RESEARCH PROJECT

IDENTIFYING CERVICAL CANCER RISK FACTORS IN R MARKDOWN

Mai Ha

Oct 16th 2020

Abstract

This project is an attempt to identify the cause(s) of cervical cancer. Many variables (factors) were recorded and most are inconclusive. Three different statistical models (linear and quadratic discriminant analysis, and logistic regression analysis) were compared to each other and the most accurate model was used to identify the cause(s) for cervical. The chosen model suggested that the most predictive factors for cervical cancer are HPV (positive) and the length of smoking (in years, with a positive correlation).

Contents

PROJECT DESCRIPTION

The project was written in the statistical programming language R version 4.0.2 in R Studio using MacTeX to write mathematical formulas in LaTeX, subsequently produced into this format using R Markdown.

```
knitr::opts_chunk$set(echo=TRUE)
options(citr.use_betterbiblatex = FALSE)
#add bibliography and the ability to make citations to the rest of the document
```

BACKGROUND INFORMATION

What is Cervical Cancer?

An abnormal growth of cells in the cervix (the lowest part of the uterus connecting to the vagina) is referred to as cervix epidermoid carcinoma, or more commonly known as cervical cancer. There are two main types of cervical cancer, squamous cell carcinoma and adenocarcinoma. These names refer to the location and types of cell growth. Squamous cell carcinoma originates in the squamous cells lining the outer part of the cervix while adenocarcinoma forms in the column-shaped glandular cells living the cervical canal (Clinic 2018).

Effects of Cervical Cancer

Cervical cancer is the most frequent neoplasia among women living in the "so-called" third world countries; it is also one of the leading cause of cancer mortality among women in developing countries (Garza-Salazar, Meneses-Garcia, and Morales-Vasquez 2017).

According to the World Health Organization, cervical cancer is the fourth most frequent cancer in women, with an estimate of 530,000 new cases in 2012 (WHO n.d.). Also in 2012, approximately 270,000 women died from cervical cancer, of which more than 85% of the deaths occurred in low- and middle- income countries (WHO 2016). In the Region of the Americas, cervical cancer claimed 83,000 diagnoses and killed almost 36,000 women in 2012 (PAHO n.d.). In 2014, cervical cancer is the eighth most common cancer the United States (Tao et al. 2014).

POSSIBLE FACTORS IN RESEARCH

Cervical Cancer and Age

The age group that were most by cervical cancer were from 50-59 and 30-49 years old in that order (Reynoso-Noveron et al. 2017).

A study on cervical neoplasia risk factors revealed that 46-55 year-old women are specifically at higher risk for high-grade squamous intraepithelial lesions (HSIL). These lesions referred to the extensive changes in squamous cells of the cervix, which serves as an indicator for progression to cervical carcinoma in at least 25% of cases (Tao et al. 2014).

Another study on the relation of age and cervical cancer showed that cervical cancer progression and survival outcomes were age independent in women that are 35 years of age or less (Pelkofski et al. 2016). The study also mentioned that results regarding age and cervical cancer are conflicting and concluded that further study of a larger cohort of women could yield different outcomes.

Cervical Cancer and Sexual Activity

A study revealed that women who had their first sexual intercourse at an age less than 18 has a significant odd ratio of 5.44 (Sharma and Pattanshetty 2017). This meant if a woman had her first intercourse at 18 years old and younger, she was 5.44 times more likely to get cervical cancer than someone who had their first intercourse at or above age 18.

Cervical Cancer and Smoking

A very large study was done on the EPIC (European Prospective Investigation into Cancer and Nutrition) cohort to study the risk factors of cervical cancer. This study was conducted on over 450,000 qualified participants, with over 300,000 women and 150,000 men. Participants were mostly between the age of 35 and 70, with no prevalent cancer or pre-cancer. In one particular paper from this study, the authors established that duration and intensity of smoking tobacco increased the risk of cervical cancer, conversely, the time lapse from quitting has a negative correlation with cervical cancer risk (Roura et al. 2013).

Cervical Cancer, (Hormonal) Contraceptives and Pregnancy

The same prospective study of the EPIC cohort above also investigated the associations between hormonal contraceptives and risk of developing cervical neoplasia or carcinoma in another paper. This paper suggested that long-term use of several types of hormonal contraceptives are risk factors for the generation of cervical cancer and pre-cancer cells. The paper further concluded a positive association between the number of full-term pregnancies and cervical cancer. In general, increasing number of full-term pregnancies and duration of oral contraceptives use would increase the risk of cervical cancer (Roura et al. n.d.).

