

Statistical Analysis of Risk Factors for Generalized Cervical Diseases

Weiwei Wu^a

School of Light Industry, Beijing Technology and Business University, Beijing, China

Keywords: Cervical Cancer, Cervical Intraepithelial Neoplasia (CIN), Logistic Regression, Venezuela.


Abstract: The aim of this study was to investigate the risk factors of generalized cervical diseases, including cervical cancer and cervical intraepithelial neoplasia (CIN). The data obtained in 2017 in Venezuela on 858 women were analysed. Logistic regression analyses were conducted to data both before and after balancing (oversampling) and data with different manipulation for missing values. 5, 4, 6 and 8 out of 9 variables were screened respectively as important risk factors of cervical diseases after backward selection in four acquired logistic regression models. Diagnosis of HPV infection, smoking and use of intrauterine device (IUD) were all screened in four models while number of sexual partners and age were included in three models separately. Diagnosis of HPV infection, use of IUD and number of sexual partners are positively correlated with cervical diseases. Smoking was negatively associated with cervical diseases based on the data. More variables were selected in model after data balancing. Model fitted to data after deleting missing values performed better than the model fitted to data with imputation. It requires high public attention to prevention of cervical diseases in Venezuela in terms of HPV infection, use of IUD and number of sex partners. Logistic regression models in our study are able to estimate patients' risks of cervical diseases and can be used as a predictive tool for prevention.

1 INTRODUCTION

Cervical cancer occurs in the cells of the cervix which links the uterus and vagina. It is a common chronic disease among females with 66% 5-year survival rate for all people (Cancer.Net 2018). Some pre-cancerous changes of cervix, including cervical intraepithelial neoplasia (CIN) and squamous intraepithelial lesion (SIL), will developed into malignant tumor if there is lack of proper treatment; therefore, early detection of pre-cancerous is necessary for prevention. Cervical cancers and its pre-cancers are able to be detected by Paps screening, Thinprep cytologic test (TCT), biopsy and etc. Similar to human immunodeficiency virus (HIV) for AIDS, human papilloma virus (HPV) is considered as a main cause of cervical cancer in vast investigations and approximately 95% of malignant cervical lesions are detected with HPV DNA (Gershenson 2004). The most common way of HPV transmission and infection is through sexual behaviours. A large body of clinical trials and studies proved that HPV vaccine is a highly and

long-term efficient prevention of HPV infection, cervical lesions and other relative anogenital warts disease in both female and male (Drolet 2019).

Cervical cancer is the fourth leading cause of mortality among cancers in women worldwide (Sung 2021). It has caused 604,127 new cases and 341,831 deaths in 2020 all around the world. Specifically, according to researches, estimated 80% of cervical cancer occurred globally are recorded in developing countries (Correnti 2011). The incidence and mortality rate of cervical cancer is extremely high in Latin America despite the availability of Paps screening since women with low socioeconomic status are less accessible to the screening and less educated with the importance of it (Villa 2012). In Venezuela, according to the GLOBOCAN 2018 data, cervical cancer deaths reached 2,210 or 1.34% of total deaths, which is a relatively high rate compared to other developed countries like the United States (Global Cancer Observatory 2018). Moreover, Venezuela currently provides nationwide cervical cancer screening program including Paps for women of 25 to 65 years old, but HPV vaccine has not been provided and incorporated into national vaccination program,

^a <https://orcid.org/0000-0001-9816-7703>

which represents there is no adequate vaccine coverage (Bardach 2017). Specifically, there is a lack of access to prevention, early diagnosis and even treatment that could be provided (Denny 2012).

Even though cervical cancer has been well studied and discussed, there are still some problems exist. To be specific, it is obscure for people to analyse risk of generalized cervical diseases, including CIN, cervicitis, cervical cancer, etc. The occurrence and development of cervical cancer has a gradual evolutionary process, which can take from several years to decades. A huge body of evidence indicates that chronic cervical diseases have potential to develop into cancer. Therefore, it is necessary to explore risk factors related to generalized cervical diseases. Consequently, it is useful for people to find this disease earlier and have a better way to prevent disease than it used to be.

