# Statistical Analysis on Relationship between HPV infection and Risk Factors

**Jiarui Chang**[1, a, †], **Jiayi Chen**[2, b, †], **Xiaoqing Chen**[3, c, †]

[1]Rothesay Netherwood School, Rothesay E2E5H1, Canada

[2]Dukekunshan University, No. 8 Duke Avenue, Kunshan 215316, China

[3]University of Toronto, 27 King's College Circle, Toronto, Ontario M5S 1A1, Canada

[a]joy.chang@rns.cc, [b]jiayi.chen364@dukekunshan.edu.cn, csherley.chen@mail.utoronto.ca

[†]These authors contributed equally.

**Abstract -** Cervical cancer is one of the most preventable cancers, and human papillomavirus (HPV) infection is one of the major factors that can cause it. A cost-effective model that helps low-income countries diagnose cervical cancer needs to be found as soon as possible. Data of 858 patients in 2017 was used to analyze possible behaviors that could cause HPV infection. The logistic regression model was built to find the most significant variables that can help estimate HPV infection. A confusion matrix was used to evaluate the reliability of the model. Logistic regression analysis showed that the final explanatory variables were age, number of sexual partners, number of pregnancies, and years of using hormonal contraceptives ($p < 0.05$). The optimized new model has an accuracy of 0.5877. Our study suggests that the behaviors commonly viewed as risk factors of HPV infection are insignificant when combined to analyze. The findings have played a great role in enlightening other studies on the infection of HPV and the pathogenic factors of cervical cancer and opened a new path.

***Keywords:*** Cervical cancer, HPV infection, Sexual behaviors

## 1. Introduction

Cervical cancer is cancer arising from the cervix. It is due to the abnormal growth of cells that can invade or spread to other parts of the body [1]. According to the World Health Organization (WHO), cervical

cancer develops in a woman's cervix and is one of the most preventable cancers when diagnosed. Although corresponding vaccines have been developed, it still kills 300,000 women worldwide every year [2]. Moreover, it is the fourth most common cancer for females globally due to the low screening rates, whether because of financial or language problems, leading to inadvertence and irreparable deterioration in cancer detection, especially in developing countries [3]. Among all risk factors, human papillomavirus (HPV) infection is the major one that almost always causes cervical cancer. Although the human immune system can resist 90% of types of HPV automatically, the persistent infection of HPV also exists, and it will lead to cervical cancer [3]. In research from Chabeda et al., there are three commercially available HPV vaccines used to prevent infection. However, these vaccines are not effective in eliminating pre-existing infections, so a large number of women who are already infected with HPV will not benefit from the current vaccine [4]. At the same time, developing countries bear the greatest burden of cervical cancer from HPV infection due to a lack of resources to implement effective vaccination and screening programs [5]. Especially in India, vaccination coverage is not extensive, which causes cervical cancer to become the second most common cancer. Although government support can improve screening rates, there will inevitably be an oversight [2]. According to Hu & Ma's research, most females are diagnosed with cervical cancer between 30-50 years old worldwide, and the age of patients tends to be younger [3].

Since cervical cancer is common and there is no effective treatment for HPV infection, a lot of studies have been done in this field and have achieved some findings. Some studies mainly research the mutation of genes after the infection of HPV that can be used to analyze cancerous tissues and non-cancerous tissues [3]. Moreover, the NGS-based HPV testing method can be the potential detection method to replace the current method in the future [3]. Because the high-risk HPV can cause cervical cancer, some studies would like to develop the vaccine to contain more high-risk types of HPV that can produce more such types of antibodies. Moreover, some studies indicate that certain behaviors will increase the infection rate. In McDermott and Bowles's research, they indicate that smoking will increase the infection rate of HPV, and the longer smoking year will cause a higher rate of HPV infection [6]. Moreover, females with unstable estrogen levels will also have a higher rate of HPV infection. In De Villiers's research, he states that contraceptives will influence females' hormones and affect females' estrogen levels [7]. Moreover, the first sexual intercourse at an early age also will cause the infection. In Kahn et al. research, they illustrate that the younger female who has sexual intercourse at an early age has more cases of HPV infection than adult females [8]. Overall, some unexpected behaviors will cause a higher infection rate. These behaviors are mentioned in some studies individually, and fewer studies research whether these unexpected behaviors combined will increase infection rate and increase significantly or not. Thus, our study aimed to research whether the population with all unexpected behaviors has a higher HPV infection probability than those who do not.