Cervical Cancer and HIV/AIDS

A cross-sectional study in Brazil of 494 HIV-infected women concluded that along with presence of HPV infection and younger age, the severity of immunosuppression induced by HIV infection was a strong predictor of cervical intra-epithelial neoplasia (Teixeira et al. 2012). In general, it could take 15 to 20 years for cervical cancer to develop in women with normal immune systems, while it could take only 5 to 10 years in women with weakened immune systems (having untreated HIV infection) to develop this cancer (WHO 2016)

Cervical Cancer and HPV

HPV stands for Human Papilloma Virus. A study claims that one out of every ten women that acquired the HPV infection will develop cervical cancer (Reynoso-Noveron et al. 2017). Certain factors that were formerly believed to be associated with an increased risk of developing cervical cancer are now also known to be risk factors for HPV infection (Reynoso-Noveron et al. 2017). These risks are as followed:

- Tobacco consumption
- Sexually Transmitted Diseases or STD's (specifically herpes and chlamydia)
- Use of oral hormonal contraceptive
- Age of onset of sexual activity with the absence of protection
- High-risk sexual behavior throughout lifetime (i.e. multiple sexual partners)

The World Health Organization reported at least 13 of the 100 types of HPV were cancer-causing, of which the types HPV 16 and 18 caused up to 70% of cervical cancer cases and precancerous cervical lesions. Although most HPV infections could clear up on their own, there was a risk for all women that HPV infection became chronic and precancerous lesions progressed into invasive cervical cancer (WHO 2016).

STATISTICAL MODELS INFORMATION

Model 1: Linear Discrimination Analysis

Discrimination analysis is a multivariate technique concerned with separating distinct sets of objects and allocating new objects to previously defined groups based on a set of features, $x_1, x_2, ..., x_p$.

Let π_1 and π_2 be two multivariate populations, and let $f_1(x)$ and $f_2(x)$ be the density functions associated with the random vector x for the density functions associated with the random vector x for the two populations, respectively.

The density functions are normally distributed with mean μ_i and covariance matrix, $\sum_i$ for $i = 1, 2$. If two populations have equal covariance matrices, $\sum_1 = \sum_2 = \sum$, then we use the linear discrimination rule (Johnson and Wichern 2007).

Linear Discrimination Rule

By the linear classification rule, an object x_0 is classified into π_1 if

$$(\mu_1 - \mu_2)' \sum^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)' \sum^{-1} (\mu_1 + \mu_2) \geq 0$$

and it is classified to π_2 otherwise (Bagui, Bagui, and Hemasinha 2016).

Model 2: Quadratic Discrimination Analysis

Quadratic Discrimination Rule

The quadratic classification rule is used when two groups have unequal covariance, $\sum_1 \neq \sum_2$; an object x_0 is classified into π_1 if

$$-\frac{1}{2} x_0' (\sum_1^{-1} - \sum_2^{-1}) x_0 + (\mu_1' \sum_2^{-1}) x_0 - k \geq 0$$

, where $k = \frac{1}{2} ln(\frac{|\sum_1|}{|\sum_2|}) + \frac{1}{2}(\mu_1 \sum_1^{-1} \mu_1 - \mu_2' \sum_2^{-1} \mu_2)$.

If $\mu_1, \mu_2, \sum_1$ and $\sum_2$ are unknown, then they may be replaced by their corresponding unbiased sample estimates, $\bar{X}, \bar{Y}, S_1$ and S_2, respectively (Bagui, Bagui, and Hemasinha 2016).

Model 3: Logistic Regression Analysis

Logistic regression is an approach to classification where some or all of the variables are qualitative (Johnson and Wichern 2007). The response variable Y is dichotomous, which means it is restricted to two values, usually assigned 0 and 1, where 1 is "success" and 0 is otherwise.

Let $\pi(x)$ be the probability of getting a "success". The odds ratio is equal to

$$odds = \frac{\pi(x)}{1 - \pi(x)}$$

and the logit is defined as the natural log of the odds. Then, linear regression is applied using the logit as the response instead of the response variable of interest. This is due to the possibility that the range of the logit may assume values between $-\infty$ to ∞ instead of just between 0 to 1 as in the case of the response variable (Bagui, Bagui, and Hemasinha 2016). Thus, the regression function is

$$ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Solving for $\pi(x)$ using the function

$$\frac{e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}{1-e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}$$

will then help estimate the coefficients $\beta_0, \beta_1, ..., \beta_p$.

