A Categorical Approach in Predicting Recurrence of Cervical Cancer in Canadian Women

Colin Mozo(1004345176), Harrick Cheong(1004403969), Lavan Rajendra(1004019251), Zhiye Luan(1004282867)

11 April 2021

Contents

Abstract

Introduction

Cervical cancer is one of the prominent chronic diseases diagnosed in Canadian women today. Currently, the disease occurs in 1 out of every 168 women in Canada, where approximately 65% of the diagnosed women are reported to have recovered from the disease (Canada 2009). It is well established that recurrence of cancer among patients whose cancer has gone into remission poses significant threat to the survival of these patients which more prominently affects older women. This study aims to determine the statistically significant main biological factors recorded upon first diagnosis in predicting the recurrence of the cancer. In this context, recurrence of the disease refers to the disease being active again and remission of the disease refers to the cancer symptoms being reduced.

Data

The data was gathered from the Toronto Sunnybrook Health Science Center with data originating from cases ranging from 1984 to 2002 which recorded key parameters of interest in cancer cases including but not limited to tumor size, cell type, grade, depth of invasion and lymph node status.

Objective

The main objective of this study is to determine which traits measured in these cancer patients play an important role in predicting the recurrence of cancer. The total recorded traits of each case provided by the hospital are of following:

- Patient number
- Diagnosis date
- Has had radiation therapy
- Age
- Status of Capillary lymphatic spaces
- State of disease
    - no evidence of disease
    - alive with disease
    - dead of disease
    - dead of complications (disease present)
    - dead of complications (disease absent)
    - dead of unrelated causes
- State of cell differentiation
- Histology of cancer cells
- Identifiable disease after surgery
- Depth of cancerous tumor
- Pelvis lymph node involvement
- Date of recurrence
- Size of tumor upon diagnosis
- Most recent follow-up date.

Methods

Data Cleanup

The following data provided by Sunnybrook hospital is loaded into a table in `R` via the `read_excel` function from the library `readxl`.

```
##       MRNO        SURGDATE                         ADJ_RAD           AGE_1
##  Min.   :  1   Min.   :1984-05-25 00:00:00   Min.   :0.0000   Min.   :19.00
##  1st Qu.:227   1st Qu.:1989-10-16 00:00:00   1st Qu.:0.0000   1st Qu.:34.00
##  Median :453   Median :1993-04-20 00:00:00   Median :0.0000   Median :40.00
##  Mean   :453   Mean   :1993-03-05 18:00:23   Mean   :0.1665   Mean   :42.13
##  3rd Qu.:679   3rd Qu.:1996-07-24 00:00:00   3rd Qu.:0.0000   3rd Qu.:49.00
##  Max.   :905   Max.   :2001-08-02 00:00:00   Max.   :4.0000   Max.   :76.00
##                                              NA's   :22
##      CLS_1           DIS_STA           GRAD_1        HISTOLOG_1
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.000
##  Median :0.0000   Median :0.0000   Median :2.000   Median :1.000
##  Mean   :0.5132   Mean   :0.1653   Mean   :1.724   Mean   :1.927
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :2.0000   Max.   :5.0000   Max.   :3.000   Max.   :6.000
##  NA's   :112      NA's   :22                       NA's   :1
##     MARGINS          MAXDEPTH_1       PELLYMPH_1
##  Min.   :0.00000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.: 2.500   1st Qu.:0.00000
##  Median :0.00000   Median : 5.000   Median :0.00000
##  Mean   :0.04646   Mean   : 7.133   Mean   :0.06298
##  3rd Qu.:0.00000   3rd Qu.:10.000   3rd Qu.:0.00000
##  Max.   :3.00000   Max.   :50.000   Max.   :1.00000
##  NA's   :1         NA's   :134
##     RECURRN1                       SIZE_1          FU_DATE
##  Min.   :1986-04-28 00:00:00   Min.   : 0.000   Min.   :1987-05-16 00:00:00
##  1st Qu.:1990-12-09 18:00:00   1st Qu.: 0.000   1st Qu.:1995-03-16 00:00:00
##  Median :1994-02-06 00:00:00   Median : 0.000   Median :1998-01-07 00:00:00
##  Mean   :1993-12-17 21:31:45   Mean   : 8.812   Mean   :1997-04-09 00:46:17
##  3rd Qu.:1996-06-22 06:00:00   3rd Qu.:18.750   3rd Qu.:1999-09-14 12:00:00
##  Max.   :2000-08-18 00:00:00   Max.   :70.000   Max.   :2001-09-25 00:00:00
##  NA's   :837                   NA's   :31       NA's   :34
```

Upon first glance it is evident that this data was not originally intended for use in the prediction of cancer relapse which leads to some of the data being unsuitable for modeling whether it was due to errors in recording data or errors caused by software comparability in translating this dataset from one language to another. It is also evident to note that not every column is suited to be an explanatory variable. The following variables will be omitted for lack of relevance/practicality as predictors.

