Development of a cervical cancer progress prediction tool for human papillomavirus-positive Koreans: A support vector machine-based approach

Jimin Kahng¹, Eung-Hee Kim², Hong-Gee Kim² and Wonbae Lee³

Abstract
Objectives: To develop a Web-based tool to draw attention to patients positive for human papillomavirus (HPV) who have a high risk of progression to cervical cancer, in order to increase compliance with follow-up examinations and facilitate good doctor–patient communication.

Methods: Records were retrospectively analysed from women who were positive for HPV on initial testing (before any treatment). Information concerning age, Papanicolaou (PAP) smear result and presence of 15 high-risk HPV genotypes was used in a support vector machine (SVM) model, to identify the patient features that maximally contributed to progression to high-risk cervical lesions.

Results: Data from 731 subjects were analysed. The maximum number of correct cancer predictions was seen when four features (PAP, HPV16, HPV52 and HPV35) were used, giving an accuracy of 74.41%. A web-based high-risk cervical lesion prediction application tool was developed using the SVM model results.

Conclusions: Use of the web-based prediction tool may help to increase patient compliance with physician advice, and may heighten awareness of the significance of regular follow-up HPV examinations for the prevention of cervical cancer, in Korean women predicted to have heightened risk of the disease.

¹Department of Laboratory Medicine, College of Medicine, Catholic University of Korea, Seoul, Republic of Korea
²Biomedical Knowledge Engineering Laboratory, Seoul National University, Seoul, Republic of Korea
³Department of Pediatrics, College of Medicine, Catholic University of Korea, Seoul, Republic of Korea

Corresponding author:
Wonbae Lee, Department of Pediatrics, Bucheon St Mary’s Hospital, College of Medicine, Catholic University of Korea, 2 Sosa-dong, Wonmi-gu, Bucheon, Gyeonggi-do 420-717, Republic of Korea.
Email: leewb@catholic.ac.kr

Creative Commons CC-BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 3.0 License (http://www.creativecommons.org/licenses/by-nc/3.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (http://www.uk.sagepub.com/aboutus/openaccess.htm).
Introduction

Cervical cancer is the second most common form of cancer in women worldwide. The development of cervical cancer requires the persistent presence of human papillomavirus (HPV). To date, ~120 HPV types have been identified and classified according to oncogenicity into high-risk types (such as HPV16 and HPV18), and low-risk types (such as HPV6, HPV11 and HPV40). The most common high-risk types worldwide are HPV16 (~57% of cases) and HPV18 (~16% of cases); however, the proportions of invasive cervical cancers related to particular HPV types vary, depending on geographical location. Not all HPV infections progress to cervical cancer; it has been reported that ~90% of HPV infections resolve within 2 years of identification. However, it remains unclear why HPV infections resolve in some cases and not in others; individual differences in susceptibility may be contributing factors. At present, surgical treatments (such as conization or loop electrosurgical excision) are the main treatment options for patients with pre-cancerous conditions or early stage cancers.

In the Republic of Korea, efforts by national and private health organizations have helped to increase the rate of cervical cancer screening to 52.5% in 2010; a positive change that may, in due course, deliver a similar decline in cervical cancer death rates in the Republic as has been seen in the USA. HPV DNA tests may be used as primary screening tests, in conjunction with Papanicolaou (PAP) smears; a number of studies have shown that HPV DNA tests have greater sensitivity for the detection of cervical intraepithelial neoplasia than PAP smear tests alone.

Not all women undergo screening tests for cervical cancer, and not all women attend regular check-ups after the detection of abnormal epithelial lesions. In view of the long precancerous stage of cervical cancer (which can be >10 years), more frequent follow-up examinations can enable precancerous lesions to be detected at an earlier stage. In the Republic of Korea, irregular screening (due to cultural reluctance or resistance and poor comprehension of the need for regular screening or follow-up) leads to the development of cervical cancers that could have been avoided.

