CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Transcriptomics-based prediction of recurrence-free survival in prostate cancer following prostatectomy
Authors	O'Donnell, Autumn
Publication date	2024
Original Citation	O'Donnell, A. 2024. Transcriptomics-based prediction of recurrence-free survival in prostate cancer following prostatectomy. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2024, Autumn O'Donnell. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-11 13:27:21
Item downloaded from	https://hdl.handle.net/10468/17534

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Transcriptomics-Based Prediction of Recurrence-Free Survival in Prostate Cancer Following Prostatectomy

Autumn O'Donnell

UCC

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

NATIONAL UNIVERSITY OF IRELAND, CORK

SCHOOL OF MATHEMATICAL SCIENCES

DISCIPLINE: STATISTICS

A thesis submitted for the degree of

Doctor of Philosophy

Successfully defended on 4^{th} December 2024

Head of School: Dr. Kevin Hayes

Head of Discipline: Dr. Eric Wolsztynski

Supervisors: Dr. Eric Wolsztynski

Dr. Michael Cronin

Dr. Shirin Moghaddam

I, Autumn O'Donnell, certify that the work I am submitting in this thesis is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within this thesis. I have read and understood the regulations of University College Cork concerning plagiarism.

Autumn O'Donnell

Contents

List of Contributions vi
Acknowledgements viii
Abstract xi

1 Introduction **1**
1.1 General Context 1
1.1.1 Prostate Cancer 2
1.1.2 Conventional Survival Analysis 3
1.1.3 Machine Learning for Survival Analysis 4
1.2 Research Contribution 5
1.2.1 Methodology 5
1.2.2 Time-point of Evaluation 6
1.2.3 The Dataset 8
1.2.4 Aims and Research Questions 9
1.3 References 11

2 Systematic Review **17**
2.1 Introduction 18
2.2 Background 19
2.2.1 Survival Analysis 19
2.2.2 Machine Learning 21

2.3 Methods 26
2.3.1 Study Registration 26
2.3.2 Search Strategy 26
2.3.3 Data Extraction 28
2.3.4 Quality Assessment 28
2.4 Results 29
2.4.1 Data 29
2.4.2 Modelling Frameworks 34
2.4.3 Models 37
2.4.4 Evaluation 38
2.4.5 Stability and Feature Selection 41
2.5 Discussion 45
2.5.1 Comparison of Traditional Methods 45
2.5.2 Stability and Modelling Structures 54
2.5.3 Bias and Quality Assessment 55
2.6 Systematic Review - Prostate Cancer 56
2.6.1 Prostate Cancer - Results 56
2.6.2 Prostate Cancer - Discussion 68
2.7 Conclusion 71
2.8 References 72

3 Post-Operative Prediction of BCR-free Survival 86
3.1 Introduction 87
3.2 Materials and Methods 89
3.2.1 Data 89
3.2.2 Statistical Analysis 90
3.2.3 Evaluation Framework 93
3.3 Results 94

3.3.1 Discrimination 94
3.3.2 Calibration 96
3.3.3 Prediction Performance 98
3.3.4 Model Stability and Feature Selection 99
3.4 Discussion 104
3.4.1 Key Findings 104
3.4.2 Other Markers of Interest 105
3.4.3 Model Performance and Key Feature Identification . . 105
3.5 Conclusion 107
3.6 References 108

4 Pre-Operative Prediction of BCR-free Survival 117
4.1 Introduction 118
4.2 Materials and Methods 121
4.2.1 Data 121
4.2.2 Models 121
4.2.3 Evaluation Framework 124
4.3 Results 126
4.3.1 Discrimination 127
4.3.2 Calibration 128
4.3.3 Predictive Performance 129
4.3.4 Model Stability and Feature Selection 132
4.4 Discussion 135
4.5 Conclusion 138
4.6 References 140

5 Discussion & Conclusion 147
5.1 Research Aims & Contributions 147

5.2 Discussion 150
5.2.1 Systematic Review 150
5.2.2 Post-operative Prediction 153
5.2.3 Pre-operative Prediction 156
5.3 Further & Future Research 158
5.3.1 Pre-operative Genetic Biomarkers 158
5.3.2 Dimensionality Reduction 160
5.3.3 Beyond Biochemical Recurrence 162
5.4 Conclusion 163
5.5 References 164

Appendix A **165**

List of Contributions

List of publications

Chapters 2, 3, and 4 of this thesis correspond respectively to the following papers:

- O'Donnell A, Cronin M, Moghaddam S, Wolsztynski E. A systematic review on machine learning techniques for survival analysis in cancer. *Submitted for review (2024)*.
- O'Donnell A, Wolsztynski E, Cronin M, Moghaddam S. Improving the post-operative prediction of BCR-free survival time with mRNA variables and machine learning. Cancers. 2023;15(4):1276.
- O'Donnell A, Cronin M, Moghaddam S, Wolsztynski E. Pre-operative prediction of BCR-free survival with mRNA variables in prostate cancer. PLoS ONE. 2024;19(10):e0311162.

List of presentations

Preliminary research from the above work has also been presented at the following seminars and conferences:

- Oral presentation, "Prognostic Modelling of Time-To-Event Endpoints In Prostate Cancer", Forty-second Conference of Applied Statistics in Ireland (CASI), 2022 (Cork, Ireland).
- Poster presentation, "Potential of mRNA Variables in Prediction of Poor Prognosis Prostate Cancer", the European Association of Cancer Research (EACR) Conference on Cancer Genomics, 2022 (Oxford, UK).
- Oral presentation, "Machine Learning Methodologies for Modelling of Time-to-Event Endpoints in Prostate Cancer", Forty-third annual conference of the International Society of Clinical Biostatistics (ISCB), 2022 (Newcastle, UK).
- Oral presentation, "Investigating Machine Learning Methods for the Inclusion of Serum Biomarkers in Time-to-BCR Prediction.", Forty-third Conference of Applied Statistics in Ireland (CASI), 2023 (Killarney, Ireland).
- Poster presentation, "A Systematic Review on Machine Learning Techniques for Survival Analysis in Cancer", Forty-fourth annual conference of the International Society of Clinical Biostatistics (ISCB), 2023 (Milan, Italy).

Acknowledgements

First and foremost, I would like to thank my supervisors, Dr. Eric Wolsztynski, Dr. Michael Cronin and Dr. Shirin Moghaddam. To Eric, thank you for your generosity with your time, no matter how busy you were, for kindly taking over the position of first supervisor and providing me with opportunities to showcase my work. Thank you, Michael, for helping me grow not only as a researcher but also as a lecturer and providing me with guidance in my career path. To Shirin, thank you for the initial project proposal and for providing me with possibilities to progress outside of my research.

I would also like to thank the bodies whose support made this research possible. This research was supported by the Higher Education Authority (HEA) Human Capital Initiative Pillar 1, 2020, and in part by Science Foundation Ireland under grant number 12/RC/2289-P2, co-funded under the European Regional Development Fund.

Thank you to Professor William Watson of the Prostate Cancer Research Consortium, University College Dublin for your collaboration and sharing of the Irish liquid-biopsy cohort. Thank you to Dr. Rebecca Taylor of Eli Lilly for your guidance during my internship and for inspiring me to explore new pathways in my research.

Finally, to my family and friends, without whose support, none of this would have been possible. To Zillah and Michael for supporting me and my

decision to return to academia. To Gareth for making sure I could return and for your guidance throughout the process. To Megan, thank you for being the best friend, mentor, remote 'colleague', and collaborator anyone could ask for. And to Rowan, thank you for being my rock and keeping everything else going when I couldn't.

I don't see how he can EVEN finish, if he doesn't begin.

– Lewis Carroll

Abstract

This thesis investigates the application of machine learning (ML) methods for survival analysis, to the prediction of biochemical recurrence (BCR) in prostate cancer patients following radical prostatectomy, using mRNA expression data. While conventional survival analysis methods remain commonplace in clinical applications, they are unfeasible with high-dimensional data such as mRNA assays and rely on assumptions that may not fully capture the complexity of cancer recurrence. This research explores the potential of ML techniques to overcome these limitations, offering more accurate and personalised predictions for prostate cancer outcomes. A systematic review of the current literature on the use of ML for survival analysis in cancer was carried out to identify key methodologies and their performance compared to traditional statistical approaches. This review highlights a need for improved reporting and standardisation in the field. The findings informed the selection of ML models for further study, including random survival forests (RSF), LASSO Cox, and boosted Cox models.

This research's first experimental phase focuses on post-operative BCR prediction, where mRNA data from tumour tissue was integrated with clinical variables with the aim of improving performance. Results demonstrated that ML models incorporating mRNA expression outperformed those with clinical variables alone and could also improve on traditional Cox proportional hazards (CPH) models. The second phase extends the analysis to pre-operative settings, exploring the potential feasibility of using mRNA data to predict BCR before surgery. This pre-operative model also showed significant promise, suggesting the potential for early intervention strategies that could better inform treatment decisions.

The study contributes to the growing field of precision oncology, offering insights into how genetic information and ML can be leveraged to improve clinical decision-making in prostate cancer.

Chapter 1

Introduction

In this chapter, we introduce the general context of biochemical recurrence (BCR) prediction in prostate cancer, and motivate the use of survival analysis based on machine learning (ML) strategies to this end. We then provide an overview of the methodology used in this work, including our consideration of particular time-points on the patient timeline, present the dataset used in our analyses, and outline the main research questions considered in this thesis.

1.1 General Context

Survival analysis is a field of statistics fundamental to cancer research. This branch of statistics is concerned with modelling and analysing the time until a specific event, such as death, disease recurrence, or progression, occurs, making it particularly relevant for understanding the dynamics of cancer outcomes and providing crucial insights into the prognosis and progression of a disease [1].

Survival analysis is thus used to evaluate the prognosis of cancer patients by estimating survival probabilities and identifying different risk groups within a population of patients [2]. It is also key in assessing the efficacy of treatments [3]. By comparing estimates of survival curves and other related quantities, researchers can determine the effectiveness of new treatments or inter-

ventions, which is particularly important in clinical trials aiming to establish the superiority of novel therapies over standard care.

Furthermore, survival analysis facilitates the identification of prognostic information [4]. Using Cox regression and other such survival models that allow including patient, clinical, and other relevant information in the characterisation of survival patterns, researchers can identify demographic (age, sex), clinical (tumour size, stage), and biological (biomarkers, genetic mutations) variables that significantly influence risk and measure the effect of these predictors on the outcome of interest. An estimate of the probability of survival derived from such techniques can thus be employed to determine if a particular treatment regime will suit the individual [5]. These insights are vital for personalised medicine, enabling tailored treatment strategies based on individual risk profiles. The use of survival analysis for this purpose is the focus of this thesis in the context of prostate cancer.

1.1.1 Prostate Cancer

Prostate cancer (PCa) remains one of the most prevalent malignancies among men worldwide, presenting a significant public health challenge due to the variability in associated clinical outcomes [6]. A critical aspect of managing prostate cancer involves predicting the likelihood of biochemical recurrence (BCR) following initial treatment [7]. BCR, defined by a rise in prostate-specific antigen (PSA) levels, often precedes clinical relapse and metastasis, making its prediction critical for timely and personalised therapeutic intervention [8].

Prostate cancer is most commonly treated through surgical intervention, usually by radical prostatectomy [9]. Despite this procedure often involving the complete removal of the prostate gland, patients may still experience

BCR [10]. The management of prostate cancer, and in particular the prediction of BCR, has traditionally relied on clinical and pathological variables such as PSA levels, Gleason scores, tumour staging, and surgical margins [8]. However, more precise prediction remains desired [11].

In recent years, there has been growing interest in the molecular underpinnings of prostate cancer, particularly the role of messenger ribonucleic acid (mRNA) expression profiles [12–14]. mRNA is a crucial intermediary between genes and proteins, reflecting real-time changes in gene activity [15]. Molecular data offers a more detailed and dynamic picture of tumour behaviour, potentially unveiling novel biomarkers for recurrence [1]. By integrating mRNA expression data, this research aims to enhance the predictive power of prognostic models, leading to better management of patients at risk of BCR.

1.1.2 Conventional Survival Analysis

Conventional survival analysis techniques, such as the Cox proportional hazards (CPH) model, have been widely used in medical research [1]. The CPH model assesses the impact of multiple covariates on survival time, assuming the hazard of any subject in the population is proportional to a baseline hazard profile, or in other words, that a change in a given covariate has a simple multiplicative effect on the hazard of that subject over the entire survival time [16]. Despite their utility, these methods have notable limitations. For instance, the CPH model assumes that covariate effects on the hazard rate are constant over time, which may not hold true in all clinical scenarios [17]. Additionally, these traditional methods are not appropriate for the analysis of high-dimensional data, such as mRNA expression profiles, where they may lead to overfitting and reduced predictive accuracy [18], and are not evaluable at all when there are more covariates than observations in the data.

1.1.3 Machine Learning for Survival Analysis

Machine learning (ML) methods have the potential to overcome many of the limitations that traditional statistical methods impose [19] by capturing complex, non-linear relationships between covariates and survival outcomes, unlike traditional statistical models [20]. ML techniques are also well-suited for analysing high-dimensional data [21]. Several strategies have already been adapted for survival analysis, including regularisation of the CPH model [22], approaches based on survival trees [23], boosting algorithms [24], support vector machines [25], shallow neural networks [26], and deep learning [27], many of which also do not rely on the proportional hazards assumption. By leveraging ML-based survival analysis methodologies, this research aims to improve BCR prediction in prostate cancer using mRNA expression profiles. A traditional CPH model was thus compared to a tree-based ML model (random survival forest (RSF) [23]), a regularisation approach (Least Absolute Shrinkage and Selection Operator (LASSO) Cox [28]), and a boosting method (boosted Cox [29]).

In order to alleviate the limitations of the CPH model, regularisation allows for the use of the model in high-dimensionality settings by forcing the contribution of less important variables in the model to be reduced (or shrunk) towards zero, effectively reducing model complexity by introducing sparsity. The approach taken in tree-based methods is different. The survival outcome is predicted by recursively partitioning the data into subgroups [30], using a criterion relevant to the characterisation of survival, until the cohort is split into sufficiently well-separated (small) clusters. Ensembles of survival trees to form RSF [23] are more commonly used than single-tree designs, as they mitigate the risk of overfitting from a single-tree model [31] and also

reduce prediction variability. Boosting is a technique that combines multiple base learners into an ensemble to improve overall predictive performance. In boosted Cox models, the learner is a CPH model [24], which yields a scheme well suited for survival analysis. Section 2.2.2 provides more details on these techniques.

1.2 Research Contribution

1.2.1 Methodology

In order to determine which machine learning methods could be used in this application, an extensive literature review was undertaken. During this initial phase of the research, many discrepancies in the methodologies and findings of published papers were found, making it difficult to determine which methods should be considered in our work. This led us to carry out a full systematic review (Chapter 2) to gain a precise and complete understanding of the literature and the potential of machine learning for survival analysis application to cancer research. Due to limited comparability in the results of the systematic review, ML models were selected for their prevalence in the literature and their performance in previous studies [32–37]. The possibility to evaluate the contributions of variables to the models' predictive performance was also considered when selecting methods for implementation. Novel methodologies with modest descriptions of their model structure and no available source code were thus deemed inappropriate (this was the case for deep-learning and multi-task implementations, in particular).

During this PhD research, a particular research question, the feasibility of a deep learning methodology, was considered. Deep learning methods notoriously are black boxes, making their functionality limited in a clinical setting

where understanding the contribution of input information is essential [38]. DeepSurv [27] is a deep neural network implementation with a Cox proportional hazard-based outcome with readily available published code in Python and add-in implementation available in R. In DeepSurv's current iteration, it is not possible to easily extract more than the C-index of the model, and no implementation of a variable importance or similar technique is available. One of this research group's Master of Science students developed a method for extracting this information. Althought the implementation was successful, there remained persistent limitations to it (e.g. in terms of the volume of data treatable), and this approach was therefore deemed unfit for purpose in our work.

1.2.2 Time-point of Evaluation

In a context where the primary aim is to identify patients at higher risk of cancer recurrence and facilitate timely intervention to improve patient outcome, the time-point at which cancer recurrence prediction is carried out considerably influences the prediction output, as it directly affects both the accuracy and clinical usefulness of a survival model.

Many survival models rely on variables that become available only after treatment intervention. These may include information on surgical margins, pathological staging, and post-operative complications. While such variables can be highly predictive of recurrence, their use limits the model to making predictions only after the treatment [39]. This introduces several significant limitations. Predictions are delayed since post-treatment variables are only available after the intervention, which may reduce the opportunity for early intervention strategies that could positively impact patient outcomes. Clinicians often need to make critical decisions about the course of treatment at

the time of diagnosis [2], not just after surgery, and a model that only uses post-treatment data cannot inform these early decisions.

Early identification of high-risk patients would allow for more aggressive or personalised treatment strategies closer to the time of initial diagnosis, potentially improving outcomes and avoiding potential patient morbidity [40]. Pre-operative prediction could guide decisions on whether to pursue surgery, what type of treatment to perform, or whether to combine surgery with other therapies, enabling a more personalised approach to patient care [2]. Moreover, providing accurate recurrence prediction at diagnosis can help significantly in counselling patients about their treatment options and expected outcomes, offering them a clearer understanding of their prognosis [41].

However, developing pre-operative prediction models is challenging because it requires identifying strong predictor variables that are available at or before diagnosis. Before surgery, the available data typically include clinical examination findings, imaging results, and possibly biopsy data, which may not be as detailed or directly correlated with outcomes as post-operative pathological data [39]. It is thus essential to identify reliable pre-operative predictors of recurrence. These predictors must be robust enough to compensate for the absence of post-treatment data. Furthermore, the biological behaviour of cancer can be complex and influenced by factors that only become apparent during or after treatment [42].

Developing effective models that can predict recurrence at diagnosis would be clinically valuable, allowing for early personalised treatment planning, and is one of the key aims of this thesis.

1.2.3 The Dataset

The dataset used for this analysis was developed in a study by Taylor et al. [14] and contains a total of 232 patients and is comprised of two parts, i.e., a set of routine clinical variables and another of more than 26,000 messenger ribonucleic acid (mRNA) variables. The data was accessed on the 1st of May 2021 and was fully anonymised at source (data accessible at NCBI GEO database, accession GSE21032). The RNA data was produced from microarray experiments using Affymetrix Human Exon 1.0 ST Array and pre-processed expression intensities using robust multichip average (RMA) and quantile normalisation followed by log_2 transformation, a common pre-processing procedure in transcriptomics data to mitigate skewness, further details of which can be found in the supplementary material of the original study [14].

In the original study, following radical prostatectomy, patients were followed up at regular intervals with physical exam and serum PSA testing every 3 months for the first year, 6 months for the second year, and annually thereafter. BCR, in this dataset, is defined as PSA $\geq 0.2ng/mL$ on two occasions post-operatively. Note that the actual time-to-BCR may be shorter than the recorded value due to the PSA measurements taking place at set intervals up to one year apart. With respect to BCR, censoring is defined as any loss to following-up including death. It was not possible to take into consideration whether death was due to cancer as this information was not available. The individual values of PSA post-operatively and the time of testing was also not available, only the time-to-BCR information was included in the dataset. A total of 133 and 135 of the retained patients in the post-operative and pre-operative sets, respectively, had mRNA information available. The analysis

was thus carried out on these sub-cohorts.

1.2.4 Aims and Research Questions

This PhD thesis aims to determine the potential of machine learning methods for survival analysis in prostate cancer when mRNA variables are included. The specific aims are as follows.

1. Identify suitable machine learning methods for survival analysis with the potential for improved performance over traditional statistical methods.

2. Determine whether the inclusion of mRNA expression information can improve predictive performance of biochemical recurrence in the **post-operative** setting.

3. Determine whether the inclusion of mRNA expression information has the potential to improve the predictive performance of biochemical recurrence in the **pre-operative** setting.

The research outlined here aimed to contribute towards the improvement of clinical decision-making in prostate cancer care with timely precision medicine, as well as providing a concise and complete summary of the use of machine learning methods for survival analysis in cancer, their performance, and common pitfalls.

This thesis contains five chapters. Chapter 2 presents a systematic review on the use of ML-based survival analysis in cancer applications. Chapters 3 and 4 present two studies on BCR prediction in PCa using these methodologies in the post-operative and pre-operative settings, respectively. These three chapters follow the formatting requirements of specific journals for publication. Each chapter is presented as a complete study, with references following

the main text in each case. Section 2.6 in Chapter 2 is an addition to the main systematic review study for publication focused on the prostate cancer papers. The conclusion in Section 2.7 also reflects this addition.

The systematic review in Chapter 2 addresses our first research aim to identify ML methods that can improve on more conventional methodologies. A secondary aim of this study was to determine which, if any, of the ML methods yielded optimal performance. Chapter 2 closes with a focused discussion on this use in prostate cancer, with the goal of informing the selection of appropriate methods in the rest of our work.

Chapter 3 presents a study on the post-operative prediction of BCR. This addresses the first and second research aims. The dataset used for this research contained mRNA expression data obtained trans-operatively from removed tumour tissue samples, thus making a post-operative analysis the natural first step. As the mRNA data is of high dimensionality, with over 26,000 variables, ML methods were implemented and compared to a traditional statistical approach.

Chapter 4 presents a study on the pre-operative prediction of BCR. This addresses the first and third research aims and is our main contribution towards the improvement of clinical decision-making in prostate cancer prognosis with timely precision medicine. Under the assumption that the mRNA data obtained trans-operatively would be equivalent to that obtained from a biopsy sample, the mRNA expression data was again used, this time in conjunction with pre-operatively available clinicopathological variables using ML methods.

Chapter 5 provides a final discussion and conclusion of the thesis, including a summary of the main findings, directions for future research, a discussion on the limitations of this study, and concluding remarks.

1.3 References

[1] Cox TF. Medical Statistics for Cancer Studies. Chapman and Hall/CRC; 2022.

[2] Kattan MW, Eastham JA, Stapleton AMF, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. JNCI: Journal of the National Cancer Institute. 1998;90:766-71.

[3] Emmerson J, Brown JM. Understanding Survival Analysis in Clinical Trials. Clinical Oncology. 2021 1;33:12-4.

[4] Armitage P, Gehan EA. Statistical methods for the identification and use of prognostic factors. International Journal of Cancer. 1974 1;13:16-36.

[5] Thurtle D, Rossi SH, Berry B, Pharoah P, Gnanapragasam VJ. Models predicting survival to guide treatment decision-making in newly diagnosed primary non-metastatic prostate cancer: a systematic review. BMJ Open. 2019 6;9:e029149.

[6] Greenberg DC, Wright KA, Lophathanon A, Muir KR, Gnanapragasam VJ. Changing presentation of prostate cancer in a UK population – 10 year trends in prostate cancer risk profiles in the East of England. British Journal of Cancer. 2013 10;109:2115-20.

[7] Roberts WB, Han M. Clinical significance and treatment of biochemical recurrence after definitive therapy for localized prostate cancer. Surgical Oncology. 2009 9;18:268-74.

[8] Remmers S, Verbeek JFM, Nieboer D, van der Kwast T, Roobol MJ. Predicting biochemical recurrence and prostate cancer-specific mortality

after radical prostatectomy: comparison of six prediction models in a cohort of patients with screening- and clinically detected prostate cancer. BJU International. 2019 10;124:635-42.

[9] Sanda MG, Cadeddu JA, Kirkby E, Chen RC, Crispino T, Fontanarosa J, et al. Clinically localized prostate cancer: AUA/ASTRO/SUO guideline. Part II: recommended approaches and details of specific care options. The Journal of urology. 2018;199:990-7.

[10] Tourinho-Barbosa R, Srougi V, Nunes-Silva I, Baghdadi M, Rembeyo G, Eiffel SS, et al. Biochemical recurrence after radical prostatectomy: what does it mean? International braz j urol. 2018;44:14-21.

[11] Ross AE, D'Amico AV, Freedland SJ. Which, when and why? Rational use of tissue-based molecular testing in localized prostate cancer. Prostate Cancer and Prostatic Diseases. 2016 3;19:1-6.

[12] Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. The lancet oncology. 2011;12:245-55.

[13] Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PloS one. 2013;8:e66855.

[14] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer cell. 2010;18:11-22.

[15] Akalin A. Computational Genomics with R. Chapman and Hall/CRC; 2020.

[16] Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34:187-202.

[17] Jiang N, Wu Y, Li C. Limitations of using COX proportional hazards model in cardiovascular research. Cardiovascular Diabetology. 2024 6;23:219.

[18] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33:1-22. Available from: `https://www.jstatsoft.org/v33/i01/`.

[19] Yan L, Gao N, Ai F, Zhao Y, Kang Y, Chen J, et al. Deep learning models for predicting the survival of patients with chondrosarcoma based on a surveillance, epidemiology, and end results analysis. Frontiers in Oncology. 2022 8;12.

[20] Chi SQ, Tian Y, Li J, Tong DY, Kong XX, Poston G, et al. Time-dependent and nonlinear effects of prognostic factors in nonmetastatic colorectal cancer. Cancer Medicine. 2017 8;6:1882-92.

[21] Kalakoti Y, Yadav S, Sundar D. Survcnn: A discrete time-to-event cancer survival estimation framework using image representations of omics data. Cancers. 2021 7;13.

[22] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1996 1;58:267-88.

[23] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The annals of applied statistics. 2008;2:841-60.

[24] Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC bioinformatics. 2008;9:1-10.

[25] Belle VV, Pelckmans K, Huffel SV, Suykens JAK. Support vector methods for survival analysis: A comparison between ranking and regression approaches. Artificial Intelligence in Medicine. 2011 10;53:107-18.

[26] Biganzoli E, Boracchi P, Mariani L, Marubini E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Statistics in medicine. 1998;17:1169-86.

[27] Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018 2;18.

[28] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of Statistical Software. 2011;39:1-13.

[29] Ridgeway G. The state of boosting. Computing science and statistics. 1999:172-81.

[30] Gordon L, Olshen RA. Tree-structured survival analysis. Cancer treatment reports. 1985 10;69:1065-9.

[31] Gilhodes J, Zemmour C, Ajana S, Martinez A, Delord JP, Leconte E, et al. Comparison of variable selection methods for high-dimensional sur-

vival data with competing events. Computers in Biology and Medicine. 2017 12;91:159-67.

[32] Chi S, Tian Y, Wang F, Wang Y, Chen M, Li J. Deep Semisupervised Multitask Learning Model and Its Interpretability for Survival Analysis. IEEE Journal of Biomedical and Health Informatics. 2021 8;25:3185-96.

[33] Dai X, Park JH, Yoo S, D'Imperio N, McMahon BH, Rentsch CT, et al. Survival analysis of localized prostate cancer with deep learning. Scientific Reports. 2022 12;12.

[34] Lee C, Light A, Alaa A, Thurtle D, van der Schaar M, Gnanapragasam VJ. Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database. The Lancet Digital Health. 2021 3;3:e158-65.

[35] Li R, Zhu J, Zhong WD, Jia Z. Comprehensive Evaluation of Machine Learning Models and Gene Expression Signatures for Prostate Cancer Prognosis Using Large Population Cohorts. Cancer Research. 2022 5;82:1832-43.

[36] Meng X, Wang X, Zhang X, Zhang C, Zhang Z, Zhang K, et al. A Novel Attention-Mechanism Based Cox Survival Model by Exploiting Pan-Cancer Empirical Genomic Information. Cells. 2022 5;11.

[37] Nezhad MZ, Sadati N, Yang K, Zhu D. A Deep Active Survival Analysis approach for precision treatment recommendations: Application of prostate cancer. Expert Systems with Applications. 2019 1;115:16-26.

[38] Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ. 2019 3:l886.

[39] Kattan MW, Wheeler TM, Scardino PT. Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. Journal of clinical oncology. 1999;17:1499.

[40] Vince RA, Jiang R, Qi J, Tosoian JJ, Takele R, Feng FY, et al. Impact of Decipher Biopsy testing on clinical outcomes in localized prostate cancer in a prospective statewide collaborative. Prostate Cancer and Prostatic Diseases. 2022 12;25:677-83.

[41] Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. The Lancet Oncology. 2015 4;16:e173-80.

[42] Kattan MW, Vickers AJ, Yu C, Bianco FJ, Cronin AM, Eastham JA, et al. Preoperative and postoperative nomograms incorporating surgeon experience for clinically localized prostate cancer. Cancer: Interdisciplinary International Journal of the American Cancer Society. 2009;115:1005-10.