1.) **`SURDATE`** - For this study, we are not interested in the date of first cancer diagnosis but rather the physical traits of the patient in question that lead to their relapse.

2.) **`FU_DATE`** - While the latest follow-up date may provide some insight to the prediction of a patient's survival rate for their cancer relapse, this study is instead mainly focused on the prediction of relapse.

3.) **`MRNO`** - Although this specifies the patient number for each patient in this study, this will not have any relevance to predicting cancer relapse.

4.) **DIS_STA** - Most of the entries do not make sense in our context of predicting relapse as the patient is considered to be deceased in half of the cases, and for the cases when the patent are not dead, knowing the according status of the patient is not helpful (e.g. How does "alive with disease" or "no evidence of disease" lead to relapse?).

The following methods will be used to prepare the data for cleanup.

We will use the package tidyverse to clean the data and omit certain observations to continue with data analysis. This is our cleaned up data set where NA values are omitted as we believe removing them will not introduce any bias towards our modeling.

```r
cc %>%
  mutate(recurrence = ifelse(is.na(RECURRN1), 0, 1),
         histology = ifelse(!(cc$HISTOLOG_1 %in% c(1,3)), 5, cc$HISTOLOG_1),
         ADJ_RAD = ifelse(ADJ_RAD >= 1, 1, 0)) %>%
  drop_na(FU_DATE) %>%
  drop_na(-RECURRN1) %>%
  subset(GRAD_1 != 0) %>%
  mutate(diffs = FU_DATE - RECURRN1) %>%
  filter(diffs >= 0 | recurrence == 0) -> cc_clean
```

In this clean up, we created a new column called **recurrence** that is binary where **recurrence** $= 0$ means *no relapse* and **recurrence** $= 1$ means *relapse*, re-coded **ADJ_RAD** as binary where **ADJ_RAD** $= 0$ means *patient didn't received radiation therapy* and **ADJ_RAD** $= 1$ means *patient received radiation therapy*, and renamed **HISTOLOG_1** column to **histology** and also re-coded it where **histology** $= 1$ means *Squamous cell carcinoma*, **histology** $= 3$ means *adenocarcinoma*, and **histology** $= 5$ means other types of cancer cells. We omitted all the NA values in every column except **RECURRN1** since they will not influence our modeling and will not make sense in our data analysis.

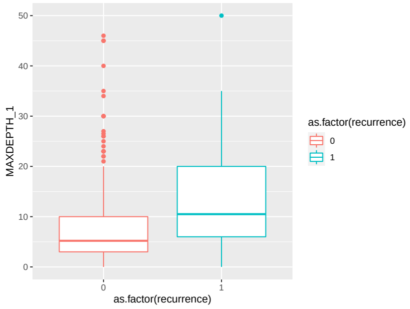
We also omitted **GRAD_1** $= 0$ because it meant a missing value. Additionally, we introduced a new column **diffs** to look at each patient and see if they had a recurrence date after their follow-up date because recurrence should be equal or before the follow-up date hence, the ones after follow-up date are considered errors/typos which we omitted.

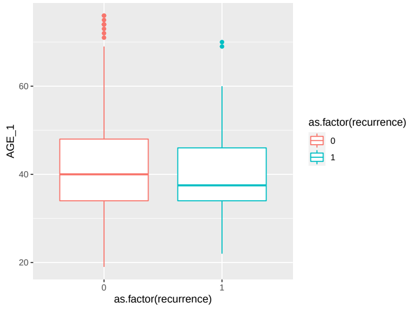
Data Diagnostics

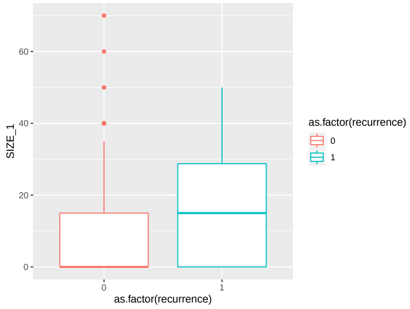
Looking at the data set *cc_clean*, this is un-grouped data and we have that recurrence is our response variable (binary) and the rest of other columns except the ones we are not considering are our explanatory variables. Looking further, we have **AGE_1**, **MAXDEPTH_1**, and **SIZE_1** as continuous variables and the rest are categorical variables (a mix of nominal and ordinal variables)

Visual Analysis

Here, we will investigate if recurrence can be predicted by our 3 continuous variables via side-by-side box plots.







