

Report Outline



- I. Background
 - A. Data Description
 - B. Research Goals
- II. Data Reprocessing
- III. Exploratory Data Analysis
 - A. Multicollinearity
- IV. Model Selection
 - A. Model and tests of fit
 - B. ROC
 - C. Machine Learning
 - D. Plots
- V. Conclusions and Final Thoughts
- VI. References

I. Background

Cancer is one of the hardest to treat ailments of humanity. According to the Canadian Cancer Society, in 2019 1.3% of 107,400 new cases of cancer in women were cases of cervical cancer. The cervix is the opening of the uterus and its health is crucial to the reproductive system's function. The area is home to a large number of blood vessels and lymphatic channels which can increase the risk of spread. In order to construct the most effective treatment plan we must understand the contributing factors and risks (Canadian Cancer Society, n.d).

The data in question has been gathered in Toronto from about 1984 to present and covers 905 cervical cancer patients. The researches recorded the following values:

- **MRNO**: Patient number
- **SURGDAT**: Date of surgery/Date of diagnosis (used interchangeably)
- **AGE**: Age
- **CLS**: Capillary Lymphatic Spaces - presence of tumor cells inside of the capillary lumens of either the lymphatic or microvascular drainage systems within the primary tumor (Chen, M., Jin, Y., Bi, Y., Li, Y., Shan, Y., & Pan, L., 2015)
- **DIS_STA**: Disease Status – whether the patient is dead of complications (disease present/absent) or unrelated causes, alive with disease or without.
- **GRAD**: Grade of cell differentiation – how much do the cancer cells differ from healthy cells.
- **HISTOLOG**: Type of cancer
- **MARGINS**: Disease left after primary surgery – none, in the para-vaginal area, vaginal area or both.
- **MAXDEPT**: Depth of tumor (mm)
- **PELLYMPH**: Pelvis lymph node involvement
- **RECURRN**: Date of recurrence of disease
- **SIZE**: Size of tumor
- **FU_DATE**: Last follow up date



Our main goals are to determine which of the factors above affect the probability of relapse of the disease and to classify the patients by individual risk of relapse.

II. Data Reprocessing

Pre-Processing

- Load packages and data
- Rename column names for easier data handling

Data cleaning

We consider the following cases and modify our dataset accordingly.

1. Patients with no FU_DATE are to be excluded, as it is unknown whether they have relapsed or not and thus their data is useless for our research purposes.
2. The original data does not include a column for a binary outcome for relapse of the cancer. We create this column, and give it a value of 0 for no relapse and 1 for relapse, depending on if the column RECURRN1 has an NA value or not.
3. Original data has some ADJ_RAD values >1, we recode this so that patients who did not receive radiation therapy have ADJ_VAL of 0, and patients that did receive radiation therapy have ADJ_VAL value of 1 (regardless of which site was radiated).
4. Patients who have died without relapse can also be excluded, as they didn't have the full opportunity to relapse.
5. Categorical variables are columns that can be grouped under common subject, we convert these data types to factor for later analysis.
6. Certain rows contain NA/missing values for certain columns (ADJ_RAD, AGE, CLS, DIS_STA, GRAD, HISTOLOG, MARGINS, MAXDEPTH, PELLYMPH, RECURRN_Y_or_N, SIZE) and will not be useful for our research purposes, thus we remove these rows.

