Association between Social Behavioral Risk Factors and Cervical Cancer

Shuyi Chen[1, a, *], Jiaoyang Li[2, b, *], Zetian Zhang[3, c, *]

[1]Department of Cognitive Science, University of California, Irvine, Irvine, California

[2]Department of Computer and Data Sciences; Department of Biology, Case Western Reserve University, Cleveland, OH

[3]Department of Biology, Grinnell College, Grinnell, IA

*Corresponding author: [a]shuyic9@uci.edu, [b]jxl1816@case.edu, [c]zhangzet@grinnell.edu

*These authors contributed equally.

Keywords: Biostatistics, R studio, Principal Component Analysis, Logistic Regression, Confusion Matrix, Cervical Cancer, Social Theory.

Abstract: Cervical Cancer is the leading medical cause of death in women globally. It cannot be cured but is preventable. When use of cancer prevention strategies remains low in developing countries like Indonesia where the poverty rate is high. Our study that focuses on exploring the association between social risk behavioral factors with cervical cancer may help women to identify their cancer risk when medical access is limited. There were 72 respondents with 22 of them having cervical cancer responded to the questionnaire with nine questions for different variables. All assumptions of Health Belief Model, Protection Motivation Theory, and Theory of Planned behavior were considered to measure the nine categorical factors. To avoid collinearity, the data was processed by Principal Component Analysis (PCA) to identify factors that are associated with cervical cancer, the first three components are selected. A confusion matrix is built upon the PCA model to examine accuracy. Variables from PC1 with the highest contributions and qualities of representation are considered most significant. The three categorical risk factors are social support, motivation, and empowerment. The logistic equation of cancer risk prediction based on the social determinants has shown an accuracy of 94% from confusion matrix. By identifying most associated social risk factors, it can be used for various purposes including education and further research studies. Also, the time and cost-efficient prediction model allows more women to detect their state of risk early before the disease begins to deteriorate. CCS CONCEPTS• Mathematics of computing ~ Mathematical software ~ Statistical software • Mathematics of Computing ~ Probability and statistics ~ Probability inference problems ~ Maximum Likelihood estimation.

1. Introduction

Cervical Cancer is a serious public health problem in women all over the world. It is the second most common cancer among women worldwide. The estimated global prevalence is 11.7% [1]. Globally, each year cervical cancer kills about 300,000 women, especially in developing countries [2]. Up to 99% of cervical cancer cases are linked with persistent high-risk human papillomaviruses (HPV), an extremely common virus transmitted through sexual contact [3]. Recently, cancer prevention is based on HPV vaccination and secondary prevention approaches screening, the pap smear test VIA test, which will prevent most cervical cancer cases [4]. Unfortunately, the scope of screening and vaccination still remains low in developing countries like Indonesia, particularly in regions with high poverty [5]. The elimination of cervical cancer in these countries is challenged with inadequate medical and social systems. Even when cancer prevention tests such as pap smear tests become available, many positive patients were “lost” in following-up treatments. Therefore, women from these regions are more likely to have cervical cancer and are less likely to get proper treatment instantly.

Some studies have investigated the relationship between social determinants and the possibility of getting cervical cancer. To increase the coverage of cervical screening and lower the rate of Ca Cervix, Williams-Brennan and his co-workers tried to identify the social determinants of health (SDH) that could potentially have an association with screening for those women who live in developing countries [6]. Because Ca Cervix, fortunately, has a great possibility to be prevented due to its slow progression, effective and potent treatments may be applied to the patients if there exists an early detection [7]. The sample in this study has a fairly large size [6]. The location of the data collection is similar to Indonesia where our data originates [6]. Their variables also share similarities with ours, since our variables were derived from social theories such as The Health Brief Model (HBM), Protection Motivation Theory (PMT), and Theory of Planned Behavior (TPB). Sari, Mudigdo, and Dermatoto focused on using multilevel analysis to analyze the social determinants of cervical cancer in Yogyakarta, a city in Indonesia [8]. The location of data collection in this study shares a high similarity with that of our data, and the researchers also explored the possible influence of social determinants on Ca Cervix [8]. However, Sari, Mudigdo, and Dermatoto employed a different approach by using Logistic Regression models and performing multilevel analysis, while our group employed the method of principal component analysis (PCA) [8]. Furthermore, Sobar, Machmud, and Wijaya examined whether machine learning algorithms would be able to accurately detect and predict the risk of getting cervical cancer according to the behavior variable and its determinants. Specifically, two popular machine learning techniques, Naïve Bayes (NB) and Logistic Regression (LR) were used as the classification methods. Drain et al. investigated the determinants of cervical cancer in developing countries as well by employing regression analysis and developing an ANCOVA model for cervical cancer incidence [9]. Besides, Santamaria-Ulloa and Valverde-Manzanares also assessed the association between social determinants and incidence of cervical cancer by conducting Poisson and spatial regression analysis with Stata and Arcmap software [10].