Thus, for a response variable with several predictor variables, the Logistic Regression Analysis creates a linear model based on the natural log of the odds ratio rather than the mean of the variables (Johnson and Wichern 2007).

Logistic Regression Classification Rule

An object x is assigned to π_1 if the estimated odds is greater than 1.

$$\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)} = e^{\hat{\beta}_0+\hat{\beta}_1 x_1+...+\hat{\beta}_p x_p} > 1$$

Object x is assigned to π_1 if the logit is greater than 0.

$$ln\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p > 0$$

(Bagui, Bagui, and Hemasinha 2016).

Model Selection using Miscalculation Rate and AIC (Stepwise-function)

The performance of the discriminant function can be evaluated by applying the rules to the data and then calculating the misclassification rate. (Johnson and Wichern 2007). Additionally, the accuracy rate is also calculated (opposite of misclassification) to aid the selection of model. Both calculations of the misclassification and accuracy rates will be applied on all three models (where appropriate) to optimize model selection.

The Akaike Information Criterion (AIC) is a measure of evaluating statistical models for a given dataset. The best model for a particular dataset is one with the smallest AIC value (Kimura and Waki 2017). This is the misclassification rate as mentioned above. The AIC is defined by

AIC = (-2) log(maximum likelihood) + 2*(number of independently adjusted parameters within the model)

(Akaike 1974).

Since there are many candidates for the best model, the stepwise function is also aplied to determine the best model out of the three models by applying local search algorithms to find the models with the smallest AIC value possible (Kimura and Waki 2017). This method is applicable to the Logistic Regression Model only.

IMPLEMENTATION AND RESULTS

Description of Dataset

The dataset was made available in the University of California in Irvine Machine Learning Repository. Data was collected at the 'Hospital Universitario de Caracas' in Caracas, Venezuela and was donated to the university repository in 2017. The dataset comprised of demographic information, habits, and historic medical records of 858 patients accross 36 attributes. There were missing values due to several patients decided not to answer some of the questions because of privacy concerns (UCI 2017). The data set can be found using the link below

https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

13 variables chosen due to lack of background research on the omitted attributes. Many of the attributes that were disregarded were specific types of STD's (besides HPV and HIV), and IUD (yes/no, years). For a complete list of all attributes, please visit the link above.

The chosen variables and their respective variable names in the dataset are:

- Age (age)
- Number of sexual partners (nosxp)
- First sexual intercourse (firstsx): the age at which each patient has their first sexual intercourse
- Number of pregnancies (npreg)
- Smokes (smokes): whether each patient smokes
- Smokes years (smokesyr): how many years each patient smokes
- Smokes packs (smokespk): how many pack each patient smoke per year
- Hormonal contraceptive (hcntr): whether each patient takes hormonal contraceptives
- Hormonal contraceptive years (hcntryr): how many years each patient takes hormonal contraceptives
- STD-HIV (stdhiv): whether each patient reports to have HIV
- STD-HPV (stdhpv): whether each patient reports to have HPV
- Cancer (cancer): whether each patient is diagnosed with cervical cancer
- HPV (hpv): whether each patient is diagnosed/confirmed to have HPV

  The dataset is splitted into two smaller datasets for use with the appropriate modeling methods. These datasets are named "ccdata1"" and "ccdata2". "ccdata1" is the set of all 13 variables mentioned above, and "ccdata2" is the set of the variables that are continuous variables (for discriminant analyses). The variables included in "ccdata2" are "age", "nosxp", "firstsx", "npreg", "smokesyr", "smokespk", "hcntryr", and "cancer".

```r
#Import datasets into R below
ccdata1 = na.omit(read.csv("risk_factors_cervical_cancer.csv",header=TRUE))
#ccdata1 is the set of chosen variables
ccdata2 = na.omit(read.csv("risk_factors_cervical_cancer1.csv",header=TRUE))
#ccdata2 is the set of all variables
library(MASS);library(RcmdrMisc) #necessary library for the models
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: sandwich
```

Results

The following analyses are done using R Version 4.0.2, released on 2020-06-22. These analyses are partly based on a tutorial from the University of Cincinnati's "An Introduction to Statistical Learning" (UC 2018) and a paper called "The Statistical Classification of Breast Cancer Data" (Bagui, Bagui, and Hemasinha 2016).