Looking at the `boxplots` above, we can see that for `AGE_1`, having a recurrence or not appears to not be different between ages. Also, for `MAXDEPTH_1`, we can see that there seems to be a difference in `MAXDEPTH_1` between those who had recurrence versus who didn't. Upon further observing, box plot for `SIZE_1` appears to be different between those who a recurrence or did not.

Because we are not certain that the data is truly normal we cannot perform any mean tests. We can investigate the stat Q-Q for each data sample. This proves to be not necessary as from visual observation, it can be easily concluded that the data is not normal when `recurrence` is zero for `SIZE_1` therefore we can verify our observations through Mood's median tests which does not take into account for the normality of the data..

```
# performing Mood's median test on each subset
library(smmr)
median_test(cc.2, MAXDEPTH_1, recurrence)$test[3,]
```

```
##      what        value
## 3 P-value 0.0003038101
```

```
median_test(cc.2, SIZE_1, recurrence)$test[3,]
```

```
##      what value
## 3 P-value     1
```

```
median_test(cc_clean, AGE_1, recurrence)$test[3,]
```

```
##      what     value
## 3 P-value 0.1001916
```

Our claims appear to be correct as P-value for `MAXDEPTH_1` < 0.05 indicating that there is a statistically significant difference in `MAXDEPTH_1` between patients having a relapse or not, P-value for `AGE_1` > 0.05 meaning no difference in `AGE_1` between having a relapse or not, and P-value for `SIZE_1` > 0.05 meaning no difference in SIZE_1 between having a relapse or not.

Now, we check if the factors `AGE_1`, `MAXDEPTH_1`, or `SIZE_1` have any correlation to each other. If these predictors are strongly correlated, then we have multicollinearity among our continuous variables which can cause bias towards our modeling. We should remove predictors that have a high P-value.

```
# Using R-based correlation function to see if we have multicollineariry
(one <- cor(cc_clean$AGE_1, cc_clean$SIZE_1))
```

```
## [1] 0.0104173
```

```
(two <- cor(cc_clean$SIZE_1, cc_clean$MAXDEPTH_1))
```

```
## [1] 0.3618578
```

```
(three <- cor(cc_clean$AGE_1, cc_clean$MAXDEPTH_1))
```

```
## [1] 0.1350133
```

A correlation near 0 indicates a relationship between two variables is weak (linearly). From the results above, our continuous variables are weakly correlated to each other, hence we can include them in our model and will instead remove them if they are deem insignificant based on our modeling.

Modelling

First, we investigate our main effect model for all of the variables considered to be a valid predictor.

```
cc %>%
  mutate(recurrence = ifelse(is.na(RECURRN1), 0, 1),
         histology = ifelse(!(cc$HISTOLOG_1 %in% c(1,3)), 5, cc$HISTOLOG_1),
         ADJ_RAD = ifelse(ADJ_RAD >= 1, 1, 0)) %>%
  drop_na(FU_DATE) -> cc.n
```

```
mem <- glm(recurrence~ ADJ_RAD + AGE_1 + CLS_1 + GRAD_1 + as.factor(histology) +
      MARGINS + SIZE_1 + MAXDEPTH_1 + PELLYMPH_1, family=binomial, data = cc.n)
summary(mem)
```

```
##
## Call:
## glm(formula = recurrence ~ ADJ_RAD + AGE_1 + CLS_1 + GRAD_1 +
##     as.factor(histology) + MARGINS + SIZE_1 + MAXDEPTH_1 + PELLYMPH_1,
##     family = binomial, data = cc.n)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2340  -0.3622  -0.2518  -0.1791   2.8764
##
```

```
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)            -4.72459    0.98604  -4.791 1.66e-06 ***
## ADJ_RAD                -0.29661    0.50477  -0.588  0.55679    
## AGE_1                  -0.01623    0.01622  -1.001  0.31692    
## CLS_1                   0.92493    0.28751   3.217  0.00130 ** 
## GRAD_1                  0.57013    0.25662   2.222  0.02631 *  
## as.factor(histology)3  0.55938    0.46008   1.216  0.22405    
## as.factor(histology)5  0.20545    0.60766   0.338  0.73529    
## MARGINS                -0.15479    0.48571  -0.319  0.74996    
## SIZE_1                  0.01699    0.01229   1.383  0.16670    
## MAXDEPTH_1              0.05437    0.01866   2.914  0.00357 ** 
## PELLYMPH_1              0.66417    0.52835   1.257  0.20873    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 304.11  on 635  degrees of freedom
## Residual deviance: 257.30  on 625  degrees of freedom
##   (235 observations deleted due to missingness)
## AIC: 279.3
##
## Number of Fisher Scoring iterations: 6
```