Considering the limited investigation of risk factors and generalized cervical disease, we investigated the risk factors of cervical diseases by analysing the data from Hospital Universitario de Caracas.

2 METHODS

2.1 Data Resource

The dataset was obtained from the UCI Machine Learning Repository, which is a collection of domain theories, data generators and databases from various fields. It was established at UC Irvine in 1987 and was widely used as a public practice source of machine learning algorithms all around the world.

This dataset was collected at Hospital Universitario de Caracas' in Caracas, Venezuela, Latin-America and the study of the data collection was published in 2017 (Fernandes 2017). The dataset contains 36 relative variables and 858 patients, including feature information of historic medical records, demographic information and habits. Due to privacy concern, some patients rejected to some questions during the data collecting, which leads to some missing values in this dataset.

2.2 Research Variables

Research variables are showed in Table 1

Dependent variable is 'Diagnosis of cervical diseases' (Dx. Cervical), which is the combination of diagnosis of cervical cancer and diagnosis of CIN.

It is a categorical variable which represents the existence of cervical disease of patients. Number '1' represents that a patient has confirmed cervical diseases while number '0' represents a patient is cervical healthy.

Nine known variables were considered as independent variables, which are 'Age', 'Number of sexual partners', 'Age of first sexual intercourse', 'Number of pregnancies', 'Smokes', 'Sexually transmitted disease infection (STDs)', 'Dx. HPV', 'Intrauterine Device (IUD)' and 'Hormonal Contraceptives'.

'Age' is a numerical variable with the minimum of 13 and maximum of 84. 'Number of sexual partners' is a numerical variable with the minimum of 1 and maximum of 28. 'Age of first sexual intercourse' is a numerical variable with the minimum of 10 and maximum of 32. 'Number of pregnancies' is a numerical variable with the minimum of 0 and maximum of 11. 'Smokes', a categorical variable, represents whether or not a patient smoke. Number '1' represents that a patient smokes or used to smoke while number '0' represents a patient has no smoking history. 'STDs', a categorical variable, represents whether a patient have sexually transmitted diseases. Number '1' represents that a patient has STDs while number '0' represents a patient does not have STDs. 'Dx. HPV', a categorical variable, represents whether a patient was diagnosed as HPV infection. Number '1' represents that a patient was diagnosed as HPV infection while number '0' represents a patient was not. 'IUD', a categorical variable, represents whether a patient have used intrauterine device – a device fitted inside uterus for birth control. Number '1' represents that a patient uses or once used IUD while number '0' represents a patient has never used IUD. 'Hormonal Contraceptives', a categorical variable, represents whether a patient have used hormonal medication for contraception. Number '1' represents that a patient uses or once used hormonal contraceptives while number '0' represents a patient has never used hormonal contraceptives.

Table 1: Distribution of selected characteristics.

Dependent Variable	Type		Number	Percent (%)
Dx. Cervical Diseases	categorical	Yes	27	3.14%
		No	831	96.86%
Independent Variable	Type		Range/ Number	Percent (%)
Number of sexual partners	numerical		1-28	
Age of first sexual intercourse	numerical		10-32	
Number of pregnancies	numerical		0-11	
Age	numerical		13-84	
Smokes	categorical	Yes	124	14.45%
		No	734	85.55%
STDs	categorical	Yes	79	9.21%
		No	779	90.79%
Dx. HPV	categorical	Yes	18	2.19%
		No	840	97.81%
Hormonal contraceptives	categorical	Yes	565	65.85%
		No	293	31.15%
IUD	categorical	Yes	83	9.67%
		No	775	90.33%