Not so many studies have explored the relationship between the possibility of getting cervical cancer and the behavioral variable with its associated social-theory-derived determinants. Moreover, most of the previous studies used Logistic Regression models and machine learning algorithms. Therefore, we aim to investigate the relationship between social determinants and the risk of getting Ca Cervix using the method of principal component analysis.

2. Method

2.1 Data Collection

This dataset is retrieved from UCI Machine Learning Repository, which is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

2.2 Research Variable

The dataset Sobar, consists of one binary dependent variable (Ca_Cervix) and 18 attributes that comes from 8 categorical independent variables. These 8 variables are the following: perception; intention; motivation; empowerment; social support; normativity; attitude; and behavior. They can be explained as such: perception is defined as the perceived threat of cancer and the evaluation of behaviors to counteract the threat; Intention of performing cancer prevention behavior; Motivation is one of determining factors of organizational prevention behavior; Empowerment is defined as a system of beliefs, referring to the ability to make decisions upon cancer prevention; Subjective norm is the presence of a significant person or group of people that shows support and approval toward the act of prevention behavior; Attitude and subjective norm interacted with perceived control, oftenly it suggests the respondents attitude toward the novel ways of treatments and prevention on cervical cancer; Social support includes emotionality and whether the respondent has access to the updated information regarding cancer; Behavior includes sexual behavior and personal hygiene. The result of the data was collected through questionnaires with nine questions with each variable. There are 72

respondents, 22 of them had cervical cancer and the rest are cancer-free. All respondents are urban citizens of Jarkata, Indonesia. All assumptions of Health Belief Model, Protection Motivation Theory, and Theory of Planned behavior were considered to measure the 8 variables. A score from 0-15 was given with each variable.

2.3 Statistical analysis

We performed logistic regression on all x variables and test for multicollinearity. Then principal component analysis (PCA) was conducted, followed by logistic regression on the first three principal components (PC).

2.4 Logistic Regression

To better explore the probability of having cervical cancer in relation to our independent variables, we use logistic regression to measure the likelihood (log of odds ratio $p/(1-p)$) of getting cervical cancer. Another rationale for choosing logistic regression is that our dependent variable is binary, Y=1 or Y=0. It measures whether people filling out this questionnaire have cervical cancer or not.

2.5 Test for multicollinearity

Because the data is gathered in a form of questionnaire, we are not sure if the data is processed with NA values removed or under what conditions the data was collected. Based on the known information, we first performed logistic regression on all x variables, but each variable has a p-value 0.999 and 1.000, suggesting there exists multicollinearity. We further tested for variance influence factor (vif). Its value suggests high collinearities among data.

2.6 PCA selection

We use PCA to reduce dimensionality and multicollinearity. Each PC is constructed based on the linear combinations of original predictor variables and completely uncorrelated from with other PCs, because the next PC is orthogonal to the previous one.

We performed PCA on the covariance matrix of our 19 predictor variables. The portion of variance is calculated and is equivalent to the eigenvalues of our covariance matrix. A scree plot is included to show the magnitude of eigenvalues for each PCs, which represent the amount of variability that is preserved from the original data set and can be explained by our newly constructed PCs (figure 1). First three PCs and six PCs can explain up to 71.2% and 83.3% of variability, respectively. The decision for using how many PCs is later determined by logistic regression. Since the logistic regression based on newly constructed six PCs does not show any significance for the third and sixth PCs, we decide three PCs are sufficient to explain the variability.