Method 1: Linear Discriminant Analysis (LDA)

The first model (named "model1") deals with linear discriminant analysis using "ccdata2". The probability table is as follows:

```r
model1 = lda(formula=cancer ~ ., data = ccdata2) #linear discriminant analysis
pred.lda = predict(model1, data = ccdata2)
table1 = table(ccdata2$cancer, pred.lda$class)  #number of prediction in each category
round(prop.table(table1),3)  #probability table model1
```

```
##
##          0     1
##   0 0.970 0.004
##   1 0.024 0.001
```

This table shows that 2.4% of the population are incorrectly categorized into the non-cancer group when they indeed have cancer, while 0.4% of the population is predicted to be in the cancer group when they actually do not have cervical cancer. Thus the misclassification rate is the sum of these two errors, which is as follows:

```r
round(mean(ccdata2$cancer != pred.lda$class),3)  #overall error rate for model1
```

```
## [1] 0.028
```

Method 2: Quadratic Discriminant Analysis (QDA)

The second model (named "model2") deals with quadratic discriminant analysis using "ccdata2". The probability table is as follows:

```
model2 = qda(formula=cancer ~ ., data=ccdata2) #quadratic discriminant analysis
pred.qda = predict(model2, data = ccdata2)
table2 = table(ccdata2$cancer, pred.qda$class)
round(prop.table(table2),3)  #probability table model2
```

```
##
##          0     1
##   0 0.959 0.016
##   1 0.024 0.001
```

This table shows that 2.4% of the population are incorrectly categorized into the non-cancer group when they indeed have cancer, while 1.6% of the population is predicted to be in the cancer group when they actually do not have cervical cancer. Thus the misclassification rate is the sum of these two errors, which is as follows:

```
round(mean(ccdata2$cancer != pred.qda$class),3)  #overall error rate model2
```

```
## [1] 0.04
```

Method 3: Logistic Regression Analysis (LRA)

This third model (named "model3") deals with logistic regression on the "ccdata2". The same dataset is used so the comparison between the three models is controlled. The probability table for this model is as follows:

```
model3 = glm(cancer ~ ., data=ccdata2, family = "binomial" (link="logit")) #logistic reg
pred.glm = predict(model3, ccdata2, type="response")
table3 = table(ccdata2$cancer, ifelse(pred.glm > 0.5, "1", "0")) #prediction table mode
round(prop.table(table3),3)
```

```
##
##          0     1
##   0 0.975 0.000
##   1 0.024 0.001
```

This table shows that 2.4% of the population are incorrectly categorized into the non-cancer group when they indeed have cancer, while none of the population is predicted to be in the cancer group when they actually do not have cervical cancer. Thus the mis-classification rate is just the error rate 2.4%.

Which model is the best?

The following list of output represent the three accuracy rates of "model1" (using LDA), "model2" (using QDA), and "model3" (using LRA) respectively:

```
round(mean(ccdata2$cancer == pred.lda$class),3)  #Accuracy rate for model1
```

```
## [1] 0.972
```

```
round(mean(ccdata2$cancer == pred.qda$class),3)  #Accuracy rate model2
```

```
## [1] 0.96
```

```
round(mean(ifelse(pred.glm > 0.5, "1", "0") == ccdata2$cancer),3) #Accuracy rate model3
```

```
## [1] 0.976
```

Off all the three models, logistic regression performs the best with 97.6% accuracy. Thus, further analyses on the full dataset "ccdata1" are done using logistic regression analysis.

Further Analyses

For the first full analysis, logistic regression is used on the full dataset "ccdata1" by regressing cancer against all other variables in a new model (named "model4"). The probability table is as follows:

```
model4 = glm(cancer ~ ., data=ccdata1, family="binomial" (link="logit"))
pred.glm1 = predict(model4, ccdata1, type="response")
table4 = table(ccdata1$cancer, ifelse(pred.glm1 > 0.5, "1", "0")) #Prediction table mod
round(prop.table(table4), 3)
```

```
##
##         0     1
##   0 0.975 0.000
##   1 0.003 0.022
```

This table shows that 0.3% of the population are incorrectly categorized into the non-cancer group when they indeed have cancer, while none of the population is predicted to be in the cancer group when they do not have cervical cancer. Thus the mis-classification rate is the same as the error rate, which is 0.3%.

