**University of Toronto Scarborough, STAC51: Winter 2021**

# Case Study: Predictors of Cervical Cancer Relapse in Adults after Treatment

Ali Krisht

**Table of Contents**



# 1 Introduction

Annually, an estimated 225,000 Canadians are diagnosed with cancer of which roughly 1500 women are diagnosed with cervical cancer (Cervical Cancer Statistics, Feb 2021). Cervical cancer is a cancer that occurs in the cells of the cervix which is located in the lower part of the uterus. The cancer develops as healthy cells on the surface of the cervix become infected with the human papillomavirus (HPV), which can expand uncontrollably creating a mass known as a tumor (Government of Canada, 2017). Radical hysterectomy (the removal of the uterus) is the standard form of treatment for the early stages of cervix cancer. While both surgery and radiation therapy provide comparable cure rates, surgery is mostly favoured for healthy patients due to a shorter recovery course, the possibility of ovarian retention, and improved post-treatment outcomes (Canadian Cancer Society, 2017). Even though there is a success rate of approximately 80%, there is still a 20% chance that a patient can relapse. This problem is very prevalent in the medical industry as there have been many prior studies on whether factors such as smoking, sexual activity, frequent pregnancy, and birth control induce relapse of cervical cancer in adults (American Cancer Society, 2020).

The data used to conduct this research is from TSRCC, Toronto Sunnybrook Health Science Center. It has been collecting data from 1984 on factors related to cervical cancer in 905 patients (Covens, 2002). Some of the variables from the data set are hypothesized to have an impact on relapse for adults include age, radiation, disease status, cell differentiation, depth, size, pelvis involvement, and CLS. The goal of this study is to evaluate covariates such as the ones above to determine if they cause relapse in adults after treatment. After determining such predictors, medical professionals may gain a better understanding of the next steps and possible improvements in the care of their patients. Also, oncologists may be able to use this information to find the root cause of cervical cancer and verify predictors of relapse.

# 2 METHODS

## 2.1 Variable Selection and Reasoning:

As the outcome of interest is RECURRN1 and the predictor variables vary from age to size, it is important to analyze these variables using univariate, bivariate, and exploratory data analysis. Other than the desired response of RECURRN1 (the relapse date if there was a relapse), this study will make use of 9 of the 11 available covariates. The variables are as follows: AGE_1 (age of the patient at time of surgery), DIS_STA (0: no evidence of disease, 1: alive with disease, 2: dead with disease, 3: dead with complications (disease present), 4: dead with complications (disease absent), 5: dead of unrelated causes), GRAD_1 (cell differentiation: 1:better, 2:moderate, 3:worst), ADJ_RAD (0: no radiation therapy, >= 1: The sites of the radiation therapy), MAXDEPTH_1 (depth of the tumor in mm), PELLYMPH_1 (pelvis involvement: 0:negative, 1:positive), SIZE_1 (size of the tumor in mm), and CLS_1 (Capillary lymphatic space: 0: negative, 1: positive) (Cervical Cancer FAQs, 2002). To tidy this data, we firstly remove all the patients that did not have follow-up dates for the variable FU_DATE. We modified the 'RECURRN1' variable in such a way that it can provide us with an output of 0 or 1, with 0 representing no recurrence of the disease in the patient, and 1 representing the date of relapse. Since our research question only pertains to adults, we modified the dataset to only contain adult patients in the age group of 25 to 64. For the variable 'ADJ_RAD' there were values that were greater than 1, representing the radiated sites in the patient's body. To avoid this issue, we set all the values greater than 1 equal to 1, where 1 now represents patients that received the adjuvant radiation treatment, regardless of the part of their body it was on. Our research is related to the predictors causing relapse in adults, which means patients that died of complications (with no disease) or died due to unrelated causes, were not fit to be in our data frame. Hence, we also removed the data in the column for 'DIS_STA' containing '4' or '5'. In order to get the best possible model, we decided to remove the observations containing 'NA' values as those would skew our data and as a result, our plots and models would not be accurately represented. After cleaning the data, there are 10 columns (covariates of interest) and a total of 597 rows from the original 905.

