UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Biomarkers for detection of relapses on patients with prostate tumour operated by radical prostatectomy

Afonso Pereira da Silva Burnay

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Professora Doutora Maria Helena Mouriño Silva Nunes
Doutora Adalgisa Catarina Reis Guerra

2023

Acknowledgments

I would like to take this opportunity to express my heartfelt gratitude to all those who have supported and encouraged me throughout the course of my research and thesis writing. Without their invaluable contributions, this work would not have been possible.

First and foremost, I would like to extend my thanks to my thesis advisors, Professor Maria Helena Nunes and Doctor Adalgisa Guerra, for their guidance during the course of this research. Although the journey was challenging, their input and feedback were valuable in shaping this thesis.

I extend my sincere appreciation to the faculty members of the Department of Statistics and Operational Investigation (DEIO) at Faculdade de Ciências Universidade de Lisboa (FCUL) for their dedication to teaching and providing an enriching academic environment that fostered my intellectual growth.

I am profoundly grateful to my family for their unwavering love, support, and understanding throughout my thesis journey. Their encouragement and belief in my abilities have been a constant source of strength, motivating me to pursue my academic goals.

To my mother Ana and dad Alberto, I struggle to find the right words to express the depth of my gratitude and pride in being your son. Your unwavering love, support, and sacrifices have made all the difference in my life, and I am forever grateful. I see the countless efforts you both made to ensure my happiness and success. Your sacrifices and support have not gone unnoticed, and I am profoundly grateful for everything you have done.

To my sister Joana and brother-in-law José, I want to express my heartfelt gratitude for being the best cheerleaders and support system I could ever ask for. Your presence in my life has been a constant source of joy, laughter, and encouragement. José, I'm grateful to have you as part of my family.

Thank you, Grandpa, for the love, support, and wisdom you bestowed upon me. Your presence may no longer be physical, but your spirit continues to shape my life in beautiful ways. I carry your love and guidance in my heart as I move forward in life.

I am deeply grateful to my friends who have been an unwavering source of support and encouragement throughout my thesis journey. Your constant belief in me, willingness to listen, and words of encouragement have kept me motivated during the challenging times.

Leonor, Gabriel, Mariana and Pedro, I cherish the lively discussions, brainstorming sessions, and constructive feedback you provided, which greatly enriched the quality of this work. Your willingness to listen to my rants, doubts, and struggles has been a tremendous source of strength. Your friendship and camaraderie have made this academic endeavor enjoyable and memorable.

André, Diogo, Catarina, Rodrigo Carvalho, Rodrigo Jesus and Duarte, You provided me with a space to relax, let loose, and appreciate the lighter side of life. Your camaraderie and understanding turned the occasional study breaks into cherished memories that I'll treasure forever. I am grateful for the countless nights we spent together, laughing, talking, and making unforgettable memories.

And finally to my beloved Girlfriend, Cláudia Monteiro, my best friend, I want you to know that this thesis is not just my accomplishment, it's a testament to the love and support you have given me. Your presence in my life has brought joy, happiness, and a sense of calm amidst the chaos of academic pursuits. The times we spent together, whether studying side by side or taking breaks to enjoy each other's company, have been precious moments that I'll forever cherish. I am grateful for the gift of your love and the prospect of building a life together. The journey ahead is filled with endless possibilities, and I can't wait to walk hand in hand with you every step of the way.

To all of the above, friends and family, thank you for being the support system that helped me recharge and refocus. Your presence in my life has been a crucial aspect of maintaining my sanity and motivation during this thesis journey.

Abstract

This study aimed to investigate the influence of clinical, histologic, and MRI biomarkers on the evaluation of biochemical relapse in patients with prostate tumors who underwent radical prostatectomy (RP). The study involved a statistical analysis of MRI data acquired from patients with prostate cancer at the Radiology Department of the Hospital da Luz, Lisbon.

The logistic regression analysis identified three significant predictors of biochemical recurrence: prostate volume, index lesion size, and smooth capsular bulging. These variables demonstrated a strong association with the likelihood of relapse. Prostate volume serves as an indicator of tumor size, while index lesion size has consistently been linked to prostate cancer prognosis and relapse prediction. The extension of the tumor beyond the prostate capsule, known as smooth capsular bulging, was also associated with an increased risk of recurrence.

Survival curve analysis showed a gradual decline in the survival probability over time, emphasizing an increased risk of relapse with extended post-surgery periods. This highlights the importance of continued surveillance and appropriate interventions to mitigate the risk of relapse in patients with prostate tumors.

The random forest model provided moderate prediction accuracy, suggesting its potential as a predictive tool for relapse outcomes. However, the overall performance of the models in predicting relapse fell short of expectations, indicating the complexity and heterogeneity of prostate cancer.

Future research should consider alternative models and analytical approaches, such as feature selection techniques, neural networks, and decision tree analysis, to further explore the intricacies of biochemical recurrence and the time until relapse.

In summary, this study contributes valuable insights into the identification of biomarkers for relapse detection in patients with prostate tumors who underwent RP. However, further research is necessary to fully unravel the complexities of biochemical recurrence and develop more accurate predictive models.

Keywords:Biochemical Recurrence, Biomarkers, Survival Analysis,Logistic Regression, Random Forest

Resumo

Este estudo teve como objetivo investigar a influência dos biomarcadores clínicos, histológicos e de ressonância magnética na avaliação da recidiva bioquímica em pacientes com cancro da próstata submetidos a prostatectomia radical (PR). A pesquisa envolveu uma análise estatística abrangente dos dados de ressonância magnética (RM) adquiridos de pacientes com cancro de próstata no Departamento de Radiologia do Hospital da Luz, em Lisboa.

A prostatectomia radical é considerada a principal intervenção cirúrgica para o tratamento do cancro de próstata localizado, envolvendo a remoção completa da próstata e tecidos circundantes. A (RM) é uma técnica de imagem não invasiva que fornece informações detalhadas, tanto anatómicas como funcionais, sobre a próstata, permitindo a avaliação das características do tumor.

Através da análise estatística abrangente dos dados de RM e da avaliação dos biomarcadores clínicos, histológicos e de RM, este estudo teve como objetivo obter uma compreensão mais aprofundada dos fatores que contribuem para a recidiva bioquímica nos pacientes após prostatectomia radical. Os resultados desta investigação têm o potencial de melhorar a estratificação de risco, auxiliar na tomada de decisões de tratamento e, em última instância, contribuir para melhores resultados para os pacientes no tratamento do cancro da próstata.

O cancro da próstata é um problema de saúde global significativo, representando uma das principais causas de mortalidade em todo o mundo. Apesar de extensas pesquisas e ensaios clínicos conduzidos ao longo de várias décadas, as estratégias de tratamento eficazes para o cancro da próstata permanecem um desafio. A natureza complexa da doença, juntamente com a sua heterogeneidade e diferentes apresentações clínicas, dificulta o desenvolvimento de intervenções direcionadas e universalmente bem-sucedidas.

Fatores de risco reconhecidos desempenham um papel crucial no desenvolvimento e progressão do cancro da próstata. A história familiar da doença, especialmente em parentes de primeiro grau, tem sido identificada como um fator de risco significativo, sugerindo um potencial componente genético. A etnia também desempenha um papel, com certas populações, como homens afrodescendentes, apresentando taxas de incidência e mortalidade mais elevadas em comparação com outros grupos étnicos. A idade avançada é outro fator de risco proeminente, uma vez que o cancro da próstata afeta principalmente indivíduos mais velhos.

A identificação precoce e o diagnóstico do cancro da próstata são vitais para obter melhores resultados no tratamento. Atualmente, três métodos principais são utilizados para a deteção: teste do antígeno específico da próstata (PSA), exame retal digital (ERD) e técnicas de imagem. O teste de PSA mede os níveis de uma proteína específica, chamada antígeno específico da próstata, no sangue. A presença de níveis elevados de PSA pode indicar a existência de anormalidades na próstata, incluindo cancro. No entanto, embora o teste de PSA seja amplamente utilizado, tem limitações, como a falta de especificidade e potencial para resultados falsos-positivos, o que tem gerado discussões sobre a forma ideal de utilização e a necessidade de abordagens diagnósticas complementares.

O ERD é um exame físico realizado por um profissional de saúde para avaliar o tamanho, forma e textura da próstata. Embora menos sensível que o teste de PSA, o toque retal pode detetar anormalidades palpáveis, como nódulos ou irregularidades, que podem indicar a presença de cancro da próstata. A combinação do teste de PSA e ERD pode aumentar a taxa de deteção do cancro da próstata, pois cada método fornece informações

complementares.

As técnicas de imagem desempenham um papel crucial no diagnóstico e estadiamento do cancro da próstata. A ultrassonografia transretal (UST) permite a visualização da próstata e a identificação de áreas suspeitas que podem requerer uma investigação mais aprofundada. A RM ganhou uma atenção considerável nos últimos anos devido à sua superior resolução de tecidos moles e capacidade de detetar e caracterizar tumores da próstata. A RM é particularmente valiosa na avaliação do tamanho, localização e extensão extracapsular do tumor, fornecendo informações essenciais para o planeamento do tratamento e prognóstico. Estas ferramentas de diagnóstico desempenham um papel vital na deteção, avaliação e tratamento do cancro da próstata, permitindo abordagens de tratamento personalizadas.

Este estudo foca-se na avaliação do impacto de vários biomarcadores clínicos, histológicos e de ressonância magnética na previsão de recidiva bioquímica após prostatectomia radical. Para selecionar os preditores mais relevantes, foi utilizada uma abordagem de seleção stepwise. Por meio desse processo iterativo, a análise de regressão logística identificou três preditores significativos: volume da próstata, índice de tamanho da lesão e protuberância capsular lisa. O volume da próstata, como indicador do tamanho do tumor, teve um impacto direto na progressão da doença e na probabilidade de recidiva. O índice do tamanho da lesão, consistentemente associado ao prognóstico do cancro da próstata e à previsão de recidiva, demonstrou importância como preditor de risco de recidiva. A presença da protuberância capsular lisa, indicando extensão tumoral além da cápsula prostática, aumentou ainda mais o risco de recidiva. Utilizando o método de seleção de variáveis stepwise, o modelo de regressão logística foi capaz de selecionar os preditores mais relevantes para prever a recidiva bioquímica em pacientes após prostatectomia radical. Esse método permitiu uma exploração eficiente e sistemática dos potenciais preditores, permitindo que apenas as variáveis mais significativas fossem retidas no modelo final.

O volume da próstata, o tamanho da lesão índice e a protuberância capsular lisa surgem como preditores significativos, fornecendo informações valiosas sobre a progressão da doença e os riscos de recidiva. Ao incorporar esses biomarcadores na tomada de decisões clínicas, os profissionais de saúde podem melhorar a estratificação de risco e personalizar as estratégias de tratamento para pacientes com cancro da próstata.

A análise de sobrevivência, uma técnica estatística valiosa, foi aplicada neste estudo para complementar os resultados da regressão logística e explicar a natureza do tempo até o evento de recidiva. Ao contrário da regressão logística, que se concentra em resultados binários, a análise de sobrevivencia considera o tempo necessário para que um evento ocorra, como a recidiva bioquímica. Ao analisar o declínio gradual na probabilidade de sobrevivência ao longo do tempo, essa análise fornece informações sobre o aumento do risco de recidiva com períodos prolongados após a cirurgia.

O declínio observado na probabilidade de sobrevivência salienta a importância da vigilância contínua e intervenções oportunas para mitigar o risco de recidiva em pacientes com tumores de próstata. O cancro da próstata pode apresentar taxas variadas de progressão e recorrência, tornando o monitoramento regular crítico para detetar quaisquer sinais de recidiva. A vigilância rigorosa permite que os profissionais de saúde intervenham prontamente e implementem estratégias de tratamento adequadas, potencialmente melhorando os resultados dos pacientes e prolongando a sobrevida.

O modelo de Random Survival Forests, usado neste estudo, mostrou uma precisão moderada na previsão da recidiva. É uma técnica de machine learning que utiliza um conjunto de decision trees para fazer previsões. A sua capacidade de lidarem com interações complexas e capturar relações não lineares entre os preditores torna-as uma ferramenta promissora para prever recidivas em pacientes com cancro da próstata.

No entanto, embora o modelo de Survival Forest tenha demonstrado potencial, o desempenho geral dos modelos na previsão de recidivas ficou aquém das expectativas. Vários fatores podem ter contribuído para essas limitações. A heterogeneidade da população-alvo, composta por pacientes com diversas características clínicas e perfis tumorais, pode introduzir complexidade e dificultar a precisão da predição. Além disso, as restrições de tamanho da amostra podem limitar a generalização dos resultados. O cancro da próstata é uma doença multifacetada com vários fatores que influenciam a recidiva, tornando assim difícil capturar todos os preditores relevantes dentro do tamanho da amostra.

Para aumentar a precisão da previsão, pesquisas futuras devem explorar abordagens alternativas de modelagem. Técnicas de seleção de recursos, que identificam o subconjunto mais informativo de preditores, podem ajudar a identificar as principais variáveis que levam à recidiva. As neural networks, com a sua capacidade de capturar padrões complexos e relacionamentos não lineares, podem fornecer mais informações sobre a previsão de recidivas. A análise de árvores de decisão, uma abordagem de modelagem intuitiva e interpretável, pode oferecer informações valiosas sobre as relações hierárquicas entre os preditores.

Ao considerar essas abordagens de modelagem alternativas, os investigadores podem expandir a compreensão da previsão de recidivas no cancro da próstata. Essa exploração pode levar ao desenvolvimento de modelos de previsão mais precisos, permitindo uma melhor estratificação de risco e estratégias de tratamento personalizadas para os pacientes.

Em conclusão, este estudo contribui com informações valiosas para a identificação de biomarcadores para a deteção da recidiva bioquímica em pacientes com tumores de próstata após PR. Os achados destacam a importância de considerar as características do tumor, incluindo o volume da próstata, o índice do tamanho da lesão e a protuberância capsular lisa, na avaliação do risco de recidiva. Além disso, o estudo realça a necessidade da realização de futuros estudos para aprofundar a compreensão da recorrência bioquímica.

Palavras Chave: Recurrencia Bioquímica, Biomarcadores, Análise de sobrevivência, Regressão Logistica, Florestas Aleatórias

Index

List of Figures

List of Tables

List of abbreviations

AIC Akaike Information Criteria.

AUC Area Under Curve.

BIC Bayesian information criterion.

BR Biochemical recurrence.

CD Cook's Distance.

DRE Digital Rectal Exam.

ECE Extracapsular extension.

FDA Food and Drug Administration.

FN False Negative.

FP False Positive.

FPR False Positive Rate.

HR Hazard Ratio.

LRA Logistic Regression Analysis.

MLE Maximum likelihood estimator.

MLR Multiple Logistic Regression.

MRI Magnetic resonance imaging.

OR Odds Ratio.

PCa Prostate Cancer.

PEC Prediction Error Curve.

PH Proportional Hazard.

PI-RADS Prostate Imaging Reporting Data System.

PSA Prostate-specific-antigen.

ROC Receiver Operating Characteristic.

RP Radical prostatectomy.

RSF Random Survival Forest.

SE Standard Error.

TN True Negative.

TP True Positive.

TPR True Positive Rate.

TRUS Transrectal Ultrasound.

VIF variance inflation factor.

VIMP Variable Importance.

Chapter 1

Introduction

The prostate is a gland that is part of the male reproductive system. It is located just below the bladder and in front of the rectum. Its function is to produce a fluid that makes up a part of semen.

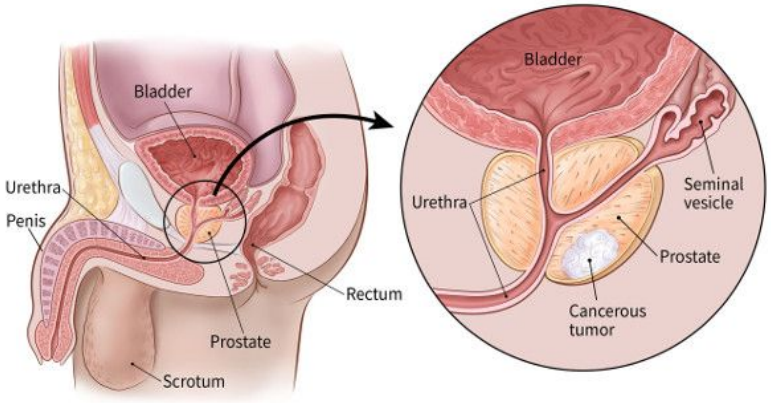


Figure 1.1: Illustrative figure of Prostate Cancer

Cancer is a disease in which cells in the body grow out of control and when it is started in the prostate it is called prostate cancer. For many years humankind has been fighting cancer. According to the International Agency for Research on Cancer, cancer was the leading cause of death worldwide in 2020, with almost 10 million deaths in that year[4].

Although there have been decades of research and clinical trials, cancer is still a very hard disease to cure because we still know very little about it [5], however there are a few well-known risk factors that lead to prostate cancer, such as a person's family history, ethnicity, old age, and genetic factors[6].

In 2018, prostate cancer was responsible for 358,989 deaths (3.8% of all deaths caused by cancer in men) with 1,276,106 new cases, making it the second most frequent malignant cancer (after lung cancer). The average age of diagnosis worldwide is 66 years, and increasing age is correlated with the incidence and mortality of

the disease. It is estimated that until 2040 there will be 2,293,818 new cases and that there will be an increase of 1.05% in mortality [6]. In Portugal 6759 new cases were diagnosed in 2020 with a total of 1917 deaths in the same year [7].

According to more recent research, when an inflammatory response occurs following an injury to epithelial cells, it can lead to prostate cancer. It is a complex process influenced by genetic changes and various processes, in which normal cells can evolve into possibly atrophic cells, these from prostatic intraepithelial neoplasia to local cancer, and the latter to metastatic cancer. The development of cancer can also be facilitated by growth factors and hormones (androgens). Thus, it is possible to say that prostate cancer is caused by complex interactions, which occur throughout life, of mutable factors inherent to the subject's life [8].

Identifying this disease early is very important, and currently there are three methods of diagnosis: Prostate-specific-antigen (PSA) testing, Digital Rectal Exam (DRE) and Imaging [9].

PSA is the most important factor for identifying men at increased risk of prostate cancer. In 1994, PSA testing gained FDA approval, recommending to perform biopsy on patients with PSA higher than 4 ng/mL, which was later found that a fifth of cancers are present with PSA values lower than 4 ng/mL [10].

A DRE is technique where a physician feels the prostate by inserting a finger into the rectum to identify if the prostate is enlarged, has lumps, areas of hardness or other types of abnormal texture.

North American and European urology guidelines recommend incorporation of a DRE with PSA screening, because when used in combination it improves cancer detection rate, as compared with either test used alone [11].

Conventional diagnostic imaging is obtained via Transrectal Ultrasound (TRUS), as it allows for structural assessment of the gland. More recently Magnetic resonance imaging (MRI) has been used for detection, characterization, and staging of prostate cancer [10] .

Figure 1.2: MRI of a man with prostate cancer

MRI is the most accurate tool and it allows us to locate and understand at what stage the cancer is, making it easier to choose the most appropriate and specific treatment for each patient[10].

The enormous technological development in MRI with the arising of equipment and methods of image acquisition and processing more and more sophisticated has led to increasing use of this technique in areas such as diagnosis support, tumour staging and therapeutic decision. The excellent spatial resolution and the diversity

of contrasts used in MRI (for example, anatomical image and diffusion) give this technique a high sensitivity and specificity in the detection of prostate tumours, being fundamental for the planning of minimally invasive robotic surgical interventions.

The interest in mapping non-invasive histological/functional/anatomical features using MRI as a possible alternative to the pathological anatomy and predictive biomarker in prostate surgeries has recently increased [12], nonetheless such methodologies need clinical validation.

One of the ways to treat prostate cancer is to do a Radical prostatectomy (RP), which aims to remove the entirety of cancer and it is used when the cancer is thought to be confined only within the prostate. During the intervention, the prostate and some tissue around it are removed.

Biochemical recurrence (BR) happens if the PSA exceeds a defined level post RP and it marks the return of the disease.

The natural history of BR after RP can be long but variable, approximately 35% of patients will develop a PSA recurrence within 10 years after surgery [13].

The main goal of this study is to analyse the influence of clinical, histologic and MRI biomarkers on the evaluation of the BR of the patients with a prostate tumour operated by RP.

This work involves the statistical analysis of MRI data of patients with prostate cancer, acquired in the radiology department of the Hospital da Luz, Lisbon. Data collection was performed between April 2022 and June 2022 and statistical analyses were performed between July 2022 and March 2023.

The methods that will be used are: Exploratory analysis; Multiple Logistic Regression and evaluation of models' performance; Survival Analysis; Random Survival Forest.

In Chapter 2, the Methodology used will be described, including all the statistical methods and the statistical software used. In Chapter 3, the data that was used to the statistical analyses will be described . In Chapter 4, the Results from all the performed statistical analyses will be described and finally, in Chapter 5, the main conclusions and the Discussion will be presented.

Chapter 2

Methodology

The present study aims to investigate the impact of various clinical, histologic, and MRI biomarkers on the prediction of BR in patients who have undergone RP for prostate cancer. A retrospective analysis was conducted on a sample of 151 patients who have undergone RP between 2013 and 2021, using both logistic regression and survival analysis methods.

Logistic regression was used to model the association between various predictor variables and the binary outcome of BR status. Several candidate predictors were included in the model, including age, PSA levels, index lesion size, prostate volume, capsular contact length, and the presence of Black striation periprostatic fat on MRI.

Survival analysis was also employed to examine the time to BR event. Specifically, the Kaplan-Meier method was used to estimate the probability of remaining free from BR over time. The log-rank test was used to compare survival curves between different groups of patients, based on their values for the predictor variables of interest.

Overall, the use of both logistic regression and survival analysis has allowed to comprehensively evaluate the role of various biomarkers in predicting BR status in patients who have undergone RP. The following sections will provide a more comprehensive description of the methodology employed in this study.

2.1 Logistic Regression Analysis

Logistic Regression Analysis (LRA) [14] is a statistical analysis method used to predict a dichotomic outcome, modeling the chance of an outcome based on individual characteristics. This binomial outcome can take two values: 0- outcome absent or 1- outcome present. Some possible examples of this are the presence or absence of a certain disease (yes/no), a yes or no question (yes/no), among others.