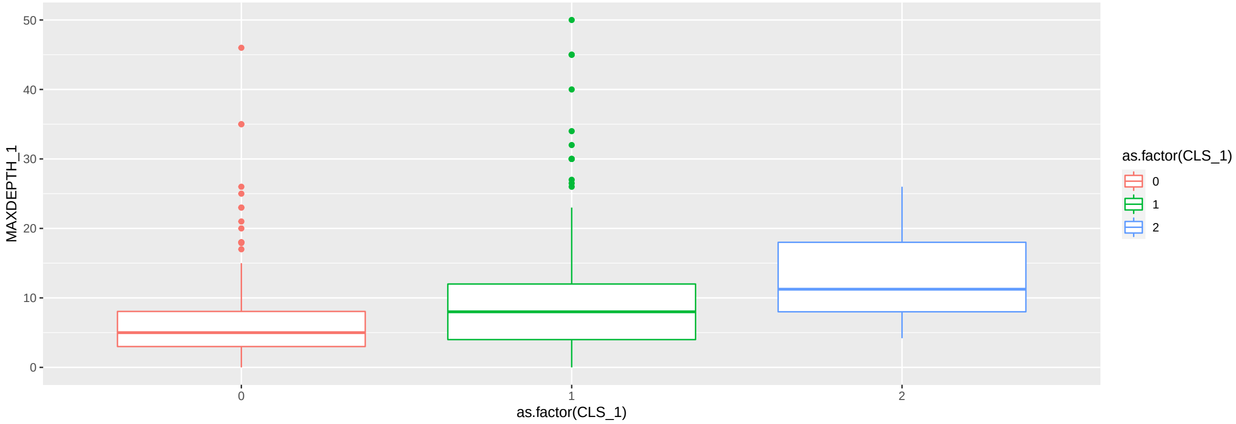
When we look at all our 871 patients, CLS_1, GRAD_1, and MAXDEPTH_1 are significant predictors. Let’s see if we can use our clean data set and see if we get the same significant predictors.

```
memr <- glm(recurrence~ADJ_RAD + AGE_1 + CLS_1 + GRAD_1 + as.factor(histology) +
             MARGINS + MAXDEPTH_1 + SIZE_1 + PELLYMPH_1, family=binomial, data = cc_clean)
summary(memr)
```

```
##
## Call:
## glm(formula = recurrence ~ ADJ_RAD + AGE_1 + CLS_1 + GRAD_1 +
##     as.factor(histology) + MARGINS + MAXDEPTH_1 + SIZE_1 + PELLYMPH_1,
##     family = binomial, data = cc_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1502  -0.3763  -0.2738  -0.2196   2.8058
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)   
## (Intercept)            -3.52992    1.08849  -3.243  0.00118 **
## ADJ_RAD                -0.33383    0.54041  -0.618  0.53676   
## AGE_1                  -0.02236    0.01719  -1.301  0.19329   
## CLS_1                   0.78295    0.29649   2.641  0.00827 **
## GRAD_1                  0.16989    0.31172   0.545  0.58575   
## as.factor(histology)3  0.38691    0.47602   0.813  0.41633   
## as.factor(histology)5  0.35507    0.61523   0.577  0.56385   
## MARGINS                -0.10168    0.49068  -0.207  0.83584   
## MAXDEPTH_1              0.05053    0.01936   2.610  0.00906 **
## SIZE_1                  0.02405    0.01269   1.895  0.05804 .
## PELLYMPH_1              0.71662    0.55852   1.283  0.19947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 275.83  on 545  degrees of freedom
## Residual deviance: 243.02  on 535  degrees of freedom
## AIC: 265.02
##
## Number of Fisher Scoring iterations: 6
```