Using stepwise analysis, the chosen model and its corresponding AIC are as follows:

```
choose$formula
```

```
## cancer ~ hpv
```

```
choose$aic
```

```
## [1] 38.62892
```

Creating a model (named "model5") using this suggestion yields this following probability table:

```
model5 = glm(cancer ~ hpv, ccdata1, family="binomial" (link="logit"))
pred.glm2 = predict(model5, ccdata1, type="response")
table5 = table(ccdata1$cancer, ifelse(pred.glm2 > 0.5, "1", "0"))
round(prop.table(table5), 3)
```

```
##
##         0     1
##   0 0.973 0.001
##   1 0.003 0.022
```

This table shows that 0.3% of the population are incorrectly categorized into the non-cancer group when they indeed have cancer, while 0.1% of the population is predicted to be in the cancer group when they, in fact, do not have cervical cancer. Thus the mis-classification rate is the sum of these two error rates, which is as follows:

```
round(mean(ifelse(pred.glm2 > 0.5, "1", "0") != ccdata1$cancer),3) #error rate model5
```

```
## [1] 0.004
```

Based on suggested models from the stepwise output, a few more variables were added to create a "model6" to acquire the same low mis-classification rate as in the model of all variables ("model4"). Its probability table is as follows:

```
model6 = glm(cancer ~ hpv + smokes + smokesyr, ccdata1, family="binomial" (link="logit")
pred.glm3 = predict(model6, ccdata1, type="response")
table6 = table(ccdata1$cancer, ifelse(pred.glm3 > 0.5, "1", "0"))
round(prop.table(table6), 3)
```

```
##
##         0     1
##   0 0.975 0.000
##   1 0.003 0.022
```

This table shows that the error rates of the new model ("model6") is the same as in the model of all variables ("model4"), which has the mis-classification rate as follows:

```
round(mean(ifelse(pred.glm3 > 0.5, "1", "0") != ccdata1$cancer),3)
```

```
## [1] 0.003
```

The following accuracy rates are of "model4", "model5", and "model6" respectively.

```
round(mean(ifelse(pred.glm1 > 0.5, "1", "0") == ccdata1$cancer),3) #accuracy rate model
```

```
## [1] 0.997
```

```
round(mean(ifelse(pred.glm2 > 0.5, "1", "0") == ccdata1$cancer),3) #accuracy rate model
```

```
## [1] 0.996
```

```
round(mean(ifelse(pred.glm3 > 0.5, "1", "0") == ccdata1$cancer),3) #accuracy rate model
```

```
## [1] 0.997
```

These accuracy rates show that model4 and model6 have the same strengths even though model6 has a significantly smaller set of variables.

CONCLUSION AND DISCUSSION

Out of the three methods of classification, the logistic regression model has the lowest overall error rate and highest accuracy levels.

Using the logistic regression model on the full dataset yields a higher accuracy levels than the reduced dataset of just the continuous factors (99.7% versus 97.6%). However, the stepwise selection function reveals that "hpv" (whether a patient is positive for HPV) alone can predict cervical cancer with up to 99.6% accuracy. Additionally, adding the variables "smokes" (whether a patient smokes) and "smokesyr" (how many years a patient smokes) slightly increases the accuracy to 99.7%, which is the same accuracy as the model using all of the factors for prediction.

As expected from the background research, HPV is the most prevalent factor in predicting cervical cancer. The next two useful predictive factors are (1) whether a patient smokes and (2) the time length that each patient smokes (if any). This is also expected given tobacco consumption is one of the factors that increases the risk of getting HPV. Thus, the three variables "hpv", "smokes", and "smokesyr" can predict cervical cancer with the highest accuracy.

Fortunately, approximately 70% of cervical cancer cases could be avoided through HPV vaccination (PAHO n.d.). Vaccine against HPV (specifically HPV 16 and 18) have been approved for use in many countries (WHO 2016). Furthermore, smoking is a factor that can be controlled, eliminated, and avoided through medication, lifestyle changes, and raising awareness.

REFERENCES

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions of Automatic Control* 19 (6): 716–23.