LRA is based on probabilities associated with the values of the dichotomous random variable Y. More precisely, Y represents a categorical outcome that only takes the value 1 (success or positive outcome) or 0 (failure or negative outcome).

The relationship between Y and the covariates is based on the logistic function,

$$Y = \log_e \left[\frac{\pi}{1-\pi} \right] = \boldsymbol{\beta}_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j X_j, \tag{2.1}$$

where β_0 represents the intercept term, signifying the log-odds when all predictors are zero. β_j represents all the coefficients that correspond to the predictor variables X_j, with j ranging from $(1, 2, ..., j)$.

The logit transformation allow one to write:

$$P\left(Y = 1 | X_1, ..., X_p\right) = \frac{exp\left(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_1 + ... + \boldsymbol{\beta}_p X_p\right)}{1 + exp\left(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_1 + ... + \boldsymbol{\beta}_p X_p\right)}, \tag{2.2}$$

where π indicates the probability of the event given the variables, $P(Y = 1|\mathbf{X})$ is the probability of the outcome variable Y being equal to 1, given the values of the predictor variables $X_1, ..., X_p$, $\boldsymbol{\beta}_0$ is the intercept term, and $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p$ are the predictor variables' coefficients. These coefficients and the intercept are estimated using the maximum likelihood method [15].

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE)[16] is a method used to estimate the parameters of a statistical model by maximizing the likelihood function. In the context of logistic regression, MLE involves finding the values of the coefficients that maximize the probability of observing the given set of outcomes. This process is often carried out by taking the natural logarithm (log) of the likelihood function, leading to the log-likelihood function:

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} y_i \cdot \ln(p_i) + (1 - y_i) \cdot \ln(1 - p_i), \tag{2.3}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ represent the coefficients representing the effects of predictor variables, y_i the observed binary outcome for the i-th observation and p_i the predicted probability of the event based on the logistic

regression model.

Maximizing the log-likelihood is equivalent to maximizing the likelihood function itself. Optimization techniques, such as gradient descent or Newton-Raphson, are commonly employed to find the values of the coefficients that result in the highest likelihood of observing the given data.

In summary, the likelihood function quantifies the probability of observing the data given a particular set of model parameters, and Maximum Likelihood Estimation seeks the parameter values that make the observed data most probable under the assumed model. This method forms the basis for estimating coefficients in logistic regression and other statistical models.

Odds ratio

In logistic regression, Odds Ratio (OR) is a measure used to quantify the relationship between the response variable and the predictors. It is defined as the ratio of the odds of the event occurring in the presence of a particular predictors to the odds of the event occurring in the absence of that predictor, while holding all other predictors constant [15].

To compare what happens when there is an increase of one unit to one of the predictor variables, the ratio of the predictions is computed:

$$\frac{Odds_{x_j+1}}{Odds} = \frac{e^{\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 X_1+...+\boldsymbol{\beta}_j(X_j+1)+...+\boldsymbol{\beta}_p X_p}}{e^{\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 X_1+...+\boldsymbol{\beta}_j X_j+...+\boldsymbol{\beta}_p X_p}} = e^{\boldsymbol{\beta}_j}. \tag{2.4}$$

The result of equation 2.4 can be interpreted as follows: a one-unit change in X_j results results in an increase in the log odds by $e^{\boldsymbol{\beta}_j}$.

Table 2.1: Odds ratio interpretation

odds ratio of 1	No association between the response variable and the predictor variable
odds ratio greater than 1	Positive association, An increase in the value of the variable is associated with an increase in the odds of the event occurring.
odds ratio less than 1	Negative association, An increase in the value of the variable is associated with a decrease in the odds of the event occurring.

In addition to estimating the logistic regression model's coefficients, it is crucial to evaluate the predictor variables' significance. This can be accomplished by conducting hypothesis testing or inspecting the coefficients' confidence intervals. The logistic regression model's overall significance can be determined using the Wald test statistic[15].

2.1.1 Diagnostics of the logistic model

The fundamental prerequisites for performing logistic regression should be fulfilled at all times. There are 4 basic assumptions: independence of errors, linearity, lack of strongly influential outliers and absence of multicollinearity.

Independence of errors assumption

The **independence of errors assumption** assumes that the errors or residuals of the model are independent of each other. This means that the observations or outcomes within the dataset are assumed to be unrelated and not influenced by each other.

When the independence of errors assumption is violated, it means that the errors or residuals exhibit some form of correlation or dependence. This can occur when there are repeated measures within the same individual, clustered data, or other forms of correlated observations. Violation of this assumption can lead to biased parameter estimates, inflated or underestimated standard errors, and incorrect statistical inferences.

Linearity assumption

The **linearity assumption** assumes that there is a linear relationship between the independent variables (predictors) and the log-odds of the outcome variable.

In logistic regression, the log-odds (logit) of the probability of the outcome variable being in a certain category are modeled as a linear combination of the predictor variables. The linearity assumption states that the effect of each predictor variable on the log-odds is constant and follows a straight-line relationship.

However, in many real-world scenarios, the relationship between the predictors and the log-odds may not be strictly linear. In such cases, the linearity assumption may be violated, leading to biased or unreliable model estimates.

To assess the linearity assumption, researchers often examine plots or evaluate statistical tests that assess the linearity between each predictor variable and the logit of the outcome. If the relationship appears to be nonlinear, transformations of the predictor variables, such as quadratic terms or splines, can be used to account for the nonlinear associations.

Lack of strongly influential outliers assumption - Cook's Distance

The **lack of strongly influential outliers assumption** refers to the expectation that extreme or atypical observations should not have a disproportionately large impact on the model's results or overall accuracy.

The assumption of the lack of strongly influential outliers suggests that the model's results should remain relatively stable and reliable even in the presence of a few extreme observations. If there are too many

outliers or if a single outlier has a strong influence on the model's outcome, the overall accuracy and validity of the model can be compromised.

Cook's Distance (CD) is a method used to evaluate the impact of individual observations on the estimated coefficients in regression analysis, identifying potential influential points (data points that have a significant impact on the results of statistical analysis). It measures the distance change in the predicted values of the response variable when a particular observation is omitted from the analysis, relative to the overall variation in the response variable.

In the follow equation, δ_i is defined as the sum of all the changes in the regression model when observation i, for $i = 1, .., n$, is removed from it:

$$\delta_i = \frac{\sum_{j=i}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{(p+1)\hat{\sigma}^2}, \tag{2.5}$$

where $\hat{\sigma}$ is the standard deviation, p represents the number of predictors in the regression model, $\hat{Y}_j$ is the jth fitted response value and $\hat{Y}_{j(i)}$ is the fitted response value obtained when excluding observation i.

The identification of influential observations is now widely recognized as a crucial aspect of statistical analysis, and Cook's distance is a common tool for detecting such observations in regression analysis. However, CD should not be applied as a sole criterion for accepting or rejecting individual cases as it is not a statistical test [17].

Absence of multicollinearity assumption - Variance inflation factor

The **absence of multicollinearity assumption** assumes that there is no high correlation or redundancy among the independent variables included in the model [18].

When multicollinearity is present, it becomes difficult to distinguish the separate effects of the correlated variables on the outcome variable. The estimated coefficients may become unstable, and the standard errors of the coefficients may become inflated. As a result, it becomes challenging to identify the true contribution of each predictor variable to the model.

The **variance inflation factor (VIF)** is a measure that shows how much the variance of the estimated coefficient is being inflated by multicollinearity, and is defined as:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{2.6}$$

where R_i^2 is the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones [19]. Values of VIF higher than 10 indicate multicollinearity[20].

2.1.2 Model Selection Methods

The objective of a model is that it can predict an outcome with the use of variables, but it is possible to have two different models that predict the same outcome. What is wanted is that the model that is chosen is the most simple, yet the one that is more accurate. There are several methods to create this ”optimal model”.

Backward elimination (step-down)

This procedure allows us to know if a model is preferred to another. The following is a list of how this process works:

1. Build a regression model that contains all the possible variables;
2. Perform a Wald test to see what variables are significant ($H_0 : \beta_j = 0$ for j = 1, ..., p). If all the coefficients are significant then all the variables are important to explain the explanatory variable and none of them should be cut from the model.
3. If at least one of the above variables is not significant, the one with the biggest p-value is cut and a new model is adjusted considering the remaining variables.
4. Repeat the second and third steps until there are only significant variables in the model.

Its important to note that for a categorical variable, an F test should be used to test the null hypothesis that all the parameters of associated dummy variables are equal to zero [21].

Forward selection (step-up)

Another common procedure is designated as **forward selection (step-up)**:

1. Begin with a model that contains no predictor variables.
2. Fit simple models for each individual predictor variable and evaluate their performance using a chosen criterion.
3. Choose the predictor variable that, when added to the current model, improves the model fit the most according to the chosen criterion.
4. Include the selected predictor variable in the model.
5. Repeat the process by evaluating the remaining predictor variables, adding the one that provides the most significant improvement in model fit.
6. Continue this process until at least one of the variables is not significant [21].

Stepwise selection

The most used procedure for the selection of variables is the **Stepwise Selection**.

This procedure is a fusion of step-down and step-up model selection. It is possible to start with all variables or just one, and each time one variable is removed or added, all the variables that are present in the model are analyzed with the objective to check if they should be removed from the model at that step [21].

Akaike Information Criteria

Akaike Information Criteria (AIC) is a measure of quality of the adjusted model. It takes into account the precision and complexity of the model, and it is given by

$$AIC = -2\log(\hat{\ell}) + 2p, \tag{2.7}$$

where, $\hat{\ell}$ represents the maximized value of the likelihood function of the model and p the number of parameters in the model.

The larger the number of parameters in the model, the more significant the likelihood, so $log(L)$ grows with the complexity of the model.

The AIC is a tool for the selection of models. The lower the AIC, the better the model [21].

For the comparison of two AICs, the absolute value of each means nothing. What can give some information is their difference:

$$\Delta_l = AIC_l - AIC_{min}, \tag{2.8}$$

where the AIC of the l-th model is denoted as AIC_l, while AIC_{min} represents the lowest AIC obtained from the set of models under consideration [22].

To interpret Δ_l the criteria is displayed in Table 2.2:

Table 2.2: AIC difference interpretation

$\Delta_l < 2$	the l-th model has substantial support and is likely to be a good description of the data
$2 < \Delta_l < 4$	strong support for the i-th model
$4 < \Delta_l < 7$	very weak support for the i-th model.
$\Delta_l > 10$	model has essentially no support

Bayesian information criterion

Bayesian information criterion (BIC) is a statistical measure that balances the goodness of fit of a model with its complexity, by penalizing models with more parameters, in order to identify the most likely model among

a set of alternatives [23]. BIC is calculated based in the following equation:

$$BIC = p \log(n) - 2 \log(\hat{\ell}), \tag{2.9}$$

where $\hat{\ell}$ represents the maximized value of the likelihood function of the model, p the number of parameters in the model and n the sample size or number of observations in the dataset[24].

Just like AIC, the model with the lowest BIC value is preferred.

2.1.3 Goodness-of-fit diagnostics for the Logistic Regression Model

Goodness-of-fit diagnostics are used to assess how well a logistic regression model fits the observed data. These diagnostics evaluate the overall performance of the model in terms of its ability to predict the outcome variable based on the independent variables. There are several goodness-of-fit diagnostics methods, including binned residuals, precision, recall and receiver operating characteristic (ROC) curves, and area under curve (AUC) that will be further described.

Binned Residuals

Binned residual plots [25] are a useful tool for assessing the fit of regression models for binary outcomes, such as logistic or probit models. Unlike residual-versus-fitted plots used for linear regression, binned residual plots can accommodate the discrete nature of residuals from binary outcome models.

The construction of binned residual plots involves ordering predicted probabilities and calculating residuals, then dividing the data into equal-sized bins and plotting the average residual against the average predicted probability for each bin. If the model is correct, approximately 95% of the points are expected to lie within the confidence limits. Departures from random scatter can indicate that the fitted model does not accurately describe the data [25].

Positive predictive value, Sensitivity and ROC curves

Positive predictive value and Sensitivity can be used as goodness-of-fit measures for evaluating the performance of a logistic regression model. In the context of logistic regression, the model predicts the probability of a binary outcome, and a threshold is applied to classify instances as positive or negative based on their predicted probabilities [26].

Table 2.3: Confusion Matrix for positive predictive value and Sensitivity

	actual positive	actual negative
predicted positive	TP	FP
predicted negative	FN	TN

True Negative (TN) represents the instances that the model correctly identified as negative. On the other hand, False Positive (FP) corresponds to instances where the model incorrectly identified them as positive. False Negative (FN) signifies instances incorrectly labeled as negative by the model. In contrast, True Positive (TP) denotes instances that the model correctly identified as positive.

Additionally, the True Positive Rate (TPR), also known as Sensitivity, is the proportion of actual positive instances correctly identified by the model. Conversely, the False Positive Rate (FPR) quantifies the proportion of actual negative instances that the model incorrectly identified as positive.

Positive predictive value measures the proportion of correctly predicted positive instances out of all instances predicted as positive. In logistic regression, positive predictive value can be interpreted as the proportion of correctly predicted events,

$$Positive\ predictive\ value = \frac{TP}{TP + FP}. \tag{2.10}$$

Sensitivity, on the other hand, measures the proportion of correctly predicted positive instances out of all actual positive instances. In logistic regression, sensitivity can be seen as the ability of the model to capture the true positive events,

$$Sensitivity = \frac{TP}{TP + FN}. \tag{2.11}$$

Furthermore, positive predictive value and sensitivity are closely related to the Receiver Operating Characteristic (ROC) curve [27]. The ROC curve is a graphical representation of the model's performance at various classification thresholds, plotting the true positive rate (TPR) (also known as Sensitivity) against the false positive rate (FPR).

The TPR measures the fraction of positive examples that are correctly labeled,

$$TPR = \frac{TP}{FP + TN}. \tag{2.12}$$

The FPR measures the fraction of negative examples that are misclassified as positive,

$$FPR = \frac{FP}{FP + TN} = 1 - Specificity. \tag{2.13}$$

Figure 2.1 displays two ROC curves, one for Algorithm 1 and another for Algorithm 2. The performances of the algorithms appear to be comparable.

Ideally, a good classification model will have a ROC curve that hugs the top-left corner of the plot, indicating a high TPR and a low FPR across different thresholds.

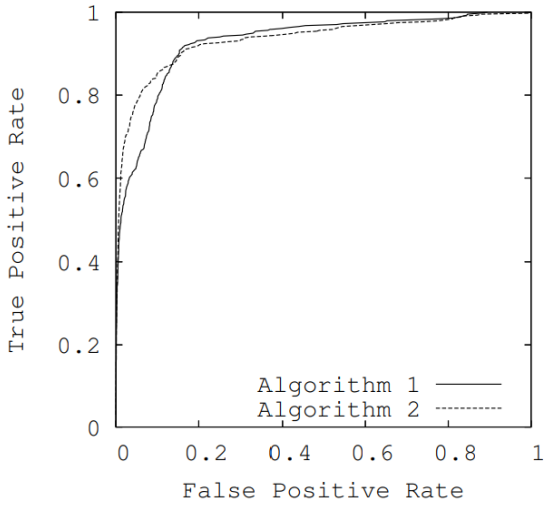


Figure 2.1: ROC Curve example for the comparison of two Algorithms

The area under the ROC curve (AUC-ROC) is commonly used as a summary measure of the model's discriminating ability. A higher AUC-ROC indicates better discrimination between positive and negative instances.

Area Under Curve

The AUC[28] refers to, as its name indicates, the area under the ROC curve.

It is a summary measure of the overall performance of a diagnostic test and represents the probability that a randomly chosen positive case will be ranked higher than a randomly chosen negative case by the test[27] and can be computed using the trapezoidal rule [29].

Table 2.4 displays the interpretation of the AUC values, ranging from 0.5 to 1 [29]:

Table 2.4: AUC values interpretation

AUC	test score
<0.7	non-useful
0.7-0.79	fair
0.8-0.89	good
0.9-0.99	excellent
1	perfect

To compare the AUC of two different ROC curves, the DeLong test is employed to check/verify if there is a significant difference between the two curves [30].

2.2 Survival Analysis

The first life table ever produced was back in 1662 when John Graunt invented it, beginning the development of Survival analysis.[31]

Survival analysis is an important methodology for the analysis of data on events observed over time, such as disease occurrence, disease recurrence, recovery, or other experiences of interest.

2.2.1 Survival Data

Two of the most important terms when referring to Survival analysis are Survival data and time. **Survival data** is data that represents the time of life or survival of individuals who belong to a specific population. **Time** represents the time taken from the initially defined instant until the occurrence of an event of interest.

Survival analysis is distinguished from other statistical areas because of the possible existence of censored data. **Censored data** happens when the event of interest does not occur to a patient or subject during the time the study is taking place [32].

There are several methods to censor data. The more common are: **right-censored**, **left-censored** or **Interval-censored**, which will be further described.

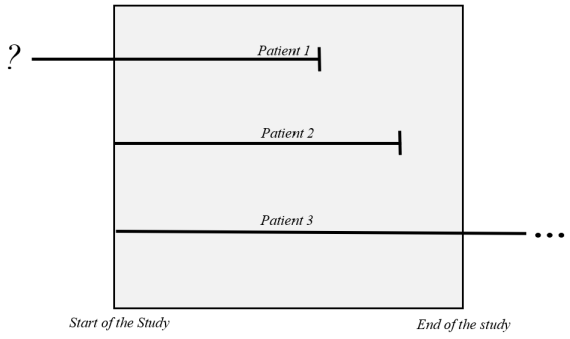


Figure 2.2: Example of Left-censored (Patient 1), Right-censored (Patient 3) and not censored (Patient 2) Data

Right-Censored Data exists when the study comes to a halt before an event occurs or when the subject simply leaves the study before it has ended. In Figure 2.2 it is possible to observe that the Patient 3 event did not happen before the end of the study, and because of that the data is right-censored.

Left-Censored Data exists when subjects enter the study when the event has already occurred and there is no information about at what time it happened. Figure 2.2 illustrates an instance of such an occurrence where Patient 1 was enrolled in the study post the occurrence of the event. Thus, the timing of the event is unknown, leading to left-censored data.

Interval-Censored Data also exists, and the only difference from Left-censoring is that an interval of when the event occurred is known[33].

There also exists **Truncation**. In survival analysis, truncation occurs when the observation period for individuals in a study is limited, and only those who survive past a certain point or experience an event within a specific time-frame are included in the analysis. Truncation is essential to consider when interpreting survival outcomes, as it influences the estimation of survival probabilities and can impact the generalizability of study findings.

It is likely that when performing a study using survival analysis there will be factors that influence the survival of the subject. Those factors are named **Risk factors**. Those factors can be inherent to the subject or external. A variable can be classified as constant if its values remain constant throughout the study, such as sex or variables that are only measured once. On the other hand, a variable is considered time-dependent if its value changes over the course of the study[32].

2.2.2 Survival and Hazard Function

Using T that represents the time of life of an individual that belongs to a homogeneous population as a non-negative, continuous variable [32], we can define the **Survival Function** on the instant t as being the probability of an individual surviving beyond the instant t:

$$S(t) = P(T > t) = 1 - F(t), \qquad t \geq 0, \tag{2.14}$$

where $F(.)$ is the distribution function of the random variable T.

Due to the characteristics of the distribution function, $S(.)$ has the following properties: $S(t)$ is a continuous, strictly decreasing function, and

$$S(0) = 1, \qquad \lim_{t \to \infty} S(t) = 0, \tag{2.15}$$

that is, the probability of survival in the instant 0 ($t = 0$) is 1, and the limit of the Survival Function when t approaches ∞ is always 0. i.e., the survival function will approach zero has the time t increases.

The Hazard function, equation 2.16, specifies the instantaneous rate of death in the instant t, that is, the probability of an individual experiencing an event in the infinitesimal interval $[t, t + \delta t[$, knowing that he/she has not experienced the event until time t. In other words, it corresponds to the instantaneous event rate for an individual who has survived until time t. The **Hazard Function** is defined as,

$$h(t) = \lim_{dt \to 0^+} \frac{P(t \leq T < t + dt | \, T \geq t)}{dt}. \tag{2.16}$$

A note about relevant properties of the hazard function:

$$h(t) \geq 0 \qquad \int_0^\infty h(t)dt = \infty. \tag{2.17}$$

The Hazard function, $h(t)$, does not take negative values, as we can see in equation 2.17, describing the evolution of the probability of instant death of an individual throughout time. Furthermore, the integral of $h(t)$ between 0 and ∞ is equal to ∞. Therefore, the hazard function is not a probability but rather a rate since its counter-domain is not restricted to the interval [0,1] but the set of non-negative real numbers. Note that the hazard function focuses on the event's occurrence, whereas the survival function relies on not experiencing the event.

2.2.3 Life tables and survival curves

Life tables (also called a mortality tables or actuarial tables) are the core of Survival analysis. They are a statistical tool used to analyze survival data and estimate mortality rates for a population. They provide information on the probability of surviving to a certain age or experiencing an event, such as death, based on the age-specific death rates observed in a population. Life tables can be used to estimate life expectancy, calculate cumulative survival rates, and compare mortality patterns across different populations. They are commonly used in demography, epidemiology, and public health research [32].

If we consider a cohort of n individuals that come from the same study population, the interval $[0, \infty[$ is divided in $k+1$ adjacent intervals and a fixed amplitude of

$$I_j = [a_{j-1}, a_j[, \qquad j = 1, ..., k+1 \tag{2.18}$$

with $a_0 = 0$, $a_k = L$ and $a_{k+1} = \infty$, where ℓ is the upper limit of observation.

Based on this results, life tables give us the following information:

- The number of people at risk at instant a_{j-1}: n_j
- The number of censured cases at I_j: m_j
- The number of deaths at I_j: d_j
- The Probability of survival beyond I_j: $P_j = S(a_j)$
- The conditional probability of death, in I_j knowing that survived beyond I_{j-1}: $\hat{q}_j$ [32].

These tables are often presented graphically as a **survival curve** that plots Time vs Cumulative Survival.

A survival curve, is a graphical representation of the probability of survival over time in a specific population or group. It is commonly used in medical and survival analysis to analyze the time it takes for an event of interest, such as death or failure, to occur.

