Factors Affecting Relapse of Cervical Cancer

STAC51 Final Case Study Winter 2023

Sanskriti Kanagala, Jerry Dang, Steven Pham and Priyal Bhasin

2023-04-10

Sanskriti(1006779540) - Exploratory Data Analysis and Model Validation
Jerry(1005838685) - Model Building, Model Selection and Conclusion
Steven - Background research, Presentation and Limitations
Priyal(1007311703) - Model Validation and Presentation

Background Information

Cervical cancer is a type of cancer that occurs in the cells of the cervix, most commonly caused by the human papillomavirus (HPV) through a sexually transmitted infection ("Cervical cancer", 2022). According to the World Health Organization (WHO), there were approximately 604,000 new cases of cervical cancer in 2020, and the disease is responsible for about 342,000 annual deaths ("Prevent cervical cancer", 2023). The relapse of cervical cancer is an important issue that requires studying, and it was found that there may be certain variables contributing to relapse. By analyzing these variables, we can identify risk factors for patients that have already received a treatment for the disease and predict the probability of the cancer coming back.

It was found that these variables were: patient number, surgery date, if the patient received radiation therapy or not, the age of the patient, the presence of capillary lymphatic spaces, the status of the disease, cell differentiation, histolog, the remaining disease after surgery, the depth of the tumor, the date of the reoccurrence of the disease, the size of the tumor, and the follow-up date. The analyses in this report are as follows: First, the individual variables of the dataset will be analyzed. Then, model selection methods will be applied to select a suitable regression model that will assist in predicting the probability of relapse of cervical cancer, and various model diagnostics will be applied. Finally, the findings of the report will be summarized, and possible limitations of the study will be addressed.

Research Question(s)

For this study, the study questions were figuring out which of the given variables are important in predicting the relapse of cervical cancer, as well as creating a method for classifying patients based on their individual risk of relapse.

Load the DataSet

```
data = read_excel("cervical_cancer.xls")
```

Exploratory Data Analysis

The Data

Our dataset contains the following variables:
-MRNO : Patient Number
-SURGDAT : Sugery Date
-ADJ_RAD : Received radiation therapy (0 = no, 1 = yes)
-AGE_1 : Patient Age
-CLS_1 : Capillary Lymphatic Spaces (0 = negative,1 or 2 = positive – 2 means more positive cells)
-DIS_STA : Disease Status (0 = no disease, 1= alive + disease, 2 = dead + disease, 3 = dead + complications (disease present), 4 = dead + complications (disease absent), 5 = dead (unrelated causes)
-GRAD_1 : Cell differentiation (1 = better, 2 = moderate, 3 = worst, 0 = missing value)
-HISTOLOG : Ranging from 0 to 6, stands for the type of histology (determined by the pathologist) of the cancer cells
-MARGINS : Disease remaining after surgery (0 = clear, 1 = para-vaginal area, 2 = vaginal area, 3 = both)
-MAXDEPTH_1 : Depth of Tumor (mm)
-PELLYMPH_1 : (0 = negative, 1 = positive)
-RECURRN1 : Date of reoccurrence of disease

-SIZE_1 : Size of tumor (mm) upon diagnosis
-FU_DATE : Follow-up date

If there is recurrence of cancer, a date is recorded otherwise there is no entry recorded. To make analysis easier, we decided to convert recurrence column to relapse column with "1" when there is a relapse of cancer and "0" when there is no relapse of cancer

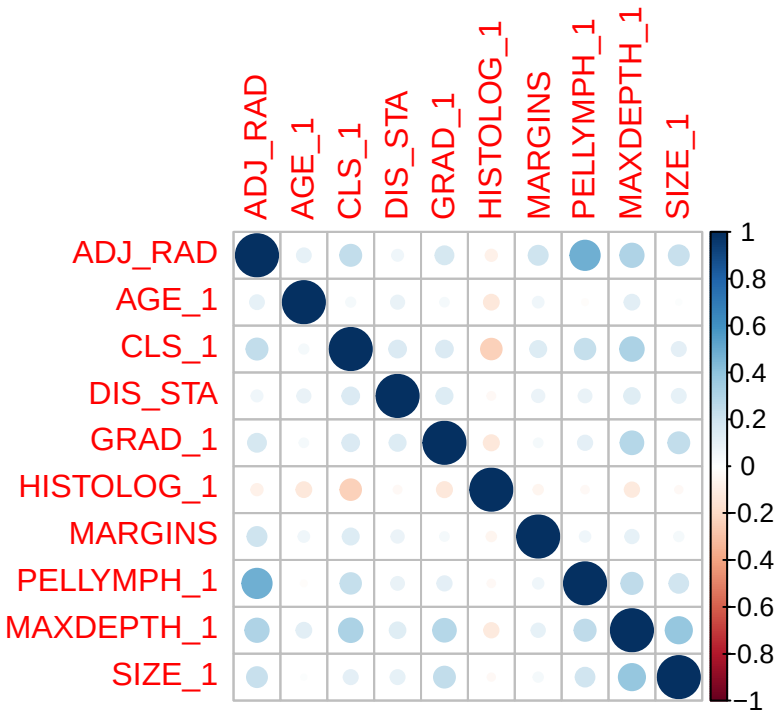
For our analysis, we will not be taking into consideration FU_DATE and SURGDATE as they are not significant factors to predict relapse of cancer.

```
data = data %>% mutate(Relapse = case_when(!is.na(data$RECURRN1) ~ 1,
                                           TRUE ~ 0))
```