## 2. Methods

### *2.1. Data Collection*

The dataset of cervical cancer risk classification in our research was obtained from UCI Repository. It

claimed the original data was from Kelwin Fernandes, Jaime S. Cardoso and Jessica Fernandes. With the full name of UC Irvine Machine Learning Repository, UCI Repository is a collection of databases, domain theories, and data generators used by the machine learning community for the empirical analysis of machine learning algorithms [9]. Kelwin Fernandes and Cardoso are from INESC TEC, and Universidade do Porto, both located in Porto, Portugal. Jessica Fernandes is from Universidad Central de Venezuela located in Caracas, Venezuela [10].

The dataset was collected at "Hospital Universitario de Caracas" in Caracas, Venezuela, comprising demographic information, habits, and historical medical records of 858 patients in 2017. Several patients decided not to answer some of the questions because of privacy concerns, which were then counted as missing values [10]. The collection methods included sending out questionnaires to patients and inquiring about the examination records saved by the hospital.

*2.2. Variable Statistics*

The original dataset includes 36 attributes, some of which are regarded as the potential factors of the infection of HPV, such as the number of sexual partners and whether smoking. In contrast, the other attributes are diagnosing diseases closely related to cervical cancer, such as genital warts, syphilis and genital herpes. Among these attributes, we selected six factors as our independent variables, one diagnosis (Dx. HPV) as the dependent variable since the infection of HPV largely increases the risk of getting cervical cancer, and two factors as the confounders. We chose these independent variables as our research objects because they are the riskiest factors of HPV infection, and they are easy to understand.

Table 1 Distribution of selected dummy variables

| Characteristics (dummy) | Label | N | Percentage (%) |
|---|---|---|---|
| Contraceptives | whether the participant has ever taken contraceptives or not | Yes: 436 | 64.49 |
| | | No: 240 | 35.50 |
| Smoke | whether the participant has ever smoked or not | Yes: 96 | 14.20 |
| | | No: 580 | 85.79 |
| HPV | whether the participant has been infected with HPV or not | Yes: 16 | 2.36 |
| | | No: 660 | 97.63 |

Table 2 Distribution of selected continuous variables

| Characteristics (continuous) | Label | Mean (SD) | Range |
|---|---|---|---|
| Age | the age of the participant | 27.24 (8.69) | 13 - 84 |
| Number of sexual partners | how many sexual partners does the participant have | 2.521 (1.633) | 1 - 28 |
| First sex | the age of the participant when she had her first sex | 17.16 (2.86) | 10 - 32 |
| Number of pregnancies | how many pregnancies does the participant have | 2.325 (1.46) | 0 - 11 |
| Smoke year | how many years has the participant smoked | 1.22 (4.17) | 0 - 37 |
| Contraceptives year | how many years has the participant taken contraceptives | 2.33 (3.86) | 0 - 30 |

SD: Standard deviation of each variable.
N: The total number
Range: The minimum value to the maximum value

According to *Table 1* and *Table 2* above, the variables comprise dummy variables and continuous variables. The mean values and standard deviations were calculated to get an overview of the dataset. First, we cleaned the data by deleting all the missing values. After that, we determined the independent variables, dependent variables and confounders, and classified them. Lastly, suitable graphs were created to inspect the independent variables and identify the relationship between independent variables and dependent variables.

*2.3. Regression Model Selection*

Since the dependent variable chosen is binary, a multi-predictor logistic model was fitted to the oversampled dataset to test the research hypothesis regarding the relationship between the likelihood that a woman is infected with HPV and her sexual experiences. Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a dataset [11]. It is a generalized linear regression analysis model commonly used to solve binary classification problems, which is widely used in data mining, automatic disease diagnosis, economic forecasting, etc. To reduce the influence of independent variables on model evaluation, the stepwise selection was used to find the most significant variables. After that, the logit formula came out to estimate the model's efficiency. Suppose there is a situation where the amount of data in the two categories of the dependent variable is

highly polarized. In that case, we will use oversampling to reduce the possible impact of data imbalance. When evaluating the model, sensitivity is always closer to 100 is better, and this is how we can handle class imbalance problems efficiently and smartly. Then we evaluated the efficiency of our model by using the confusion matrix. This specific table layout allows visualization of the performance of an algorithm, typically a supervised learning one. The logistic procedure carried out the logistic regression analysis in RStudio version 1.4.1106 in the Windows 10 environment.