## 2.2 Exploratory Data Analysis

```
    relapse             age          radiation         pellyMPH          disSTA            grad          maxDepth
 Min.   :0.00000   Min.   :25.00   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000   Min.   : 0.000
 1st Qu.:0.00000   1st Qu.:34.00   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 3.000
 Median :0.00000   Median :39.00   Median :0.0000   Median :0.00000   Median :0.0000   Median :2.000   Median : 5.000
 Mean   :0.06365   Mean   :40.85   Mean   :0.1139   Mean   :0.07035   Mean   :0.0938   Mean   :1.754   Mean   : 7.304
 3rd Qu.:0.00000   3rd Qu.:47.00   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:10.000
 Max.   :1.00000   Max.   :64.00   Max.   :1.0000   Max.   :1.00000   Max.   :2.0000   Max.   :3.000   Max.   :50.000
      size              cls           agegroup
 Min.   : 0.000   Min.   :0.0000   Length:597
 1st Qu.: 0.000   1st Qu.:0.0000   Class :character
 Median : 0.000   Median :0.0000   Mode  :character
 Mean   : 7.462   Mean   :0.4539
 3rd Qu.:15.000   3rd Qu.:1.0000
 Max.   :70.000   Max.   :1.0000
```

*Figure 0*. Summaries of the covariates

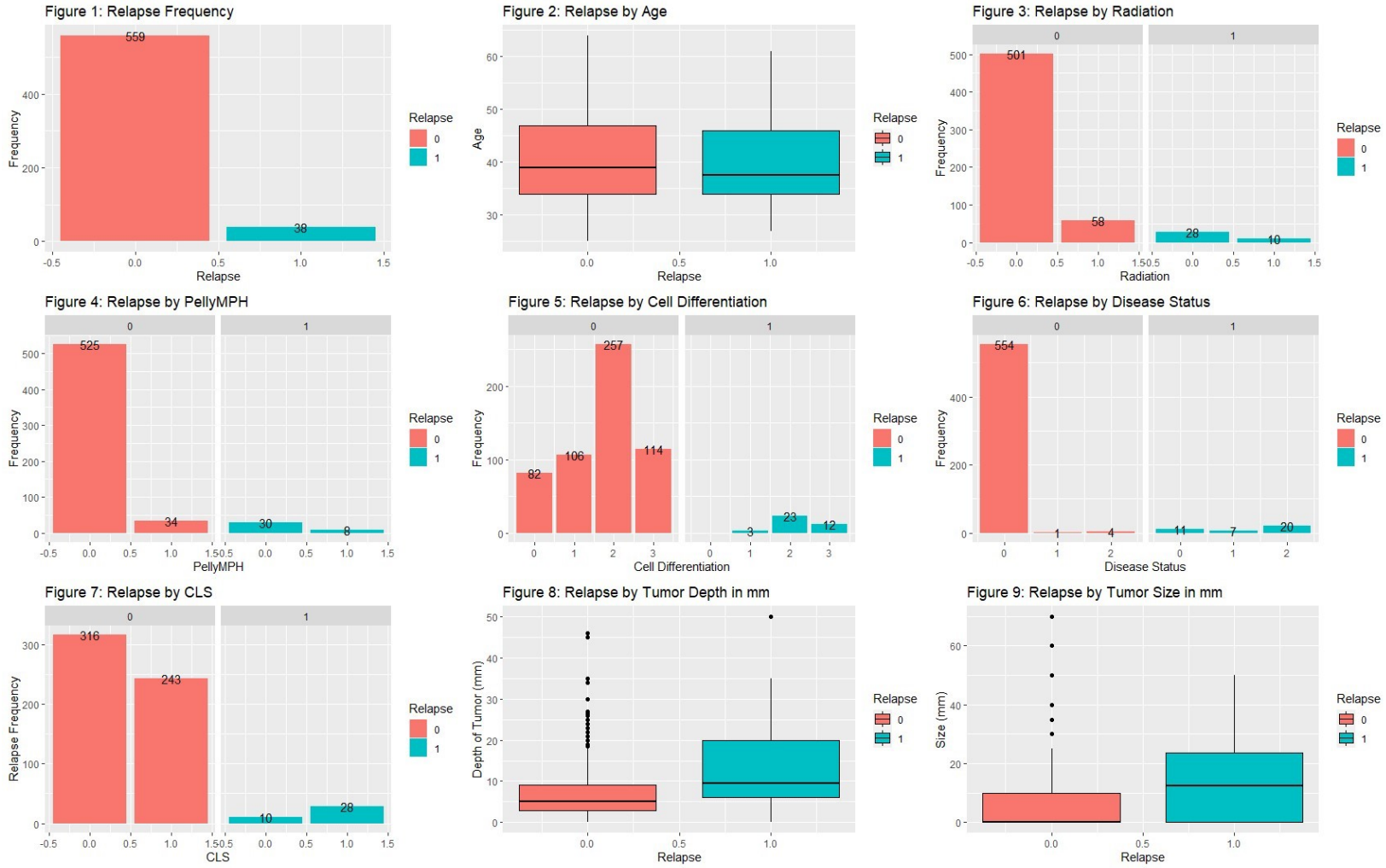
As seen in *Figure 0* there are only three continuous variables; age, maxDepth, and size. The mean and median age of the adult patients is 39 and 41 respectively, with minimum and maximum ages of 25 and 64 respectively. The average max depth of a tumor is 7.304 mm and the median depth is 5 mm. The average size of a tumor is 7.46 mm, with a minimum of 0 mm and maximum of 70 mm.

