University of Helsinki
Faculty of Science
Department of Mathematics and Statistics

Master's thesis

Prostate Cancer Relapse Prediction with Biomarkers and Logistic Regression

Pyry Halonen

Statistics

Supervisor: Leena Kalliovirta March 18, 2022

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty Faculty of Science	Laitos — Institution — Department Department of Mathematics and Statistics

Tekijä — Författare — Author

Pyry Halonen

Työn nimi — Arbetets titel — Title

Prostate Cancer Relapse Prediction with Biomarkers and Logistic Regression

Oppiaine — Läroämne — Subject

Statistics

Työn laji — Arbetets art — Level Master's thesis	Aika — Datum — Month and year March 2022	Sivumäärä — Sidoantal — Number of pages 51

Tiivistelmä — Referat — Abstract

Prostate cancer is the second most common cancer among men and the risk evaluation of the cancer prior the treatment can be critical. Risk evaluation of the prostate cancer is based on multiple factors such as clinical assessment. Biomarkers are studied as they would also be beneficial in the risk evaluation. In this thesis we assess the predictive abilities of biomarkers regarding the prostate cancer relapse.

The statistical method we utilize is logistic regression model. It is used to model the probability of a dichotomous outcome variable. In this case the outcome variable indicates if the cancer of the observed patient has relapsed. The four biomarkers AR, ERG, PTEN and Ki67 form the explanatory variables. They are the most studied biomarkers in prostate cancer tissue.

The biomarkers are usually detected by visual assessment of the expression status or abundance of staining. Artificial intelligence image analysis is not yet in common clinical use, but it is studied as a potential diagnostic assistance. The data contains for each biomarker a visually obtained variable and a variable obtained by artificial intelligence. In the analysis we compare the predictive power of these two differently obtained sets of variables. Due to the larger number of explanatory variables, we seek the best fitting model. When we are seeking the best fitting model, we use an algorithm *glmulti* for the selection of the explanatory variables. The predictive power of the models is measured by the receiver operating characteristic curve and the area under the curve.

The data contains two classifications of the prostate cancer whereas the cancer was visible in the magnetic resonance imaging (MRI). The classification is not exclusive since a patient could have had both, a magnetic resonance imaging visible and an invisible cancer. The data was split into three datasets: MRI visible cancers, MRI invisible cancers and the two datasets combined. By splitting the data we could further analyze if the MRI visible cancers have differences in the relapse prediction compared to the MRI invisible cancers.

In the analysis we find that none of the variables from MRI invisible cancers are significant in the prostate cancer relapse prediction. In addition, all the variables regarding the biomarker AR have no predictive power. The best biomarker for predicting prostate cancer relapse is Ki67 where high staining percentage indicates greater probabilities for the prostate cancer relapse. The variables of the biomarker Ki67 were significant in multiple models whereas biomarkers ERG and PTEN had significant variables only in a few models. Artificial intelligence variables show more accurate predictions compared to the visually obtained variables, but we could not conclude that the artificial intelligence variables are purely better. We learn instead that the visual and the artificial intelligence variables complement each other in predicting the cancer relapse.

Avainsanat — Nyckelord — Keywords

logistic regression model, biomarker, prostate cancer, artificial intelligence

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1 Introduction

In 2020, prostate cancer caused 375 000 deaths worldwide and it is the second most common cancer among men (Sung et al. [1]). Risk evaluation and accurate diagnosis of prostate cancer is needed before effective treatment. Nowadays risk evaluation is based on four indicators: Gleason score, level of prostate-specific antigen, clinical stage assessment and imaging stage assessment. Gleason score tells how aggressive the cancer seems and it is given by a pathologist from samples of cells. A biomarker is a biological molecule found in tissues that indicates normal or abnormal condition (Strimbu & Tavel [2]). Biomarkers may give added value when evaluating the risk of prostate cancer or give additional information to the clinical parameters (Moldovan et al. [3], Schoots et al. [4]).

In this thesis we analyze if biomarkers have relation to the prostate cancer relapse by analysing the predictive abilities of biomarkers using logistic regression model. Relapse also known as biochemical recurrence of prostate cancer is the outcome variable in the statistical models. It is defined as two sequential measurements of high prostate-specific antigen ($>$0.2 ng/mL). In the analysis we have four biomarkers to explain the relapse: AR, ERG, PTEN and Ki67. They are the most studied in prostate cancer tissue and are possibly implemented to clinical practice (Guo et al. [5], Wang et al. [6], Troyer et al. [7], Berlin et al. [8]).

The four biomarkers are usually detected by immunohistochemistry and visual assessment of the expression status or abundance of staining. In the analysis we have a variable measured visually and a variable measured by artificial intelligence (AI) for each biomarker. Artificial intelligence image analysis is not yet in common clinical use, but it is studied as a potential diagnostic assistance (Fourcade & Khonsari [9]). The artificial intelligence uses neural network in this study and it is trained to analyze histological images for potential biomarkers. The objective of the used artificial intelligence is to help standardize the analysis, to catch the findings a pathologist may miss and to increase diagnostic accuracy (Niazi [10]). One task of the thesis is to assess the applicability of artificial intelligence in this field of study.

The data consist of 387 observations which are patients who all underwent robot-assisted laparoscopic prostatectomy as primary therapy at the Helsinki University Hospital. The study period was from January 2014 to September 2015. All patients underwent preoperative magnetic resonance imaging at discretion of urologist. The cancers were classified into two groups based on whether the cancer was visible in the magnetic resonance imaging. The classification was not exclusive for observations since a patient could have had both magnetic resonance imaging visible and invisible cancer. The classification of cancer was taken into consideration in the analysis.

The objective of this thesis is to model prostate cancer relapse with the given biomarkers. We want to know if biomarkers have any relation to the prostate cancer relapse so it could be predicted. In order to find out the applicability of artificial intelligence in this field of study, we compare artificial intelligence to visual assessment of human observers in the relapse prediction. Additionally we want to examine if the classification of cancer affects the relapse prediction.

The logistic regression model used in this thesis is a regression model used for describing a dichotomized outcome variable with one or multiple explanatory variables. The unknown parameters of the predictors are estimated using the maximum likelihood estimation and the predictive power of the models is measured mainly with receiver operating characteristic curve. Pursuing the models with the best fit, we used algorithm *glmulti* (Calcagno & de Mazancourt [11]) to select the explanatory variables in the environment of R software (R Core Team [12]).

This thesis begins with theoretical part where we define the logistic regression model and the practices affiliated to it: parameter estimation, significance testing and selection of covariates. There we also explain the important concept of odds ratio. From model theory we advance to the model diagnostics where we go through issues that have to be checked before the model can be used for analysis. After the theoretical part comes the data chapter which explains each variable and the dataset. Finally, there is the analysis and the conclusions. Appendix contains all the model summaries used in the analysis.

2 Logistic Regression Model

The logistic regression model is a generalized linear model. Generalized linear models have three following components. The first is the random component that describes the randomness of the process. It defines the outcome variable y and its probability distribution. Generalized linear models assume that observations $\mathbf{y} = (y_1, \ldots, y_n)^T$ are independent. The second component is the linear predictor that describes the fixed part of the process. A parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and a matrix $\mathbf{X}$ forms the linear predictor $\boldsymbol{\beta}\mathbf{X}$.

$$\text{The matrix } \mathbf{X} \text{ is a } n \times p \text{ model matrix that contains values of explanatory variables } \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \text{ for the } i = 1, \ldots, n \text{ observations.} \tag{2.1}$$

The third component is the link function that describes the relation between the random component and the linear predictor with

$$g[\mathrm{E}(\mathbf{y})] = \boldsymbol{\beta}\mathbf{X},$$

where g is the link function.

In the case of logistic regression model the outcome variable is Bernoulli distributed with probabilities of $P(y_i = 1) = \pi_i$ and $P(y_i = 0) = 1 - \pi_i$. Note also that $\mathrm{E}(y_i) = \pi_i$. It can be viewed as binomial distribution where the number of trials $n_i = 1$. The natural parameter for the binomial distribution is $\ln[\pi_i/1 - \pi_i]$, so the link function for logistic regression is the *logit-link*

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right), \tag{2.2}$$

where ln is the natural logarithm.

Logistic regression model is used to model the probability of a dichotomous outcome variable. In this case the outcome variable is the *Relapsed* and is marked as

$$y_i = \begin{cases} 0, & \text{if the cancer of observed patient } i \text{ has not relapsed,} \\ 1, & \text{if the cancer of observed } i \text{ patient has relapsed.} \end{cases} \tag{2.3}$$

Logistic regression model can be written using two equivalent formulas when the explanatory variables are continuous. They are

$$\operatorname{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{j=1}^{p} \beta_j x_{ij} \tag{2.4}$$

and

$$\pi_i = \frac{\exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)}, \tag{2.5}$$

where $i = 1, \ldots, n$. For categorical explanatory variables we use dummy variables.

The part of $\ln(\pi_i/1 - \pi_i)$ in the model formula is considered as log odds. The practical problem with the log odds is that a change in the scale of the log odds is rather hard to explain. To have a more useful interpretation we need to convert the log odds to odds ratio (OR) (Yule [13]).

For example, let x be a dichotomous explanatory variable in a logistic regression model. Now the odds ratio for x is the ratio of the odds for $x = 1$ to the odds for $x = 0$, which is notated as

$$\text{OR} = \frac{\frac{\pi(1)}{[1 - \pi(1)]}}{\frac{\pi(0)}{[1 - \pi(0)]}}. \tag{2.6}$$

The odds ratio describes how much more likely or unlikely the outcome variable y_i is going to get value 1 with $x = 1$ compared to $x = 0$. If the variable x is continuous the odds ratio is the ratio of odds for $\pi(x)$ to the odds for $\pi(x + 1)$. The relationship between the coefficient and the odds ratio makes the logistic regression model such a powerful analytic research tool (Hosmer et al. [14]).

Odds ratio of 1 means that the explanatory variable x_j shows no discrimination on the prediction of the outcome variable. In this case the outcome variable y_i expresses the relapse of cancer. Less than 1 odds ratio of an explanatory variable means that the variable decreases the occurrence of the cancer relapse. Greater than 1 odds ratio of an explanatory variable means that the variable increases the occurrence of the cancer relapse.

For example, if one dichotomized explanatory variable x_1 denotes whether a specimen does ($x_1 = 1$) or does not ($x_1 = 0$) have a staining from a specific biomarker. Let the variable x_1 have odds ratio of 0.5 in the logistic regression model. Then the odds of cancer relapse among those patients whose specimen have a staining from the biomarker is one-half the odds of cancer relapse for those patients who did not get stained from the biomarker. An another example, where a continuous variable x_2 has odds ratio of 1.2

in the logistic regression model and the variable x_2 indicates the percentage of an other biomarker staining valued between 1-100. Now the odds of relapse multiply by 1.2 per percentage increase of staining.

2.1 Estimation

Hosmer, Lemeshow, Sturdivant [14] is the main source for this section. We need to estimate the values of the unknown parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ in the equation (2.5) in order to fit the logistic regression model to a set of data. The values of β_j, $j = 1, \dots, p$ are called coefficients. They describe the change in log odds of having the outcome per unit change in the explanatory variable x_j, $j = 1, \dots, p$. The general method used for parameter estimation in logistic regression model is *maximum likelihood*. The method of maximum likelihood produces values for the unknown parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ that maximize the probability of obtaining the observed set of data. Usually the logistic regression model contains an intercept coefficient β_0 as well though it is not very conventional, because it describes the log odds of $y = 1$ while all explanatory variables x_j, $j = 1, \dots, p$ are equal to zero. A model containing only the intercept coefficient and no explanatory variables is called a null model.

The maximum likelihood method uses maximum likelihood function which expresses the probability of the observed data as a function of the unknown parameters. There is no analytical solution for maximizing the maximum likelihood function in the logistic regression model so maximizing must be done numerically. The maximum likelihood function of model (2.5) is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i} [1 - \pi(\mathbf{x}_i)]^{1-y_i}. \tag{2.7}$$

However, it is simpler to use the logarithm of the maximum likelihood function, so we define it as

$$l(\boldsymbol{\beta}) = \ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(\mathbf{x}_i)] + (1 - y_i) \ln[1 - \pi(\mathbf{x}_i)]\}. \tag{2.8}$$

We call the logarithm of the maximum likelihood function the *log-likelihood function*.

To proceed finding the maximum value of the log-likelihood function, we differentiate the log-likelihood function with respect to $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_j$, $j = 1, \dots, p$, and setting the result to zero. This yields equations

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] \tag{2.9}$$

and for $j = 1, \ldots, p$

$$\sum_{i=1}^{n} \mathbf{x}_i [y_i - \pi(\mathbf{x}_i)]. \tag{2.10}$$

The values of $\boldsymbol{\beta}$ obtained as the solution to the log-likelihood equations are called the maximum likelihood estimates and are denoted with $\hat{\boldsymbol{\beta}}$. We report the values of $\hat{\boldsymbol{\beta}}$ under the title *Estimate* in tables in the analysis chapter.