In doctor–patient communication, a patient’s need for information is particularly great when cancer is involved. Almost half of cancer patients reported that no information had been given to them on how to deal with their disease. This knowledge gap can widen in time-constrained environments such as secondary and tertiary hospitals, where each patient encounter is hardly long enough to ensure patient satisfaction. This, in turn, can influence patient compliance with physician advice.

The goal of the present study was to develop an intuitive, web-based, visual application that can be used during patient–doctor encounters to quickly draw attention to those patients with a high risk of progression to cervical cancer. Use of the tool should help increase patient compliance with physician advice, and should also heighten awareness of the significance of regular follow-up HPV examinations. The application aims to visualize potential progression or regression of HPV infection,
according to patient genotype and PAP test results, based on a support vector machine model.

**Subjects and methods**

**Subjects**

Records were retrospectively analysed from women who had undergone HPV DNA tests at Bucheon St Mary’s Hospital, Catholic University of Korea, Bucheon, Republic of Korea, during the period between 1 January 2006 and 31 December 2012; the women were HPV positive before any treatment was administered and their HPV genotype did not change during follow-up. If a cervical lesion suspicious of HPV infection was seen on colposcopic examination, HPV testing was performed. If testing proved to be HPV-negative and the lesion persisted in the immediate follow-up period, another HPV test was recommended. Subjects were followed up with colposcopic examination, PAP smear and HPV test every 3 months; all patients had an initial examination and more than one follow-up check. Subjects were included in the present study if they had undergone HPV DNA chip tests and had PAP smear or biopsy results available. Biopsies were performed if pathology worse than a low-grade squamous intraepithelial lesion was detected on PAP smear. A second HPV test was carried out in order to confirm the clearance of HPV infection after treatment for cervical intraepithelial neoplasia (CIN).

For each subject, their age when first tested for HPV, HPV, PAP smear and biopsy results, and the duration of follow-up were recorded. For the purposes of the present study, a high-risk lesion was defined as high-grade squamous intraepithelial lesion (HSIL), atypical squamous cells that cannot exclude HSIL, carcinoma in situ, squamous cell carcinoma or adenosquamous carcinoma, atypical glandular cells-favour neoplasia, adenocarcinoma or neuroendocrine carcinoma on PAP smear, or cervical intraepithelial neoplasia (CIN) 2, CIN3, carcinoma in situ or invasive carcinoma on biopsy.

Due to the retrospective nature of the study, the need for ethical approval and for patient consent was waived by the Catholic Medical Centre Institutional Review Board.

**HPV testing**

The HPV tests were performed using the HPVDNAChip kit (BioMedLab, Seoul, Republic of Korea) as previously described.\(^{16}\) Using a sterilized vaginal speculum, cervical specimens were collected during colposcopic examination by inserting a cytobrush attached to an HPVDNAChip Sampler (BioMedLab) into the endocervical and exocervical areas. After removal, the cytobrush was placed into transport medium (2 ml phosphate buffered saline) and immediately refrigerated at 4°C for ≤72 h before analysis. DNA extracted in compliance with the manufacturer’s instructions was immediately used or stored at 4°C; samples kept for >24 h were stored at −70°C until analysis. The HPVDNAChip kit is able to identify 22 HPV genotypes: 15 high-risk types (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68 and 69) and seven low-risk types (HPV6, 11, 34, 40, 42, 43 and 44).

**Support vector machine model**

The support vector machine (SVM), developed by Vapnik,\(^{17}\) has been used to model various biomedical situations, including protein–protein interaction predictions,\(^{18,19}\) gene expression data analysis\(^{20}\) and decision support in the diagnosis of heart valve disease.\(^{21}\) The SVM classifies data into two categories of membership by constructing a high dimensional hyperplane, while being optimized to maximize the margin that best divides membership. For the present study, an SVM model was developed using
LibSVM (http://weka.wikispaces.com/LibSVM) with the radial basis function as a kernel function. A total of 17 features were used to build the model: age, PAP result and the 15 high-risk HPV genotypes. Each subject in the study population was classified by the model as either regression ($\omega_1$) or progression to cancer ($\omega_2$). To identify features that deliver the highest classification performance of the model, a forward feature selection process was used.\textsuperscript{22} The model focused on reducing the rates of false-positive and false-negative results.