We found that there are lot of missing values or na values in our dataset. Removing those values that contains NA in them. In ADJ_rad we see that max is 4 which does not match with our description of data, so we decided to remove that observations from our data.

```
df = data %>% dplyr::select(!starts_with("REC"))
df = df[complete.cases(df), ]
df = df %>% filter(!ADJ_RAD>1)
```

Correlation plot for Categorical Variables

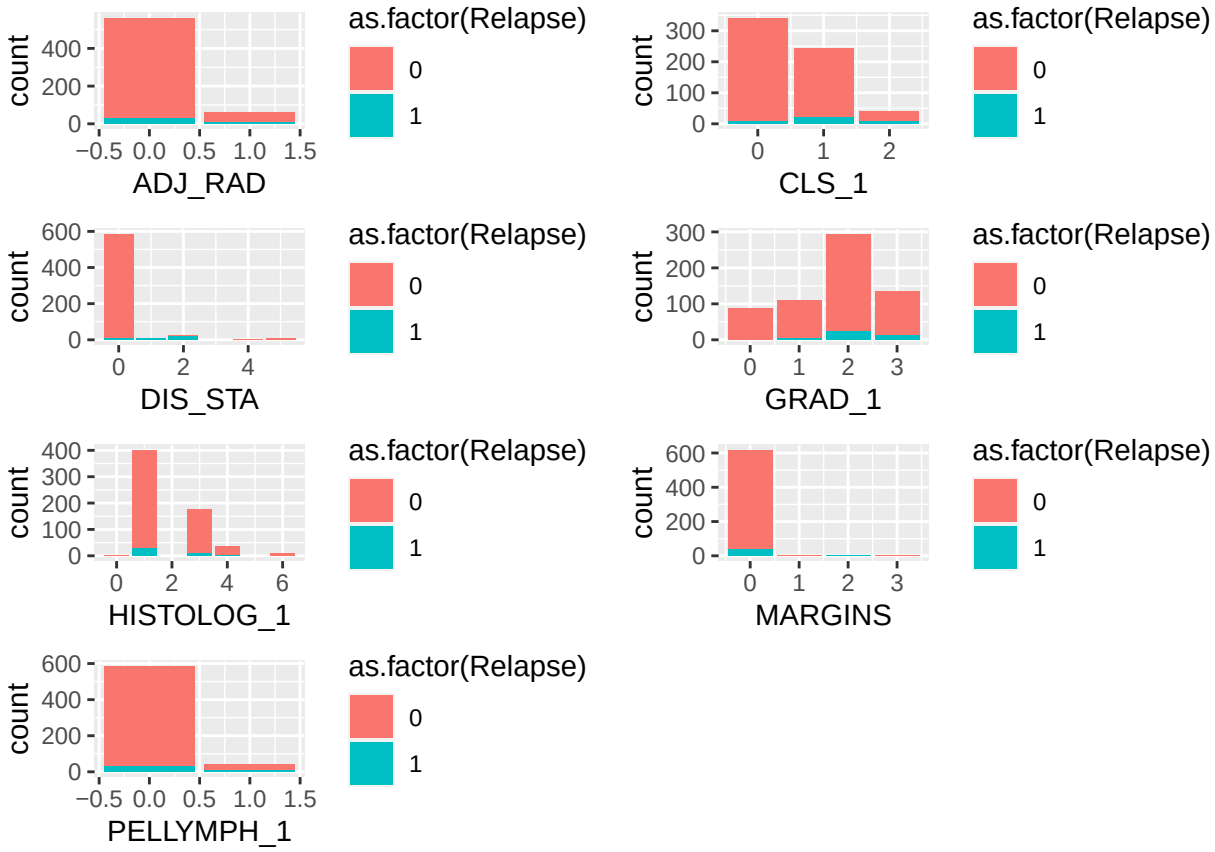



```
##                ADJ_RAD       AGE_1       CLS_1     DIS_STA      GRAD_1
## ADJ_RAD     1.00000000  0.10781644  0.24751835  0.06817011  0.17976714
## AGE_1       0.10781644  1.00000000  0.04748652  0.09723163  0.04209166
## CLS_1       0.24751835  0.04748652  1.00000000  0.15785668  0.15644435
## DIS_STA     0.06817011  0.09723163  0.15785668  1.00000000  0.14469238
```

```
## GRAD_1       0.17976714  0.04209166  0.15644435  0.14469238  1.00000000
## HISTOLOG_1 -0.07177537 -0.12309954 -0.23728233 -0.03434127 -0.12764610
## MARGINS      0.20174365  0.06442542  0.14374000  0.08553784  0.04250065
## PELLYMPH_1   0.48014951 -0.01966405  0.23446245  0.09270200  0.11376457
## MAXDEPTH_1   0.30372866  0.12338446  0.31057670  0.13728625  0.28929219
## SIZE_1       0.22201351  0.01516203  0.11068605  0.10906076  0.24543860
##              HISTOLOG_1     MARGINS  PELLYMPH_1 MAXDEPTH_1      SIZE_1
## ADJ_RAD    -0.07177537  0.20174365  0.48014951  0.3037287  0.22201351
## AGE_1      -0.12309954  0.06442542 -0.01966405  0.1233845  0.01516203
## CLS_1      -0.23728233  0.14374000  0.23446245  0.3105767  0.11068605
## DIS_STA    -0.03434127  0.08553784  0.09270200  0.1372862  0.10906076
## GRAD_1     -0.12764610  0.04250065  0.11376457  0.2892922  0.24543860
## HISTOLOG_1  1.00000000 -0.05025874 -0.03200655 -0.1153369 -0.03013871
## MARGINS    -0.05025874  1.00000000  0.06166294  0.1001509  0.04942014
## PELLYMPH_1 -0.03200655  0.06166294  1.00000000  0.2504058  0.19697545
## MAXDEPTH_1 -0.11533690  0.10015093  0.25040580  1.0000000  0.38417138
## SIZE_1     -0.03013871  0.04942014  0.19697545  0.3841714  1.00000000
```