2.2 Significance Testing of Coefficients

Again, this section utilizes literature of Hosmer et al. [14] unless informed otherwise. After estimating parameters we want to asses the significance of the explanatory variables in the model. In order to check if an explanatory variable has a significant relation to the outcome variable, we need to perform a statistical hypothesis test. We set the null hypothesis to be

$$H_0 : \beta_j = 0 \tag{2.11}$$

and the alternative hypothesis

$$H_1 : \beta_j \neq 0. \tag{2.12}$$

The p-value significance threshold is set to be 0.05.

There are several ways to determine whether the explanatory variables in the model are significantly related to the outcome variable. Hosmer et al. [14] suggests that the usual way to test the significance of the variable is with the *likelihood ratio test* (Wilks [15]). It is favoured since it shares the same principles with significance tests of linear regression models. The likelihood ratio test is formed using two models, where one is estimated with the specific variable and one is estimated without it. Then the predicted values of the two models are compared to the observed values of the response variable. In the case of logistic regression, the comparison of observed to predicted values utilizes the log-likelihood function defined in equation (2.8). Likelihood ratio test uses the deviance statistic

$$\begin{aligned} D &= -2 \ln \left[\frac{\text{(likelihood of the fitted model)}}{\text{(likelihood of the saturated model)}} \right] \\ &= -2 \sum_{i=1}^{n} \left[y_i \ln \frac{\hat{\pi}_i}{y_i} + (1 - y_i) \ln \frac{1 - \hat{\pi}_i}{1 - y_i} \right]. \end{aligned} \tag{2.13}$$

The saturated model in (2.13) is a perfect fitting model, which is obtained when the model contains as many parameters as there are observations in the data. A perfect fitting model is not useful in practice whereas a parsimonious model is useful in estimating the true relation.

Using deviances we can compute the G test statistic

$$\begin{aligned} G &= D(\text{model without the variable}) - D(\text{model with the variable}) \\ &= -2\ln\left[\frac{(\text{likelihood without the variable})}{(\text{likelihood with the variable})}\right]. \end{aligned} \tag{2.14}$$

Under the null hypothesis, the G test statistic follows a chi-square distribution with 1 degrees of freedom where we can obtain the p-value for the hypothesis deciding if the explanatory variable is significant in the model.

We can use statistically equivalent *Wald test* (Wald [16]) for significance testing. Under the null hypothesis, Wald test statistic follows a standard normal distribution, where the p-value can be obtained. The Wald test statistic is defined as

$$W = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}, \tag{2.15}$$

where $\widehat{SE}(\hat{\beta}_j)$ is the estimate of standard error of $\hat{\beta}_j$. We can obtain an estimated standard error of $\hat{\beta}_j$ from the inverse of the estimated information matrix $\hat{\mathbf{I}}$ which is

$$\hat{\mathbf{I}}^{-1} = (\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}, \tag{2.16}$$

where $\hat{\mathbf{V}}$ denotes $n \times n$ diagonal matrix having the $\hat{\pi}_i(1-\hat{\pi}_i)$ as the main element and $\mathbf{X}$ is defined in equation (2.1). The square roots of the main diagonal elements of the $\hat{\mathbf{I}}^{-1}$ are the estimates of standard errors of $\hat{\beta}_j$.

2.3 Selecting Covariates

The section is based on Calcagno & de Mazancourt [11] except for the parts where it is stated otherwise. In some cases data can contain a large amount of possible explanatory variables that may also be irrelevant or unnecessary to describe the outcome variable. As a result we may need to drop some of them from the model. In the empirical analysis we start from a benchmark model containing all explanatory variables. Using the benchmark model we can seek for the best fitting model as well as various models with specific sets of potential explanatory variables to obtain comparison.

One method to select only the important explanatory variables for the model is a stepwise method. It begins with the benchmark model and then one proceeds to drop a non-significant explanatory variable or an explanatory variable which reduces the fit of the model least. The procedure of dropping non-significant variables can be repeated to the point where all the explanatory variables are significant. The stepwise method can also be performed the other way around by starting from a null model without any explanatory variables and then adding the most significant ones.

When using the stepwise method, the decision of dropping or adding a variable can be justified with t-test or some other similar hypothesis testing tool together with a specified significance level. Choosing a proper significance level can cause problems since the number of tests may rise to unexpectedly high (Harrell [17]). Using the information criteria can avoid the use of hypothesis testing tools and the adjustments required for the proper significance level. Then we can compare the models during the procedure of adding or dropping the explanatory variable with information criteria.

The stepwise method is not robust, since the outcome is dependent on the starting model. Starting from the benchmark model and starting from the null model can lead to dissimilar models. Using significance level as a decisive factor and as a stopping rule can naturally lead to different models just by adjusting the significance level.

We decide to use only information criteria when selecting the explanatory variables for the models. Giving each model an information criterion value allows us to rank the models. Thereby the model that holds the smallest value of information criterion is considered to have the best fit. The important difference in using only information criterion to stepwise method is that the information criterion ensures that the best fitting model can always be identified.

Our goal at selecting covariates involves seeking the best model that can still accurately reflect the true outcome experience of the data. According to Hosmer et al. [14], the commonly used information criterion utilized to compare models with different numbers of covariates is the Akaike information criterion (AIC) (Akaike [18]). The measure of Akaike information criterion for a model M can be obtained from

$$\mathrm{AIC}_M = -2[l(\hat{\boldsymbol{\beta}}_M) - k], \tag{2.17}$$

where $l(\hat{\boldsymbol{\beta}}_M)$ is the maximized log-likelihood for model M and k is the number of parameters in the model M.

However, we use the Bayesian information criterion (BIC) (Schwarz [19]) for the selection of the covariates because it is known to punish for increasing the number of parameters more severely than Akaike information criterion (Agresti [20]). The more parsimonious the model is, the better. This is why we chose the Bayesian information criterion and its nature of repelling any unnecessary covariates over the Akaike information criterion. Minimizing the number of covariates in the model has its advantages: The model is more likely to be numerically stable and is less dependant on the observed data. Increasing the number of covariates in the model increases the estimated standard errors (Hosmer et al. [14]). Due to these facts we also use the Bayesian information criterion on the model ranking. The Bayesian information criterion is defined for model M as

$$\mathrm{BIC}_M = k\ln(n) - 2[l(\hat{\boldsymbol{\beta}}_M)], \tag{2.18}$$

where $l(\hat{\boldsymbol{\beta}}_M)$ is the maximized log-likelihood for model M, k is the number of parameters in the model M and n is the number of observations in the data.

2.3.1 Using Glmulti of R Software

We do not settle for manually browsing through candidate models searching for the best fitting one. To get further from benchmark models with all covariates we use the algorithm *Glmulti* for finding the best fitting model. The algorithm has two different settings: brute force and genetic. The brute forcing type goes trough every possible variation of given set of explanatory variables to predict the given outcome variable. After going through all possible models we can obtain the best fitting model by the value of the Bayesian information criterion. The algorithm itself does not fit any models. It just produces model formulas and passes them onto desired model fitting function of R software.

The genetic version of the algorithm is more complicated than simply going through all the possible combinations. However, it can be crucial when the number of candidate models is too large for brute forcing. Going through every single combination of variables can quickly get over the computing limitations especially when we allow the algorithm to form models with interacting variables.

The genetic algorithm picks an adjustable size of population of models. In every generation, models are fitted and the values of Bayesian information criterion are used to calculate fitness of the model, ω. The fitness of the ith model is

$$\omega_i = \exp(-(IC_i - IC_{best})), \tag{2.19}$$

where IC_{best} is the best value of Bayesian information criterion in the current population of models. Higher IC value means lower fitness for the model.

The genetic algorithm contains three different methods that produce models for the next generation: asexual reproduction, sexual reproduction and immigration. The rates of sexual reproduction and immigration can be controlled via given parameters. A model which is a product of asexual reproduction is a copy of its parent which will contain a component of mutation. Mutation probability of a component in the model is an adjustable parameter for asexual reproduction. Asexual reproduction models are drawn randomly from the parent generation with a probability proportional to fitness. A model which is a product of sexual reproduction has two parent models which are selected also randomly from the parent generation with a probability proportional to fitness. This selection of parent models guarantee the convergence of the algorithm to the best fitting model. Components of parent models are combined to the model produced by sexual reproduction. Mutation is possible too in the method of sexual reproduction. A model which is a product of immigration has the state of each component chosen randomly with equal probability. Immigration produced models play important part in the genetic version of the algorithm, because they have the biggest changes in the structure of the models which are fitted. This means that immigration is a way to avoid being stuck around a local optimum of Bayesian information criterion value and improves convergence in many cases (Yang [21]).

The genetic version of the algorithm has three different adjustable stopping rules as parameters. The first is a target improvement in the best values of Bayesian information criterion. The second is target improvement in the average values of Bayesian information criterion. The third parameter is the number of every consecutive 20 generations the targets can be failed to be fulfilled until stopping the algorithm. This means that if the observed improvements are below the given target values then the genetic algorithm is declared not to have significantly improved. This routine is checked every 20 generations. If during the amount of given consecutive 20 generations the routine has not found any significant improvement, the algorithm stops.

In the analysis where the genetic version of the algorithm was used, we ran it through a few hundred times since it does not always converge to the same outcome. Following parameter values were chosen randomly between two values. For population size of 100-150 the mutation probability was 0.001-0.1, the sexual reproduction rate was 0-0.2, the immigration rate was 0-0.5, the stopping rule of target best IC was 0-0.5, the stopping rule of target average IC was 0-0.5, and the stopping rule of every consecutive 20 generations the targets can be failed was 2-5. These parameter values were suggested by Calcagno & de Mazancourt [11].

Figure 1 illustrates the operation of the algorithm when applied to dataset A, where each dot is one fitted model. The algorithm has brute forced through all possible 256 candidate models and we can see that there exists one model with truly lower value of the Bayesian information criterion than the other models. The red line indicates the two units of the Bayesian information criterion away from the best fitting model.

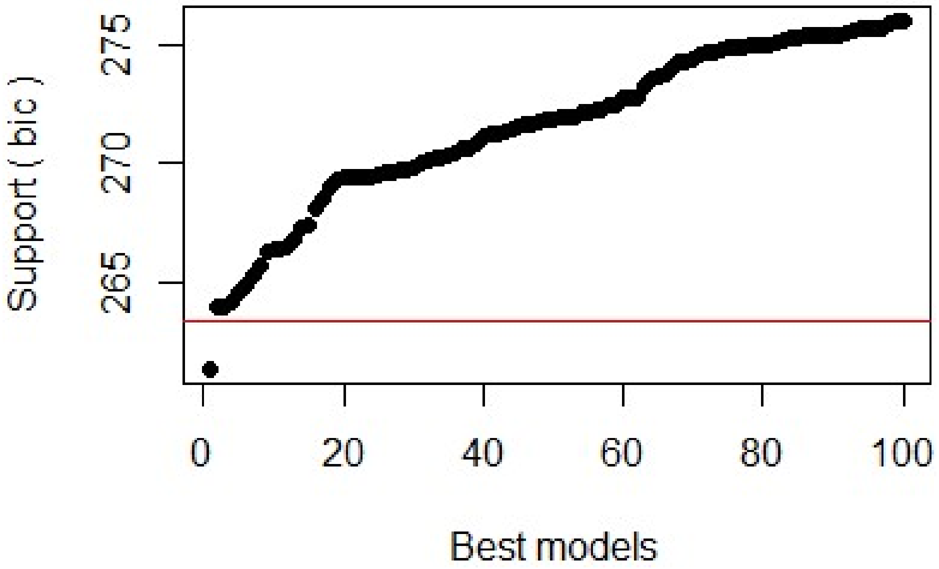

Figure 1: The values of the Bayesian information criterion for models formed from dataset A using the algorithm *glmulti* on brute force setting.

3 Model Diagnostics

In the previous chapter we already introduced a method to find a model that is considered to have the best fit using the value of the Bayesian information criterion. However, the fit is not the only measurable feature a model can have. In this chapter we demonstrate methods to further evaluate the competency of the logistic regression model.

3.1 Checking Collinearity

This section is based on Agresti [20]. In the logistic regression model with multiple covariates can happen a phenomenon called collinearity. It means that one covariate has exact linear dependence to one or multiple other covariates. In this case we use the term collinearity also when there is a near linear dependency. The source of collinearity is the data and it is not caused by the formed model.

Logistic regression is sensitive to significant collinearity between any of the independent explanatory variables in the model. However, collinearity does not reduce the predictive power or correctness of the model. Collinearity mainly affects inference regarding individual covariates. For example an explanatory variable can lose its significance if another predictor in the model has collinearity with it. This is critical to the analysis, because we specifically want to treat the variables as individuals to get more accurate results.