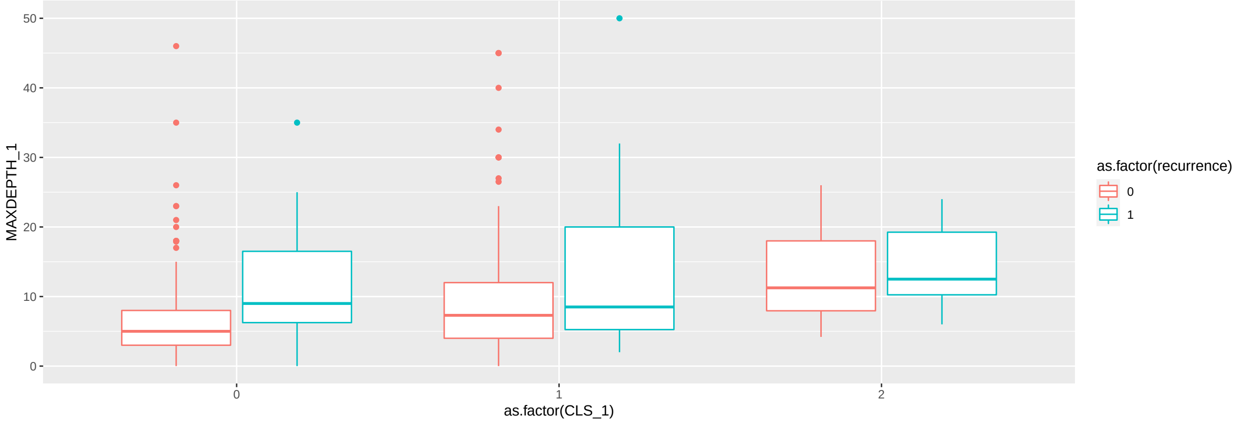
When we use our cleaned up dataset, we lose `GRAD_1` as a significant predictor. This happened because when `GRAD_1` $= 0$, it indicated a missing value and so it would be appropriate to omit this. However, we still have `CLS_1` and `MAXDEPTH_1` as significant predictors which tells us using our cleaned data set will not introduce any bias for significance of predictors. So, our significant predictors are `CLS_1` and `MAXDEPTH_1` which means that these should be in our final model to predict recurrence.

Before that, let's see if there is any interaction between `CLS_1` and `MAXDEPTH_1`.



Considering all patients regardless of whether they had a relapse or not, based on the box plot, there seems to be a difference in `MAXDEPTH_1` for each `CLS_1` group.

Let's now consider patients between relapse or not and see if there is a difference within a particular CLS_1.



Based on the boxplot, it looks like `MAXDEPTH_1` is different among those who had relapse and no relapse within each `CLS_1` group.

```
cc_clean %>% drop_na(CLS_1, MAXDEPTH_1) -> cc.3
cc.3 %>% filter(CLS_1 == 0) -> cc.cls0
cc.3 %>% filter(CLS_1 == 1) -> cc.cls1
cc.3 %>% filter(CLS_1 == 2) -> cc.cls2

median_test(cc.cls0, MAXDEPTH_1, recurrence)$test[3,]
```

```
##      what      value
## 3 P-value 0.02960688
```

```
median_test(cc.cls1, MAXDEPTH_1, recurrence)$test[3,]
```

```
##      what     value
## 3 P-value 0.3970001
```

```
median_test(cc.cls2, MAXDEPTH_1, recurrence)$test[3,]
```

```
##      what value
## 3 P-value     1
```

MAXDEPTH_1 for CLS_1 $= 0$ is different among those who had a relapse and those who didn't since the P-value is significant while MAXDEPTH_1 for CLS_1 $= 1$ and CLS_1 $= 2$ do not have a difference in MAXDEPTH_1 among those who had a relapse and those who didn't since it's P-value is nonsignificant. This shows that there is no interaction between CLS_1 and MAXDEPTH_1 since in CLS_1 $= 1$ and CLS_1 $= 2$ groups, MAXDEPTH_1 showed no difference when one patient had a relapse and one who didn't. We should also see in the final model that interaction is deemed insignificant.

