Introduction

In this study, we will create a model that predicts how various cancer-linked activities and features affect one's likelihood of cancer. This dataset contains medical and lifestyle information for 1500 patients, designed to predict the presence of cancer based on various cancer-linked attributes. It is structured to provide a realistic challenge for predictive modeling in the medical domain. The data samples include the following information for each person:

1. **Age**: Integer values representing the patient's age, ranging from 20 to 80.
2. **Gender**: Binary values representing gender, where 0 indicates Male and 1 indicates Female.
3. **BMI**: Continuous values representing Body Mass Index, ranging from 15 to 40.
4. **Smoking**: Binary values indicating smoking status, where 0 means No and 1 means Yes.
5. **GeneticRisk**: Categorical values representing genetic risk levels for cancer, with 0 indicating Low, 1 indicating Medium, and 2 indicating High.
6. **PhysicalActivity**: Continuous values representing the number of hours per week spent on physical activities, ranging from 0 to 10.
7. **AlcoholIntake**: Continuous values representing the number of alcohol units consumed per week, ranging from 0 to 5.
8. **CancerHistory**: Binary values indicating whether the patient has a personal history of cancer, where 0 means No and 1 means Yes.
9. **Diagnosis**: Binary values indicating the cancer diagnosis status, where 0 indicates No Cancer and 1 indicates Cancer.

Justification of the Model

We chose a logistic regression model because the response variable is binary (cancer diagnosis: 0 or 1), making it appropriate for predicting the probability of a categorical outcome based on multiple predictor variables. Logistic regression is suitable for this setting as it estimates the relationship between the predictors and the log-odds of the response, providing interpretable coefficients and a probabilistic interpretation of the model's predictions.

Overview of the paper's structure.

The report is divided into 4 sections: Introduction, Data Description, Results and Interpretation, and Discussion. In the Introduction, we will briefly describe the context of the problem we are trying to solve and the model we will be using. Data Description will describe the distribution of our predictor variables and their relationships. Results and Interpretation will find and fit the optimal model and interpret it in the context of our study. Lastly, Discussion will summarize the overall report and go over further areas that this project can go in terms of application and relevance.

Data Description

We will begin by analyzing the summary statistics of the data set. Due to the nature of the data, some variables are binary, labeled with (B). We note that the diagnosis (response variable) is binary. Per Figure 1, of the non-binary variables, Age and BMI are more uniformly distributed whereas Physical Activity and Alcohol Intake appear somewhat normally distributed. Lastly, Genetic risk is right skewed. To summarize the binary variables, we can see that ~60% of

patients did not have cancer, there is a roughly even split by gender, non-smokers were around 70% of patients and lastly those without a history of cancer around 90%.

Variable	Mean	Standard Deviation
Age	50.32	17.641
BMI	27.513	7.23
GeneticRisk	0.509	0.679
PhysicalActivity	4.898	2.866
Alcohol Intake	2.418	1.419

Table 1. Variable means and standard deviations of continuous variables

Table 1 gives each variable's mean and standard deviation for the continuous variables in the dataset. We observe that the average patient statistics are 50.3 years of age, a BMI of 27.5, a genetic risk of either 1 or 2, around 5 hours each week working out, and about 2.4 alcoholic drinks per week. Age has the largest standard deviation at 17.6, likely due to the large scope this data covers (20-80 years old), which is the biggest integer range in the dataset.