Chapter 2

A Systematic Review on Machine Learning Techniques for Survival Analysis in Cancer

Abstract

Machine learning (ML) has the potential to overcome limitations of conventional survival analysis, which are commonly found in cancer studies. However, the performance of these methods varies across studies. As a result, it is unclear if ML methods outperform traditional statistical methods and which have the best performance. The present study aimed to review the literature.

This systematic review was conducted using the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines. Electronic databases were systematically searched using keywords related to Machine Learning, Survival Analysis and Cancer.

The review included 196 studies, from which 39 comparable studies were used for the analysis. The predictive performance of the different machine learning methods varied across the cancer types, but overall, multi-task and deep learning methods generally appeared to yield superior performance. However, due to the lack of consistency in reporting across studies and frequent omission of important methodological descriptions and detailed results, these findings are limited to a small sample of the studies investigated.

2.1 Introduction

Cancer is one of the most common diseases worldwide, according to the World Health Organisation (WHO), with one in five people being diagnosed with cancer in their lifetime and more than 19 million new cases of cancer diagnosed each year. One in ten people die from cancer, with approximately 10 million cancer-related deaths each year [1]. Various factors such as genetic predisposition, environmental influences, lifestyle choices, and infections contribute to the development and progression of cancer [2]. Many cancers are preventable and treatable if detected early. Diagnostic tools like the Papanicolaou test (pap smear) or human papillomavirus (HPV) detection test for cervical cancer [3], mammograms for breast cancer [4], and digital rectal examinations (DRE) for prostate cancer [5] are vital for early detection. Other tools such as nomograms can be used to help determine the risk of death or cancer progression by using various factors and help to determine suitable treatment pathways [6].

For this reason, much research has been undertaken to determine accurate survival prediction based on various patient information. From a clinical standpoint, this evaluation could be based on a clinician's medical expertise alone or using a clinical nomogram that combines diagnostic or prognostic tools to aid in the decision-making process [7]. Such models are developed using fine-tuned mathematical and statistical techniques, with time to an event such as death or recurrence often being the outcome of interest. As such, survival analysis is a predominant methodological tool in cancer research. Due to the limitations of the traditional statistical methodologies for survival analysis, such as linearity assumptions and issues pertaining to high dimensionality, machine learning (ML) methods have been developed and adapted

to overcome these issues and improve prediction. There is now a large array of different ML methodologies for survival analysis implemented in cancer research, and while their performance varies, it is not clear whether any, or which ML techniques may yield superior performance, or whether they are all capable of improving on traditional methods. This motivated this systematic review, which aimed to evaluate published studies on machine learning models for survival analysis in cancer.

2.2 Background

2.2.1 Survival Analysis

Survival analysis involves statistical methods to analyse time-to-event data where the outcome variable is a clinical endpoint, such as the time to death or to recurrence of the disease T, with probability density function $f(t)$ [8]. One of the main functions of interest in survival analysis is the survival function, which models the probability that a subject survives beyond time $t > 0$,

$$S(T) = \Pr(T > t)$$

It is a non-increasing function with $S(0) = 1$ and $S(\infty) = 0$. Rather than directly estimating the survival function, the hazard function

$$h(t) = \frac{p(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

is often considered instead, which represents the instantaneous rate at which events occur, given no prior event, with corresponding cumulative hazard function

$$H(t) = \int_0^t h(u)\, du$$

such that $S(t) = exp(-H(t))$.

Direct estimation of the survival function often relies on parametric models, assuming specific distributions for survival times (e.g., exponential, Weibull), which can be restrictive and may not fit the data well [9]. Also, in survival analysis, data are often right-censored, in which case the event of interest (e.g., death) has not occurred by the end of the study period. This makes direct estimation of the survival function more complex, as the exact survival times for censored observations are unknown. A common methodology used for survival analysis that takes censoring into account is the Cox proportional hazards model (CPH) [10]. This semi-parametric approach does not require a parameterised characterisation of the hazard function, which makes it more robust and widely applicable compared to parametric models of $S(t)$. However, the assumption of proportional hazards that it relies upon implies that the hazard ratios between different levels of covariates are constant over time, i.e., the effect of covariates on the hazard rate is multiplicative and does not change. The model may provide biased or misleading results if this assumption is violated. The CPH model also relies on the assumption of a log-linear relationship between the covariates and the hazard function. The model may not fit the data well if the true relationship between covariates and the hazard is non-linear. While this can be partially addressed by transforming covariates or including interaction terms, it still requires careful consideration and validation [11]. The addition of time-varying covariates or stratification can alleviate proportional hazards assumption [12]; however, it remains restricted to low dimensional data. In situations with a large number of covariates relative to the number of events, the CPH model may overfit the data, leading to unstable estimates [13]. Cancer research often involves high-dimensional data, having many more variables (p) than observations (n), leading to potential

overfitting and unstable hazard ratio estimates in the CPH model [14].

2.2.2 Machine Learning

Machine learning methods have the potential to overcome many of the limitations that traditional statistical methods impose [15] by capturing complex, non-linear relationships between covariates and survival outcomes, unlike traditional statistical models [16]. ML techniques are also well-suited for analysing high-dimensional data often encountered in survival analysis, such as genomic data [17], radiology data [18], or electronic health records (EHR) [19]. Several ML strategies have already been adapted for survival analysis and were found throughout this systematic review. These include, but are not limited to; regularisation of the CPH model, approaches based on survival trees, boosting algorithms, support vector machines, shallow neural networks, and deep learning.

Regularisation methods

In order to alleviate the limitations of the CPH model, regularised alternatives have been developed to allow for the use of the model in high-dimensionality settings by adding a penalty term to the likelihood function. The least absolute shrinkage and selection operator (LASSO) adds an L_1 penalty to the CPH model, i.e. penalizes the sum of the absolute values of the coefficients

$$\lambda \sum_{j=1}^{p} \left|\beta_j\right|, \ \lambda > 0.$$

This approach encourages sparsity in the coefficients [20] by automatically selecting important covariates and shrinking other coefficients toward zero, effectively reducing model complexity. Ridge regression is another regularised

approach that instead adds an L_2 penalty to the CPH model, i.e. penalizing the sum of the squared values of the coefficients

$$\lambda \sum_{j=1}^{p} \left|\beta_j^2\right|, \ \lambda > 0.$$

Although this approach also penalizes large coefficients [21], it does not allow shrinkage to make any of the coefficients equal to zero, unlike the LASSO. Elastic Net (EN) combines L_1 and L_2 penalties linearly into

$$\lambda \sum_{j=1}^{p} \left|\beta_j\right| + (1-\lambda) \sum_{j=1}^{p} \left|\beta_j^2\right|, \ \lambda \in (0,1)$$

thus allowing both variable selection and coefficient shrinkage [21]. The regularisation hyperparameter λ thus controls the strength of the penalty applied to the conventional CPH criterion.

Tree-based methods

In tree-based approaches, the survival outcome is predicted by recursively partitioning the data into subgroups [22]. At each split, the covariate that maximizes a chosen separation criterion is selected to make the split, and the same covariate may be considered for more than one split. Commonly implemented criteria to determine child nodes include:

- maximising the log-rank test statistic to yield the greatest difference between survival curve estimates of the two subsets [23];
- maximising the likelihood ratio test statistic for exponential survival times [24];
- maximizing the Cox partial likelihood [24].

Ensembles of survival trees by random survival forests (RSF) [25] are more commonly used than a single tree design. This ensemble relies on bootstrap aggregation, whereby a subset of features is randomly selected at each node to perform the next split. The predicted survival time (or hazard) for each observation is obtained by aggregating predictions from all trees in the forest, which mitigates the risk of overfitting of a single-tree model [26], and also reduces prediction variability.

An alternative tree-based method uses conditional inference tests to determine the best split [27]. Thus, the output of the resulting conditional inference tree is the set of conditional survival probabilities determined by the node memberships that observations fall into. A conditional survival forest (CSF) is an ensemble of conditional trees.

Boosting methods

Boosting is an ensemble technique that combines the predictions from multiple base or weak learners to improve overall predictive performance in a sequential manner. Two main methodologies are implemented for boosting in survival analysis: model-based boosting [28] and likelihood-based boosting [29]. In model-based boosting, after an initial model is derived, the residuals are computed, and a weak learner is fit to the residual. This new learner is scaled by the learning rate parameter and linearly combined to the initial model. Each subsequent learner is fit to the errors (residuals) of the previous model, and the process is repeated for a predefined number of iterations or until convergence. In likelihood-based boosting, the weak learners are fitted by the negative gradient of the log-likelihood with respect to the current model's parameters.

Boosting in survival analysis can follow the traditional tree-based struc-

ture similar to the original boosting methods, or it can be based on another modelling structure, such as the CPH model. Hyperparameters for such models include the learning rate parameter (ν) and either the number of iterations or the stopping criterion. With an increasing number of boosting iterations, performance can be improved but may also lead to overfitting if not combined with regularisation. An early stopping criterion can be included to stop training early if performance on a validation set does not improve for a set number of iterations. A tolerance for the minimum improvement required to continue training can also be implemented to mitigate these common issues.

Support Vector Machines

Support Vector Machines (SVM) for classification involve the creation of hyperplane decision boundaries to separate data into classes. In Support Vector Regression (SVR), the implemented hyperplane is the regression function (e.g., a regression line) [30]. The data points on either side of the hyperplane closest to the hyperplane are called Support Vectors and define the boundary line. SVR consists of fitting the best boundary within an acceptable distance of the hyperplane.

There are two main implementations of survival SVM (SSVM): the regression approach and the ranking approach [31]. The former considers censoring when formulating the inequality constraints of the support vector problem. A linear function that assumes a linear relationship between the covariates and the survival time can be used in the regression approach. Kernel functions can be used to capture non-linear relationships between covariates and survival time. SVMs use kernel functions to map input features into a higher-dimensional space where a hyperplane separates the data non-linearly. Common kernels include linear, polynomial, radial basis function (RBF), and sig-

moid. In the ranking approach, the inequality constraints set the objective to maximize the concordance index for comparable pairs of observations. The objective is to minimise ranking loss, ensuring that the predicted risk scores respect the observed survival order. A hybrid approach can also be implemented, combining the regression and ranking constraints in the same model. The hyperparameters for SSVM include the regularisation parameter, kernel type and kernel parameters.

Neural Networks

Neural networks (NN) for survival analysis leverage the potential of artificial neural network architectures to predict time-to-event outcomes and model complex, non-linear relationships in data. An NN is composed of interconnected nodes (neurons) arranged into an input layer, a number of hidden layers, and an output layer. The input layer can take covariates of any type; some NN architectures (convolutional neural networks, CNN) allow raw image data to be taken as input. Each connection between neurons has an associated weight, and each neuron applies an activation function to its input eventually culminating in a output layer of one or more neurons. The activation function can take one of several forms, the ReLU, sigmoid, or tanh functions being common choices, to introduce non-linearity into the model. Cox-nnet is an example of an artificial neural network (ANN) for survival with a single hidden layer, and produces an output "proportional hazards" layer, comparable to the CPH model [32]. Typical hyperparameters in NNs are the number of hidden layers, the number of neurons in each layer and the activation function used.

Deep learning is a subset of machine learning focused on neural networks with a large number of hidden layers (deep architectures) and are capable of

learning hierarchical features from data. A popular deep feed-forward neural network, DeepSurv [33], is similar in design to the previously described Coxnnet. DeepSurv combines CPH regression with neural networks, training the model so as to optimise the average negative log partial likelihood [33]. In this setting, the final layer is a single node, and the output of the network is taken as the predicted log-risk function. DeepSurv, like many deep learning methods, also has the potential to be implemented in shallow frameworks. An alternative methodology involves direct modelling of the discrete-time hazard function, such as the DeepHit model [34]. The output of DeepHit provides the probabilities of different events occurring at specific discrete time-points.

2.3 Methods

2.3.1 Study Registration

The present study was registered with the International Prospective Register of Systematic Reviews (the full protocol can be found on PROSPERO; Protocol ID: CRD42023391624) and was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [35].

2.3.2 Search Strategy

Electronic databases were searched from inception to 24th February 2023, including PubMed, Scopus, Web of Science, Embase and Association of Computing Machinery (ACM) database. The searching methodology included terms and synonyms relating to machine learning techniques for survival analysis in cancer:

$$(Neoplasm\ OR\ Cancer\ OR\ Carcinoma)\ AND$$
$$(Survival\ Analysis\ OR\ time\ to\ event)\ AND$$
$$(Machine\ Learning\ OR\ Deep\ Learning)$$

Studies of patients with cancer or being assessed for risk of cancer development were eligible for this review. Studies were eligible if they evaluated machine learning methodologies and reported an estimate of either the survival function $S(t) = 1 - P(t)$, or the hazard function $h(t) = \frac{f(t)}{S(t)}$. Papers were excluded if the studies only reported on classification models. Comparisons to traditional statistical methodologies for survival analysis (e.g. Cox proportional hazard regression, parametric techniques, etc.) were analysed if present. The primary outcome was to assessed how machine learning methodologies for survival analysis in cancer compared to traditional statistical methods and thus, studies were included if they implemented ML methods for survival and compared them, analysing the results, with respect to predictive performance, stability, and repeatability, where applicable. For predictive performance, acceptable metrics included the concordance index, area under the curve (from a receiver operating characteristic curve), decision curve analysis, etc. For stability, acceptable metrics included the mean or median number of variables included across bootstrapping or cross-validation, the inclusion of a variable in more than a certain cut-off threshold of bootstraps of cross-validation sets, etc. For repeatability, acceptable frameworks included internal or external validation. Information was synthesised where possible. The secondary outcome was to assessed what machine learning methodology had the greatest predictive performance for survival analysis in cancer, studies were also included without comparison to traditional methods for assessment of this outcome.

The following study designs were included: randomized controlled trials, observational studies (all types), cohort (longitudinal) studies, cohort studies, case-control studies, before-after studies or cross-sectional studies.

2.3.3 Data Extraction

We created a data extraction form for study characteristics and outcome data piloted by three investigators on ten randomly selected studies in the review. Data extraction was conducted by AOD. The following data for study characteristics and outcomes were extracted from each included study: Study authors, year, title, study type, data type, data accessibility, software, replicability, pre-filtering methods applied, additional feature selection methods, sampling framework, cancer/carcinoma type, patient inclusion/exclusion criteria, sample size, interventions and comparators, survival outcome(s) of interest, ML method(s) implemented for survival analysis, reason(s) for the use of ML, type of comparator, traditional method(s) implemented for survival analysis, quantitative comparison of machine learning methodologies for survival analysis in cancer against conventional statistical methods with respect to predictive performance, stability, and repeatability, where applicable. Information was synthesised where possible.

2.3.4 Quality Assessment

Due to the nature of this review, the risk of bias was assessed utilising the Checklist for Assessment of Medical Artificial Intelligence (ChAMAI checklist) [36]. The ChAMAI checklist assessment considers thirty requirements across problem and data understanding, data preparation, modelling and deployment with respect to quality and bias. Each requirement is scored between 0 and 2 with two possible ranges 0, 0.5, 1 and 0, 1, 2, for less important and

more important considerations, respectively. Being focused on a methodological review, the risk of bias was assessed based on the validity of the statistical approach implemented in a paper, the framework it was built upon, and the results garnered.

2.4 Results

In total, 4456 studies were identified from the database searches. After the removal of 1488 duplicates, and another 2309 studies excluded based on their title and abstract, a total of 659 studies were selected for full-text review. Of these, 463 were excluded (full exclusion reasons are broken down and can be seen in the PRISMA flow diagram, Figure 2.1). A total of 196 studies were thus included following full-text review. Table 2.1 provides a general view of the included studies. Due to inherent differences in cancer types, we have reported on each cancer type individually as appropriate. Thirty-nine of the included studies considered more than one cancer type in their analysis, 5 of which considered pan-cancer. The most prevalent single cancer type studied was lung cancer, accounting for 15% of studies and was also included in 23 studies that considered multiple cancer types. All studies were published between 2007 and -2023, with approximately 75% of the publications from 2020 onwards (Figure 2.2).

2.4.1 Data

Across all cancer types, most of the studies were retrospective (188), but eight studies were prospective. The sample size used for analysis varied considerably across the studies, ranging from 17 to over 250,000.

Table 2.1: Extracted characteristics of the included articles.

Cancer Type	Count		Retro-spective		Accessible Data		Available Source Code		Stated time-point of evaluation		Outcome of Interest					
											Multiple Outcomes		Overall Survival		Other	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Overall	196	100.0%	188	95.9%	139	70.9%	59	30.1%	56	28.6%	38	19.4%	146	74.5%	81	41.3%
Bladder	6	3.1%	6	100.0%	6	100.0%	2	33.3%	1	16.7%	4	66.7%	3	50.0%	4	66.7%
Blood	8	4.1%	8	100.0%	5	62.5%	2	25.0%	1	12.5%	2	25.0%	8	100.0%	2	25.0%
Bone	2	1.0%	2	100.0%	1	50.0%	1	50.0%	1	50.0%	0	0.0%	2	100.0%	0	0.0%
Brain	16	8.2%	15	93.8%	8	50.0%	2	12.5%	7	43.8%	4	25.0%	15	93.8%	5	31.3%
Breast	22	11.2%	22	100.0%	15	68.2%	11	50.0%	4	18.2%	3	13.6%	15	68.2%	10	45.5%
Cervical	5	2.6%	4	80.0%	4	80.0%	0	0.0%	2	40.0%	1	20.0%	4	80.0%	2	40.0%
Colorectal	11	5.6%	11	100.0%	8	72.7%	0	0.0%	5	45.5%	2	18.2%	4	36.4%	8	72.7%
Esophageal	2	1.0%	2	100.0%	1	50.0%	0	0.0%	0	0.0%	1	50.0%	1	50.0%	2	100.0%
Gastric	8	4.1%	8	100.0%	4	50.0%	3	37.5%	2	25.0%	1	12.5%	7	87.5%	2	25.0%
Head & Neck	20	10.2%	17	85.0%	13	65.0%	2	10.0%	8	40.0%	6	30.0%	9	45.0%	16	80.0%
Kidney	4	2.0%	4	100.0%	3	75.0%	0	0.0%	1	25.0%	1	25.0%	3	75.0%	1	25.0%
Liver	6	3.1%	6	100.0%	2	33.3%	0	0.0%	4	66.7%	0	0.0%	4	66.7%	2	33.3%
Lung	23	11.7%	21	91.3%	19	82.6%	6	26.1%	10	43.5%	3	13.0%	16	69.6%	10	43.5%
Melanoma	6	3.1%	6	100.0%	5	83.3%	4	66.7%	1	16.7%	1	16.7%	5	83.3%	2	33.3%
Ovarian	2	1.0%	2	100.0%	1	50.0%	1	50.0%	0	0.0%	0	0.0%	1	50.0%	1	50.0%
Pancreatic	3	1.5%	3	100.0%	1	33.3%	0	0.0%	1	33.3%	1	33.3%	3	100.0%	1	33.3%
Prostate	8	4.1%	7	87.5%	6	75.0%	4	50.0%	4	50.0%	2	25.0%	3	37.5%	6	75.0%
Sarcoma	2	1.0%	2	100.0%	0	0.0%	1	50.0%	2	100.0%	1	50.0%	2	100.0%	1	50.0%
Uterine	2	1.0%	2	100.0%	1	50.0%	0	0.0%	1	50.0%	0	0.0%	2	100.0%	0	0.0%
Other	1	0.5%	1	100.0%	0	0.0%	0	0.0%	1	100.0%	1	100.0%	1	100.0%	1	100.0%
MCT	39	19.9%	39	100.0%	36	92.3%	20	51.3%	0	0.0%	4	10.3%	38	97.4%	5	12.8%

MCT: Multiple Cancer Type.

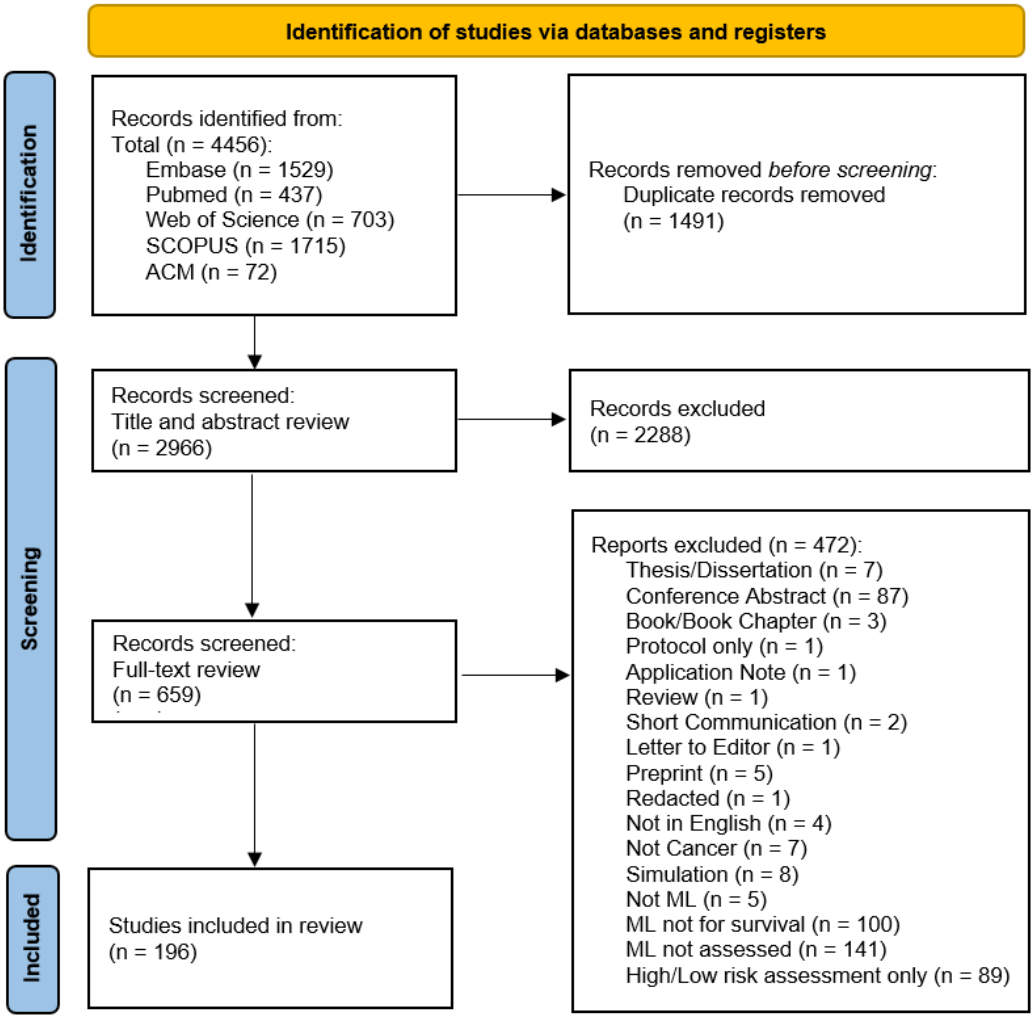

Figure 2.1: PRISMA flowchart of included studies.

Data availability

Data availability is vital for advancing the research field and thus improving clinical practice. As previously discussed, with an increase in the dimensionality of data, it becomes more challenging to produce models without an increase in the sample size. By making data available, more accurate models can be created. Approximately 70% of the studies used publicly available data or would be made available by the authors upon request. Twenty percent of

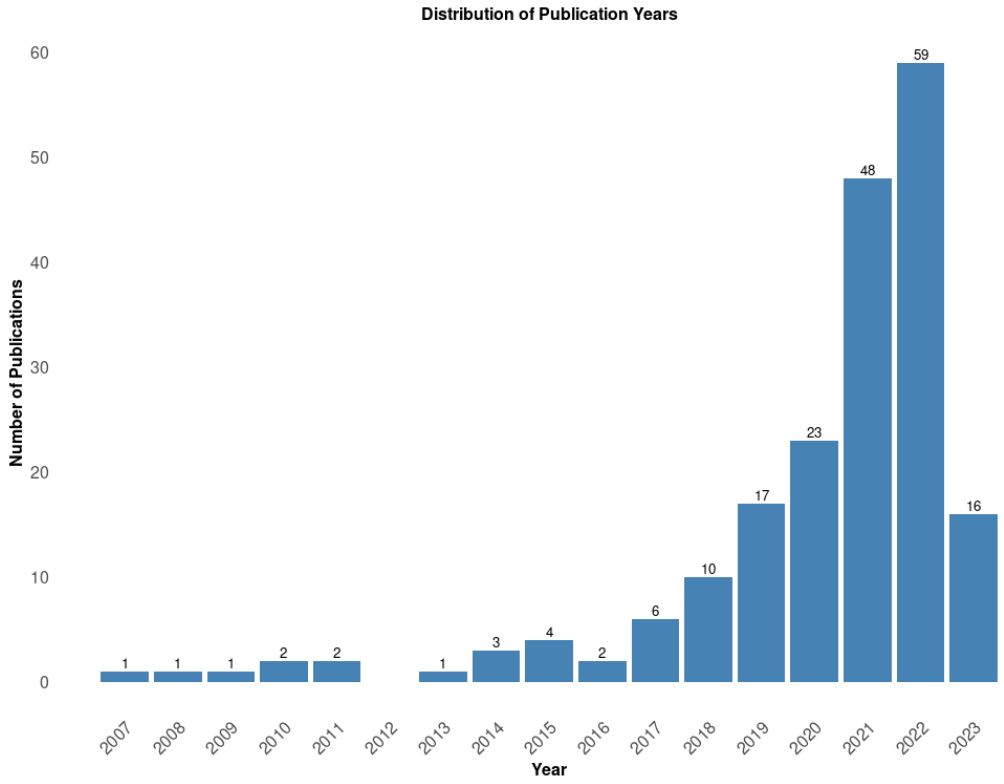

Figure 2.2: Number of papers by publication year.

studies did not make a statement regarding data availability. Another 3% of studies stated that the data would not be available.

Data Sources

Approximately 50% of the studies used datasets from at least one of the following major public databases: the Cancer Genome Atlas (TCGA), National Center for Biotechnology Information's Gene Expression Omnibus (NCBI GEO), the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) database, The Cancer Imaging Archive (TCIA) and The National Cancer Database (NCDB). Thirty percent of the studies used hospital-level data, 7% used other national or global databases as their sole data source,

and 3% used clinical trial data. Ten percent of the studies used datasets available in R and Python from packages such as `survival`, `Bioconductor`, `ipred`, `DeepSurv`, `penalized`, `rotsf`, `biospear` and `RankDeepSurv`. Eight studies did not state where their data was acquired from.

Data Types

A variety of data types were used across the included studies. 50% of the studies included a single data type, while the remaining contained two or more. Clinicopathological variables were the most prevalent data type across the studies; used in 157 studies, 68 alone, and 89 in combination with other data types. Seventy studies included genomic, transcriptomic or proteomic data alone or in combination with each other or other data types. Forty studies included radiomics data, and 23 studies included raw imaging data.

Study Populations

In order to be able to show the benefit of a designed model, it is necessary to understand the study population on which it has been built and, if assessed, the population it has been validated with. Across all cancer types, approximately 50% of studies reported patient-level inclusion and exclusion criteria, 25% of studies provided detail at a cohort-level, and the remaining 25% did not report details of the study population.

Evaluation Time-points

When developing models for survival analysis in cancer, the use case of the produced model should be considered as the variables considered applicable will depend on the planned clinical application of the proposed model. If, for example, the aim of the model is to determine what further treatment, if

any, may be required following surgical intervention, then pathological variables extracted during the surgery can be included in the model. However, if the model aimed to determine if the surgery was likely to be successful in eradicating the cancer, then only variables available pre-operatively should be included. Over 70% of the studies did not state the time-point of evaluation being considered by their models. Of the remaining 30%, the most common evaluation time-points were post-operative and pre-operative.

Outcomes of Interest

Approximately 20% of the studies considered multiple outcomes of interest in their analysis. Seventy-five percent of studies investigated overall survival as an outcome of interest. Forty-one percent of studies considered other outcomes, with progression-free and recurrence-free survival being the most common. Three studies did not state their outcome of interest directly.

Source Code

In an age of reproducible research, code availability is paramount. When both the data and source code are available, accurate replication of results to validate findings becomes possible. Seventy percent of the studies did not state whether the code was available. Of the remaining 30% of studies, the code was publicly available via depositories such as Github. The remaining five studies stated that they would make the code available upon request.