*2.4. Model Diagnostics*

We used the logistic model and calculated the interaction of all independent variables and the deviance residuals. The deviance residuals are the measurement of deviance of observations. It can be used to check the fitness of the model. If the model is good, the deviance will be smaller. Otherwise, the deviance will be larger.

## 3 Results

*3.1. Exploratory Data Analysis*

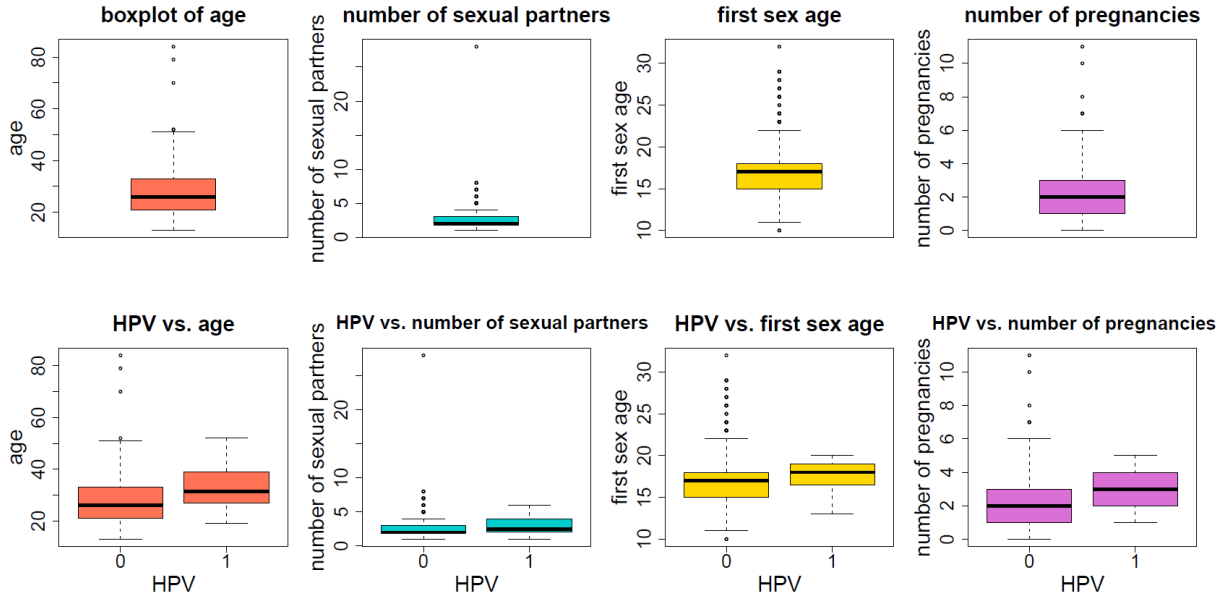


Figure 1 Boxplots of several continuous variables (the first row) & boxplots of HPV against the continuous variables (the second row)

The first row of *Figure 1* involves four density plots that illustrate the continuous independent variables, age, number of sexual partners, age having the first sex, and number of pregnancies. The boxplot of age shows that the age of subjects is concentrated between 22 and 34. The most frequent age is about 27 years old. The outliers are spread out in the range between 53 and 85. Based on the boxplot of the number of sexual partners of people infected, we can see that most subjects have 2 to 3 sexual partners.

The outliers are below 10, yet an extreme outlier shows that one subject has 28 sexual partners, but it cannot be counted as an error until the further examination is examined. The boxplot of first sex age proves that most subjects had their first sex at around 17 years old. More outliers are concentrated between 22 and 34, compared to only one outlier below the lower quartile 11. The fourth boxplot of the number of pregnancies shows that the most frequent pregnancy time is 2. The less frequent pregnancy numbers are 1 and 3. The four outliers are all located from 7 and 11, which make the range above the higher quartile bigger than the range below the lower quartile.