Our final model.

```
sat <-glm(recurrence ~ CLS_1*MAXDEPTH_1, family = binomial, data = cc_clean)
summary(sat)
```

```
##
## Call:
## glm(formula = recurrence ~ CLS_1 * MAXDEPTH_1, family = binomial,
##     data = cc_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2398  -0.4080  -0.3004  -0.2198   2.8348
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.99991    0.43530  -9.189  < 2e-16 ***
## CLS_1             1.15109    0.45694   2.519  0.01176 *
## MAXDEPTH_1        0.09012    0.02957   3.047  0.00231 **
## CLS_1:MAXDEPTH_1 -0.03957    0.03096  -1.278  0.20130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 275.83  on 545  degrees of freedom
## Residual deviance: 251.77  on 542  degrees of freedom
## AIC: 259.77
##
## Number of Fisher Scoring iterations: 6
```

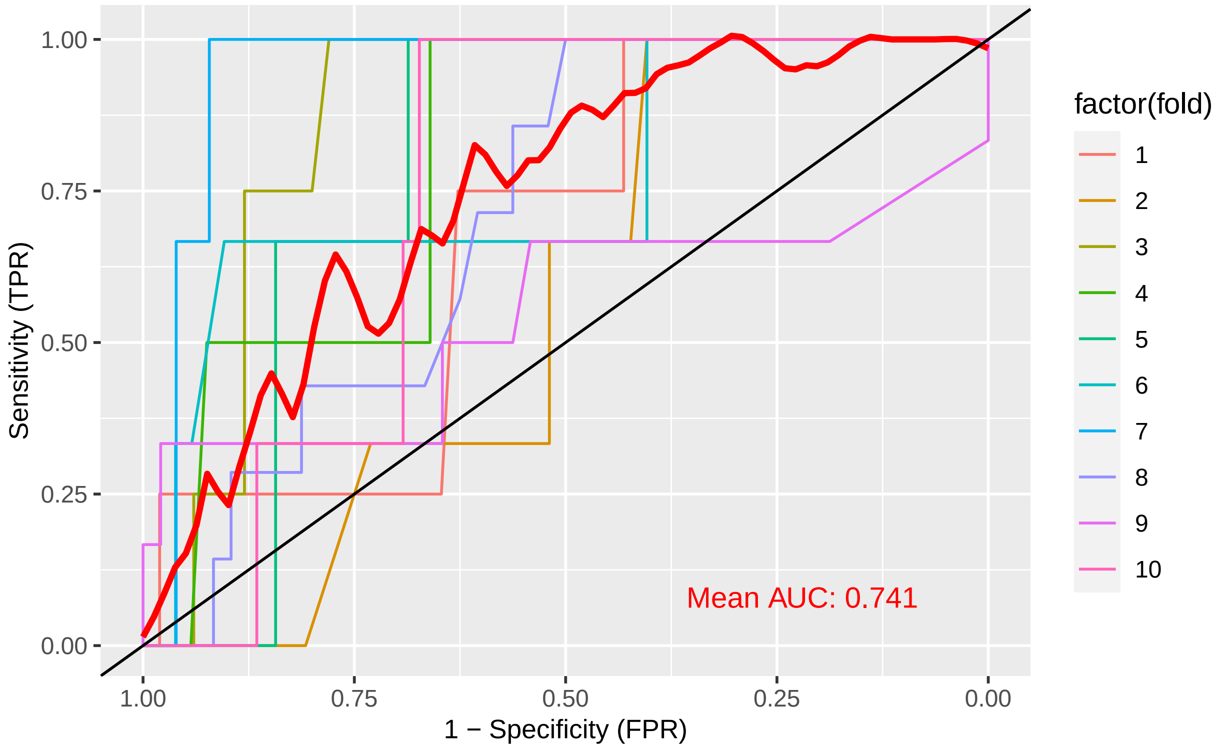
```
fm <- glm(recurrence ~ CLS_1 + MAXDEPTH_1, family = binomial, data = cc_clean)
summary(fm)
```

```
##
## Call:
## glm(formula = recurrence ~ CLS_1 + MAXDEPTH_1, family = binomial,
##     data = cc_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0290  -0.3934  -0.3019  -0.2467   2.7163
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.66375    0.32367 -11.319  < 2e-16 ***
## CLS_1        0.67667    0.26206   2.582 0.009819 **
## MAXDEPTH_1   0.05839    0.01687   3.461 0.000538 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 275.83  on 545  degrees of freedom
## Residual deviance: 253.40  on 543  degrees of freedom
## AIC: 259.4
##
## Number of Fisher Scoring iterations: 6
```

As expected, the interaction between `CLS_1` and `MAXDEPTH_1` appears to be insignificant so we will omit it out of the final model. We can conclude that our final model that predicts recurrence will contain the main effect predictors predictors `CLS_1` and `MAXDEPTH_1`.