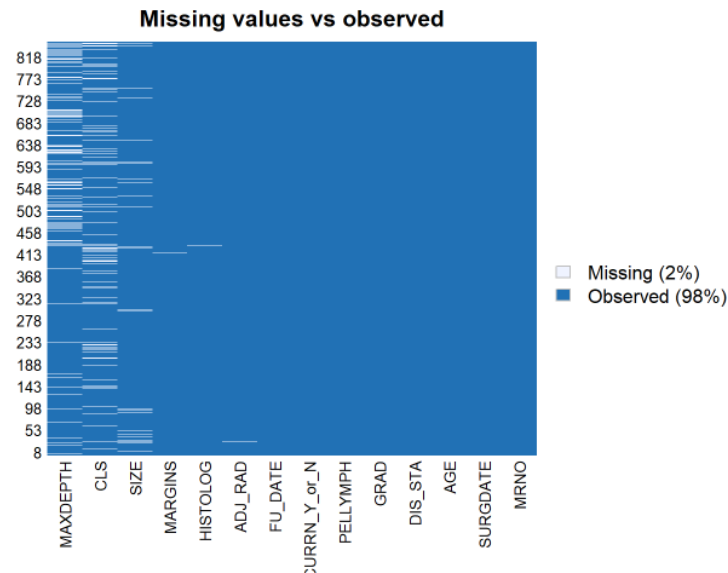
After processing, we are able to remove 369/905 rows where the data is not useful.

```
glimpse(cervical_cancer_data_clean)
```

```
## Rows: 536
## Columns: 15
## $ MRNO      <dbl> 2, 8, 15, 16, 19, 22, 24, 30, 31, 32, 34, 36, 37, 38...
## $ SURGDATE  <dtm> 1984-10-30, 1985-01-17, 1985-07-25, 1985-07-26, 198...
## $ ADJ_RAD   <fct> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0...
## $ AGE       <dbl> 35, 57, 61, 34, 41, 38, 36, 32, 26, 31, 27, 32, 53, ...
## $ CLS       <fct> 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 2, 0...
## $ DIS_STA   <fct> 0, 0, 0, 2, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ GRAD      <fct> 2, 2, 3, 2, 3, 2, 2, 2, 2, 3, 1, 3, 3, 3, 3, 2, 2, 3...
## $ HISTOLOG  <fct> 1, 1, 1, 4, 1, 1, 4, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1...
## $ MARGINS   <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ MAXDEPTH  <dbl> 15, 3, 15, 2, 9, 25, 5, 2, 6, 15, 0, 8, 35, 4, 0, 7,...
## $ PELLYMPH  <fct> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ RECURRN1  <dtm> NA, NA, NA, 1986-04-28, NA, 1989-02-22, NA, NA, NA,...
## $ RECURRN_Y_or_N <dbl> 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ SIZE      <dbl> 20, 0, 25, 0, 0, 30, 0, 0, 10, 0, 0, 10, 10, 0, 10, ...
## $ FU_DATE   <dtm> 1994-08-29, 1995-01-20, 1995-02-01, 1987-05-16, 199...
```

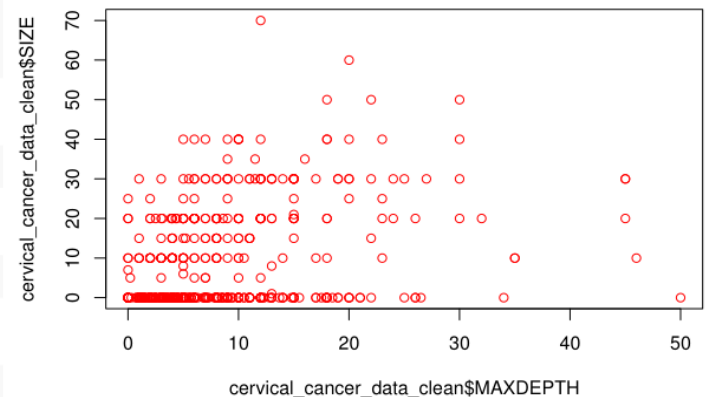
III. Exploratory Data Analysis

Observation of missing values in our data after cleaning



A. Multicollinearity

```
#Checking if any predictor variables are correlated  
  
# Corr size and age:  
cor(cervical_cancer_data_clean$SIZE, cervical_cancer_data_clean$AGE)  
  
## [1] 0.01647459  
  
# Corr depth and age:  
cor(cervical_cancer_data_clean$MAXDEPTH, cervical_cancer_data_clean$AGE)  
  
## [1] 0.1154779  
  
# Corr depth and size:  
cor(cervical_cancer_data_clean$MAXDEPTH, cervical_cancer_data_clean$SIZE)  
  
## [1] 0.362087  
  
plot(cervical_cancer_data_clean$MAXDEPTH,  
      cervical_cancer_data_clean$SIZE, col="red")
```