To check for any possible collinearity among the covariates in the model we use *the variance inflation factor* (VIF). The measure of the variance inflation factor for covariate x_i is

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}, \tag{3.1}$$

where R_i^2 is the coefficient of determination for the regression of x_i on all remaining independent variables included in the model. The variance inflation factor is the multiple by which the variance increases, because the other covariates are correlated with the covariate x_i. The minimum value of the variance inflation factor is 1, which would imply no collinearity between the covariate x_i and the remaining covariates. Values exceeding 10 indicate serious collinearity which should be handled (Menard [22]).

If collinearity occurs in a model, one approach is to remove those covariates that have significant collinearity. Removing such redundant covariates can reduce the standard errors of the other estimated effects. Another method to treat collinearity is to combine the collinear covariates especially when they are indicators of a common feature.

3.2 Predictive Power

In this section we introduce two methods to measure predictive power of a model: *classification table* and *receiver operating characteristic curve*. Even though a model can have a good fit it does not mean the model classifies well. Also, accurate or inaccurate classification does not address the criteria for a good fit of the model (Hosmer et al. [14]).

3.2.1 Classification Table

This subsection is based on Hosmer et al. [14]. In the method of classification table, the built model cross-classifies the dichotomized observed outcome y_i with a prediction $\hat{y}_i$ of whether $y_i = 0$ or $y_i = 1$. The prediction $\hat{y}_i$ for observation i is $\hat{y}_i = 1$, when $\hat{\pi}_i > \pi_0$ and $\hat{y}_i = 0$, when $\hat{\pi}_i < \pi_0$. The probability π_0 is the pre-selected cut-off value. Typically the cut-off point is set $\pi_0 = 0.50$.

The classification table forms a 2×2 table as an outcome where we see the number of correct and false classifications (Table 1). The classification table yields measures of *sensitivity* and *specificity*. Sensitivity is defined as $P(\hat{y} = 1 \mid y = 1)$, which is obtained from Table 1 using formula $a/(a + c)$. Specificity is defined as $P(\hat{y} = 0 \mid y = 0)$, which is obtained from Table 1 using formula $d/(b + d)$. A model with high sensitivity classifies observations with positive outcome well and a model with high specificity classifies observations with negative outcome well. These are not exclusive, so a model can classify the observed outcome well despite the value of the outcome.

Table 1: An example of a classification table.

		Observed outcome y		
		1	0	Total
Prediction $\hat{y}$	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	

Measures such as sensitivity and specificity derived from a 2×2 classification table depend heavily on the distribution of the estimated probabilities in the observed data. That is why the fit of a model should not be relayed on a classification table. These measures might depend entirely on the ratio of the observed outcome y rather than a correctness of a model.

The disadvantage of the classification table is that the predictions are highly dependant from the value of the cut-off term π_0. Using only the classification table when measuring predicting power of a model can give biased results, but it is suitable when the only goal of an analysis is classification. In the case of logistic regression model, classification table can be used as a supplement tool when assessing predicting power of the model.

3.2.2 Receiver Operating Characteristic Curve

The classification table decided the outcome of the prediction $\hat{y}$ by only one cut-off point whereas the receiver operating characteristic curve is more informative method for measuring predictive power of a model. It takes into consideration the estimated sensitivity and specificity for all the possible values of cut-off point π_0. The receiver operating characteristic curve is normally displayed in a plot as a concave line connecting points (0, 0) and (1, 1). The bigger the area under the curve (AUC) in the plot is, the better predictive power the model has. These principles are illustrated in Figure 2 (Agresti [20]).

The sensitivity is also known as the true positive rate. The false positive rate is $P(\hat{y}_i = 1 \mid y_i = 0) = (1 - \text{specificity})$. The receiver operating characteristic curve is defined as a plot of the true positive rate as a function of the false positive rate when the values of the cut-off point π_0 decreases from 1 to 0. For example, when the cut-off point π_0 is close to the value 1, then almost all predictions are $\hat{y}_i = 0$. In contrast, when the cut-off point π_0 is close to the value 0, then almost all predictions are $\hat{y}_i = 1$ (Agresti [20]).

The area under curve takes values between 0 and 1. In general, if the area under curve is equal to 0.5, then prediction shows no discrimination. There exists no strict thresholds, but only guidelines to categorize the values of the area under curve. We consider

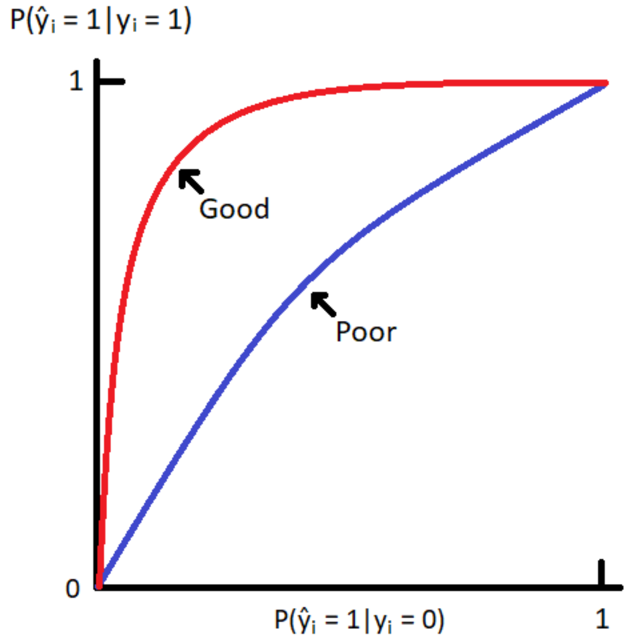

Figure 2: An example of two receiver operating characteristic curves. One indicating good predictive power and other indicating poor predictive power.

$0.7 \leq \text{AUC} < 0.8$ to be acceptable discrimination and $0.8 \leq \text{AUC} < 0.9$ to be excellent discrimination (Hosmer et al. [14]).

3.3 Goodness of Fit

The section is based on Hosmer et al. [14] unless stated otherwise. Even though we have found a well fitted model with purposefully selected covariates using the Bayesian information criterion, we still do not know whether the probabilities produced by the model accurately reflect the observed outcome variable. To answer that, we use a statistical goodness of fit test. It gives us a summary measure for a model whereas selecting covariates is examination of the individual components of a model.

To find out whether the predicted values of the model are accurate compared to the observed values, we chose to use *the Hosmer–Lemeshow goodness of fit test* (Hosmer & Lemeshow [23]). There are a few limitations in the Hosmer–Lemeshow goodness of fit test. The test does not measure the actual amount of goodness of fit, it just detects if there is significant lack of fit. Additionally, if there is occurrence of poor fit, the test does not tell the cause of it. The test can only be performed for the fitted values determined by the covariates in the model and not for the set of all available covariates. This means that each model has to be tested separately after the desirable set of covariates has been found.

To obtain the test we set null hypothesis

$$H_0 : \text{Cannot conclude that model does not fit}$$

and the alternative hypothesis

$$H_1 : \text{Model does not fit.}$$

So the desirable outcome from the the Hosmer–Lemeshow goodness of fit test is a large p-value that indicates adequacy of the model.

Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ be the i:th observation of the explanatory variables, $i = 1, \dots, n$. It can happen that $x_i = x_l$ for some $i \neq l$, so that the vector of explanatory variables are equal in different observations. We denote with S the number of unique values of observed x_i, $i = 1, \dots, n$, and note that it may happen that $S < n$. We adopt notation m_S for all the observations $x_i = x_s$, $i = 1, \dots, n$ and $s = 1, \dots, S$. Further, let y_S be the number of $y = 1$ outcomes among the m_S observations. From this we can conclude that $\sum_{s=1}^{S} m_s = n$ and $\sum_{s=1}^{S} y_s = n_1$, where n_1 denotes the number of $y = 1$ outcomes in the data. After this modification, we proceed to group the data into g groups. For each observation i we assign an estimated probability $\hat{\pi}_s = \hat{P}(y_i = 1 \mid x_i = x_s)$ and then order the observations from the smallest estimated probability to the largest. The g groups are formed then with the first group containing $n_1' = \frac{n}{g}$ observations with the smallest values of $\hat{\pi}_s$ and the last group $n_g' = \frac{n}{g}$ observations with the largest values of $\hat{\pi}_s$, $s = 1, \dots, S$. The Hosmer-Lemeshow goodness of fit statistic, $\widehat{C}$, can be obtained from the following formulas

$$\begin{aligned}
\widehat{C} &= \sum_{k=1}^{g} \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right] \\
&= \sum_{k=1}^{g} \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{n_k' \bar{\pi}_k} + \frac{(n_k' - o_{1k} - (n_k' - \hat{e}_{1k}))^2}{n_k'(1 - \bar{\pi}_k)} \right] \\
&= \sum_{k=1}^{g} \left[\frac{(o_{1k} - n_k' \bar{\pi}_k)^2}{n_k' \bar{\pi}_k (1 - \bar{\pi}_k)} \right],
\end{aligned} \tag{3.2}$$

where

$$o_{1k} = \sum_{s=1}^{c_k} y_s,$$

$$o_{0k} = \sum_{s=1}^{c_k}(m_s - y_s),$$

$$\hat{e}_{1k} = \sum_{s=1}^{c_k} m_s\hat{\pi}_s,$$

$$\hat{e}_{0k} = \sum_{s=1}^{c_k} m_s(1 - \hat{\pi}_s),$$

$$\bar{\pi}_k = \frac{1}{n'_k}\sum_{s=1}^{c_k} m_s\hat{\pi}_s$$

and c_k is is the number of covariate variations in the kth group. The Hosmer-Lemeshow goodness of fit test statistic $\widehat{C}$ asymptotically follows a χ^2 distribution with $g-2$ degrees of freedom.

The parameter g is the number of groups and it is commonly set to be $g = 10$. The Hosmer-Lemeshow goodness of fit test is practical, because it outputs a single value which is not hard to interpret. It gets the most accurate results when the S is large and both outcomes ($y_i = 0$ and $y_i = 1$) are frequently occurring. The probability of missing an important deviation from the fit is higher in the the process of grouping with a small data. The simulation results reported in the literature of Canary et al. [24] indicate that the Hosmer-Lemeshow goodness of fit test is not especially accurate for data sizes $n < 400$.

The p-value obtained from the Hosmer-Lemeshow goodness of fit test should not be used for selecting covariates or comparing models. If two models both have p-value over 0.05 from the Hosmer-Lemeshow goodness of fit test. One is not better than the other, the conclusion is that the both models fit. Only if a third model has p-value under 0.05 then we would favour the other two models.

3.4 Influence Diagnostics

In this section the main source is Agresti [20] unless stated otherwise. After assessing the goodness of fit, the last phase in building a logistic regression model is to check individual observations that are extremely influential on the model. Maximum likelihood estimation method is sensitive to influential observations and so only a few exceptional observations can undermine the correctness of the model (Pregibon [25]). We try to identify leverage and outliers that are significant in influence diagnostics. Then we decide whether the specific observations should be included in the model.

If an outcome value of an observation does not follow the general trend of the model, we call that observation an outlier. If an observation i has extreme values of predictor variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, we say that the particular observation has high leverage. In this case extreme means unusually low or high values. Even abnormal combination of values of predictors can cause high leverage.

There are multiple ways to identify influential observations and in this case we use the *standardized residuals* and the *Cook's distance.* Residual is the measure of difference between the observed value and the estimated value. The standardized residuals have similarities with the Pearson's residuals, but they take leverage into account. The Cook's distance uses leverage too, because it employs standardized residuals.

3.4.1 Residuals

First we introduce the Pearson's residual. For observation i with observed outcome value y_i the Pearson's residual e_i is

$$e_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\widehat{\text{var}}(y_i)}} = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}}, \tag{3.3}$$

where $\hat{\pi}_i$ is the estimated probability for the outcome value y_i.

For the standardized residual we need to acquire the measure of leverage $\hat{h}_{ii}$. It can be obtained from the so-called hat matrix for logistic regression

$$\hat{\mathbf{H}}_{\mathbf{W}} = \hat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}^{1/2}, \tag{3.4}$$

where the weight matrix $\hat{\mathbf{W}}$ is $n \times n$ diagonal matrix with element $\hat{w}_{ii} = n_i\hat{\pi}_i(1-\hat{\pi}_i)$, $i = 1, \ldots, n$ and $\mathbf{X}$ is defined in equation (2.1). The standardized residual for ith observation is

$$\begin{aligned} r_i &= \frac{e_i}{\sqrt{1-\hat{h}_{ii}}} \\ &= \frac{y_i - \hat{\pi}_i}{\sqrt{[\hat{\pi}_i(1-\hat{\pi}_i)(1-\hat{h}_{ii})]/n_i}}. \end{aligned} \tag{3.5}$$

The advantage of the standardized residuals compared to the Pearson's residuals is that the standardized residual approximate $N(0,1)$ and more appropriately recognise redundancies. Absolute standardized residual values of larger than about 3 provide evidence of significant influence.