**Statistical analyses**

Fisher’s exact test and $\chi^2$-test were used to analyse relationships between variables. For two-group comparisons, two sample $t$-tests (unpaired $t$-tests) were used for parametric analyses and Wilcoxon’s rank sum test for nonparametric analyses. For comparisons of more than two groups, analysis of variance was used for parametric analyses and the Kruskal–Wallis test for nonparametric analyses. A $P$-value $< 0.05$ was considered to be statistically significant. All statistical analyses were performed using SAS\textsuperscript{®} software, version 9.2 (SAS Institute, Cary, NC, USA).

**Results**

A total of 731 subjects were included in the study: 75 (10.26%) were aged 10–29 years; 169 (23.12%) were aged 30–39 years; 290 (39.67%) were aged 40–49 years; 123 (16.83%) were aged 50–59 years; 74 (10.12%) were aged $\geq$60 years. A total of 310 subjects (42.41%) were diagnosed with high-risk lesions at the initial examination had a follow-up period of $<6$ months. Of the remaining 421 subjects, 251 (59.62%) had a follow-up period of $<6$ months, 102 (24.23%) had a follow-up period of $\geq6$ months but $<2$ years and 68 (16.15%) had a follow-up period of $\geq2$ years.

**Relationship between age and high-risk lesions**

Of the 197 subjects aged $\geq50$ years, 114 (57.87%) progressed to high-risk lesions, compared with 231 out of 534 subjects aged $<50$ years (43.26%). This difference was statistically significant ($P = 0.0004$, $\chi^2$-test). Logistic regression analysis gave an odds ratio of 1.023; however, this may not indicate a linear increase with age.

**PAP smear and biopsy results**

Table 1 shows the distribution of subjects based on PAP smear and biopsy results. Of the 587 subjects who underwent biopsy, 311 (52.98%) had high-risk lesions, including 69 with CIN2, 60 with CIN3, 120 with carcinoma in situ, three with adenocarcinoma in situ, 52 with squamous cell carcinoma, one with adenosquamous cell carcinoma and six with adenocarcinoma. Comparison of the PAP smear and biopsy results showed that PAP tests had a specificity of 97.0% and a sensitivity of 38.5% in the present study.

**Relationship between genotype and high-risk lesions**

The seven most frequently occurring genotypes, in decreasing order of occurrence, were: HPV16 ($n=207$; 28.32%); HPV52 ($n=103$; 14.09%); HPV58 ($n=90$; 12.31%); HPV51 ($n=65$; 8.89%); HPV56 ($n=57$; 7.80%); HPV18 ($n=51$; 6.98%); HPV35 ($n=40$; 5.47%). A total of 594 subjects (81.26%) were infected with a
single genotype; 137 subjects (18.74%) were infected with multiple genotypes. HPV infection regressed in 312 subjects (42.68%) and persisted in 419 subjects (57.32%).

The seven most frequently occurring genotypes classified according to the development of high-risk lesions are shown in Table 2. Table 3 shows the specificity, sensitivity, and the negative and positive predictive values of the seven most frequently occurring HPV genotypes for the development of high-risk cervical lesions.

**SVM model and high-risk cervical lesion prediction tool**

If the SVM model was 100% accurate, 386 subjects would be classified as $\omega_1$ and 345 as $\omega_2$. Table 4 shows construction of the SVM model in terms of feature combinations, accuracy and the confusion matrix at each feature selection step. Accuracy was confirmed using 10-fold cross-validation. The maximum number of correct cancer predictions (228/345) is seen when four features (PAP, HPV16, HPV52 and HPV35) are used, with an accuracy of 74.41%.