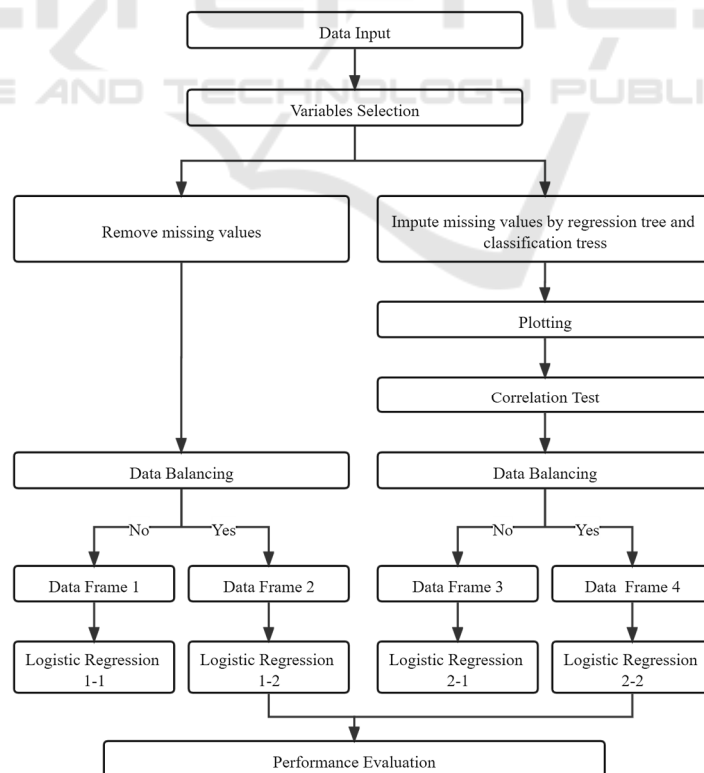


Figure 1: Implementation flow chat.

2.3 Statistical Method

We opted the CSV file from UCI after a comprehensive searching for well-quality data. The process and manipulation of the data is presented in Figure 1.

The CSV file was inputted to R Studio (R version 4.1.1). The missing values in the dataset were treated by two kinds of manipulations and two new data frames were formed. One data frame removed all missing values while another imputed missing values by decision tree. Then, visualization and correlation tests, including Chi-square test and t-test, were performed to the second data frame. Data balancing (Oversampling) was conducted to both two data frames and two extra data frames were acquired.

Logistic regression (LR) was used to analyse all four data frames. Logistic regression is a multivariate method which was invented for binary outcomes (labelled as '0' and '1'). It is used when studies are concerned with whether an event happened or not, which is appropriate for models for decision making and estimate of disease occurrence

and thus is widely used in health and medical researches. In logistic regression, the logarithm of the odds ratio (log-odd) was converted into the probability of outcomes by an algebraic manipulation (Boateng2019, Glantz 2017, Hosmer 1989). Variables were selected by backward stepwise method during analysis.

Then, Performance evaluation was conducted to the results of logistic regression. 75% of samples was selected as training data and 25% was selected as testing data at random. Evaluation process was conducted 100 times for each model to reduce error and average values was taken as results.

3 RESULTS

3.1 Study Population

858 participants were included in the data, 27 participants were diagnosed as Cervical diseases and 831 participants were cervical healthy in 'Data Frame 3'.