Having visualizations of all the covariates and the response variable is vital in gaining a better understanding of the data.

Figure 1-9. Exploratory data analysis plots

*Figure 1* represents the patient relapse frequency for all adults. As seen from the plot, the

93.64% patients did not relapse, however, 6.37% have relapsed. *Figure 2* displays the bivariate




analysis of patient relapse by age. The side-by-side boxplots for both adults who relapsed and who did not are quite similar in median and variability where the relapsed adults variability is slightly smaller. The median age for the no relapse adults is 39 whereas the median age relapsed adults is 38. There are also no outliers in the data. *Figure 3* is a faceted bar plot representing whether adult patients had relapsed if did and did not receive radiation therapy. It can be determined that of the 93.64% patients that did not relapse, 11.58% did not receive radiation therapy. Subsequently, of the 6.37% of patients that relapsed, 35.71% received radiation therapy.

*Figure 4* represents whether adults had relapsed if they had pelvic involvement. 93.64% of patients who did not relapse, 6.48% patients  did not have any involvement. On the other hand, 6.37% of adults who relapsed, 26.67% had some sort of pelvic involvement. *Figure 5* represents the relapse of patients based on the cell-differentiation. Patients that have a missing value for`GRAD_1` have a no relapse record. 97% of the adults with a better cell differentiation did not relapse. However, 8.2% of patients relapsed having a moderate cell-differentiation and 10.5% having worst cell-differentiation. *Figure 6* displays the relapse of patients by the status of disease. It can be determined that 98% of the patients did not relapse as they had no evidence of

the disease during their follow up. However, 87.5% of patients were found with evidence of disease had relapsed. 83.3% of patients that relapsed died with the disease present.

*Figure 7* represents a grouped bar plot as we have two categorical variables, relapse and CLS. It shows us adult patients if they relapsed, and their frequency based on the prognostics of CLS. 97% and 89.6% of patients did not relapse having tested negative and positive, respectively. *Figure 8* shows the bivariate analysis of patient relapse and the depth of tumor. This side-by-side boxplot tells us that the median of tumor depth is much lower for patients who did not relapse (~5 mm) than it is for patients who did (~9mm). Outliers are present for both the boxplots, but they are mostly for patients with no relapse, and only 1 outlier for patients that relapsed. *Figure 9* displays whether the patients relapsed or not, depending on the size of tumor. The median of the tumor for patients who did not relapse is miniscule, just above 0. This could be due to the fact that the doctors found this tumor extremely early, and diagnosed the patients correctly so they did not relapse again. On the other hand, the median of tumor for patients that relapsed is higher, approximately 12 mm. Outliers are present for patients that did not relapse, in fact the biggest tumor (~70 mm) was found in a patient that did not relapse.

## 2.3 Variable Significance

Before creating and selecting a complex model it is helpful to check the significance of the covariates in a simple model. To perform this task, the vglm() method from the VGAM package is used, and likelihood ratio test (lrtest() method) is applied to the vglm model. The purpose is to check the significance for each covariate with the p-values from the LRT.

| Variables: | P-value | Significance |
|---|---|---|
| age | 0.09584 | $< \alpha = 0.1$, fair predictor |
| disSTA | < 2.2 x 10^-16 | $<\alpha = 0.001$, very good predictor |
| maxDepth | 0.0178 | $<\alpha = 0.01$, good predictor |
| size | 0.8337 | $<\alpha = 1$, poor predictor |
| radiation | 0.842 | $<\alpha = 1$, poor predictor |
| pellyMPH | 0.2579 | $<\alpha$ = poor predictor |
| grad | 0.4044 | $<\alpha$ = poor predictor |

*Figure 10.* P-values and significance of the predictors

Based on *figure 10*, we can conclude that age, disSTA, and maxDepth are good predictors for the model as those p-values are significant. We should expect to see these 3 variables when selecting our final model.

## 2.4 Check for Correlation

Checking for multicollinearity between the three continuous covariates (age, maxDepth, size), may result in a simpler model since a strong correlation value (> 0.80) causes troubles in GLMs.

*Figure 11* shows that none of the covariates have a correlation value > 0.80, so we may conclude that there is no correlation among the variables.