We see that there is no pattern that signifies correlation between MAXDEPTH and SIZE. Furthermore, correlation values for SIZE and AGE, MAXDEPTH and AGE, are also small. Thus multicollinearity does not appear to be a problem.

B. Model Selection

By looking at the data it is evident that the response variable recurrence is binary, therefore the most likely model for the data is binomial. We explore the logit, probit, and cloglog links and compare AIC values:

- Logit:

```
# Logit
cervical_cancer.fit_all_pred_logit <-
  glm(RECURRN_Y_or_N ~
    ADJ_RAD + AGE + CLS +
    GRAD + HISTOLOG + MARGINS + MAXDEPTH + PELLYMPH + SIZE,
    data=cervical_cancer_data_clean, family=binomial(link = "logit"))

cervical_cancer.fit_all_pred_logit$aic

## [1] 283.4876
```

- Probit:

```
# probit
cervical_cancer.fit_all_pred_binomial_probit <-
  glm(RECURRN_Y_or_N ~
    ADJ_RAD + AGE + CLS +
    GRAD + HISTOLOG + MARGINS + MAXDEPTH + PELLYMPH + SIZE,
    data=cervical_cancer_data_clean, family=binomial(link="probit"))

cervical_cancer.fit_all_pred_binomial_probit$aic

## [1] 282.8241
```

- Cloglog

```
# cloglog
cervical_cancer.fit_all_pred_binomial_cloglog <- glm(RECURRN_Y_or_N ~
  ADJ_RAD + AGE + CLS +
  GRAD + HISTOLOG + MARGINS + MAXDEPTH + PELLYMPH + SIZE,
  data=cervical_cancer_data_clean, family=binomial(link="cloglog"))

cervical_cancer.fit_all_pred_binomial_cloglog$aic

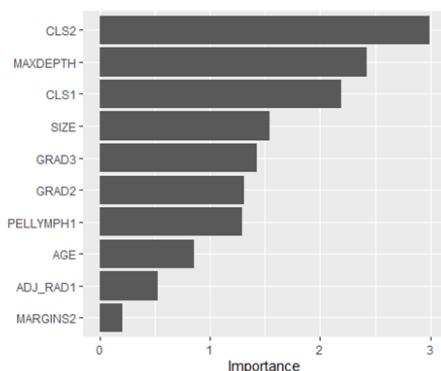
## [1] 283.712
```

Binomial with probit link has the lowest AIC value and looks like the better model. However, the difference in AIC values is extremely small and logit link is easier to interpret.

Therefore we select the binomial model with logit link:

```
# Select binomial logit
cervical_cancer.fit_all_pred <-
  cervical_cancer.fit_all_pred_logit
```

Due to the large number of independent variables our goal is to eliminate the insignificant predictors. We examine impact of the variables on the model via a plot of importance scores and the odds ratios for model predictors:



The table of
confirm the
the importance

	OR	2.5 %	97.5 %
## (Intercept)	1.962485e-08	NA	2.150278e+203
## ADJ_RAD1	7.656840e-01	2.673418e-01	2.010062e+00
## AGE	9.857561e-01	9.526798e-01	1.017575e+00
## CLS1	2.536872e+00	1.128219e+00	6.044667e+00
## CLS2	5.870095e+00	1.812835e+00	1.879277e+01
## GRAD2	2.452423e+00	7.242464e-01	1.141262e+01
## GRAD3	2.888919e+00	7.383090e-01	1.472568e+01
## HISTOLOG1	6.928306e+05	4.531688e-207	NA
## HISTOLOG3	1.197060e+06	1.073836e-206	NA
## HISTOLOG4	9.914412e+05	1.494724e-206	NA
## HISTOLOG6	7.139802e-01	5.087200e-237	1.211925e+21
## MARGINS1	1.292189e+00	5.014067e-02	1.461104e+01
## MARGINS2	1.291386e+00	5.841115e-02	1.136508e+01
## MARGINS3	3.631242e-07	NA	9.686591e+107
## MAXDEPTH	1.049562e+00	1.007981e+00	1.090728e+00
## PELLYMPH1	1.994552e+00	6.806558e-01	5.654591e+00
## SIZE	1.019702e+00	9.942810e-01	1.045240e+00

odds ratios
results from
table where it

is evident that CLS and MAXDEPTH are associated with higher odds of relapse.

Model with no interactions: We fit the model with no interactions to find the most significant variables.

```
# Fit no predictors
cervical_cancer.fit_no_pred <- glm(RECURRN_Y_or_N ~ 1,
  data=cervical_cancer_data_clean, family=binomial)

cervical_cancer.fit_no_interaction_backward <-
  step(cervical_cancer.fit_all_pred, direction="backward", test = "Chisq")

cervical_cancer.fit_no_interaction_forward <-
  step(cervical_cancer.fit_no_pred, scope =~ ADJ_RAD + AGE + CLS + GRAD +
    HISTOLOG + MARGINS + MAXDEPTH + PELLYMPH + SIZE,
    direction="forward", test = "Chisq")

cervical_cancer.fit_no_interaction_both = step(cervical_cancer.fit_all_pred,
  direction = "both", test="Chisq")
```

We perform forward, backward and both step() and conclude that the resulting models are the same.

Therefore we pick the following model: RECURRN_Y_or_N ~ CLS + MAXDEPTH + SIZE.

Considering the summary for the model:

The variables that we previously found significant are in the model as expected.

```
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.89571 0.37278 -10.450 < 2e-16 ***
## CLS1 0.89439 0.40369 2.216 0.02672 *
## CLS2 1.70673 0.53438 3.194 0.00140 **
## MAXDEPTH 0.04898 0.01840 2.662 0.00776 **
## SIZE 0.02393 0.01207 1.983 0.04740 *
```

Model with second degree interactions: Due to the

complexity of the data it makes sense to consider the two-way interactions. For example, radiation therapy can affect depth and size of the tumor or the effect of type of tumor on cell differentiation. To inspect the relations we use step() starting with the model with all no-interaction variables. Backward step() (starting from all the variables and interactions) was insufficient since it simply did not remove enough variables to simplify the model, and the forward step() (starting from no variables and going up to the interactions) did not add enough terms.

```
cervical_cancer.fit_all_interactions_forward <-
  step(cervical_cancer.fit_no_pred, scope =~ (ADJ_RAD + AGE + CLS + GRAD +
    MAXDEPTH + PELLYMPH + SIZE)^2,
    direction="forward", test = "Chisq")
```

Output: RECURRN_Y_or_N ~ MAXDEPTH + CLS + SIZE + CLS:SIZE

```
cervical_cancer_fit_all_interactions_backward <-
  step(glm(RECURRN_Y_or_N ~ ADJ_RAD + AGE + CLS + GRAD + MAXDEPTH +
    PELLYMPH + SIZE + (ADJ_RAD + AGE + CLS + GRAD + MAXDEPTH +
    PELLYMPH + SIZE)^2,
    data=cervical_cancer_data_clean,
    family=binomial(link="logit")), test = "Chisq")
```