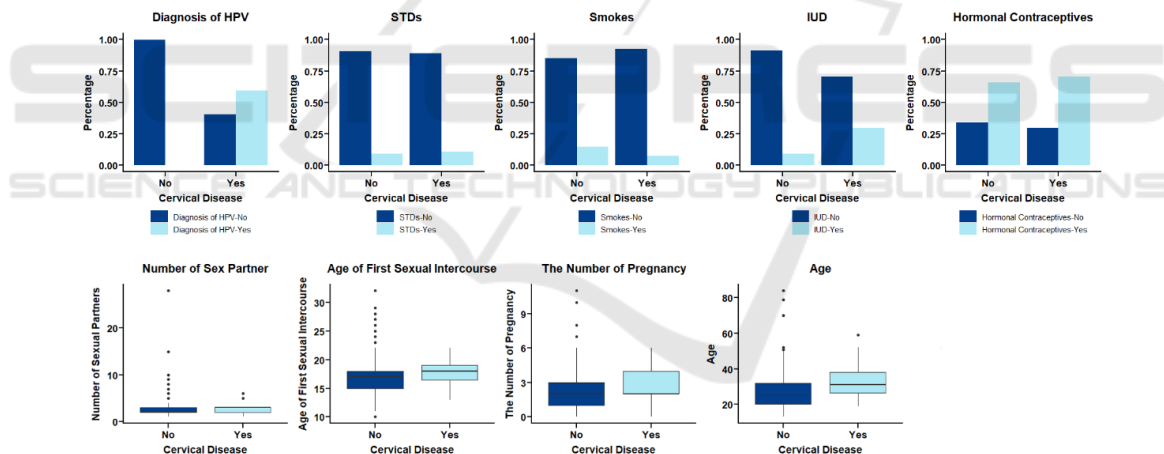


Figure 2: The bar chats and boxplot plots of association between cervical diseases and risk factors

Distribution and ranges of selected variables of study population were showed in the Table 1. Most of the women were nonsmokers (85.55%), not diagnosed with STDs (90.79%), not infected with HPV (97.81%), and never used IUD (90.33%). More than half of them took hormonal contraceptives (65.85%). The mean of age of study population is 26.82. The average number of pregnancies is 2.27. Average age of first intercourse is 17.00. The mean of the number of sexual partners of all participants is 2.53.

3.2 EDA

The association between cervical diseases and risk factors were presents in the Figure 2. In the first bar chart, a large proportion of people with HPV infection also have Cervical Diseases, which proved previous studies about HPV and these maladies. In the bar chart of IUD, the percentage of IUD use is greater in cervical diseases diagnosed patients than in cervical health patients. In bar chart of Smokes, cervical diseases patients have less proportion of smokers. For STDs and Hormonal Contraceptives,

the distribution of cervical diseases is nearly even. From the box plot, the average age, number of pregnancies and age of first sexual intercourse are

all slightly greater in confirmed cervical diseases patients.

Table 2: Correlation tests between cervical diseases and each risk factor.

CHISQUARE TEST					
Variable	STDs	Dx. HPV	Hormonal Contraceptives	IUD	Smokes
x-square	<0.01	415.21	0.09	10.46	0.61
df	1	1	1	1	1
p-value	0.99	<0.01	0.78	<0.01	0.44
T-TEST					
Variable	Age	Number of sexual partners	Age of first sexual intercourse	Number of pregnancies	
t score	6.19	2.25	3.13	2.25	
df	87	23	41	21	
p-value	<0.01	0.03	<0.01	0.04	

3.3 Correlation Tests

Results of correlation tests were presents in the Table 2.

According to the p-value, ‘Dx HPV’, ‘IUD’, ‘Number of sexual partners’, ‘Age of first sexual intercourse’, ‘Number of pregnancies’, ‘Age’ are significant parameters for cervical diseases. ‘STDs’, ‘Hormonal Contraceptives’ and ‘Smokes’ are insignificant.

3.4 Data Balance

For ‘Data Frame 1’, the data frame that removed all missing values, after over-sampling, 648 diagnosed cervical diseases patients and 648 none diagnosed patients were obtained and formed ‘Data Frame 2’.

For ‘Data Frame 3’, the data frame that imputed missing values by decision tree, after over-sampling, 831 diagnosed cervical diseases patients and 831 none diagnosed patients were obtained and formed ‘Data Frame 4’.

3.5 Logistic Regression

Four logistic regressions were calculated to predict the occurrence of cervical diseases. Table 3 shows the results of four logistic regression models for data frames with different manipulation. LR 1-1, LR 1-2, LR2-1, LR 2-2 were fitted to ‘Data Frame 1’, ‘Data

Frame 2’, ‘Data Frame 3’, ‘Data Frame 4’ separately. Data used for LR 1-1 and 2-1 are imbalanced and for LR 1-2 and LR 2-2 are balanced by oversampling.