2.4.2 Modelling Frameworks

To better understand the structure of the analysis undertaken in the studies, it was important to consider the modelling framework used. Table 2.2 provides an overview of the frameworks used. The initial element of the pipeline

is the decision to implement pre-filtering of the data prior to the model building. Next is the sampling framework under which the modelling will occur, which can be a simple train/test split for more robust methods such as cross-validation or bootstrapping. Finally, once the model has been built, validation should be considered.

Pre-filtering

As discussed in the introduction, ML is often implemented in cases where $p > n$ due to its greater ability to handle high dimensionality [17]. However, when the number of predictors greatly exceeds the sample size, i.e. when $p >> n$, some form of pre-filtering can be beneficial to improve model performance [15]. Forty-six percent of the studies implemented pre-filtering into their modelling pipeline. The most common methods used included univariate Cox proportional hazard feature selection and correlation-based pre-filtering, implemented by 29 and 18 studies, respectively. Other pre-filter methods included the use of pre-defined or hand-selected feature sets, pathway analysis, variance or expression-based thresholding and clustering or principal component analysis.

Sampling

Across all cancer types, 53% of the studies used a train/test split as their sampling framework, while 27% used a k-fold cross-validation framework. Another eight studies used bootstrapping for their sampling framework. Another nine studies used other sampling methods, such as Monte-Carlo CV and modified versions of the cross-validation framework. More than 10% of studies did not implement any resampling framework or even perform a single data split.

Table 2.2: Modelling frameworks of the included articles.

Cancer Type	Pre-filtering		Sampling Framework										Validation					
			Train/test split		k-fold CV		Boot-strapping		Other		None		Internal Validation		External Validation		None	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Overall	90	46.4%	104	53.1%	53	27.0%	8	4.1%	9	4.6%	22	11.2%	169	86.2%	39	19.9%	11	5.6%
Bladder	3	50.0%	2	33.3%	3	50.0%	0	0.0%	0	0.0%	1	16.7%	5	83.3%	1	16.7%	0	0.0%
Blood	5	62.5%	1	12.5%	2	25.0%	2	25.0%	0	0.0%	3	37.5%	6	75.0%	2	25.0%	0	0.0%
Bone	1	50.0%	1	50.0%	1	50.0%	0	0.0%	0	0.0%	0	0.0%	2	100.0%	1	50.0%	0	0.0%
Brain	9	56.3%	11	68.8%	1	6.3%	0	0.0%	2	12.5%	2	12.5%	14	87.5%	3	18.8%	0	0.0%
Breast	13	59.1%	9	40.9%	7	31.8%	1	4.5%	1	4.5%	4	18.2%	18	81.8%	2	9.1%	4	18.2%
Cervical	3	60.0%	3	60.0%	1	20.0%	0	0.0%	0	0.0%	1	20.0%	4	80.0%	1	20.0%	0	0.0%
Colorectal	6	54.5%	5	45.5%	2	18.2%	1	9.1%	1	9.1%	2	18.2%	8	72.7%	1	9.1%	3	27.3%
Esophageal	1	50.0%	2	100.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	100.0%	1	50.0%	0	0.0%
Gastric	3	37.5%	4	50.0%	2	25.0%	0	0.0%	2	25.0%	0	0.0%	6	75.0%	3	37.5%	0	0.0%
Head & Neck	9	45.0%	9	45.0%	7	35.0%	1	5.0%	1	5.0%	2	10.0%	17	85.0%	7	35.0%	1	5.0%
Kidney	1	25.0%	1	25.0%	2	50.0%	0	0.0%	0	0.0%	1	25.0%	3	75.0%	1	25.0%	1	25.0%
Liver	4	66.7%	2	33.3%	2	33.3%	0	0.0%	1	16.7%	1	16.7%	5	83.3%	2	33.3%	0	0.0%
Lung	12	52.2%	17	73.9%	5	21.7%	1	4.3%	0	0.0%	0	0.0%	22	95.7%	6	26.1%	0	0.0%
Melanoma	3	50.0%	2	33.3%	2	33.3%	0	0.0%	0	0.0%	2	33.3%	4	66.7%	3	50.0%	1	16.7%
Ovarian	1	50.0%	0	0.0%	1	50.0%	1	50.0%	0	0.0%	0	0.0%	2	100.0%	0	0.0%	0	0.0%
Pancreatic	1	33.3%	2	66.7%	1	33.3%	0	0.0%	0	0.0%	0	0.0%	2	66.7%	1	33.3%	0	0.0%
Prostate	1	12.5%	4	50.0%	2	25.0%	0	0.0%	0	0.0%	2	25.0%	6	75.0%	2	25.0%	0	0.0%
Sarcoma	0	0.0%	2	100.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	100.0%	0	0.0%	0	0.0%
Uterine	1	50.0%	2	100.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	100.0%	0	0.0%	0	0.0%
Other	1	100.0%	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%
MCT	12	30.8%	25	64.1%	12	30.8%	0	0.0%	1	2.6%	1	2.6%	38	97.4%	2	5.1%	1	2.6%

MCT: Multiple Cancer Type.

Validation

Due to the nature of the sampling frameworks implemented, the majority (85%) of studies included internal validation of their results. Twenty percent of the studies included external validation, half of which also had internal validation. Eleven studies did not provide any validation of their findings, however, and only reported on the training outcomes.

2.4.3 Models

Comparison with Traditional Models

Approximately 55% of the studies compared ML models to a traditional statistical methodology or a current state-of-the-art model (SoA). Half of these included a single ML model, and the other half carried out a comparison involving multiple modelling approaches. Across all studies, 25% considered a single ML model framework and so also did not make any comparisons to a traditional or SoA method (these were thus not analysed in the primary outcome in this review).

Machine Learning Models

The primary outcome of interest in this study is the comparison of machine learning models to traditional statistical methods for survival analysis. Figure 2.3 shows an overview of the ML methods used across the studies, and Table 2.3 details the breakdown of the implemented modelling strategies. The most prevalent ML modelling method was random survival forest (RSF), which was considered in more than 55% of the studies and as the only ML in 37 studies. Fourteen studies also considered other tree-based methods, such as single survival trees and conditional survival forests (CSF). Neural

network-based methods were the next most prevalent, most of which being deep learning structures (approximately 40% of all the studies), while shallow architectures were considered in 15% of studies. Regularised models were also implemented in more than 35% of the studies. Boosting and SVM structures were considered less frequently (20% and 12% of studies, respectively), and 21 studies considered other modelling methods, such as Bayesian frameworks and multi-task models based on a linear model.

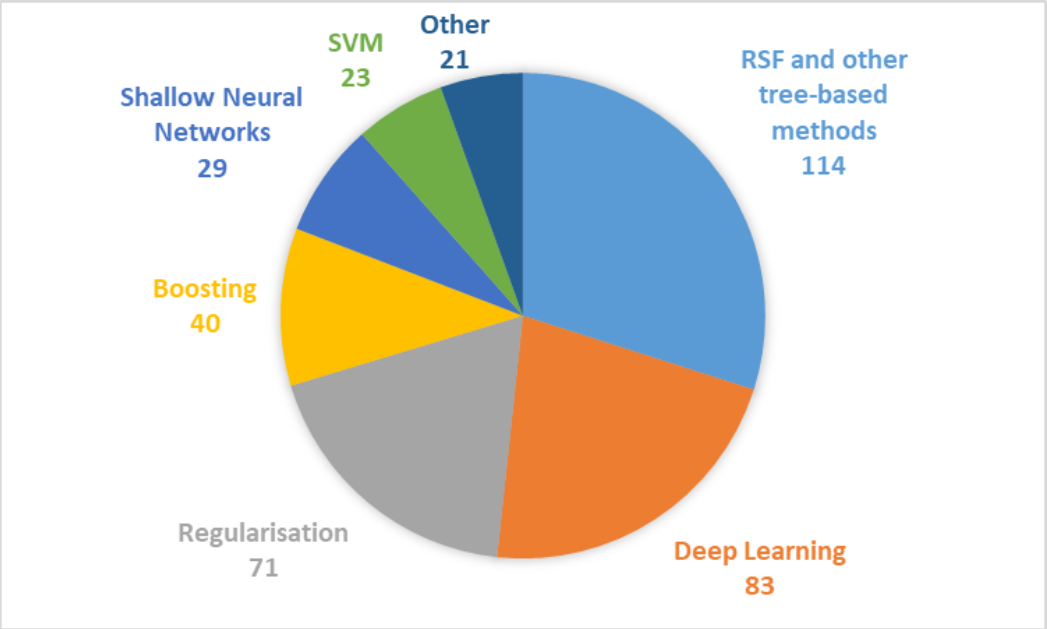

Figure 2.3: Distribution of Machine Learning Methods.

2.4.4 Evaluation

Evaluation metrics were considered under three main headings: discrimination, calibration and predictive performance. These three groups are described below, and Table 2.4 summarises how the most common metrics in each group were used in the literature reviewed by cancer type.

Table 2.3: Models of the included articles.

Cancer Type	Traditional Comparison		Multi ML Models implemented		Machine Learning													
					RSF and other tree based methods		Deep Learning		Regular-isation		Boosting		Shallow Neural Network		SVM		Other	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Overall	104	53.1%	110	56.1%	114	58.2%	83	42.3%	71	36.2%	40	20.4%	29	14.8%	23	11.7%	21	10.7%
Bladder	2	33.3%	5	83.3%	2	33.3%	0	0.0%	3	50.0%	4	66.7%	0	0.0%	0	0.0%	2	33.3%
Blood	4	50.0%	3	37.5%	6	75.0%	0	0.0%	3	37.5%	2	25.0%	0	0.0%	0	0.0%	1	12.5%
Bone	2	100.0%	1	50.0%	2	100.0%	1	50.0%	0	0.0%	0	0.0%	1	50.0%	0	0.0%	0	0.0%
Brain	9	56.3%	4	25.0%	10	62.5%	5	31.3%	3	18.8%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Breast	14	63.6%	17	77.3%	17	77.3%	9	40.9%	6	27.3%	5	22.7%	2	9.1%	4	18.2%	3	13.6%
Cervical	4	80.0%	3	60.0%	4	80.0%	1	20.0%	2	40.0%	1	20.0%	0	0.0%	0	0.0%	1	20.0%
Colorectal	6	54.5%	3	27.3%	8	72.7%	3	27.3%	3	27.3%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Esophageal	1	50.0%	0	0.0%	0	0.0%	1	50.0%	1	50.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Gastric	4	50.0%	4	50.0%	3	37.5%	6	75.0%	1	12.5%	1	12.5%	1	12.5%	1	12.5%	0	0.0%
Head & Neck	7	35.0%	11	55.0%	11	55.0%	9	45.0%	9	45.0%	4	20.0%	5	25.0%	1	5.0%	1	5.0%
Kidney	2	50.0%	2	50.0%	2	50.0%	3	75.0%	1	25.0%	1	25.0%	0	0.0%	0	0.0%	0	0.0%
Liver	2	33.3%	4	66.7%	6	100.0%	3	50.0%	3	50.0%	1	16.7%	0	0.0%	2	33.3%	0	0.0%
Lung	10	43.5%	10	43.5%	9	39.1%	11	47.8%	10	43.5%	4	17.4%	3	13.0%	3	13.0%	0	0.0%
Melanoma	4	66.7%	4	66.7%	4	66.7%	0	0.0%	1	16.7%	3	50.0%	3	50.0%	1	16.7%	0	0.0%
Ovarian	0	0.0%	1	50.0%	2	100.0%	1	50.0%	1	50.0%	1	50.0%	0	0.0%	0	0.0%	1	50.0%
Pancreatic	1	33.3%	1	33.3%	2	66.7%	0	0.0%	0	0.0%	1	33.3%	0	0.0%	1	33.3%	0	0.0%
Prostate	8	100.0%	4	50.0%	4	50.0%	6	75.0%	1	12.5%	1	12.5%	0	0.0%	1	12.5%	2	25.0%
Sarcoma	2	100.0%	1	50.0%	1	50.0%	1	50.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Uterine	2	100.0%	1	50.0%	1	50.0%	1	50.0%	1	50.0%	0	0.0%	0	0.0%	1	50.0%	0	0.0%
Other	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
MCT	20	51.3%	31	79.5%	19	48.7%	22	56.4%	22	56.4%	11	28.2%	14	35.9%	8	20.5%	10	25.6%

MCT: Multiple Cancer Type.

Discrimination

In the context of survival analysis, discrimination metrics refer to measures that assess the ability of a model to distinguish between individuals with different survival times correctly. As we are generally interested in survival across a period of time rather than at one specific time-point, the discrimination metrics used to evaluate the model must allow assessment of the cohort over time and, therefore, commonly rely on an evaluation of either $S(t)$ or $h(t)$ for each subject. The most commonly implemented metric was Harrell's concordance index (C-index) [37] defined as the proportion of all feasible pairs of patients whose predicted and observed outcomes are concordance, was used in over 80% of studies to evaluate their models' discriminatory ability. Approximately 15% of which also consider other discriminatory metrics such as the integrated Brier Score (IBS) [38] which is the average of the Brier score [39] values for an interval of time [40]. Four other studies used IBS as the sole metric for discrimination, and four more used other novel evaluation methods. The remaining 25 studies did not evaluate their models using a discrimination metric applicable to longitudinal survival data.

Calibration & Predictive Performance

Calibration metrics evaluate how well the predicted survival probabilities agree with the actual observed outcomes. While discrimination metrics assess the model's ability to distinguish between different survival times, calibration metrics focus on the accuracy of the predicted probabilities. A quarter of the studies included an evaluation of model calibration; most implemented the Brier Score or calibration curves.

Performance metrics that focus on assessing model performance at specific time-points rather than over the entire duration of follow-up allow for evalu-

ation of how well a model predicts the probability of an event occurring by a particular clinical time-point of interest in the patient follow-up (e.g., at the 3-year survival horizon) [41]. This is an important concept in medical applications, and pre-specified time-points can be determined for each cancer type to consider someone disease-free and inform subsequent follow-up. Other forms of model performance can also be considered, such as classification metrics, including the none time-dependent measures of receiving operator characteristics (ROC) curve analysis to report on the area under the ROC curve (AUC) [8], as a binary metric evaluated only on uncensored observations.

Almost 40% of studies included additional forms of performance evaluation. Most of these used a form of AUC, either in a time-dependent assessment or as a binary metric evaluated only on uncensored observations. Six studies used decision curve analysis (DCA) [42] for net benefit evaluation and specific time-points. The net benefit considered in the DCA is an indication of the increased or decreased number of patients who could see benefit in their treatment if the model was used for evaluation, where a positive net benefit can be viewed as the proportion of patients who would benefit the from use of the model in a clinical decision making setting.

2.4.5 Stability and Feature Selection

Most ML modelling pipelines include pre-filtering, inbuilt dimensionality reduction, and/or other feature selection techniques. The primary goal of these steps is usually either to enhance model-building performance or identify relevant biomarkers (or both). The stability of this feature elimination and selection process should be assessed to evaluate (and ensure) future usability, reliability and interpretability of a given model in clinical settings. The resampling framework, which feature selection also depends upon, can be

Table 2.4: Distribution of metrics used in the 196 articles reviewed by cancer type.

Cancer Type	Discrimination						Calibration		Performance			
	C-Index		Other		None				AUC		Decision Curve Analysis	
	N	%	N	%	N	%	N	%	N	%	N	%
Overall	163	83.2%	34	17.3%	25	12.8%	49	25.0%	61	31.1%	6	3.1%
Bladder	3	50.0%	2	33.3%	2	33.3%	2	33.3%	4	66.7%	0	0.0%
Blood	6	75.0%	2	25.0%	1	12.5%	2	25.0%	2	25.0%	0	0.0%
Bone	1	50.0%	1	50.0%	1	50.0%	1	50.0%	2	100.0%	0	0.0%
Brain	13	81.3%	2	12.5%	2	12.5%	2	12.5%	5	31.3%	1	6.3%
Breast	18	81.8%	5	22.7%	3	13.6%	5	22.7%	6	27.3%	0	0.0%
Cervical	4	80.0%	1	20.0%	1	20.0%	1	20.0%	1	20.0%	0	0.0%
Colorectal	10	90.9%	1	9.1%	1	9.1%	4	36.4%	4	36.4%	0	0.0%
Esophageal	2	100.0%	0	0.0%	0	0.0%	1	50.0%	1	50.0%	0	0.0%
Gastric	8	100.0%	0	0.0%	0	0.0%	2	25.0%	4	50.0%	1	12.5%
Head & Neck	17	85.0%	4	20.0%	3	15.0%	6	30.0%	7	35.0%	0	0.0%
Kidney	4	100.0%	0	0.0%	0	0.0%	0	0.0%	1	25.0%	0	0.0%
Liver	6	100.0%	0	0.0%	0	0.0%	0	0.0%	2	33.3%	1	16.7%
Lung	19	82.6%	3	13.0%	3	13.0%	5	21.7%	6	26.1%	1	4.3%
Melanoma	3	50.0%	0	0.0%	3	50.0%	2	33.3%	2	33.3%	0	0.0%
Ovarian	1	50.0%	1	50.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Pancreatic	3	100.0%	0	0.0%	0	0.0%	0	0.0%	1	33.3%	0	0.0%
Prostate	4	50.0%	1	12.5%	3	37.5%	3	37.5%	1	12.5%	2	25.0%
Sarcoma	0	0.0%	0	0.0%	2	100.0%	1	50.0%	2	100.0%	0	0.0%
Uterine	2	100.0%	1	50.0%	0	0.0%	1	50.0%	0	0.0%	0	0.0%
Other	0	0.0%	1	100.0%	0	0.0%	1	100.0%	1	100.0%	0	0.0%
MCT	39	100.0%	9	23.1%	0	0.0%	10	25.6%	9	23.1%	0	0.0%

MCT: Multiple Cancer Type.

leveraged to assess this aspect of the model by determining selection rates for each of the candidate features [43], or evaluate specific selection performance metrics [44, 45]. Table 2.5 reports on the breakdown of the feature selection and assessment of variables. Less than 50% of the studies considered in this review reported on the stability of their models or provided an assessment of included features.

The nature of the feature selection and/or dimensionality reduction techniques used imposes constraints on their implementation. If the filtering or selection approach is unsupervised, i.e. the outcome variable is not taken into consideration, the filter can be applied to the entire dataset before the implementation of the sampling framework. Considering these methods within the sampling framework or on a separate hold-out set is also possible, but this is not a requirement. When feature selection/filtering is supervised, however, the outcome variable is used to make feature selection/filtering decisions, and such techniques should be applied only on the training samples (or a separate hold-out set for preliminary analysis), i.e. only on data points not used for model testing or validation, to avoid introducing selection bias into the model [46]. Approximately 50% of the 196 studies reviewed implemented one or more forms of feature selection. Sixty-one studies used supervised techniques, and 55 used unsupervised methods, 24 of which used both. Note that the number of studies that implement some form of feature selection does not directly relate to those that use pre-filtering, as some of the supervised techniques include recursive feature elimination methods.

Table 2.5: Distribution of feature selection approaches used in the 196 articles reviewed by cancer type.

			Feature Selection					
Cancer Type	**Variable Assessment**		**Supervised**		**Unsupervised**		**None**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Overall	87	44.4%	61	31.1%	55	28.1%	104	53.1%
Bladder	5	83.3%	1	16.7%	2	33.3%	3	50.0%
Blood	7	87.5%	5	62.5%	2	25.0%	3	37.5%
Bone	2	100.0%	1	50.0%	1	50.0%	1	50.0%
Brain	8	50.0%	6	37.5%	3	18.8%	7	43.8%
Breast	7	31.8%	7	31.8%	11	50.0%	8	36.4%
Cervical	2	40.0%	3	60.0%	1	20.0%	2	40.0%
Colorectal	7	63.6%	6	54.5%	1	9.1%	5	45.5%
Esophageal	1	50.0%	1	50.0%	0	0.0%	1	50.0%
Gastric	4	50.0%	2	25.0%	1	12.5%	5	62.5%
Head & Neck	11	55.0%	9	45.0%	5	25.0%	11	55.0%
Kidney	1	25.0%	0	0.0%	1	25.0%	3	75.0%
Liver	3	50.0%	2	33.3%	4	66.7%	2	33.3%
Lung	4	17.4%	10	43.5%	6	26.1%	11	47.8%
Melanoma	1	16.7%	2	33.3%	3	50.0%	3	50.0%
Ovarian	1	50.0%	0	0.0%	1	50.0%	1	50.0%
Pancreatic	2	66.7%	1	33.3%	0	0.0%	2	66.7%
Prostate	3	37.5%	1	12.5%	0	0.0%	7	87.5%
Sarcoma	0	0.0%	0	0.0%	0	0.0%	2	100.0%
Uterine	1	50.0%	1	50.0%	1	50.0%	0	0.0%
Other	1	100.0%	1	100.0%	1	100.0%	0	0.0%
MCT	16	41.0%	2	5.1%	11	28.2%	27	69.2%

MCT: Multiple Cancer Type.

2.5 Discussion

In this systematic review, we have summarised the outcomes of 196 studies relating to the use of machine learning methods for survival analysis in cancer. The primary outcome of interest is how these ML methods compare to traditional statistical methods. An additional outcome of interest being which ML method has the greatest predictive performance. In order to make this evaluation relevant to the broadest community, we inspected each cancer type individually and then summarised their combined findings. For studies which looked at multiple cancer types, they were included in the analysis of each relevant cancer.

2.5.1 Comparison to traditional methods and predictive performance

Due to the high variability of the study frameworks, we focused the analysis for our primary outcome on a subset of studies that evaluated the C-index of models used for the prediction of overall survival. This consisted of 57 studies, 17 of which considered multiple cancer types. Four of the multiple cancer studies only reported on pan-cancer analysis, so results could not be separated by cancer type. Fourteen further studies did not report the numeric values of the results, only including a graphical representation of the discussion; thus, these were excluded from the comparative analysis. The remaining 39 studies were analysed to evaluate the outcomes of interest. Table 2.6 shows the breakdown of comparisons by cancer type.

Table 2.6: Description and distribution of outcomes discussed in the 39 articles reviewed by cancer type.

				C-Index							
					ML Methods						
Cancer Type	**N**	**Framework**	**Reference**	**CPH**	**RSF**	**Deep Learning**	**Regular-isation**	**Boosting**	**Shallow Neural Network**	**SVM**	**Other**
Bladder	318	CV	Ning, Zhenyuan (2023) [47]	0.566	0.585		0.586			0.598	0.721
	396	CV	Zhao, Zhangxin (2023) [48]	0.549		0.625	0.605				0.679
Blood	116	CV	Gu, Wangrong (2021) [49]	0.644			0.526				0.644
	92	CV	Gu, Wangrong (2021) [49]	0.488			0.684				0.753
	523	CV	Li, Chunyang (2021) [50]	0.640	0.730		0.730				
	129	Train/ Test split	Van Belle, Vanya (2011) [31]	0.690						0.690	
Bone	3145	Train/ Test split	Yan, Lizhao (2022) [15]	0.773	0.803	0.832			0.821		
Brain	502	CV	Ning, Zhenyuan (2023) [47]	0.676	0.728		0.712			0.759	0.784
	502	CV	Zhao, Zhangxin (2023) [48]	0.596		0.685	0.696				0.708
	407	CV	Audureau, Etienne (2018) [51]	0.698	0.701	0.700					

Table 2.6 continued: Description and distribution of outcomes discussed in the 39 articles reviewed by cancer type.

				C-Index							
					ML Methods						
Cancer Type	**N**	**Framework**	**Reference**	**CPH**	**RSF**	**Deep Learning**	**Regular-isation**	**Boosting**	**Shallow Neural Network**	**SVM**	**Other**
Brain	286	Train/Test split	Jiang, Shuai (2021) [52]	0.774		0.715					
	95	Train/Test split	Chen, Haiyan (2021) [53]	0.811	0.847					0.823	
	209	Train/test split	Tang, Bo (2019) [54]	0.506		0.670	0.600	0.598			0.704
	181	Train/Test split	Lu, Yiping (2020) [55]	0.630	0.908						
	260	Train/Test split	Moradmand, Hajar (2021) [56]	0.674	0.668						
Breast	115	CV	Gu, Wangrong (2021) [49]	0.396			0.557				0.697
	78	CV	Gu, Wangrong (2021) [49]	0.603			0.606				0.722
	295	CV	Gu, Wangrong (2021) [49]	0.536			0.771				0.760
	686	CV	Sundrani, Sameer (2021) [57]	0.737				0.747			
	583	CV	Huang, Zhi (2019) [58]	0.653	0.629	0.692			0.73		

Table 2.6 continued: Description and distribution of outcomes discussed in the 39 articles reviewed by cancer type.

				C-Index							
					ML Methods						
Cancer Type	**N**	**Framework**	**Reference**	**CPH**	**RSF**	**Deep Learning**	**Regular-isation**	**Boosting**	**Shallow Neural Network**	**SVM**	**Other**
Breast	456	CV	Fu, Xiaohang (2023) [59]	0.632	0.620	0.748			0.643		
	573	CV	Goli, Shahrbanoo (2016) [60]	0.610						0.610	
	1673	Monte Carlo CV	Bice, Noah (2020) [61]	0.625	0.737	0.749					
	686	Train/ Test split	Van Belle, Vanya (2011) [31]	0.670						0.670	
	154401	Train/ Test split	Chi, Shengqiang (2021) [62]	0.791	0.801	0.809		0.764	0.807		
	1980	Train/ Test split	Katzman, Jared L. (2018) [33]	0.632	0.620	0.654					
	2232	Train/ Test split	Katzman, Jared L. (2018) [33]	0.659	0.648	0.676					
	322348	Train/ Test split	Park, Jung In (2023) [63]	0.828				0.813		0.834	0.804
	1980	Train/ Test split	Jing, Bingzhong (2018) [64]	0.631	0.619	0.661					
	2232	Train/ Test split	Jing, Bingzhong (2018) [64]	0.658	0.648	0.688					

Table 2.6 continued: Description and distribution of outcomes discussed in the 39 articles reviewed by cancer type.

				C-Index							
					ML Methods						
Cancer Type	**N**	**Framework**	**Reference**	**CPH**	**RSF**	**Deep Learning**	**Regular-isation**	**Boosting**	**Shallow Neural Network**	**SVM**	**Other**
Cervical	768	CV	Matsuo, Koji (2018) [65]	0.607	0.600	0.616	0.606	0.606			
	85	Train/ Test split	Carlini, Gianluca (2022) [66]	0.707	0.517						
Colorectal	46	Bootstapping	Karatza, Eleni (2022) [67]	0.606	0.473		0.699				
	888	CV	Sundrani, Sameer (2021) [57]	0.647				0.663			
	11098	Train/ Test split	Chi, Shengqiang (2021) [62]	0.771	0.774	0.782		0.753	0.780		
	128061	Train/ Test split	Tian, Yu (2018) [68]	0.731	0.898						
	49275	Train/ Test split	Yu, Haohui (2022) [69]	0.788		0.824					
Gastric	3505	Train/ Test split	Donghia, Rossella (2023) [70]	0.764	0.773					0.545	
	2935	Train/ Test split	Huang, Tao (2022) [71]	0.693	0.716	0.731		0.709			

Table 2.6 continued: Description and distribution of outcomes discussed in the 39 articles reviewed by cancer type.

				C-Index							
					ML Methods						
Cancer Type	**N**	**Framework**	**Reference**	**CPH**	**RSF**	**Deep Learning**	**Regular-isation**	**Boosting**	**Shallow Neural Network**	**SVM**	**Other**
Gas	1061	Train/ Test split	Hao, Degan (2022) [72]	0.696		0.849					
Kidney	405	CV	Ning, Zhenyuan (2023) [47]	0.594	0.630		0.608			0.632	0.699
	230	CV	Schulz, Stefan (2021) [73]	0.747		0.779					
Liver	333	CV	Ning, Zhenyuan (2023) [47]	0.543	0.595		0.545			0.601	0.719
Lung	403	CV	Ning, Zhenyuan (2023) [47]	0.536	0.557		0.540			0.568	0.708
	438	CV	Ning, Zhenyuan (2023) [47]	0.570	0.579		0.588			0.577	0.695
	86	CV	Gu, Wangrong (2021) [49]	0.432			0.708				0.711
	98	Train/ Test split	Tang, Bo (2019) [54]	0.517		0.673	0.449	0.524			
	109	Train/ Test split	Yang, Bin (2020) [74]	0.729			0.737				

Table 2.6 continued: Description and distribution of outcomes discussed in the 39 articles reviewed by cancer type.

