, Pollock KG, Cuschieri K, Robertson C, Kavanagh K, et al. Prevalence of cervical disease at age 20 after immunisation with bivalent HPV vaccine at age 12-13 in Scotland: retrospective population study. *BMJ* (2019) 365:l1161. doi: 10.1136/bmj.l1161

42. Hall MT, Simms KT, Lew JB, Smith MA, Brotherton JM, Saville M, et al. The projected timeframe until cervical cancer elimination in Australia: a modelling study. *Lancet Public Health* (2019) 4(1):e19–27. doi: 10.1016/S2468-2667(18)30183-X

43. Serrano B, Ibáñez R, Robles C, Peremiquel-Trillas P, de Sanjosé S, Bruni L. Worldwide use of HPV self-sampling for cervical cancer screening. *Prev Med* (2022) 154:106900. doi: 10.1016/j.ypmed.2021.106900