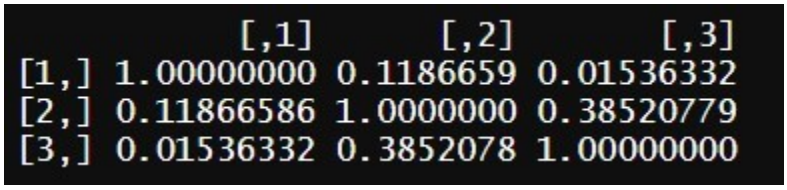



*Figure 11.* Covariance matrix for age, maxDepth, and size

## 3 Model Selection

The method of fitting for this data is best suited with generalized linear-regression models. Such models are the ideal way to predict the strength and significance of many predictor variables. The binomial GLM model with default link for the response relapse contains the covariates: age, radiation, disease status, cell differentiation, depth, size, pelvis involvement, and CLS. There is also an interest for an interaction between size and the depth of the tumor as it is prevalent in staging patients for choosing relevant treatments. From the summary, the predictors of disSTA, maxDepth, size, and maxDepth:size have significant p-values of 2.38e-14, 0.000153, 0.026255, and 0.018045 respectively. The last step in the model selection process is running the step() function in both forward and backwards directions, and determining which one provides the lowest AIC value. The backwards direction step gives an AIC value of 139.01 and the final model as seen in *figure 11.1* below, whereas the forward direction gives a value of 145.02 and the final model stated in *figure 11.2*. The AIC for the backward direction is smaller thus we can choose the model containing age, disSta, maxDepth, size and the interaction maxDepth:size.

$$y = -3.661985 - 0.041491age + 3.356980disSta + 0.144841maxDepth + 0.077010size - 0.005459maxDepth : size$$

*Figure 11.1*. Backwards model for Lowest AIC of 139.01

$$y = -4.006846 - 0.043369age + 0.124326radiation - 1.305302pellyMPH + 3.390985disSta + 0.114930grad + 0.151124maxDepth + 0.078939size + 0.343449cls - 0.00560maxDepth : size$$

*Figure 11.2*. Forwards model for AIC of 145.02

### 3.1 Model Validation

The model with the lowest AIC value must be validated. Since our data is ungrouped, the deviance will not fit the chi-squared distribution, alternatively we can use a test called HosmerLemeshow test that will identify whether or not the observed data matches the expected one. From *figure 12* below, p-value is 0.2768, we fail to reject the null, and the current model that we have is sufficient (fits the data well). Furthermore, the classification table shows us the

number of successes (y = 1) predicted by the model, with the number actually observed, and same thing for the number of failures (y = 0). In this 4x4 matrix we will focus on 2 outcomes sensitivity (True Positive) and specificity (True Negative) (Zainotz, 2020).

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  mod.2$y, fitted(mod.2)
X-squared = 5.1035, df = 4, p-value = 0.2768
```

*Figure 12.* Output from Hosmer and Lemeshow goodness of fit test including the p-value

### 3.2 Classification table