3.4.2 Cook's Distance

The Cook's distance is based on the change in the estimated parameter values $\hat{\boldsymbol{\beta}}$, when the specific observation i is removed from the data. As said, the Cook's distance for observation i uses the leverage $\hat{h}_{ii}$ obtained from the hat matrix $\hat{\mathbf{H}}_{\mathbf{W}}$ and the standardized residual r_i. The measure of Cook's distance D_i for observation i is

$$D_i = r_i^2 \left[\frac{\hat{h}_{ii}}{u(1 - \hat{h}_{ii})} \right], \tag{3.6}$$

where u is the number of coefficients in the model.

Martin & Pardo [26] suggests using $(2u/n)$ as the threshold value when deciding whether an observation is extremely influential. If an observation has the evidence to be extremely influential by either of the shown methods, we can proceed to remove the observation from the model. After the removal we need to refit the model again. Then we can make comparisons and evaluate the robustness of the models.

4 Data

After radical prostatectomy, all the patients are checked routinely to measure their prostate-specific antigen. The outcome variable *Relapse* is true if the patient has had prostate cancer relapse. The relapse is defined as two sequential measurements of high prostate-specific antigen (>0.2 ng/mL) after surgery. We try to explain whether the relapse can be predicted with biomarkers.

The data contains two different cancer classifications. If the cancer was already spotted by a radiologist on the magnetic resonance imaging (MRI), the cancer is classified as MRI visible. The area where the cancer was found in MRI is called a region of interest. Otherwise the cancer was MRI invisible and was found afterwards by a pathologist. We split the data in three different datasets: Dataset A, MRI visible exclusively (n=274), Dataset B, MRI invisible exclusively (n=168), and the third Dataset C is MRI visible and invisible combined (n=130). Observations in dataset C must have had cancer lesions spotted in the MRI by the radiologist and afterwards by the pathologist elsewhere than in the region of interest.

An observation is dropped from the analysis if it has at least one null value in any used variable. In this data, there are less observations for MRI invisible cancers. The combined dataset C is naturally the smallest as the patient must have had a MRI visible cancer and a MRI invisible cancer to be in it.

The dataset contains four biomarkers for which we have both a visually provided value by the pathologist and an artificial intelligence produced value. In total we have 8 variables to use as logistic regressions predictors in datasets A and B. Dataset C contains 16 explanatory variables which are all the explanatory variables from datasets A and B. The biomarkers are called: AR, ERG, PTEN and Ki67. Each biomarker behaves and reacts differently which is why the values of each biomarker in some of variables have different quantity.

If a patient have had a spotted cancer by the radiologist, the region of interest is inspected in 1-3 tissue microarray cores. The MRI invisible cancers have 1 core for examination. The cores are then cut to 4 μm-thick sections for the biomarker staining. While some patients have multiple values for a single MRI visible cancer variable and some only one, we have to modify them into a single value without losing information.

Biomarker AR, androgen receptor, is known for its ability to be the driver of prostate cancer progression. Usual non-surgical treatment option for advanced prostate cancer is androgen deprivation therapy (Wadosky & Koochekpour [27]). The variable AR visual has a fuse score given for each 1-3 inspected core. Fuse score is a product of strength and percentage scores. Strength is the staining intensity of the biomarker AR. It takes the following values: 0 indicates negative value, 1 indicates low value, 2 indicates intermediate value and 3 indicates strong values. The strength value is then multiplied by the percentage of the biomarker AR positive nuclei. To have only one value from these multiple fuse scores, we use the average of them in the analysis. For AR visual to be relative to other variables we divided the final value by three to be in the range of percentages and not 0-300. In the variable AR AI the artificial intelligence has counted the amount of biomarker AR positive and negative nucleus. The value we use in the variable AR AI is positive nucleus percent.

Biomarker ERG, ETS-related gene, is an indicator for tumour carrying the transmembrane protease serine 2 gene and ERG fusion. It is the most common gene alteration in prostate cancer, but there is no prognostic value for its tissue-based detection proven (Wang et al. [6]). The variable ERG visual is dichotomized and biomarker ERG positive in one of the three cores means that the variable ERG visual gets value 1. The variable ERG AI sums biomarker ERG positive and negative nuclei of all cores and then takes ratio of positive to negative on a $\log_{10}$ scale.

Biomarker PTEN, phosphatase and tensin homolog, is a tumour suppressor gene. The inactivation of PTEN changes gene expression in prostate cancer and is related to higher Gleason score, lower disease-specific survival time and greater probability of secondary therapies after radical prostatectomy (Lahdensuo et al. [28]). The variable PTEN visual is dichotomized and gets value 1 if all the cores are biomarker PTEN positive. If even one of the cores is biomarker PTEN negative, the value of PTEN visual is then 0. The variable PTEN AI does not count nuclei but biomarker PTEN positive and negative areas. The variable PTEN AI sums biomarker PTEN positive areas and negative areas of all cores and then takes ratio of positive to negative in mm^2 on a $\log_{10}$ scale.

Biomarker Ki67 is regularly used to measure cell proliferation. Earlier analysis proved that high expressions of biomarker Ki67 is associated with unfortunate outcome of prostate cancer for example death of metastasis (Berlin et al. [8]). The variable Ki67 visual is mean percentage of all biomarker Ki67 positive nuclei in cores. Estimating percentages visually is more demanding task than obtaining the values for the other visual variables. The variable Ki67 visual has interesting side note linked to it in the data: "Mostly a quick test to have something to measure against the AI output". The variable Ki67 AI is mean percentage of nuclei in cores that are biomarker Ki67 positive.

Table 2: Summary of the Dataset A.

n=274, Relapsed=51 (19%)				
Variable	Min	Max	Median	Mean
AR visual	0.0	98	54	53
AR AI	12	98	91	85
ERG visual	0	1	0	0.27
ERG AI	-3.8	1.2	-2.3	-1.8
PTEN visual	0	1	1	0.62
PTEN AI	-2.5	6.5	2.2	1.9
Ki67 visual	0	12	1.7	2.0
Ki67 AI	0.026	21	3.0	3.8

5 Analysis

The analysis is done using R software (R Core Team [12]) including packages: *broom* ([29]), *caret* ([30]), *car* ([31]), *pROC* ([32]) and *ROCit* ([33]).

We want to know if the biomarkers could predict prostate cancer relapse. Besides the predictive power of the biomarkers, which is the main question, we compare MRI visible cancers to MRI invisible and artificial intelligence to visual. We use logistic regression model to predict the dichotomized relapse variable with the four biomarkers and the total of 16 explanatory variables, which 8 concern Dataset A and 8 concern Dataset B. These variables are explained in chapter 4.

First we analyze the dataset A, then we analyze the dataset B, and lastly we analyze the dataset C. In the analysis of dataset A we use only variables from MRI visible cancers (n=274) in the statistical models. In the analysis of dataset B we use only variables that are from MRI invisible cancers (n=168) in the statistical models. The third case, analysis of dataset C (n=130), uses variables from both dataset A and dataset B in the statistical models. The models formed from dataset A are considered the main models, because of the lack of data, we cannot afford to lose any more observations.

In all three cases we begin the analysis by building a benchmark model with all explanatory variables included. For comparison of artificial intelligence and visual we use models with separated artificial intelligence and visual covariates. Possible collinearity is checked via variance inflation factor. We use values of Bayesian information criterion to rank the models. The covariates in the best fitting models are chosen by algorithm *glmulti* by lowest possible values of Bayesian information criterion. To check the goodness of fit we use the Hosmer–Lemeshow test with $g = 10$ and for predictability we use receiver

Table 3: The estimates of Model A1, the benchmark model of dataset A.

Variable	Estimate	OR (95% CI)	p-value
AR visual	0.0024	1.00 (0.98 - 1.02)	0.819
AR AI	0.0029	1.00 (0.98 - 1.03)	0.839
ERG visual	-0.77	0.46 (0.11 - 2.03)	0.304
ERG AI	-0.14	0.87 (0.52 - 1.42)	0.587
PTEN visual	0.31	1.37 (0.47 - 4.17)	0.570
PTEN AI	-0.30	0.74 (0.54 - 1.00)	0.052
Ki67 visual	0.16	1.18 (0.99 - 1.40)	0.062
Ki67 AI	0.14	1.15 (1.04 - 1.28)	0.005*

operating characteristic curve and its area under the curve. We use standardized residuals and Cook's distance for checking if the models contain extremely influential observations. These methods we employ in the analysis are explained in sections 2 and 3. The p-value significance threshold is set to be 0.05 in the logistic regression models.

Section 5.1 holds the analysis of dataset A. In subsections 5.1.1 and 5.1.2 are the best two competing models for dataset A. Comparison of visual and artificial intelligence is found in subsection 5.1.3. Dataset B is treated in section 5.2. Lastly the analysis of the combined dataset C is in section 5.3. Appendix A contains the R software printouts of the estimated models.

5.1 Logistic Regression for Dataset A

In this section we analyze dataset A which has all MRI visible cancer variables as predictors. After seeing how each variable performs in the benchmark model A1 we let the *glmulti* algorithm calculate us a model with the best predictors. For dataset A we build the additional models using artificial intelligence and visual variables.

Next we interpret the Model A1, see Table 3 for reference. The variables for biomarker AR seem to be bad predictors on this model. The estimates for the variables AR AI and AR visual are very close to zero and the p-values indicate that neither of them has no effect on the relapse. The estimates for the ERG AI and ERG visual variables are negative which indicates that positive values of ERG variables lead to low probabilities of relapse. The biomarker PTEN is the only biomarker that has conflict between the artificial intelligence and visual variables. These variables of biomarker PTEN differ in the effect as well as in the predictive ability. The variable PTEN visual indicates that positive values of it lead to high probabilities of relapse whereas variable PTEN AI indicates that positive values of it lead to low probabilities of relapse. The variable PTEN visual has the largest odds ratio confidence interval so the odds ratio has the lowest precision.

The only variable with statistically significant p-value is Ki67 AI. In this model it is a very good predictor for the probability of relapse compared to the others. It has small confidence interval of odds ratio and a very small p-value. The estimate and the odds ratio are strongly towards relapse as the variables of Ki67 are in percentages. One percent of Ki67 AI positive has 1.15 higher odds than zero percent of Ki67 AI. The estimates suggest that the variable Ki67 visual has more effect in the model than the variable Ki67 AI per percent of positive biomarker Ki67. The variable Ki67 visual has conservative values within a small gap whereas the values of variable Ki76 have wider range. Therefore, we observe a larger estimate and larger odds ratio for the former variable.

The benchmark model of dataset A with all covariates, Model A1, have area of 0.708 under the receiver operating characteristic curve. That is considered to be mediocre predictive power. The value of Bayesian information criterion of Model A1 is 288 and the Hosmer-Lemeshow goodness of fit test suggests that there is no evidence for poor fit with p-value of 0.50. There was no significant collinearity, because when the highest value of variance inflation factor is 3.5.

5.1.1 Best Model for Dataset A

When the algorithm *glmulti* chooses the covariates in the dataset A by the lowest value of Bayesian information criterion, the outcome is Model A2 (Table 4) that contains only one covariate Ki67 AI with Bayesian information criterion value 261. Figure 3 illustrates the regression curve of the Model A2. Also, the second best model would have had also only one covariate, Ki67 visual with Bayesian information criterion value 267. We decide that the second best model is not worthwhile, because the difference between values of Bayesian information criterion is greater than two units (Calcagno & de Mazancourt [11]). The algorithm with the brute force setting generates all possible 2^p candidate models, where p is the number of the explanatory variables. A single run of the algorithm takes only a few seconds on average with 8 explanatory variables.

Figure 4 gives the distribution of the variable Ki67 AI in Model A2 between the non-relapsed (0) and the relapsed (1). Figure 5 depicts the distribution of the variable Ki67 visual. We observe that the values of the biomarker Ki67 (both visual and artificial intelligence) are a tiny bit higher for the relapsed than for the non-relapsed. The difference between the relapsed and non-relapsed is larger in the variable Ki67 AI.