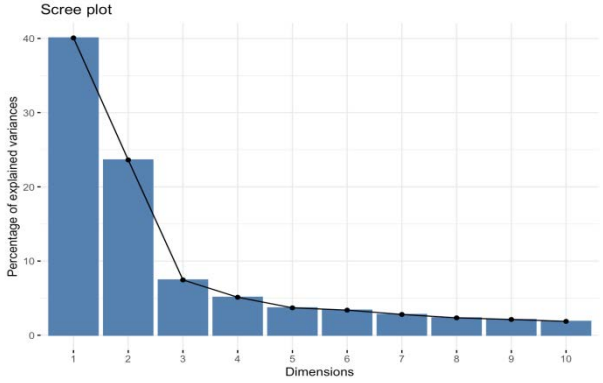

Figure 1. Screen plot. Only the first 10 dimensions are shown. The percentages of explained variances are in decreasing order as the number of dimensions increases

2.7 Confusing Matrix

In order to visualize the performance of our algorithm, we make a confusion matrix based on our PCA regression result. We are using the first three-dimension data. The reason we are using PCA regression results instead of the five variables we select is there exists severe multicollinearity in our original data set. Both the logistic regression result and the confusion matrix result based on original data indicate we shall change our algorithm and build a new confusion matrix.

3. Result

3.1 Correlations Among Variables

The quality of representation shows a decreasing trend as we increase our dimensions. We use loadings to better visualize the weight each variable has on each PC (Table 1). Loadings represent the coefficients of linear combinations of our original *x* variables projected onto the PC. Therefore, positive loading represents positive correlation between our x variable and the PC. We project our original *x* values on plots formed axes constructed by two PCs that are orthogonal to each other (Figure 2). The arrows for each of our original predictor variables point away from the point of origin intersected by both PCs. The longer the arrow, and the bigger the loading, the greater the variable that is represented on the PC plot. Based on the locations of two arrows and their loadings in the same PC, we can infer the correlation between two variables. As expected, all variables on PC1 are positively correlated, because our predictor variables are measurements of positive social theories relevant to cervical cancer. If two variables are negatively correlated, this suggests an increase in the value of one causes a decrease in the value of another. All variables are positively correlated with each other, except for attitude_spontaneity. Because attitude_spontaneity does not have loadings in first three PCs, we need better explanation why it appears on PC2, though it is poorly represented on PC2.

Table.1. Loadings of Variables in each PC

Loadings			
	PC1	PC2	PC3
behavior_sexualRisk			
behavior_eating			
behavior_personalHygine	0.143	0.125	-0.430
intention_aggregation			-0.419
intention_commitment			
attitude_consistency			
attitude_spontaneity			
norm_significantPerson		0.181	
norm_fulfillment		0.656	
perception_vulnerability		0.526	0.324
perception_severity		0.432	
motivation_strength	0.138		-0.424
motivation_willingness	0.343		-0.307
socialSupport_emotionality	0.386		-0.121
socialSupport_appreciation	0.243		
socialSupport_instrumental	0.362	-0.171	0.350
empowerment_knowledge	0.406		
empowerment_abilities	0.403		