Secondly, four boxplots in the second row of *Figure 1* are graphed to investigate the relationship between the categorical dependent variables (HPV) and the continuous independent variables (age, number of sexual partners, first sex age, and number of pregnancies). The boxplot of "HPV vs age" shows that people infected by HPV are concentrated between 30 and 38, and the subjects not infected are younger than the infected. The age of the most infected is about 34 years old. Rather than older women, younger women tilt to getting HPV more since the medium is closer to the lower quartile and the lower range. The outliers are spread out in the range between 53 and 85. Based on the boxplot of the number of sexual partners, we can see that the infected subjects have more sexual partners than the uninfected ones. Most of the infected subjects have 2 to 4 sexual partners, and the three sexual partners are the most frequent. The outliers of the uninfected are below 10. Yet, an extreme outlier shows that one subject has 28 sexual partners – but it cannot be counted as an error until examining it further. The boxplot of first sex age proves that most infected subjects had their first sex at around 17 years old. The fourth boxplot of the number of pregnancies shows that the most frequent pregnancy time of the infected subjects is 3. The less frequent pregnancy numbers are 2 and 4. Noticeably, the boxplots of all the infected subjects' data have no outliers. It might be because the number of infected subjects is smaller than the number of uninfected ones, but we still need further investigation to explain that.

### *3.2 Performance evaluation*

After analyzing the distribution of the dependent variable, it was clearly evident that over 90% of the data are in class 0, and the remaining less than 10% are in class 1. Such a big difference is observed in the amount of data available, and balancing the data is necessary. After oversampling, there were 660 uninfected cases and 640 infected cases in the dataset. From the confusion matrix of the oversampled dataset, the accuracy was 0.3905, the sensitivity was 0.2308, and precision and specificity were both 1.

### *3.3 Logistic Regression*

Table 3 Coefficients of independent variables before stepwise selection

| Coefficients: | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -3.53681 | 0.52266 | -6.767 | 1.32e-11 | *** |
| Age | 0.02108 | 0.01184 | 1.781 | 0.07497 | . |

| | | | | | |
|---|---|---|---|---|---|
| n_sexual_partners | 0.21846 | 0.04467 | 4.890 | 1.01e-06 | *** |
| first_sex | 0.08168 | 0.02937 | 2.782 | 0.00541 | ** |
| n_pregnancies | 0.19288 | 0.06003 | 3.213 | 0.00131 | ** |
| smoke1 | 0.20979 | 0.22337 | 0.939 | 0.34761 | |
| smoke_year | 0.01083 | 0.01495 | 0.724 | 0.46890 | |
| contraceptives1 | 0.17817 | 0.14742 | 1.209 | 0.22682 | |
| contraceptives_year | 0.06392 | 0.01719 | 3.719 | 0.00020 | *** |

Table 4 Coefficients of independent variables after stepwise selection

| Coefficients: | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -3.46009 | 0.50996 | -6.785 | 1.16e-11 | *** |
| Age | 0.02498 | 0.01032 | 2.420 | 0.01552 | * |
| n_sexual_partners | 0.22147 | 0.04472 | 4.953 | 7.32e-07 | *** |
| first_sex | 0.07761 | 0.02899 | 2.677 | 0.00743 | ** |
| n_pregnancies | 0.18162 | 0.05858 | 3.100 | 0.00193 | ** |
| smoke1 | 0.34310 | 0.17142 | 2.002 | 0.04534 | * |
| contraceptives_year | 0.07087 | 0.01531 | 4.630 | 3.66e-06 | *** |

Before any modification for the oversampled dataset, there were 8 independent variables, and many of them had a p value larger than 0.1, which means no significance. If they were kept, the final estimation would be largely affected. The method backwards and forward selection using AIC removed the irrelevant variables. Results showed that the final explanatory variables chosen were age, number of sexual partners, number of pregnancies, and years of using hormonal contraceptives ($p < 0.05$).

After fitting the best, the formula of the logistic model came out:

$$\text{Logit (HPV)} = -3.46 + 0.02x_1 + 0.22x_2 + 0.08x_3 + 0.18x_4 + 0.34x_5 + 0.07x_6 \quad (1)$$

where

$x_1$ stands for age,

$x_2$ stands for the number of sexual partners,

$x_3$ means the age of first sex,

$x_4$ means the number of pregnancies,
$x_5$ means whether one smokes or not,
$x_6$ means the year of using hormonal contraceptives.