The area under the receiver operating characteristic curve for Model A2 with algorithm chosen covariates is 0.64 and the Hosmer-Lemeshow goodness-of-fit test suggests no evi-

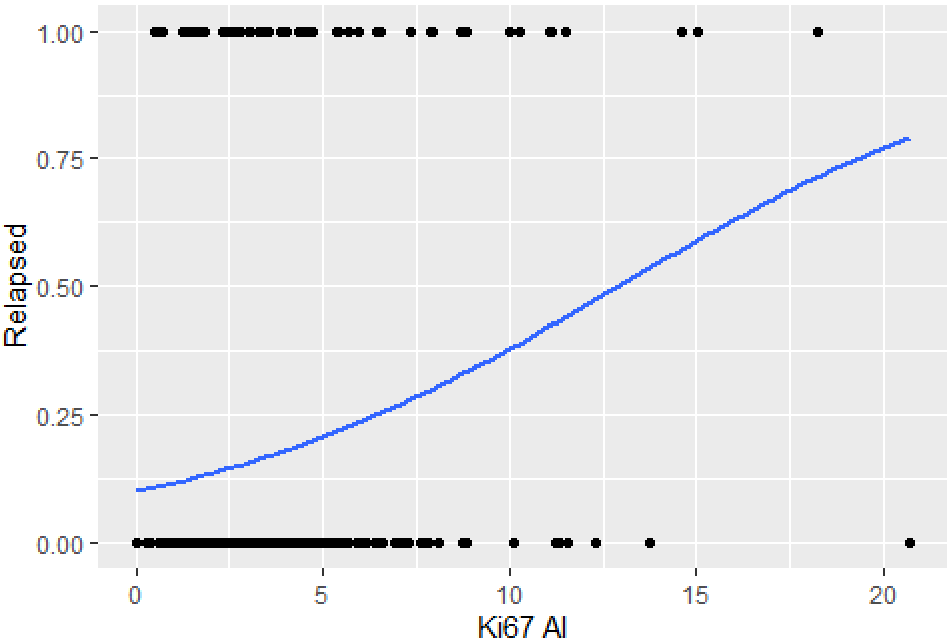

Figure 3: Regression curve of the Model A2 which contains only the variable Ki67 AI.

Table 4: The estimates of the Model A2.

Variable	Estimate	OR (95% CI)	p-value
Ki67 AI	0.17	1.19 (1.08 - 1.31)	0.00033*

dence of poor fit with p-value of 0.79.

We begin to check if the model contains any extremely influential observations. Figure 6 shows that non-relapsed samples have very small spread of standardized residuals and almost all are packed on the same line between 0 and -1. The relapsed samples on the top however have notably higher standardized residuals as they are settled between 1 and 2. We cannot conclude that the model contains any influential observations by the standardized residuals since none of them exceeds the limit of value 3 (Hosmer et al. [14]).

The situation of extremely influential observations is different when we consider the Cook's distances in Figure 7. One observation stands out heavily and we can proceed to remove it from the model. The same observation indexed 202 can be seen at the bottom in Figure 6. It is a non-relapsed observation with the largest value of variable Ki67 AI (20.7) in the

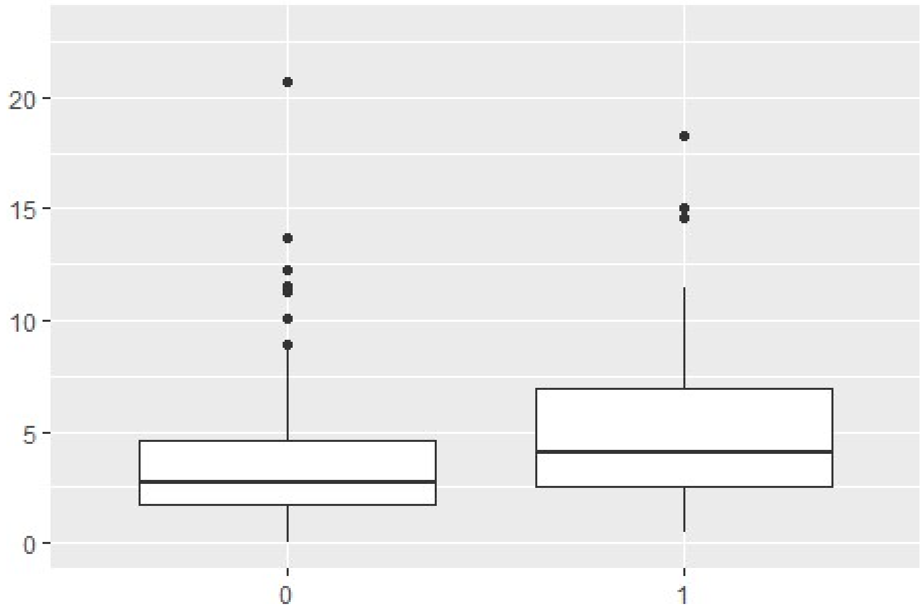

Figure 4: The distribution of the variable Ki67 AI in Model A2 with non-relapsed (0) and relapsed (1).

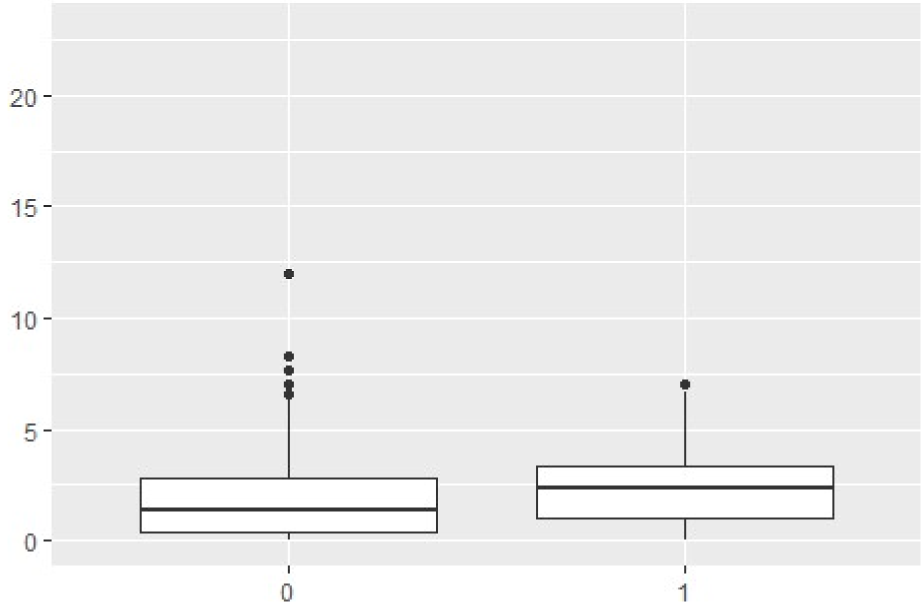

Figure 5: The distribution of the variable Ki67 visual with non-relapsed (0) and relapsed (1).

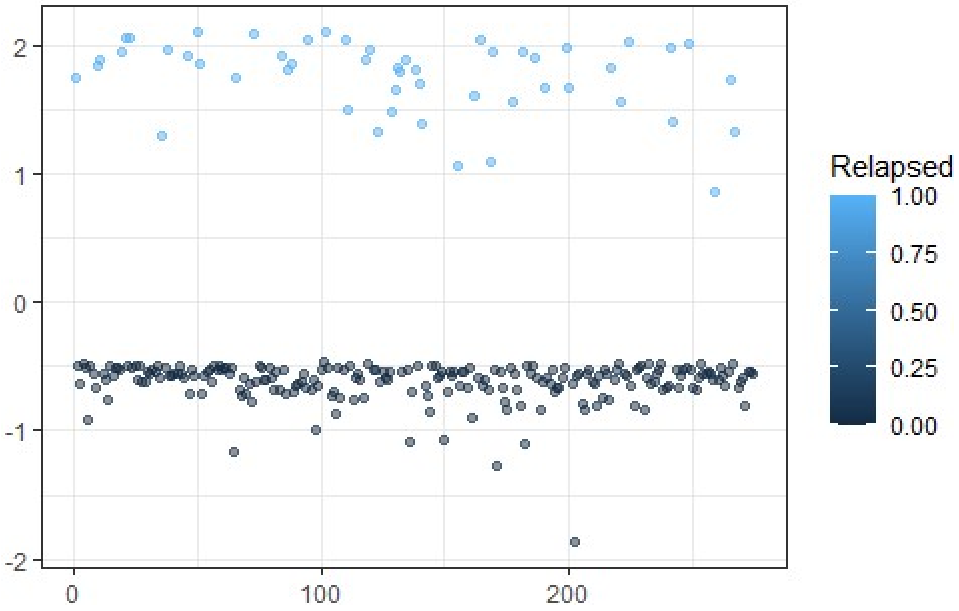

Figure 6: Standardized residuals for Model A2.

data. When the observation is removed from the Model A2, we fit the model again. The new Bayesian information criterion value is 258 and the area under the receiver operating characteristic curve is 0.65.

The Bayesian information criterion value suggests that the algorithm chosen Model A2 is indeed better compared to the Model A1. It may not be as usable because the area under the receiver operating characteristic curve is drastically better in Model A1 with all covariates. Even the benchmark Model A1 has the area under the receiver operating characteristic curve of 0.71, which is acceptable. The literature states that models with area under the curve equal or larger to 0.8 predict excellently (Hosmer et al. [14]). However, the value of area under the receiver operating characteristic curve gets punished in Model A2 for the low number of covariates. This is especially true in this data where many relapses still happen with low values of Ki67 and relapsed observations have higher standardized residuals than others. We conclude that the relapse can not be trustfully explained by only one biomarker or covariate.

5.1.2 Best Model for Dataset A With Interacting Covariates

We return back to the whole dataset A and modify the set of covariates by allowing all possible two variable interaction terms. Interaction terms increase the number of candidate models to 2^{p^2}, where p is the number of the explanatory variables so we use the

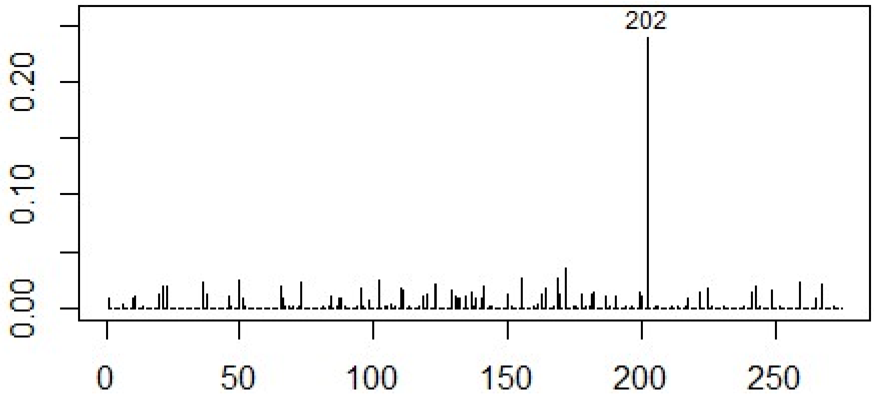

Figure 7: Cook's distance for Model A2.

Table 5: The estimates of Model A3.

Variable	Estimate	OR (95% CI)	p-value
Ki67 AI	0.21	1.23 (1.10 - 1.38)	0.00043*
Ki67 visual : ERG AI	-0.11	0.90 (0.84 - 0.95)	0.00069*
Ki67 AI : PTEN AI	-0.034	0.97 (0.94 - 0.99)	0.022*

genetic version of the algorithm *glmulti*. While $p = 8$, the number 2^{8^2} is so large that we would have never gone through all the candidate models with the brute force setting of the algorithm. A single run of the algorithm with the genetic setting takes 15 seconds on average. The algorithm was run through a few hundred times with different parameter values since it does not always converge to the same outcome. All in all, the genetic setting made it possible that the computing could be done in the matter of hours rather than taking almost an infinity.

When the algorithm has two-level covariates enabled, it no longer results in a model with only the variable Ki67 AI as its only predictor. From the previous models it is no surprise that the variable Ki67 AI is also present in this best two-level model, Model A3 (Table 5). Model A3 has Bayesian information criterion value of 258 and area under the receiver operating characteristic curve is 0.71. There is no evidence of poor fit by the Hosmer-Lemeshow goodness-of-fit test with p-value 0.39. No significant collinearity is detected by the variance inflation factor, highest value is 1.5.

All the three predictors in Model A3 have p-value under the 0.05 threshold. The variable Ki67 AI being the only one-level predictor and with positive odds ratio. The other two two-level predictors are interaction variables Ki67 AI : PTEN AI and Ki67 visual : ERG

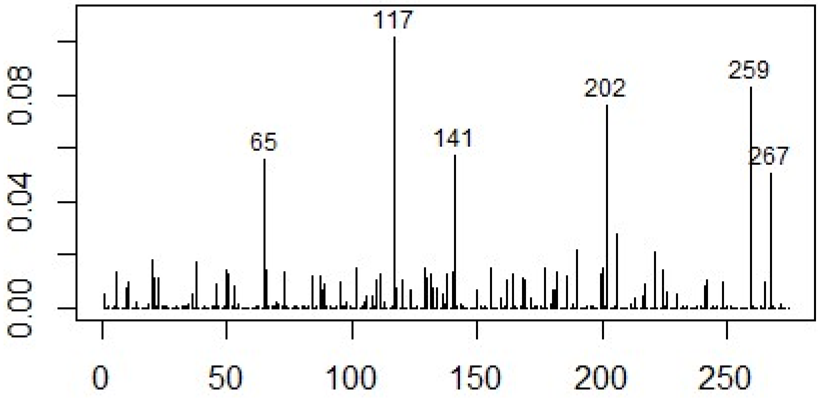

Figure 8: Cook's distance for Model A3.