```
                predicted
response    0    1
       0  555    4
       1   16   22
```

*Figure 13.* Classification table used to sensitivity, specificity, and concordance rate

From *figure 13*, we can conclude that the sensitivity is 0.5789, specificity is 0.9928, and the concordance rate is 0.96649. We then move on to find the ROC (receiver operating characteristic) curve, it is the concordance index which estimates the probability that both the predictions and the outcome are concordant. The bigger the area under the ROC curve the better the model is.

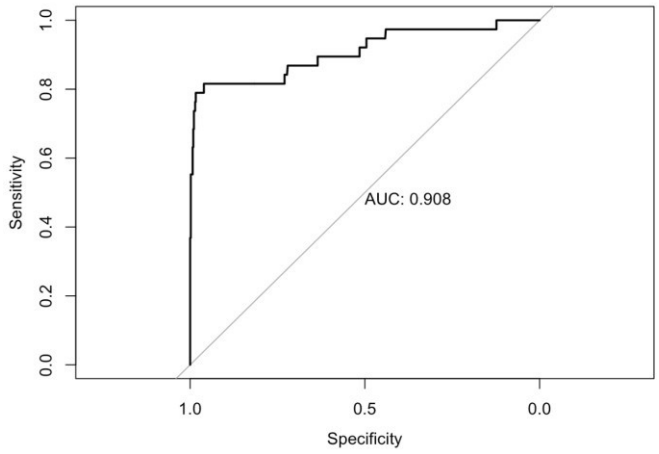



*Figure 14.* Receiver Operating Characteristic (ROC) Curve

In *figure 14* we can see that the area under the ROC curve is 0.908 (90.8%), and since it's very close to 100, this also confirms that the model we found is sufficient.

### 3.3 Final Model

Final model contains the covariates y = -3.661985 -0.041491 age + 3.356980 disSta + 0.144841 maxDepth + 0.077010 size -0.005459 maxDepth:size.

## 4 Conclusion

The goal of this research was to evaluate the covariates used throughout this study to determine if they cause relapse in adults after treatment. It was surprising that when performing the likelihood-ratio test, size was found to be insignificant. However, when included in a more complex model with interaction terms, it was determined to be prevalent. Through our analysis, it has been proven that age, disSTA, maxDepth, size, and the interaction term between size and maxDepth are the major factors that play a critical role in patients relapsing after surgery. This resulting model contradicts the initial hypothesis which included age, radiation, disease status, cell-differentiation, depth, size, pelvis involvement, and CLS. This was proven by that fact that we got the lowest AIC of 139.01, Hosmer and Lemeshow p-value of 0.28, high sensitivity and specificity values of 0.58 and 0.99 respectively, and the ROC AUC of 90.8% all indicating a good model fit.

This study can be very beneficial to the field as scientists are trying to find new methods to detect cervical cancer in its early stages, so they can diagnose it and remove the tumor and decrease the chance of the relapse. (Cervical Cancer Research, 2020). From this research, we found useful predictors in determining adult relapse, and scientists would be able to use this information in finding cancer in such patients. The data did provide us with quite a few limitations, as the original data consisted of 905 observations. However, there was a lot of missing data, and after tidying it up, we were left with 597 rows, which is ~37% lower than original data. Also, the sample size for the relapsed patients is quite small which can hinder the final results. Another limitation is that the data was collected in an environment with good healthcare and thus our findings cannot be generalized for relapsed patients in places with no access to good healthcare. For future research, the data also could have included other factors, such as whether the patients smoke, their sexual history, full term pregnancies, and usage of contraceptives (Chen X, 2011). Using such factors, researchers can also help distinguish between similar conditions such as Prolapsed Uterine Fibroid which has similar symptoms of abnormal bleeding and large masses (Al-Shukri, et al, 2019). It would also be more relevant to check the importance of covariates when we split the response variable, and to check whether they relapsed again, this might allow us to generate an even better model. We can also perform a more advanced level of analysis, such as using classification trees to expose the structure of data to get a better understanding on how different variables impact the data, and how removing unwanted values such as ‘NA’ causes the tree to be different. Overall, predictors such as age, disease status, max depth, size, and the interaction between max depth and size have an impact on relapse.

# 5 References


- Cervical cancer - statistics. (2021, February 08). Retrieved April 09, 2021, from https://www.cancer.net/cancer-types/cervical-cancer/statistics
- Canada, P. (2017, October 23). Government of Canada. Retrieved April 10, 2021, from https://www.canada.ca/en/public-health/services/chronicdiseases/cancer/cervicalcancer.html
- What is cervical cancer? - canadian cancer society. (n.d.). Retrieved April 10, 2021, from https://www.cancer.ca/en/cancer-information/cancertype/cervical/cervicalcancer/?region=on
- "Cervical Cancer Risk Factors: Risk Factors for Cervical Cancer." *American Cancer Society*, The American Cancer Society, 3 Jan. 2020, www.cancer.org/cancer/cervicalcancer/causes-risks-prevention/risk-factors.html.
- Cervical cancer research: Latest research in cervical cancer. (2020, January 03). Retrieved April 11, 2021, from https://www.cancer.org/cancer/cervicalcancer/about/newresearch.html
- Government of Canada, S. (2017, May 08). Age categories, life cycle groupings. Retrieved April 09, 2021, from https://www.statcan.gc.ca/eng/concepts/definitions/age2
- Zainotz, C. (2020, June 30). Daniel Debbarma. Retrieved April 12, 2021, from https://www.real-statistics.com/logistic-regression/classification-table/
- Chen, X., Jiang, J., Shen, H., & Hu, Z. (2011, May 25). Genetic susceptibility of cervical cancer. Retrieved April 11, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3597058
- York University. (2002, May 3). Cervical Cancer Case Study FAQs. Retrieved from http://www.math.yorku.ca/Who/Faculty/Ng/ssc2002/CervicalFAQ.htm

- Covens, A. (2002). Cervical Cancer. Retrieved April 6, 2021, from https://ssc.ca/en/casestudy/cervical-cancer
- Al-Shukri, M., Al-Ghafri, W., Al-Dhuhli, H., & Gowri, V. (2019, November). Vaginal Myomectomy for Prolapsed Submucous Fibroid: It is Not Only About Size. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6851067/