Time can be presented in any measure unit which is pertinent to the study (seconds, years, millenniums...) and cumulative survival is a measure that ranges from 0 to 1, with 1 being the first instance (meaning that at time 0, survival is 100%) and 0 the last (meaning that at the last instance of time, survival is null).

The estimation of the survival curve requires the initial estimation of the survival function, which is accomplished using the non-parametric estimator of the survival function, a statistical method [34].

Kaplan-Meier Estimator is the generalization for censured data from the empirical survival function. Let:

- $t_{(1)}, ..., t_{(r)}$ distinct death instants in a sample with dimension n $(r \leq n)$
- d_i: number of deaths that have occurred in $t_{(i)}, i = 1, ..., r$
- n_i: number of individuals in the study immediately before $t_{(i)}, i = 1, ..., r$

then,

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{2.19}$$

where $\hat{S}(t)$ is equal to 1 for $0 \leq t \leq t_{(1)}$.

It is possible to compute $\hat{S}(t)$ recursively:

$$\hat{S}(t_{(1)}) = 1 - \frac{d_1}{n_1}, \tag{2.20}$$

$$\hat{S}(t_{(i)}) = \hat{S}(t_{(i-1)}) \left(1 - \frac{d_i}{n_i}\right). \tag{2.21}$$

Based on this results it is possible to compute the **Kaplan–Meier curve**.

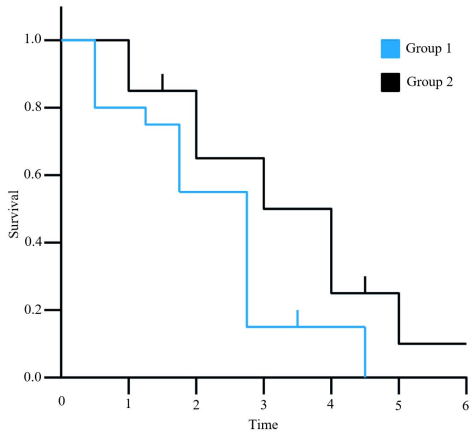


Figure 2.3: Illustrative Example of Two Kaplan-Meier curves, with n subjects for two groups

The x-axis represents Time, and the y-axis represents the cumulative probability of survival. Steps in the graph correspond to the times at which events were observed. As failures (events) begin to occur during the

follow-up period, the survival probability decreases. The survival probability decreases as failures (events) occur during the follow-up period. The study's accuracy is higher at the left of the curve (higher sample size) and reduces with time [34].

In figure 2.3, is represented a plot with two Kaplan-Meier Curves, Group 1 (Blue) and Group 2 (Black). By visually comparing the two curves it is possible to see the survival experience of two different groups of subjects and determine whether there is a significant difference in survival time between them. If the two curves are significantly different, this suggests that the factor being compared (e.g. treatment group, disease severity, etc.) impacts survival. If the two curves are similar, it suggests that the factor being compared does not have a significant effect on survival.

It is important to note that Kaplan-Meier curves can be affected by several factors, such as censoring, follow-up time, and sample size. Therefore, it is important to carefully consider these factors when interpreting the results of a Kaplan-Meier analysis.

In summary, the Kaplan-Meier estimator serves as an exploratory method in survival analysis by allowing researchers to visually explore and compare survival curves, identify time-dependent patterns, examine differences between groups, and assess the impact of censoring. It provides a foundation for further analysis and inference in survival studies and aids in generating hypotheses and insights into the underlying dynamics of the data.

2.3 Cox proportional hazards model

The most used method for establishing the relationship between the survival of a patient and several explanatory variables is the Cox proportional hazards model [32].

Using T as a continuous random variable that represents time of life, the **risk function** at the instant t, for an individual that is associated with a vector of variables $\mathbf{z} = (z_1, ..., z_p)'$ in the form:

$$h(t; \mathbf{z}) = h_0(t) \exp(\beta_1 z_1 + ... + \beta_p z_p) \tag{2.22}$$

where $\beta_1, ..., \beta_p$ are the unknown coefficients of the regression that represent the effect of the p covariates on survival and $h_0(t)$ is an arbitrary non negative function.

Considering two individuals that are associated to the vectors $\mathbf{z}_1$ and $\mathbf{z}_2$, that only differ in the values of the co-variable $\mathbf{z}_j$. Given a proportional risk function, there is

$$\frac{h(t; \mathbf{z}_1)}{h(t; \mathbf{z}_2)} = \frac{h_0(t) \exp(\beta_1 z_{11} + ... + \beta_j z_{1j} + ... + \beta_p z_{1p})}{h_0(t) \exp(\beta_1 z_{21} + ... + \beta_j z_{2j} + ... + \beta_p z_{2p})} = \exp(\beta_j (z_{1j} - z_{2j})), \tag{2.23}$$

where $\exp(\beta_j)$ represents the **relative risk** of the occurrence of the event for the two individuals that differ a unit in the values of the co-variable $\mathbf{z}_j$, with the remaining variables respective values being equal [32].

Maximum Likelihood Estimation

In a study with n individuals that were observed at k distinctive times of life $t_{(1)} < ... < t_{(k)}, k < n$. The risk at the instant $t_{(i)}$ is

$$R_i = R(t_{(i)}) = \{j : t_j \geq t_{(i)}\}, \tag{2.24}$$

with $\mathbf{z}_{(i)}$ as the vector of variables associated with the individual who experienced an event at time $t_{(i)}$. The inference in $\boldsymbol{\beta}$ is based on this function [32]:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{exp(\boldsymbol{\beta}' \mathbf{z}_{(i)})}{\sum_{l \in R_i} exp(\boldsymbol{\beta}' \mathbf{z}_l)}. \tag{2.25}$$

This likelihood function does not depend on $h_0(t)$ (the baseline hazard function), so it allows for inference on the vector of parameters $\boldsymbol{\beta}$. Under general conditions, the partial Maximum likelihood estimator of $\boldsymbol{\beta}$ is consistent, asymptotically normal with a mean value $\boldsymbol{\beta}$ and a covariance matrix $I(\boldsymbol{\beta})^{-1}$ [32], where

$$I_{jk}(\boldsymbol{\beta}) = -E\left(\frac{\partial^2 log(L)}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k}\right). \tag{2.26}$$

2.3.1 Model Selection Methods

One of the most important information in survival analysis is knowing which variables significantly influence survival time. Using the **Wald test**, given a coefficient $\boldsymbol{\beta_j}$ that represents the effect of a co-variable $\mathbf{z}_j$ in the survival of a given patient, it is possible to test:

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0, \tag{2.27}$$

in which, under H_0, the test statistic $\frac{\hat{\boldsymbol{\beta}}_j^2}{var(\hat{\beta}_j)}$ has asymptotic distribution χ_1^2.

If the results from the test are not statistically significant, there is no evidence that $\boldsymbol{\beta_j}$ is different than 0, so the variable does not influence survival time in the presence of the other variables [32].

Aiming to identify the most relevant variables to include in a model, **Stepwise selection** is one of the widely used techniques for automated variable selection. It iteratively adds or removes variables based on specific criteria, such as statistical significance or goodness-of-fit measures.

Stepwise is comprised of 3 main types of variable selection: forward selection, backward elimination, and bidirectional elimination. These stepwise methods provide a systematic approach to variable selection, but it's important to be mindful of their limitations, such as the potential for overfitting and sensitivity to initial conditions

2.3.2 Goodness-of-fit diagnostics for the Cox model - Proportional Hazard Assumption

The Cox model has one fundamental Assumption, the Hazards are proportional. Proportional Hazard (PH) assumption assumes that the hazard functions for different groups or levels of a predictor variable are proportional over time. It is important because violating this assumption can lead to biased estimates and incorrect inferences in survival analysis. When the assumption holds, the Cox PH model provides a flexible and powerful tool for analyzing survival data [35].

In the subsequent chapter, we will delve into various techniques for assessing the proportional hazards assumption. One commonly employed approach is the Schoenfeld residuals test, which offers a formal assessment of the proportional hazards assumption. Additionally, we will explore the evaluation of the Schoenfeld residuals plot, another method utilized to examine the assumption. By discussing these techniques in detail, we will gain a comprehensive understanding of how to effectively assess the proportional hazards assumption in survival analysis.

Residuals play a key role in the diagnostic methods for the Cox model. Various types of residuals are defined

for it and serve for different objectives[36].

Schoenfeld residuals

As referenced before, Schoenfeld Residuals are a fundamental part in testing PH assumption [2].

Considering p covariates and n independent observations consisting of time, covariates, and censoring information. These observations can be represented as (t_i, z_i, c_i), where i ranges from 1 to n, and c_i equals 1 for uncensored observations and 0 otherwise. It is possible to derive the Schoenfeld residuals as explained next, consider the kth covariate:

$$\frac{\partial L_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = \sum_{i=1}^{n} c_i \left\{ z_{ki} - \frac{\sum_{j \in R(t_i)} z_{kj} e^{\mathbf{z}_j' \boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\mathbf{z}_j' \boldsymbol{\beta}}} \right\} = \sum_{i=1}^{n} c_i \left\{ z_{ki} - \bar{z}_{kw_i} \right\}, \tag{2.28}$$

where

$$\bar{z}_{kw_i} = \frac{\sum_{j \in R(t_i)} z_{kj} e^{\mathbf{z}_j' \boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\mathbf{z}_j' \boldsymbol{\beta}}}. \tag{2.29}$$

Based on equation 2.28, the Schoenfeld residual estimator on the kth covariate for the ith individual is obtained by the substitution of the partial likelihood estimator of the coefficient, $\hat{\beta}$:

$$\hat{r}_{kS_i} = c_i \left(z_{ki} - \widehat{\bar{z}}_{kw_i} \right) \tag{2.30}$$

where

$$\hat{\bar{z}}_{kw_i} = \frac{\sum_{j \in R(t_i)} x_{kj} e^{\mathbf{z}_j' \boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\mathbf{z}_j' \boldsymbol{\beta}}}. \tag{2.31}$$

Based on this understanding, the discussion will now focus on the Schoenfeld residuals test and plot.

The Schoenfeld residuals test is a statistical procedure that evaluates the relationship between the Schoenfeld residuals and time. By assessing the correlation between these residuals and time, this test provides a formal assessment of the PH assumption. If the resulting p-value is significant, it indicates a violation of the PH assumption [37].

Schoenfeld residuals are also useful because plots of these residuals against time can reveal whether or not a covariate coefficient is time-dependent [2].

The plot of Schoenfeld residuals against time for each covariate should exhibit consistent residuals without any discernible pattern. Any noticeable pattern in the residuals suggests that the corresponding covariate is time-dependent. As a general guideline, a non-zero slope in the plot indicates a violation of the proportional hazard assumption.

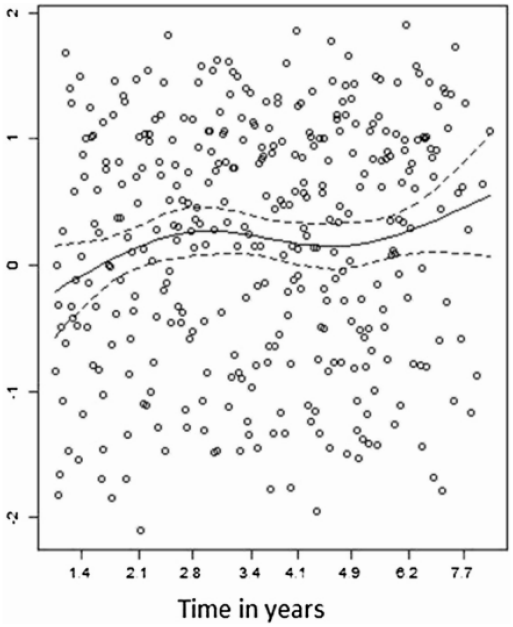


Figure 2.4: Schoenfeld residuals obtained for waist circumference in the multivariate Cox model for the example case-cohort study of colorectal cancer risk which utilized waist circumference levels as the primary exposure variable. [1].

The plot in Figure 2.4 does not show a pattern of changing residuals for waist circumference, indicating that the covariate is not time-dependent and suggesting that it does not follow the PH assumption. [1].

Cox-Snell residuals

The Cox-Snell residual for an individual i, where $i = 1, 2, ..., n$ is given by [2]:

$$r_{CS_i} = exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_i)\hat{H}_0(t_i), \tag{2.32}$$

where $\hat{H}_0$ is the estimated baseline Hazard and if the correct model has been fit to the data, then the r_{CS_i} will have a unit exponential distribution with a hazard ratio of 1 and $\hat{\beta}$ and $\hat{H}_0(t)$ are the partial maximum likelihood estimations. To test if the Cox-Snell residuals are approximately unit exponentially distributed, a residual plot can be constructed.

In this residual plot, if the estimate of the integrated hazard rate based on r_{CS_i} when plotted against r_{CS_i} yields a straight line with a slope equal to 1 then the Cox-Snell residuals are approximately unit exponentially distributed [2].

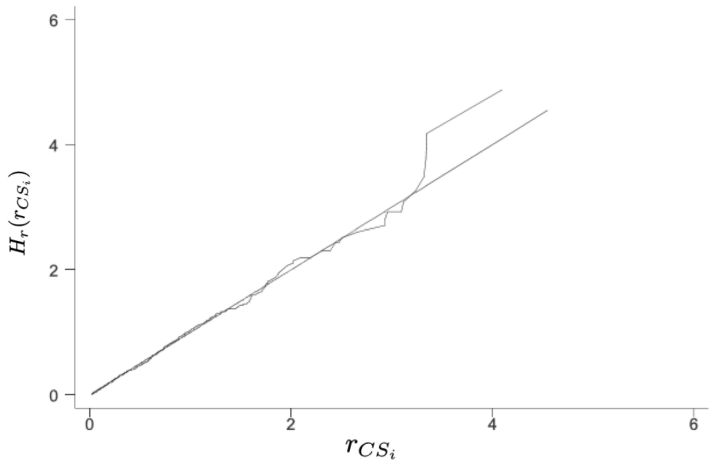


Figure 2.5: Cox-Snell residual plot using data on cabinet durations of 314 European cabinet governments.[2]

Cox-Snell residual plots are useful devices to quickly gauge the adequacy of the suggested model.

The 45-degree line through the origin in the figure 2.5 is a reference line. If Cox model holds, Estimated $Hr(r_{CS_i})$ versus r_{CS_i} should fall roughly on the line, meaning that the correct model has been fit to the data. In this case, it is possible to conclude that the correct model was fit to the data.

Martingale and Deviance Residuals

Martingale residuals are calculated as the disparity between the observed event indicator, represented by the censoring indicator $\delta_i(t)$, and the expected number of events, denoted by the integrated hazard $H_i(t)$:

$$\hat{M}_i(t) = \delta_i(t) - e^{\hat{\beta' z_i}} \hat{H}_0(t_i) = \delta_i(t) - R_i \tag{2.33}$$

The residual commonly used to evaluate outlying observations is the deviance residual. The deviance residual is simply a martingale residual re-scaled to have approximate symmetry around 0. Plots of deviance residuals against the duration time or against the observation number can be used to identify aberrant observations or clusters of observations.

Based on this information, the deviance residuals can be computed:

$$R_i = sign(\hat{M}_i(t))\{-2[M_i(t) + \delta_i log(\delta_i - \hat{M}_i(t))]\}^{1/2}, \tag{2.34}$$

where $M_i(t)$ is the martingale residual for the ith observation and δ_i the observed event for the ith observation [2].

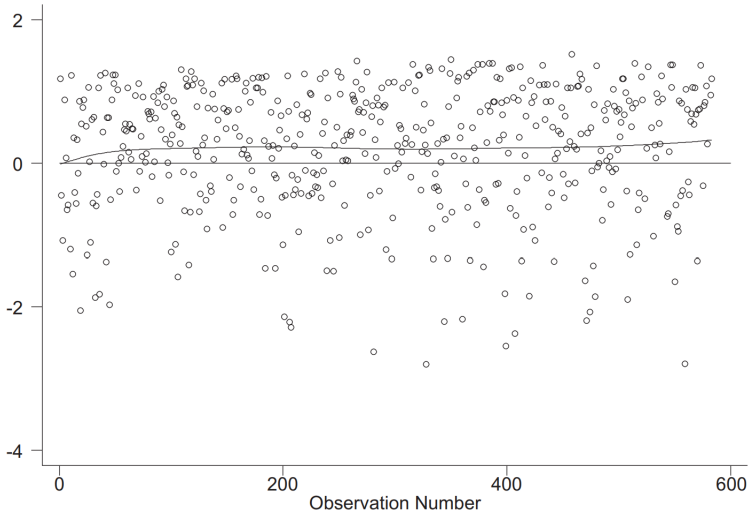


Figure 2.6: Plot of observation numbers against deviance residuals for results from a cox model adjusted to military intervention data[2]

In figure 2.6, by plotting the deviance residuals against the observation numbers of military intervention data it is possible to evaluate the presence or absence of outliers. In this plot, the residuals exhibit a fairly uniform distribution around zero. However, it is worth noting that a few observations have significantly large negative residuals, which may require additional investigation or scrutiny. This suggests that for these particular observations, the estimated hazard rate is likely overestimated [2].

There are several methods to assess the predictive performance of models, that will be further described.

Harrell's C-Index

The usual method to assess the predictive performance of the ensemble is to compute the C statistic for survival data, commonly referred to as Harrell's C-Index [38].

The appeal of the C-index lies in its interpretation as a probability of misclassification, making it a preferred choice for prediction error assessment. Moreover, it does not rely on a predetermined time point for evaluation, distinguishing it from other survival performance metrics that primarily account for censoring [39].

To calculate the C-index, the model's predicted survival times are compared with the actual survival times for each individual.

$$C = \frac{\sum_{i \neq j} I(\tilde{T}_i > \tilde{T}_j) \cdot I(\eta_i > \eta_j) \cdot \Delta_j}{\sum_{i \neq j} I(\tilde{T}_i > \tilde{T}_j) \cdot \Delta_j}, \tag{2.35}$$

where the indices i and j refer to pairs of observations in the sample, $(\tilde{T}_i, \tilde{T}_j)$ represent the "time-to-event" response of each individual, (η_j, η_i) the predicted risks, Δ_j the factor that is equal to zero if the pairs of observations are not comparable because the smaller survival time is censored, or 1 otherwise, and $I(\cdot)$ represents an indicator function that takes a value and returns 1 if the argument is true, and 0 otherwise [38].

Harrell's C-index is a useful metric for assessing the predictive power of an ensemble in survival analysis. A value of 0.5 indicates a non-informative model, while a value of 1 indicates perfect prediction. Harrell's C-index accounts for the full range of observed survival times, making it an easily interpretable coefficient [38].

Another way to access the performance of a model are Prediction Error Curve (PEC). The curve shows the relationship between time and the difference between the predicted and actual survival probabilities [40].

Prediction Error Curve

In the context of model selection, the Prediction Error Curve (PEC) can be a valuable tool to compare and assess the performance of different predictive models.

When evaluating multiple models, the PEC can be calculated for each model using the **Brier score** or any other appropriate performance metric. It allows for the visualization and comparison of the prediction accuracy and calibration of the models across different predicted probability ranges.

By examining the PECs of different models, patterns or trends can be identified in their performance. Models with lower Brier scores, indicating better prediction accuracy and calibration, are generally preferred. The shape of the PECs can also be compared to determine if certain models consistently outperform others across different predicted probability ranges [3].

The expected Brier score is defined as

$$BS(t) = E\left[(Y(t) - \pi(t|X))^2\right], \tag{2.36}$$

where X represents the covariate vector ,$Y(t)$ indicates the vital status of an individual at a given time t (taking values of 0 if the event as occurred or 1 if it has not), and $\pi(t|X)$ as the conditional event-free probabilities at time t, given the covariate combination $X = x$ [3].

The expected Brier score is a measure that represents the average Brier score one would expect to obtain if the predictive model's predicted probabilities were used to make predictions on new, unseen data. It provides an estimate of the model's overall prediction accuracy and calibration.

Furthermore, the Brier score can be considered as a function of time, which leads to prediction error curves in the whole observation period and allows transient assessment over time. The curves are primarily thought of as explanatory tools and are used to display the predictive ability of different prognostic classification schemes.

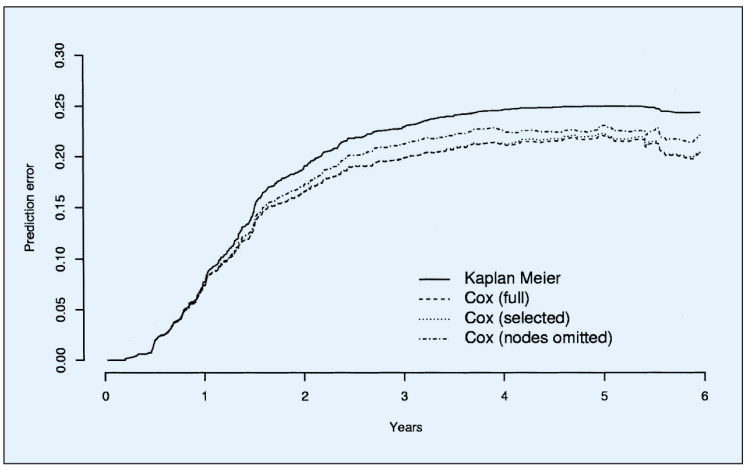


Figure 2.7: Example of Estimated Prediction Error Curves for Cox models with varying covariate settings where the Kaplan-Meier curve yields a benchmark value [3].

In figure 2.7, Prediction Error Curve (PEC)s obtained from 3 different Cox models are displayed. These methods are applied to data from node-positive breast cancer patients as collected by the German Breast Cancer Study Group [3].

The prediction error curves of the full Cox model, which includes all six predictors, and the selected Cox model, which includes tumor grade, number of lymph nodes, and progesterone receptor, show a striking similarity. This indicates that the three remaining covariates in the full model do not have substantial effects in the presence of the three selected covariates.

However, if the most important predictor, which is the number of lymph nodes, is excluded from the selected Cox model, the prediction error of this incorrectly specified model significantly increases. It becomes more similar to the prediction error obtained from the naive prediction using the pooled Kaplan-Meier estimate, which ignores all covariate information.

In summary, the Cox model with only three variables appears to be the most suitable model based on the analysis conducted.

The techniques demonstrated can assist clinical researchers in identifying the suitable statistical model when the primary objective is to predict events during the progression of a patient's disease [3].