Table 6: Confusion matrix of Model A3 after outlier removal with accuracy 0.86, sensitivity 0.27 and specificity 0.99.

		Observed outcome y	
		1	0
Prediction $\hat{y}$	1	13	3
	0	35	217

AI. So the Model A3 is not dominated by only artificial intelligence variables. Both the two-level predictors have odds ratio below 1.

To get the absolutely best outcome from this model we can remove extremely influential observations. This model has six of them illustrated in Figure 8 using the Cook's distance. When the model is fitted again without the influential observations, the new Bayesian information criterion value is 238 and the area under the curve is 0.72. Model fits now better but the predictability has almost no change even when compared to the benchmark Model A1 with area under the receiver operating characteristic curve of 0.71. The process of removing as much as six observations may not be absolutely necessary if they are relapsed observations which we already have very few. However, we obtained here a slightly better fitting model even though 3 of the removed 6 observations were relapsed observations.

The confusion matrix in Table 6 tells us how well the Model A3 predicts the observations with cut-off point set to $\pi_0 = 0.50$. In the best model, after removing influential observations, the brutal classification accuracy on the same data is 0.86. The model classifies the non-relapsed observations better than relapsed. This can be seen from high specificity

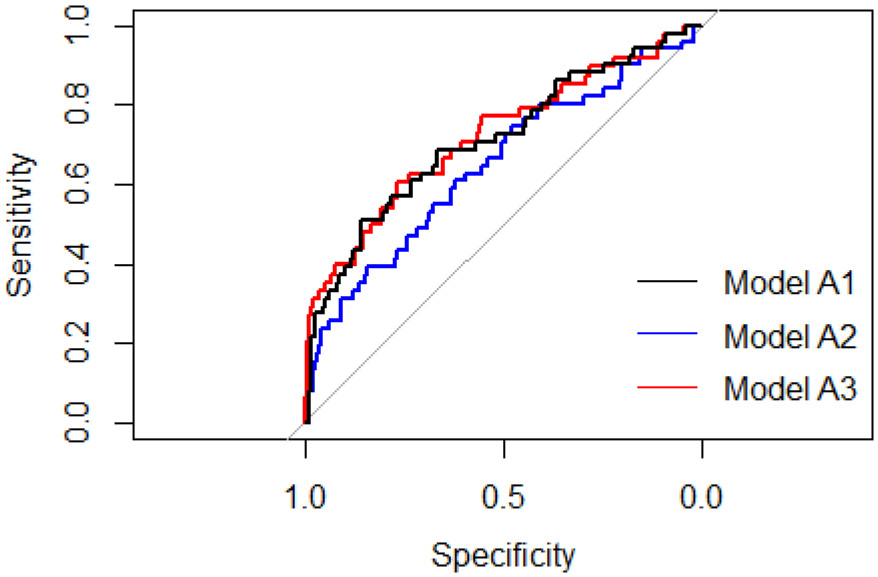

Figure 9: The receiver operating characteristic curves for Models A1, A2 and A3.

(0.99). This can be due to low amount of relapsed samples and the fact that relapses still happen at low percentages of positive Ki67 which is the only predictor which increases chances of relapse. The sensitivity of 0.27 indicates that 27 percent of relapsed observations were classified correctly.

From the Model A2 and the Model A1 with all MRI visible cancer covariates we could say that artificial intelligence is useful and predicts better the relapse. From this Model A3 we cannot conclude that artificial intelligence is purely better at predicting relapse compared to visual because the variable Ki67 visual is present in the model. Altough we could say that the visual and AI complement each other. In Figure 9 we see the receiver operating characteristic curves for Models A1, A2 and A3 where models A2 and A3 have their influential observations removed. It is visible that the predictive power of the Model A2 is being punished by the low number of covariates. The reason why the difference in predictive power between the benchmark model A1 and the best fitting Model A3 is so tiny, can be due to the imbalance of the outcome variable.

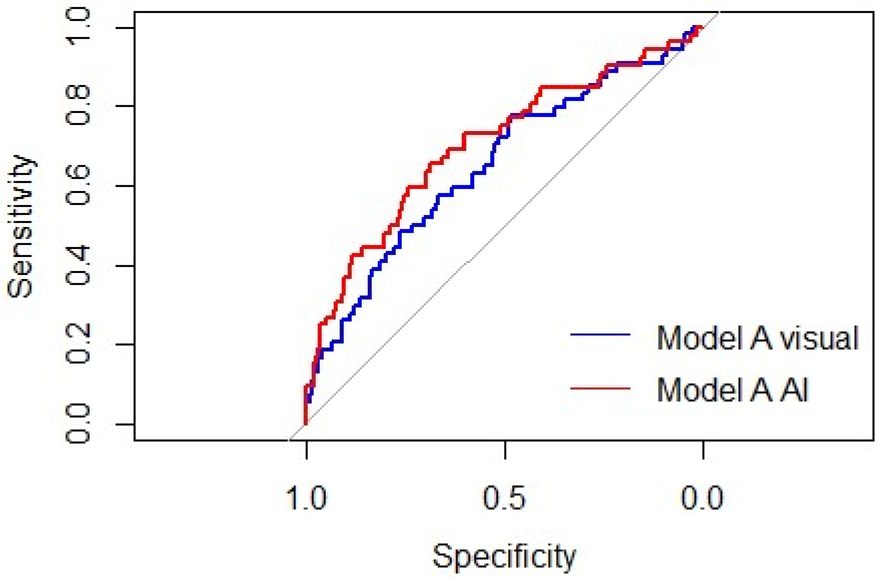

Figure 10: The receiver operating characteristic curves for Models A visual and A AI.

5.1.3 Visual Versus Artificial Intelligence in Dataset A

To even further compare the visual variables to artificial intelligence variables, we form models with only their distinctive covariates. First we consider the visual covariates case i.e., AR visual, ERG visual, PTEN visual and Ki67 visual (Table 7). The Bayesian information criterion value of the model is 278 and area under the receiver operating characteristic curve is 0.66.

Second we consider artificial intelligence covariates, i.e., AR AI, ERG AI, PTEN AI and Ki67 AI (Table 8). The Bayesian information criterion value of the model is 271 and area under the receiver operating characteristic curve is 0.69. In Figure 10 are the receiver operating characteristic curves for Models A visual and A AI. The difference between these two models is not huge but still the artificial intelligence performs better according to the Bayesian information criterion and the receiver operating characteristic curve.

5.2 Logistic Regression for Dataset B

For this part of analysis we use only the variables that are exclusive to dataset B so the values are formed from MRI invisible cancer only. The sample size of dataset B is 168. We consider that the results from this analysis are not as reliable as the ones from the dataset A analysis due to smaller sample size.

Table 7: The estimates of Model A Visual.

Variable	Estimate	OR (95% CI)	p-value
AR visual	0.01	1.01 (0.99 - 1.02)	0.283
ERG visual	-1.04	0.35 (0.14 - 0.81)	0.018*
PTEN visual	-0.46	0.63 (0.32 - 1.24)	0.178
Ki67 visual	0.22	1.24 (1.06 - 1.46)	0.009*

Table 8: The estimates of Model A Artificial Intelligence.

Variable	Estimate	OR (95% CI)	p-value
AR AI	0.01	1.01 (0.98 - 1.03)	0.631
ERG AI	-0.31	0.73 (0.54 - 0.97)	0.037*
PTEN AI	-0.22	0.80 (0.66 - 0.98)	0.028*
Ki67 AI	0.17	1.19 (1.08 - 1.31)	0.0004*

Table 9: The estimates of Model B1.

Variable	Estimate	OR (95% CI)	p-value
AR visual	0.0049	1.00 (0.98 - 1.03)	0.694
AR AI	-0.0012	1.00 (0.96 - 1.06)	0.961
ERG visual	1.24	3.45 (0.37 - 38.0)	0.287
ERG AI	-0.58	0.56 (0.21 - 1.26)	0.198
PTEN visual	2.08	8.00 (0.95 - 187)	0.098
PTEN AI	-0.31	0.73 (0.40 - 1.30)	0.298
Ki67 visual	0.13	1.13 (0.90 - 1.45)	0.274
Ki67 AI	0.066	1.07 (0.80 - 1.37)	0.619

Again the benchmark model, Model B1, contains all covariates from dataset B to explain relapse. None of the covariates are statistically significant by the p-value with 0.05 threshold. The Bayesian information criterion value of the model is 164 and area under the receiver operating characteristic curve is 0.66 (Figure 11). The Hosmer-Lemeshow Goodness-of-fit test gives no evidence for poor fit with p-value of 0.38. No significant collinearity is detected by the variance inflation factor, the highest value is 3.0.

Next we interpret the Model B1, see Table 9. Biomarker AR has almost no effect at all to the model as the odds are almost equal to 1. The difference here to the benchmark model of the dataset A, Model A1, is that AR AI has now odds less than 1.

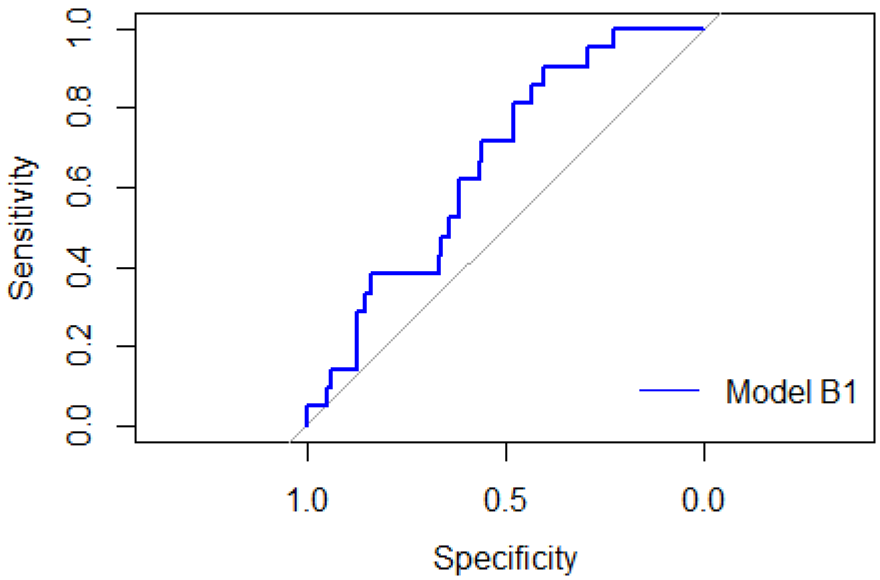

Figure 11: The receiver operating characteristic curves for Model B1.

In the Model A1 both ERG variables have negative effect to relapse. Nevertheless in this Model B1 the dichotomized variable ERG visual has odds over 3 and a wide confident interval. The value of the variable ERG AI is towards non-relapse with odds ratio of 0.56. The biomarker PTEN is similar to the biomarker ERG: The dichotomized visual variable has large odds ratio, 8.0, with wide confidence interval and the AI variable is against the relapse with odds ratio of 0.73. Both Ki67 variables have positive odds similar to the Model A1. Both AI and visual variables of Ki67 have almost the same confidence interval of odds ratio but visual have larger odds ratio, 1.13 versus 1.07.

The Model B1 with all covariates seems not to fit very well. There is not even a single one covariate with p-value lower than the 0.05 threshold. Regardless, the Hosmer-Lemeshow test gives no evidence for poor fit. Some of the visual variables disagree with artificial intelligence variables and a few have wide odds ratio confidence intervals.

Again, we search for better models using the algorithm *glmulti*. Resulting models for the best Bayesian information criterion values were same for normal covariates and interacting covariates. Both of them were null models with the Bayesian information criterion values of 132. It means that only the intercept is explanatory variable for relapse. This leads to a conclusion that no variable or biomarker from dataset B explains the relapse.

Table 10: The estimates of Model C1.