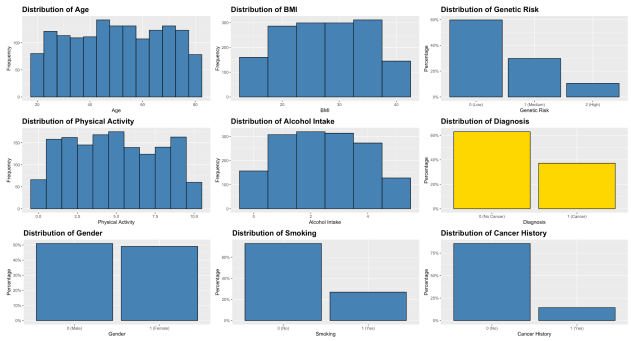

Figure 1. Distribution of variables

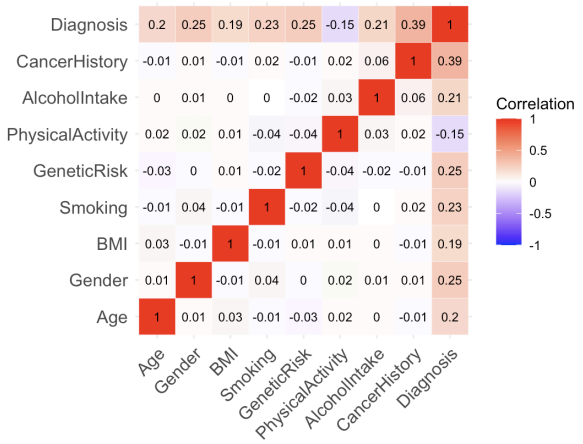

Figure 2. Heatmap of Variable Relationships

When examining figure 2, we find that generally, the correlations are relatively weak, suggesting that they are largely independent of each other, which is a good sign that we do not have multicollinearity. In terms of relation to the response variable Physical Activity stands out as it has a negative correlation to a diagnosis which will be further analyzed in this paper. Based on the dataset serving as a predictor of whether or not a patient is diagnosed given their circumstances, we can see a clear success or fail outcome here. Therefore to begin our analysis we will develop a logistic regression model for a binomial distribution.

Results and Interpretation

Running the full model in R, we achieve the output shown in Figure 3. Calculating R^2_{dev} for logistic regression models using G^2_0 and G^2_A (null deviance and residual deviance, respectively), we can find the percent variation in cancer diagnoses explained by our eight predictor variables.

```
Call:
glm(formula = Diagnosis ~ Age + Gender + BMI + Smoking + GeneticRisk +
    PhysicalActivity + AlcoholIntake + CancerHistory, family = binomial,
    data = Cancer)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -10.065670   0.628897 -16.005   <2e-16 ***
Age               0.051966   0.005002  10.388   <2e-16 ***
Gender            2.073929   0.176044  11.781   <2e-16 ***
BMI               0.116519   0.011810   9.866   <2e-16 ***
Smoking           1.955750   0.184162  10.620   <2e-16 ***
GeneticRisk       1.550032   0.124658  12.434   <2e-16 ***
PhysicalActivity -0.241617   0.029223  -8.268   <2e-16 ***
AlcoholIntake     0.620628   0.061152  10.149   <2e-16 ***
CancerHistory     4.247498   0.296407  14.330   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1965.1  on 1499  degrees of freedom
Residual deviance: 1031.4  on 1491  degrees of freedom
AIC: 1059.7
```

Figure 3. R Output for the full model

Interpreting R results

$R^2_{dev} = 1 - (G^2_A / G^2_0)$
$= 1 - (1031.4 / 1965.1) = 0.475.$
Therefore, 47.5% of the variation in cancer diagnoses is explained by the predictor variables.

Calculating the p-value for the model with predictors vs the intercept-only model, we use $G^2_0 - G^2_A$ to approximate the $\chi^2_{df1-df2}$ distribution.
$G^2_0 - G^2_A = 1965.1 - 1031.4 = 933.7$
$P(\chi^2_{1499-1491} > 933.7) = 3.032e\text{-}196 < .05$
Therefore, we have sufficient evidence to reject the null hypothesis that the intercept-only model is appropriate and instead find the model with predictors to be more appropriate.