A web-based high-risk cervical lesion prediction application was developed using the SVM model; this is accessible at http://147.46.70.245:8181/HPV_Prediction/. The application interactively displays a chart that shows the probability of progression to high-risk lesions depending on feature values selected by the user. The high-risk lesion-based SVM model uses a set of four classification features: PAP smear result, HPV16, HPV35 and HPV52. Each genotype is recorded as negative (1) or positive (2). The application supports Ajax-based client–server interaction, and the Google® Web Toolkit (http://www.gwtproject.org) was used to develop the user interface. The charting library HighCharts (http://www.highcharts.com) was used to produce the interactive charts.

**Discussion**

In the present study the SVM model identified four features (PAP smear result, HPV16, HPV35 and HPV52) that best classified subjects using available data (age, HPV genotype and PAP smear result). As further datasets become available, the model and web application has the potential to dynamically reflect any changes by adjusting the classification features.

In clinical practice it is very difficult to enforce strict adherence to follow-up examinations, and subjects in the present study were no exception to this. In addition, PAP results are subjective in nature. The current application focused on developing a visual consultation tool from available data to help facilitate patient understanding during the doctor–patient encounter, and
emphasize the potential for cervical high-risk lesion development.

Age has a known relationship with cervical cancer development, but is not a sole indicator.\textsuperscript{23} Data on the social and health care settings of subjects, such as sexual behaviour and the absence of screening, would have made the role of age more meaningful in the present study.

It is known that the longer the HPV infection has been present, the higher the chance of progression to a high-risk lesion.\textsuperscript{2} In practice it is difficult to know when a patient first became infected with HPV, and therefore it is also difficult to know how long the infection has lasted. In the present study, data on follow-up periods from the time when subjects were first tested and found to be HPV positive were available, but were likely to have marginal value in predicting progression, and so were not included in the SVM model.

Individual susceptibility factors (such as host-genome variability, lifestyle, and infection with \textit{Chlamydia trachomatis} or HIV)\textsuperscript{2} have been shown to be strong determinants in relation to the development of cervical high-risk lesions.\textsuperscript{2} Due to the limitations of outpatient settings in the Republic of Korea, information about host genome variability and lifestyle was rarely documented in the subject records used for the present study.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|c|}
\hline
Genotype & Subjects, \(n\) & Group I \(n=35\) & Group II \(n=386\) & Group III \(n=310\) \\
\hline
HPV16 & 207 (28.32) & 17 (48.57)\textsuperscript{a} & 70 (18.13) & 137 (44.19)\textsuperscript{a} \\
HPV18 & 51 (6.98) & 1 (2.86) & 28 (7.25) & 23 (7.42) \\
HPV35 & 40 (5.49) & — & 20 (5.18) & 20 (6.45) \\
HPV51 & 65 (8.89) & 2 (5.71) & 49 (12.69)\textsuperscript{a} & 16 (5.16) \\
HPV52 & 103 (14.09) & 7 (20.00)\textsuperscript{a} & 64 (16.58) & 39 (12.58) \\
HPV56 & 57 (7.80) & 1 (2.86) & 44 (11.40)\textsuperscript{a} & 13 (4.19) \\
HPV58 & 90 (12.31) & 5 (14.29) & 42 (10.88) & 48 (15.48) \\
\hline
\end{tabular}
\caption{The seven most frequently occurring genotypes in subjects \((n=731)\) positive for human papillomavirus (HPV), classified according to the development of high-risk lesions.}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|c|}
\hline
Genotype & Specificity & Sensitivity & Negative predictive value & Positive predictive value \\
\hline
HPV16 & 333/386 (86.27) & 154/345 (44.64) & 333/524 (63.55) & 154/207 (74.40) \\
HPV18 & 359/386 (93.01) & 24/345 (6.96) & 359/680 (52.79) & 24/51 (47.06) \\
HPV35 & 366/386 (94.82) & 20/345 (5.80) & 366/691 (52.97) & 20/40 (50.00) \\
HPV51 & 339/386 (87.82) & 18/345 (5.22) & 339/666 (50.90) & 18/65 (27.69) \\
HPV52 & 329/386 (85.23) & 46/345 (13.33) & 329/628 (52.39) & 46/103 (44.66) \\
HPV56 & 343/386 (88.86) & 14/345 (4.06) & 343/674 (50.89) & 14/57 (24.56) \\
HPV58 & 349/386 (90.41) & 53/345 (15.36) & 349/641 (54.45) & 53/90 (58.89) \\
\hline
\end{tabular}
\caption{Specificity, sensitivity and negative and positive predictive values for the development of high-risk lesions for the seven most frequently occurring genotypes in subjects positive for human papillomavirus (HPV).}
\end{table}
Co-infection with *Chlamydia trachomatis* was recorded in six subjects in the present study; however, since very few people are required to undergo *Chlamydia* tests within the current national health programme, such data are of little statistical value and were not analysed further. No subjects in the present study were co-infected with HIV.