Variable	Estimate	OR (95% CI)	p-value
A AR Visual	0.0085	1.01 (0.97 - 1.05)	0.681
B AR Visual	0.0088	1.01 (0.98 - 1.04)	0.562
A AR AI	0.031	1.03 (0.96 - 1.13)	0.440
B AR AI	-0.024	0.98 (0.92 - 1.04)	0.425
A ERG Visual	-0.042	0.96 (0.05 - 19.3)	0.977
B ERG Visual	0.73	2.07 (0.04 - 75.8)	0.682
A ERG AI	-0.53	0.59 (0.19 - 1.61)	0.321
B ERG AI	-0.65	0.52 (0.14 - 1.50)	0.274
A PTEN Visual	-1.29	0.28 (0.03 - 2.36)	0.240
B PTEN Visual	3.03	20.7 (1.41 - 781)	0.051
A PTEN AI	0.14	1.15 (0.63 - 2.13)	0.657
B PTEN AI	-0.71	0.49 (0.22 - 1.02)	0.067
A Ki67 Visual	0.53	1.70 (1.22 - 2.45)	0.0025*
B Ki67 Visual	0.039	1.04 (0.77 - 1.52)	0.806
A Ki67 AI	-0.0021	1.00 (0.78 - 1.22)	0.985
B Ki67 AI	0.054	1.06 (0.73 - 1.45)	0.750

5.3 Logistic Regression for Dataset C

The combined model is built from variables of datasets A and B. This requires each observation to have values in A and B datasets. It means that the sample size gets even smaller from the previous dataset B. We consider this model, Model C1, to be the least reliable due to its small sample size. Now n=130 and only 18 of them are relapsed observations. The objective of this model is to see if there are more differences between A and B dataset variables when B variables already failed to predict relapse.

Next we interpret the Model C1, see Table 10. The letter A or B before the name of the variable defines which dataset the variable is originally from. From the benchmark model, Model C1, with all covariates we see that the variable A Ki67 visual is the only variable to pass the 0.05 p-value threshold. It has also higher odds ratio (1.7) than in the dataset A models. The value of the Bayesian information criterion of the Model C1 is 163 and area under the receiver operating characteristic curve is 0.81. The Hosmer-Lemeshow Goodness-of-fit test gives no evidence for poor fit with p-value of 0.96. No significant collinearity is detected by the variance inflation factor, the highest value is 5.3.

In the Table 10, the variable AR in all its four forms has not changed from previous

Table 11: The estimates of Model C2.

Variable	Estimate	OR (95% CI)	p-value
A ERG AI	-0.47	0.63 (0.38 - 0.97)	0.046*
A Ki67 Visual	0.48	1.61 (1.20 - 2.13)	0.00041*

models and has almost no effect to the model. The odds ratios are extremely close to value 1 and confidence intervals are very compact. The variable A ERG visual and the variable B ERG visual are both dichotomized and have oddly large odds ratio confidence intervals. The variable B ERG visual being towards relapse and the variable A ERG visual against. Both the variable A ERG AI and the variable B ERG AI are against relapse and have more precise odds ratio confidence intervals. The dichotomized visual variables of PTEN, A PTEN visual and B PTEN visual, have mixed stances. The variable A PTEN visual is against the relapse and the variable B PTEN visual is full on relapse with the most imprecise odds ratio confidence interval of the analysis, but it still almost has the p-value (0.051) under the threshold. The artificial intelligence variables of PTEN are mixed too, the variable A PTEN AI being towards relapse. In the biomarker Ki67 variables, the variable A visual has the largest odds ratio. Noticeable thing is that the variable A Ki67 AI has slightly less than one odds ratio unlike the other Ki67 variables increase the probability of relapse.

We again use the algorithm to obtain a more parsimonious model and we receive the parsimonious model, Model C2, which has two significant covariates: the variable A ERG AI and the variable A Ki67 visual. From Table 11 we see that both of them have similar odds ratios as in the model C1 although the variable A ERG AI was not significant in the Model C1. After building the Model C2 we see that not a single variable from dataset B was significant and the algorithm did not suggest any of them either.

Model C2 has the value of the Bayesian information criterion of 103 and the area under the receiver operating characteristic curve of 0.74. The receiver operating characteristic curves for Models C1 and C2 are illustrated in Figure 12. The interacting covariates were not possible due to computational reasons even for the genetic version of the algorithm.

5.4 Results

The main dataset in the analysis is A, which is the most reliable due to its sample size. We find that the variable Ki67 AI is statistically significant to predict cancer relapse in the Model A1 and ends up being the only covariate chosen by the algorithm in the Model A2. When the algorithm is allowed to use interacting covariates, we get Model

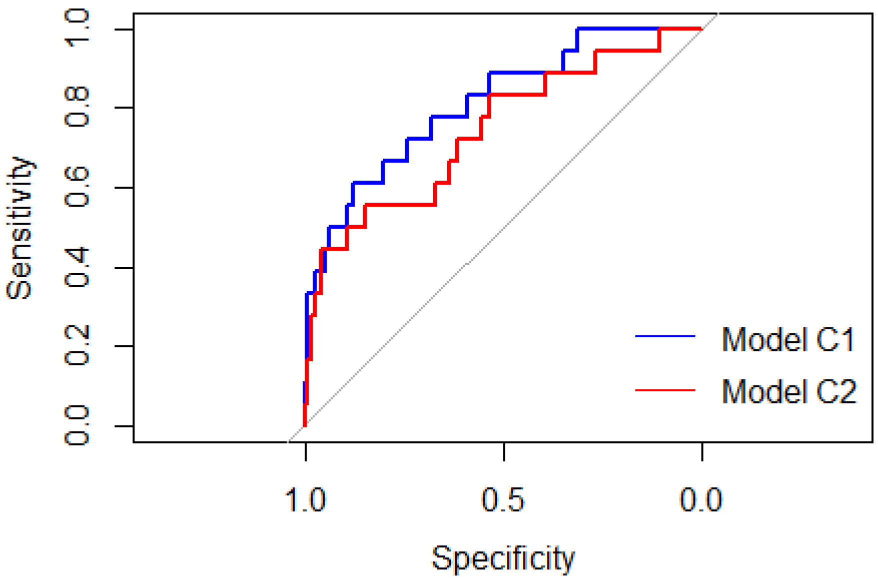

Figure 12: The receiver operating characteristic curves for Models C1 and C2.

A3 and then the covariates are Ki67 AI, Ki67 AI : PTEN AI and Ki67 visual : ERG-AI. The Model A3 is considered as the best model. Goodness of fit of the models were tested with the Hosmer-Lemeshow test using $g = 10$ and none of the three dataset A models had evidence of poor fit. Predictability of the models were measured with area under the receiver operating characteristic curve. Even the largest area under the receiver operating characteristic curve (0.72) for Dataset A models was not able to reach the desired boundary of 0.8.

To compare artificial intelligence to visual in the dataset A, two models were built, one with only artificial intelligence variables and other with only visual variables. Model that contained artificial intelligence variables had lower value of the Bayesian information criterion (271 vs 278) and larger area under the receiver operating characteristic curve (0.69 vs 0.66). The results drawn from Model A1 and Model A2 imply that artificial intelligence variables could be better predictors. From the Model A3 we cannot conclude that artificial intelligence is purely better at predicting relapse, because the variable Ki67 visual is present in it. We could say that visual and artificial intelligence variables complement each other.

From dataset A we moved on to the smaller dataset B. It has its own values formed from MRI invisible cancers. Benchmark model, Model B1 with all dataset B explanatory variables had zero statistically significant predictors. The algorithm *glmulti* suggests that the lowest value of the Bayesian information criterion for dataset B variables is a null model for normal covariates and interacting covariates. This means that no variable in dataset B explains the relapse.

Purpose of the dataset C model is to compare dataset A variables to dataset B variables although it may be the least reliable model due to the lack of observations. The benchmark model, Model C1, finds the variable A Ki67 visual statistically significant. The algorithm *glmulti* chooses the best fitting model with two predictors: the variable A Ki67 visual and the variable A ERG AI. This suggests same what we encountered with the dataset A exclusive models that artificial intelligence variables are not distinctly better. The absence of the dataset B variables in the algorithm chosen models implies that they are weaker in predicting the relapse than the dataset A variables.

To answer the question whether the relapse can be predicted using biomarkers, we can conclude that the biomarker Ki67 seems to be the best biomarker of the four. The variables of Ki67 are statistically significant in multiple models. Because high percent of positive biomarker Ki67 indicates high probability of the relapse, we need to remember that the values of the variables of Ki67 in the data are only between 0-20 percent.

All models in this analysis have high specificity meaning they predict non-relapsed better than relapsed. This can be due to a low amount of relapsed samples and the fact that relapses still happen at low percentages of positive variable Ki67, which is the only predictor that increases chances of relapse.

The variables of the biomarker AR are the worst predictors. Not a single model suggests them and odds ratios are almost equal to one in all benchmark models. The variables of the biomarker PTEN are only present in the Model A3 as PTEN AI. The variables of the biomarker ERG are a part of the Model A3 and also variable A ERG AI is in the Model C2. As a result, the biomarkers ERG and PTEN have some added value to the analysis unlike the biomarker AR.

6 Conclusion

The motive of this analysis was to find out if biomarkers have predictive abilities on prostate cancer relapse. Additionally we expanded the interest to comparing predicting power of visually obtained values to values which were acquired from artificial intelligence. At first we were not certain how well the artificial intelligence had performed in acquiring the values from the sample cores.

The data contained cancers with two different classifications: MRI visible and MRI invisible. We decided to split the data to three datasets where the classification was the decisive factor. Third dataset was the two datasets combined. Predictive abilities of the biomarkers could then be further investigated if the classification of cancer had impact. Each 4 biomarkers had an artificial intelligence variable and a visual variable, total of 8 explanatory variables for the first two datasets and 16 explanatory variables for the combined dataset.

Logistic regression model was the chosen method to examine how well the biomarkers could predict the dichotomized outcome of relapse. The parameters for explanatory variables were estimated using the maximum likelihood method. Significance of the explanatory variables were determined using the Wald test.

The selection of explanatory variables was done using the algorithm *glmulti* (Calcagno & de Mazancourt [11]) for models other than benchmark models with all variables. The algorithm was either operating with brute force or with the genetic setting. In all cases the explanatory variables were decided by the lowest value of the Bayesian information criterion of the model. Low values of the Bayesian information criterion indicate good fit.

After models were built we assured that there were no significant collinearity using the variance inflation factor. Predictive power of models were measured with receiver operating characteristic curve and the area under it. Goodness of fit was tested for each model with the Hosmer-Lemeshow goodness of fit test. We checked if there were any extremely influential observations with the standardized residuals and the Cook's distance. If there were any significantly influential observations, we removed them and refitted the model and made comparisons to learn their effects on the results.

The logistic regression models on MRI visible data, dataset A, showed that artificial intelligence variable for biomarker Ki67 has explanatory power for relapse. Further, investigating the modelling problem with interacting covariates, it turned out that artificial intelligence variables of biomarkers ERG and PTEN are significant components in explaining the relapse. The visual variable of biomarker Ki67 also appeared significant among the interacting covariates.

Analysis of the MRI invisible data, dataset B, showed no significant predictors in the logistic regression model. The same is observed in the logistic regression model of the combined data, dataset C: none of the variables from dataset B were significant. All in all no variables from dataset B were significant or chosen by the algorithm. That implies that MRI invisible cancer relapse is harder to predict or that MRI invisible variables are bad predictors. In the model of dataset C the only two significant explanatory variables were from dataset A; the visual variable of biomarker Ki67 and the artificial intelligence variable of biomarker ERG.

All the fitted logistic regression models had values of area under the receiver operating characteristic curve from 0.64 to 0.81. The value of area under the curve can roughly be translated to correct classification probability. The models classified non-relapsed observations better. These values indicate that the biomarkers do have some predictive power regarding the relapse.

As already mentioned, some of the biomarkers were found significant in some of the logistic regression models. However, all variables of biomarker AR had no relation to the relapse in any of the models. The biomarker Ki67 is the best predictor of the 4 biomarkers, because it is significant in greater number of models than the other biomarkers. The biomarkers ERG and PTEN are significant only in the interacting covariates with Ki67. Thus, biomarkers ERG and PTEN are not as useless as AR, but mostly seem to just support the biomarker Ki67.

When splitting dataset A to artificial intelligence and visual variables we noticed slight superiority of the artificial intelligence. The model that contained only the artificial intelligence variables had better fit and more accurate classification according to the receiver operating characteristic curve. However, we cannot conclude that artificial intelligence variables are purely better than visual variables, because the models contain significant visual variables. We could say that visual and artificial intelligence variables complement each other. To conclude, we observe that the artificial intelligence work and may give better accuracy in the relapse prediction.

These results apply only to the used data. To obtain more reliable results, we would need more observations in particular for the relapsed and MRI invisible cases. This could be solved by data imputation, but was decided not to be pursued in this analysis.

We particularly focused only to biomarkers applied on MRI visible and MRI invisible cancers. We did not include the variables of biomarkers applied to healthy area. For further studies we could add clinical variables such as Pre-operative prostate-specific antigen or Gleason score. Including variables of biomarkers applied to healthy area could deepen the analysis. Statistical analysis was solely done with logistic regression. There exist alternative methods to logistic regression for relapse prediction such as a machine learning method called random forest classifier.

References

[1] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[2] Tavel JA Strimbu K. What are biomarkers? *Curr. Opin. Hiv Aids*, 5(6):463–466, 2010.