				C-Index							
					ML Methods						
Cancer Type	**N**	**Framework**	**Reference**	**CPH**	**RSF**	**Deep Learning**	**Regular-isation**	**Boosting**	**Shallow Neural Network**	**SVM**	**Other**
Lung	165	Train/ Test split	Hong, Duo (2021) [75]	0.698			0.677				
	139	Train/ Test split	Van Belle, Vanya (2011) [31]	0.680						0.690	
	167	Train/ Test split	Van Belle, Vanya (2011) [31]	0.610						0.620	
	28927	Train/ Test split	Chi, Shengqiang (2021) [62]	0.707	0.694	0728		0.690	0.714		
	1137	Train/ Test split	Wang, Jianyong (2021) [76]	0.561		0.637			0.563		
	2687	Train/ Test split	Oh, Seungwon (2023) [77]	0.747					0.747		
	16608	Train/ Test split	Yang, Linlin (2022) [78]	0.629	0.635	0.665					
Melanoma	367	CV	Ning, Zhenyuan (2023) [47]	0.493	0.515		0.505			0.538	0.653
Ovarian	295	CV	Ning, Zhenyuan (2023) [47]	0.549	0.571		0.555			0.556	0.682

Table 2.6 continued: Description and distribution of outcomes discussed in the 39 articles reviewed by cancer type.

				C-Index							
					ML Methods						
Cancer Type	**N**	**Framework**	**Reference**	**CPH**	**RSF**	**Deep Learning**	**Regular-isation**	**Boosting**	**Shallow Neural Network**	**SVM**	**Other**
Pancreatic	12905	CV	Kim, Hyunsuk (2022) [79]	0.644	0.634					0.623	
Prostate	482	Train/ Test split	Van Belle, Vanya (2011) [31]	0.770						0.760	
	181249	Train/ Test split	Chi, Shengqiang (2021) [62]	0.801	0.802	0.805		0.783	0.806		
Uterine	539	CV	Ning, Zhenyuan (2023) [47]	0.566	0.614		0.603			0.626	0.711
	539	CV	Zhao, Zhangxin (2023) [47]	0.544		0.659	0.597				0.711
	1751	Train/ Test split	Yang, Bin (2020) [74]	0.729			0.737				
	165	Train/ Test split	Mysona, D. P. (2020) [80]	0.730	0.730		0				
	797	Train/ Test split	Qu, Wenjie (2021) [81]	0.726					0.774		
Other	499	CV	Ning, Zhenyuan (2023) [47]	0.577	0.655		0.664			0.705	0.862

Across all cancer types, an improved predictive performance can be seen from machine learning methods, with the exception of cervical cancer. Multi-task-based methods were seen to perform well across many cancer types. Deep learning approaches also performed well, though when compared to the multitask method, they only had improved performance in kidney cancer. The most common deep learning model implemented was DeepSurv, although when multiple deep learning models were implemented in a study, it was generally outperformed by the other approaches. Of note, with both deep learning and multitask and other methods, the novel approaches introduced by the studies usually showed the best predictive performance. This may be suggestive of bias in the study design and could be partly due to publication bias. Most of the novel methods have also not been used in other studies and so their performance should be investigated outside of the original study for a better understanding of their capabilities.

Although the regularised models generally outperformed the traditional CPH, they had the lowest performance of the ML methods. This may be due to the assumptions of linearity and proportionality that the regularised models must adhere to being too restrictive for the data. LASSO regularisation was the most common technique implemented, though when compared to Elastic Net, it was generally outperformed. This may indicate that the L_1 penalisation is overly stringent and potentially reducing too many important coefficients to zero while the more relaxed L_2 penalty element of the elastic net model counteracts this. RSF, boosting methods, and SVM all showed varying predictive capabilities across the cancer types. RSF was the most commonly implemented ML method, with at least one study for each cancer type. It was shown to have the best performance across the blood cancer studies and in pancreatic cancer, but other ML methods outperformed it in

the other cancer types. It showed improved performance compared to the traditional CPH in the majority of studies; however, in breast and cervical cancer, the CPH model generally remained superior. SVM showed particularly good performance in brain cancer studies, but it was comparable to the CPH model in the other cancer types. Although different from CPH, the assumptions of SVM can still be restrictive depending on the method used and the kernel function implemented. Boosting methods were implemented in very few of the studies. It showed particularly good performance in breast cancer and generally outperformed the CPH model across all the cancer types it was considered in, though other ML methods outperformed it. The boosting methods implemented included mainly GBM and XGBoost but Boosted trees, CoxBoost and BoostCI were each implemented in a study also. Thus, some of the variability seen in the performance of the boosted models may be due to the variability in the modelling structures used.

2.5.2 Stability and Modelling Structures

For a model to be used in a clinical setting, it is necessary to understand what variables are being considered so they can be measured. When models are developed in a form of repeat framework such as repeated train/test split, k-fold CV or bootstrapping, different models are created on each iteration, and thus, there is potential that the same variables are not selected every time. Determining the stability of the model, therefore, is important to understanding which variables are being considered consistently. This can be done by inspecting the selection rates across the repetition, or a formal index of similarity could be considered, such as the Jaccard index [82]. Although more than 80 of the studies used a repeated model building framework, only five inspected their models' stability. Three investigated the selection

rates [26, 83, 84]. One uses the Jaccard Index [85], and one uses interclass correlation to assess the stability [86]. Evaluation of model stability should be carefully considered when evaluating the performance of the models, as without stability, it is not possible to understand what variables have true clinical relevance. Particularly in the case of genomic-based variables, as was discussed in detail by Venet et al. 2011, [87], the significance of a variable in a single model does not necessarily mean it is of clinical importance and should be investigated further.

2.5.3 Bias and Quality Assessment

Some particular issues of bias were noted across the studies concerning statistical modelling related to validation and feature selection. As previously mentioned, 11 studies did not implement any form of internal or external validation. Thus, the results from these studies are from the model building and are not able to give a bearing on the models' ability to perform on unseen data. Another issue found in several studies was the use of supervised feature selection techniques, such as univariate CPH, on the entire dataset prior to splitting or CV. This was seen in nine studies, and a further 13 studies did not clearly report on when the supervised feature selection took place; thus, there is potential that these may also have taken place outside the sampling framework. This method of pre-filtering will result in inflated validation results as the included variables are already known to be predictive of survival in the testing dataset. Other quality assessment issues relate to the clinical usefulness of the models built in the studies. As per the ChAMAI checklist, the description of the study population and analysis of net benefit are key requirements when reporting on machine learning in medical settings. Over 25% of the studies did not report on the study population demographics or their

inclusion/exclusion criteria. Without knowledge of the included subjects, the viability in other populations becomes more challenging to determine, and a potential lack of model predictive ability may be due to an unreported underlying characteristic of the study population. Also, only six studies reported on net benefit or decision curve analysis. These metrics can directly report on the clinical usefulness of implementing a model into routine practice, and without them, it is difficult to show the true potential of a novel model.

2.6 Systematic Review - Prostate Cancer

Eleven studies from the systematic review investigated prostate cancer. A detailed review of these papers was undertaken to answer the research questions of this thesis, so they are included in this section, although they did not feature in the manuscript submitted to the journal.

2.6.1 Prostate Cancer - Results

The eleven papers were published between 2011 and 2022, with the majority of the studies being published since 2021. Only one study was prospective [88] with the other ten being retrospective [19,30,31,62,89–94]. More than half of the studies only considered clinicopathological variables [30, 31, 62, 89, 90, 92, 93]. Two studies investigated the use of genomic or transcriptomic material in conjunction with clinical variables [91] or alone [19]. Table 2.7 provides a general view of these 11 prostate cancer studies. More detailed characteristics of these can be found in Table 2.8.

Table 2.7: Description and distribution of outcomes discussed in the 39 articles reviewed, by cancer type.

Characteristics	Categories	Number of Studies	References(s)
Cancer Type	Prostate	8	[19, 30, 88–90, 92–94]
	MCT	3	[31, 62, 91]
Data Type	Clinicopathological	10	[30, 31, 62, 88–94]
	Genomics	1	[19]
	Radiomics	2	[88, 94]
	Transcriptomics	1	[91]
Data Source	GEO	1	[19]
	Hospitals	2	[88, 94]
	Package	1	[31]
	SEER	3	[62, 90, 92]
	TCGA	2	[19, 91]
	Other database	1	[19]
	Not stated	2	[30, 93]
Data Availability	Public	6	[19, 31, 62, 90–92]
	Available upon request	3	[30, 88, 94]
	Not available	1	[89]
	Not stated	1	[93]
Outcome of Interest	Overall Survival	6	[19, 31, 62, 89, 91, 92]
	Progression-free Survival	1	[88]
	Recurrence-free Survival	2	[30, 94]
	Cancer-specific Survival	1	[90]
	Other	4	[30, 62, 89, 93]
Time-point of Evaluation	Pre-treatment	1	[88]
	Pre-operative	2	[90, 94]
	Post-operative	1	[30]
	Not stated	7	[19, 31, 62, 89, 91–93]
Sampling Framework	Single Train/Test split	3	[62, 89, 90]
	Repeated train/test split	3	[31, 88, 92]
	CV	3	[19, 91, 93]
	None	2	[30, 94]
Pre-filtering and Feature Selection	Unsupervised	3	[62, 91, 93]
	Supervised	1	[94]
	None	7	[19, 30, 31, 88–90, 92]
Number of Models	One	4	[30, 88, 93, 94]
	Two or more	7	[19, 31, 62, 89–92]
Model Types	Boosting	2	[62, 89]
	Neural Network	7	[62, 88–90, 92–94]
	Regularisation	1	[19]
	Trees	5	[19, 62, 89, 90, 92]
	SVM	2	[30, 31]

Table 2.7 continued: Description and distribution of outcomes discussed in the 39 articles reviewed, by cancer type.

Characteristics	Categories	Number of Studies	References(s)
Traditional Model	Yes	10	[19, 30, 31, 62, 88–90, 92–94]
Comparison	No	1	[91]
Evaluation Metrics	C-Index	8	[19, 30, 31, 62, 91–94]
	Integrated Brier Score	1	[62]
	Calibration Curve	2	[90, 94]
	Brier Score	2	[89, 90]
	tC-Index	3	[88–90]
	AUC	1	[94]
	Decision Curve Analysis	2	[90, 94]
	Other	2	[30, 94]
Validation	Internal	9	[19, 31, 62, 88–93]
	External	2	[30, 94]

Data

In order to show the benefit of a designed model, it is necessary to understand the study population on which it has been built and, if assessed, the population used for validation. In prostate cancer, four studies reported patient-level criteria [88, 90, 92, 94], three provided cohort-level information [19, 89, 93], and four did not report details of the study population [30, 31, 62, 91].

The majority of the studies used publicly available data [19, 31, 62, 90–92] or stated the data would be made available upon request [30, 88, 94]. The remaining two studies either did not make a statement regarding the data availability [93] or stated that the data would not be made available [89].

Two studies included genomic or transcriptomic data, and both used the Cancer Genome Atlas (TCGA) [19, 91], one of which also used data from the National Center for Biotechnology Information's Gene Expression Omnibus (NCBI GEO) and other publicly available database sources [19].

Table 2.8: Extracted characteristics of the included Prostate Cancer articles.

Cancer	Author, Year	Data Source	Data Type(s)	Sample Size	Survival	Framework	Pre-Filtering	Feature Selection	ML	Comparison	Validation	Evaluation
Prostate	Dai et al., 2022 [89]	DB	Clinico-pathologic	112276	Overall Survival, Composite Outcome	Train/Test split	None	None	DSM, GBM, RDSM, RSF	CPH	Internal	BS, tC-Index
Prostate	Khan, 2018 [30]	Not stated	Clinico-pathologic	Data 1 = 262, Data 2 = 342, Data 3 = 340	Recurrence-free Survival, Other	None	None	None	SVRc	CPH	External	C-Index, Novel Criterion
Prostate	Lee et al., 2021 [90]	SEER	Clinico-pathologic	171942	Cancer-specific Survival	Train/Test split	None	None	CSF, DeepHit, RSF, Survival Quilts	SoA	Internal	tC-Index, DCA, Calibration Curve, BS
Prostate	Lee et al., 2022 [88]	Hospital-level	Clinico-pathologic, Radiomics	585	Progression-free Survival	Repeated Train/Test split	None	None	DDHL	CPH	Internal	tC-Index

BS: Brier Score,

Table 2.8 Continued: Extracted characteristics of the included Prostate Cancer articles.

Cancer	Author, Year	Data Source	Data Type(s)	Sample Size	Survival	Framework	Pre-Filtering	Feature Selection	ML	Comparison	Validation	Evaluation
Prostate	Li et al., 2022 [19]	TCGA, GEO, Other database	Genomics	1558	Overall Survival	CV	None	None	Cox PLS, LASSO Cox, Ridge Cox, RSF, SuperPC	CPH	Internal	C-Index
Prostate	Nezhad et al., 2019 [92]	SEER	Clinico-pathologic	approx. 6000	Overall Survival	Repeated Train/Test split	None	None	DASA, RSF	CPH	Internal	C-Index
Prostate	Trinh et al., 2022 [93]	Not stated	Clinico-pathologic	115	Other	CV	Hand-selected	None	DeepSurv	CPH	Internal	C-Index
Prostate	Yan et al., 2021 [94]	Hospital-level	Clinico-pathologic, Radiomics	368	Recurrence-free Survival	None	None	Univariate CPH	DNN for survival	CPH, SoA	External	C-Index, Calibration Curve, Hosmer-Lemeshow Calibration test, AUC, DCA

BS: Brier Score, IBS: Integrated Brier Score

Table 2.8 Continued: Extracted characteristics of the included Prostate Cancer articles.

Cancer	Author, Year	Data Source	Data Type(s)	Sample Size	Survival	Framework	Pre-Filtering	Feature Selection	ML	Comparison	Validation	Evaluation
Prostate*	Chi et al., 2021 [62]	SEER	Clinico-pathologic	181249	Overall Survival, Competing Risks	Train/Test split	Hand-selected	None	DeepHit, DeepSurv, GBM, Logistic-Hazard, MTLR, PMF, RSF, SSMTL	CPH, Parametric, Fine-Gray	Internal	C-Index, IBS
Prostate*	Meng et al., 2022 [91]	TCGA	Clinico-pathologic, Transcriptomics	477	Overall Survival	CV	Expression thresholding, Correlation, PCA, Other	None	Coxnnet,LASSO Cox, Ridge Cox, SAVAE-Cox, VAECox	None	Internal	C-Index
Prostate*	Van Belle et al., 2011 [31]	Package	Clinico-pathologic	482	Overall Survival	Repeated Train/Test split	None	None	SSVM	CPH	Internal	C-Index

BS: Brier Score, *: MCT study, IBS: Integrated Brier Score

Seven of the studies used other large databases, hospital-level data or those available in R packages [31, 62, 88–90, 92, 94] while two studies did not state the source of their data [30, 93].

Five of the studies made their code available via GitHub or a similar depository [19, 62, 88, 90, 91] or stated that the code would be made available upon request [89]. The other studies did not make a statement regarding code availability [30, 31, 92–94].

Outcome of Interest

Six prostate cancer studies considered overall survival as the outcome of interest [19, 31, 62, 89, 91, 92]. Two studies were interested in recurrence-free survival [30, 94], one of which also considered clinical failure (CF), including metastasis and/or death from cancer [30]. The remaining three studies considered progression-free survival [88], cancer-specific survival [90] and the prediction of urinary continence after robot-assisted radical prostatectomy (RARP) [93], respectively.

Time-point of evaluation

A particular limitation of seven of the studies was their lack of statement regarding the time-point of evaluation being considered by their models [19, 31, 62, 89, 91–93]. The remaining four studies investigated the pre-treatment [88, 90, 94] and post-operative [30] time-points.

Sampling Framework

Six of the studies used a train/test split [31, 62, 88–90, 92] of which three did so with repetition [31, 88, 92]. Another three studies used a cross-validation

framework [19,91,93]. The final two studies did not use a sampling framework in their analysis; however, both used an external dataset for validation [30,94].

Pre-filtering

More than half of the prostate cancer studies did not implement pre-filtering or other feature selection techniques [19, 30, 31, 88–90, 92]. Two studies hand-selected variables for inclusion in their modelling frameworks [62, 93] and one study used a selection of pre-filtering techniques, including expression thresholding, correlation-based filtering, principal component analysis (PCA) mutual information and maximal information coefficient (MIC) methods [91]. One final study implemented a supervised feature selection method using univariate CPH [94].

ML Methods and Traditional Models

Only one study did not compare modelling performance to a traditional methodology [91]. Traditional CPH was considered in most of the studies [19, 30, 31, 62, 88, 89, 92–94] and a final study considered state of the art (SoA) models as a comparison [90]. Each of the eleven prostate cancer studies considered a different selection of ML methods. Four considered a single ML model, including regression-based SVM [30], DeepSurv [93], the novel Dynamic-DeepHit-Lite (DDHL) [88] and a novel DNN for survival (DSNN) [94]. Seven considered more than one ML model [19, 31, 62, 89–92]. Overall, the ML techniques spanned a wide range of base methodologies, including regularisation, boosting, tree-based methods and neural networks. Further details on the model type are described in the discussion section (2.6.2).

Discrimination

The majority of studies used discrimination metrics capable of considering survival across a period of time. Eight implemented the concordance index (C-index) for evaluation [19, 30, 31, 62, 91–94], one of which also considered the integrated Brier Score (IBS) [62] and another used a novel criterion which incorporated the C-index, sensitivity and specificity into one value [30]. The other three studies also used a variation of the C-index, evaluated at specific time-points (tC-index); thus, these are not considered discrimination metrics in this study [88–90].

Calibration

Several of the prostate cancer studies investigated the calibration of their models, two of which used the Brier Score (BS) for the evaluation [62,89], one used calibration curves and the Hosmer-Lemeshow calibration test [94] and one used both calibration curves and BS [90]. The other seven studies did not evaluate calibration [19, 30, 31, 88, 91–93].

Predictive Performance Metrics

Four prostate cancer studies included additional forms of performance evaluation. Three of the studies are those previously mentioned to have used tC-index. One also considered the BS [89], and the other evaluated BS and decision curve analysis (DCA) [90]. The other study, which considered additional evaluation metrics, used a form of ROC Curve and AUC (Area Under the Curve) along with DCA [94].

Variable selection and model stability

None of the prostate cancer studies considered the stability of their feature selection. Both studies without a sampling framework reported the number of variables considered for model building, and one of these also reported on the weight of features in the model [30]. Of the studies that implemented a single train/test split, all three reported the number of variables considered, but only one reported their importance [62]. Of the studies which implemented a repeated train/test split or CV, three reported the selected variables but did not report on their importance [19, 31, 88], one reported the selected variables and reported the feature-rank [93], one reported the number of included variables and inspected the hidden layer nodes that contributed the most to the prognosis according to the mean and variance [91]. One provided no information regarding the selected features [92].

Comparison of traditional and ML methods

As was reported in the main body of the systematic review, only Chi et al., 2021 [62] and Van Belle et al., 2011 [31] compared the traditional CPH to ML models evaluating the C-index for the prediction of overall survival and report the numeric values of the results.

Chi et al., 2021

Chi et al., 2021 [62] compared novel semisupervised multi-task learning (SSMTL) to traditional CPH, accelerated failure time (AFT) as well as several ML methods RSF, GBM, DeepSurv, DeepHit, multi-task logistic regression (MTLR), a survival neural network (LogisticHazard) and the probability mass function (PMF) method using SEER data from breast, colorectal, lung and prostate cancer datasets, respectively. Competing risk methods were compared to

the SSMTL, traditional Fine-Gray, RSF, DeepHit, and its simplified version, SimpleMLP. In prostate cancer, the SSMTL model performed the best in the competing risk, and PMF performed the best in the non-competing risk models. On inspection of the performance values given in this study, we noted that the confidence intervals were unusually narrow, suggesting incorrect standard error normalisation and that a potential risk of bias may be present due to calculation error.

Van Belle et al., 2011

Van Belle et al., 2011 [31] compared multiple SSVM methods on eight datasets available in R package `survival` and the StatLib database (five low dimensional datasets: the veteran's administration lung cancer trial (VLC), leukaemia (LCR, LD), the Mayo Clinic lung cancer data (MLC), prostatic cancer (PC), (the German breast cancer study) (GBSG) (BC), and three high dimensional datasets: Norway/Stanford breast cancer data (NSBCD), Dutch Breast Cancer Data set (DBCD), and the diffuse large B-cell lymphoma data (DLBCL)). Four SSVM methods were used, including one regression-based method (SVCR), two ranking-based methods (RANKSVMC and MODEL 1), and one hybrid method (MODEL 2). They were compared to a traditional CPH model in the low-dimensional datasets. In `PC`, the CPH model had the highest C-index. CPH had the highest C-index (C-Index +/- SD: 0.77 +/- 0.03) but the difference was not statistically significant between it and any of the SSVM models; MODEL 1 (0.75 +/- 0.04), MODEL 2 (0.76 +/- 0.03), RANKSVNC (0.75 +/- 0.03), and SVCR (0.76 +/- 0.03).

The two studies made opposite conclusions regarding the comparative performance of traditional and ML methods, and as they did not use the same ML models, it was not possible to consolidate their findings. As the outcome of interest in this thesis research was BCR-free survival, the two studies interested in recurrence-free survival (Khan, 2018 [30]; Yan et al.,2021 [94]) were considered.

Khan (2018)

Khan (2018) [30] compared regression-based SSVM with CPH, both of which were considered with and without transduction, in which survival analysis algorithms are modified by estimating appropriate target times in a semi-supervised regression context. Using a novel criterion that combines sensitivity, specificity, and the C-index, they found that transduction improved the predictive performance when used with a CPH or SSVM model. The CPH models outperformed the SSVM models for RFS, and the results varied when metastasis-free and overall survival were considered.

Yan et al., 2021

Yan et al., 2021 [94] compared a deep survival neural network (DSNN) to traditional CPH using data from hospitals in Beijing. The CPH models were each based on a SoA method for biochemical recurrence (BCR) prediction (Gleason grade group (GG-RP), National Comprehensive Cancer Network (NCCN), and the Cancer of the Prostate Risk Assessment (CAPRA/CAPRA-S)) while the DSNN was based on radiomics. The DSNN model outperformed all SoA models. For both validation cohorts (VC), the DSNN outperformed the CPH models. In VC1, all CPH models had a mean C-Index less than 0.7, GG-RP (C-Index (CI): 0.564 (0.419,0.709)), CAPRA-S (0.654, (0.490,

0.818)), NCCN (0.535, (0.408, 0.662)), and CAPRA (0.552, (0.385, 0.719)) for which the deep radiomic model (DRS-BCR) had a C-Index 0.811 (0.722, 0.900). In VC2 DRS-BCR model had a C-Index of 0.794 (0.718, 0.870) while the CPH models all have C-index of less than 0.7; GG-RP (C-Index (CI): 0.689 (0.601, 0.777)), CAPRA-S (0.654, (0544, 0.764)), NCCN (0.583, (0.498, 0.668)), and CAPRA (0.614, (0.509, 0.719)).

As with the studies of overall survival, the two RFS studies made opposite conclusions regarding comparative performance, with the SVM-based method again being inferior. Also, as they did not consider the same sets of ML models, it was not possible to consolidate their findings. Furthermore, as neither study implemented a resampling framework or even a simple train/test split, the results quoted are likely biased.

2.6.2 Prostate Cancer - Discussion

Quality Assessment

Most of the studies had moderate quality scores based on the ChAMAI checklist, with Khan, 2018 [30] having the lowest at 33.3% and Yan et al., 2021 [94] having the highest at 63.3%. Although the ChAMAI checklist is rigorous in its evaluation, other specific considerations were also noted as they inform the consideration of the results. Table 2.9 reports on the score of the ChAMAI checklist, represented as a percentage, with the higher value being of higher quality and a lower risk of bias. Specific considerations with respect to the quality of the studies are also presented in this table.

Stating the data source and the provision of source code can aid in the reproducibility of a study. Khan, 2018 and Trinh et al., 2022 [30, 93] did

Table 2.9: Quality assessment of Prostate Cancer studies.

	Chi et al., 2021 [62]	Dai et al., 2022 [89]	Khan, 2018 [30]	Lee et al., 2021 [90]	Lee et al., 2022 [88]	Li et al., 2022 [19]	Meng et al., 2022 [91]	Nezhad et al., 2019 [92]	Trinh et al., 2022 [93]	Van Belle et al., 2011 [31]	Yan et al., 2021 [94]
ChAMAI Score	50.0%	50.0%	33.3%	60.5%	52.2%	54.8%	42.9%	39.8%	44.6%	42.9%	63.3%
Data source stated	✓	✓	X	✓	✓	✓	✓	✓	X	✓	✓
Source code availability stated	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	✓
Sample size reported	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓
Number of features reported	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	X
Feature importance reported	X	X	✓	X	X	X	X	X	✓	X	✓
Modelling framework clearly described	✓	✓	✓	X	✓	✓	✓	X	✓	✓	✓
Hyper-parameter tuning described	✓	X	X	X	X	X	✓	X	X	X	X
Hyper-parameter values stated	X	X	X	X	X	X	✓	X	✓	X	X
Resampling framework implemented	X	X	X	X	✓	✓	✓	✓	✓	✓	X
Discrimination metric evaluated	✓	X	✓	X	X	✓	✓	✓	✓	✓	✓
Variability indicator for the evaluation metric calculated	✓	✓	X	✓	✓	✓	✓	X	✓	✓	X

not clearly state their data source, and Trinh et al., 2022 also did not make a statement regarding the data availability. Also, in Lee et al., 2021 and Nezhad et al., 2019 [90,92], the modelling frameworks were slightly ambiguous, making it more challenging to get a complete understanding of their structure. Meng et al., 2022 [91] describe their method for hyperparameter tuning and the selected values of these parameters, but this was not common across the other studies. In Nezhad et al., 2019 [92] it was not entirely clear what the final sample used was, which may bring into question the power of the study's findings.

The number of features to be considered for inclusion in a model, as well as reporting on those actually included, is important to understanding the dimensionality of the data set and what factors are being considered. This leads to the concept of feature importance and understanding how the features included are contributing to the model. In Chi et al., 2021 and Yan et al., 2021 [62, 94], the number of features was not reported at any point in the model-building process. While only Khan, 2018, Trinh et al., 2022 and Yan et al., 2021 [30, 93, 94] reported on feature importance.

The use of a resampling framework aids in producing robust results and so is preferable for validation purposes. More than half of the studies did not use a resampling framework [30, 62, 89, 90, 94] of which Khan, 2018 and Yan et al., 2021 [30, 94] did not even implement a single train/test split.

As previously described, the use of an appropriate metric for discrimination is of paramount importance for assessing the survival model. Dai et al., 2022; Lee et al., 2021; Lee et al., 2022 [88–90] did not report on an appropriate metric. To make an inference on a reported evaluation metric, it is necessary to report on a measure of variability or spread, such as a standard error of confidence interval. Khan, 2018; Nezhad et al., 2019; Yan et al., 2021 [30, 92, 94]

did not provide a measure of spread for their evaluation metrics.

2.7 Conclusion

In considering the quality assessment and risk of bias, it is clear that consistency is required in reporting and an improved understanding of the requirements to be met for the ChAMAI checklist. In doing so, the clinical usefulness of the model can be determined, the modelling framework will be sound, and it allows for a thorough comparison of methods being used to allow an improved understanding of the benefits of ML methods for survival analysis in cancer in the future. The use of machine learning methods for survival analysis has been increasing rapidly in recent years. In cancer research, the popularity of these methods is due to their ability to handle high dimensionality and non-linear effects, which allows the use of larger and more complex data. This systematic review aimed to determine which machine learning methods were being implemented for survival analysis in cancer studies and, from this, garner insight as to whether machine learning had the ability to outperform traditional models. A second objective of this review was to identify which methods, if any, showed the best performance.

Due to a lack of consistency in reporting and study designs, gaining such insight across all studies is difficult. Studies did not consistently compare the machine learning methods to traditional models or always make comparisons between several machine learning models, limiting the number of studies possible to draw conclusions from.

Direct comparisons between machine learning models were also prohibited by the large variations in outcomes of interest and performance metrics used and reported on, again limiting the number of comparable studies. In the sub-

set of 39 studies which quoted the C-index for models concerned with overall survival, it was found that machine learning methods could outperform the Cox proportional hazard model in most cancer types. The predictive performance of the different machine learning methods varied across the cancer types, but overall, multi-task and deep learning methods generally appeared to yield superior performance.