2.4 Random Survival Forest model

Random Survival Forest (RSF) is a machine learning algorithm for time-to-event outcomes that can effectively capture intricate relationships between predictors and survival time without needing pre-specification, whilst showing superior predictive performance [41].

This technique is used to build a risk prediction model in survival analysis, where the predictor is an aggregate formed by consolidating the outcomes of numerous survival trees. This approach has 3 main Steps [40]:

1. Draw B bootstrap samples.
2. Grow a survival tree based on the data of each of the bootstrap samples b = 1, . . . , B:
   (a) At each tree node select a subset of the predictor variables.
   (b) Among all binary splits defined by the predictor variables selected in (a), find the best split into two subsets (the daughter nodes) according to a suitable criterion for right censored data, like the log-rank test.
   (c) Repeat (a)–(b) recursively on each daughter node until a stopping criterion is met.
3. Aggregate information from the terminal nodes (nodes with no further split) from the B survival trees to obtain a risk prediction ensemble.

The Nelson-Aalen estimator is used to estimate the conditional cumulative hazard function in each terminal node of a tree, adding to the ensemble survival function from random survival forest:

$$\hat{S}^{rsf}(t|x) = exp\left(-\frac{1}{B}\sum_{b=1}^{B}\hat{H}_b(t|x)\right), \tag{2.37}$$

where $\hat{H}_b(t|x)$ is the Nelson-Aalen estimator, as seen in [40].

2.4.1 Variable importance

Variable Importance (VIMP) in a RSF model refers to the measure of the relative importance of each predictor variable in predicting survival outcomes. This measure is obtained by calculating the decrease in the model's performance when a particular predictor variable is randomly permuted while holding all other variables constant. The greater the decrease in performance, the more influential the variable is. At a given time, a VIMP of zero or negative values shows that the predictor does not contribute to predicting the response variable [41].

The RSF model holds several advantages over traditional survival models like the Cox proportional hazards model. RSF excels in handling non-proportional hazards, capturing non-linear relationships and interactions naturally, managing high-dimensional data without requiring feature selection, exhibiting robustness

to outliers, automating variable selection, accommodating missing data without imputation, and demonstrating competitive predictive accuracy, especially in scenarios with complex or non-linear predictor-survival relationships. The choice between survival models depends on the dataset characteristics, assumptions, and analysis goals, with RSF proving valuable in situations where traditional model assumptions may be violated or when a flexible, data-driven approach is desired.

2.5 Statistical software

All statistical analyses were performed with R software (version 2023.03.0+386) with packages readxl, tidyverse, gridExtra, mlbench, MASS, pROC, arm, broom, car, BSDA, ggpubr, reshape2, survival, ranger, ggplot2, dplyr, ggfortify, survminer, and contsurvplot.

Chapter 3

Data

This study aims to analyze the influence of clinical, histologic and Magnetic Resonance Imaging (MRI) biomarkers on the evaluation of the biochemical relapse of patients with a prostate tumour operated by radical prostatectomy.

The radiology department of the *Hospital da Luz*, Lisbon, collected MRI data of patients with PCa operated by radical prostatectomy available for the development of this thesis.

3.1 Variables

The Response Variable for this study is BR which is a categorical variable that states whether BR occurred or not (Yes or No).

A total of **22** variables were provided for the study.

Age at MRI - Age of the Individual, in years, at the moment he underwent the MRI.

Prostate Volume (gr) - Volume of the Prostate, in grams.

PSA at MRI (ng/ml) - PSA amount, in ng/ml, at the moment of the MRI.

Index lesion size (mm) - The size of the index lesion, in mm, which refers to the size of the lesion in a patient that have the highest cancer suspicion score according to the initial multiparametric magnetic resonance imaging (MP-MRI) [42].

Capsular contact length TLC (mm) - The capsular contact length is defined as the amount of prostate tumour that is in contact with the prostatic capsule, in mm [43].

PI-RADS V2 - Prostate Imaging Reporting Data System (PI-RADS) is a structured reporting scheme for multiparametric prostate MRI in the evaluation of suspected prostate cancer. The second version of this reporting scheme was used in the present study[44]. A score from 1 to 5 is assigned to each lesion, indicating the likelihood of clinically significant cancer:

Table 3.1: PI-RADS scale

PI-RADS level	clinical significance
1	very low
2	low
3	intermediate
4	high
5	very high

According to table 3.1, PI-RADS first level represents cancer is highly unlikely to be present. On the contrary, PI-RADS fifth level, shows cancer is highly likely to be present.

Smooth capsular bulging - Binary variable that represents whether there is a protuberance of the prostate margin continuous with the tumour (yes), or not (no).

Capsular disruption - Binary variable that represents if the prostate capsule has been disrupt (yes), or not (no).

Unsharp margin - Binary variable that represents if the margin (the edge or border of the tissue removed in cancer surgery) was clean (no), or not (yes), which means that cancer cells were found at the edge of the tissue.

Irregular contour - Binary variable that indicates if the prostate was considered to be abnormal (yes), or not (no), during a digital rectal exam.

Black striation periprostatic fat - Binary variable that shows if a black adipose tissue with deformities surrounding the prostate exists (yes), or not (no).

Retoprostatic angle obliteration - Binary variable that indicates if a loss of fatty space between prostate and rectum occurred (yes), or not (no).

Measurable ECE - Binary variable that indicates if there is a presence of the tumor outside of the prostate capsule on the MRI (yes), or not (no).

ECE in prostatectomy specimen gold standard - Binary variable that shows if there is a presence of tumor cells in the periprostatic tissue, outside the prostate gland (yes), or not (no).

Gleason score - Grading system with two Gleason grades that were assign to the most and second most predominant pattern in the patient's biopsy (from 1 to 5, where 1 indicates normality in the prostate's tissue and 5 a cell so mutated that barely reassembles normal cells). These two grades are added together to create the Gleason score that ranges normally from 6 to 10, six meaning the lowest grade of cancer [45], as it is possible to see in table 3.2.

Table 3.2: Gleason score and Grade Grouping explanation

Risk Group	Grade Group	Gleason Score
low	1	< 7
intermediate	2	7 (3+4)
	3	7 (4+3)
high	4	8
	5	9-10

Gleason score (binary value) - Binary value that represents whether the patient has a Gleason score with a Grade Group of 2 or bellow (Gleason score (binary value) = 0), or a Grade Group scoring from 3 to 5 (Gleason score (binary value) = 1).

MRI Date - The date in which the patient underwent the MRI.

Date of Birth - The patient's date of Birth.

Surgery Date - The date where Radical prostatectomy was performed, in order to remove the cancer.

Recurrence date or last follow-up appointment - Date representing the last follow-up appointment (whether Biochemical occurrence did not occur), or Recurrence (if it did in fact occur).

PSA - Continuous variable that represents Prostate-specific-antigen (PSA) found in each patient's blood at the moment of the last appointment.

3.2 Exclusion Criteria

This study included, initially, 228 participants, all of whom were patients diagnosed with PCa between 2013 and the start of 2021.

Patients who did not display data about BR (57 patients in total) were removed from the study. Some of the reasons for the lack of data from the patients can be that a patient changed hospitals, passed away, or a simple input error by the doctor.

For this study, only patients who were cancer free for a year after RP were considered. That excluded 20 more patients from the study, narrowing the total to 151, as seen in figure 3.1.

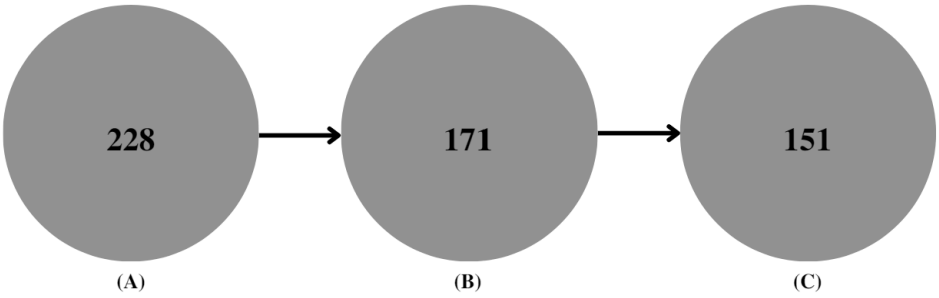


Figure 3.1: Number of patients in the study: (A) 228 patients without the application of the exclusion criteria. (B) 171 patients after the removal of 57 patients without data of BR. (C) 151 patients after the removal of 20 patients that were not cancer free for at least a year.

Chapter 4

Results

4.1 Exploratory Analysis

In this chapter, the results of the exploratory data analysis are presented. The distribution and relationships among the variables are explained and a summary of the patterns and trends observed in the data are also presented. These findings provide important insights that guide the development of the research hypotheses and inform subsequent analyses. Overall, the exploratory analysis serves as a foundation for the remainder of this thesis.

In order to answer the research question there is a need to understand the available data first, so table 4.1 was created.

Table 4.1: Characteristics of patients by Biochemical Recurrence in prostatectomy specimen (sample size = 151)

Variables		**BR** (nº of patients = 32)	**No BR** (nº of patients = 119)	**p-value**
Continuous Variables				
Age at MRI (years)		$61.29 \pm 5.72(55.58, 67.01)$	$61.25 \pm 6.63(54.62, 67.88)$	0.641
Prostate Volume (gr)		$35.69 \pm 11.97(23.72, 47.66)$	$44.67 \pm 20.45(24.22, 65.13)$	0.011
PSA at MRI (ng/ml)		$7.16 \pm 3.47(3.69, 10.63)$	$6.49 \pm 3.34(3.15, 9.84)$	0.189
Index lesion size (mm)		$17.16 \pm 6.57(10.59, 23.73)$	$13.18 \pm 5.06(8.12, 18.25)$	0.000
Capsular contact length TLC (mm)		$16.59 \pm 10.72(5.88, 27.31)$	$10.54 \pm 7.52(3.02, 18.06)$	0.001
Categorical Variables				
PI-RADS V2	3	1 (3.125%)	8 (6.72%)	0.000
	4	8 (25.00%)	76 (63.87%)	
	5	23 (71.875%)	35 (29.41%)	
Smooth capsular bulging	No	7 (21.88%)	65 (54.62%)	0.002
	Yes	25 (78.13%)	54 (45.38%)	
Capsular disruption	No	10 (31.25%)	73 (61.34%)	0.005
	Yes	22 (68.75%)	46 (38.66%)	
Unsharp margin	No	11 (34.38%)	69 (57.98%)	0.03
	Yes	21 (65.63%)	50 (42.02%)	
Irregular contour	No	10 (31.25%)	80 (67.23%)	0.001
	Yes	22 (68.75%)	39 (32.77%)	
Black striation peripro-static fat	No	21 (65.63%)	97 (81.51%)	0.091
	Yes	11 (34.38%)	22 (18.49%)	
Retoprostatic angle oblit-eration	No	27 (84.38%)	115 (96.64%)	0.021
	Yes	5 (15.63%)	4 (3.36%)	
Measurable ECE	No	24 (75.00%)	107 (89.92%)	0.055
	Yes	8 (25.00%)	12 (1.01%)	
ECE gold standard	No	18 (56.25%)	92 (77.31%)	0.031
	Yes	14 (43.75%)	27 (22.69%)	
Gleason score	No	22 (68.75%)	85 (71.43%)	0.939
	Yes	10 (31.25%)	34 (28.57%)	

Table 4.1 contains four columns (Variables, Biochemical Recurrence, No Biochemical Recurrence and p-value), and is divided by two sections (Continuous Variables and Categorical Variables).

The **Variables** column holds all the variables that were used. This study is comprised of 5 continuous variables, 10 categorical variables and 1 categorical response variable as previously mentioned.
The **Continuous variables** are represented by the average $\pm$ the standard deviation and the **Categorical variables** are represented by the frequency of each of their categories (in percentage).
The **BR** and the **No BR** columns separate this results in two groups: patients where biochemical recurrence occurred and the contrary, respectively, showing the mean ± the standard deviation for continuous variables and and the frequency for categorical variables.

The **p-value** column represents the p-value for the test of comparison of means between those with biochemical recurrence and those without biochemical recurrence.

For all the continuous variables the normality was rejected (via the Shapiro-Wilk normality test) so the Wilcoxon test was used (the non-parametric equivalent of the paired t-test).
In the case of the categorical variables, two different types of tests were used, the Chi-square test and Fisher's exact test. Fisher's exact test is used when at least one cell in the contingency table has an expected frequency less than 5. Otherwise, the chi-square test is used. For this reason, Fisher's exact test was only used for the PI-RADS V2 and Retoprostatic angle obliteration variables.

The test of Independence between the biochemical recurrence variable and some of the variables had a p-values greater than 0.05, meaning that the Independence hypothesis was rejected. There was insufficient evidence to suggest that they were not independent of the response variable, indicating a lack of relationship between those variables and the occurrence of BR.

The Dependent variables to BR are:

- Age at MRI
- PSA at MRI
- Black striation periprostatic fat
- Measurable ECE
- Gleason Score

4.1.1 Continuous Variables

The age at MRI did not exhibit a significant difference between patients with or without biochemical recurrence (61.29 years vs 61.25 years, respectively), nor did the PSA at MRI (7.16 ng/ml vs 6.49 ng/ml) as can be checked in the p-values recorded in table 4.1.

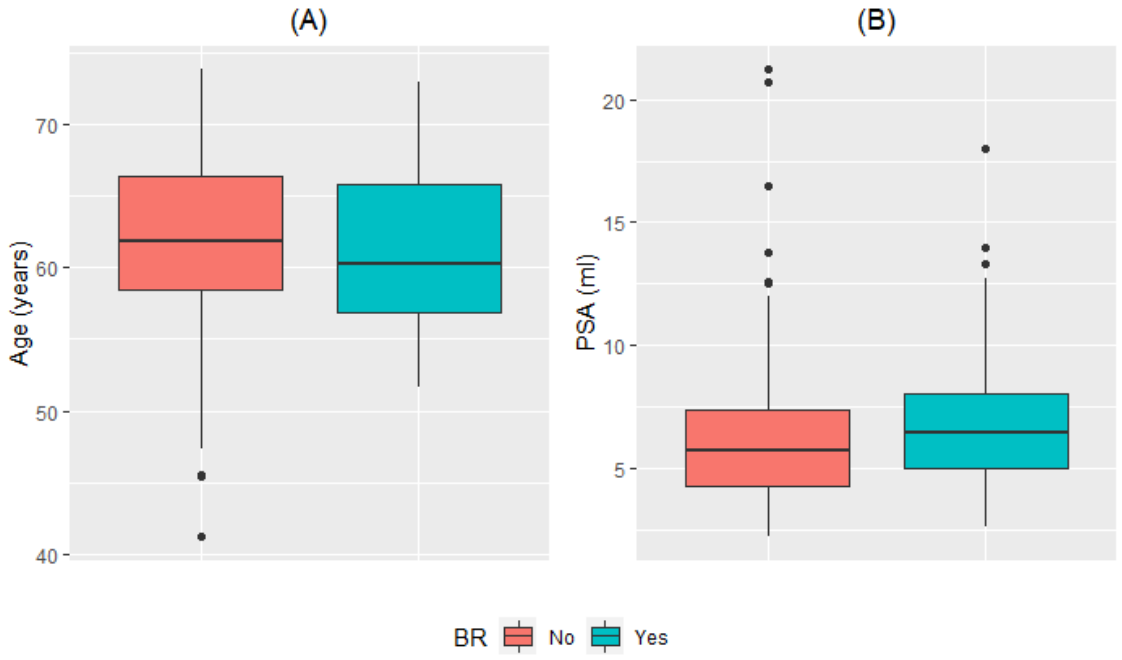


Figure 4.1: Box-plots of patients with or without biochemical recurrence by: (A) Age at MRI in years and (B) PSA at MRI in ng/ml

In Figure 4.1, the box plots in (A) depict the distribution of age at MRI for patients with and without biochemical recurrence. The median age for patients with Biochemical recurrence (BR) is 60.25 years, while for non-BR patients is 61.9 years. Three outliers are presented in the non-BR group (41.2, 45.4, and 45.5 years).
The box plots in (B) depict the distribution of PSA at MRI for patients with and without biochemical recurrence (BR). The non-BR group has a median PSA at MRI of 5.7 ng/ml, while the BR group has a median of 6.455 ng/ml. Both groups have outliers, with the non-BR group having 7 outliers ranging from 12.50 to 21.20, and the BR group having 3 outliers at 13.3, 14.0, and 18.0.

In contrast to the variables shown in Figure 4.1, the index lesion size, prostate volume, and capsular contact length display significant differences between patients who experienced biochemical recurrence (BR) and those who did not, as seen in table 4.1. The respective values for the mean of the variables are 17.16 mm index lesion size for BR patients compared to 13.18 mm for non-BR patients, 35.69 grams of prostate volume for BR patients compared to 44.67 grams for non-BR patients, and 16.59 mm of capsular contact length for BR patients compared to 10.54 mm for non-BR patients.

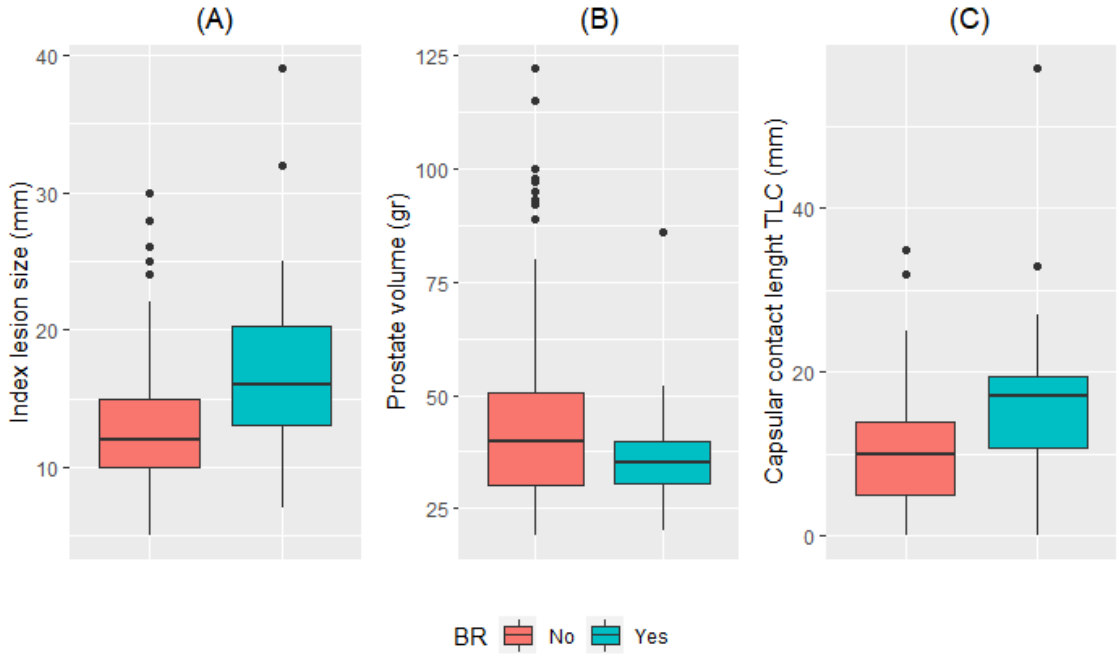


Figure 4.2: Box-plots of patients with or without biochemical recurrence by: (A) Index lesion size in millimeters, (B) Prostate Volume in grams and (C) Capsular contact length TLC in millimeters

In figure 4.2, box plots in (A) represent the index lesion size for patients who did and did not experience BR. The non-BR group has more outliers, ranging from 24 to 30 mm (a total of 7 outliers), while the BR group has 2 outliers (32 and 39 mm). The median, quartiles, upper and lower limits of the index lesion size are also different between the two groups.
Box plots in (B) belong to the Prostate volume, in grams, of patients where BR did or did not occur. The non-BR group has outliers ranging from 89 to 122 gr (nine in total), while the BR group has just one outlier (86 grams). The two box plots show a difference in median, upper limit and the third quartil, but have similar lower limits and first quartiles.
The Box plots in (C) belong to the Capsular contact length for patients that BR did occur or not. Both of the groups show outliers, four outliers for the non-BR group (32 mm three times and 35 mm) and two outliers for the BR group (33 mm and 57 mm). The median and quartiles appear to differ.

It is important to emphasize that this analysis is primarily visual in nature. It relies on visual observations and interpretations to examine the data and draw conclusions. Statistical analysis or formal tests are not employed in this visual analysis approach.

4.1.2 Categorical Variables

Black striation periprostatic fat, Measurable ECE and Gleason score did not show significant differences between patients where BR had occurred and the non-BR group, as seen in table 4.1.

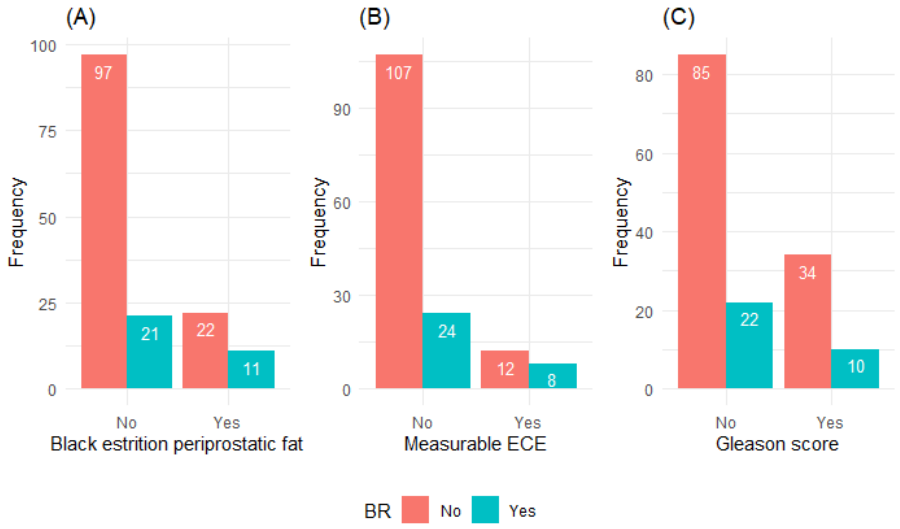


Figure 4.3: Bar plots grouped by BR occurrence of (A) Black striation periprostatic fat, (B) Measurable ECE and (C) Gleason score.