Bagui, Subhash, Sikha Bagui, and Rohan Hemasinha. 2016. "The Statistical Classification of Breast Cancer Data." *International Journal of Statistics and Application 2016* 6 (1): 15–22. https://doi.org/10.5923/j.statistics.20160601.03.

Clinic, Mayo. 2018. "Cervical Cancer Overview." Mayo Clinic. 2018. https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501.

Garza-Salazar, Jaime, Abelardo Meneses-Garcia, and Flavia Morales-Vasquez. 2017. *Cervical Cancer*. Switzerland: Springer.

Johnson, Richard, and Dean Wichern. 2007. "Discrimination and Classification." In *Applied Multivariate Statistical Analysis*, edited by Joanne Wendelken, Sixth Edition, 584–644. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Kimura, Keiji, and Hayato Waki. 2017. "Minimization of Akaike's Information Criterion in Linear Regression Analysis via Mixed Integer Nonlinear Program." *Optimization Methods and Software* 33 (3): 633–49. https://doi.org/https://doi.org/10.1080/10556788.2017.1333611.

PAHO. n.d. "Cervical Cancer." Pan American Health Organization. Accessed April 18, 2018. http://www.paho.org/hq/index.php?option=com_topics&view=article&id=348&Itemid=40936&lang=en.

Pelkofski, Elizabeth, Jessica Stine, Nolan Wages, Paola Gehrig, Kenneth Kim, and Leigh Cantrell. 2016. "Cervical Cancer in Women Aged 35 Years and Younger." *Clinical Therapeutics* 38 (3): 459–66.

Reynoso-Noveron, Nancy, Adriana Pena-Nieves, Maryori Rodriguez, and Alejandro Mohar-Betancourt. 2017. "Cervical Cancer Epidemiology." In *Cervical Cancer*. Switzerland: Springer.

Roura, Esther, Noemie Travier, Tim Waterboer, Silvia de Sanjose, F. Xavier Bosch, Michael Pawlita, Valeria Pala, et al. 2013. "Smoking as a Major Risk Factor for Cervical Cancer and Pre-Cancer: Results from the EPIC Cohort." *International Journal of Cancer* 135 (2). https://doi.org/https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.28666.

———. n.d. "The Influence of Hormonal Factors on the Risk of Developing Cervical Cancer and Pre-Cancer: Results from the EPIC Cohort." *PLOS ONE*. Accessed April 20, 2018. https://doi.org/https://doi.org/10.1371/journal.pone.0147029.s003.

Sharma, Pragati, and Sanjay Pattanshetty. 2017. "A Study on Risk Factors of Cervical Cancer Among Patients Attending a Tertiary Care Hospital: A Case-Control Study." *Clinical Epidemiology and Global Health*, October. http://www.sciencedirect.com/science/article/pii/S2213398417300702.

Tao, Lixin, Lili Han, Xia Li, Qi Gao, Lei Pan, Lijuan Wu, Yanxia Luo, Wei Wang, Zihe Zheng, and Xiuhua Guo. 2014. "Prevalence and Risk Factors for Cervical Neoplasia: A

Cervical Cancer Screening Program in Beijing." *BioMed Central Public Health*, November. https://doi.org/https://doi.org/10.1186/1471-2458-14-1185.

Teixeira, Nara Chartuni Pereira, Angela Cristina Labanca Araujo, Christine Miranda Correa, Claudia Teixeira da Costa Lodi, Maria Ines Miranda Lima, Nara de Oliveira Carvalho, Dora Mendez del Castillo, and Victor Hugo Melo. 2012. "Prevalence and Risk Factors for Cervical Intraepithelial Neoplasia Among HIV-Infected Women." *Brazillian Journal of Infectious Diseases* 16 (2): 164–69. https://doi.org/https://doi.org/10.1016/S1413-8670(12)70299-4.

UC. 2018. "Linear and Quadratic Discriminant Analysis." UC Business Analytics R Programming Guide. April 2018. http://uc-r.github.io/discriminant_analysis#eval.

UCI. 2017. "Cervical Cancer (Risk Factors) Data Set." UCI Marchine Learning Repository. 2017. https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29.

WHO. 2016. "Human Papillomavirus (HPV) and Cervical Cancer." Fact sheet. World Health Organization. June 2016. http://www.who.int/mediacentre/factsheets/fs380/en/.

———. n.d. "Cervical Cancer." World Health Organization. Accessed April 16, 2018. http://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/.