Variables in these four models were all selected by backward method. 5, 4, 6 and 8 variables were screen respectively as meaningful risk factors of cervical diseases in LR 1-1, LR 2-1, LR 1-2 and LR 2-2. The equation of LR 1-1 is

$$\log \frac{P}{1-P} = -5.6317 + 31.7832X_{Dx.HPV} + 2.2513X_{IUD} + 0.5737X_{No.of.sexual.partners} - 0.5929X_{No.of.pregnancies} - 26.3415X_{Smokes} \quad (1)$$

and Akaike information criterion (AIC) of the model is 66.13. ‘Age of first sexual intercourse’, ‘Age’, ‘STDs’, ‘Hormonal Contraceptives’ were removed in LR 1-1. The equation of LR 2-1 is

$$\log \frac{P}{1-P} = -5.7685 + 6.9936X_{Dx.HPV} + 1.3372X_{IUD} + 0.0449X_{Age} - 2.3251X_{Smokes} \quad (2)$$

and AIC of the model is 127.06. ‘Number of sexual partners’, ‘Age of first sexual intercourse’, ‘Age’, ‘STDs’, ‘Hormonal Contraceptives’ are removed in LR 2-1. The equation of LR 1-2 is

$$\log \frac{P}{1-P} = -2.2062 + 47.2360X_{Dx.HPV} + 2.4577X_{IUD} + 0.8233X_{No.of.sexual.partners} - 0.0718X_{Age} - 3.4499X_{STDs} - 38.0888X_{Smokes} \quad (3)$$

and AIC of the model is 595.04. ‘Age of first sexual intercourse’, ‘Number of sexual partners’, ‘Hormonal Contraceptives’ were removed in LR 1-2.

Table 3: Logistic regression analyses for cervical diseases.

LR 1-1			LR 2-1		
AIC = 66.13			AIC = 127.06		
Variable	Estimate	Pr (> z)	Variable	Estimate	Pr (> z)
(Intercept)	-5.6317	<0.0001	(Intercept)	-5.7685	<0.0001
Number of sexual partners	0.5717	0.0659	Age	0.0449	0.1062
Number of pregnancies	-0.5929	0.1771	Smokes	-2.3251	0.0916
Smokes	-26.3415	0.9912	IUD	1.3372	0.0468
IUD	2.2513	0.0320	Dx. HPV	6.9936	<0.0001
Dx. HPV	31.7832	0.9894			

LR 1-2			LR 2-2		
AIC = 595.04			AIC = 1182.8		
Variable	Estimate	Pr (> z)	Variable	Estimate	Pr (> z)
(Intercept)	-2.2062	<0.0001	(Intercept)	-0.1264	0.8409
Age	-0.0718	0.0002	Age	0.0770	<0.0001
Number of sexual partners	0.8233	<0.0001	Number of sexual partners	0.3330	<0.0001
Smokes	-38.0888	0.9949	Age of first sexual intercourse	-0.1895	<0.0001
IUD	2.4577	<0.0001	Number of pregnancies	-0.3767	<0.0001
STDs	-3.4499	0.0014	Smokes	-6.8497	<0.0001
Dx. HPV	47.2360	0.9937	Hormonal Contraceptives	0.2313	0.1475
			IUD	1.1140	<0.0001
			Dx. HPV	9.7419	<0.0001

Table 4: The performance of two logistic regression for cervical diseases.

Model	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	AUC*(%)
LR 1-2	89.18(3.58)	92.79(2.27)	90.94(1.96)	92.66(2.23)	96.74(0.70)
LR 2-2	73.31(2.25)	90.15(2.35)	81.73(1.56)	88.15(2.83)	90.71(0.10)

The equation of LR 2-2 is

$$\logit(P) = \ln\left(\frac{P}{1-P}\right) = -0.1264 + 9.7419X_{Dx.HPV} + 1.1140X_{IUD} + 0.3330X_{No.of.sexual.partners} + 0.2313X_{Hormonal.Contraceptives} + 0.0770X_{Age} - 0.1895X_{Age.of.first.sexual.intercourse} - 0.3767X_{No.of.pregnancies} - 6.8497X_{Smokes} \quad (4)$$

and AIC of the model is 1182.8. ‘STDs’ were removed in LR 2-2. Partial variables in models are not significant (Pr > 0.05).