Multiple linear regression was calculated to predict weight based on their height and sex. A significant regression equation was found ($F(2, 13)= 981.202$, $p < .000$), with an R2 of .993. Participants' predicted probability of HPV infection is equal to $-3.46 + 0.02\ x_1 + 0.22\ x_2 + 0.08\ x_3 + 0.18\ x_4 + 0.34\ x_5 + 0.07\ x_6$, where x5 is coded as 1 = smoke, 2 = do not smoke. According to the model, the logarithm of a woman being infected with HPV was positively related to age, number of sexual partners, age of first sex, number of pregnancies, whether the woman smokes, and year of using contraceptives. Participant's infection probability increased by 0.22 pounds for each sexual partner, and smokers weighed 0.34 pounds more than non-smokers. In other words, the higher these independent variables, the more likely it is that a woman would be infected with HPV. After evaluation, the accuracy was 0.5877, the sensitivity was 0.5, the precision was 0.5896, and the specificity was 0.6707.

Table 5 Confusion matrix for testing dataset after regression

| | True | | |
|---|---|---|---|
| Predicted | 0 | 1 | Total |
| 0 | 114 | 58 | 172 |
| 1 | 53 | 100 | 153 |
| Total | 167 | 158 | 325 |

*3.4 Model Diagnostics*

Table 6 Deviance residuals of the logistic model

| **Deviance Residuals** | | | | |
|---|---|---|---|---|
| Min | 1Q | Median | 3Q | Max |
| -3.9876 | -0.7853 | 0.00 | 0.8103 | 1.5511 |
| **Null deviance:** 1801.9 of 1299 degrees of freedom | | | | |
| **Residual deviance:** 1140.91212.9 of 1264 degrees of freedom | | | | |
| **AIC: 1659.9** | | | | |

Two types of deviance performed are null deviance and residual deviance. The null deviance distributes the prediction of HPV infection by the model, which only includes the intercept without the predictors. The residual deviance distributes how well the prediction of HPV is when the predictors are included. The residual deviance (1140.91212.9 of 1264 degrees of freedom) is small for the model; thus, the model's fitness is acceptable. Moreover, the Akaike Information Criterion (AIC), model fitness and accuracy value, is 1140.9. Through the diagnostics of the model, the model fitness is good, and our results meet the requirements.

Table 7 R2 statistic values of each variable

| | Age | Number of sexual partners | First sex | Number of pregnancies | Smoke | Smoke year | Contraceptives | Contraceptives year |
|---|---|---|---|---|---|---|---|---|
| R2 statistic | 0.059 | 0.093 | 0.12 | 0.12 | 0.12 | 0.13 | 0.13 | 0.13 |

## 4. Discussion

Our study is distinct from others' works because we use data on cervical cancer to clarify the joint effects of several unexpected behaviors for HPV infection, such as smoking and contraceptives use. If these behaviors' effectiveness is significant, the public can use the model to predict the HPV infection for individuals and plan for future treatments if one is infected; otherwise, the public can use this result to confirm their research direction. As Chabeda's research, other research mainly focuses on demonstrating the causality between the infection of HPV and having cervical cancer [4] but neglect the importance of educating the public about the essential issues, unhealthy behaviors that lead to disease. In contrast, our study is more suitable for developing countries where the low vaccine and screening rates. It is a feasible way to reduce the infection rate of HPV by analyzing the risk factors and establishing the research direction.

Compared with the initial evaluation, the sensitivity of testing data after optimization has increased from 0.23 to 0.5, and the accuracy also increases. It shows that balancing the dataset is useful. However, the accuracy is 0.5877, which is not high enough to help with estimation. Based on the accuracy of our logistic model, we conclude that we cannot predict HPV infection depending on the joint effects of these unexpected behaviors. This conclusion makes sense because these behaviors are not the major factors of HPV infection, and they can only be considered as the cofactors instead of essential conditions.

In Hu and Ma's research, they indicate the persistence infection is the weak immune system of the host, and a normal immune system can eliminate the virus automatically in one year [6]. Moreover, as Kovachev states in the research, host immunity and HPV infection are mutually dependent [12]. Thus, we can consider the weak immune system cannot ward off the virus as usual, and the unexpected cofactors accelerate the infection and even exacerbate the disease. Then, we can conclude that the rate

of infection among the population with these unexpected behaviors was not significantly higher than that population without these unexpected behaviors.