In figure 4.3 there are three different bar plots. In (A) it is possible to see the frequency of patients with and without Black striation periprostatic fat, grouped by the occurrence,or not, of BR. In the BR group 97 patients did not have this fat versus the 22 that did have. In the non-BR group 21 did not have it versus 11.
(B) corresponds to the frequency of patients with Measurable ECE, grouped by the occurrence, or not, of BR. In the BR group 107 patients did not have an ECE that was Measurable versus the 12 that had. In the non-BR group 24 did not have it versus 8 that did.
(C) corresponds to the frequency of patients with Gleason score higher then 2, grouped by the occurrence,or not, of BR. In the BR group 85 patients did not have a Gleason score with grade higher then 2 versus the 34 that had. In the non-BR group 22 did not have it versus 10 that did.

ECE gold standard, Smooth capsular bulging, Capsular disruption, Unsharp margin and Irregular contour are variables where there is a significant difference between patients where BR had occurred and patients where it did not, as seen on table 4.1.

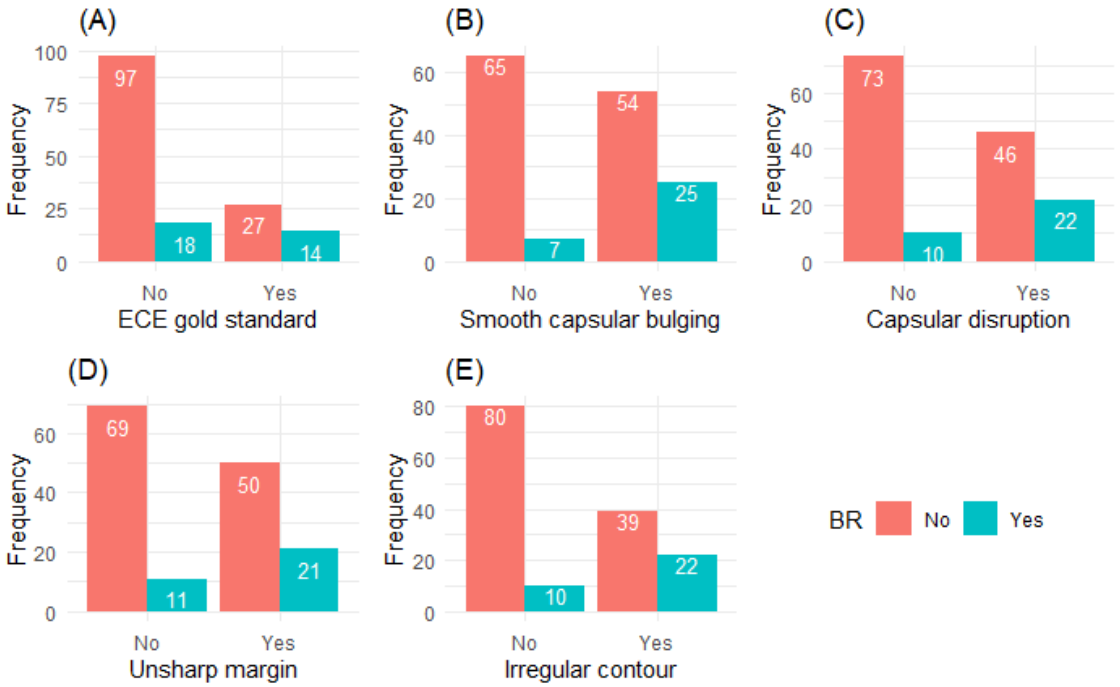


Figure 4.4: Bar plots grouped by BR occurrence of (A) ECE gold standard, (B) Smooth capsular bulging, (C) Capsular disruption, (D) Unsharp margin and (E) Irregular contour.

In Figure 4.4 there are 5 different bar-plots. In (A), the frequency of patients with a presence of tumor cells in the periprostatic tissue, outside the prostate gland, grouped by the occurrence,or not, of BR. In the BR group 18 patients did not display the presence of this tumor cells versus 14 that did. In the non-BR group, 97 patients did not display them versus the 27 that did.
(B) represents the frequency of patients with and without a smooth capsular bulging grouped by BR. The BR group displayed 7 patients without a smooth capsular bulging versus 25 that did. In the non-BR group, 65 patients did not have smooth capsular bulging and 54 did have it.
(C) corresponds to the frequency of patients with and without a disrupted capsule, grouped by BR. The BR group has 10 patients with no capsular disruption and 22 with it. The non-BR group as 73 patients with no capsular disruption versus 46 with it.
(D) expresses the frequency of patients with a unsharp margin, or not, of the tissue removed in the cancer surgery, grouped by BR. In the BR group 11 patients did not have the unsharp margin versus 21 who did. In the non-BR group 69 did not have a unsharp margin and 50 did.
(E) corresponds to the frequency of patients with and without a irregular contour of the prostate during a digital rectal exam, grouped by BR. The BR group has 10 patients without an irregular contour and 22 with. The non-BR group displays 80 patients without a irregular contour and 39 with.

PI-Rads V2 and Retoprostatic angle obliteration are also variables where there is a significant difference between patients where BR had occurred and patients where it did not, also seen in table 4.1.

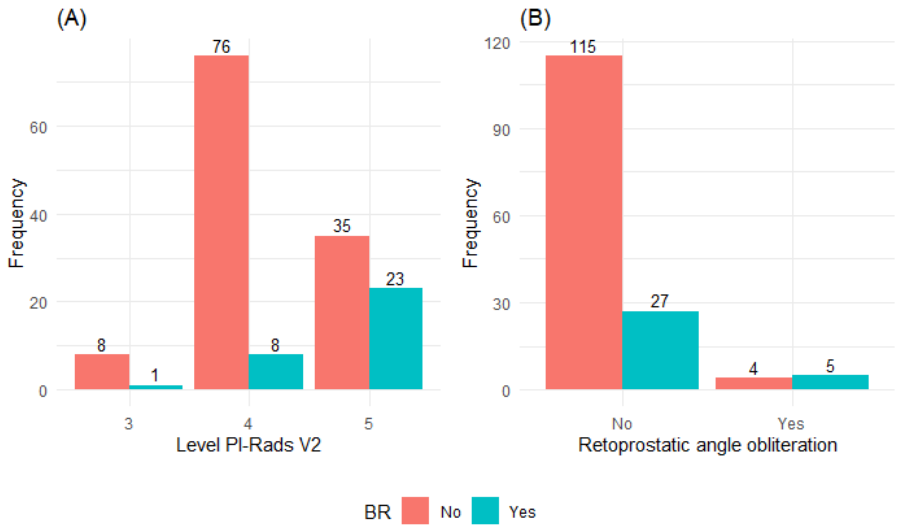


Figure 4.5: Bar plots grouped by BR occurrence of (A) PI-Rads V2 level and (B) Retoprostatic angle obliteration.

In Figure 4.5 2 bar-plots exist. Bar-plot (A) is the frequency of patients cataloged with 3 different levels of PI-rads, grouped by BR. In the BR group 1 patient is level 3, 8 are level 4 and 23 are level 5. For the non-Br group there are 8 patients level 3, 76 level 4 and 35 level 5.
(B) represents the frequency of patients with and without a loss of fatty space between prostate and rectum, grouped by BR. The BR group of patients were 27 for the non Retoprostatic angle obliteration and 5 for it. For the non-BR group there were 115 patients with a Retoprostatic angle obliteration, and 4 who did not have it.

4.2 Logistic Regression

In this sub-chapter, the results of the logistic regression analysis are presented, conducted to explore the relationship between each predictor variables and the binary outcome variable.

4.2.1 Univariate Logistic Regression

Table 4.2: Results from univariate Logistic regression

Variables		Coefficients	SE	OR (95% CI)	p-value
Continuous Variables					
Age at MRI (years)		0.001	0.031	1.001 (0.942, 1.066)	0.974
Prostate Volume (gr)		-0.036	0.016	0.964 (0.932, 0.992)	0.023
PSA at MRI (ng/ml)		0.055	0.055	1.056 (0.943 , 1.176)	0.323
Index lesion size (mm)		0.115	0.035	1.122 (1.05, 1.207)	0.001
Capsular contact length TLC (mm)		0.079	0.025	1.083 (1.034, 1.141)	0.002
Categorical Variables					
PI-RADS V2	3	Reference			
	4	-0.172	1.124	0.842 (0.128, 16.64)	0.879
	5	1.66	1.094	5.257 (0.879, 100.885)	0.129
Smooth capsular bulging	No	Reference			
	Yes	1.458	0.466	4.299 (1.806 , 11.474)	0.002
Capsular disruption	No	Reference			
	Yes	1.25	0.425	3.491 (1.552 , 8.336)	0.004
Unsharp margin	No	Reference			
	Yes	0.969	0.416	2.635 (1.186 , 6.133)	0.02
Irregular contour	No	Reference			
	Yes	1.507	0.428	4.513 (1.994 , 10.844)	0
Black striation peripro-static fat	No	Reference			
	Yes	0.837	0.441	2.31 (0.954 , 5.44)	0.058
Retoprostatic angle oblit-eration	No	Reference			
	Yes	1.672	0.704	5.324 (1.326 , 22.794)	0.018
Measurable ECE	No	Reference			
	Yes	1.089	0.509	2.972 (1.063 , 8.014)	0.032
ECE gold standard	No	Reference			
	Yes	0.975	0.418	2.65 (1.158 , 6.028)	0.02
Gleason score	No	Reference			
	Yes	0.128	0.432	1.136 (0.472 , 2.603)	0.767

Table 4.2 contains the results from univariate logistic regressions of of all variables under study. The table is comprised of 5 columns (Variables, Coefficients, Standard Error, Odds ratio(95% CI)and p-value).

The **Variables** column displays all the variables that were subjected to the uni-variate logistic regression.
The **Coefficients** column shows the estimated coefficients from the logistic regression model.
The **Standard Error (SE)** column gives a measure of uncertainty associated to the logistic regression coefficient.
The **Odds Ratio (OR)** column displays the OR (and its confidence interval of 95%) that associates each variable to the risk of BR.
The **p-value** column represents the p-value of the Wald test, for individual independence significance, of each variable to the response variable.

All variables are statistically significant in explaining the existence of BR, except for **Age at MRI**, **PSA at MRI**, **PI-RADS v2**, **Black striation periprostatic fat** and **Gleason score**.

4.2.2 Multiple Logistic Regression

To adjust the Multiple Logistic Regression (MLR) model various factors were considered.

The variables that were rejected in univariate logistic regression were not considered for this model as well as 2 other variables: **Capsular contact length TLC** and **Measurable ECE**.

Capsular contact length TLC has a high correlation with Index lesion size ($r = 0.72$), and between the 2 variables the former has a higher collinearity between all the other variables. Measurable ECE was independent of BR, as seen on table 4.1 ($p - value = 0.55$), so it was also removed.

Table 4.3: Results from multiple Logistic regression with all the variables

Variables		**Coefficients**	**SE**	**OR (95% CI)**	**p-value**	**VIF**
Intercept		-1.721	0.910	0.179 (0.030, 1.097)	0.058	
Prostate Volume (gr)		-0.037	0.018	0.964 (0.926, 0.995)	0.043	1.083
Index lesion size (mm)		0.073	0.045	1.076 (0.986, 1.178)	0.100	1.409
Smooth capsular bulging	No	Reference				
	Yes	0.944	0.573	2.570 (0.854, 8.284)	0.100	1.415
Capsular disruption	No	Reference				
	Yes	0.241	0.702	1.273 (0.316, 5.111)	0.731	2.415
Unsharp margin	No	Reference				
	Yes	-0.556	0.677	0.574 (0.144, 2.109)	0.412	2.279
Irregular contour	No	Reference				
	Yes	0.718	0.657	2.050 (0.579, 7.865)	0.275	2.152
Retoprostatic angle obliteration	No	Reference				
	Yes	0.578	0.859	1.783 (0.328, 10.095)	0.501	1.354
ECE gold standard	No	Reference				
	Yes	-0.218	0.565	0.804 (0.257, 2.388)	0.700	1.560

The following model was obtained:

$$\begin{aligned} logit(p) = -1.721 \ &- 0.037 \cdot Prostate\ Volume + 0.073 \cdot Index\ lesion\ size + \\ &0.944 \cdot Smooth\ capsular\ bulging + 0.241 \cdot Capsular\ disruption - \\ &0.556 \cdot Unsharp\ margin + 0.718 \cdot Irregular\ contour\ + 0.578 \cdot \\ &Retoprostatic\ angle\ obliteration - 0.218 \cdot ECE\ gold\ standard, \end{aligned} \tag{4.1}$$

where p is the probability of the patient having BR.

The provided logistic regression model represents a mathematical equation used for binary classification.

Certain assumptions that must be met in order to ensure accurate and reliable results.

Linearity

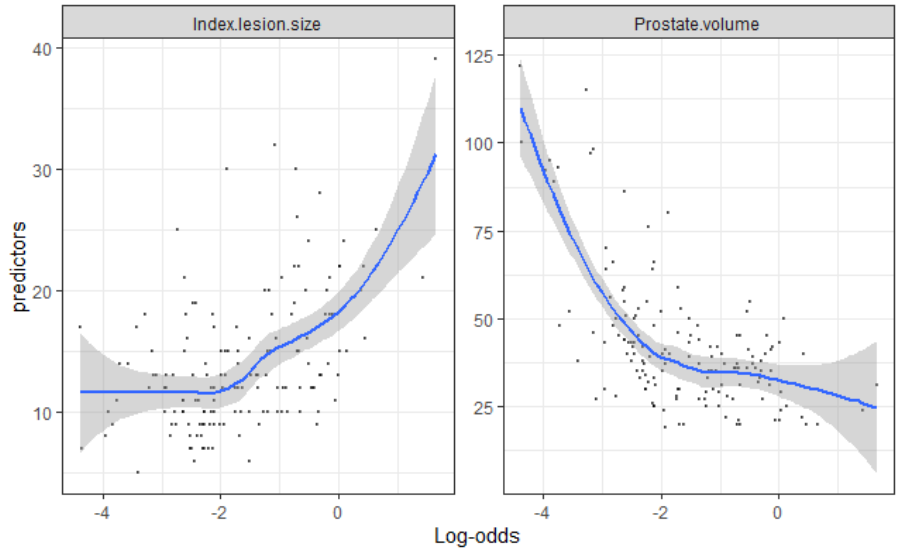


Figure 4.6: Scatter plot between Index lesion size (left), and Prostate Volume (right), and the logit values

In figure 4.6, the smoothed scatter plots indicate that index lesion size and prostate volume are roughly linearly associated with the time until BR in the logit scale. Improvements might be carried out in future research by including other type of functions to describe the relation between each of the above predictors with the outcome under consideration.

Influential values

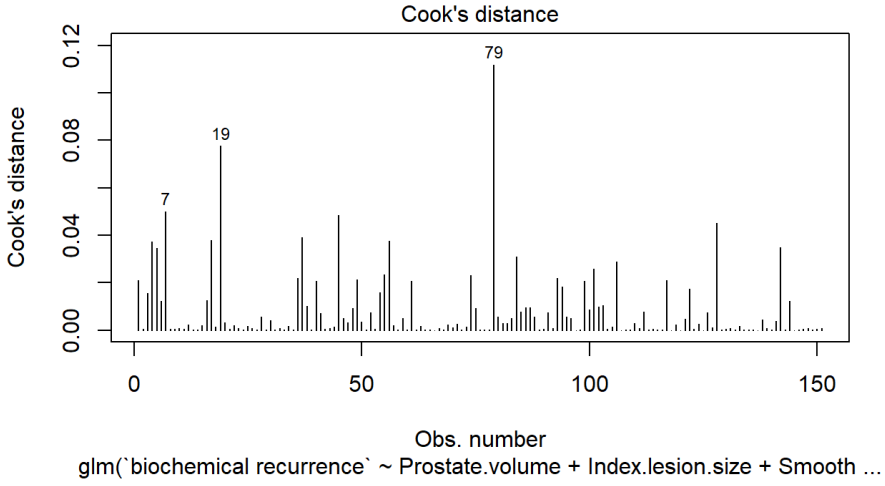


Figure 4.7: Cook's Distance of the adjusted MLR model

Potential influential points can be observed in figure 4.7, (with observation 7,19 and 79), but this alone is not

sufficient information to make conclusions so a standardized residual plot was adjusted to the model:

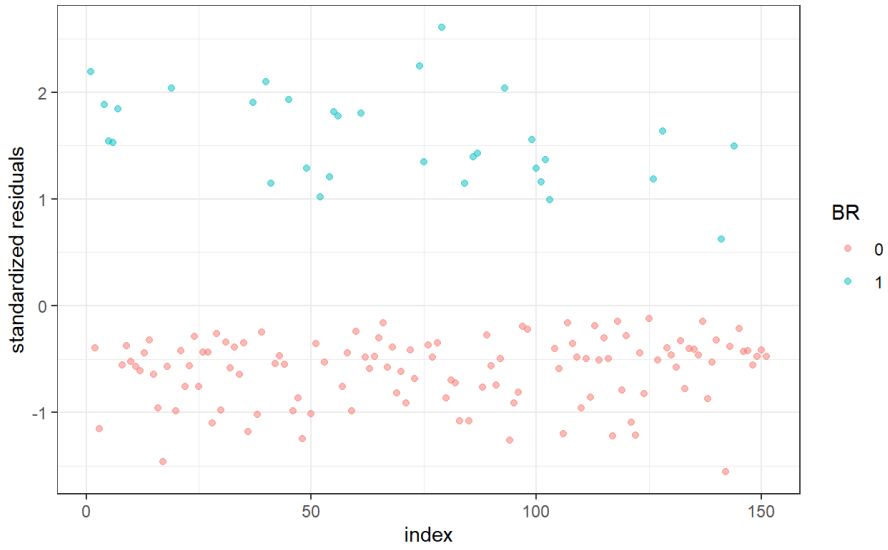


Figure 4.8: Standardized residuals plot of the adjusted MLR model

There are no influential observations in our data as indicated by the standard residuals being less than or equal to 3 [46].

Multicollinearity

To test for the absence of multicollinearity the calculation of the variance inflation factor (VIF) was implemented:

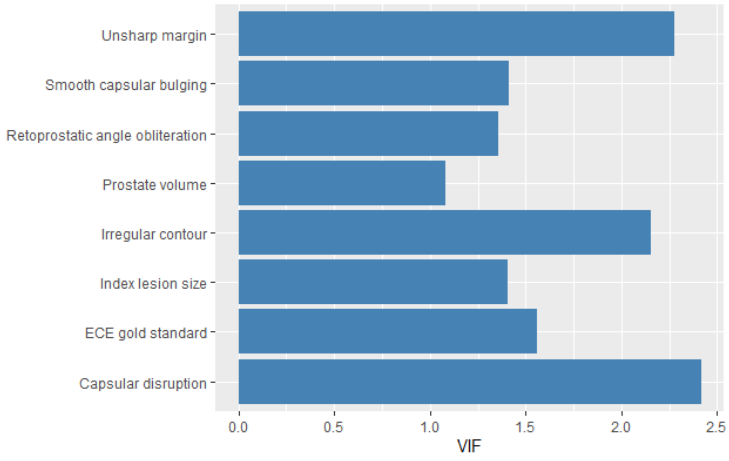


Figure 4.9: VIF for the predictor variables in adjusted MLR model

VIFs are ranging from 1 to 2.5, indicating that there is no significant multicollinearity among the predictor variables in the model. This is generally considered acceptable and suggests that the coefficients can be reliably interpreted.

All of the assumptions are confirmed, but it is also important to assess how good the fit of a model is. A well-fitted model is expected to yield precise results by accurately approximating the output when presented with new data.

To assess the overall fit of a logistic regression model, a binned residual plot was adjusted to the MLR model:

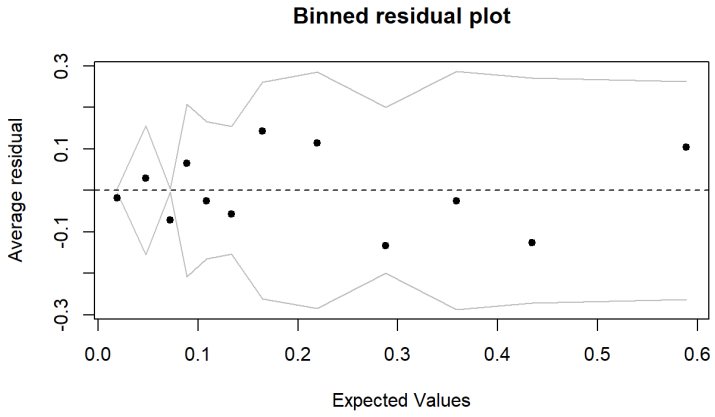


Figure 4.10: Binned residual plot for the adjusted MLR model

In figure 4.15, visually most values fall between the lines, so it is possible to say that the binned residual plot does not show any major problem with the fit of the model.

As for the performance of the model, it can be accessed via a ROC curve plot:

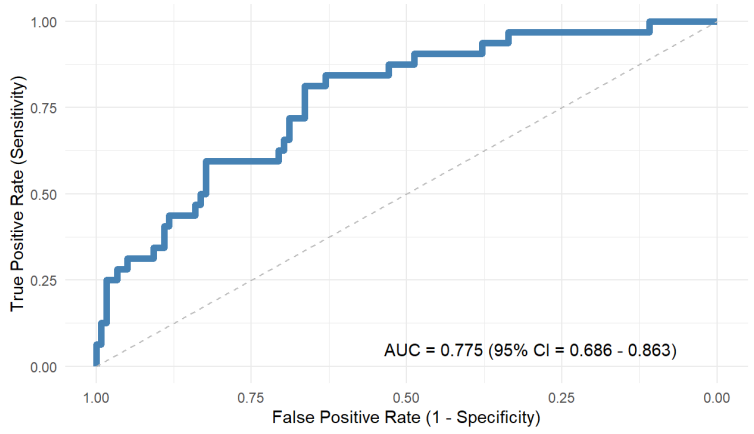


Figure 4.11: ROC curve for the adjusted MLR model

Figure 4.11 contains the ROC curve for the adjusted MLR model and from it an AUC of 0.775 was obtained, suggesting that the model has a moderate discriminatory power.