empowerment_desires	0.406		0.312

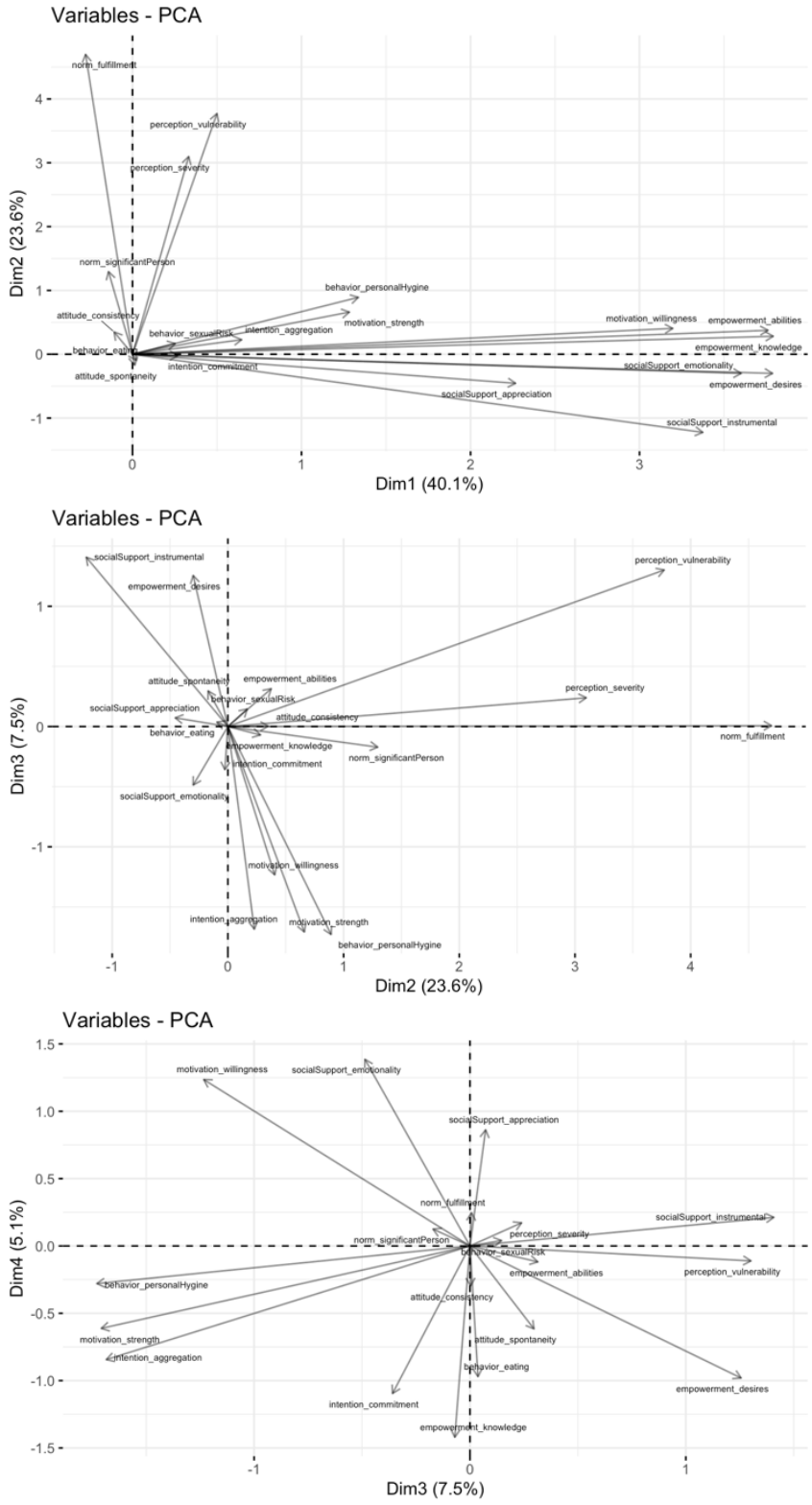

Figure 2. PCA plot, quality of representation of predictor variables based on loadings of original predictor variables rotated for newly constructed axes

3.2 Contributions of x variables

Contributions of each variable to the construction of each PC are measured and shown in figure 3. Variables above the red line indicate strong contributions; if the same variable is always below the red line in each of the three contribution plots, we exclude them for simplicity because the impacts of

these variables are not significant. Variables with low impacts, namely intention_commitment, behavior_sexualRisk, norm_significantPerson, attitude_consistency, attiude_spontaniety should be excluded from future questionnaires.

