[Reivew]Prediction of Cervical Cancer Risk from Taking Hormone Contraceptivese

Su jeong RU[1], Kyung-A KIM[2], Myung-Ae CHUNG[3], Min Soo KANG[5]

Abstract

In this study, research was conducted to predict the probability of cervical cancer occurrence associated with the use of hormonal contraceptives. Cervical cancer is influenced by various environmental factors; however, the human papillomavirus (HPV) is detected in 99% of cases, making it the primary attributed cause. Additionally, although cervical cancer ranks 10th in overall female cancer incidence, it is nearly 100% preventable among known cancers. Early-stage cervical cancer typically presents no symptoms but can be detected early through regular screening. Therefore, routine tests, including cytology, should be conducted annually, as early detection significantly improves the chances of successful treatment. Thus, we employed artificial intelligence technology to forecast the likelihood of developing cervical cancer. We utilized the logistic regression algorithm, a predictive model, through Microsoft Azure. The classification model yielded an accuracy of 80.8%, a precision of 80.2%, a recall rate of 99.0%, and an F1 score of 88.6%. These results indicate that the use of hormonal contraceptives is associated with an increased risk of cervical cancer. Further development of the artificial intelligence program, as studied here, holds promise for reducing mortality rates attributable to cervical cancer.

1. Introduction

Cancer is the number one cause of death in Korea. In 2019, the total number of cancer cases in women was 120,538 of which a total of 3,273 were cervical cancer cases. It ranked 10th in the cancer incidence ranking with 2.7% of all cancers that women get (Seoul Asan Medical Center,

2022). Every year, more than 570,000 women are diagnosed with cervical cancer worldwide, and about 310,000 dies. This is because one woman dies of cervical cancer every two minutes, and cervical cancer is one of the most threatening diseases to women's health. There are so many cases of cervical cancer and death worldwide. 85% of cervical cancer deaths occur in low-income countries occur in low-income

1 First Author Office assistant, Dept. of BigData medical, Eulji University, Korea. Email: tina7765@naver.com
2 Second Author. Doctoral student, Dept. of Medical Artificial Intelligence, Eulji University, Korea.
  Email: kyungakim@naver.com
3 Third Author, Professor, Dept. of BigData Medical Convergence, Eulji University, Korea. Email: machung@eulji.ac.kr
4 Corresponding Author, Professor Dept. of Medical IT, Eulji University, Korea. Email: mskang@eulji.ac.kr

countries. The WHO proposed strategy plan to require global practice for cervical cancer in May 2018 World Health Assistant for cervical cancer retirement in May 2018. It emphasized the political will to make cervical cancer a reality, and urged all stakeholders to unite for this common goal (Korea National Institute of Health, 2020).

Figure 1: Percentage of patients with cervical cancer by age in 2020

According to statistics from the Health Insurance Review and Assessment Service, 16,461 patients in their 40s accounted for about 25.7% of cervical cancer patients in 2020, followed by patients in their 50s and 30s with more than 20% each. In addition, compared to 2016, the increase rate of patients in their 20s and 30s was high. It increased by about 47% from 2,606 in 2016 to 3,836 in 2020, and by about 16.7% from 11,966 in 2016 to 13,970 in 2020 (South Korea Policy Briefing, 2020).

Contraceptive pills contain two types of estrogen preparations and progesterone preparations, and are classified into generations 1 to 4 depending on the type and content of these two preparations. The side effects and efficacy of each generation of contraceptives vary slightly (Hwang, 2022). In this way, data from patient samples is identified and data preprocessing is used to predict whether the risk of cervical cancer really increases depending on the taking and duration of hormone contraceptives. In addition, after constructing a model using patient sample variables, we learned and predicted with the Logistic Regression algorithm to predict whether the patient would develop cervical cancer.