With all this, it is important to say that most variables were rejected and such a model would not be admissible. To try and create an admissible model, the Stepwise selection method was adjusted for the selection of variables, and the following model was the result of that:

Table 4.4: Results from multiple logistic regression after stepwise regression

Variables		coefficients	SE	OR(95% IC)	p-value	VIF
Intercept		-1.795	0.855	0.166 (0.031, 0.896)	0.0358	
Prostate volume		-0.039	0.017	0.961 (0.925, 0.991)	0.024	1.019
Index lesion size		0.089	0.040	1.093 (1.013, 1.187)	0.025	1.133
Smooth capsular bulging	Yes	1.098	0.506	2.997 (1.147, 8.555)	0.030	1.116

$$logit(p) = -1.795 - 0.039 \cdot Prostate\ Volume + 0.089 \cdot Index\ lesion\ size + 1.098 \cdot Smooth\ capsular\ bulging \tag{4.2}$$

The Stepwise model has only 3 variables, 2 continuous (Prostate volume and index lesion size) and 1 binary (Smooth capsular bulging).

Has it was done in the first MLR, the assumptions must be met:

The linearity was already proven in figure 4.6.

Influential values

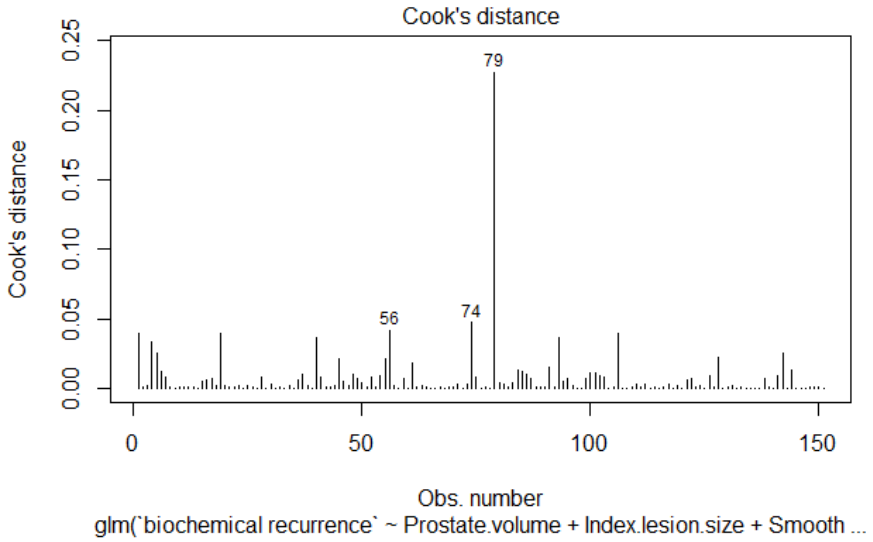


Figure 4.12: Cook's Distance of the adjusted MLR model after stepwise selection method

Potential influential points can also be observed in figure 4.12, (with observation 56, 74 and 79), but this alone is also not sufficient information to make conclusions so a standardized residual plot was adjusted to the stepwise model:

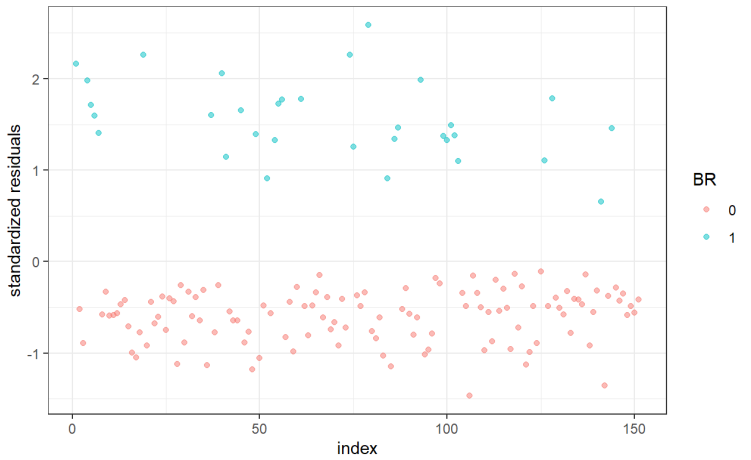


Figure 4.13: Standardized residuals plot of the adjusted MLR model after stepwise selection method

There are no influential observations in this model, as indicated by the standard residuals being less than or equal to 3 as expected[46].

VIF

The calculation of the variance inflation factor (VIF) was also implemented:

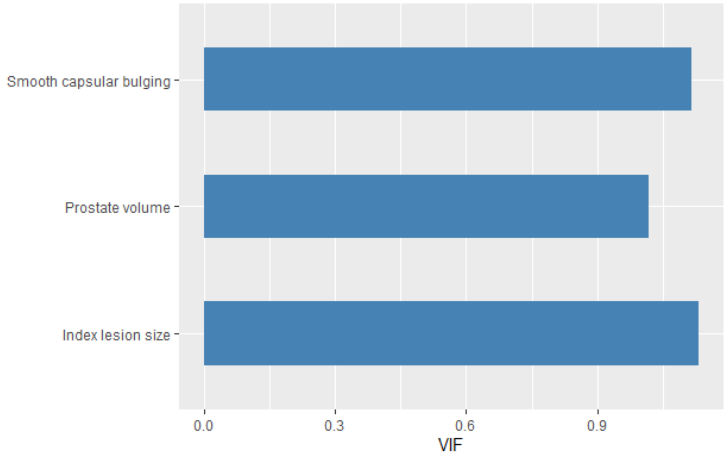


Figure 4.14: VIF for the predictor variables adjusted to the MLR model after stepwise selection method

In figure 4.14, theVIFs round the values of 1, indicating also that there is no multicollinearity among the predictor variables in the model.

As expected, all of the assumptions are also confirmed.

A binned residual plot was adjusted to the stepwise model:

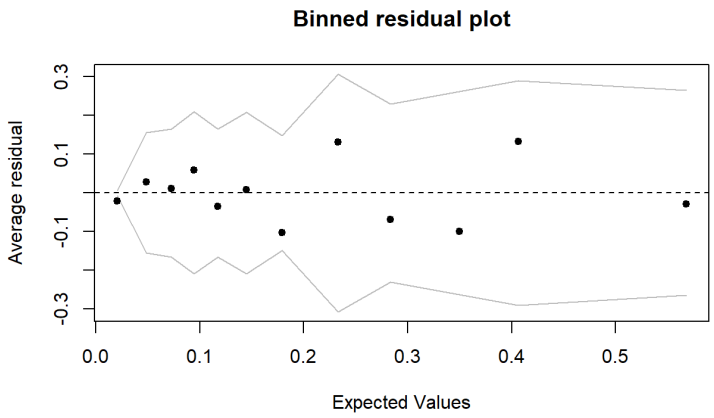


Figure 4.15: Binned residual plot for the adjusted MLR model after stepwise selection method

Visually most values also fall between the lines, so it is possible to say that the binned residual plot does not show any major problem with the fit of the model.

The ROC curve to access the performance of the model:

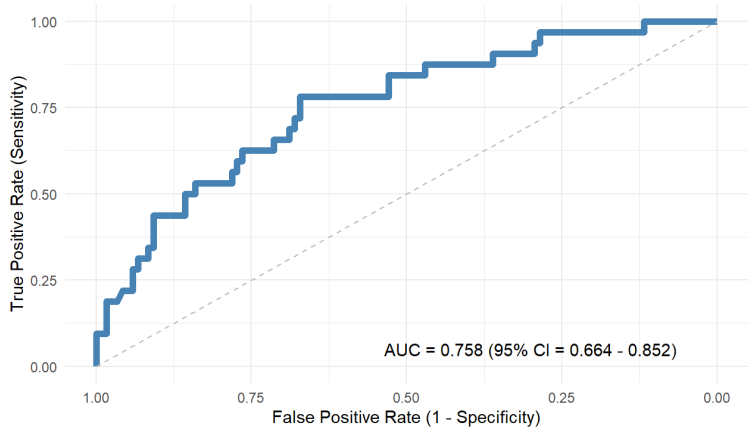


Figure 4.16: ROC curve for the adjusted MLR model after stepwise selection method

The AUC of the Stepwise model suggests the same as the AUC of the first model, that the model has a moderate discriminatory power.

Comparison of the two models

The comparison of the two models shown above provides valuable insights into which model performs better in predicting the outcome of interest, in the case of this study Biochemical recurrence.

When comparing two models using AIC or BIC, the model with the lower AIC or BIC value is considered to be a better fit for the data.

Table 4.5: AIC and BIC values for both models

	AIC	BIC
MLR model	147.707	175.015
Stepwise model	140.143	152.212

The AIC as a difference of almost 8 ($\Delta_i = 7.564$) showing a that the stepwise model is much stronger and a BIC difference of more that 22, which makes sense since BIC penalizes models with more parameters.

To compare both AUCs, the Delong test was employed, as it can be utilized to demonstrate statistically significant differences between theAUC values of two models.[47].

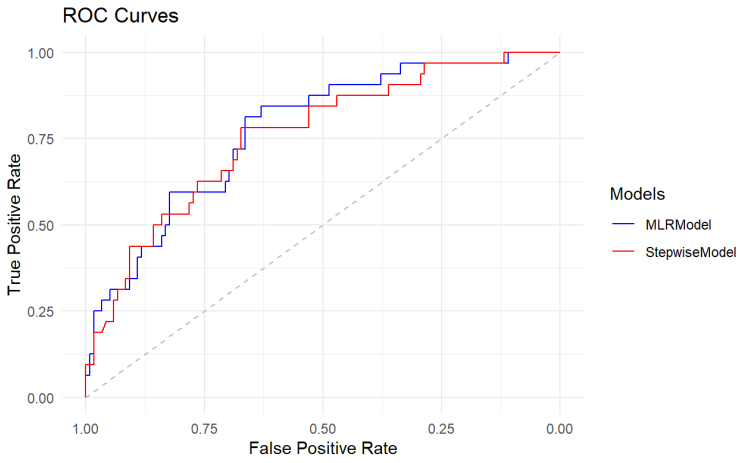


Figure 4.17: ROC curves for both of the models

The DeLongs test for the two ROC curves produced a p-value of 0.3637, which means that it cannot be concluded that there is a significant difference in the AUCs of the two models.

In summary, based on the analysis conducted, the stepwise model is selected and retained as the final model. Although there were no significant differences observed between the models AUCs, the evaluation of the BIC and AIC indicated that the stepwise model performed better.

4.3 Survival Analysis

This sub-chapter presents the results of the survival analysis, which examined the time-to-event data and the associated factors.

The main findings of the study related to time-to-event outcomes will be presented. Specifically, the survival table will be analysed, as well as a description and comparison of the survival curves. The Cox proportional hazards model will be accessed in order to collect the hazard ratios and the associated p-values for each predictors choose the best survival model.

4.3.1 Kaplan-Meier

One of the most commonly used tools in survival analysis is the Kaplan-Meier estimator, which is used to estimate the survival function of a group of individuals.

Table 4.6: Survival table

Time	Survival (CI 95%)	SE	Cases at risk
433	0.993 (0.979, 1.000)	0.00717	139
452	0.986 (0.966, 1.000)	0.01014	137
455	0.978 (0.954, 1.000)	0.01239	136
491	0.971 (0.943, 0.999)	0.01431	133
611	0.963 (0.932, 0.995)	0.01623	123
687	0.955 (0.920, 0.991)	0.01802	118
688	0.947 (0.909, 0.986)	0.01963	117
697	0.939 (0.898, 0.981)	0.02109	116
704	0.930 (0.887, 0.975)	0.02243	115
706	0.922 (0.877, 0.970)	0.02367	114
709	0.914 (0.867, 0.964)	0.02485	112
744	0.906 (0.856, 0.958)	0.02600	109
810	0.897 (0.845, 0.952)	0.02717	104
899	0.887 (0.833, 0.945)	0.02848	95
959	0.878 (0.821, 0.938)	0.02976	92
1181	0.866 (0.806, 0.930)	0.03169	73
1199	0.854 (0.791, 0.922)	0.03345	72
1392	0.840 (0.774, 0.913)	0.03549	64
1394	0.827 (0.757, 0.904)	0.03735	63
1486	0.812 (0.738, 0.893)	0.03958	55
1607	0.796 (0.719, 0.883)	0.04179	52
1681	0.780 (0.699, 0.871)	0.04398	49
1744	0.762 (0.676, 0.859)	0.04655	43
1873	0.740 (0.648, 0.845)	0.05005	35
1894	0.718 (0.621, 0.830)	0.05333	33
1935	0.695 (0.592, 0.815)	0.05641	31
2009	0.670 (0.563, 0.798)	0.05960	28
2025	0.645 (0.534, 0.780)	0.06235	27
2096	0.616 (0.499, 0.760)	0.06605	22
2131	0.585 (0.463, 0.738)	0.06955	20
2508	0.501 (0.342, 0.735)	0.09767	7
2593	0.401 (0.224, 0.717)	0.11895	5

The Survival table for the BR data is displayed in table 4.6. The table is composed of 4 columns: Time, Survival (CI 95%), SE and Cases at Risk.

The **Time** column refers to the length of time, in days, between the start of the study and the occurrence of the event of interest (BR).

The **Survival (CI 95%)** column indicates the probability of survival, with a confidence interval of 95%, for an individual to not experience biochemical recurrence at a given interval of time.

The **SE** (Standard Error) column refers to the measure of the variability or uncertainty associated with the estimated survival probability at a given time.

The **Cases at risk** column represents the number of individuals who are still at risk of experiencing the event of interest at a given time.

As time progresses it is possible to see the confidence interval widening, indicating greater uncertainty in the estimate. It is important to recall that only patients who did not experience BR within the first 365 days after RP were included in the study.

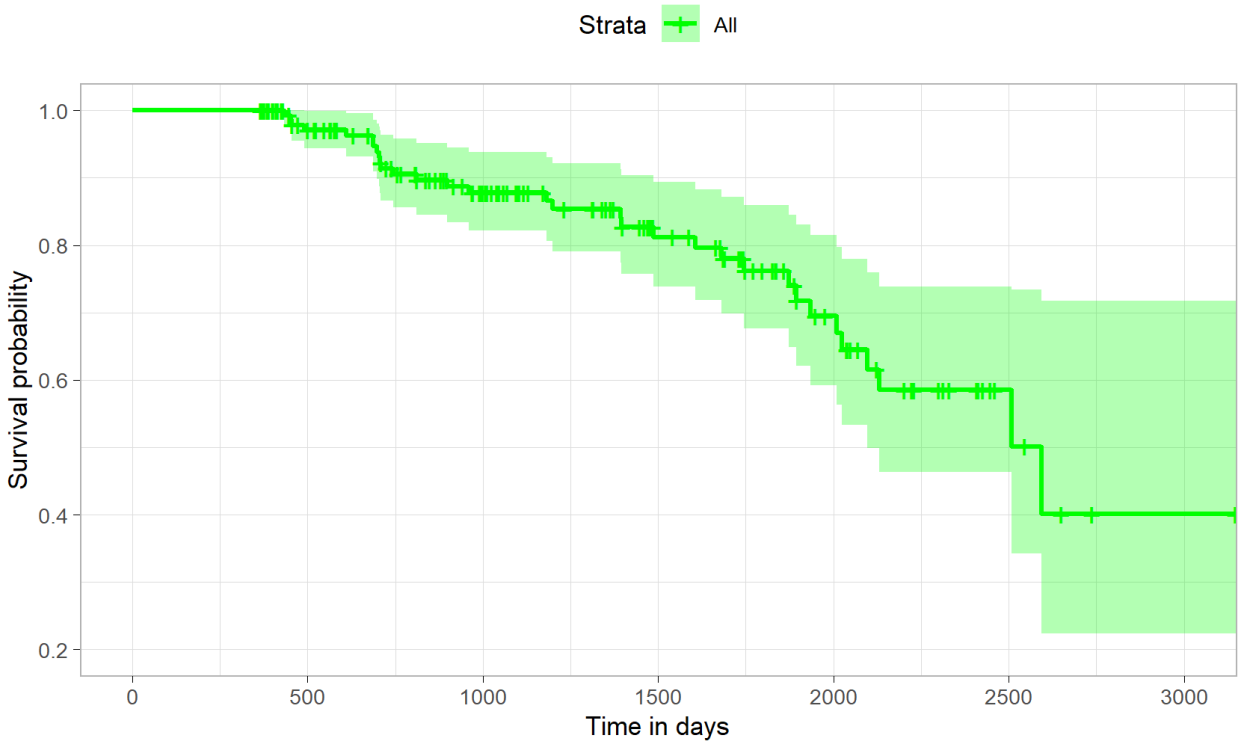
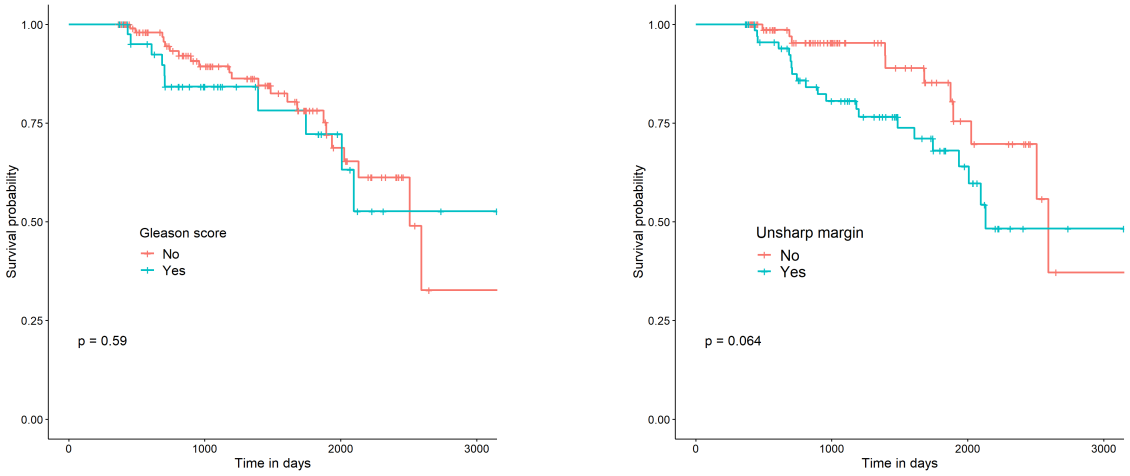


Figure 4.18: Plot of the Survival curve for the probability of a patient at a given time experiencing BR

In figure 4.18, the survival curve is a good way to visually access the computed data in table 4.6. The survival probability starts at around 0.993 at 433 days and decreases to 0.401 at 2593 days.

The curve shows a steep drop in survival probability in the first few years, and then a more gradual decline in later years, suggesting that the first 3 years post-treatment are a critical time period for monitoring patients for biochemical recurrence.

For a better understanding of each variable in terms of survival, Kaplan-Meier curves were computed to analyse which predictor did in fact impact the survival of each patient as well as a log-rank test to access if the survival distribution differs between the groups for each predictor.



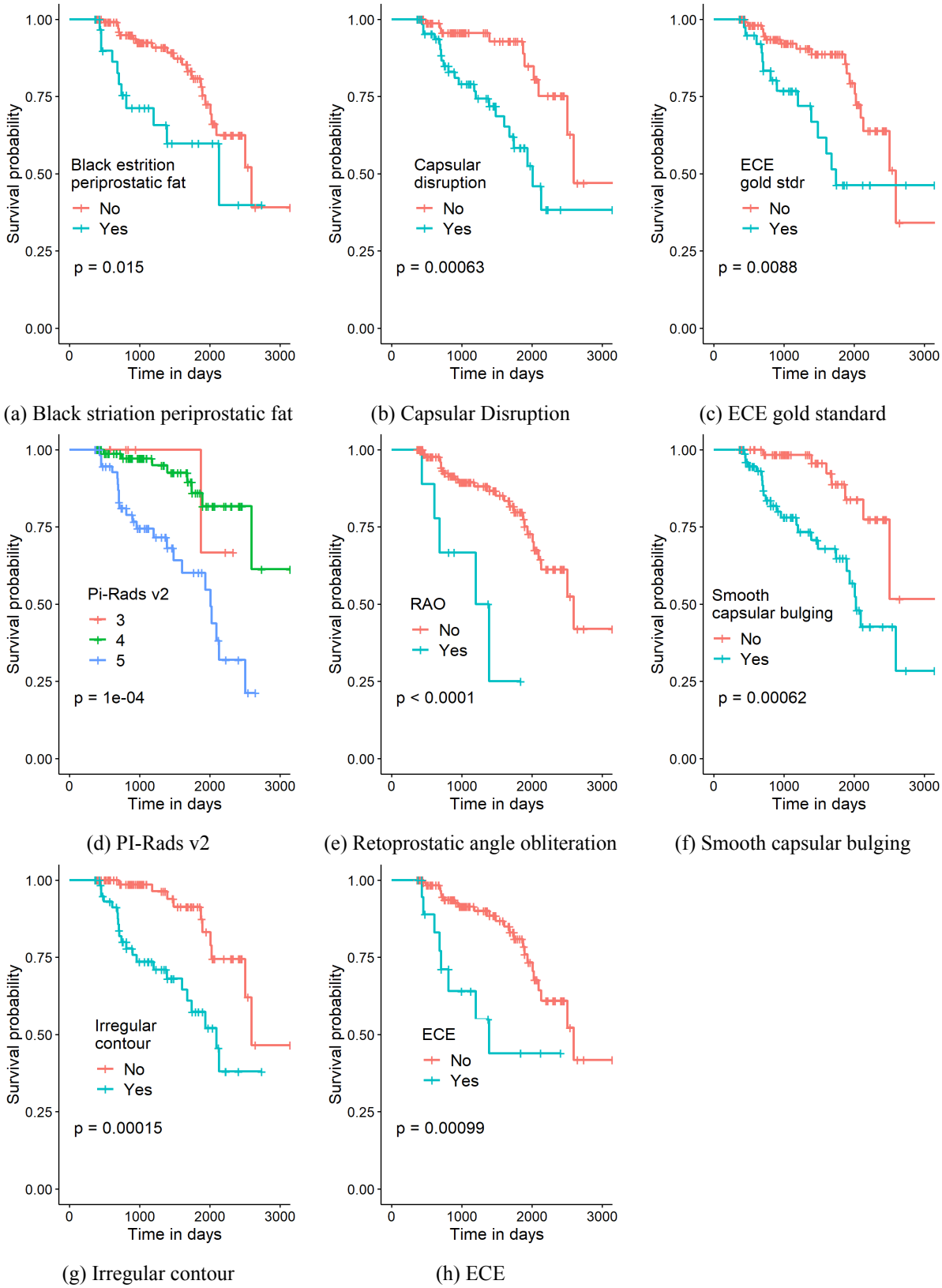
(a) Kaplan-Meier Curves for Gleason score

(b) Kaplan-Meier Curves for Unsharp Margin

Figure 4.19: Kaplan-Meier curves for the variables that are not statistically significant based on log-rank tests.