With respect to the PCa studies, in both the overall survival and RFS, when the traditional CPH model was compared to SVM models, CPH either outperformed or was comparable to the ML method [30, 31]. For all other comparisons, at least one of the ML models outperformed CPH [62, 94]. As only two studies could be considered for comparison for each outcome of interest, and given the lack of overlap in the sets of machine learning models considered, it was not possible to draw further conclusions from this analysis. Furthermore, as with the full systematic review results, the lack of consistency in the reporting and evaluating of results and in the frameworks and methods implemented limit the ability to infer any substantial findings in a more general sense about the use of ML methodologies for survival analysis in PCa, at the time of writing.

2.8 References

[1] Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer statistics for the year 2020: An overview. International Journal of Cancer. 2021 8;149:778-89.

[2] Anand P, Kunnumakara AB, Sundaram C, Harikumar KB, Tharakan ST, Lai OS, et al.. Cancer is a preventable disease that requires major lifestyle changes. Springer New York LLC; 2008.

[3] Mayrand MH, Duarte-Franco E, Rodrigues I, Walter SD, Hanley J, Ferenczy A, et al.. Human Papillomavirus DNA versus Papanicolaou Screening Tests for Cervical Cancer; 2007. Available from: `www.nejm.org`.

[4] Njor S, Nystrom L, Moss S, Paci E, Broeders M, Segnan N, et al. Breast cancer mortality in mammographic screening in Europe: A review of incidence-based mortality studies. Journal of Medical Screening. 2012 9;19:33-41.

[5] Smith DS, Catalona WJ. Interexaminer variability of digital rectal examination in detecting prostate cancer*. Urology. 1995;45:70-4.

[6] Thangue NBL, Kerr DJ. Predictive biomarkers: a paradigm shift towards personalized cancer medicine. Nature Reviews Clinical Oncology. 2011 10;8:587-96.

[7] Simmons CPL, McMillan DC, McWilliams K, Sande TA, Fearon KC, Tuck S, et al.. Prognostic Tools in Patients With Advanced Cancer: A Systematic Review. Elsevier Inc.; 2017.

[8] Cox TF. Medical Statistics for Cancer Studies. Chapman and Hall/CRC; 2022.

[9] Crowther MJ, Lambert PC. A general framework for parametric survival analysis. Statistics in Medicine. 2014 12;33:5280-97.

[10] Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34:187-202.

[11] Mahmud M, Abrahamowicz M, Leffondré K, Chaubey YP. Selecting the Optimal Transformation of a Continuous Covariate in Cox's Regres-

sion: Implications for Hypothesis Testing. Communications in Statistics - Simulation and Computation. 2006 1;35:27-45.

[12] Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME, Groothuis-Oudshoorn CG. Time-varying covariates and coefficients in Cox regression models. Annals of translational medicine. 2018;6(7).

[13] Wang H, Shen Z, Tan Z, Zhang Z, Li G. Fast Lasso-type safe screening for Fine-Gray competing risks model with ultrahigh dimensional covariates. Statistics in Medicine. 2022 10;41:4941-60.

[14] Ramirez R, Chiu YC, Zhang SY, Ramirez J, Chen Y, Huang Y, et al. Prediction and interpretation of cancer survival using graph convolution neural networks. Methods. 2021 8;192:120-30.

[15] Yan L, Gao N, Ai F, Zhao Y, Kang Y, Chen J, et al. Deep learning models for predicting the survival of patients with chondrosarcoma based on a surveillance, epidemiology, and end results analysis. Frontiers in Oncology. 2022 8;12.

[16] Chi SQ, Tian Y, Li J, Tong DY, Kong XX, Poston G, et al. Time-dependent and nonlinear effects of prognostic factors in nonmetastatic colorectal cancer. Cancer Medicine. 2017 8;6:1882-92.

[17] Kalakoti Y, Yadav S, Sundar D. Survcnn: A discrete time-to-event cancer survival estimation framework using image representations of omics data. Cancers. 2021 7;13.

[18] Lombardo E, Kurz C, Marschner S, Avanzo M, Gagliardi V, Fanetti G, et al. Distant metastasis time to event analysis with CNNs in independent head and neck cancer cohorts. Scientific Reports. 2021 12;11.

[19] Li Y, Brendel M, Wu N, Ge W, Zhang H, Rietschel P, et al. Machine learning models for identifying predictors of clinical outcomes with first-line immune checkpoint inhibitor therapy in advanced non-small cell lung cancer. Scientific Reports. 2022 12;12.

[20] Tibshirani R. The lasso method for variable selection in the cox model. Statistics in Medicine. 1997 2;16:385-95.

[21] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of Statistical Software. 2011;39:1-13.

[22] Gordon L, Olshen RA. Tree-structured survival analysis. Cancer treatment reports. 1985 10;69:1065-9.

[23] Ciampi A, Thiffault J, Nakache JP, Asselain B. Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. Computational Statistics Data Analysis. 1986 10;4:185-204.

[24] H C, S H, A MSC, Chang. In: J G, Allan D, B JVKMI, Umphrey, editors. Recursive Partition: A Versatile Method for Exploratory-Data Analysis in Biostatistics. Springer Netherlands; 1987. p. 23-50. Available from: `https://doi.org/10.1007/978-94-009-4794-8_2`.

[25] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Annals of Applied Statistics. 2008 9;2:841-60.

[26] Gilhodes J, Zemmour C, Ajana S, Martinez A, Delord JP, Leconte E, et al. Comparison of variable selection methods for high-dimensional survival data with competing events. Computers in Biology and Medicine. 2017 12;91:159-67.

[27] Pang H, Datta D, Zhao H. Pathway analysis using random forests with bivariate node-split for survival outcomes. Bioinformatics. 2010;26:250-8.

[28] Bühlmann P, Hothorn T. Boosting algorithms: Regularization, prediction and model fitting. Statistical Science. 2007 11;22:477-505.

[29] Binder H, Schumacher M. Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. BMC Bioinformatics. 2009 1;10.

[30] Khan FM. Predicting Advanced Prostate Cancer from Modeling Early Indications in Biopsy and Prostatectomy Samples via Transductive Semi-Supervised Survival Analysis. BioMed Research International. 2018;2018.

[31] Belle VV, Pelckmans K, Huffel SV, Suykens JAK. Support vector methods for survival analysis: A comparison between ranking and regression approaches. Artificial Intelligence in Medicine. 2011 10;53:107-18.

[32] Ching T, Zhu X, Garmire LX. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. PLoS Computational Biology. 2018 4;14.

[33] Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018 2;18.

[34] Lee C, Zame W, Yoon J, der Schaar MV. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. Proceedings of the AAAI Conference on Artificial Intelligence. 2018 4;32.

[35] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. International Journal of Surgery. 2021 4;88.

[36] Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. Elsevier Ireland Ltd; 2021.

[37] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543-6.

[38] Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine. 1999;18(17-18):2529-45.

[39] Brier GW. Verification of forecasts expressed in terms of probability. Monthly weather review. 1950;78(1):1-3.

[40] Park SY, Park JE, Kim H, Park SH. Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). Korean Journal of Radiology. 2021;22(10):1697.

[41] Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of-year predicted risks. Biostatistics. 2019;20(2):347-57.

[42] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making. 2006;26(6):565-74.

[43] Haury AC, Gestraud P, Vert JP. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PloS one. 2011;6(12):e28210.

[44] Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences. 2022;34(4):1060-73.

[45] Davis CA, Gerick F, Hintermair V, Friedel CC, Fundel K, Küffner R, et al. Reliable gene signatures for microarray classification: assessment of stability and performance. Bioinformatics. 2006;22(19):2356-63.

[46] Gareth J, Daniela W, Trevor H, Robert T. An introduction to statistical learning: with applications in R. Spinger; 2013.

[47] Ning Z, Zhao Z, Feng Q, Chen W, Xiao Q, Zhang Y. Mutual-Assistance Learning for Standalone Mono-Modality Survival Analysis of Human Cancers. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023 6;45:7577-94.

[48] Zhao Z, Feng Q, Zhang Y, Ning Z. Adaptive risk-aware sharable and individual subspace learning for cancer survival analysis with multi-modality data. Briefings in Bioinformatics. 2023 1;24.

[49] Gu W, Zhang Z, Xie X, He Y. An Improved Muti-Task Learning Algorithm for Analyzing Cancer Survival Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021 3;18:500-11.

[50] Li C, Patil V, Rasmussen KM, Yong C, Chien HC, Morreall D, et al. Predicting survival in veterans with follicular lymphoma using structured electronic health record information and machine learning. International Journal of Environmental Research and Public Health. 2021 3;18:1-19.

[51] Audureau E, Chivet A, Ursu R, Corns R, Metellus P, Noel G, et al. Prognostic factors for survival in adult patients with recurrent glioblastoma:

a decision-tree-based model. Journal of Neuro-Oncology. 2018 2;136:565-76.

[52] Jiang S, Zanazzi GJ, Hassanpour S. Predicting prognosis and IDH mutation status for patients with lower-grade gliomas using whole slide images. Scientific Reports. 2021 12;11.

[53] Chen H, Li C, Zheng L, Lu W, Li Y, Wei Q. A machine learning-based survival prediction model of high grade glioma by integration of clinical and dose-volume histogram parameters. Cancer Medicine. 2021 4;10:2774-86.

[54] Tang B, Li A, Li B, Wang M. CapSurv: Capsule Network for Survival Analysis With Whole Slide Pathological Images. IEEE Access. 2019;7:26022-30.

[55] Lu Y, Patel M, Natarajan K, Ughratdar I, Sanghera P, Jena R, et al. Machine learning-based radiomic, clinical and semantic feature analysis for predicting overall survival and MGMT promoter methylation status in patients with glioblastoma. Magnetic Resonance Imaging. 2020 12;74:161-70.

[56] Moradmand H, Aghamiri SMR, Ghaderi R, Emami H. The role of deep learning-based survival model in improving survival prediction of patients with glioblastoma. Cancer Medicine. 2021 10;10:7048-59.

[57] Sundrani S, Lu J. Computing the Hazard Ratios Associated With Explanatory Variables Using Machine Learning Models of Survival Data; 2021. Available from: `https://doi.org/10.`

[58] Huang Z, Johnson TS, Han Z, Helm B, Cao S, Zhang C, et al. Deep learning-based cancer survival prognosis from RNA-seq data: Approaches and evaluations. BMC Medical Genomics. 2020 4;13.

[59] Fu X, Patrick E, Yang JYH, Feng DD, Kim J. Deep multimodal graph-based network for survival prediction from highly multiplexed images and patient variables. Computers in Biology and Medicine. 2023;154:106576.

[60] Goli S, Mahjub H, Faradmal J, Mashayekhi H, Soltanian AR. Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression. Computational and Mathematical Methods in Medicine. 2016;2016.

[61] Bice N, Kirby N, Bahr T, Rasmussen K, Saenz D, Wagner T, et al. Deep learning-based survival analysis for brain metastasis patients with the national cancer database. Journal of Applied Clinical Medical Physics. 2020 9;21:187-92.

[62] Chi S, Tian Y, Wang F, Wang Y, Chen M, Li J. Deep Semisupervised Multitask Learning Model and Its Interpretability for Survival Analysis. IEEE Journal of Biomedical and Health Informatics. 2021 8;25:3185-96.

[63] Park JI, Bozkurt S, Park JW, Lee S. Evaluation of race/ethnicity-specific survival machine learning models for Hispanic and Black patients with breast cancer. BMJ Health and Care Informatics. 2023 1;30.

[64] Jing B, Zhang T, Wang Z, Jin Y, Liu K, Qiu W, et al. A deep survival analysis method based on ranking. Artificial Intelligence in Medicine. 2019 7;98:1-9.

[65] Matsuo K, Purushotham S, Jiang B, Mandelbaum RS, Takiuchi T, Liu Y, et al. Survival outcome prediction in cervical cancer: Cox models vs

deep-learning model. American Journal of Obstetrics and Gynecology. 2019 4;220:381.e1-381.e14.

[66] Carlini G, Curti N, Strolin S, Giampieri E, Sala C, Dall'olio D, et al. Prediction of Overall Survival in Cervical Cancer Patients Using PET/CT Radiomic Features. Applied Sciences (Switzerland). 2022 6;12.

[67] Karatza E, Papachristos A, Sivolapenko GB, Gonzalez D. Machine learning-guided covariate selection for time-to-event models developed from a small sample of real-world patients receiving bevacizumab treatment. CPT: Pharmacometrics and Systems Pharmacology. 2022 10;11:1328-40.

[68] Tian Y, Li J, Zhou T, Tong D, Chi S, Kong X, et al. Spatially varying effects of predictors for the survival prediction of nonmetastatic colorectal Cancer. BMC Cancer. 2018 11;18.

[69] Yu H, Huang T, Feng B, Lyu J. Deep-learning model for predicting the survival of rectal adenocarcinoma patients based on a surveillance, epidemiology, and end results analysis. BMC Cancer. 2022 12;22.

[70] Donghia R, Guerra V, Pesole PL, Liso M. Contribution of macro- and micronutrients intake to gastrointestinal cancer mortality in the ONCONUT cohort: Classical vs. modern approaches. Frontiers in Nutrition. 2023 1;10.

[71] Huang T, Huang L, Yang R, Li S, He N, Feng A, et al. Machine learning models for predicting survival in patients with ampullary adenocarcinoma. Asia-Pacific Journal of Oncology Nursing. 2022 12;9.

[72] Hao D, Li Q, Feng QX, Qi L, Liu XS, Arefan D, et al. SurvivalCNN: A deep learning-based method for gastric cancer survival prediction us-

ing radiological imaging data and clinicopathological variables. Artificial Intelligence in Medicine. 2022 12;134.

[73] Schulz S, Woerl AC, Jungmann F, Glasner C, Stenzel P, Strobl S, et al. Multimodal Deep Learning for Prognosis Prediction in Renal Cancer. Frontiers in Oncology. 2021 11;11.

[74] Yang B, Ji H, Zhong J, Ma L, Zhong J, Dong H, et al. Value of 18F-FDG PET/CT-Based Radiomics Nomogram to Predict Survival Outcomes and Guide Personalized Targeted Therapy in Lung Adenocarcinoma With EGFR Mutations. Frontiers in Oncology. 2020 11;10.

[75] Hong D, Zhang L, Xu K, Wan X, Guo Y. Prognostic Value of Pre-Treatment CT Radiomics and Clinical Factors for the Overall Survival of Advanced (IIIB–IV) Lung Adenocarcinoma Patients. Frontiers in Oncology. 2021 5;11.

[76] Wang J, Chen N, Guo J, Xu X, Liu L, Yi Z. SurvNet: A Novel Deep Neural Network for Lung Cancer Survival Analysis With Missing Values. Frontiers in Oncology. 2021 1;10.

[77] Oh S, Kang SR, Oh IJ, Kim MS. Deep learning model integrating positron emission tomography and clinical data for prognosis prediction in non-small cell lung cancer patients. BMC Bioinformatics. 2023 12;24.

[78] Yang L, Fan X, Qin W, Xu Y, Zou B, Fan B, et al. A novel deep learning prognostic system improves survival predictions for stage III non-small cell lung cancer. Cancer Medicine. 2022 11;11:4246-55.

[79] Kim H, Park T, Jang J, Lee S. Comparison of survival prediction models for pancreatic cancer: Cox model versus machine learning models. Genomics and Informatics. 2022 6;20.

[80] Mysona DP, Tran LKH, Tran PMH, Gehrig PA, Le LV, Ghamande S, et al. Clinical calculator predictive of chemotherapy benefit in stage 1A uterine papillary serous cancers. Gynecologic Oncology. 2020 1;156:77-84.

[81] Qu W, Liu Q, Jiao X, Zhang T, Wang B, Li N, et al. Development and Validation of a Personalized Survival Prediction Model for Uterine Adenosarcoma: A Population-Based Deep Learning Study. Frontiers in Oncology. 2021 2;10.

[82] Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: A review. Journal of King Saud University - Computer and Information Sciences. 2022 4;34:1060-73.

[83] Chen Z, Wei Q. Developing an Improved Survival Prediction Model for Disease Prognosis. Biomolecules. 2022 11;12:1751.

[84] Gilhodes J, Dalenc F, Gal J, Zemmour C, Leconte E, Boher JM, et al. Comparison of Variable Selection Methods for Time-to-Event Data in High-Dimensional Settings. Computational and Mathematical Methods in Medicine. 2020 7;2020:1-13.

[85] Schlosser P, Knaus J, Schmutz M, Dohner K, Plass C, Bullinger L, et al. Netboost: Boosting-Supported Network Analysis Improves High-Dimensional Omics Prediction in Acute Myeloid Leukemia and Huntington's Disease. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021 11;18:2635-48.

[86] Adeoye J, Hui L, Koohi-Moghadam M, Tan JY, Choi SW, Thomson P. Comparison of time-to-event machine learning models in predicting oral cavity cancer prognosis. International Journal of Medical Informatics. 2022 1;157:104635.

[87] Venet D, Dumont JE, Detours V. Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLoS Computational Biology. 2011 10;7:e1002240.

[88] Lee C, Light A, Saveliev ES, van der Schaar M, Gnanapragasam VJ. Developing machine learning algorithms for dynamic estimation of progression during active surveillance for prostate cancer. npj Digital Medicine. 2022 12;5.

[89] Dai X, Park JH, Yoo S, D'Imperio N, McMahon BH, Rentsch CT, et al. Survival analysis of localized prostate cancer with deep learning. Scientific Reports. 2022 12;12.

[90] Lee C, Light A, Alaa A, Thurtle D, van der Schaar M, Gnanapragasam VJ. Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database. The Lancet Digital Health. 2021 3;3:e158-65.

[91] Meng X, Wang X, Zhang X, Zhang C, Zhang Z, Zhang K, et al. A Novel Attention-Mechanism Based Cox Survival Model by Exploiting Pan-Cancer Empirical Genomic Information. Cells. 2022 5;11.

[92] Nezhad MZ, Sadati N, Yang K, Zhu D. A Deep Active Survival Analysis approach for precision treatment recommendations: Application of prostate cancer. Expert Systems with Applications. 2019 1;115:16-26.

[93] Trinh L, Mingo S, Vanstrum EB, Sanford DI, Aastha, Ma R, et al. Survival Analysis Using Surgeon Skill Metrics and Patient Factors to Predict Urinary Continence Recovery After Robot-assisted Radical Prostatectomy. European Urology Focus. 2022 3;8:623-30.

[94] Yan Y, Shao L, Liu Z, He W, Yang G, Liu J, et al. Deep learning with quantitative features of magnetic resonance images to predict biochemical recurrence of radical prostatectomy: A multi-center study. Cancers. 2021 6;13.

Chapter 3

Improving the Post-Operative Prediction of BCR-Free Survival Time with mRNA Variables and Machine Learning

Abstract

Predicting the risk of, and time to biochemical recurrence (BCR) in prostate cancer patients post-operatively is critical in patient treatment decision pathways following surgical intervention. This study aimed to investigate the predictive potential of mRNA information to improve upon reference nomograms and clinical-only models, using a dataset of 187 patients that includes over 20,000 features. Several machine learning methodologies were implemented for the analysis of censored patient follow-up information with such high-dimensional genomic data. Our findings demonstrated the potential of inclusion of mRNA information for BCR-free survival prediction. A random survival forest pipeline was found to achieve high predictive performance with respect to discrimination, calibration, and net benefit. Two mRNA variables, namely ESM1 and DHAH8, were identified as consistently strong predictors with this dataset.

3.1 Introduction

Radical prostatectomy (RP) is the primary treatment for prostate cancer [1]. Approximately 30% of patients undergo RP [2], of which up to 70% remain disease-free after 10 years [3], while 20–40% fail and experience biochemical recurrence (BCR) [4]. In the latter case, the time from surgical intervention until BCR varies across patients, with over 60% of BCR cases occurring in the first two years post-operatively [5]. The probability of BCR-free survival is a widely investigated topic as it plays a role in patient treatment decision making [6, 7]. The probability of recurrence can be used to determine if and when adjunctive therapies, such as radiotherapy, hormone therapy, and chemotherapy, should be initiated [8]. Current state-of-the-art nomograms such as that produced by the Memorial Sloan Kettering Cancer Centre (MSKCC) use routine clinical variables to make the prediction [9].

Originally developed by Kattan et al. in 1999, a post-operative nomogram using prostate specific antigen (PSA) at diagnosis, Gleason sum score, prostatic capsular invasion, surgical marginal status, seminal vesicle invasion, and lymph node status are used to make prediction of BCR at seven years [10] with updates on this in 2005 and 2009 [11, 12]. Other models such as the Walz et al., 2009, are designed for prediction at 2-years post-operatively [5]. Two potential limitations of these reference nomograms are the modelling structure they are based on and the nature of the variables used. Nomograms can predict only for the probability of an event occurring at a specific point in time; the original Kattan nomogram [10], for example, considers a 7-year post-RP horizon. This limits the ability to make treatment decisions at different time-points. Both nomograms also consider only clinical variables, although further information may become routinely available for prediction.

This study considered models that allowed for less stringent assumptions on the variable effects, and the inclusion of a number of genetic variables, to overcome these limitations.

The Cox proportional hazard model [13] is a statistical approach of reference for the analysis of censored survival data. Two key assumptions of this model that may be unrealistic are that the independent risk predictors interact linearly in quantifying the relative risk of patients, and that all hazards are proportional within the cohort. Any continuous predictor variable with a non-linear relationship to the risk would need to be transformed to fit the model. However, this would require prior knowledge of the distribution of all variables to be considered for inclusion [14]. Furthermore, such traditional regression models are prone to overfitting, in that the addition of (possibly redundant) variables to the model can lead to an improved fit whilst reducing the generalisability of the model to unseen patients [14, 15]. Other important limitations of Cox models in the context of modern, high-dimensional clinical datasets, are their inability to adjust for collinearity (i.e., redundancy) in the feature set, and their non-applicability to datasets containing more candidate predictors than observations [15, 16].

Machine learning methodologies offer solutions to overcome these limitations. Model regularisation can be applied to traditional models including Cox models [17] to handle collinearity and reduce the potential of overfitting, yielding the likes of the LASSO Cox and Cox ridge regression models [15]. Gradient boosting may also be applied to Cox modelling in order to mitigate the risk of overfitting, and so, is naturally applicable in high-dimensional environments [18, 19]. Other alternative models include the random survival forest, a non-parametric, tree-based strategy that does not impose linearity constraints and thus allows the inclusion of covariates with associations that

are not linear in nature [20].

Improvements may be achieved also in terms of the nature of the information used in current clinical nomograms, since these use routine clinical variables only. Over the past 15 years, the utility of gene expression biomarkers has been widely researched and demonstrated [21–25]. The assessment of mRNA variables has become more cost-reasonable and feasible, offering new opportunities to improve on current models. Currently, mRNA data typically becomes available post-operatively when clinicians have large biological samples of diseased tissue to extract the genetic component from. The primary goal of this study is, firstly, to determine whether the inclusion of mRNA variables can lead to an improvement in the predictive performance of models determining time-to-BCR, and secondly, to identify relevant statistical or machine learning methodologies to perform this prediction.

3.2 Materials and Methods

3.2.1 Data

The dataset used for this analysis was developed in a study by Taylor et al., and contains a total of 232 patients [25] and is comprised of 2 parts, i.e., a set of routine clinical variables and another of messenger ribonucleic acid (mRNA) variables. Following pre-processing for duplicates and missing information, a final 187 patients with complete information including BCR-free duration were retained. BCR in this dataset, is defined as PSA $\geq$ 0.2 ng/mL on two occasions post-operatively [25]. Any records of a Gleason score of 5 were upgraded to a score of 6 per the current grading system [26]. Of these 187 remaining patients, 133 had mRNA information, and so any mRNA analysis was carried out on this sub-cohort. Although the primary

and secondary Gleason scores were available for all remaining individuals, they were not considered for use in our model pipelines as their combined information is already captured in the combined Gleason sum score. Pre-operative variables such as clinical stage and biopsy Gleason score were also excluded as these are seen as approximations of the pathological variables available for analysis in the post-operative setting. As extracapsular extension (ECE), seminal vesicle invasion (SVI), and lymph node involvement (LNI) were included in the modelling, the pathological stage was also removed as the combined effect of these variables should account for all of its effect [27]. All of the aforementioned variable exclusions are in line with the method of the original Kattan nomogram and the current, state-of-the-art MSKCC nomogram [9, 10].

The final set of routine clinical variables consisted of both continuous and categorical features. The continuous variables were PSA (mean = 10.4, standard deviation (SD) = 16.13) and age (mean = 58.1, SD = 6.67) at diagnosis and the categorical variables considered for inclusion in a model can be seen in Table 3.1 along with their descriptive statistics. The set of mRNA variables was not pre-processed and was comprised of over 20,000 features.

3.2.2 Statistical Analysis

Four distinct predictive modelling strategies were considered: the traditional Cox proportional hazard (CPH) model [13]; an L_1-penalised regularisation of this model named LASSO Cox [28]; a gradient boosting approach consisting of a sequential ensemble of CPH models, named Boosted Cox [29]; and a random survival forest (RSF) [20], as an alternative to Cox-based modelling of the risk of BCR. Each modelling approach was implemented with either routine

Table 3.1: Descriptive statistics of the categorical clinical variables.

Variable	N	%
BCR-free Survival		
Censored	135	72
BCR	52	28
Ethnicity		
White	147	79
Black	30	16
Asian	4	2
Unknown	6	3
Pathological Gleason Score		
6	52	28
7	103	55
8	17	9
9	15	8
Extracapsular Extension (ECE)		
Absent	52	28
Present	135	72
Seminal Vesicle Invasion (SVI)		
Negative	156	83
Positive	31	17
Lymph Node Involvement (LNI)		
Normal	140	75
Abnormal	19	10
Not Done	28	15
Surgical Marginal Status (SMS)		
Negative	137	73
Positive	50	27

clinical variables only, or with a combination of clinical and mRNA variables.

Before inclusion, the set of mRNA variables was reduced using pre-filtering and feature selection techniques outlined hereafter, applied either separately or combined, prior to model building. A correlation-based pre-filter was considered in order to remove variables with a high Pearson correlation with

other variables from the dataset in order to prevent multicollinearity, using an absolute cut-off value of 0.6. This cut-off was selected as it had previously been shown to be optimal by Goh et al. on similar data [30]. The feature selection technique consisted of performing univariate Cox PH analysis on each variable separately and ordering the output with respect to their adjusted log-rank score, whilst also adjusting all p-values for false discovery rate (FDR) [31]. The 50 variables with an FDR-adjusted p-value less than 5% and yielding highest log-rank scores were retained for inclusion for multivariate modelling. This method was implemented in line with that outlined by Beer et al. on their assessment of gene expression and the prediction of lung cancer [32]. Such multiple testing methods can be prone to high false positive rates [33], and FDR p-value correction is a popular method used to mitigate this risk. All traditional CPH models underwent further feature selection by means of a forward stepwise selection with a maximum of 10 variables chosen for the models with mRNA variables. All RSF models underwent recursive feature elimination [34] with the selected number of variables being evaluated based on the model yielding the highest concordance index (C-index), which is defined as the proportion of all usable pairs of patients whose predicted and observed outcomes are concordant [35, 36].

R (version 2022.07) [37] was used for all statistical and machine learning analyses. The regularisation parameter (λ) was selected using cross-validation for the LASSO Cox models [38], using a final value for λ corresponding to one standard deviation of the cross-validated error, as a standard approach to achieve parsimony [34]. Hyperparameters for the boosted Cox and RSF models were kept to default settings [29, 39]. Thus, boosted Cox models contained 100 boosting iterations and a fixed shrinkage parameter of 0.1. The RSF models were built using 500 trees, considering $\sqrt{P}$ variables at each

split (where P denotes the number of features considered for modelling) and with a minimum size of 15 for the terminal nodes.