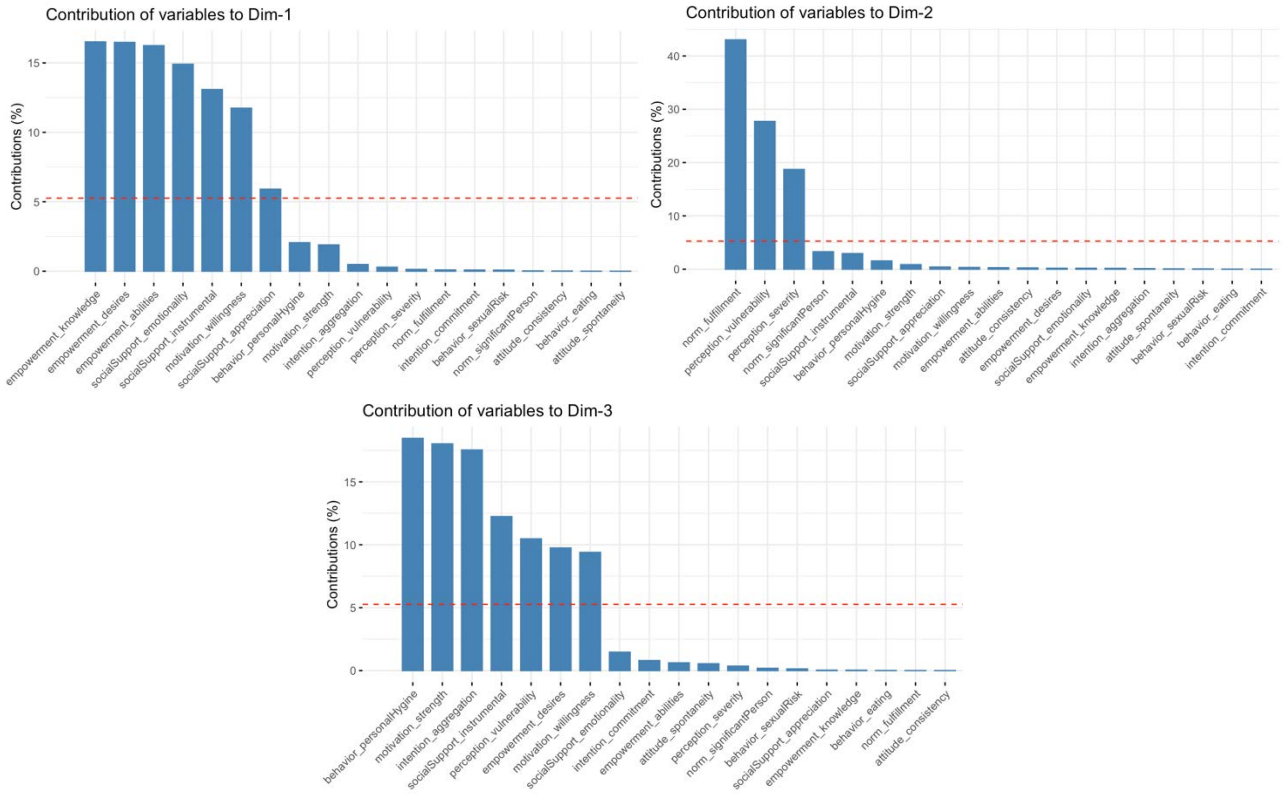

Figure 3. Contribution plots of predictor variables

3.3 Selection of Most Correlated Variables

We now select the most correlated variables based on the cross-comparison between contributions and qualities of representation from figure 2 and figure 3. Because PC3 is not significant from logistic regression, PC3 plays a less important role as compared to PC1 and PC2. (Table 2) Based on the percent of variability that can be explained by our PCs, variables from PC1 with the highest contributions and qualities of representation are considered most significant. These variables are Instrumental, Willingness, Abilities, Knowledge, Desires, Emotionality, and Appreciation. They belong to subcategories of social support, motivation, and empowerment, suggesting these social variables have the biggest impact on the risk of getting cervical cancer.

3.4 PCA Model in Determining Cancer Risk

The equation is:

$$\text{logit}(p) = -3.3389 - 0.38 * X_{\text{component1}} - 0.66 * X_{\text{component2}} + 0.095 * X_{\text{component3}} \quad (1)$$

Where components are created based on reconstruction of original x variables. Logistic regression based on decomposition of PCs shows the correlation between cervical cancer with our social theory variables (not shown).

Table.2. Summary statistics of logistic models based on PCs. Asterisks indicate the level of significance. The first and second PCs are significant. The AIC score is 35, which is the best score we get after comparison with other models. The multicollinearity and dimensionality are reduced. *** p < 0.0005 ** p <0.005

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.84511	1.20174	-3.200	0.001376**
PC1	-0.38121	0.10670	-3.573	0.000353***
PC2	-0.65943	0.21233	-3.106	0.001898**
PC3	0.09581	0.12944	0.740	0.459186

Table.3. Confusion Matrix

Predicted	0	1	Total
1	22	0	22
2	2	12	14
Total	24	12	36

The confusion matrix we build gives an accuracy of 94%. Usually, we are expecting the accuracy to be as high as possible. Here, we believe an accuracy of 94% is convincing and could prove the efficiency of our model and algorithm.

4. Conclusion

According to the corresponding contributions and qualities, our results indicate that the most significant social behavioral risk variables are instrumental social support, motivation of willingness, empowerment of abilities, empowerment of knowledge, empowerment of desires, social support associated with emotionality, and social support associated with appreciation. These variables belong to the three categories: social support, motivation, and empowerment. Therefore, we can conclude that among all other social theory derived factors, these three are the most influential components in determining the risk of cervical cancer in developing nations like Indonesia, which means that women with lower scores on these categories exhibit higher risk of getting cervical cancer. We believe the reason why social support is highly associated with the risk of cervical cancer is that the collectivist culture practiced in Indonesia teaches social conformity. Moreover, motivation represents a human's inner drive to succeed over life-challenges, which shapes a person’s behavior when it comes to life-challenges such as cancer. Social empowerment is understood as a system of belief and ability to develop self-confidence and self-autonomy. In fact, we believe that women who have higher scores of empowerment have more self-awareness toward cancer, thus reducing an individual's risk in getting the disease.