Comparing Models

Running forward stepwise logistic regression with the Akaike Information Criterion (AIC) allows us to compare the best model at each size, starting with the null model.

```
## Start:  AIC=-2183.22
## Diagnosis ~ 1
##
##                   Df Sum of Sq    RSS     AIC
## + CancerHistory    1    53.752 295.72 -2431.7
## + GeneticRisk      1    22.464 327.00 -2280.9
## + Gender           1    21.900 327.57 -2278.3
## + Smoking          1    18.008 331.46 -2260.6
## + AlcoholIntake    1    15.821 333.65 -2250.7
## + Age              1    13.508 335.96 -2240.3
## + BMI              1    12.294 337.17 -2234.9
## + PhysicalActivity 1     7.872 341.60 -2215.4
## <none>                         349.47 -2183.2
##
## Step:  AIC=-2431.73
## Diagnosis ~ CancerHistory
##
##                   Df Sum of Sq    RSS     AIC
## + GeneticRisk      1    23.227 272.49 -2552.4
## + Gender           1    21.379 274.34 -2542.3
## + Smoking          1    17.008 278.71 -2518.6
## + Age              1    14.109 281.61 -2503.1
## + BMI              1    12.858 282.86 -2496.4
## + AlcoholIntake    1    12.794 282.92 -2496.1
## + PhysicalActivity 1     8.639 287.08 -2474.2
## <none>                         295.71 -2431.7
##
## Step:  AIC=-2552.44
## Diagnosis ~ CancerHistory + GeneticRisk
##
##                   Df Sum of Sq    RSS     AIC
## + Gender           1   21.5854 250.90 -2674.2
## + Smoking          1   17.8547 254.63 -2652.1
## + Age              1   15.1191 257.37 -2636.1
## + AlcoholIntake    1   13.3653 259.12 -2625.9
## + BMI              1   12.4728 260.02 -2620.7
## + PhysicalActivity 1    7.5673 264.92 -2592.7
## <none>                         272.49 -2552.4
```

```
## Step:  AIC=-2674.23
## Diagnosis ~ CancerHistory + GeneticRisk + Gender
##
##                   Df Sum of Sq    RSS     AIC
## + Smoking          1   16.5207 234.38 -2774.4
## + Age              1   14.8642 236.04 -2763.8
## + AlcoholIntake    1   13.0543 237.85 -2752.4
## + BMI              1   12.8845 238.02 -2751.3
## + PhysicalActivity 1    8.1737 242.73 -2721.9
## <none>                         250.90 -2674.2
##
## Step:  AIC=-2774.4
## Diagnosis ~ CancerHistory + GeneticRisk + Gender + Smoking
##
##                   Df Sum of Sq    RSS     AIC
## + Age              1   15.3276 219.06 -2873.8
## + BMI              1   13.2324 221.15 -2859.6
## + AlcoholIntake    1   13.1498 221.23 -2859.0
## + PhysicalActivity 1    7.1579 227.22 -2818.9
## <none>                         234.38 -2774.4
##
## Step:  AIC=-2873.85
## Diagnosis ~ CancerHistory + GeneticRisk + Gender + Smoking +
##     Age
##
##                   Df Sum of Sq    RSS     AIC
## + AlcoholIntake    1   13.0572 206.00 -2964.0
## + BMI              1   12.3928 206.66 -2959.2
## + PhysicalActivity 1    7.4715 211.58 -2923.9
## <none>                         219.06 -2873.8
```

```
## Step:  AIC=-2964.03
## Diagnosis ~ CancerHistory + GeneticRisk + Gender + Smoking +
##     Age + AlcoholIntake
##
##                   Df Sum of Sq    RSS     AIC
## + BMI              1   12.2540 193.74 -3054.0
## + PhysicalActivity 1    8.1239 197.87 -3022.4
## <none>                         206.00 -2964.0
##
## Step:  AIC=-3054.03
## Diagnosis ~ CancerHistory + GeneticRisk + Gender + Smoking +
##     Age + AlcoholIntake + BMI
##
##                   Df Sum of Sq    RSS     AIC
## + PhysicalActivity 1    8.3518 185.39 -3118.1
## <none>                         193.74 -3054.0
##
## Step:  AIC=-3118.12
## Diagnosis ~ CancerHistory + GeneticRisk + Gender + Smoking +
##     Age + AlcoholIntake + BMI + PhysicalActivity
```

Figure 4. R output of forward stepwise logistic regression with AIC

As we can see from the results, out of the nine models compared here, the model with all eight predictors results in the lowest AIC value, corresponding to the best model.

If we run the same procedure but with a different criterion (Bayesian information criterion), we can observe that removing any of the predictors will both raise RSS and AIC which indicate that it is a worse model. Therefore the starting model (the full model) is optimal.

```
backwardAIC <- step(model, direction = "backward")
```

```
## Start:  AIC=-3118.12
## Diagnosis ~ Age + Gender + BMI + Smoking + GeneticRisk + PhysicalActivity +
##     AlcoholIntake + CancerHistory
##
##                    Df Sum of Sq    RSS     AIC
## <none>                          185.39 -3118.1
## - PhysicalActivity  1     8.352 193.74 -3054.0
## - BMI               1    12.482 197.87 -3022.4
## - AlcoholIntake     1    13.576 198.97 -3014.1
## - Age               1    14.715 200.11 -3005.6
## - Smoking           1    16.293 201.69 -2993.8
## - Gender            1    20.686 206.08 -2961.4
## - GeneticRisk       1    24.272 209.66 -2935.6
## - CancerHistory     1    51.812 237.20 -2750.4
```

Figure 5. R output of backward stepwise logistic regression with BIC

Assessing the model

Since we determined the model with all eight predictors is the optimal model through AIC and BIC criterion, we can run diagnostic tools to determine if the model breaks any of the statistical assumptions and directions to proceed from there.