There is no significant problem of multi-correlation between our variables. We will need to use all the variables/predictors in our analysis.

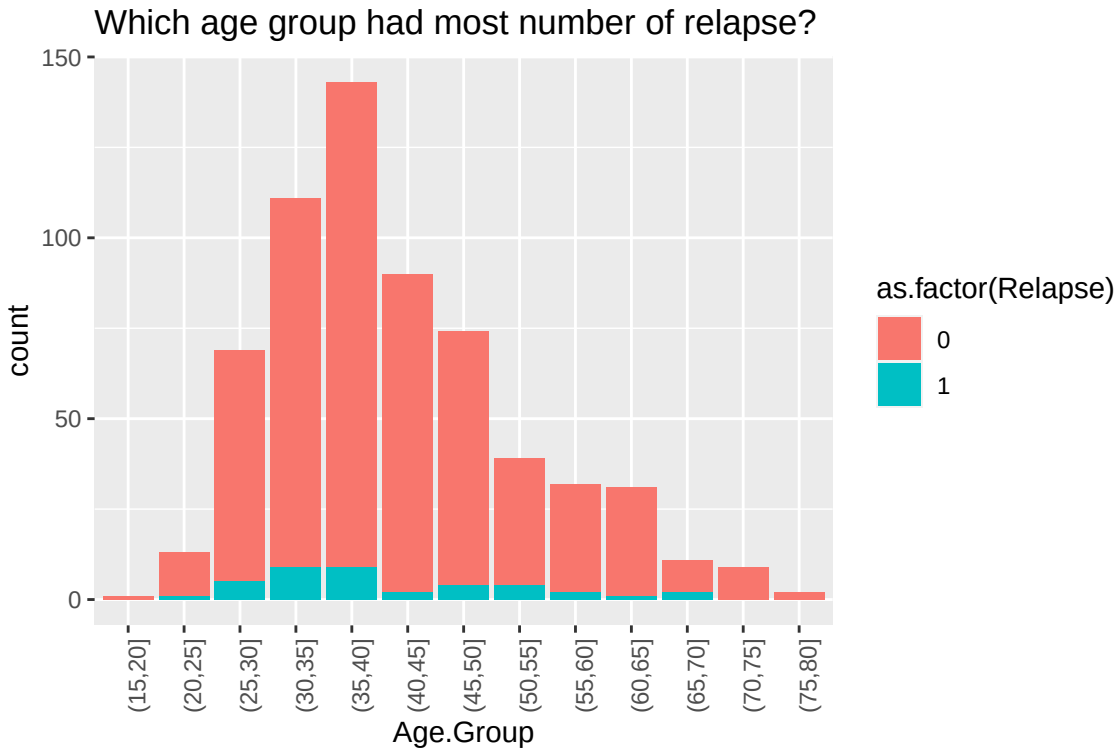
Analysis of Categorical Variables



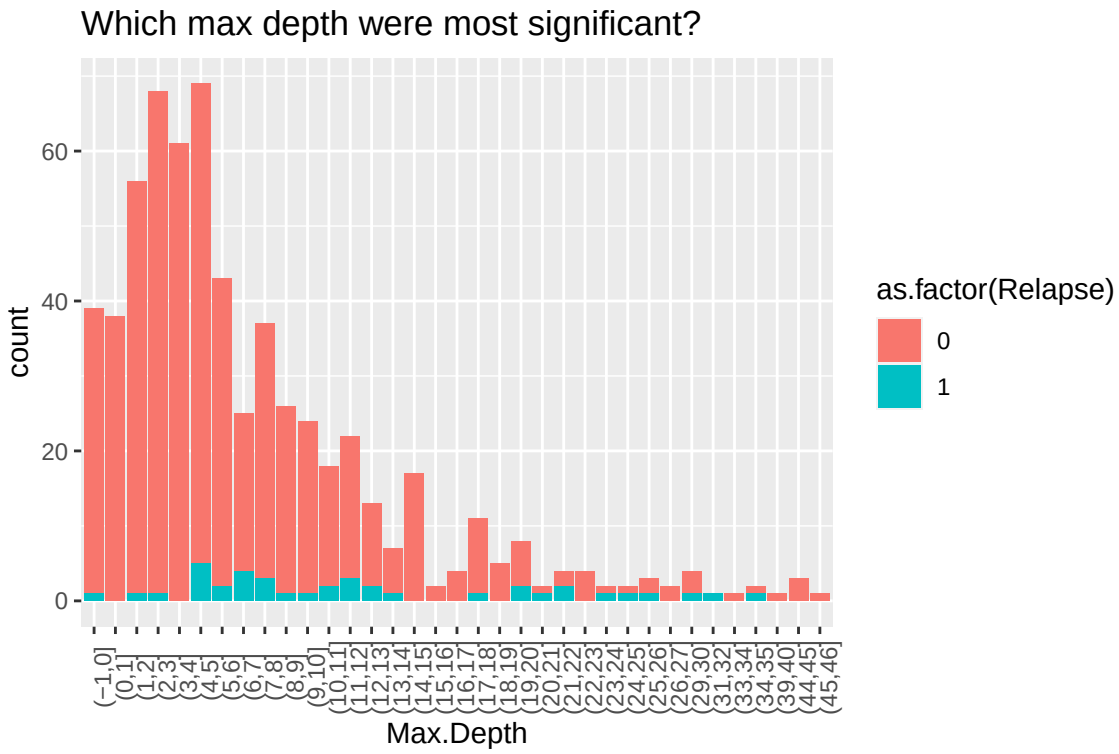
We notice that there is higher proportion of no-relapse among the patients, we will need to do further analysis using stepAIC to see