Furthermore, after PCA selection that helps to reduce multicollinearity, we can build a logistic model that aims to determine an individual's risk of getting cervical cancer through calculating their score of the questionnaire with the social behavioral risk factors. Proven by the confusion matrix, our model exhibits a high efficiency with an accuracy of 94%. In application, this model could be especially useful for women in low-income and middle-income countries like Indonesia. Not only because these countries share a highly similar social environment, but also this model is able to act as a free and instant preventive test for women who live in regions with difficulties reaching proper medical treatment. Moreover, since the constraint of the relatively stagnant economic condition, the expenses of regular cervical screening could often be unaffordable for women living in those countries. The model we built is the “first checkup”, and women would be able to obtain a general

idea of whether they possess a high possibility of getting cervical cancer. Regardless of the score, periodic screening and other physical examinations are recommended. Nevertheless, if a high risk is suggested by the model, one should plan to complete further cervical screening to determine whether they may indeed have cervical cancer. Thus, for patients who have cervical cancer which they are unknown of, the predictive result of our model may allow the patients to detect one's condition as early as possible so that there is a much higher chance for patients with cervical cancer to be treated right away, thus increasing the survival rate.

Nevertheless, there definitely exist some limitations in our study. For future research, we plan to include a larger sample size that can represent more regions of the world. It may be also advantageous if additional variables are considered when building the model.

References

[1] M. Urasa and E. Darj, "Knowledge of cervical cancer and screening practices of nurses at a regional hospital in Tanzania," Afr. Health Sci., vol. 11, no. 2, pp. 48–57, 2011.

[2] Cervical Cancer. (2020). Global Surgery Foundation. https://www.globalsurgeryfoundation.org/cervical-cancer

[3] Cervical cancer. (2019, December 2). World Health Organization. https://www.who.int/health-topics/cervical-cancer#tab=tab_1

[4] M. R. Balogun, O. O. Odukoya, M. a Oyediran, and P. I. Ujomu, "Cervical cancer awareness and preventive practices: a challenge for female urban slum dwellers in Lagos, Nigeria.," Afr. J. Reprod. Health, vol. 16, no. 1, pp. 75–82, Mar. 2012.

[5] E. J. Domingo, R. Noviani, M. R. M. Noor, C. a. Ngelangel, K. K. Limpaphayom, T. Van Thuan, K. S. Louie, and M. a. Quinn, "Epidemiology and Prevention of Cervical Cancer in Indonesia, Malaysia, the Philippines, Thailand and Vietnam," Vaccine, vol. 26, 2008.

[6] Williams-Brennan, L., Gastaldo, D., Cole, D. C., & Paszat, L. (2012). Social determinants of health associated with cervical cancer screening among women living in developing countries: A scoping review. Archives of Gynecology and Obstetrics, 286 (6), 1487–1505. https://doi.org/10.1007/s00404-012-2575-0

[7] C. Banura, F. M. Mirembe, A. R. Katahoire, P. B. Namujju, and E. K. Mbidde, "Universal routine HPV vaccination for young girls in Uganda: a review of opportunities and potential obstacles.," Infect. Agent. Cancer, vol. 7, no. 1, p. 24, Jan. 2012.

[8] Sari, H. E., Mudigdo, A., & Dermatoto, A. (2016). Multilevel analysis on the social determinants of cervical cancer in Yogyakarta. Journal of Epidemiology and PublicHealth, 01 (02), 100–107. https://doi.org/10.26911/jepublichealth.2016.01.02.03

[9] Unlersen, M.F., Sabanci, K., & Özcan, M. (2017). Determining cervical cancer possibility by using machine learning methods. Research Gate. https://www.researchgate.net/publication.322233711_Determing_Cervical_Possibility_by _Using_Machine_Learning_Methods.

[10] Sobar, Machmud, R., & Wijaya, A. (2016). Behavior determinant based cervical cancer early detection with machine learning algorithms. Advanced Science Letters, 22 (10), 3120–3123. https://doi.org/10.1166/asl.2016.7980