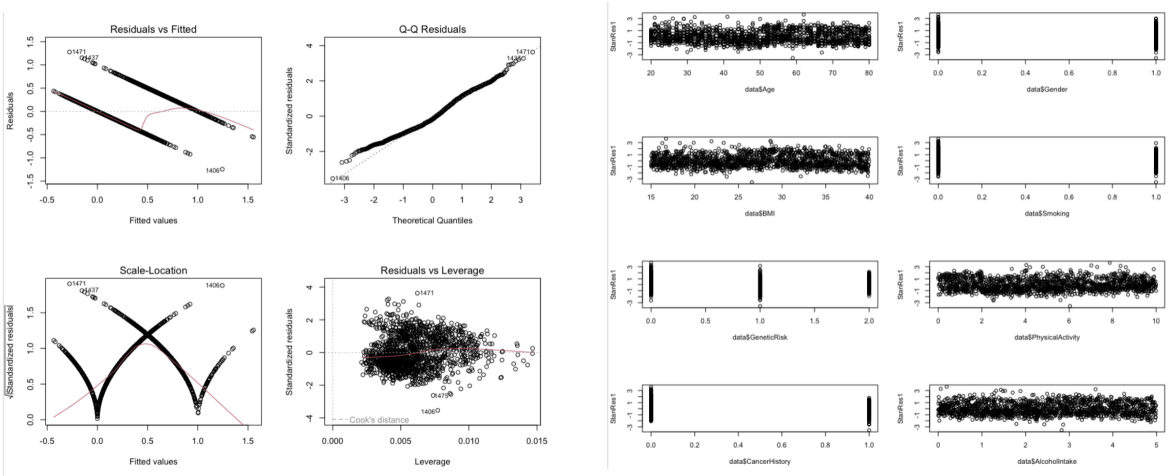

Figure 6. Diagnostic Graph outputs for the full model

From the graphs, we can see that there is a non-linear pattern in the Residuals vs Fitted Plot, suggesting that the linearity assumption may not be satisfied. Furthermore, the spread of standardized residuals across the range of fitted values indicates heteroscedasticity, violating the assumption of constant variance. However, the Standardized Residuals vs Predictors plots for each feature appear randomly scattered along 0. Therefore, future action to be taken with the model may include non-linear transformations and weighted least squares.

```
bcPower Transformations to Multinormality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.8302        0.83       0.6870       0.9735
Y2   -0.0162        0.00      -0.0416       0.0092
Y3    0.7598        0.76       0.5580       0.9616
Y4   -0.4651       -0.47      -0.4968      -0.4333
Y5   -0.1661       -0.17      -0.1916      -0.1405
Y6    0.6597        0.66       0.5994       0.7200
Y7    0.6935        0.69       0.6354       0.7516
Y8   -0.9983       -1.00      -1.0501      -0.9465
Y9   -0.2330       -0.23      -0.2600      -0.2060

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```

Figure 7. Transformation output in R

Figure 7 indicates the following transformations that should be taken to better fit our model and mitigate the issues shown in the Residuals vs Fitted plot in the diagnostic test. The new R^2_{dev} value for the updated full model is:

```
Call:
glm(formula = Diagnosis ~ Age1 + Gender + BMI1 + Smoking + GeneticRisk1 +
    PhysicalActivity1 + AlcoholIntake1 + CancerHistory, family = binomial,
    data = cancer)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.674842   0.981709  -3.743 0.000182 ***
Age1                1.911271   0.208911   9.149  < 2e-16 ***
Gender              1.798084   0.158595  11.338  < 2e-16 ***
BMI1              -26.446033   2.811420  -9.407  < 2e-16 ***
Smoking             1.673176   0.166207  10.067  < 2e-16 ***
GeneticRisk1        1.353401   0.128786  10.509  < 2e-16 ***
PhysicalActivity1  -0.457182   0.058456  -7.821 5.24e-15 ***
AlcoholIntake1     -0.011690   0.006665  -1.754 0.079452 .
CancerHistory       3.809792   0.263941  14.434  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1979.0  on 1499  degrees of freedom
Residual deviance: 1192.4  on 1491  degrees of freedom
AIC: 1210.4

Number of Fisher Scoring iterations: 5
```

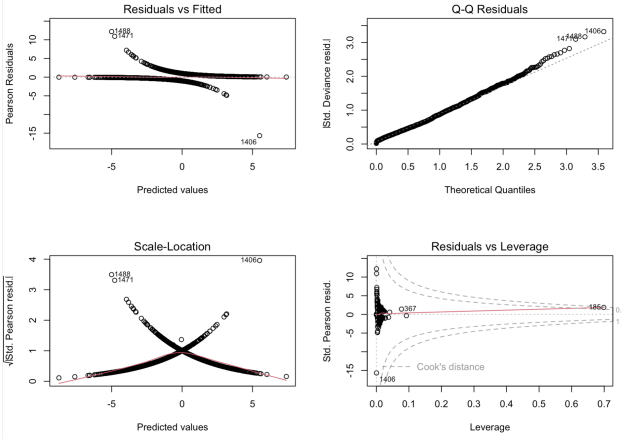

Figure 8. R Output for the transformed logistic regression model and the diagnostic plots corresponding to the transformed model

$R^2_{dev} = 1 - (G^2_A / G^2_0)$
$= 1 - (1192.4 / 1970.0) = 0.395$
Therefore, 39.5% of the variation in cancer diagnoses is explained by the predictor variables.

Although this is a decrease from the original, untransformed model, we can see improvements in the diagnostic plots, specifically the tails of the Normal Q-Q plot, the average around 0 in the Residuals vs Fitted plot, and the less obvious shape in the Scale-Location plot.

Discussion

Our final model demonstrates several interesting findings. Notably, the odds of being diagnosed with cancer can be examined through individual variables. For our categorical variables, we found that the odds of getting cancer increased by a factor of 7.925 for females, 7.0639 for smokers, and 69.895 for those with a history of cancer. For our continuous variables, we found that a one-unit increase in age, alcohol intake, BMI, and physical activity changed the odds of being diagnosed with cancer by factors of 1.053, 1.124, 0.785, and 1.860, respectively. Lastly, our multinomial variable for genetic risk showed that a one-unit increase in risk increased the odds of getting cancer by a factor of 4.711. We can conclude that smoking and having a history of cancer are the most significant predictors of cancer.

These findings are consistent with available research on cancer diagnosis. For example, a study conducted by Scott Kulm et al. used a regression model to predict cancer diagnosis, utilizing similar health predictors along with additional variables. They found that their model accurately predicted their test data. However, our logistic model is more appropriate than a linear one, given that our response variable has only two possible outcomes.

The primary limitations of this model stem from potential violations of model assumptions. However, since we used a logistic model instead of a linear one, issues such as normality and linearity are not relevant. Additionally, we may want to consider adding more predictors in the future. By including other predictors, such as dietary measures or more accurate assessments of body composition, we may achieve better results. If we consider new predictors, we must ensure they are significant to prevent overfitting the model. Our final logistic model is

$$\hat{y}\,\textbf{(Diagnosis}^{0.83}\textbf{)} = 1/(\,1 + \exp(-\{\,-10.066 + 0.052(\log(\text{Age})) + 2.074(\text{Gender}) \;\; 0.1176(\text{BMI}^{-0.47}) + 1.956(\text{Smoking}) + 1.550(\text{Genetic Risk}^{0.66}) - 0.2427(\text{Physical Activity}^{0.69}) + 0.621(\text{Alcohol Intake}^{-1.00}) + 4.247(\text{Cancer History})\,\}\,)$$

Works Cited

Kulm, Scott et al. “Simple Linear Cancer Risk Prediction Models With Novel Features Outperform Complex Approaches.” *JCO clinical cancer informatics* vol. 6 (2022): e2100166. doi:10.1200/CCI.21.00166