which categorical variables can affect the relapse of cancer.

Analysis of Quntative Variables

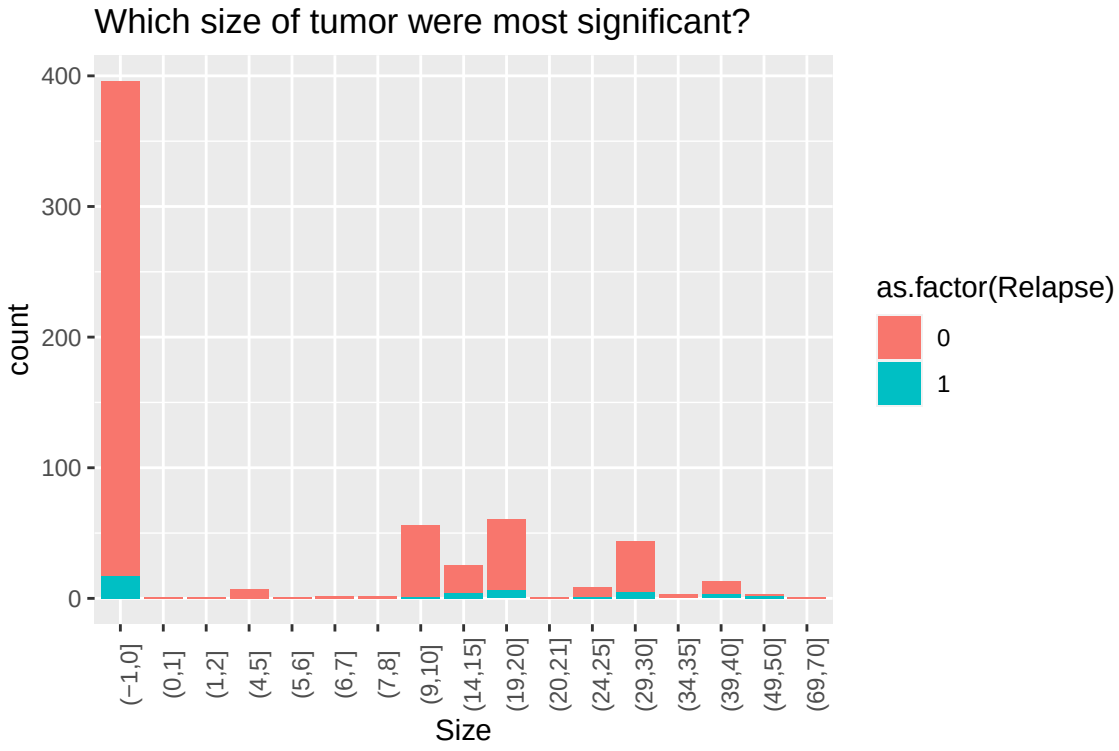
The following graphs will try to answer 3 questions:
* Which age group had most number of relapse?
* What is the max tumor depth and how it relates to relapse of cancer?
* What is the most prominent size of tumor and how it relates to relapse of cancer?