[3] Moldovan PC, Van den Broeck T, Sylvester R, et al. What is the negative predictive value of multiparametric magnetic resonance imaging in excluding prostate cancer at biopsy? a systematic review and meta-analysis from the european association of urology prostate cancer guidelines panel. *Urol Oncol.*, 72(2):250–266, 2017.

[4] Schoots IG, Roobol MJ, Nieboer D, Bangma CH, Steyerberg EW, Hunink MG. Magnetic resonance imaging-targeted biopsy may enhance the diagnostic accuracy of significant prostate cancer detection compared to standard transrectal ultrasound-guided biopsy: a systematic review and meta-analysis. *Urol Oncol.*, 68(3):438–450, 2015.

[5] Richeng Jiang, Douglas E. Linn, Hege Chen, Hegang Chen, Xiangtian Kong, Jonathan Melamed, Clifford G. Tepper, Hsing-Jien Kung, Angela M.H. Brodie, Joanne Edwards, Yun Qiu Zhiyong Guo, Xi Yang, Feng Su,. A novel androgen receptor splice variant is up-regulated during prostate cancer progression and promotes androgen depletion–resistant growth. *Cancer Res.*, 69(6):2305—-2313, 2009.

[6] Wang Z, Wang Y, Zhang J, et al. Significance of the tmprss2:erg gene fusion in prostate cancer. *Mol Med Rep.*, 16(4):5450–5458, 2017.

[7] Troyer, Dean A., Jamaspishvili, Tamara, Wei, Wei, Feng, Ziding, Good, Jennifer, Hawley, Sarah, Fazli, Ladan, McKenney, Jesse K., Simko, Jeff, Hurtado-Coll, Antonio, Carroll, Peter R., Gleave, Martin, Lance, Raymond, Lin, Daniel W., Nelson, Peter S., Thompson, Ian M., True, Lawrence D., Brooks, James D., and Squire, Jeremy A. A multicenter study shows pten deletion is strongly associated with seminal vesicle involvement and extracapsular extension in localized prostate cancer. *The Prostate*, 75(11):1206–1215.

[8] Berlin A, Castro-Mesta JF, et al. Prognostic role of ki-67 score in localized prostate cancer: A systematic review and meta-analysis. *Urol Oncol.*, 35(8):499–506, 2017.

[9] Fourcade A, Khonsari RH. Deep learning in medical image analysis: A third eye for doctors. *J Stomatol Oral Maxillofac Surg.*, 120(4):279–288, 2019.

[10] Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol.*, 20(5):e253–e261, 2019.

[11] Vincent Calcagno, Claire de Mazancourt. *glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models.* Journal of Statistical Software, Volume 34, Issue 12. American Statistical Association, 2010.

[12] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020.

[13] G Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652, 1912.

[14] David W. Hosmer, Stanley Lemeshow, Rodney X. Sturdivant. *Applied Logistic Regression.* Wiley. 2013.

[15] S.S. Wilks. *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses.* Institute of Mathematical Statistics, 1938.

[16] Abraham Wald. *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large.* American Mathematical Society, 1943.

[17] FE Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer-Verlag, 2001.

[18] H Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[19] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[20] Alan Agresti. *Foundations of Linear and Generalized Linear Models.* Wiley. 2015.

[21] WX Yang. An Improved Genetic Algorithm Adopting Immigration Operator. *Intelligent Data Analysis*, 8:385–401, 2004.

[22] Scott Menard. *Applied logistic regression analysis*, volume 106. Sage, 2002.

[23] David W Hosmer and Stanley Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10):1043–1069, 1980.

[24] Jana D Canary, Leigh Blizzard, Ronald P Barry, David W Hosmer, and Stephen J Quinn. A comparison of the hosmer–lemeshow, pigeon–heyse, and tsiatis goodness-of-fit tests for binary logistic regression under two grouping methods. *Communications in Statistics-Simulation and Computation*, 46(3):1871–1894, 2017.

[25] Daryl Pregibon. Logistic regression diagnostics. *The annals of statistics*, 9(4):705–724, 1981.

[26] Nirian Martin and Leandro Pardo. On the asymptotic distribution of cook's distance in logistic regression models. *Journal of Applied Statistics*, 36(10):1119–1146, 2009.

[27] Kristine M Wadosky and Shahriar Koochekpour. Androgen receptor splice variants and prostate cancer: From bench to bedside. *Oncotarget*, 8(11):18550–18576, 2017.

[28] Lahdensuo K, Erickson A, Saarinen I, et al. Loss of pten expression in erg-negative prostate cancer predicts secondary therapies and leads to shorter disease-specific survival time after radical prostatectomy. *Mod Pathol*, 29(12):1565–1574, 2016.

[29] David Robinson, Alex Hayes, and Simon Couch. *broom: Convert Statistical Objects into Tidy Tibbles*, 2020. R package version 0.7.2.

[30] Max Kuhn. *caret: Classification and Regression Training*, 2021. R package version 6.0-90.

[31] John Fox and Sanford Weisberg. *An R Companion to Applied Regression.* Sage, Thousand Oaks CA, third edition, 2019.

[32] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

[33] Md Riaz Ahmed Khan and Thomas Brandenburger. *ROCit: Performance Assessment of Binary Classifier with Visualization*, 2020. R package version 2.1.1.

A Model Summaries

Model A1, benchmark model of Dataset A

```
Call:
glm(formula = Relapsed ~ ., family = "binomial", data = Dataset A)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5243  -0.6214  -0.4990  -0.3437   2.4709

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -2.561178   1.238819  -2.067  0.03869 *
AR Visual                     0.002407   0.010527   0.229  0.81918
AR AI                         0.002930   0.014391   0.204  0.83868
ERG Visual                   -0.766382   0.745655  -1.028  0.30405
ERG AI                       -0.137765   0.253464  -0.544  0.58677
PTEN Visual                   0.313852   0.553377   0.567  0.57061
PTEN AI                      -0.301045   0.155089  -1.941  0.05225 .
Ki67 Visual                   0.163944   0.087854   1.866  0.06203 .
Ki67 AI                       0.143370   0.051199   2.800  0.00511 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 263.35  on 273  degrees of freedom
Residual deviance: 237.40  on 265  degrees of freedom
AIC: 255.4

Number of Fisher Scoring iterations: 5
```

Model A2, algorithm chosen model, Dataset A

```
Call:
glm(formula = Relapsed ~ 1 + Ki67 AI, family = "binomial",
    data = Dataset A)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7709  -0.6383  -0.5517  -0.4927   2.1069

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.1904     0.2670  -8.203 2.34e-16 ***
Ki67 AI             0.1702     0.0474   3.590  0.00033 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 263.35  on 273  degrees of freedom
Residual deviance: 249.99  on 272  degrees of freedom
AIC: 253.99

Number of Fisher Scoring iterations: 4
```

Model A3, algorithm chosen model with interacting variables, Dataset A

```
Call:
glm(formula = Relapsed ~ 1 + Ki67 AI + Ki67 VIsual : ERG AI +
    Ki67 AI : PTEN AI, family = "binomial", data = Dataset A)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7486  -0.5926  -0.4879  -0.4208   2.2110

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.50241    0.29635  -8.444  < 2e-16 ***
Ki67 AI                   0.20553    0.05841   3.519 0.000434 ***
Ki67 Visual : ERG AI     -0.10840    0.03196  -3.392 0.000694 ***
Ki67 AI : PTEN AI        -0.03380    0.01476  -2.290 0.022047 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 263.35  on 273  degrees of freedom
Residual deviance: 235.09  on 270  degrees of freedom
AIC: 243.09

Number of Fisher Scoring iterations: 4
```

Model with only Visual variables, Dataset A

```
Call:
glm(formula = Relapsed ~ 1 + AR Visual + ERG Visual + PTEN Visual
 + Ki67 Visual, family = "binomial", data = Dataset A)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5557  -0.6574  -0.5625  -0.4429   2.2289

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.934040   0.425003  -4.551 5.35e-06 ***
AR Visual               0.009007   0.007521   1.198   0.2311
ERG Visual             -0.891016   0.419765  -2.123   0.0338 *
PTEN Visual            -0.380026   0.333683  -1.139   0.2548
Ki67 Visual             0.196616   0.079540   2.472   0.0134 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 277.13  on 285  degrees of freedom
Residual deviance: 264.96  on 281  degrees of freedom
AIC: 274.96

Number of Fisher Scoring iterations: 4
```

Model with only AI variables, Dataset A

```
Call:
glm(formula = Relapsed ~ 1 + AR AI + ERG AI + PTEN AI + Ki67 AI,
 family = "binomial", data = Dataset A)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7778  -0.6200  -0.5058  -0.3866   2.4956

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.262731   0.990155  -3.295 0.000984 ***
AR AI                        0.007725   0.010428   0.741 0.458825
ERG AI                      -0.349407   0.151443  -2.307 0.021045 *
PTEN AI                     -0.215956   0.098867  -2.184 0.028940 *
Ki67 AI                      0.186155   0.048280   3.856 0.000115 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 275.15  on 295  degrees of freedom
Residual deviance: 251.41  on 291  degrees of freedom
AIC: 261.41

Number of Fisher Scoring iterations: 4
```

Model B1, benchmark model of Dataset B

```
Call:
glm(formula = Relapsed ~ ., family = "binomial", data = Dataset B)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9581  -0.5799  -0.4744  -0.2854   2.3767

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -5.012202   2.462442  -2.035   0.0418 *
AR Visual                0.004886   0.012415   0.394   0.6939
AR AI                   -0.001240   0.025039  -0.050   0.9605
ERG Visual               1.239734   1.163491   1.066   0.2866
ERG AI                  -0.581499   0.451763  -1.287   0.1980
PTEN Visual              2.079708   1.256608   1.655   0.0979 .
PTEN AI                 -0.310432   0.298148  -1.041   0.2978
Ki67 Visual              0.125131   0.114466   1.093   0.2743
Ki67 AI                  0.065865   0.132393   0.497   0.6188
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 126.59  on 167  degrees of freedom
Residual deviance: 118.32  on 159  degrees of freedom
AIC: 136.32

Number of Fisher Scoring iterations: 6
```

Model B2, algorithm chosen model with and without interacting variables, Dataset B

```
Call:
glm(formula = Relapsed ~ 1, family = "binomial", data = Dataset B)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5168  -0.5168  -0.5168  -0.5168   2.0393

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9459     0.2333  -8.341   <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: 126.59  on 167  degrees of freedom
Residual deviance: 126.59  on 167  degrees of freedom
AIC: 128.59

Number of Fisher Scoring iterations: 4
```

Model C1, benchmark model of Dataset C

```
Call:
glm(formula = Relapsed ~ ., family = "binomial", data = Dataset C)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5783  -0.5232  -0.3170  -0.1408   2.6104

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -8.016777   3.972594  -2.018  0.04359 *
A AR Visual                  0.008453   0.020539   0.412  0.68066
B AR Visual                  0.008827   0.015237   0.579  0.56237
A AR AI                      0.030709   0.039775   0.772  0.44007
B AR AI                     -0.023631   0.029592  -0.799  0.42456
A ERG Visual                -0.041990   1.483874  -0.028  0.97742
B ERG Visual                 0.728099   1.778203   0.409  0.68220
A ERG AI                    -0.529924   0.533539  -0.993  0.32060
B ERG AI                    -0.646120   0.590285  -1.095  0.27370
A PTEN Visual               -1.285169   1.094102  -1.175  0.24014
B PTEN Visual                3.031611   1.556352   1.948  0.05143 .
A PTEN AI                    0.136085   0.306037   0.445  0.65656
B PTEN AI                   -0.711462   0.387943  -1.834  0.06666 .
A Ki67 Visual                0.532114   0.175661   3.029  0.00245 **
B Ki67 Visual                0.039431   0.160191   0.246  0.80557
A Ki67 AI                   -0.002063   0.111038  -0.019  0.98518
B Ki67 AI                    0.054183   0.169803   0.319  0.74966
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 104.562  on 129  degrees of freedom
Residual deviance:  79.965  on 113  degrees of freedom
AIC: 113.96

Number of Fisher Scoring iterations: 6
```

Model C2, algorithm chosen model without interacting variables, Dataset C

```
Call:
glm(formula = Relapsed ~ 1 + A ERG AI + A Ki67 Visual,
    family = "binomial", data = Dataset C)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4187  -0.5351  -0.3931  -0.3031   2.5184

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.9019     0.7217  -5.407 6.43e-08 ***
A ERG AI                    -0.4676     0.2348  -1.991 0.046446 *
A Ki67 Visual                0.4761     0.1348   3.532 0.000413 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 104.562  on 129  degrees of freedom
Residual deviance:  88.885  on 127  degrees of freedom
AIC: 94.885

Number of Fisher Scoring iterations: 5
```