3.2.3 Evaluation Framework

The models were assessed and bench-marked via bootstrapping, using 100 bootstrap resamples for model building and the corresponding out-of-bag (OOB, i.e., test) points for independent testing. Thus, training samples of size N = 133, sampled from the original dataset with replacement, were used. Those observations not used in the training sample were used in the independent test samples, each of size n = 49. The predictive performance of the models was assessed in terms of discrimination and calibration. In this study, discrimination refers to the model's ability to correctly rank patients relative to each other, using a bootstrap-corrected c-index and corresponding confidence intervals (CI), in order to obtain unbiased estimates of external predictive discrimination [40]. Calibration analysis provides a way to assess model bias for various risk profiles, by comparing the observed and predicted probabilities [41]. Here, two approaches to calibration were implemented within the bootstrapping framework. The method of Harrell et al. compares quintiles of the predicted survival of patients at a specific time-point to their Kaplan–Meier estimates [40]. Two- and five-year BCR-free survival endpoints were considered in keeping with calibration of the reference MSKCC nomogram. The second method involves the comparison of quintiles of the predicted overall BCR-free survival probability with their observed Kaplan–Meier estimates [42]. This second method allows for inspection of the model's ability to predict across all time-points, as opposed to the 2-year or 5-year survival horizons.

Receiver operating characteristic (ROC) curve and decision curve analy-

sis (DCA) were also assessed at both 2-year and 5-year horizons to further assess the predictive performance. ROC analysis was summarised in terms of the corresponding bootstrapped area under the ROC curve (AUC) values calculated on the OOB points [36, 43]. DCA [44] provided complementary assessment in terms of the clinical value and potential net benefit to patients of the models' predictions.

In complement to predictive performance, the feature selection stability of the best-performing modelling pipelines across the 100 bootstraps was assessed in terms of the overall feature selection rates, and measured using the Jaccard, Davis, and Dice indices [45, 46]. These indices range between 0 and 1, a higher index value indicating a more stable model [47]. Further, individual feature selection rates were analysed to extract information about potentially relevant predictors for BCR.

3.3 Results

The results from the evaluation of the predictive pipelines described in Section 3.2.2 using the frameworks outlined in Section 3.2.3 are reported in this section.

3.3.1 Discrimination

The models' discriminatory performance based on the bootstrapped C-indices are shown in Table 3.2 along with their corresponding confidence intervals. The boxplots of best the performing pipelines are shown in Figure 3.1.

Pairwise two-sided Wilcoxon signed-rank tests of the null hypothesis of no difference in the C-index distributions reported in Table 3.2 were carried out on all pairs of the models, in order to compare the discriminative per-

Table 3.2: Bootstrapped C-indices. The top performing pipelines for each statistical model (namely CPH, LASSO Cox, Boosted Cox, and RSF) are highlighted in bold.

Model	**C-Index**	**95% Confidence Interval**
MSKCC	0.770	(0.675, 0.844)
Clinical Variables Only		
Cox	0.798	(0.724, 0.857)
LASSO	0.761	(0.710, 0.824)
Boosted	0.770	(0.709, 0.829)
RSF	0.795	(0.727, 0.852)
Clinical Variable and Correlation Prefiltered mRNA variables		
Cox	**0.828**	**(0.720, 0.909)**
LASSO	0.762	(0.696, 0.826)
Boosted	**0.816**	**(0.757, 0.866)**
RSF	0.749	(0.695, 0.811)
Clinical Variables and Univariate Cox Feature Selected mRNA variables		
Cox	0.765	(0.688, 0.827)
LASSO	**0.768**	**(0.703, 0.835)**
Boosted	0.798	(0.737, 0.845)
RSF	0.782	(0.730, 0.846)
Clinical Variables and Correlation Prefiltered and Univariate Cox Feature Selected mRNA variables		
Cox	0.792	(0.671, 0.865)
LASSO	0.763	(0.464, 0.918)
Boosted	0.814	(0.753, 0.863)
RSF	**0.821**	**(0.747, 0.874)**

formance of the pipelines relative to each other, with FDR correction for p-values [31]. The traditional CPH model and the RSF using only clinical variables outperformed the MSKCC nomogram ($p - value < 0.001$). No significant difference was found between the MSKCC and either the LASSO or Boosted models (p-value = 0.210 and 0.915, respectively). There was also no significant difference between the traditional CPH model and the RSF (p-value = 0.280). These tests also indicated that at least one of the models

that included mRNA variables outperformed each of the clinical-only models. Furthermore, the inclusion of mRNA variables yielded improved performance for the CPH, Boosted Cox, and RSF models (all $p-values < 0.001$), but not for the LASSO Cox model (all $p-values > 0.300$). For all modelling strategies, the pipeline yielding the highest bootstrapped C-index was identified as the best-performing pipeline (shown in bold in Table 3.2) and retained for further analyses.

3.3.2 Calibration

The MSKCC nomogram, the current state-of-the-art post-operative model, is designed with a 2-year or 5-year horizon as the predictive endpoints, and all models were tested for both 2- and 5-year survival calibration, as shown in Figure 3.2. For this data, the MSKCC nomogram showed reasonable calibration for both these endpoints. Some under-prediction occurred for those with low observed survival and some over-prediction occurred for those with higher observed survival. For overall survival, however, the nomogram appeared mis-calibrated for this data, yielding under-prediction at all time-points of reference.

In the clinical-only models, overall survival was slightly over-predicted but with much closer adherence to the 45-degree line overall (thus, closer to the observed values) than the MSKCC model. All clinical-only models also showed reasonable calibration at the 2- and 5-year endpoints. It can be noted that in both 2- and 5-year endpoint curves the LASSO and Boosted Cox models yielded very little separation between the top four quintiles.

As for calibration of the best-performing mRNA models; RSF, Boosted, and LASSO models all yielded similar calibration curves with close adherence to the 45-degree line for overall survival as well as 2- and 5-year survival.

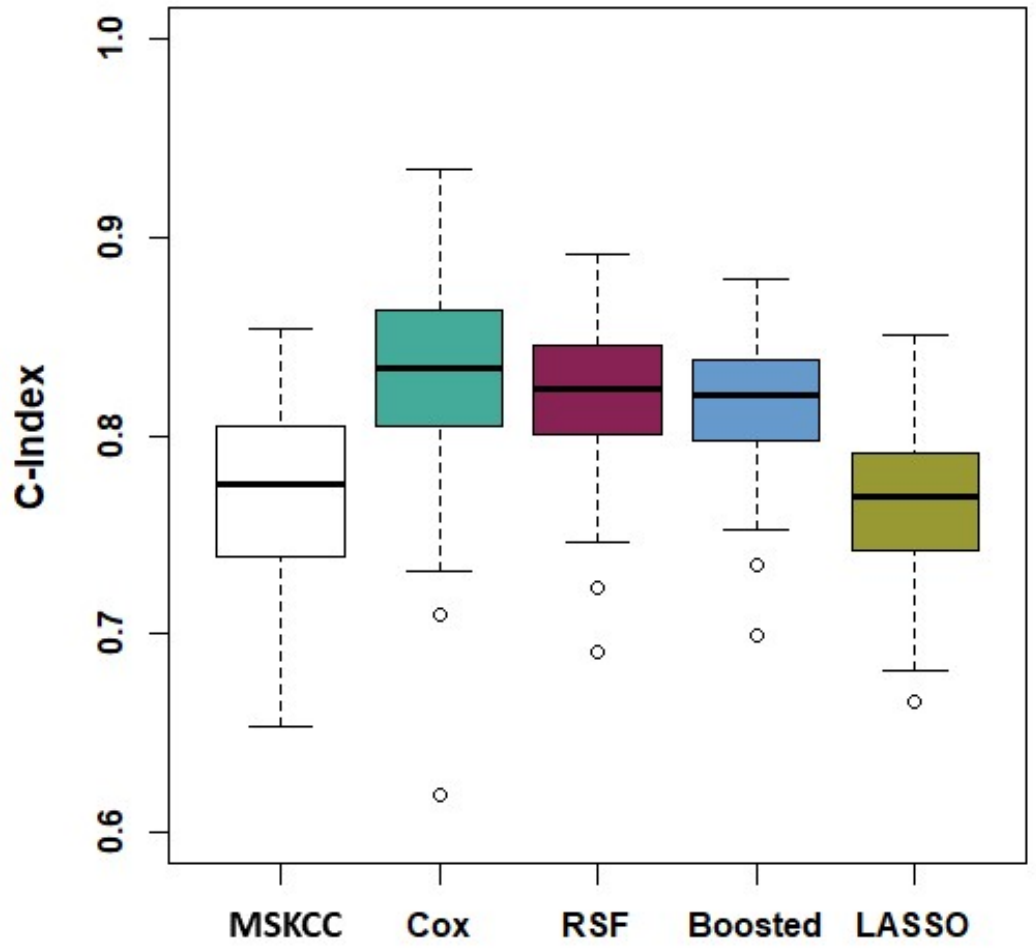

Figure 3.1: Bootstrap C-Index distributions of the MSKCC nomogram and the most competitive model pipelines using mRNA data. From **left** to **right**; MSKCC, Cox model with clinical variables and correlation prefiltered mRNA variables, RSF model with clinical variables and correlation prefiltered and univariate Cox feature selected mRNA, Boosted Cox model with clinical variables and correlation prefiltered mRNA variables, LASSO Cox model with clinical variables and univariate Cox feature selected mRNA.

The CPH model, however, showed poorer calibration for all three endpoints. Thus, optimisation of the CPH model with respect to the C-index, with the

inclusion of mRNA variable, led to a decrease in model calibration.

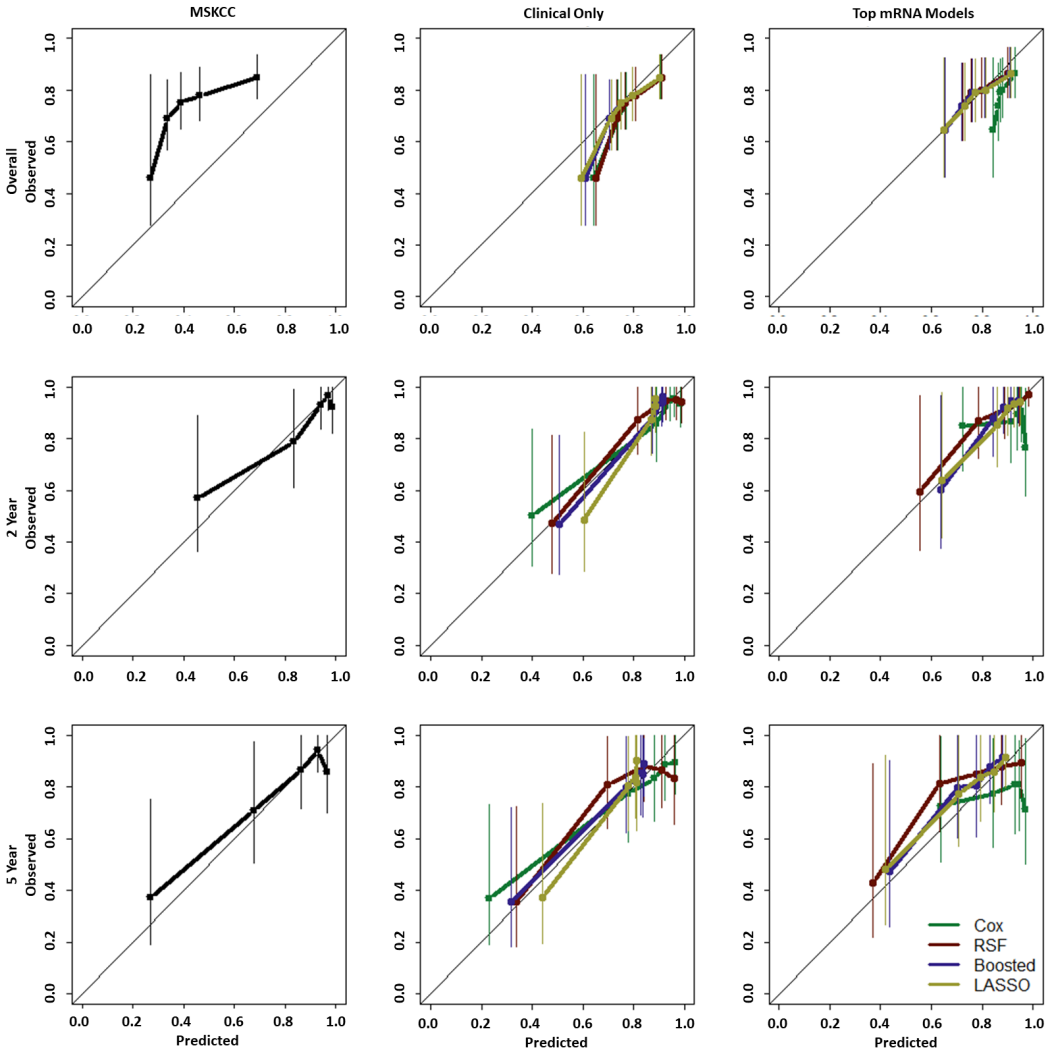

Figure 3.2: Calibration curves for MSKCC (**left**), clinical-only models (**centre**) and best-performing mRNA-based models (**right**) for (from **top** to **bottom**) overall, 2-year and 5-year survival.

3.3.3 Prediction Performance

ROC and decision curve analysis (DCA) were also both carried out for the 2- and 5-year survival endpoints, with the output shown in Figure 3.3. Optimal

out-of-bag prediction [48] was achieved by the RSF at the 2-year horizon (AUC = 0.824, CI of 2,000 bootstrap resamples = (0.716, 0.933)) and 5-year horizon (AUC = 0.772, CI = (0.651, 0.893)). DCA indicated that only the RSF model showed a continued net benefit across all threshold probabilities for both the 2- and 5-year endpoints. All models improved upon the "Treat All" method, but none of the other methodologies showed benefit at 2 years. At 5 years both the LASSO and Boosted Cox models retained some net benefit across all thresholds but were below that of the RSF model.

3.3.4 Model Stability and Feature Selection

Overall model stability was evaluated on all clinical-only pipelines and on the best performing pipelines for each modelling strategy when including the entire set of over 20,000 mRNA variables, identified in Table 3.2 in terms of their C-index performance. A first assessment consisted of statistical summaries of the sizes of the final predictive subsets selected by each pipeline, reported in Table 3.3. All clinical-only pipelines showed little deviation around a typical 5-variable model structure. In terms of the make-up of these models, however, the LASSO and Boosted Cox pipelines selected 14 clinical variables across all bootstraps, compared to 8 for traditional CPH and the RSF (Table 3.3), suggesting higher stability of the latter 2 for the clinical-only models. When including mRNA data, all models typically tended to contain around 10 to 16 variables for prediction. Remarkably, the typical final feature set size of Cox-based mRNA-inclusive pipelines had a coefficient of variation (CoV) either lower than, or comparable to that of clinical-only pipelines, indicating reasonably low variability in final model sizes for mRNA-inclusive models. An increase in CoV, however acceptable, was noted for RSF after introduction of mRNA information. This first analysis thus suggested that mRNA-inclusive

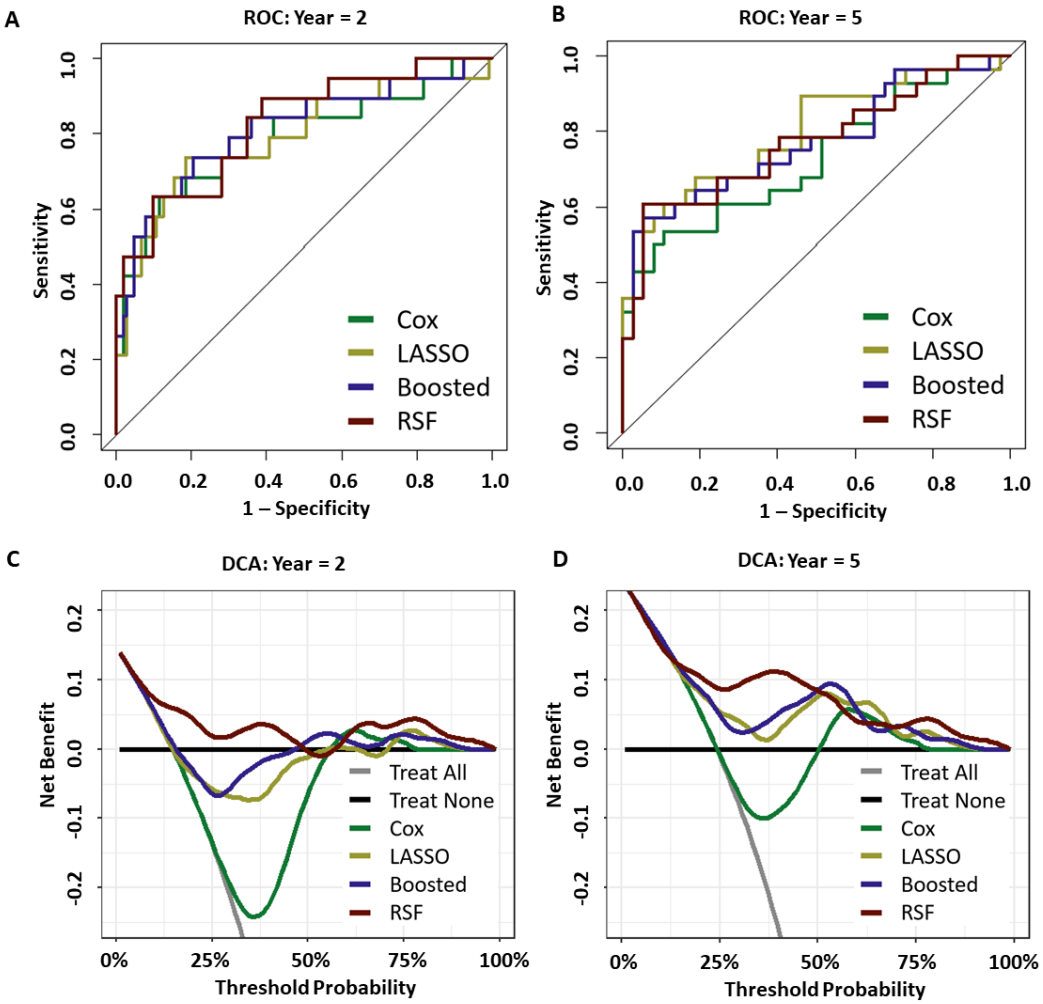

Figure 3.3: (**A**,**B**): ROC for 2-year (**A**) and 5-year (**B**) horizons. (**C**,**D**): DCA curves at the same horizons.

pipelines tended to include a reasonable number of variables in their models.

Model stability was further measured by considering the overall individual feature selection rates across bootstraps, as well as on the basis of three overall stability indices [47], namely the Jaccard, Davis, and Dice indices, as shown in Figure 3.4. The Davis index in Figure 3.4A confirmed increased stability of the traditional Cox and RSF pipelines over the LASSO and Boosted Cox alternatives for clinical-only modelling. As for mRNA-inclusive pipelines,

Table 3.3: Summaries of final predictive model sizes (i.e., number of variables used) selected within the clinical-only models and best-performing pipelines, in terms of their mean, median, standard deviation, and coefficient of variation (CoV), as well as the overall number of features selected at least once across 100 bootstraps.

	Mean	**Median**	**SD**	**CoV**	**Overall**
Clinical-only					
Cox	4.55	5	1.250	27.5%	8
LASSO	3.51	3	1.541	43.9%	14
Boosted	5.10	5	0.870	17.1%	14
RSF	4.96	5	1.524	30.7%	8
Top performing pipelines					
Cox	9.46	10	0.926	9.8%	623
LASSO	11.88	12	3.891	32.8%	343
Boosted	14.98	15	3.052	20.4%	323
RSF	17.85	16	9.057	50.7 %	335

the comparison of metrics shown in Figure 3.4A with those of Figure 3.4B indicated lower stability, as would be expected given the sheer size of the genomic dataset. Finally, RSF outperformed the CPH pipeline when using mRNA, as shown in Figure 3.4B. In complement, the traces of overall feature selection rates shown in Figure 3.4C,D illustrate the tendency of RSF pipelines to yield higher model stability and feature selection rates overall compared to the LASSO and Boosted Cox models, whether using clinical-only or combined clinical and mRNA information.

In summary, this set of results clearly indicates that the RSF is a more stable pipeline than the LASSO and Boosted Cox alternatives. Whether considering clinical-only or clinical and mRNA features, the RSF yielded the most stable pipeline overall in the latter case. Moreover, the RSF and Boosted Cox methodologies, which were the most competitive overall, do not rely on a proportional hazard assumption. This may suggest that the Cox PH assumption is a possible limitation for the inclusion of mRNA information. Similarly,

RSF does not constrain to linear interactions between the predictors, which may be a reason for its overall higher predictive performance. The overall lower performance seen with LASSO Cox, on the other end, may be due to the inability of this model to cope well with multi-collinearity or confounding, both of which are possibly present among the mRNA feature set. The differences in performance observed between the models must be confirmed in independent studies before any reliable insights can be derived on the specific aspects behind the performance of each model.

Overall, the pathological Gleason Score was the single-most prevalent predictor across all pipelines (Figure 3.4C,D; Table 3.4). It was the only variable that was selected more than 50% of the time in conventional Cox modelling. The other three modelling approaches had more stable individual feature selection rates. Figure 3.4C illustrates the consistent tendency by all clinical-only models to use pathological Gleason score and SVI in particular. This trend remained present in LASSO and Boosted Cox models following the introduction of mRNA data (Figure 3.4D and Table 3.4), whereas RSF tended to use LNI and PSA mainly alongside the Gleason score. Finally, two mRNA variables, namely DNAH8 and ESM1, appeared consistently in a number of pipelines. Both of these have been noted in other reports for their association with prostate cancer. DNAH8 has been shown to be a marker for poor prognosis in prostate cancer [49]. Elevated ESM1 has been shown to be associated with prostate cancer progression and metastasis [50].

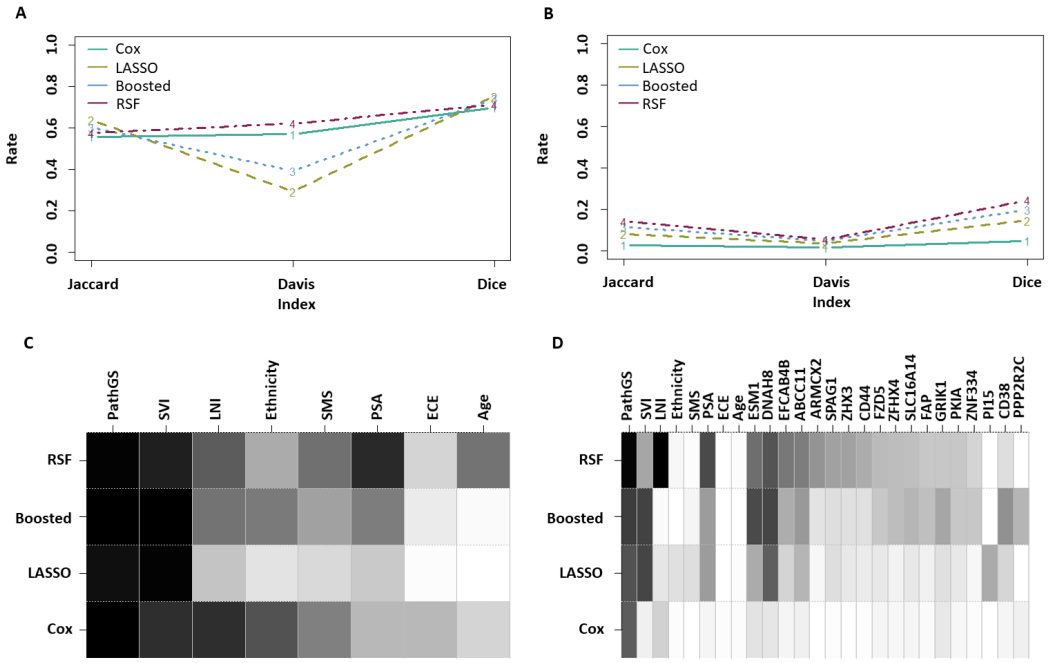

Figure 3.4: (**A**) Stability indices of the clinical-only models. (**B**) Stability indices of the top mRNA-inclusive pipelines. (**C**) Feature selection rates in clinical-only models (a darker shade indicates a higher rate). (**D**) Feature selection rates in the best-performing mRNA-inclusive pipelines, for features with a selection rate ≥ 0.2 in at least one of these top pipelines. For comparative purposes, the shade of grey used for ESM1 in RSF corresponds to a 50% selection rate.

Table 3.4: Selection rates of all variables with a selection rate $\geq$ 50% in at least one of the best-performing mRNA-inclusive pipelines.

	Pathological Gleason Score	**LNI**	**SVI**	**PSA**	**DNAH8**	**ESM1**
Cox	56%	16%	5%	4%	9%	10%
LASSO	60%	9%	64%	34%	55%	29%
Boosted	67%	2%	65%	34%	64%	62%
RSF	86%	88%	31%	62%	59%	50%

3.4 Discussion

3.4.1 Key Findings

In this study, the machine learning pipelines developed for prediction of BCR-free survival using Cox proportional hazards and random survival forest models using clinical variables only outperformed the reference MSKCC nomogram with respect to model discrimination and calibration. This is in line with previous findings by Jeffers et al., where their RSF model with clinical variables has only been shown to improve on the Kattan nomogram [51]. Additionally, predictive performance of the models was enhanced by the addition of mRNA variables. The three best performing pipelines included both mRNA information and clinical variables, indicating the potential for genetic information to improve the predictive performance when combined with routine clinical variables. The mRNA features most frequently included in the models were DNAH8 (Dynein axonemal heavy chain 8) and ESM1 (Endothelial Cell Specific Molecule 1). The main function of DNAH8 is currently understood to be related to sperm motility [52] and has been shown to be a marker for poor prognosis in prostate cancer [49] and to be associated with metastasis [53]. Elevated ESM1 readings have been shown to be associated with prostate cancer progression and metastasis [50], ESM1 over-expression correlates to a higher Gleason Score [54]. Both of these variables were also found to be useful predictors in the pre-operative setting by this group in other analyses.

3.4.2 Other Markers of Interest

Four clinical variables were repeatedly retained in the models that included mRNA information; pathological Gleason sum, seminal vesicle invasion, pre-operative PSA, and lymph node involvement. PSA was not consistently selected in the Cox-based models, but was in the RSF pipelines, which suggests that this antigen level may interact with predictors in a nonlinear way with respect to the endpoint of interest. None of the mRNA variables identified by Taylor et al. as tumour suppressors [25] were selected in our models. This may indicate a lack of association between tumour suppression and biochemical recurrence. Similarly, none of the variables used in the Decipher classifier for the prediction of metastasis following RP [22] were selected. These mRNA variables and their pathways, which are associated with metastasis, may be different to those involved in biochemical recurrence itself. Other non-genetic markers such as measures of systemic inflammation such as neutrophil-to-lymphocyte ratio (NLR) have also been investigated recently for their potential as predictive markers for BCR. Results on the prognostic potential of NLR for BCR prediction presented in the literature are mixed [55–61]. As the dataset used in this study did not contain this information, it was not possible to investigate the potential of NLR in conjunction with mRNA variables at this time, and it is therefore an avenue for future work.

3.4.3 Model Performance and Key Feature Identification

While the traditional CPH model with prefiltered mRNA variables demonstrated the best discrimination performance, its poorer calibration at specific endpoints as well as overall calibration suggests overfitting. On the other

end, the best performing RSF-based pipeline outperformed other modelling strategies in terms of discrimination and calibration for 2-year, 5-year, and overall BCR-free survival prediction. The RSF model was also shown to have the highest AUC at both 2-year and 5-year horizons and gave rise to the most consistent net benefit at these time-points in the decision curve analysis. Among the most common methodologies available for survival analysis in continuous time, it may thus be a more opportune strategy overall compared to Cox-based approaches. Although the machine learning models were tested internally using a bootstrapping framework, these findings require external validation, which was not possible given the relatively small size of this dataset. Further improvement may also be possible using novel deep learning techniques, such as the Cox-based deep neural network recently described by Hao et al. 2021 [62], for application to high-dimensional data with small sample sizes. Experimentation of this recent innovation to BCR prediction for prostate cancer will be done in future work by this group.

Not all mRNA-inclusive pipelines yielded an increase in the predictive performance over their clinical-only counterpart. Whilst including mRNA variables in Boosted Cox models always led to a statistically significant increase in performance, regardless of feature pre-filtering and selection technique used, none of the LASSO Cox alternatives did. As for RSF pipelines, improvement required the combined use of correlation prefiltering and univariate feature selection. These findings are in line with those found by Gade et al., who investigated the inclusion of microRNA and mRNA data by using a bipartite graph as a single predictor [63]. The differences across the pipelines may be due to a large extent to the lack of statistical power associated with the relatively small sample size available, but it also indicates that the model building process is non trivial. Less than 2% of the features available were

used at least once by a pipeline within the bootstrapping framework, and a gain in both model stability and predictive performance may be achieved with feature pre-processing strategies that are better suited to the nature of these data.