Of the 10 variables tested, Gleason score and Unsharp margin (figure 4.19) were the only two predictor variables where the p-value were higher than 0.05, meaning that, for those 2 variables, the survival curves for their groups were identical, so it can be concluded that having a Gleason score of grade 3 to 5 does not significantly impact the survival of the patients. It can also be concluded that having a Unsharp margin also does not significantly impact the survival of patients.

For the statistically significant variables, the opposite can be stated. In figure 4.20 are the remaining 8 variables and even without looking at the log-rank test, the differences between the curves are possible to see visually. The fact that the curves mostly do not cross each other is a visual confirmation that these variables have a different survival distribution between their groups.



(a) Black striation periprostatic fat (b) Capsular Disruption (c) ECE gold standard

(d) PI-Rads v2 (e) Retoprostatic angle obliteration (f) Smooth capsular bulging

(g) Irregular contour (h) ECE

Figure 4.20: Kaplan-Meier curves for the variables that are statistically significant based on log-rank tests.

4.3.2 Cox proportional hazards model

The Cox regression model was used to investigate the association between various covariates and the time-to-event outcome. To access which variables are significant, univariate Cox regressions were applied to the dataset.

Table 4.7: Results from univariable Cox regression

Variables		Coefficients	SE	HR (95% CI)	p-value
Continuous Variables					
Age at MRI (years)		0.035	0.028	1.035 (0.980, 1.094)	0.217
Prostate volume (gr)		-0.020	0.014	0.980 (0.953, 1.008)	0.165
PSA at MRI (ng/ml)		0.013	0.043	1.014 (0.932, 1.102)	0.753
Index lesion size (mm)		0.097	0.025	1.102 (1.050, 1.156)	0.000
Capsular contact lenght TLC (mm)		0.093	0.018	1.097 (1.06, 1.136)	0.000
Categorical Variables					
Index lesion PI-Rads V2	3	Reference			
	4	-0.311	1.064	0.733 (0.091, 5.896)	0.770
	5	1.258	1.024	3.518 (0.473, 26.168)	0.219
Unsharp margin	No	Reference			
	Yes	0.681	0.374	1.976 (0.949, 4.113)	0.069
Black striation peripro-static fat	No	Reference			
	Yes	0.876	0.373	2.401 (1.156, 4.987)	0.019
Measurable ECE	No	Reference			
	Yes	1.281	0.416	3.600 (1.594, 8.13)	0.002
Smooth capsular bulging	No	Reference			
	Yes	1.363	0.429	3.908 (1.686, 9.058)	0.001
Capsular disruption	No	Reference			
	Yes	1.253	0.389	3.502 (1.634, 7.502)	0.001
Irregular contour	No	Reference			
	Yes	1.362	0.386	3.904 (1.833, 8.315)	0.000
Retoprostatic angle oblit-eration	No	Reference			
	Yes	1.778	0.512	5.919 (2.169, 16.150)	0.001
ECE gold standard	No	Reference			
	Yes	0.910	0.359	2.484 (1.229, 5.017)	0.011
Gleason score	No	Reference			
	Yes	0.209	0.383	1.232 (0.582, 2.608)	0.586

Table 4.7 contains results from the uni-variate cox regression for each predictor variable and its comprised of 5 columns:

The **Variables** column displays all the variables that were subjected to the uni-variate cox regression.

The **Coefficients** column shows the estimated regression coefficients for each variable.

The **Standard Error (SE)** column gives a measure of uncertainty associated to the estimated regression coefficient.

The **Hazard Ratio (HR)** column displays the HR (and its confidence interval of 95%) that quantifies the association between the presence or absence of a specific variable and the hazard of Biochemical Recurrence..

The **p-value** column represents the p-value of the Wald test, for individual independence significance, of each variable to the response variable.

Looking at the p-value column, it is possible to see that 6 variables were rejected: **Age at MRI, Prostate Volume, PSA at MRI, Index lesion PI-Rads V2, Unsharp margin** and **Gleason score**, meaning that these variables were not significantly associated with BR.

Despite these results, it was considered to keep **Unsharp margin** but as the variable was highly correlated with capsular disruption ($r = 0.69$) and Irregular contour ($r = 0.66$) this idea was discarded.

For the construction of the Cox regression model, the variable **Capsular contact length TLC** was rejected for multicollinearity (with lesion size, $r = 0.72$). Additionally, variables such as **Black striation periprostatic fat** and **Measurable ECE** were not included in the model as they were found to be unrelated to BR, as observed in Table 4.1.

Table 4.8: Results from Cox regression

Variables	**Coefficients**	**SE**	**HR (95% CI)**	**p-value**
Index lesion size	0.060	0.031	1.062 (0.999, 1.129)	0.054
Smooth capsular bulging	0.589	0.498	1.803 (0.679, 4.784)	0.237
Capsular disruption	0.562	0.487	1.755 (0.675, 4.559)	0.248
Irregular contour	0.620	0.475	1.859 (0.733, 4.713)	0.192
ECE gold standard	-0.101	0.468	0.904 (0.361, 2.263)	0.830
Retoprostatic angle obliteration	0.640	0.625	1.896 (0.557, 6.452)	0.306

Table 4.8 contains 5 columns. Each column represents the same as in table 4.7 but for multiple Cox regression instead of uni-variable Cox regression.

Every single variable is rejected from the model, indicating that more variables should be removed.

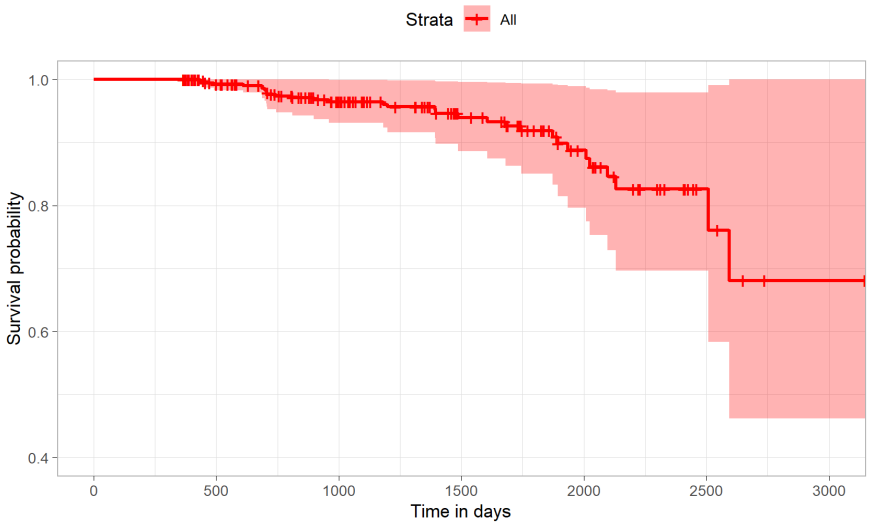


Figure 4.21: Survival curve for the Cox model

The validation of the Cox model is imperative for reliable model inference and interpretation.

One of the assumptions adjacent to the cox model is the proportional hazards (PH) assumption:

Table 4.9: Results from Schoenfeld residuals test for the Cox model

	rho	chisq	p-value
Index lesion size	0.086	0.045	0.831
Smooth capsular bulging	-0.111	1.406	0.236
Capsular disruption	0.151	0.287	0.592
Irregular contour	-0.201	2.678	0.102
Retoprostatic angle obliteration	0.133	0.025	0.875
ECE gold standard	-0.253	3.776	0.052
GLOBAL		5.015	0.542

Rho corresponds to the scaled Schoenfeld residuals for each variable. Based on the above output, none of the covariates have a statistically significant test result, and neither does the global test. As a result, it can be conclude that the proportional hazards assumption holds.

The Schoenfeld residuals are designed to be time-independent. However, if a plot of these residuals exhibits a non-random pattern with respect to time, it suggests a breach of the PH assumption.

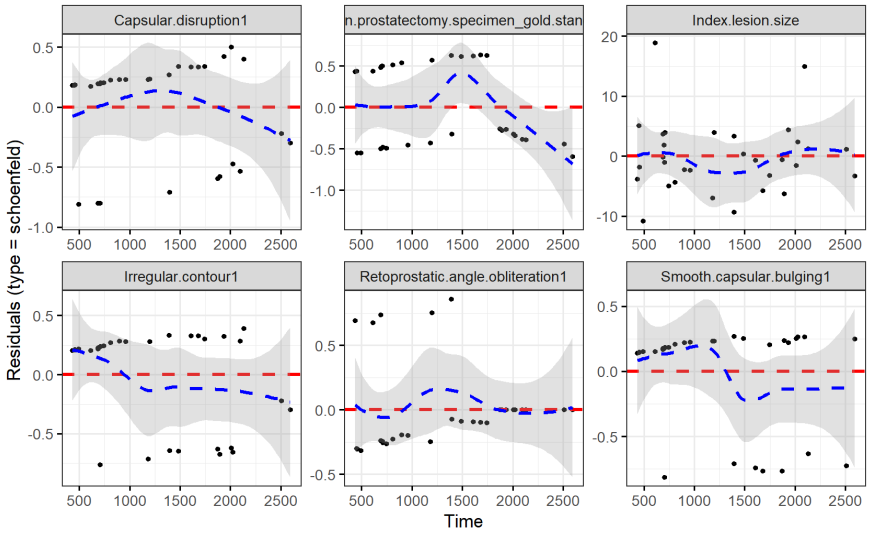


Figure 4.22: Schoenfeld residual for the Cox model

In Figure 4.28, the Schoenfeld residual plot for the Cox model is depicted. To provide a reference, a horizontal line at y=0 (colored in red) has been included. The plot also showcases a solid line (colored in blue), which represents a smoothing spline fit, and a grey area indicating a +/- 2-standard-error band around the fit.

Upon visual inspection, there appears to be no pattern between the Schoenfeld residuals and time, suggesting that the proportional hazards assumption holds for each of the six variables in these Cox model.

One way to identify influential observations or outliers is to examine the deviance residuals graphically:

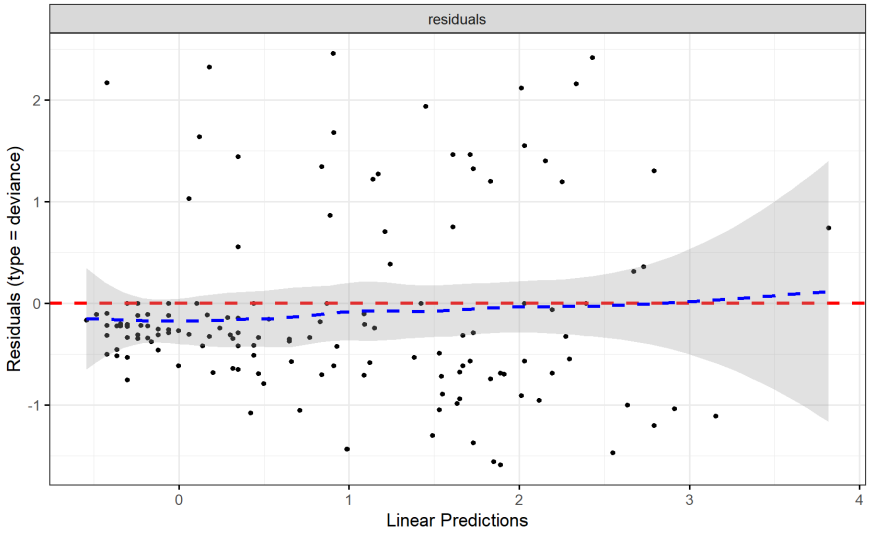


Figure 4.23: Deviance residuals for the Cox model

In Figure 4.23, the deviance residual plot for the Cox model is illustrated. A horizontal line at y=0 (highlighted in red) is included as a reference point. The plot showcases a solid line (displayed in blue) that represents a

smoothing spline fit. Additionally, a grey area is shown, indicating a +/- 2-standard-error band around the fit.

AS the pattern of the data in figure 4.23 appears to be relatively symmetric around zero, it can be concluded that there are no major outliers or influential observations that are affecting the model's performance.

As a final test for the performance of the Cox model, a Cox-Snell residual plot was employed, as it is a effective tool to assess the appropriateness of the proposed model.

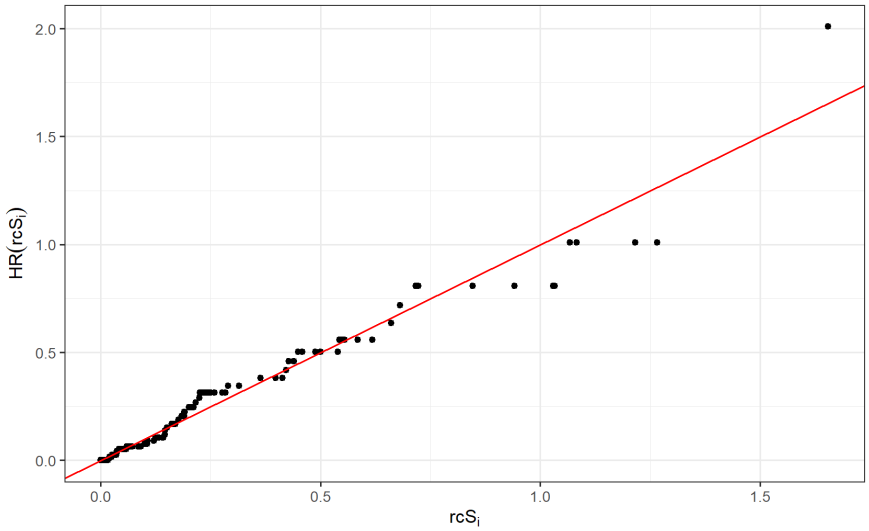


Figure 4.24: Cox-Snell residuals for the Cox model

As the points fall roughly on the line, it can be concluded that the data was fitted with an appropriate model.

stepwise model

To reach the best model possible, the stepwise method was employed.

Table 4.10: Results from cox regression after stepwise method

Variables	**Coefficients**	**SE**	**HR (95% CI)**	**p-value**
Index lesion size	0.080	0.028	1.084 (1.027, 1.144)	0.004
Capsular disruption	0.760	0.451	2.139 (0.884, 5.174)	0.092
Irregular contour	0.729	0.455	2.073 (0.849, 5.061)	0.109

Table 4.10 contains the same variables as in table 4.8. Only left with 3 variables, this model rejects 2 of them (**Capsular disruption** and **Irregular contour**).

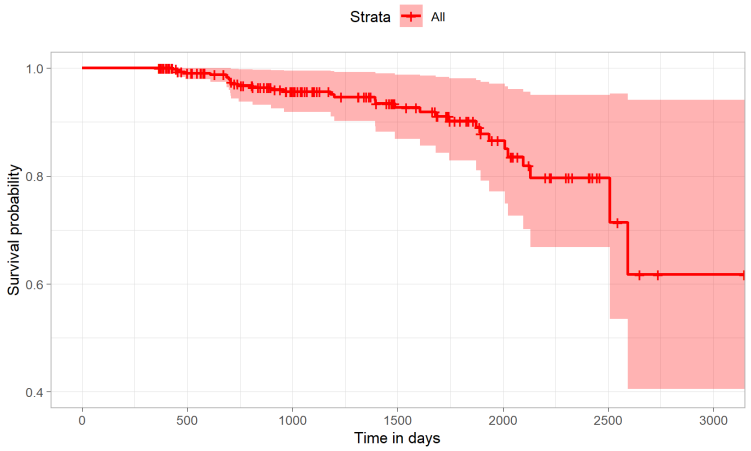


Figure 4.25: Survival curve for the Cox model after the stepwise method

The Cox model after the stepwise method should also be validated, firstly the proportional hazards (PH) assumption:

Table 4.11: Results from Schoenfeld residuals test for the Cox model after the stepwise method

	rho	chisq	p-value
Index lesion size	0.072	0.000	0.999
Capsular disruption	0.036	0.344	0.558
Irregular contour	-0.297	2.780	0.096
GLOBAL		2.840	0.417

Rho corresponds to the scaled Schoenfeld residuals for each variable. Like on the cox model before the stepwise method (table 4.9), none of the covariates or the global test show a statistically significant result. Therefore, it can be concluded that the proportional hazards assumption holds.

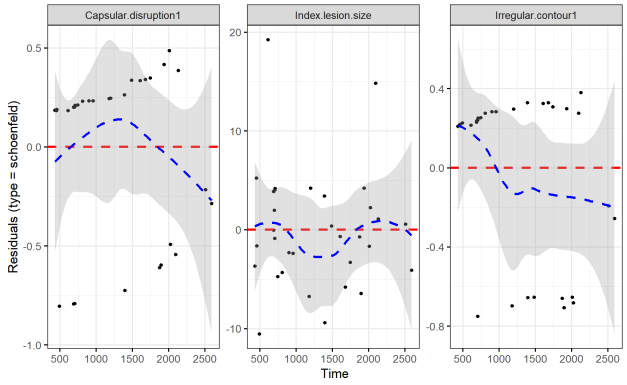


Figure 4.26: Schoenfeld residual for the Cox model after the stepwise method

In Figure 4.26, the Schoenfeld residual plot for the Cox model is depicted. To provide a reference, a horizontal line at y=0 (colored in red) has been included. The plot also showcases a solid line (colored in blue), which represents a smoothing spline fit, and a grey area indicating a +/- 2-standard-error band around the fit. For the 3 variables left in the Cox model after the stepwise method, there appears to be no pattern between the Schoenfeld residuals and time, therefor proportional hazards assumption holds.

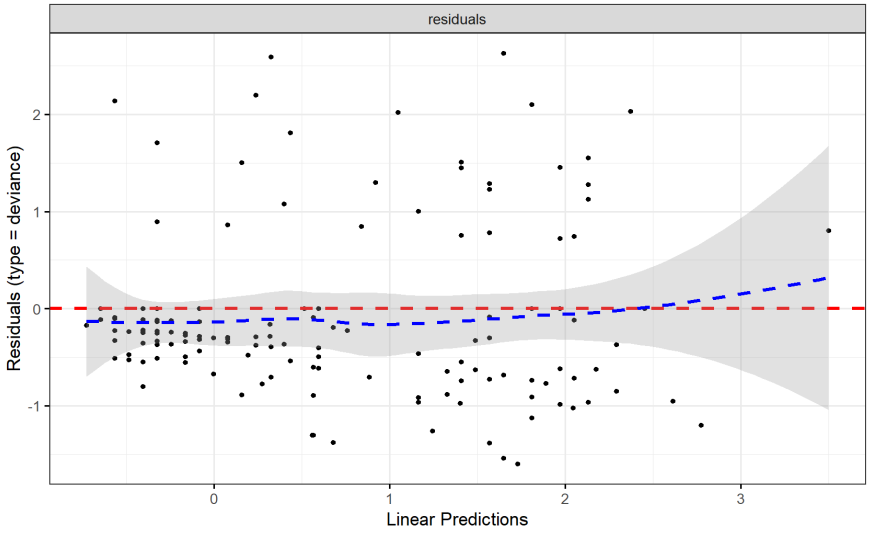


Figure 4.27: Deviance residuals for the Cox model after the stepwise method

In Figure 4.27, the deviance residual plot for the Cox model is illustrated. A horizontal line at y=0 (highlighted in red) is included as a reference point. The plot showcases a solid line (displayed in blue) that represents a smoothing spline fit. Additionally, a grey area is shown, indicating a +/- 2-standard-error band around the fit. To access for influential data points, the points in deviance residuals plot (figure 4.27) also are shown to be relatively symmetric around zero. It can be concluded that there are no major outliers or influential observations that are affecting the model's performance.

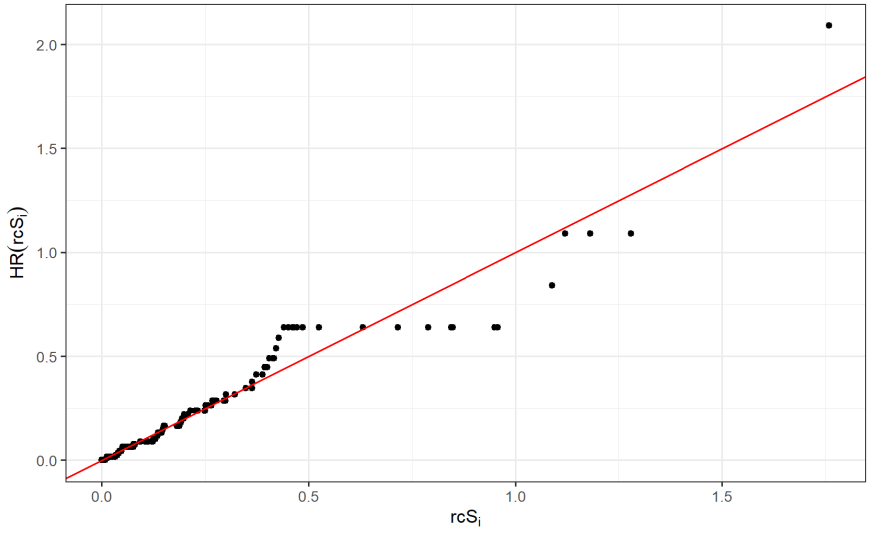


Figure 4.28: Cox-Snell residuals after stepwise method

And finally, to access the performance of the model, the Cox-Snell residual plot was employed and as the

points fall more or less on the line, it can be concluded that the data was also fitted with an appropriate model.

4.3.3 Random Survival Forest model

Random Survival Forest (RSF) model was selected as the final model in this thesis, has it can handle complex relationships between predictors and survival, high-dimensional data, and missing/censored data without specifying functional forms of predictors.