It looks like that Age.Group between 30 to 40 had most number of relapse among others.We can also see that the age group from 35 to 40, also have the highest number of cases for cervical cancer.



The most common tumor depth was found to be between 4mm to 5mm, and we can also see that this group shows most number of relapse of cancer, compared to other groups.



For size variable we need to take into consideration that some of them are labeled as 0 due to lack of consistency in coding and that's why it is also highest in our data set.

Model Analysis

```
A <- as.factor(df$ADJ_RAD)
C <- as.factor(df$CLS_1)
D <- as.factor(df$DIS_STA)
G <- as.factor(df$GRAD_1)
H <- as.factor(df$HISTOLOG_1)
M <- as.factor(df$MARGINS)
P <- as.factor(df$PELLYMPH_1)
R <- as.factor(df$Relapse)

D1 <- as.numeric(df$DIS_STA==1)
D2 <- as.numeric(df$DIS_STA==2)
```

Main Effect Model

First, we fit our full model with all our chosen data. We used all the variables mentioned from before except for the dates of surgery and follow-ups since it was not relevant for our purposes to build the main effect model.

```
full.model <- glm(R ~ A + AGE_1 + C + D + G + H + M + MAXDEPTH_1 + P + SIZE_1, data = df, family = bino
summary(full.model)
```

```
##
## Call:
## glm(formula = R ~ A + AGE_1 + C + D + G + H + M + MAXDEPTH_1 +
##     P + SIZE_1, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5520  -0.1899  -0.1168  -0.0459   3.4935
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.446e+01  1.152e+04  -0.003  0.99761
## A1          -6.713e-01  9.788e-01  -0.686  0.49280
## AGE_1       -3.869e-02  2.978e-02  -1.299  0.19387
## C1           6.631e-01  7.036e-01   0.942  0.34596
## C2           8.986e-01  1.016e+00   0.885  0.37636
## D1           7.400e+00  1.453e+00   5.094 3.51e-07 ***
## D2           6.045e+00  9.163e-01   6.597 4.18e-11 ***
## D4          -1.694e+01  1.231e+04  -0.001  0.99890
## D5          -1.649e+01  6.676e+03  -0.002  0.99803
## G1           1.406e+01  1.821e+03   0.008  0.99384
## G2           1.610e+01  1.821e+03   0.009  0.99295
## G3           1.535e+01  1.821e+03   0.008  0.99327
## H1           1.474e+01  1.137e+04   0.001  0.99897
## H3           1.517e+01  1.137e+04   0.001  0.99894
## H4           1.563e+01  1.137e+04   0.001  0.99890
## H6          -1.690e-01  1.232e+04   0.000  0.99999
## M1          -1.804e+01  1.163e+04  -0.002  0.99876
```

```
## M2          -4.319e-01  3.669e+00  -0.118  0.90630
## M3          -1.738e+01  1.773e+04  -0.001  0.99922
## MAXDEPTH_1   8.426e-02  2.904e-02   2.902  0.00371 **
## P1           4.397e-01  9.670e-01   0.455  0.64929
## SIZE_1       2.436e-02  1.986e-02   1.226  0.22015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 291.90  on 624  degrees of freedom
## Residual deviance: 106.04  on 603  degrees of freedom
## AIC: 150.04
##
## Number of Fisher Scoring iterations: 19
```

There are some variables which are non significant in our model, so we decided to use backwards elimination with stepAIC to find a better model for our data.

```
reduced.model <- stepAIC(full.model)
```

```
summary(reduced.model)
```

```
##
## Call:
## glm(formula = R ~ AGE_1 + D + G + MAXDEPTH_1, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6583  -0.1999  -0.1370  -0.0558   3.5391
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -18.13071 1136.38973  -0.016 0.987271
## AGE_1         -0.04310    0.02846  -1.514 0.129903
## D1             7.42249    1.41954   5.229 1.71e-07 ***
## D2             6.12535    0.85669   7.150 8.68e-13 ***
## D4           -15.84386 7592.38658  -0.002 0.998335
## D5           -15.66253 4090.61596  -0.004 0.996945
## G1            13.25702 1136.38990   0.012 0.990692
## G2            15.41964 1136.38933   0.014 0.989174
## G3            14.69903 1136.38946   0.013 0.989680
## MAXDEPTH_1     0.09122    0.02512   3.632 0.000282 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 291.90  on 624  degrees of freedom
## Residual deviance: 109.99  on 615  degrees of freedom
## AIC: 129.99
##
## Number of Fisher Scoring iterations: 18
```