Moreover, unknown potential predictors of BCR may have been inadvertently removed from the pool of candidates prior to model building, e.g., due to the inability of feature selection methods based on multiple univariate selection to identify variables that may provide additional predictive potential only when used in conjunction with other variables. Other methods of pre-filtering and feature elimination will be investigated in future work including the potential benefit of clustering and other relevant methods such as principal component analysis.

3.5 Conclusion

The ability to predict BCR-free survival is critical for therapeutic decision-making. After comparing several machine learning methodologies with the reference Memorial Sloan Kettering nomogram, the main finding of this study was that the inclusion of mRNA variables increased the predictive performance. The best-performing pipeline combined correlation based pre-filtering and univariate feature selection of mRNA variables with a random survival forest model. It yielded high discriminative performance while retaining satisfactory calibration at 2-year, 5-year and overall survival, as well as the highest AUC and net benefit across all pipelines benchmarked. mRNA variables ESM1 and DNAH8 were consistently found to be strong predictors for 2-year, 5-year, and overall BCR-free survival.

3.6 References

[1] Sanda MG, Cadeddu JA, Kirkby E, Chen RC, Crispino T, Fontanarosa J, et al. Clinically localized prostate cancer: AUA/ASTRO/SUO guideline. Part II: recommended approaches and details of specific care options. The Journal of urology. 2018;199(4):990-7.

[2] Wang J, Xia HHX, Zhang Y, Zhang L. Trends in treatments for prostate cancer in the United States, 2010-2015. American journal of cancer research. 2021;11(5):2351.

[3] John Hopkins Medicine. Prostate Cancer Prognosis. John Hopkins Medicine; 2022. `https://www.hopkinsmedicine.org/health/conditions-and-diseases/prostate-cancer/prostate-cancer-prognosis`.

[4] Tourinho-Barbosa R, Srougi V, Nunes-Silva I, Baghdadi M, Rembeyo G, Eiffel SS, et al. Biochemical recurrence after radical prostatectomy: what does it mean? International braz j urol. 2018;44:14-21.

[5] Walz J, Chun FKH, Klein EA, Reuther A, Saad F, Graefen M, et al. Nomogram predicting the probability of early recurrence after radical prostatectomy for prostate cancer. The Journal of urology. 2009;181(2):601-8.

[6] Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. Cancer: Interdisciplinary International Journal of the American Cancer Society. 2008;113(11):3075-99.

[7] Moghaddam S, Jalali A, O'Neill A, Murphy L, Gorman L, Reilly AM, et al. Integrating serum biomarkers into prediction models for biochemical recurrence following radical prostatectomy. Cancers. 2021;13(16):4162.

[8] Diblasio CJ, Kattan MW. Use of nomograms to predict the risk of disease recurrence after definitive local therapy for prostate cancer. Urology. 2003;62:9-18.

[9] Memorial Sloan Kettering Cancer Center. Prostate Cancer Nomograms. Memorial Sloan Kettering Cancer Center; 2022. `https://www.mskcc.org/nomograms/prostate`.

[10] Kattan MW, Wheeler TM, Scardino PT. Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. Journal of clinical oncology. 1999;17(5):1499-9.

[11] Stephenson AJ, Scardino PT, Eastham JA, Bianco Jr FJ, Dotan ZA, DiBlasio CJ, et al. Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2005;23(28):7005.

[12] Kattan MW, Vickers AJ, Yu C, Bianco FJ, Cronin AM, Eastham JA, et al. Preoperative and postoperative nomograms incorporating surgeon experience for clinically localized prostate cancer. Cancer: Interdisciplinary International Journal of the American Cancer Society. 2009;115(5):1005-10.

[13] Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187-202.

[14] Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. JNCI: Journal of the National Cancer Institute. 1998;90(10):766-71.

[15] Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in medicine. 1997;16(4):385-95.

[16] Sohn I, Kim J, Jung SH, Park C. Gradient lasso for Cox proportional hazards model. Bioinformatics. 2009;25(14):1775-81.

[17] Joffe E, Coombes KR, Qiu YH, Yoo SY, Zhang N, Bernstam EV, et al. Survival Prediction In High Dimensional Datasets–Comparative Evaluation Of Lasso Regularization and Random Survival Forests. Blood. 2013;122(21):1728.

[18] De Bin R. Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. Computational Statistics. 2016;31(2):513-31.

[19] Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC bioinformatics. 2008;9(1):1-10.

[20] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The annals of applied statistics. 2008;2(3):841-60.

[21] Sager M, Yeat NC, Pajaro-Van der Stadt S, Lin C, Ren Q, Lin J. Transcriptomics in cancer diagnostics: developments in technology, clinical research and commercialization. Expert Review of Molecular Diagnostics. 2015;15(12):1589-603.

[22] Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PloS one. 2013;8(6):e66855.

[23] Cuzick J, Berney D, Fisher G, Mesher D, Møller H, Reid J, et al. Prognostic value of a cell cycle progression signature for prostate cancer death in a conservatively managed needle biopsy cohort. British journal of cancer. 2012;106(6):1095-9.

[24] Klein EA, Cooperberg MR, Magi-Galluzzi C, Simko JP, Falzarano SM, Maddala T, et al. A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. European urology. 2014;66(3):550-60.

[25] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer cell. 2010;18(1):11-22.

[26] Kryvenko ON, Epstein JI. Prostate cancer grading: a decade after the 2005 modified Gleason grading system. Archives of pathology & laboratory medicine. 2016;140(10):1140-52.

[27] Wittekind C, Asamura H, Sobin LH. TNM Atlas: Illustrated Guide to the TNM Classification of Malignant Tumours. Wiley-Blackwell; 2014.

[28] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of statistical software. 2011;39(5):1.

[29] Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B. mboost: Model-Based Boosting; 2022. R package version 2.9-7. Available from: `https://CRAN.R-project.org/package=mboost`.

[30] Goh L, Song Q, Kasabov N. A novel feature selection method to improve classification of gene expression data. In: Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29; 2004. p. 161-6.

[31] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289-300.

[32] Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature medicine. 2002;8(8):816-24.

[33] Witten DM, Tibshirani R. Survival analysis with high-dimensional covariates. Statistical methods in medical research. 2010;19(1):29-51.

[34] Hastie T, R T, J F. The Elements of Statistical Learning. Springer; 2009.

[35] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543-6.

[36] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.

[37] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2021. Available from: `https://www.R-project.org/`.

[38] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1-22. Available from: `https://www.jstatsoft.org/v33/i01/`.

[39] Ishwaran H, Kogalur UB. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). manual; 2021. R package version 2.13.0. Available from: `https://cran.r-project.org/package=randomForestSRC`.

[40] Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996;15(4):361-87.

[41] Remmers S, Verbeek JF, Nieboer D, van der Kwast T, Roobol MJ. Predicting biochemical recurrence and prostate cancer-specific mortality after radical prostatectomy: comparison of six prediction models in a cohort of patients with screening-and clinically detected prostate cancer. BJU international. 2019;124(4):635-42.

[42] Royston P. Tools for checking calibration of a Cox model in external validation: prediction of population-averaged survival curves based on risk groups. The Stata Journal. 2015;15(1):275-91.

[43] Bewick V, Cheek L, Ball J. Statistics review 13: receiver operating characteristic curves. Critical care. 2004;8(6):1-5.

[44] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making. 2006;26(6):565-74.

[45] Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences. 2022;34(4):1060-73.

[46] Davis CA, Gerick F, Hintermair V, Friedel CC, Fundel K, Küffner R, et al. Reliable gene signatures for microarray classification: assessment of stability and performance. Bioinformatics. 2006;22(19):2356-63.

[47] Bommert A, Lang M. stabm: Stability Measures for Feature Selection. Journal of Open Source Software. 2021;6(59):3010.

[48] Breiman L. Out-of-bag estimation. 1996.

[49] Wang Y, Ledet RJ, Imberg-Kazdan K, Logan SK, Garabedian MJ. Dynein axonemal heavy chain 8 promotes androgen receptor activity and associates with prostate cancer progression. Oncotarget. 2016;7(31):49268.

[50] Pan KF, Lee WJ, Chou CC, Yang YC, Chang YC, Chien MH, et al. Direct interaction of β-catenin with nuclear ESM1 supports stemness of metastatic prostate cancer. The EMBO journal. 2021;40(4):e105450.

[51] Jeffers A, Sochat V, Kattan MW, Yu C, Melcon E, Yamoah K, et al. Predicting prostate cancer recurrence after radical prostatectomy. The Prostate. 2017;77(3):291-8.

[52] Liu C, Miyata H, Gao Y, Sha Y, Tang S, Xu Z, et al. Bi-allelic DNAH8 variants lead to multiple morphological abnormalities of the sperm flagella and primary male infertility. The American Journal of Human Genetics. 2020;107(2):330-41.

[53] Song F, Zhang Y, Pan Z, Hu X, Yi Y, Zheng X, et al. Identification of novel key genes associated with the metastasis of prostate cancer based on bioinformatics prediction and validation. Cancer cell international. 2021;21(1):1-16.

[54] Lai CY, Chen CM, Hsu WH, Hsieh YH, Liu CJ. Overexpression of endothelial cell-specific molecule 1 correlates with Gleason score and expression of androgen receptor in prostate carcinoma. International Journal of Medical Sciences. 2017;14(12):1263.

[55] Bahig H, Taussky D, Delouya G, Nadiri A, Gagnon-Jacques A, Bodson-Clermont P, et al. Neutrophil count is associated with survival in localized prostate cancer. BMC cancer. 2015;15:1-8.

[56] Jang W, Cho K, Kim K, Yoon C, Kang Y, Lee J, et al. Prognostic impact of preoperative neutrophil-to-lymphocyte ratio after radical prostatectomy in localized prostate cancer. Prostate Cancer and Prostatic Diseases. 2016;19(3):298-304.

[57] Kwon YS, Han CS, Yu JW, Kim S, Modi P, Davis R, et al. Neutrophil and lymphocyte counts as clinical markers for stratifying low-risk prostate cancer. Clinical genitourinary cancer. 2016;14(1):e1-8.

[58] Lee H, Jeong SJ, Hong SK, Byun SS, Lee SE, Oh JJ. High preoperative neutrophil–lymphocyte ratio predicts biochemical recurrence in patients with localized prostate cancer after radical prostatectomy. World journal of urology. 2016;34:821-7.

[59] Poyet C, Adank JP, Keller E, Mortezavi A, Rabgang T, Pfister B, et al. Mp66-05 Pretreatment Systemic Inflamatory Response Parameters Do

Not Predict the Outcome in Men With Prostate Cancer Undergoing Radical Prostatectomy. The Journal of Urology. 2015;193(4S):e817-7.

[60] Sharma V, Cockerill PA, Viers BR, Rangel LJ, Carlson RE, Karnes RJ, et al. Mp6-05 the association of preoperative neutrophil to lymphocyte ratio with oncologic outcomes following radical prostatectomy for prostate cancer. The Journal of Urology. 2015;193(4S):e55-6.

[61] Zhang GM, Zhu Y, Ma XC, Qin XJ, Wan FN, Dai B, et al. Pretreatment neutrophil-to-lymphocyte ratio: a predictor of advanced prostate cancer and biochemical recurrence in patients receiving radical prostatectomy. Medicine. 2015;94(41).

[62] Hao L, Kim J, Kwon S, Ha ID. Deep learning-based survival analysis for high-dimensional survival data. Mathematics. 2021;9(11):1244.

[63] Gade S, Porzelius C, Fälth M, Brase JC, Wuttig D, Kuner R, et al. Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction in prostate cancer. BMC bioinformatics. 2011;12:1-10.

Chapter 4

Pre-Operative Prediction of BCR-Free Survival with mRNA Variables in Prostate Cancer

Abstract

Technological innovation yielded opportunities to obtain mRNA expression data for prostate cancer (PCa) patients even prior to biopsy, which can be used in a precision medicine approach to treatment decision-making. This can apply in particular to predict the risk of, and time to biochemical recurrence (BCR). Most mRNA-based models currently proposed to this end are designed for risk classification and post-operative prediction. Effective pre-operative prediction would facilitate early treatment decision-making, in particular by indicating more appropriate therapeutic pathways for patient profiles who would likely not benefit from a systematic prostatectomy regime. The aim of this study is to investigate the possibility to leverage mRNA information pre-operatively for BCR-free survival prediction. To do this, we considered time-to-event machine learning (ML) methodologies, rather than classification models at a specific survival horizon.

We retrospectively analysed a cohort of 135 patients with clinical follow-up data and mRNA information comprising over 26,000 features (data accessi-

ble at NCBI GEO database, accession GSE21032). The performance of ML models including random survival forest, boosted and regularised Cox models were assessed, in terms of model discrimination, calibration, and predictive accuracy for overall, 3-year and 5-year survival, aligning with common clinical endpoints.

Results showed that the inclusion of mRNA information could yield a gain in performance for pre-operative BCR prediction. ML-based time-to-event models significantly outperformed reference nomograms that used only routine clinical information with respect to all metrics considered. We believe this is the first study proposing pre-operative transcriptomics models for BCR prediction in PCa.

External validation of these findings, including confirmation of the mRNA variables identified as potential key predictors in this study, could pave the way for pre-operative precision nomograms to facilitate timely personalised clinical decision-making.

4.1 Introduction

Although radical prostatectomy (RP) has been the primary treatment for Prostate Cancer (PCa) for the last forty years, post-operative recurrence remains high [1]. This event is determined by a rise in Prostate Specific Antigen (PSA) and is termed a biochemical recurrence (BCR). BCR occurs in 20-40% of patients following RP [2]. Thus, determining factors that influence the time-to-BCR is critical for treatment decision-making. In particular, the ability to predict treatment failure pre-operatively could facilitate earlier decision-making regarding primary treatments and adjuvant therapies [3], including as to whether to recommend RP to patients at higher risk of recurrence. Recent

technological improvements have created the opportunity to obtain patient genetic data routinely, which could be used in a precision medicine approach to inform therapeutic decisions. Identifying clinically useful variables with a strong potential for prediction of BCR is the main objective of this study. Specifically, we aim to examine the potential for inclusion of messenger ribonucleic acid (mRNA) transcriptomic information. In doing so, this work also evaluates the potential of machine learning (ML) methodologies over traditional methods and state-of-the-art nomograms. To date, most mRNA-based models proposed to this end are designed for risk classification and post-operative prediction. Effective pre-operative prediction would facilitate early treatment decision-making, in particular by indicating more appropriate therapeutic pathways for patient profiles who would likely not benefit from a systematic prostatectomy regime.

The Kattan nomogram for time-to-BCR prediction has been widely validated and is still commonly used for baseline comparison due to its high AUC (0.80) when initially assessed [4]. The Memorial Sloan Kettering Cancer Center (MSKCC) web-based nomogram is an update on the Kattan nomogram, with the most recent version dated April 2021 yielding a C-index of 0.80 on validation of a single-centred site [5]. While these nomograms have good performance metrics, they are subject to limitations. As the Kattan model was created in 1998, some criteria in its design are outdated. For instance, the tumour grading protocol used for Gleason scoring has changed so that the lowest score (originally 5) now assigned is 6 [6]. The PSA level deemed to indicate BCR has also changed from ≥ 0.4 ng/mL to having two PSA measurements ≥ 0.2 ng/mL [7].

Both models utilise routine preoperative clinical variables only, despite advances being made in this field to generate and integrate high-throughput

patient-specific information. To date, predictive models that use mRNA information have been considered mainly in post-operative settings for the prediction of BCR [8, 9] and metastasis [10]. As the method currently implemented for mRNA extraction relies on tumour samples, genetic information generally only becomes available post-operatively. However, it is now possible to obtain this information from pre-operative needle biopsy, or more recent techniques again such as liquid biopsy, which gives rise to potential inclusion of mRNA information earlier in the patient care timeline. A number of studies have shown that the genetic profile at biopsy is representative of the entire tumour [11–13]. With this in mind, we assume that pre-operative mRNA information aligns strongly with that obtained during RP used in our models. In this study, we used mRNA data acquired post-operatively as a proxy for pre-operative modelling, together with pre-operative routine clinical variables, to produce the first pre-operative model for prediction of time-to-BCR. Other pre-operative models found in the literature are classifiers designed for binary prediction of recurrence or metastasis [14, 15], rather than time-to-event prediction models. Moreover, the selection of relevant mRNA biomarkers is performed here from within a set of over 26,000 mRNA variables, independently of previous findings, unlike in other (post-operative) mRNA-based models [8] that used predefined sets of variables. As such, the proposed study also proposes a benchmark of ML methodologies against conventional clinical modelling strategies for pre-operative time-to-BCR prediction.

4.2 Materials and Methods

4.2.1 Data

The dataset was developed in a study by Taylor *et al* [16] containing 232 patients, of whom 198 had BCR-free survival time, along with censoring status, routine pre-operative clinical variables (namely patient age, PSA at diagnosis, Gleason score, ethnicity and clinical stage), and over 26,000 mRNA transcriptome variables. The data was accessed on the 1st of May 2021 and was fully anonymised at source (data accessible at NCBI GEO database, accession GSE21032). The mRNA data was obtained from surgical specimens during RP. Six patients with missing information were removed from the dataset, imputation not being feasible due to the large amount of missing information for each case, the bias introduced by imputation would likely outweigh that of selection bias. A further two patients had identical values for all variables except metastatic status and were thus both removed. Gleason score of 5 (n = 2), and Gleason score of 6 (n = 100) were combined to align with current grading protocols [6]. Of the 190 remaining patients, 135 had mRNA information. Our analyses were thus carried out on this sub-cohort.

4.2.2 Models

In this study, follow-up times varied between patients, and the sole event of interest was BCR. Censored patients were included for analysis up until censoring occurred. Let the right-censored survival information recorded for the N patients in this cohort be denoted by

$$(T, \delta)_i, i = 1, \ldots, N$$

where $T \geq 0$ and $\delta \in \{0, 1\}$ are respectively BCR-free survival time and censoring status ($\delta = 0$ indicating the patient observation was censored at time T). Letting T^0 be the true BCR-free survival time, and τ^0 the true censoring time, then $T = min(T^0, \tau^0)$. Let also $m \leq N$ be the number of BCR events in this cohort, at distinct times

$$t_1 < t_2 < \cdots < t_m,$$

and d_j BCR events recorded at time t_j, $j = 1, \ldots, m$. Given an $N \times P$ set $\mathbf{X} = (X_1, \ldots, X_P)$ of P variables recorded for each patient, four distinct modelling strategies with the capacity to adjust for censoring were applied to the data for survival analysis of the time to BCR ($\{(T, \delta, \mathbf{X})_i\}_{i=1}^N$). From this analysis it was possible to derive the survival curve $S_i(t) = Pr(T_i > t)$ and associated survival characteristics of each patient. As several subsets of $\mathbf{X}$ were considered in our analyses, hereafter P loosely denotes the number of features in the relevant subset. One of the methods we considered for survival analysis was the conventional Cox proportional hazard model [17]. A limitation of this model is its inapplicability to wide datasets, where the number of predictors P far exceeds the number of observations N. To overcome this limitation, a forward stepwise selection (FS) was used, selecting only the most important features. As ML methodologies are widely recognised as being better suited for this type of high dimensionality, LASSO Cox [18] and boosted Cox [19] models were considered as alternatives to FS. The fourth strategy considered was the random survival forest (RSF), which consists of bagging survival trees [20]. The chosen ML models were selected for their prevalence in the literature and their shown improved performance in previous studies in cancer [21–23]. Different models were produced for each of these techniques based on different variable subsets, with sixteen models created. Firstly, Cox,

RSF, LASSO Cox and boosted Cox models were developed using routine pre-operative clinical variables only, and compared to the current MSK nomogram for reference. The other twelve models were built using combined clinical and mRNA variable subsets obtained following either (unsupervised) correlation pre-filtering, or (supervised) univariate Cox feature selection, or both. The correlation filter was applied to the entire dataset, using an absolute cut-off value of 0.6 on Pearson's correlation [24], a threshold deemed optimal in previous studies on similar data [25]. Correlation was used as a pre-filtering technique as it not only allows for dimensionality reduction but provides an unsupervised method to reduce multicollinearity in the data [26]. Univariate Cox feature selection of mRNA variables was performed on each of the resampled training sets, and the 50 variables with the best log-rank scores were retained in line with methods used by Beer *et al* on their assessment of gene expressions and the prediction of lung cancer [27]. Feature selection based on univariate Cox proportional hazard analyses is a supervised technique commonly used in clinical applications, that consists in electing standalone predictors for multivariate modelling [28]. FS was performed on all Cox models i.e. with or without feature pre-filtering/selection, allowing for a maximum of 10 variables to ensure a reasonable N/P ratio. Recursive feature elimination was performed on all RSF models. The hyperparameters of the RSF and boosted Cox models were kept to conservative settings [29,30]. Thus, the RSF models were obtained using 500 trees, considering $\sqrt{P}$ variables at each split and with a minimum size of 15 for the terminal nodes and boosted Cox models were built using 100 boosting iterations and with a fixed shrinkage parameter of 0.1. The regularisation parameter (λ) was selected using nested cross-validation for the LASSO Cox models [31], using a final value for λ corresponding to one standard deviation of the cross-validated

error, as a standard approach to achieve parsimony [32].

4.2.3 Evaluation Framework

The 16 models described above were evaluated with respect to model discrimination, calibration, predictive performance and feature selection stability, via bootstrapping (using 100 iterations, i.e. where data was repeatedly sampled with replacement to create a dataset for training of equal size to the full dataset and the unselected samples, or out-of-bag (OOB) points, were used for testing). All analyses were carried out using the statistical software environment R [33] version 4.0.2.

Discrimination

In this study, discrimination refers to a model's ability to correctly rank patients relative to each other based on their predicted time-to-BCR. A bootstrap-corrected concordance index (C-index), defined as the proportion of all feasible pairs of patients whose predicted and observed outcomes are concordant [34], and associated confidence interval (CI), were derived for each model for unbiased estimation of external predictive discrimination.

$$\mathscr{C} = \hat{P}(\hat{S}_i(t) > \hat{S}_j(t) | T_i > T_j)$$

Bootstrap-corrected C-indices take into consideration the difference between the bootstrap sample models performance and the performance of the model built using the full dataset penalising for overfitting.

Calibration

Two forms of calibration analyses were used to assess model bias [35]. The first method is the more commonly implemented specific time-point method

which aims to assess the conformity of a model to a binary outcome and is described in more detail in Harrell et al [35]. The second evaluates the models conformity of a continuous model response typical of survival analysis. Both have been derived based on the OOB survival curve estimates $\{\hat{S}_i(t)\}_{i=1}^N$ yielded by each model for all N patients. Since the MSK nomogram is specifically calibrated to 5-year survival, calibration curves at the 5-year horizon were obtained with respect to quintiles $\{\tilde{p}_1, \ldots, \tilde{p}_5\}$ of the sample of values $\{\hat{p}_i^5\}_{i=1}^N$, where $\hat{p}_i^5 = \hat{S}_i(5 \text{ years}) = \hat{Pr}(T_i > 5 \text{ years})$ denotes the probability for patient i to survive at least 5 years. Points on the calibration curves have coordinates $(\bar{S}_q^5, \hat{S}_{KM,q}^5), q = 1, \ldots, 5$, where $\bar{S}_q^5$ denotes the sample mean survival likelihood

$$\bar{S}_q^5 = \frac{1}{|\mathscr{S}_q^5|} \sum_{l \in \mathscr{S}_q^5} \hat{S}_l(5)$$

for each subset $\mathscr{S}_q^5 = \{i \in \{1, \ldots, N\} | \tilde{p}_{q-1} < p_i^5 \leq \tilde{p}_q\}, q = 1, \ldots, 5$ (setting $\tilde{p}_0 = 1$) [36], and $\hat{S}_{KM,q}^5$ denotes the value of the Kaplan-Meier estimate of BCR-free survival $\hat{S}_{KM,q}(t)$ obtained for each subset $\mathscr{S}_q^5$ and evaluated at $t = 5$ years. A second approach was used to assess model calibration for overall BCR-free survival estimation. This "overall survival" calibration curve was defined with respect to the quintiles $\{\tilde{t}_1, \ldots, \tilde{t}_5\}$ of the time-to-event data $\{t_j\}_{j=1}^m$ by coordinates $(\bar{S}_q, \hat{S}_{KM}(\tilde{t}_q)), q = 1, \ldots, 5$, where $\bar{S}_q$ denotes the sample mean survival likelihood

$$\bar{S}_q = \frac{1}{|\mathscr{S}_q|} \sum_{l \in \mathscr{S}_q} \hat{S}_l(t_l)$$

in each subset $\mathscr{S}_q = \{i \in \{1, \ldots, N\} | \tilde{t}_{q-1} < t_i \leq \tilde{t}_q\}, q = 1, \ldots, 5$ (setting $\tilde{t}_0 = 0$) [36]. The confidence bands for each calibration curve were derived directly from the corresponding Kaplan-Meier estimates.

Predictive Performance

Bootstrapped receiver operating characteristic (ROC) curves and corresponding area under the ROC curve (AUC) values were evaluated for all models to assess their predictive performance at specific, clinically relevant time-points [37]. ROC is used to assess the correct classification of subjects as remaining BCR-free or not at a specific time-point. As certain points in time are considered clinically important for determination further treatment this analysis can give an indication of the performance of a model for this classification task. Decision curve analysis (DCA) [38] was carried out to provide complementary assessment in terms of the clinical value of the predictions from each model. DCA is also considered at specific time-points and again relates to the clinical importance of the model's performance. The net benefit considered in the DCA is an indication of the increased or decreased number of patients who could see benefit in their treatment if the model was used for evaluation. Where a positive net benefit can be viewed as the proportion of patients who would benefit the from use of the model in a clinical decision making setting.

Lastly, feature selection rates from all model fits were recorded and analysed for all pipelines, to determine model stability in terms of their tendency to use the same features across all bootstrap resamples. This also allowed examination of what mRNA material was deemed relevant for BCR prediction.

4.3 Results

Results from the quantitative analysis of the 16 modelling pipelines are presented hereafter with respect to each assessment criterion successively. For discrimination, the model outputs were compared by way of multiple pairwise

comparisons using Wilcoxon signed-rank tests which are a nonparametric alternative to the t-test that allows for comparison of the centrality of the distributions [39]. 5% significance, after p-value adjustment for false discovery rate (FDR) was used as the level of significance [40]. These are the p-values discussed hereafter. For the predictive performance of the most discriminating pipelines, the AUC was compared by way of multiple pairwise comparisons using DeLong tests also at 5% significance, after p-value adjustment for FDR.

4.3.1 Discrimination

Bootstrap C-indices and CIs for each model are shown in Table 4.1. The MSK model outperformed all other models using only clinical variables ($p < 0.025$). An important finding is that inclusion of mRNA variables significantly increased model discrimination performance for all modelling strategies ($p < 0.001$), compared to their baseline clinical counterpart and MSK. The gain obtained from inclusion of genetic information is observed in terms of the other performance aspects reported on hereafter.

Table 4.1 further shows that pipelines using univariate Cox feature selection yielded significantly higher discrimination compared to those using correlation-based filtering only ($p < 0.001$), and combining both filtering strategies was optimal for boosted Cox and RSF models ($p < 0.001$). Overall, RSF achieved the highest discrimination performance, yielding statistically significant improvement over all other models ($p < 0.001$). All further analyses for each modelling strategy were carried out on their most discriminating pipelines.

Table 4.1: Bootstrap-corrected model discrimination performance in terms of C-index and associated 95% bootstrap confidence intervals (CI). The best-performing pipeline for each modelling strategy is highlighted in bold.