After applying the model to the data, 20 random curves were created along with a black curve that represents the average for each of the patients:

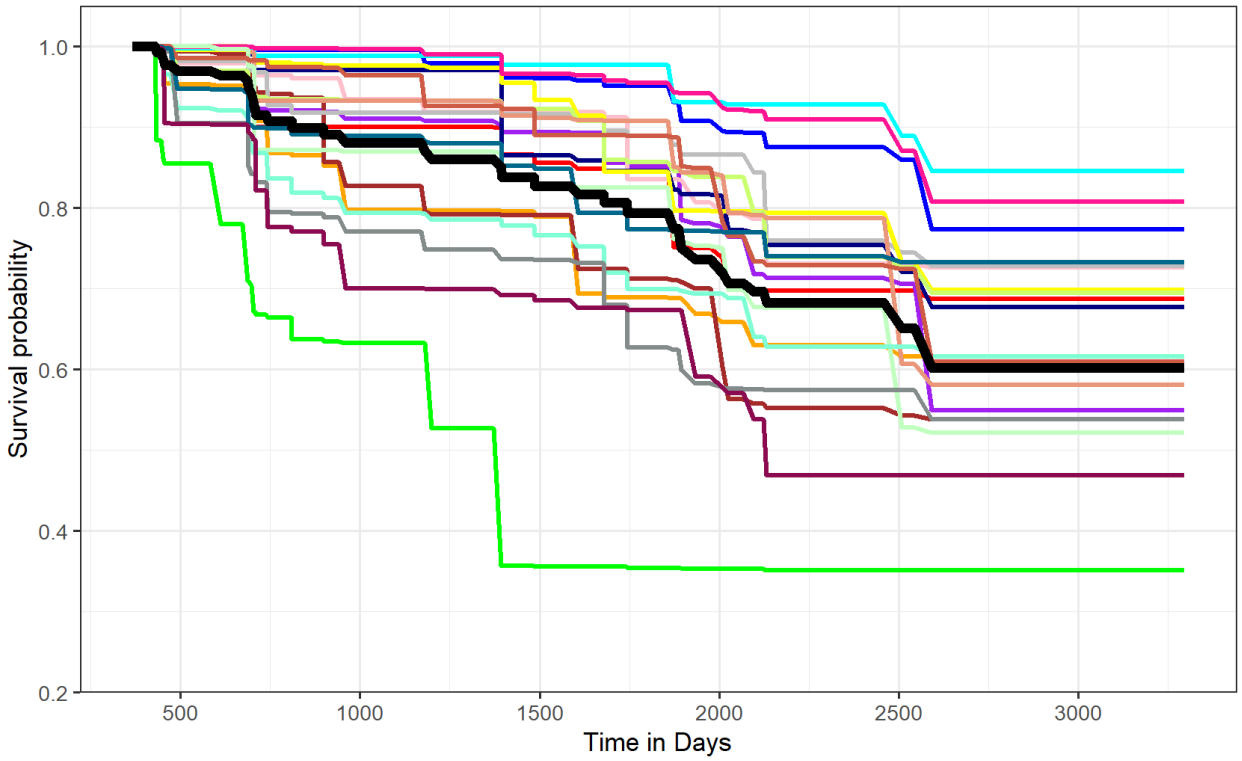


Figure 4.29: 20 random survival curves and a black curve that represents the average, generated by RSF model

This model has a prediction error of 0.293 (1-Harrel's C-index), this indicates that the RSF model has a moderate predictive performance.

Table 4.12: Importance of each variable for a Random Survival Forest Model

Variables	**Importance**
Irregular contour	0.024
Index lesion PI-RADS V2	0.020
Smooth capsular bulging	0.020
Capsular disruption	0.012
Capsular contact length TLC	0.009
Measurable ECE	0.009
Retoprostatic angle obliteration	0.008
ECE gold standard	0.008
Prostate volume	0.004
Index lesion size	0.003

Based on the variable importance measures for the random survival forest model, the top predictors of the outcome are Irregular contour, Index lesion PI-RADS V2, and Smooth capsular bulging, with importance values of 0.0239, 0.0202, and 0.0196, respectively. Capsular disruption and Capsular contact length TLC also showed some importance with values of 0.0117 and 0.0091, respectively. Other predictors such as Measurable ECE, Retoprostatic angle obliteration, and Prostate volume also had some importance but to a lesser extent. The Index lesion size had the lowest importance value among all the predictors. The variables Age at MRI, PSA value at MRI, Unsharp margin, and Black striation periprostatic fat did not show significant importance in predicting survival outcomes in the RSF model. It is important to note that the Standard error was not estimated.

4.3.4 Comparison between RSF and Cox model

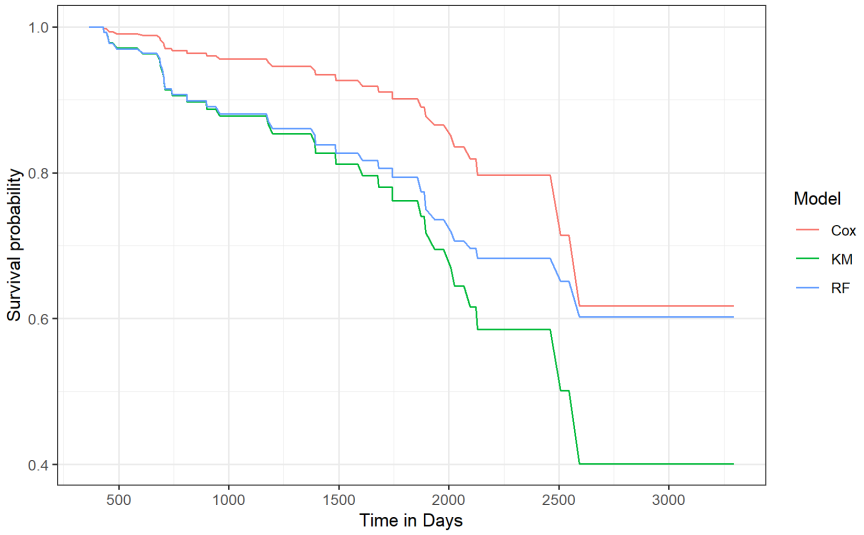


Figure 4.30: Survival probability curves of the 3 models used in the study

In figure 4.30 3 survival curves for the 3 different models are displayed: Cox regression (red), Kaplan-Meier (green) and Random Forest (blue).

It can be inferred that the Cox survival curve exhibits a more optimistic survival trend, whereas the Kaplan-Meier curve shows the opposite. The random survival forest curve falls in between these two, suggesting a moderate outlook in terms of survival probabilities.

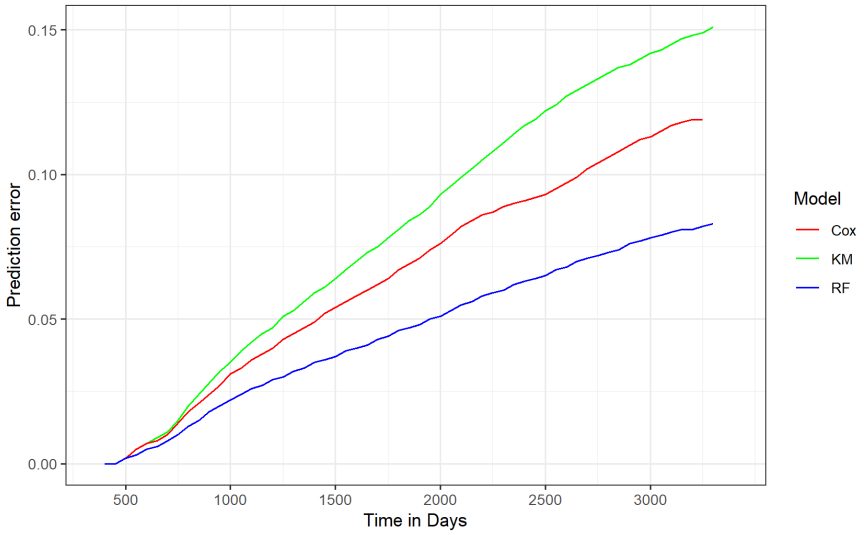


Figure 4.31: Estimated Prediction Error Curves for the Cox model (red) and RSF model (blue) with Kaplan-Meier as reference (green).

In figure 4.31, the estimated Prediction Error Curve (PEC)s obtained from the Cox model and the RSF model can be seen. The RF prediction error curve consistently lies below the Cox curve, suggesting that the RF model has a lower cumulative prediction error, indicating superior predictive performance. Therefore, based

on the Prediction Error Curves, the RSF may be preferred over the Cox model for predicting the event outcomes.

However, Harrell's C-index values were calculated for both the Cox model and the RSF model, resulting in a value of 0.771 for the Cox model and 0.717 for the RSF model. This suggests that the Cox model has a higher discriminatory ability in predicting the event outcomes compared to the RSF model. The higher C-index for the Cox model indicates better overall performance and a stronger ability to correctly rank the survival times of individuals in terms of the event occurrence. Therefore, based on the C-index results, the Cox model may be preferred over the RSF model for predicting the event outcomes.

Chapter 5

Conclusions and Discussion

This study aimed to identify biomarkers for detecting relapses in patients with prostate tumors who underwent radical prostatectomy. Multiple modeling approaches, including logistic regression, survival analysis (Cox proportional hazards model), and random forest modeling, were employed to analyze the relationship between various variables and relapse outcomes.

The analysis of the logistic regression findings revealed that out of the initially considered variables, a total of 10 variables were found to be significant in the simple logistic regression analysis. However, after applying a more rigorous multivariate approach, only three variables, prostate volume, index lesion size, and smooth capsular bulging, remained significant predictors of biochemical recurrence.

Including these three variables in the final logistic regression model suggests their strong association with the likelihood of relapse. The Prostate volume serves as an indicator of tumor size, which can have a direct impact on disease progression and the probability of recurrence. Index lesion size has been consistently identified as a crucial factor in prostate cancer prognosis and relapse prediction[48]. The larger the index lesion size, the higher the risk of relapse post-surgery. Lastly, smooth capsular bulging, which refers to the extension of the tumor beyond the prostate capsule, has also been associated with an increased risk of recurrence.

These findings align with previous research and provide valuable insights into the factors contributing to relapse in prostate cancer patients [49][50]. They emphasize the importance of considering tumor characteristics and their impact on disease progression when assessing relapse risk.

Furthermore, the survival curve analysis demonstrated a gradual decline in the survival probability over time, highlighting an increased risk of relapse as the time since surgery extends. These results underscore the importance of continued surveillance and appropriate interventions to mitigate the risk of relapse in patients with prostate tumors. The Cox proportional hazards model complemented the logistic regression analysis by considering the time-to-event nature of relapse outcomes.

Additionally, the random forest model yielded a moderate prediction error of 0.293 (1-Harrel's C-index). However, as stated in the results section regarding the rsf model, the Variable Importance (VIMP) of the rsf

model did not have a estimated standard error, which could have been helpful in assessing the precision of the estimates. In future approaches, this can be addressed by employing a subsampling approach to estimate the standard error [51].

It is important to interpret the results of the logistic regression analysis alongside the survival analysis and random forest modeling. Each approach contributes unique insights into the relationship between variables and relapse outcomes, taking into account different statistical and analytical considerations.

The prediction error curves suggest that the Random Survival Forest (RSF) model performs better, as it consistently has a lower cumulative prediction error compared to the Cox model. This indicates superior predictive performance for the RSF model.

On the other hand, the Harrell's C-index values favor the Cox model, as it demonstrates a higher discriminatory ability and better overall performance in ranking survival times.

The findings of this study indicate that the predictive models, including logistic regression and Cox regression, yielded results that were not as favorable as anticipated. While some significant variables were identified, the overall performance of the models in predicting relapse in patients with prostate tumors post-radical prostatectomy fell short of expectations. Several factors may have contributed to these less-than-optimal outcomes.

On the clinical side of the issue, many factors could be responsible for these results like the heterogeneity of the target population, the lack of additional predictors, the complexity of the disease or simply the size of the sample [52].

There are several alternative models and analytical approaches that could have been considered for analyzing the data in the study like Feature Selection Techniques [53], Neural networks [54] and Decision tree analysis [55] to name a few.

In summary, the combination of logistic regression, survival analysis, and random forest modeling provided valuable insights into the identification of biomarkers for relapse detection in patients with prostate tumors post-radical prostatectomy. However, this works has to be interpreted as a first approach to this complex problem. Future research on this issue shall be carried out to shed light on the intricacies of biochemical recurrence and the time until relapse.

Bibliography

[1] Xiaonan Xue, Xianhong Xie, Marc Gunter, Thomas E Rohan, Sylvia Wassertheil-Smoller, Gloria YF Ho, Dominic Cirillo, Herbert Yu, and Howard D Strickler. Testing the proportional hazards assumption in case-cohort analysis. *BMC Medical Research Methodology*, 13:88, 12 2013.

[2] Janet M. Box-Steffensmeier and Bradford S. Jones. *Event History Modeling- Diagnostic Methods for the Event History Model*. Cambridge University Press, 3 2004.

[3] M. Schumacher, E. Graf, and T. Gerds. How to assess prognostic models for survival data: A case study in oncology. *Methods of Information in Medicine*, 42:564–571, 2 2003.

[4] Jacques Ferlay, Murielle Colombet, Isabelle Soerjomataram, Donald M. Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. Cancer statistics for the year 2020: An overview. *International Journal of Cancer*, 149:778–789, 8 2021.

[5] John C. Bailar and Heather L. Gornik. Cancer undefeated. *New England Journal of Medicine*, 336:1569–1574, 5 1997.

[6] Prashanth Rawla. Epidemiology of prostate cancer. *World Journal of Oncology*, 10:63–89, 2019.

[7] International Agency for Research on Cancer. Global cancer organization, 2020.

[8] Ann C. Klassen and Elizabeth A. Platz. What can geography tell us about prostate cancer? *American Journal of Preventive Medicine*, 30:S7–S15, 2 2006.

[9] Allison S. Glass, K. Clint Cary, and Matthew R. Cooperberg. Risk-based prostate cancer screening: Who and how? *Current Urology Reports*, 14:192–198, 6 2013.

[10] Gillian Murphy, Masoom Haider, Sangeet Ghai, and Boraiah Sreeharsha. The expanding role of mri in prostate cancer. *American Journal of Roentgenology*, 201:1229–1238, 12 2013.

[11] John R. Prensner, Mark A. Rubin, John T. Wei, and Arul M. Chinnaiyan. Beyond psa: The next generation of prostate cancer biomarkers. *Science Translational Medicine*, 4, 3 2012.

[12] Esmée C.A. van der Sar, Veeru Kasivisvanathan, Mrishta Brizmohun, Alex Freeman, Shonit Punwani, Rifat Hamoudi, and Mark Emberton. Management of radiologically indeterminate magnetic resonance imaging signals in men at risk of prostate cancer. *European Urology Focus*, 5:62–68, 1 2019.

[13] Stephen J. Freedland, Elizabeth B. Humphreys, Leslie A. Mangold, Mario Eisenberger, Frederick J. Dorey, Patrick C. Walsh, and Alan W. Partin. Risk of prostate cancer–specific mortality following biochemical recurrence after radical prostatectomy. *JAMA*, 294:433, 7 2005.

[14] Sandro Sperandei. Understanding logistic regression analysis. *Biochemia Medica*, pages 12–18, 2014.

[15] Chao-Ying Joanne Peng and Tak-Shing Harry So. Logistic regression analysis and reporting: A primer. *Understanding Statistics*, 1:31–70, 2 2002.

[16] S. A. Czepiel. Maximum likelihood estimation of logistic regression models: theory and implementation. 2002.

[17] R. Dennis Cook. Cook's distance, 2011.

[18] Jill C. Stoltzfus. Logistic regression: A brief primer. *Academic Emergency Medicine*, 18:1099–1104, 10 2011.

[19] CFI Team. Variance inflation factor (vif), 6 2020.

[20] N. A. M. R. Senaviratna and T. M. J. A. Cooray. Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics*, pages 1–9, 10 2019.

[21] Patrícia Bermudez and Marília Antunes. Generalized linear models class notations. 2021.

[22] Kenneth P. Burnham and David R. Anderson. Multimodel inference. *Sociological Methods Research*, 33:261–304, 11 2004.

[23] V.A. Profillidis and G.N. Botzoris. Trend projection and time series methods. *Modeling of Transport Demand*, pages 225–270, 2019.

[24] Julie Lorah and Andrew Womack. Value of sample size for computation of the bayesian information criterion (bic) in multilevel modeling.

[25] Jessica Kasza. Stata tip 125: Binned residual plots for assessing the fit of regression models for binary outcomes. *Stata Journal*, (199-2018-3586), 2015.

[26] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. pages 233–240. ACM Press, 2006.

[27] Jane V. Carter, Jianmin Pan, Shesh N. Rai, and Susan Galandiuk. Roc-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*, 159:1638–1645, 6 2016.

[28] John Muschelli. Roc and auc with a binary predictor: a potentially misleading metric. *Journal of Classification*, 37:696–708, 10 2020.

[29] Jayawant N. Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5:1315–1316, 9 2010.

[30] Olga V. Demler, Michael J. Pencina, and Ralph B. D'Agostino. Misuse of delong test to compare aucs for nested models. *Statistics in medicine*, 31:2577, 10 2012.

[31] Liberato Camilleri. History of survival analysis. *The Sunday Times of Malta*, page 53, 3 2019.

[32] Cristina Rocha and Ana Luísa Papoila. *Análise de Sobrevivência*. Sociedade Portuguesa de Estatística, 10 2009.

[33] John P. Klein and Melvin L. Moeschberger. Censoring and truncation, 2003.

[34] Ahmed Barakat, Aaina Mittal, David Ricketts, and Benedict A Rogers. Understanding survival analysis: actuarial life tables and the kaplan–meier plot. *British Journal of Hospital Medicine*, 80:642–646, 11 2019.

[35] Ilari Kuitunen, Ville T. Ponkilainen, Mikko M. Uimonen, Antti Eskelinen, and Aleksi Reito. Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review. *BMC Musculoskeletal Disorders*, 22:489, 12 2021.

[36] Yishu Xue and Elizabeth D. Schifano. Diagnostics for the cox model. *Communications for Statistical Applications and Methods*, 24:583–604, 11 2017.

[37] UCLA-Advanced Research Computing Statistical methods and Data Analytics. Applied survival analysis, chapter 6 | r textbook examples.

[38] Matthias Schmid, Marvin N. Wright, and Andreas Ziegler. On the use of harrell's c for clinical risk prediction via random survival forests. *Expert Systems with Applications*, 63:450–459, 11 2016.

[39] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2, 9 2008.

[40] Ulla B. Mogensen, Hemant Ishwaran, and Thomas A. Gerds. Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software*, 50:1, 2012.

[41] Kaci L Pickett, Krithika Suresh, Kristen R Campbell, Scott Davis, and Elizabeth Juarez-Colunga. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC Medical Research Methodology*, 21:216, 10 2021.

[42] Rafael R. Tourinho-Barbosa, Jean de la Rosette, and Rafael Sanchez-Salas. Prostate cancer multifocality, the index lesion, and the microenvironment. *Current Opinion in Urology*, 28:499–505, 11 2018.

[43] Aslıhan Onay, Metin Vural, Ayse Armutlu, Sevda Ozel Yıldız, Murat Can Kiremit, Tarık Esen, and Barıs Bakır. Evaluation of the most optimal multiparametric magnetic resonance imaging sequence for determining pathological length of capsular contact. *European Journal of Radiology*, 112:192–199, 3 2019.

[44] Jeremy Jones and Marcin Czarniecki. Prostate imaging-reporting and data system (pi-rads), 3 2014.

[45] Prostate Cancer Foundation. Gleason score and grade group.

[46] James P. Stevens. Outliers and influential data points in regression analysis. *Psychological Bulletin*, 95:334–344, 3 1984.

[47] Michael J. Pencina, Ralph B. D'Agostino, Ralph B. D'Agostino, and Ramachandran S. Vasan. Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in medicine*, 27:157–172, 1 2008.

[48] Anthony V. D'Amico. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *JAMA*, 280:969, 9 1998.

[49] Stephen J. Freedland, Elizabeth B. Humphreys, Leslie A. Mangold, Mario Eisenberger, Frederick J. Dorey, Patrick C. Walsh, and Alan W. Partin. Risk of prostate cancer–specific mortality following biochemical recurrence after radical prostatectomy. *JAMA*, 294:433, 7 2005.

[50] Thomas Van den Broeck, Roderick C.N. van den Bergh, Nicolas Arfi, Tobias Gross, Lisa Moris, Erik Briers, Marcus Cumberbatch, Maria De Santis, Derya Tilki, Stefano Fanti, Nicola Fossati, Silke Gillessen, Jeremy P. Grummet, Ann M. Henry, Michael Lardas, Matthew Liew, Olivier Rouvière, Jakub Pecanka, Malcolm D. Mason, Ivo G. Schoots, Theo H. van Der Kwast, Henk G. van Der Poel, Thomas Wiegel, Peter-Paul M. Willemse, Yuhong Yuan, Thomas B. Lam, Philip Cornford, and Nicolas Mottet. Prognostic value of biochemical recurrence following treatment with curative intent for prostate cancer: A systematic review. *European Urology*, 75:967–987, 6 2019.

[51] Hemant Ishwaran and Min Lu. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine*, 38:558–582, 2 2019.

[52] John R. Prensner, Mark A. Rubin, John T. Wei, and Arul M. Chinnaiyan. Beyond psa: The next generation of prostate cancer biomarkers. *Science Translational Medicine*, 4, 3 2012.

[53] Ivan Jambor, Ugo Falagario, Parita Ratnani, Ileana Montoya Perez, Kadir Demir, Harri Merisaari, Stanislaw Sobotka, George K. Haines, Alberto Martini, Alp Tuna Beksac, Sara Lewis, Tapio Pahikkala, Peter Wiklund, Sujit Nair, and Ash Tewari. Prediction of biochemical recurrence in prostate cancer patients who underwent prostatectomy using routine clinical prostate multiparametric mri and decipher genomic score. *Journal of Magnetic Resonance Imaging*, 51:1075–1085, 4 2020.

[54] Yalei Chen, Ian M. Loveless, Tiffany Nakai, Rehnuma Newaz, Firas F. Abdollah, Craig G. Rogers, Oudai Hassan, Dhananjay Chitale, Kanika Arora, Sean R. Williamson, Nilesh S. Gupta, Benjamin A. Rybicki, Sudha M. Sadasivan, and Albert M. Levin. Convolutional neural network quantification of gleason pattern 4 and association with biochemical recurrence in intermediate-grade prostate tumors. *Modern Pathology*, 36:100157, 7 2023.

[55] Nathan C. Wong, Cameron Lam, Lisa Patterson, and Bobby Shayegan. Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy. *BJU International*, 123:51–57, 1 2019.