2. Related Researches

2.1. Related Research

In March 2002, "Lancet" also published a paper stating that the use of oral contraceptives over a long period of time increases the incidence of cervical cancer in women infected with the human papilloma virus by four times. Although the incidence of cervical cancer has not increased compared to women who have been using oral contraceptives for less than five years and women who have never taken oral contraceptives at all, women who have taken oral contraceptives for five to nine years were nearly three times more likely to develop cervical cancer, the paper. It was also reported that women who used oral contraceptives for more than a decade were more than four times more likely to develop cervical cancer. In developed countries, women over the age of 50 who rarely use oral contraceptives have a 3.8-person risk of developing cervical cancer per 1,000 people. However, for women who used oral contraceptives for five years, the figure rose to four per 1,000 people, and for women who used them for 10 years, it rose to 4.5 people. Meanwhile, in underdeveloped countries, 7.3 and 8.3 per 1,000 people, respectively (Lee, 2007).

2.2. Cervical Cancer Analysis

The most common symptom of invasive cervical cancer is postcoital vaginal bleeding, occurring after sexual intercourse. Additionally, irregular vaginal bleeding may also be observed. As the cancer progresses, pain and swelling in the legs may be present. Cervical cytology, a method where cells are collected from the cervix using a small instrument and examined to assess their condition, is commonly performed. Furthermore, individuals may undergo HPV virus infection testing through genetic analysis. If high-risk HPV infection is detected or abnormal cells are found in cervical cytology, colposcopy, an examination of the cervix with magnification, may be conducted. In severe cases, imaging techniques such as CT or MRI scans are used to assess the extent of metastasis. (Hu, 2023). According to an analysis of stage-related factors of cervical cancer patients at Yonsei University's graduate school, the factors influencing cervical cancer were identified as major variables such as age, education level, marriage age, past cytology, early pregnancy age, first sex age, menopause, and oral contraceptive use. It was found that 23.8% of the age, 7.7% of the past cytology test, 3.7% of the use of contraceptives, and 1.6% of the first sexual age, and the age of the subject's age, past cytology test, oral contraceptive use, and the first sexual age were suggested as major influences factors for cervical cancer. (Lee, 2002)

According to the study, 67.1% of subjects with intraepithelial cancer underwent cytology tests, and 23.9% of subjects with advanced stage underwent cytology tests, and fewer subjects with advanced stage underwent cytology tests in the past ($x^2 = 26.124$, P=.000). The first sexual

intercourse age of those diagnosed with intraepithelial cancer is 7.2% for those under the age of 19, 11.4% for early stages, and 23.9% for advanced stages, and the first sexual intercourse age is faster for those diagnosed with advanced stages (2 =20.461, P=.000). In terms of contraception, 52.4% of the subjects had contraception in intraepithelial cancer, but 15.6% of the subjects were pregnant in the advanced stage, and the more those diagnosed with the advanced stage, the more they did not (2 =18.466, P=.000). On the other hand, 16.9% of those who were diagnosed with intraepithelial cancer used contraceptives, but 34.8% of those who were diagnosed with advanced diseases used contraceptives, and the more advanced they were (2 =16.029, P=.000) (Lee, 2002).

3. Body

There are many papers on the effects of hormonal contraceptives on cervical cancer. Most papers and studies used statistical analysis methods using the SPSS program through surveys. In this paper, for more accurate research, we learned and predicted the probability of hormonal contraceptives developing cervical cancer using the Logistic Regression algorithm using artificial intelligence.

4. Experiment

4.1. Experimental Environment

Among the machine learning platforms are Rapidminer, Dataiku, Dataiku DSS, Kraken, Altarless, Matlab, SAS, Azure ML Studio, IBM Watson Studio, Google Cloud AI, Amazon Web Service AWS, etc. (Park, 2021)

In this experiment, Microsoft Azure was used. Microsoft Azure provides multiple writing environments depending on the project type and ML environment level without the need to install. Machine learning models can be learned and distributed without writing code, and experiments can be analyzed and optimized using visualization (Mirosoft Azure manual, 2022).

4.2. Generation of Data Set

Although I wanted to use patient record data in domestic hospitals, I tried to produce data because it was difficult to use patient data due to personal information protection, but it was difficult to produce it. While searching for cervical cancer risk factor data, we used data collected from the "Hospital Universitario de Caracas" in Caracas, Venezuela, from the Interpretable (Molnar, 2023). The data consists of 36 attributes, including demographic information, habits,

and past medical records of 998 patients. To improve the performance of the analysis, except for 29 attributes such as sex-mediated diseases, non-hormone contraception, diagnosis, and smoking, the data was composed of a total of seven attributes: age, number of sex partners, first sex (age), number of pregnancies, hormonal contraceptives, and hormonal contraceptives (year).