Using stepAIC for backwards elimination, the resulting predictors from the reduced model without interaction terms are age, disease status, grad which is the state of the cervix cell, depth of the tumor and size of the tumor upon diagnosis glm(R ~ AGE_1 + D + G + MAXDEPTH_1) AIC: 129.99

Model with Interaction Term

To make the model better, we took some of the predictors from the reduced model and put the quantitative variables as the main interaction predictors. There were other ways to get a better model but after several tries with other models, we decided to continue with this model based off the reduced model from the main effect model.

```
# Interaction model - < 0.05 (significant interaction, otherwise no significance)
int.model <- glm(R ~ (AGE_1 + MAXDEPTH_1) * G + D, data = df, family = binomial)
reduced.model1 <- stepAIC(int.model)
```

```
summary(reduced.model1)
```

```
##
## Call:
## glm(formula = R ~ AGE_1 + MAXDEPTH_1 + G + D + AGE_1:G, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3209  -0.2006  -0.1500  -0.0114   3.6064
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.073e+01  7.211e+03  -0.003 0.997707
## AGE_1       -3.037e-03  1.674e+02   0.000 0.999986
## MAXDEPTH_1   9.086e-02  2.589e-02   3.509 0.000449 ***
## G1           5.343e+00  7.211e+03   0.001 0.999409
## G2           1.649e+01  7.211e+03   0.002 0.998176
## G3           2.138e+01  7.211e+03   0.003 0.997634
## D1           9.749e+00  2.284e+00   4.268 1.97e-05 ***
## D2           7.208e+00  1.295e+00   5.567 2.59e-08 ***
## D4          -1.674e+01  1.251e+04  -0.001 0.998933
## D5          -1.616e+01  6.468e+03  -0.002 0.998006
## AGE_1:G1     1.784e-01  1.674e+02   0.001 0.999149
## AGE_1:G2    -4.809e-03  1.674e+02   0.000 0.999977
## AGE_1:G3    -1.554e-01  1.674e+02  -0.001 0.999259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 291.90  on 624  degrees of freedom
## Residual deviance: 102.11  on 612  degrees of freedom
## AIC: 128.11
##
## Number of Fisher Scoring iterations: 19
```

glm(R ~ AGE_1 + MAXDEPTH_1 + G + D + AGE_1:G) AIC = 128.11 Using stepAIC for backwards elimination, the reduced model with interaction terms includes the age, depth of tumor, state of cell, disease status, and an interaction term age:state of the cell.

The reduced interaction model has a slightly lower AIC value compared to othe main effect reduced model without interaction, we decided to go with the model with an interaction term.

Model Validation/Diagnostics

Splitting data into train and test data

For the purposes of model validation, we split our dataset into 20-80 randomly selected observations used to create the training set and the other used as our validation set.

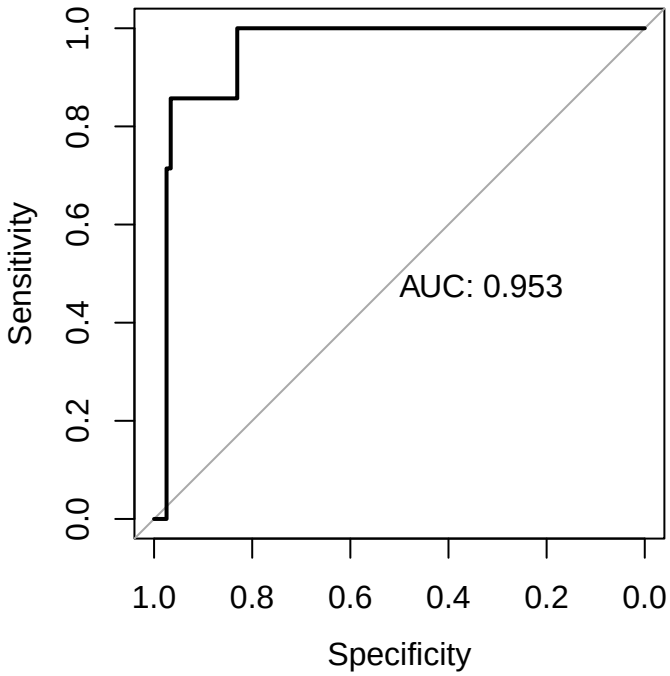
```
set.seed(1006779540)
n = nrow(df)
dv.sample <- sample(1:nrow(df),floor(nrow(df)*.8),replace = FALSE)
df.train <- df[dv.sample,]
df.test <- df[-dv.sample,]
```

Model on training

```
model1 <- glm(Relapse ~ AGE_1 + MAXDEPTH_1 + GRAD_1 + DIS_STA + AGE_1:GRAD_1, data = df.train , family =
```

ROC curve and AUC

```
y_hat_int = predict(model1,type = "response",newdata = df.test)
mean1 <- mean(y_hat_int)
roc_logit = roc(df.test$Relapse~y_hat_int,plot = TRUE,print.auc = TRUE)
```



The AUC = 0.9528 which is very close to 1, the model fits the data very well.