‘Diagnosis of HPV infection’, ‘Smokes’, ‘IUD’ were all screen as important risk factors in these four models. Among them, ‘Diagnosis of HPV’ and ‘Smokes’ are highly associate with cervical diseases. ‘Smokes’ are negatively correlated with cervical diseases. ‘Number of sexual partners’ and ‘Age’ were included in three models separately.

After an inverse operation of logistic transaction, the value of P, which represents the probability of occurrence of cervical diseases, was acquired. A value of P that is closer to 0 was regarded as less likely of cervical diseases, while that is closer to 1 was regarded as more likely of cervical diseases.

Comparison of LR 1-1 and LR 1-2 and comparison of LR 2-1 and LR2-2 shows that more variables were included in model after data balancing. Comparison of models with different manipulation, LR 1-2 and LR 2-2, shows that the correlation of ‘Diagnosis of HPV infection’ and

cervical diseases are lower in model for imputed data.

3.6 Performance Evaluation of Models

Results of the performance evaluation of LR 1-2 and LR 2-2 are presented in Table.4.

For LR 1-2, 89.13% sensitivity, 92.79% specificity, 90.94% accuracy, 92.66% precision were observed. AUC of LR 1-2 is 96.74%. For LR 2-2, 73.31% sensitivity, 90.15% specificity, 81.73% accuracy, 88.15% precision were observed. AUC of LR 2-2 is 90.71%.

4 DISCUSSION

According to the results, HPV infection, use of IUD and number of sex partners need high attention for prevention of cervical diseases. Our results also support the conclusion of previous studies that there is a consistent correlation between HPV and cervical. From the result of logistic regression, it is apparent that there exists a negative correlation between Smokes and cervical diseases. However, according to previous study, it is apparent that smokers have high risk of developing cervical

cancer in US (Sierra-Torres 2003). To be specific, there is a positive relationship between smoking and diagnosis of cervical disease. The difference is probably due to the varied situation of each country that Venezuela is a low-income developing country, but the United States is a developed country. Specifically, not every woman in Venezuela may have access to smoking due to the financial issue and the high percent of excise tax in cigarette. Compared to Venezuela, people in the U.S may have easier access to smoking no matter what income they have received. Moreover, Venezuela execute more extensive and stricter ban on smoking and enforce more bans on advertising than the U.S, which may result in lower rate of smoking among women (Venezuela 2019, United States Tobacco Atlas 2021). Therefore, it can possibly explain the negative correlation in Venezuela and the positive correlation in the U.S. Further evidence and comprehensive researches are needed to prove this inference.

Different from previous studies that consider the effects of risk factors on CIN or cervical cancer separately, we focused on generalized cervical diseases including both CIN and cervical cancers. The combination of CIN and cervical cancer might contribute to the early control and prevention of generalized cervical diseases. We also compared different models that were fitted to data both before and after balancing (oversampling) and data with different manipulation of missing values.

Nevertheless, our studies still have some limitations to be considered. Due to the limitation of our dataset, we only consider the diagnosis of CIN and diagnosis of cervical cancer. If there is access to data including more other cervical diseases, like cervical polyp, cervical cyst etc., models are able to be further improved and optimized. Moreover, since our dataset was collected from Venezuela, it needs to be cautious when generalizing the results and conclusions to other regions. Venezuela is a low-income country, so the data may only represent the conditions in low-income country rather than other developed or developing countries. In addition, because of the privacy concerns of some women that they did not share complete information in data collection, biases were introduced into analyses. Lastly, risk factors were screened in our study by using logistic regression, the results can be further confirmed by using random forest subsequently.