The deficiency of infected cases is a limitation of our study. In oversampling, "new" data is replicated from the minority category of data. Although it increases the amount of data, it does not provide any new information or changes to the machine learning model [13]. As a result, although oversampling was used to narrow the gap between different categories in the dependent variable, it may increase the likelihood of occurring overfitting for the regression model.

Another limitation occurs when selecting appropriate variables to examine - a few important risk factors of cervical cancer were not contained. For instance, genetic inheritance is a vital circumstance that can cause cervical cancer. Still, it was not counted in our dataset due to the difficulty of testing the genes of so many patients and collecting them. Hence, we acknowledged this limitation and focused on other notable, assessable factors.

## 5. Conclusion

Our study suggests that the behaviors commonly viewed as risk factors of HPV infection are insignificant for HPV infection estimation when combined. It may be explained by the limitation of the raw dataset since it did not mention the methodology used for selecting participants. Our study states that these risk factors have little effect on HPV infection when they are combined. We provide a more specific research direction for other researchers – future research can be done by considering the multiplicative effects of either two risk factors of HPV infection. Moreover, this research provides more information about risk factors instead of therapies. Those non-peer group readers can get insights into those investigated risk factors that may not be significant factors of HPV and the importance of screening for females. However, the population group investigated and collected in this dataset may cause bias because only patients with cervical cancer were surveyed in the hospital. We need to include more population groups such as different age groups, different workgroups (with different pressure), and groups in different living environments. Thus, the bias can be reduced, and the research will be more specific to each population group and more meaningful.

## References


[1] National Cancer Institute. (2015). What Is Cancer? Retrieved from National Cancer Institute website: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[2] *Human papillomavirus (HPV) and cervical cancer*. (2020, November 11). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer

[3] Hu, Z., & Ma, D. (2018). The precision prevention and therapy of hpv-related cervical cancer: New concepts and clinical implications. *Cancer Medicine, 7*(10), 5217-5236. doi:10.1002/cam4.1501

[4] Chabeda, A., Yanez, R. J., Lamprecht, R., Meyers, A. E., Rybicki, E. P., & Hitzeroth, I. I. (2018). Therapeutic vaccines for high-risk HPV-associated diseases. *Papillomavirus Research*, *5*, 46–58. https://doi.org/10.1016/j.pvr.2017.12.006

[5] Parkin, D. M., & Bray, F. (2006). Chapter 2: The burden of HPV-related cancers. *Vaccine*, *24*, S11–S25. https://doi.org/10.1016/j.vaccine.2006.05.111

[6] McDermott, J. D., & Bowles, D. W. (2019). Epidemiology of head and neck squamous cell carcinomas: Impact on staging and prevention strategies. *Current Treatment Options in Oncology, 20*(5). doi:10.1007/s11864-019-0650-5

[7] De Villiers, E. (2003). Relationship between steroid hormone contraceptives and hpv, cervical intraepithelial neoplasia and cervical carcinoma. *International Journal of Cancer, 103*(6), 705-708. doi:10.1002/ijc.10868

[8] Kahn, J. A., Rosenthal, S. L., Succop, P. A., Ho, G. Y., & Burk, R. D. (2002). Mediators of the association between age of first sexual intercourse and subsequent human papillomavirus infection. *PEDIATRICS, 109*(1). doi:10.1542/peds.109.1.e5

[9] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: About. Retrieved from archive.ics.uci.edu website: https://archive.ics.uci.edu/ml/about.html

[10] Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017). Transfer learning with partial observability applied to cervical cancer screening. Pattern Recognition and Image Analysis, 243–250. https://doi.org/10.1007/978-3-319-58838-4_27

[11] Rosencrance, L. (2019, May). What is logistic regression? - Definition from WhatIs.com. Retrieved from SearchBusinessAnalytics website: https://searchbusinessanalytics.techtarget.com/definition/logistic-regressio

[12] Kovachev, S. M. (2021). A review On INOSINE Pranobex immunotherapy for Cervical HPV-Positive Patients. Infection and Drug Resistance, Volume 14, 2039-2049. doi:10.2147/idr.s296709.

[13] Wijaya, C. Y. (2021, May 24). 5 SMOTE techniques for oversampling your imbalance data. Medium. https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5