The effect of co-infection in the prediction of cancer progression requires further study and analysis.

The prediction application tool developed in the present study is a systematic attempt to display the risk of progression to cervical high-risk lesions, with the goal of increasing informed compliance from patients, with respect to regular follow-up examinations. This tool is not a comprehensive source of information for the patient but is rather a means to increase patient awareness about high-risk HPV genotypes. It must be used in conjunction with expert medical advice, derived from clinical experience and up-to-date knowledge of cervical cancer.

### Declaration of conflicting interest

The authors declare that there are no conflicts of interest.

### Table 4. Support vector machine model construction using age, Papanicolaou smear result (PAP) and 15 high-risk genotypes in subjects positive for human papillomavirus (HPV): at each selection step a further feature is added and the predictive accuracy of the steps up to that point is calculated.

<table>
<thead>
<tr>
<th>FSS/AF/accuracy(%)</th>
<th>Confusion matrix</th>
<th>FSS/AF/accuracy(%)</th>
<th>Confusion matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/PAP/71.13</td>
<td>$\omega_1$</td>
<td>$\omega_2$</td>
<td>10/HPV33/71.54</td>
</tr>
<tr>
<td></td>
<td>358</td>
<td>28</td>
<td>11/HPV51/71.81</td>
</tr>
<tr>
<td></td>
<td>183</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>2/HPV16/75.10</td>
<td>$\omega_1$</td>
<td>$\omega_2$</td>
<td>13/HPV45/72.36</td>
</tr>
<tr>
<td></td>
<td>335</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>3/HPV52/74.55</td>
<td>$\omega_1$</td>
<td>$\omega_2$</td>
<td>15/HPV69/72.09</td>
</tr>
<tr>
<td></td>
<td>131</td>
<td>214</td>
<td></td>
</tr>
<tr>
<td>4/HPV35/74.41</td>
<td>$\omega_1$</td>
<td>$\omega_2$</td>
<td>17/Age/65.80</td>
</tr>
<tr>
<td>5/HPV58/74.14</td>
<td>$\omega_1$</td>
<td>$\omega_2$</td>
<td></td>
</tr>
<tr>
<td>6/HPV31/72.64</td>
<td>$\omega_1$</td>
<td>$\omega_2$</td>
<td></td>
</tr>
<tr>
<td>7/HPV56/72.22</td>
<td>$\omega_1$</td>
<td>$\omega_2$</td>
<td></td>
</tr>
<tr>
<td>8/HPV39/72.22</td>
<td>$\omega_1$</td>
<td>$\omega_2$</td>
<td></td>
</tr>
<tr>
<td>9/HPV59/72.09</td>
<td>$\omega_1$</td>
<td>$\omega_2$</td>
<td></td>
</tr>
</tbody>
</table>

FSS, forward selection step; AF, added feature.
Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References