Validation

The method chosen to validate the model is **K-Folds Cross Validation**. An advantage to using this method is its ability to prevent training an overfit model (e.g. a poor model for generalization). This technique ceases to partition a shuffled dataset into k partitions, where for k-folds (or k-iterations), one partition is held out to be the validation dataset, and the remaining $k-1$ partitions are assigned to be training datasets. This way, the model is trained on k distinct datasets and validated k times to obtain a mean evaluation score. In terms of the evaluation score, the referenced metric is the mean AUC value.

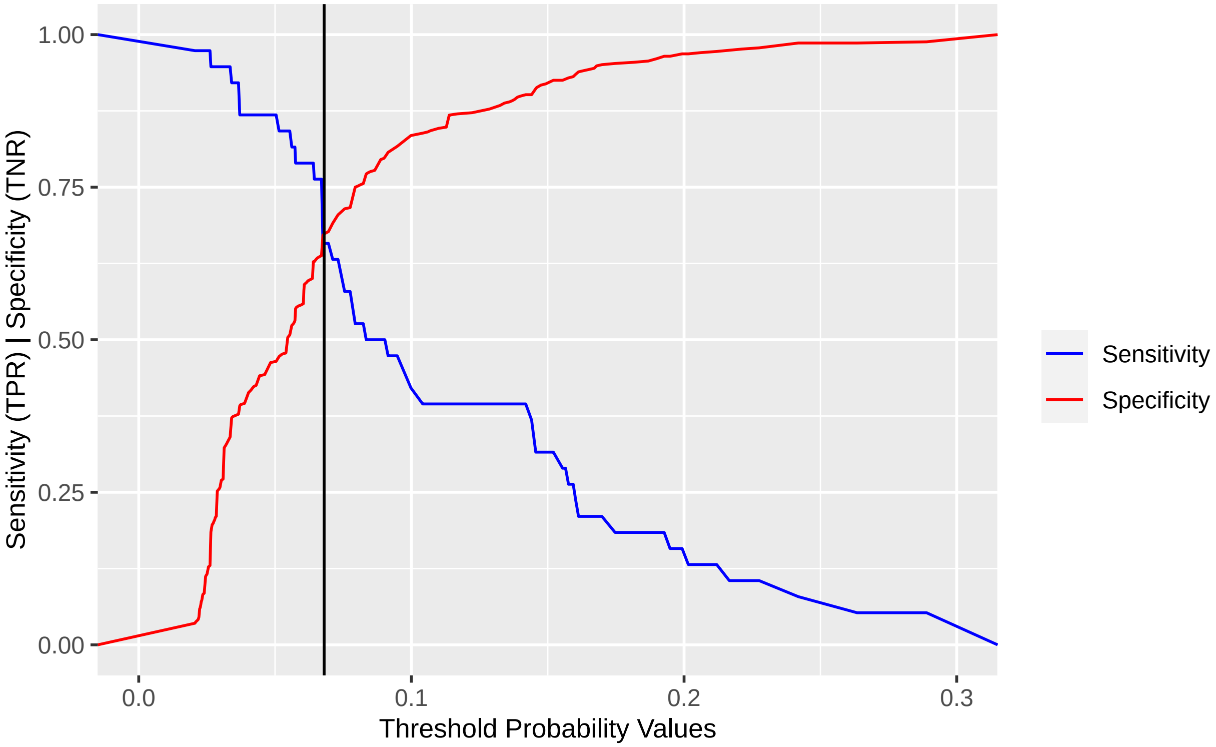


By choosing $k=10$, a ROC curve and AUC value is computted on each fold, where taking an average of the individual AUCs equate to a mean AUC of 0.741. This is evidence that the model is considered *acceptable* in regards to its ability to classify a recurrence of cervical cancer.

Fold	1	2	3	4	5	6	7	8	9	10	Mean
AUC	0.671	0.567	0.872	0.797	0.790	0.762	0.947	0.708	0.545	0.743	0.741

Classification

As for identifying the optimal cutoff for relapse classification of cervical cancer, it is crucial to optimize the decision process as the wrong diagnosis can result with significant consequences in terms of a patient's life. As such, it is recommended to maximize both true postive rate and true negative rate (e.g. sensitiveiy and specificity respectively), in context of this study. To maximize the true positive rate (TPR) means to maximize the probability of classifying a postive relapse when it is observed. As well, to maximize the true negative rate (TNR) means to maximize the probability of classifying a negative relapse when no relapse is observed. Results of this process mitigate the implications of misdiagnosising a patient. In terms of a graphical visualization, the optimized threshold point is outlined by the intersection of the specificity and sensitivity, which is held at a threshold probability of 0.068 (or 6.8%).