4.3. Data Preprocessing

Data preprocessing is carried out before model learning. I used the Azure Machine Learning Studio. The column was converted from English to Korean with the Edit Metadata block. If you look at the data, the format of the result is 0 or 1, but using the Edit Metadata block, 0 is defined as false, meaning 1 is true, and if hormone contraceptives are not taken, it is converted to false and 1 is true. Except for age and results, the remaining five attributes have missing values, which were omitted because they were not answered due to privacy concerns. The missing values were filled with empty data by replacing Int-type data and Bool data values with average using the Clean Missing Data block. Figure 2 shows the data preprocessing results.
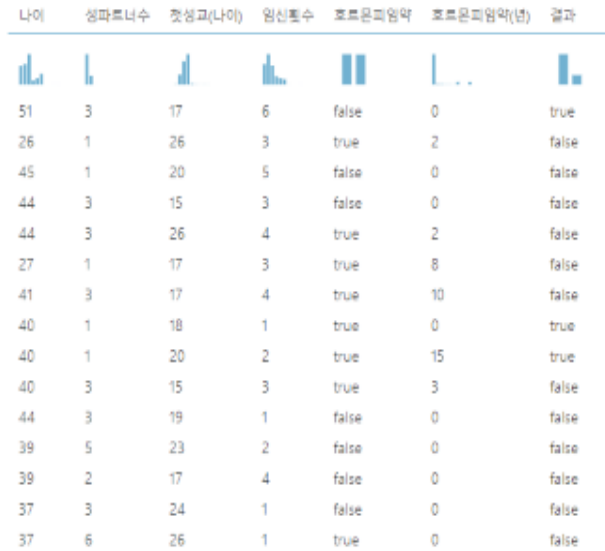
나이	성파트너수	첫성교(나이)	임신통수	호르몬피임약	호르몬피임약(년)	결과
51	3	17	6	false	0	true
26	1	26	3	true	2	false
45	1	20	5	false	0	false
44	3	15	3	false	0	false
44	3	26	4	true	2	false
27	1	17	3	true	8	false
41	3	17	4	true	10	false
40	1	18	1	true	0	true
40	1	20	2	true	15	true
40	3	15	3	true	3	false
44	3	19	1	false	0	false
39	5	23	2	false	0	false
39	2	17	4	false	0	false
37	3	24	1	false	0	false
37	6	26	1	true	0	false

Figure 2: Visualize data preprocessing

Examining how well one predicts the dependent variable values of a sample data set that are not used for learning is called cross-validation (Kim, 2019). For cross-validation, the ratio is set to train set and test set. Using the Split Data block, each data set was analyzed by setting it to 70% and 30%.

5. Results

The learned results show that the weight is 0.6 when not taking hormone contraceptives, and the weight is -0.6 when taking hormone contraceptives, which further increases the risk of cervical cancer. If the weight is 0, it is marked as 0 if it is determined that it does not affect the probability of developing cervical cancer.