Output: RECURRN_Y_or_N ~ AGE + CLS + GRAD + MAXDEPTH + PELLYMPH + SIZE + AGE:CLS
+ AGE:MAXDEPTH + GRAD:MAXDEPTH + GRAD:SIZE + PELLYMPH:SIZE

```
cervical_cancer.fit_all_interactions_both <-
  step(cervical_cancer.fit_all_pred,
    scope =~ (ADJ_RAD + AGE + CLS + GRAD + HISTOLOG +
    MARGINS + MAXDEPTH + PELLYMPH + SIZE)^2, direction = "both",
    test="Chisq")
```

Output: RECURRN_Y_or_N ~ AGE + CLS + MAXDEPTH + PELLYMPH + SIZE + AGE:CLS
+ PELLYMPH:SIZE + AGE:MAXDEPTH

Therefore both step() is the obvious best choice and gives us:

RECURRN_Y_or_N ~ AGE + CLS + MAXDEPTH + PELLYMPH + SIZE + AGE:CLS + PELLYMPH:SIZE + AGE:MAXDEPTH.

In order to be able to compare the simpler and the more complex model, add CLS:SIZE into the model resulting from both step().

However, in the summary of the model the p-value for the AGE:MAXDEPTH interaction is higher than 0.05, so we use the AIC value and decide to drop the term. The AIC of the model with AGE:MAXDEPTH is 258.23 and AIC for the model without it is 258.93. The benefits of having the term in the model are outweighed by the difficulties that a more complex model brings.

Assessing risk of relapse:

Odds ratios:

AGE CLS0	0.972115
AGE CLS1	1.008929
AGE CLS2	0.940599
MAXDEPTH	1.05129
SIZE PELLYMPH0	1.03787
SIZE PELLYMPH1	0.891136

An odds ratio > 1 indicates occurrence of an event indicating a positive association, with the ratio < 1 it indicates a higher number for the predictor (negative association). Furthermore, we can see that age plays a key role when Cls has positive level 1, and that size plays a key role when pellymph has level 0. Both reflect moderate positive associations.

C. Machine Learning

We can use machine learning to train our model on some part of the available data and test for the accuracy of the model on the part of the data. We do this by splitting our data into Training data sets (where 80 percent of the data serves to train the model) and Test data sets (where 20 percent of our data will be used to make predictions).

```
# Train/Set Test
set.seed(1234)
create_train_test <- function(data, size = 0.8, train = TRUE) {
  n_row = nrow(data)
  total_row = size * n_row
  train_sample <- 1: total_row
  if (train == TRUE) {
    return (data[train_sample, ])
  } else {
    return (data[-train_sample, ])
  }
}

# split the data between a train set and a test set
data_train <- create_train_test(cervical_cancer_data_clean_final, 0.8, train = TRUE)
data_test <- create_train_test(cervical_cancer_data_clean_final, 0.8, train = FALSE)
```

We can predict on our test dataset with `predict()`. We also make use of a confusion matrix to count the number of times True instances are classified as False, and use this table to evaluate for model accuracy.

```
# Predicted values
predict <- predict(cervical_cancer.fit_selected, data_test, type = 'response')

# confusion matrix to count the number of times True instances are classified as False
table_cancer <- table(data_test$RECURRN_Y_or_N, predict > 0.5)
table_cancer

##
##      FALSE
##  0    122
##  1      3

# Model accuracy calculated by summing the true pos + true neg over the total obervation
accuracy_Test <- sum(diag(table_cancer)) / sum(table_cancer)
accuracy_Test

## [1] 0.976
```