Interpretation

After the validation process, we have confirmed that the parameters in our model were found to be statistically significant. If we take a look at the final model, we can gather that we have only two explanatory variables that contributes to our binomial model which is the level which corresponds to the amount of cancer cells found in capillary lymphatic spaces, and the depth of the cancer when last diagnosed. Our model looks like:

$$log\left(\frac{p_{remission}}{1-p_{remission}}\right) = -3.66375 + 0.67667x_{CLS} + 0.05839x_{DEPTH}$$

Here, $p_{remission}$ stands for the probability of remission, x_{CLS} stands for the level of cancers cells, and x_{depth} stands for depth (in millimeters) of the cancer tumors.

Our results suggest the odds of cervical cancer relapse is multiplied by around $e^{0.67667} \approx 1.97$ times for each level of cancer cells found in capillary lymphatics spaces given that depth stays consistent. Similarly, the odds of relapse also multiplied by around $e^{0.05839} \approx 1.06$ times for each millimeter increase of cancer depth upon initial measurement given that the level of cancer found in capillary lymphatic spaces stays constant.

The baseline odds increase for cancer relapse is around $e^{-3.66375} \approx 0.026$ where the amount of cancer cells found in C.L. spaces was deemed to be level zero and the penetrative depth of the cancer to be under 1 millimeter.

No statistically significant interaction terms was found for the two main effects and therefore are not included in this model.

Conclusion

The final model that was concluded only had two main predictors of cancer remission which was the level of cancer cells in C.L. spaces and the depth of the cancerous tumor upon diagnosis which suggests that individuals with higher amounts of cancerous cells found in capillary lymphatic spaces are more likely to be at risk of cancer relapse compared to those that have less cancerous cells and to a less extent, those whose cancer tumors are embedded deeper are also more at risk compared to those whose tumors are less invasive. The main source of the odds of relapse is the level of cancer cells in capillary lymphatic spaces. While the depth of the tumor definitely contributes to the relapse of cancer, the multiplicative odds increase for each millimeter increase is only around 1.06. This number seems small in comparison, however one has to keep in mind that this is a millimeter increase. From data cleanup, the average cervical cancer depth was around 5~7 millimeters which leads to a significant increase in odds.

Surprisingly, our data only had the two significant predictors for our final model among all the potential inter-related explanatory variables. The most surprising discovery was that cancer cell gradation/differentiation (the process of cells transforming to generalized forms) was not a main factor in predicting remission. Intuitively, one would assume that if a cancer cell is more generalized, it would be easily enter bloodstreams and enter the heavily traversed capillary lymphatic spaces used for removal of excess fluid from interstitial spaces (Telinius and Hjortdal 2019). This factor was deemed to be a main predictor before methods were implemented to clean up the entries with missing/incorrectly recorded data which would suggest that either the implemented data-cleaning methods were flawed or that the factor was not supposed to be significant to begin with as the significance of cell gradation as a factor fell sharply when the data was cleaned up.

Another non-expected result was that the size of the cancerous tumor did not play a significant main factor in predicting relapse of cancer as one would assume that it would be correlated with tumor depth. However, this may have been due to errors when recording the data at the time of diagnosis since even after cleaning up the data set, the significance of size almost surpassed the $\alpha = 0.05$ level of significance.

References

Biodiversity and Climate Change Virtual Laboratory. SDM - Interpretation of Model Outputs. 15 Jan. 2020, support.bccvl.org.au/support/solutions/articles/6000127046-sdm-interpretation-of-model-outputs.

Hazlegreaves and Steph. “The Emotional Impact of False Positive Cancer Diagnosis.” Open Access Government, 1 Dec. 2019, www.openaccessgovernment.org/false-positive-cancer-diagnosis/66256/.

NCSS Research and Development Team. “ROC Curves in NCSS.” NCSS, n.d., www.ncss.com/software/ncss/roc-curves-ncss/.

Canada, Public Health Agency of. 2009. “Cervical Cancer.” Education and awareness;navigation page. *Aem*. https://www.canada.ca/en/public-health/services/chronic-diseases/cancer/cervical-cancer.html.

Telinius, Niklas, and Vibeke Elisabeth Hjortdal. 2019. “Role of the Lymphatic Vasculature in Cardiovascular Medicine.” *Heart* 105 (23): 1777–84. https://doi.org/10.1136/heartjnl-2018-314461.