Investigating cervical cancer risk factors

Proposal for analysis

Kristie Kooken

Preliminary Analysis

Cervical cancer is a type of cancer that starts in the cervix. The cervix connects the vagina (birth cannel) to the upper part of the uterus (womb). Anyone with a cervix is at risk for cervical cancer (www.mayoclinic.org). In the United States, there are about 11,500 cases diagnosed with 4,000 women dying from this type of cancer each year. Stated another way, for every 100,000 women, there are 7 new Cervical cancer cases reported, and 2 women will die of this cancer (www.cdc.gov). Two characteristics of this cancer are it most often occurs in women over 30 years old as well the main cause for the cancer is long lasting infection with certain types of human papillomavirus (HPV). It would be possible to prevent about 93% of cervical cancer cases by screening and HPV (human papillomavirus) vaccination (www.cdc.gov).

The goal of this project is to develop a model to predict which method of screening has the best performance to detect who would have cervical cancer as well as HPV and use the results of this model to develop a groundbreaking intervention that would reduce the burden of this cancer. This intervention would focus on statistically meaningful characteristics of predictors of cervical cancer. This is important because lives can be saved by an affordable intervention which would deliver the best medical practices without the burden of treatment for cervical cancer and allow the patient to better manage their health and wellbeing. Such a low-cost intervention would be highly desirable to all types of health insurance carriers as it can lower costs associated with cancer care and treatment for these companies.

The data that will support this model development was downloaded from the UC Irvine Machine Learning Repository and contains cervical cancer risk factors (Fernandes, K., *et. al*., 2017). The data focuses on the prediction of indicators/diagnosis of cervical cancer. Features on

this dataset include demographics, habits, and historic medical records. This data, collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela, has 36 columns in total and 858 rows of data. In the initial dataset, several patients decided not to answer some questions due to privacy concerns thus there are missing values for these patients. The dataset is useful for this project because of robust demographic and previous health information that has been collected in addition to habits around sexuality. Because cervical cancer's strong relationship to HPV, it is critical to understand the individual's sexuality and sexual habits. Without data of this nature, researchers are left in the dark without understanding the social component to the causes of this cancer and not being able to access information on habits that could put an individual at risk.

These data have both a cancer diagnosis as well as a HPV diagnosis, and thus I will develop two logistic regression models to determine if the demographic, medical history as well as habits can predict either outcome. The modeling technique will be logistic regression as the targets are discrete class labels. To check the performance of each model, accuracy scores, confusion matrixes, F1 scores and ROC curves will be generated. These will be compared across the different models to determine if the features in the datasets predict either cervical cancer or HPV well and the strength of that model. Likewise, there are 4 diagnostic tests on these data for cervical cancer and I am interested in exploring how these tests perform with a cancer diagnosis. I may run additional statistics to assess this, but I am still considering how I might better understand the diagnostic tests and their ability to determine if a patient has cancer. Likewise, I would like to include an interpretation on the coefficients or key coefficients. I think this will shed light on which features have a lot of influence on the outcome of cervical cancer or HPV.

Using the results from these models, I am hoping to learn which factors help predict cervical cancer as well as HPV. I believe that once these features can be better known, and assuming the resulting model has satisfactory performance then an intervention could be developed that is not a medical test. This would be ideal as a way to reach a broader audience and, if in fact this type of intervention could be developed, it would save lives and money in every country in the world.

Ethical concerns for this project are important to address. Because there are many different questions regarding one's sexuality from the number of partners to the sexually transmitted diseases, it is particularly important to safeguard the confidential nature of these responses. In the medical field, it is critical that patients have a trusting relationship with their medical provider for that provider to give the best and most informed medical care possible. However, in this analysis, neither the subject nor the medical provider has any say in the analysis, how the results are used or what conclusions are made. Thus, it is critical that whoever performs the analysis of this data have the same integrity that any medical provider would have. Likewise, though there are few demographics in this dataset, it is especially important that the subjects who participated for this dataset do not have their privacy violated. In the United States, I do not think we would be able to know the hospital location according to our HIPAA Privacy Rule. This provides federal protections for personal health information held by covered entities and gives patients an array of rights with respect to that information (www.hhs.gov). Data privacy is extremely critical and will only continue to be more so in the future. Without trust or the ability to be anonymous, people will not give important health information that could be critical to solving different health concerns, diseases, and cancers. This would be very

unfortunate; medical providers as well as data science folks need to work to ensure privacy is always kept.

I have discussed the dataset I will be analyzing, the method of analysis, what I hope to learn and ethical concerns. However, what will I do if this plan does not work out? If my analysis does not yield a good model, I will investigate using another type of model such as tree-based methods which can be a suitable alternative for assessing risk factors (www.datasciencecentral.com). If this second method does not work out, then I would have to assume that the features collected are not good predictors for cervical cancer or HPV. However, I am hopeful that the current dataset has robust data and will be able to shed light on factors that can predict cervical cancer or HPV.

References

Cervical Cancer is Preventable (Accessed 2023, June18). Centers for Disease Control and Prevention. https://www.cdc.gov/vitalsigns/cervical-cancer/index.html

Cancer Statistics At a Glance (Accessed 2023, June18). Centers for Disease Control and Prevention. https://gis.cdc.gov/Cancer/USCS/#/AtAGlance/

Cervical cancer (Accessed 2022, May 22). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501

Fernandes, Kelwin, Cardoso, Jaime, and Fernandes, Jessica. (2017). *Cervical cancer (Risk Factors)*. UCI Machine Learning Repository. https://doi.org/10.24432/C5Z310.

Glen, S. (2019). *Alternatives to Logistic Regression.* Data Science Central. https://www.datasciencecentral.com/alternatives-to-logistic-regression/#:~:text=For%20example%2C%20tree-based%20methods%20are%20a%20good%20alternative,work%20well%20for%20propensity%20score%20estimation%20and%20Categorization%2FClassification.

What is PHI? (Accessed 2023, June 18). U.S. Department of Health and Human Services. https://www.hhs.gov/answers/hipaa/what-is-phi/index.html