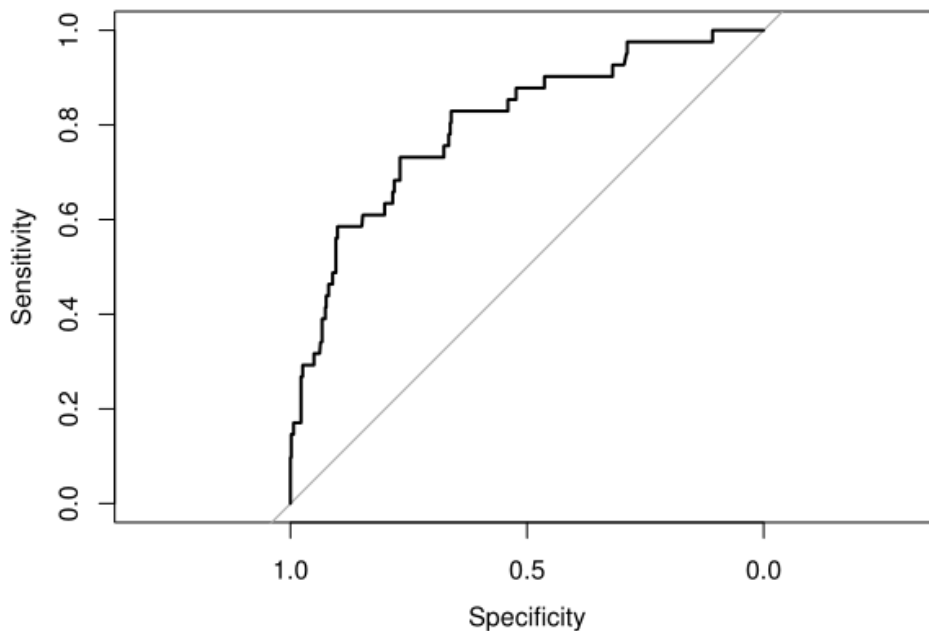
Therefore our model is 97.6% accurate.

D. ROC

The Receiving Operating Characteristic (ROC), measures the classifier's performance using the proportion of positive data points correctly considered as positive (True Positive Rate or Sensitivity) and the proportion of negative data points that are mistakenly considered as positive (False Positive Rate or Fall-out).

We get the predicted values of our selected model of response type and compare it with actual recurrence values (RECURRN_Y_or_N) using the ROC method and generate the ROC plot.


```
# Plot ROC
plot(roc1)
```



```
# Area under curve
roc1$auc
```

```
## Area under the curve: 0.8017
```

The area under the curve is 0.80 meaning our model is decent enough.

E. Plots

We are interested in a patient's age at time of diagnosis vs. whether or not they relapse.

```
# Age vs Relapse
age_vs_relapse <-
  cervical_cancer_data_clean_final %>%
  count(AGE, RECURRN_Y_or_N) %>%
  mutate(AGE=factor(AGE)) %>%
  ggplot(aes(AGE, n, fill=RECURRN_Y_or_N))+
  geom_col()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=10,
                                     size = 7))+
  labs(subtitle = 'Distribution of Records of Relapse According to Age') +
  geom_vline(aes(xintercept=median(as.numeric(AGE), na.rm = TRUE)), color="red", linetype="solid")+
  scale_x_discrete(breaks = seq(0, 100, by = 5))

cowplot::plot_grid(age_vs_relapse)
```

Similarly, we will look at patients' CLS classification vs. relapse.

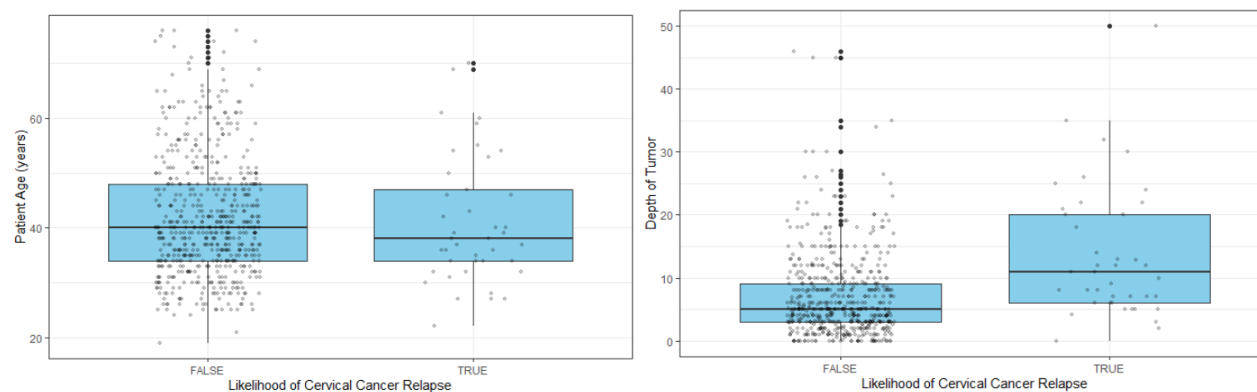


Interpretations:

Age for Relapse v Non-Relapse Patients

- The majority of patients observed were between the ages of 35 and 50
- No significant difference between relapse and non-relapse patients
- Huge disparity exist between the means and variability of relapse and non-relapse patients
- Relapse patients has larger tumor sizes upon diagnosis making it a important factor

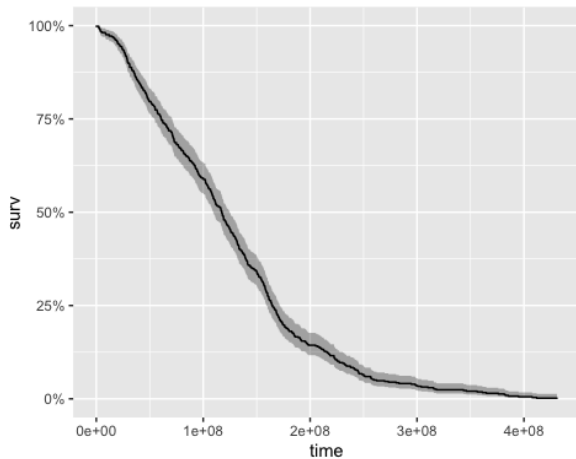
We also create a box plot of likelihood of cervical cancer relapse vs. Age and again vs. depth of tumor (MAXDEPTH).



Thus the likelihood of relapsing is higher for patients between the ages of 35-54 years old, with the median being around 38 years old. We also see this likelihood is higher for patients with depth of tumor between 6-20mm, with the median being around 12mm.

For MAXDEPTH, we can recognize that the means are similar. However, the missing values in the relapse group may have affected the outcome. Furthermore, the dissimilar boxplots are likely due to outliers.

SURVIVAL PLOT



- The importance of using Survival analysis is to test the probability of relapse over time
- The survival curve goes to 0, so it has a poor survival (low likelihood of relapse) as it shows the likelihood decreases over time

V. Conclusions and Final Thoughts

Goal: Categorize risk of relapse by available attributes.

Risk of relapse based on model:

- **Low risk (~3%):** Older with CLS=0 or 2, low MAXDEPTH, large size with PELLYMPH=1
- **Moderate risk (~10%):** Any age, CLS=1, mid-MAXDEPTH, small size with PELLYMPH=0
- **High risk(~30%):** Younger with CLS=0 or 2, high MAXDEPTH, large size with PELLYMPH=0

Note: baseline risk of relapse is 6% for the sample (included for the model). This is much lower than the typical 20% risk of relapse.

Future work

It would be resourceful to check the importance of covariates when separating the response variable (non-relapse, relapse) before specific time and after that period of time. As well, it would have been beneficial to try other statistical techniques such as the use of trees instead of explanatory tools.

Limitations

Additionally, we find that the nature of the study is on the assumption that prediction of relapse would be done right after surgery, and variables observed after surgery were not taken into account. These were the status of patients at the last follow-up date and if patients received radiation.

I. References

All references are properly cited, both in the text and at the end of the report, in the APA (6th Edition) citation style.

Chen, M., Jin, Y., Bi, Y., Li, Y., Shan, Y., & Pan, L. (2015). Prognostic significance of lymphovascular space invasion in epithelial ovarian

cancer. *Journal of Cancer*, 6(5), 412–419. <https://doi.org/10.7150/jca.11242>.