Hosmer-Lemeshow Test

Since our data is ungrouped data, we need to perform hosmer-lemeshow test.

```
hoslem.test(model1$y, fitted(model1), g=7)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model1$y, fitted(model1)
## X-squared = 5.6803, df = 5, p-value = 0.3386
```

Since p-value > 0.05, we fail to reject the null hypothesis and can say that model with interaction terms fit the data well.

Confusion Matrix

```
predictions <- ifelse(y_hat_int>0.09, 1, 0)
confusionMatrix(as.factor(predictions), as.factor(df.test$Relapse))
```

```
## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction   0   1
##          0 106   1
##          1  12   6
##
##                Accuracy : 0.896
##                  95% CI : (0.8287, 0.9435)
##     No Information Rate : 0.944
##     P-Value [Acc > NIR] : 0.989512
##
##                   Kappa : 0.4344
##
##  Mcnemar's Test P-Value : 0.005546
##
##             Sensitivity : 0.8983
##             Specificity : 0.8571
##          Pos Pred Value : 0.9907
##          Neg Pred Value : 0.3333
##              Prevalence : 0.9440
##          Detection Rate : 0.8480
##    Detection Prevalence : 0.8560
##       Balanced Accuracy : 0.8777
##
##        'Positive' Class : 0
##
```

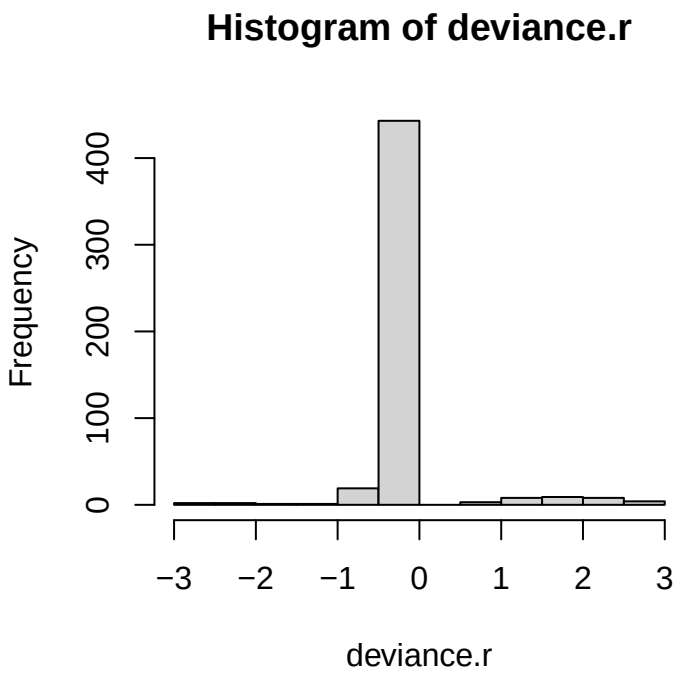
We can see that accuracy, sensitivity and specificity all have very high value and pretty close to 1, thus making our model pretty accurate.

Residual Diagnostics

```
deviance.r <- rstandard(model1)
mu.fit <- fitted(model1)
deviance.r[which(abs(deviance.r)>3)]
```

```
## named numeric(0)
```

```
hist(deviance.r)
```