Evaluating the performance of a model is described as Matrix. Metrics refer to numerical expressions of the performance of a learned model. Depending on the type of model you have learned, the indicators you need to look at also vary (Microsoft Azure Machine Learning, 2024). True Positive means that the True value, which has the same evaluation data and the model prediction, means 381 True values, on the contrary, the model prediction is True, but the evaluation data has 94 other data. The model predictions were evaluated negatively, but if the original data were True, the model evaluated the prediction negatively at the end of the four, and 31 predictions were found for the data where the actual data was also evaluated negatively. Data values calculated by metrics, and performance evaluation indicators of the classification model include accuracy, precision, Recall, and F1 Score. Accuracy is a prediction rate that exactly matches the actual class label, and the closer it is to 1, the better. Precision refers to the ability of a model to not label a negative sample as positive, and the closer it is to 1, the better. Precision refers to the ability of a model to not label a negative sample as positive, and the closer it is to 1, the better. Recall refers to the function of a model that searches for both positive samples, and the closer it is to 1, the better, and F1 score is the harmonic average of precision and reproduction rates, and the closer it is to 1, the higher the performance evaluation (Mirosoft Azure Machine Learning, 2024). The classification model for this experiment was 80.8% Accuracy, 80.2% Precision, 99.0% Recall, and 88.6% F1 Score, showing high performance classification models. Figure 3 visualizes the model performance.
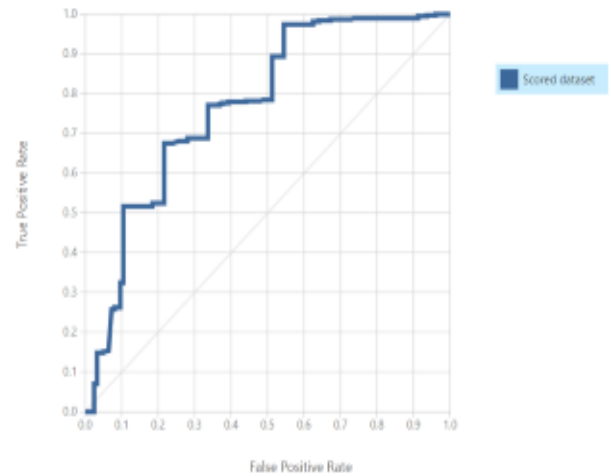

Figure 3: Visualize model performance

6. Conclusion

In this paper, we conducted a study to predict the probability of developing cervical cancer with data collected from "Hospital Universitario de Caracas". Studies have shown that taking hormone contraceptives increases the risk of cervical cancer. The predictive model uses a logistic regression model to further reduce the incidence of cervical cancer by adding new independent variables related to smoking, IUD use, and sexually transmitted diseases as well as hormonal contraceptives.

References

Hu, K., Wang, J., Sparén, P., Herweijer, E., Sjölander, A., Adami, H. O., ... & Fang, F. (2023). Invasive cervical cancer, precancerous lesions, and cervical screening participation among women with mental illness in Sweden: a population-based observational study. *The Lancet Public Health*, *8*(4), pp. e266-e275.

Hwang W, H. (2022). Infertility, Cancer Effects Misunderstanding and Truth About Contraceptive Pills. *Healtip*. Retrieved From http://www.healtip.co.kr/news/articleView.html?idxno=2433#rs.

Kim, D, H. Cross-validation [online blog]. Retrieved from https://datascienceschool.net/03%20machine%20learning/06.03%20%EA%B5%90%EC%B0%A8%EA%B2%80%EC%A6%9D.html.

Korea National Institute of Health. (2020). Women's Health in Figures 2020. *Korea centers for Disease Control and Prevention*. pp.1

Lee Y, S. (2007). Oral contraceptives also increase the risk of cervical cancer. Medigonews. Retrieved from https://www.medifonews.com/mobile/article.html?no=35578

Lee, E, H. (2002). Analysis of stage-related factors in cervical

cancer patients (Doctoral dissertation, Yonsei University Graduate School). Retrieved from https://ir.ymlib.yonsei.ac.kr/handle/22282913/128290

Microsoft Azure Machine Learning. (2024). Retrieved From https://docs.microsoft.com/ko-kr/azure/machine-learning/studio/what-is-ml-studio

Mirosoft Azure Machine Learning. (2024). Retrieved from https://learn.microsoft.com/ko-kr/azure/machine-learning/how-to-understand-automated-ml?view=azureml-api-2#roc-curve

Mirosoft Azure Machine Learning. (2024). Two-Class Logistic Regression, Article from https://learn.microsoft.com/en-us/previous-versions/azure/machine-learning/studio-module-reference/two-class-logistic-regression?WT.mc_id=AI-MVP-5003262

Molnar, C. (2023). Interpretable Machine Learning, *leanpub*. pp.4

Park H, S. (2021). Top 10 Best Machine Learning Platforms in 2021. Airtimes. Retrieved from https://www.aitimes.com/news/articleView.html?idxno=137999

SNUH blog. (2022). Medical information by cancer type-uterine cervical cancer. Retrieved from http://cancer.snuh.org/info/classification/list11.do?searchType=&tab=1&seq_no=1

South Korea Policy Briefing (2020) A common disease in women; cervical cancer. Retrieved from https://www.korea.kr/news/healthView.do?newsId=148895283