5 CONCLUSIONS

‘Diagnosis of HPV infection’, ‘IUD’, ‘Number of sexual partners’ and ‘Age’ are risk factors of cervical cancer in Venezuela. Logistic regression models in our study can estimate patients’ risks of cervical diseases and can be used as a tool for prevention. In the future, we will employ the technique of random forest to analyse statistical correlation between cervical diseases and all independent variables discussed in this paper and make comparison on these two statistical methods.

REFERENCES

- Bardach, A. E., Garay, O. U., Calderón, M., Pichón-Riviére, A., Augustovski, F., Martí, S. G., Cortiñas, P., Gonzalez, M., Naranjo, L. T., Gomez, J. A., & Caporale, J. E. (2017). Health Economic Evaluation of human papillomavirus vaccines in women from Venezuela by a lifetime markov cohort model. *J. BMC Public Health*, 17, 152.
- Boateng, E. Y., Abaye, D. A. (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research. *J. Journal of Data Analysis and Information Processing*, 07, 190–207.
- Cancer.Net. - Cervical cancer. (2021). Retrieved from <https://www.cancer.net/cancer-types/cervical-cancer/statistic>.
- Correnti, M., Medina, F., Cavazza, M. E., Rennola, A., Ávila, M., & Fernández, A. (2011). Human papillomavirus (HPV) type distribution in cervical carcinoma, low-grade, and high-grade squamous intraepithelial lesions in Venezuelan women. *J. Gynecologic Oncology*, 121, 527–531.
- Denny, L. (2012). Cervical Cancer: Prevention and Treatment. Retrieved from <https://www.discoverymedicine.com/Lynette-Denny/2012/08/27/cervical-cancer-prevention-and-treatment/>.
- Drolet, M., Bénard, É., Pérez, N., Brisson, M., Ali, H., Boily, M.-C., Baldo, V., Brassard, P., Brotherton, J. M., Callander, D., Checchi, M., Chow, E. P., Cocchio, S., Dalianis, T., Deeks, S. L., Dehlendorff, C., Donovan, B., Fairley, C. K., Flagg, E. W., ... Yu, B. N. (2019). Population-level impact and herd effects following the introduction of human papillomavirus vaccination programmes: Updated systematic review and meta-analysis. *J. The Lancet*, 394, 497–509.
- Fernandes, K., Cardoso, J. S., Fernandes, J. (2017). Transfer learning with partial observability applied to cervical cancer screening. *J. Pattern Recognition and Image Analysis*, 10255, 243–250.
- Gershenson, D. M., McGuire, W. P., Gore, M., Quinn, M. A., & Thomas, G., 2004. *Gynecologic cancer: Controversies in management*. Elsevier Ltd. Philadelphia.

- Glantz, S., Slinker, B., Neilands, B., 2017. *Primer of Applied Regression and Analysis of Variance*. McGraw Hill. 3rd edition.
- Global Cancer Observatory. (2018). Retrieved from <https://gco.iarc.fr/>.
- Hosmer, D. W., Lemeshow, S., 1989. *Applied logistic regression*. New York: Wiley.
- Sierra-Torres, C. H., Tyring, S. K., Au, W. W. (2003). Risk contribution of sexual behavior and cigarette smoking to cervical neoplasia. *J. International Journal of Gynecological Cancer*. 13, 617–625.
- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *J. CA: A Cancer Journal for Clinicians*. 71, 209–249.
- United States. Tobacco Atlas. (2021). Retrieved from <https://tobaccoatlas.org/country/usa/>.
- Venezuela. Tobacco Atlas. (2021). Retrieved, from <https://tobaccoatlas.org/country/venezuela/>.
- Villa, L. L. (2012). Cervical cancer in Latin America and the Caribbean: the problem and the way to solutions. *J. Cancer Epidemiol Biomarkers Prev*. 21, 1409–1413.