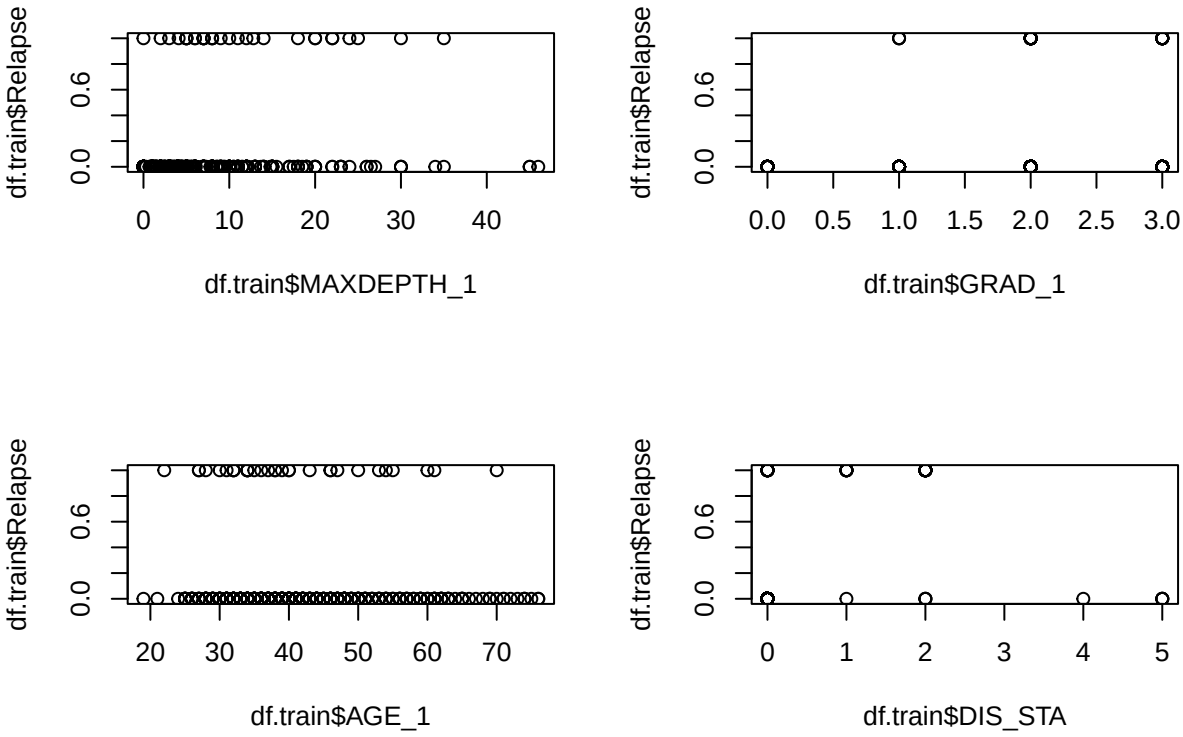
The majority of the values are between -2 and 2. Therefore we can conclude that selected model fits data well.

Limitations

In the study, there were a large number of variables included in the dataset. However, the study may be limited due to the absence of certain variables, such as side effects that the patient received following treatment, or the overall health of the patient. In addition, the dataset includes patient information from approximately 3 decades ago. Our predictions that we make based on this information may not accurately predict relapse considering the exponential advancements made in medical technology over the years. Some of the values included in the dataset were also empty (NA values) and had to be replaced.

Some of these limitations are out of our control due to the dataset being rather old and although some missing values are filled with constant values to denote that those values are indeed missing, it does limit our ability to reliably say that these results are concrete

Discussions and Conclusion



The goal of this report was to determine which factors were most significant in determining whether a cervical cancer patient would have a relapse and to classify patients as 'Low Relapse', 'Moderate Relapse', and 'High Relapse'. From our analysis, we can conclude that Age (AGE_1), the max depth (MAXDEPTH_1) of the tumor, the condition/state of the cervix cell(s) (GRAD), the disease status of the patient (DIS_STA), are factors that are most likely to affect the relapse probability of a patient diagnosed with cervical cancer. In order to categorize the patients with cervical cancer into the three categories mentioned previously, we have to look at the probability depending on each value of every predictor. Although it is difficult to classify very precisely since there are quite a few variables still in consideration, we can see that the age group of 30-40 years old, and a tumor max depth from 1-5mm have had the most relapses, so patients with those statistics can be classified as 'High Relapse'. 'Moderate Relapse' can be applied to age groups around 20-30 and 40-50 and 'Low Relapse' for all other age groups. A disease status of 1, and 2 can be considered in 'High Relapse' and 'Moderate Relapse' categories and even with no disease, we would classify this as 'Low Relapse' but the chance of relapsing is not zero. Looking at the model, the coefficients most contributing to the probability of relapse seems to be the disease status, and status of the cervix cell while keep all other predictors constant, therefore we can conclude that the classification of a patient have cervical cancer depends to a certain extent on those factors as well. Of course, the healthier the cell, the lower the classification. 1 - better to 3 - worst differentiation in cells meaning that someone with a 3 in the GRAD_1 category (cell status) would have a higher chance to be classified 'High Relapse' and someone with a 1 would be more likely to be classified 'Low Relapse'.

Appendix

All libraries used

```
library(readxl)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(pROC)
library(caret)
library(ResourceSelection)
library(corrplot)
library(MASS)
```

References

Cervical cancer. World Health Organization. (2022, February 22). Retrieved April 9, 2023, from https://www.who.int/news-room/fact-sheets/detail/cervical-cancer

Prevent cervical cancer. Centers for Disease Control and Prevention. (2023, January 24). Retrieved April 9, 2023, from https://www.cdc.gov/healthequity/features/cervical-cancer/index.html#:~:text=Cervical%20cancer%20is%20the%20fourth,cancer%20and%20342%2C000%20deaths%20worldwide.