Method (Pipeline)	Model	C-Index	CI
Clinical Variables Only	MSK	0.727	(0.644, 0.813)
	Cox	0.705	(0.631, 0.770)
	LASSO	0.708	(0.642, 0.760)
	Boosted	0.707	(0.629, 0.771)
	RSF	0.672	(0.606, 0.752)
Clinical Variables and Correlation Prefiltered mRNA variables	Cox	0.779	(0.440, 0.862)
	LASSO	0.669	(0.586, 9.753)
	Boosted	0.819	(0.748, 0.871)
	RSF	0.743	(0.677, 0.795)
Clinical Variables and Univariate Cox Feature Selected mRNA variables	**Cox**	**0.798**	**(0.710, 0.872)**
	LASSO	**0.819**	**(0.741, 0.892)**
	Boosted	0.798	(0.729, 0.859)
	RSF	0.767	(0.709, 0.826)
Clinical Variables and Correlation Prefiltered and Cox Feature Selected mRNA variables	Cox	0.788	(0.685, 0.853)
	LASSO	0.687	(0.500, 0.840)
	Boosted	**0.821**	**(0.748, 0.872)**
	RSF	**0.849**	**(0.770, 0.930)**

4.3.2 Calibration

The MSK nomogram reports on the probability of BCR-free survival at the 5-year horizon. All models were thus assessed for their calibration at 5 years and for overall BCR-free survival. Fig 4.1 shows calibration curves for the

MSK nomogram. These indicate that the model performs adequately for 5-year survival, but severely underestimates the probability of overall BCR-free survival. Fig 4.2 illustrates how the most discriminating model pipelines from the Discrimination section yielded similar 5-year calibration compared to the MSK, except the Cox model, which yielded under-prediction and over-prediction at lower and higher survival probabilities, respectively. For overall survival, all models outperformed MSK with the Cox model having the best calibration.

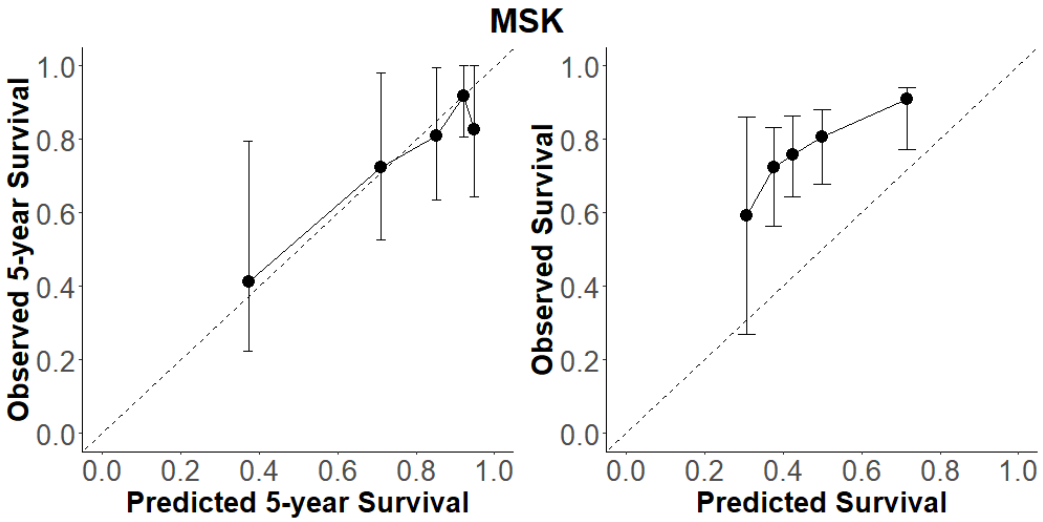

Figure 4.1: Calibration curves from MSK nomogram. Calibration curves for BCR-free survival estimation from the MSK nomogram at the 5-year horizon (left) and overall (right) with 95% confidence intervals.

4.3.3 Predictive Performance

ROC and DCA were carried out for 3-year and 5-year survival. The 3-year cut-off was chosen as it has been shown to be the time-frame with the highest recurrence rate [41], while the 5-year horizon was chosen in keeping with the current state-of-the-art cut-off [5]. ROC analysis output is shown in Fig 4.3

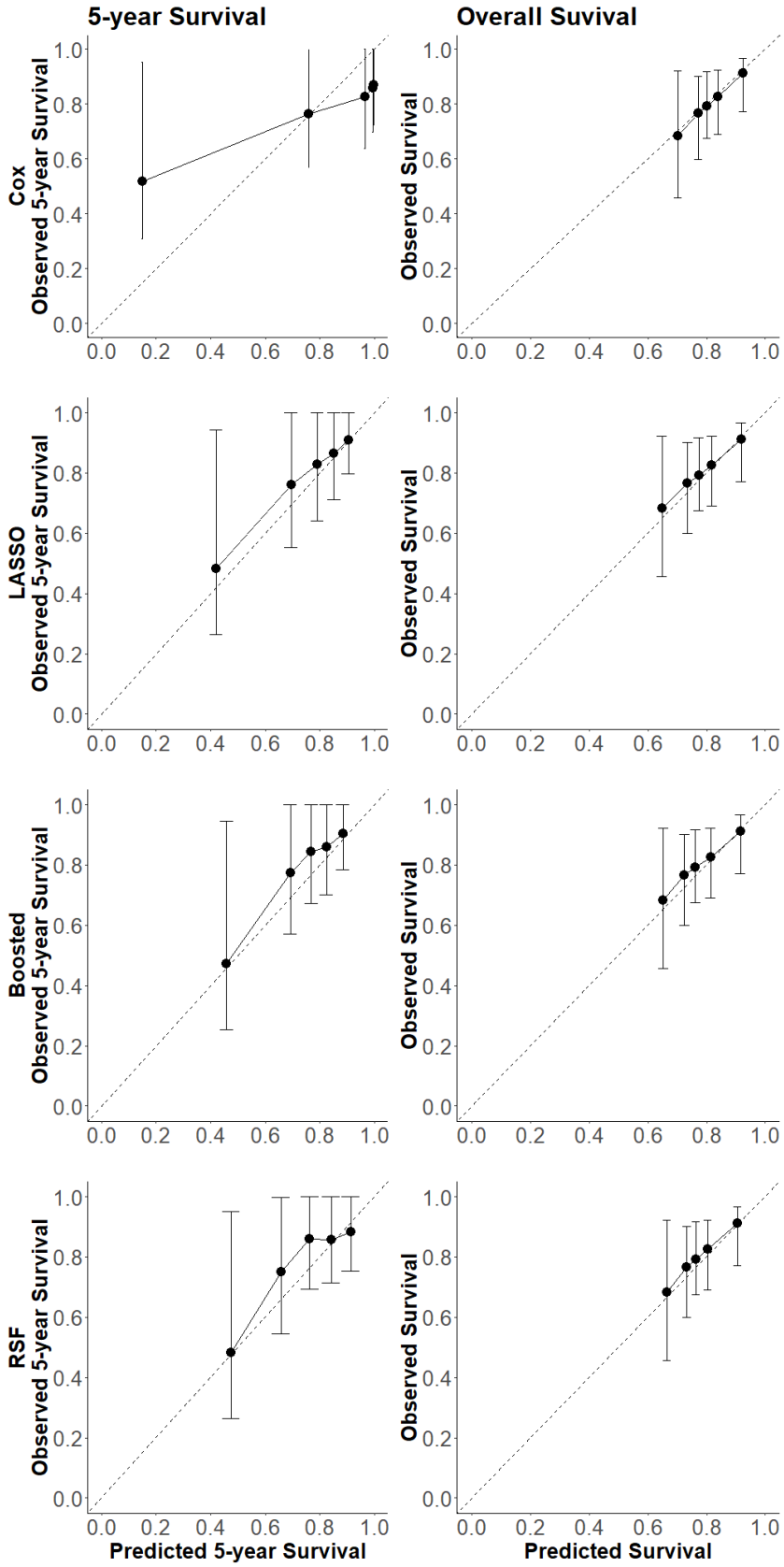

Figure 4.2: Model calibration curves. Model calibration curves at the 5-year horizon (left) and overall (right) BCR-free survival for the Cox, LASSO Cox, boosted Cox and RSF models from top to bottom, respectively with 95% confidence intervals.

and Table 4.3. The highest OOB prediction was achieved by RSF (AUC = 0.812, CI = (0.702, 0.921)) at the 3-year horizon, and boosted Cox at the 5-year horizon (AUC = 0.786, CI = (0.672, 0.899)). Using the DeLong test for comparing AUC, there was no statistically significant difference in the methods at the 3- or 5-year horizon. DCA output, shown in Fig 4.3, indicated that only the RSF model showed a continuous net benefit across all threshold probabilities for both the 3- and 5-year horizons. Although all models improved upon the "Treat All" method, no other methodologies showed benefit between the thresholds of 30% and 45% at 3 years. At 5 years, the LASSO Cox model retained some net benefit across most thresholds but was below that of the RSF model.

Table 4.2: OOB predictive performance in terms of AUC and associated 95% bootstrap confidence intervals (CI). 3-year and 5-year endpoints for the best-performing pipeline for each modelling strategy.

Endpoint	Model	AUC	CI
3-Years	Cox	0.804	(0.700, 0.907)
	LASSO	0.789	(0.684, 0.895)
	Boosted	0.794	(0.684, 0.905)
	RSF	0.812	(0.702, 0.921)
5-Years	Cox	0.781	(0.668, 0.894)
	LASSO	0.782	(0.669, 0.895)
	Boosted	0.786	(0.672, 0.899)
	RSF	0.779	(0.660, 0.899)

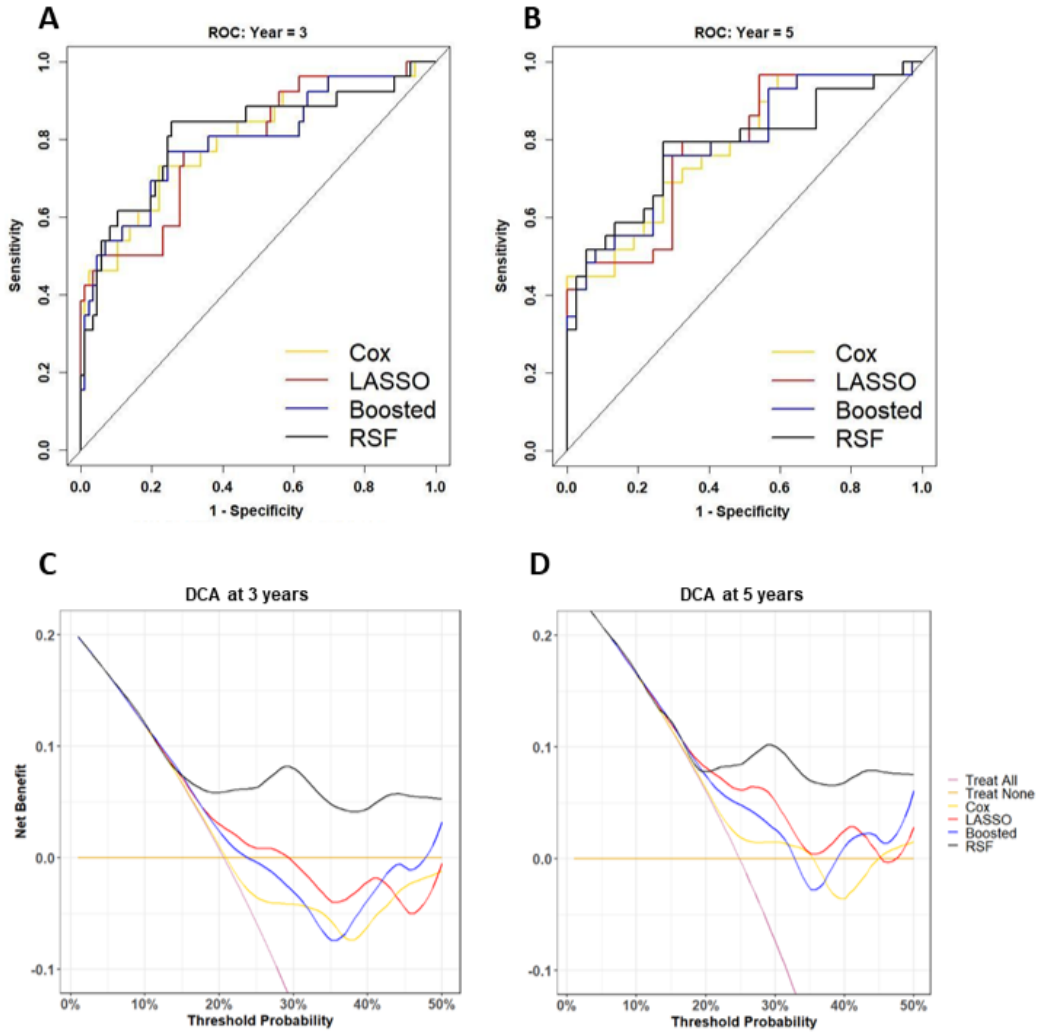

Figure 4.3: ROC and DCA. Model ROCs at the 3-year (A) and 5-year (B) BCR-free survival horizons for the models highlighted in bold in Table 4.1. Model DCAs at the 3-year (C) and 5-year (D) BCR-free survival horizons for the models highlighted in bold in Table 4.1.

4.3.4 Model Stability and Feature Selection

Stability of variable selection was assessed for all models that included mRNA variables. Table 4.3 summarises the most frequently selected variables for the

top performing models. Overall, PSA level was the only variable consistently picked by all models, but some of the mRNA candidates were also selected frequently. Boosted Cox models yielded the highest selection rates for mRNA features, i.e. the most stable pipeline. The otherwise relatively low selection rates may indicate possible information overlap or similarities in interactions between different groupings of mRNA features.

Table 4.3: In-bag variable selection rates (%) of the top 10 most frequently selected variables in the models highlighted in Table 4.1. Variables explicitly linked to genetic information found in at least 3 of these 4 lists are highlighted in bold.

Cox		LASSO		Boosted		RSF	
PSA	**70**	**PSA**	**57**	**ESM1**	**75**	**PSA**	**97**
Clinical stage	35	**ESM1**	**46**	**DNAH8**	**54**	Gleason sum	83
Gleason sum	34	**DNAH8**	**40**	**PSA**	**49**	Clinical stage	62
DNAH8	**28**	PI15	36	**ABCC11**	**47**	**DNAH8**	**45**
ABCC11	**24**	Clinical stage	32	CD38	37	DUSP18	38
ESM1	**18**	**ABCC11**	**31**	EFCAB4B	35	ARMC4	32
PI15	18	Gleason sum	22	FZD5	31	NEU3	32
SOCS2	12	XBP1	22	ZHX3	29	GRM8	31
Ethnicity	11	PLA2R1	17	FAP	28	ERO1LB	29
HELB	11	SPAG1	16	SLC16A14	27	FAM122C	29

The RSF model consistently selected PSA (97%), aggregate biopsy Gleason score (83%) and clinical stage (62%) along with mRNA variables including DNAH8, which was also highly selected in the other top modelling approaches. The predictive ability of DNAH8 in PCa has been observed previously for assessing poor prognosis [42].

mRNA variables which occurred most frequently across the top models also included ESM1 and ABCC11. Although it has not been seen to predict for PCa survival previously, ABCC11 has shown predictive ability in recurrence of colorectal cancer [43]. Increased levels of ESM1 were previously linked

to progression and development of metastasis [44]. Another mRNA variable selected in both the LASSO and Cox approaches was PI15 which has been identified as a biomarker for discrimination of metastatic progression [45].

PSA remained in the majority of the mRNA-including models, with above 49% selection rate for all the top models. Though it is considered a routine clinical variable as its testing is in the recommended guidelines [46], PSA is a protein produced from mRNA translation and can be considered a genetic biomarker itself [47]. Therefore, the most frequently used variables in our models for prediction of BCR were all genetic-based biomarkers.

As this analysis was undertaken without prefiltering of the mRNA variables with prior clinical knowledge of their association with BCR-Free survival or PCa, post-hoc analysis was undertaken to inspect the most frequently selected variables. Fig 4.4 highlights the mRNA variables in the top ten most frequently selected variables in the most discriminative models. None of the mRNA variables meet the commonly implemented criteria for the determination of differentially expressed genes (DEG), with respect to a fold change ≥ 2 and adjusted p-value < 0.05 [48]. PI15 was the only variable with an absolute fold change greater than 2 and only three of the top selected variables (CD38, FZD5 and HELB) had an adjusted p-value less than 0.05. In order to investigate the potential clinical significance of the selected variables, they were tested for association with clinical variables; namely, aggregate biopsy Gleason score and clinical stage via one-way ANOVA. FAP, GRM8, ESM1, PLA2R1, EFCAN4B, FZD5, PI15, and CD38 where all statistically significant in biopsy Gleason score (adjusted p-value < 0.05). None of the mRNA variables where statistically significant in clinical stage.

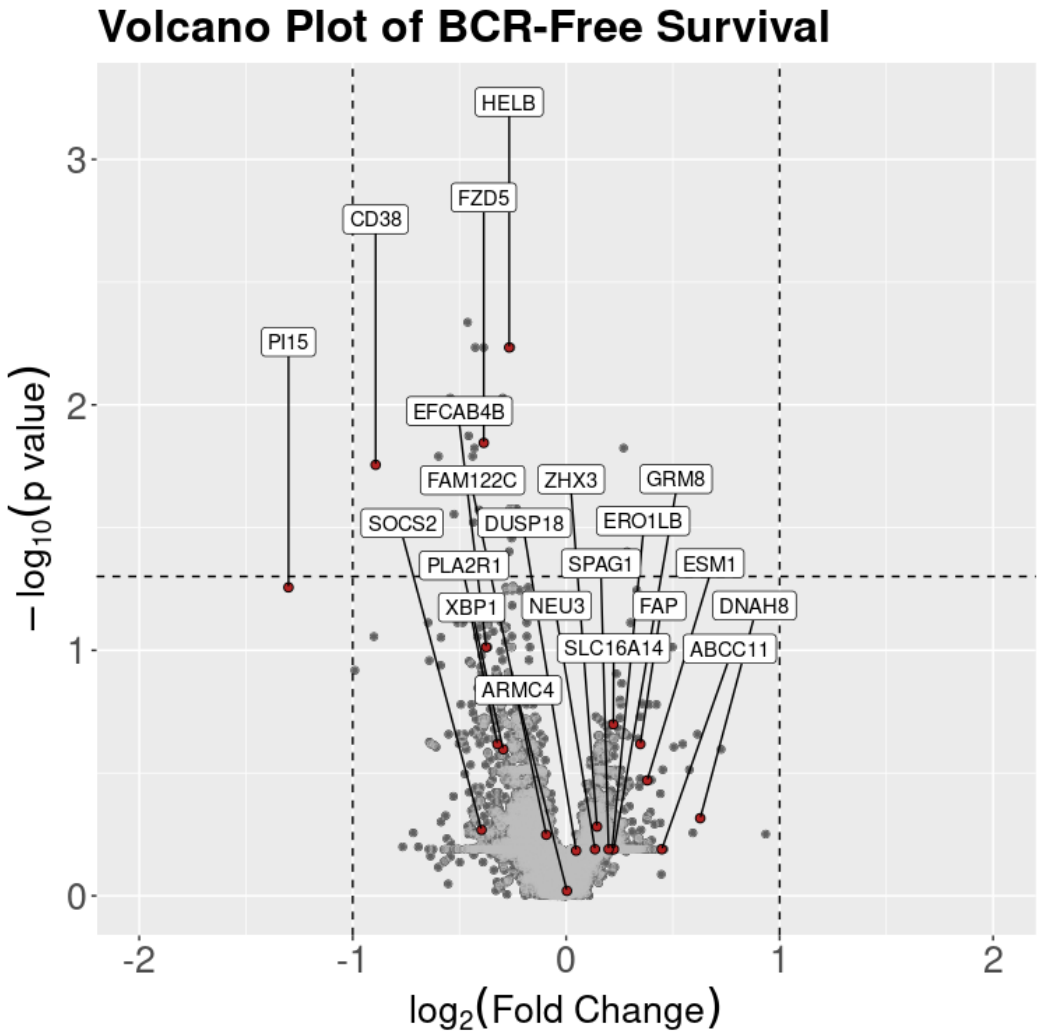

Figure 4.4: Volcano plot for BCR-Free survival-related differentially expressed genes (dashed lines indicating fold change ≥ 2 and adjusted p-value < 0.05).

4.4 Discussion

Although none of the ML models outperformed the MSK nomogram when using only pre-operative clinical variables, all yielded significant performance improvement when including mRNA variables compared to their baseline clinical counterpart and the MSK model. The RSF yielding best discriminatory

performance overall with good calibration as well as showing a continuous net benefit across all threshold probabilities. This demonstrates the potential of genetic information for prediction of BCR pre-operatively and is one of our key findings. The ability of the RSF to capture nonlinear associations and complex interactions may explain some of the observed gain in performance over Cox-based models. The models where also compared to the post-operative models from our previous study [9], and it was found that the best performing model pipelines which included both clinical and mRNA variables also yielded significant performance improvement over the post-operative MSK nomogram and post-operative models using only clinical variables ($p < 0.001$). The best performing pre-operative RSF model was comparable in predictive performance to its post-operative counterpart as well as the post-operative boosted Cox model ($p > 0.05$) and showed statistically significant improvement over the best performing post-operative Cox and LASSO models ($p < 0.001$).

To the best of our knowledge, this study is the first to propose mRNA-based models for time-to-BCR prediction that are applicable pre-operatively. Incorporating mRNA measurements in pre-operative assessment would require an extraction from tumour-biopsy samples. As these samples are already required for routine clinical diagnosis the additional workload would relate solely to the mRNA analysis which has become more cost-reasonable and feasible in recent years. Another alternative would be to use a liquid-biopsy analysis of the blood-based measurements of these mRNA variables. This would however require substantial additional research into whether these variables are present in blood and if their levels are relatable to those in a tumour sample. ML methodologies are needed to leverage the predictive capabilities of mRNA data and thus a shift is needed from conventional nomograms. Improving their interpretability will be key in enabling clinical inte-

gration of such alternatives.

DNAH8, ABCC11 and ESM1 were among the list of mRNA variables with an important contribution to at least some of the models considered. These biomarkers were found in other studies to have a role in various forms of risk characterisation for prostate cancer and/or other diseases. DNAH8 and ESM1 both had selection rate $\geq 50\%$ boosted Cox and RSF best-performing mRNA-inclusive pipelines in post-operative setting while ABCC11 had selection rate $\geq 35\%$. As the methodological approach implemented in this study did not undertake feature selection based on DEG investigation or association with clinical variables it was possible for previously uninvestigated variables to be selected and thus potentially new biomarkers for BCR-free survival prediction to be found. Though DNAH8 and ESM1 have previously been associated with prostate cancer they were not found to be BCR-free survival related DEGs in the post-hoc analysis and only ESM1 was found to be associated with a clinical variable (biopsy Gleason score). ABCC11 has not previously been found to have an association with prostate cancer and was also not a BCR-free survival related DEG, nor was in associated with the clinical variable in post-hoc analysis.

We aimed at assessing the potential for pre-operative BCR prediction, building upon the assumption that mRNA data used in the models is unlikely to have changed significantly between diagnosis and RP time-points, as was observed in several studies on both PCa or other cancers [11–13]. An aspect that may have the potential to alter the mRNA expressions is neoadjunctive therapy. Of the 135 patients with mRNA information, five received neoadjunctive hormone therapy; however, on investigation, their mRNA expression for the variables most frequently selected by the models appeared in line with those of other patients (Fig 4.5). Validation of the findings on mRNA data

acquired pre-operatively will be further explored in follow-on work.

The cohort size used in this study is relatively modest and comes from a single centre. This limitation has the potential of bias which may reduce the generalisability of the model to more diverse datasets however it also alleviates other challenges typically found in multi-centre datasets, such as heterogeneity in demographic representation and discrepancies in clinical protocols that are not unusual in cohort sizes commonly available for this kind of study. External validation of the findings on a larger multi-centre dataset or multiple single-centred datasets including more varied populations will be key for future development and general clinical application. Specifically, making the findings more generalisable.

4.5 Conclusion

This work demonstrated the potential of mRNA information (including in particular DNAH8, ABCC11, ESM1 and PI15) for improved pre-operative prediction of time-to-BCR in PCa, and is, to the best of our knowledge, the first to propose such models. Benchmarking of a number of machine learning methodologies that allow modelling of censored patient follow-up information indicated that RSF and boosted Cox models, in particular, were suitable candidates for the design of dedicated nomograms for this task. These implementations allowed leveraging relevant mRNA variables blindly from a large pool of features, without the reliance on previous findings about specific genetic biomarkers of interest. The pipelines that were implemented yielded promising performance on a relatively small cohort, in terms of model discrimination, calibration, and predictive performance of overall survival as well as 3- and 5-year survival, which are common current clinical endpoints. Following val-

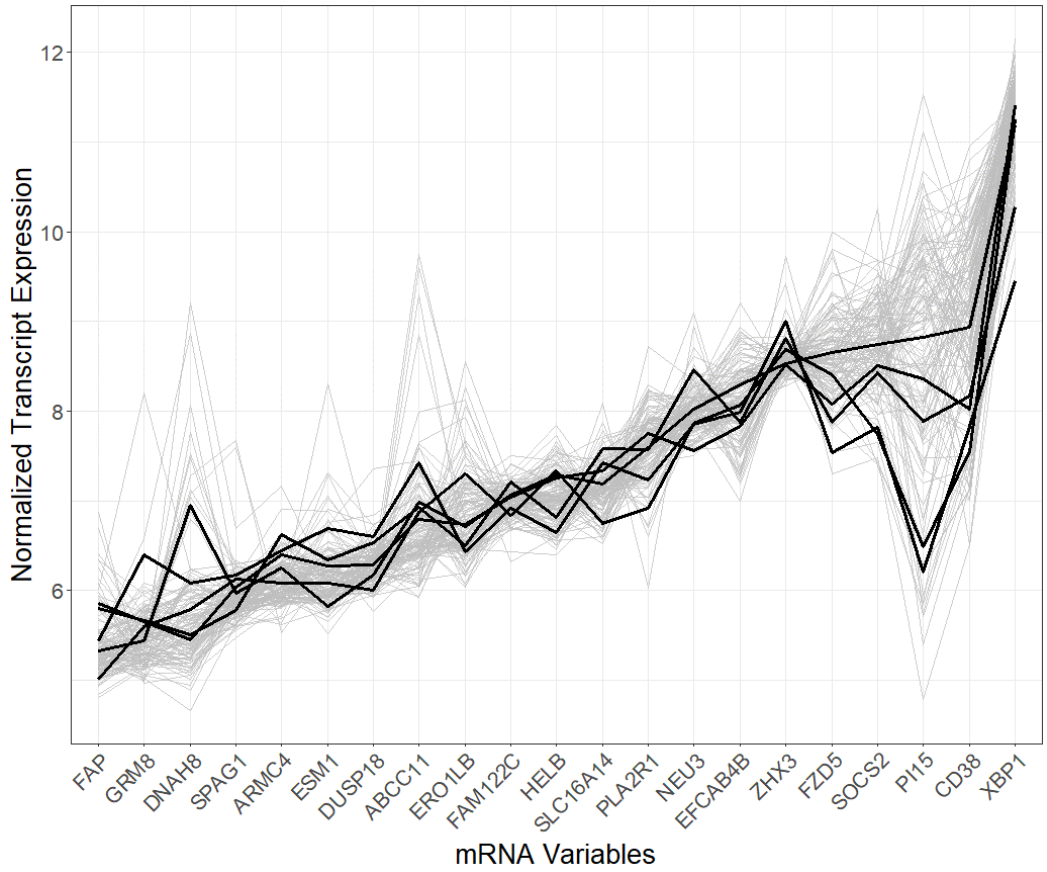

Figure 4.5: Normalised transcript expression levels. Normalised transcript expression levels for the most frequently selected mRNA variables (arranged with respect to increasing median value). Each line represents an individual patient's expression with the patients receiving no neoadjunctive therapy in grey, and those who receive neoadjunctive hormone therapy in black. There is no evident departure from the grey pattern observed in the neoadjuvant hormone therapy signatures.

idation of these findings these models could be implemented into a clinical setting using biopsy samples or potentially liquid-biopsy with the appropriate regulatory requirements.

4.6 References

[1] McClintock TR, Wang Y, Cole AP, Chung BI, Kibel AS, Chang SL, et al. Contemporary trends in the utilisation of radical prostatectomy. BJU international. 2018;122(5):726-8.

[2] Tourinho-Barbosa R, Srougi V, Nunes-Silva I, Baghdadi M, Rembeyo G, Eiffel SS, et al. Biochemical recurrence after radical prostatectomy: what does it mean? International braz j urol. 2018;44:14-21.

[3] Moghaddam S, Jalali A, O'Neill A, Murphy L, Gorman L, Reilly AM, et al. Integrating serum biomarkers into prediction models for biochemical recurrence following radical prostatectomy. Cancers. 2021;13(16):4162.

[4] Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. JNCI: Journal of the National Cancer Institute. 1998;90(10):766-71.

[5] Memorial Sloan Kettering Cancer Center. Prostate Cancer Nomograms. Memorial Sloan Kettering Cancer Center; 2022. `https://www.mskcc.org/nomograms/prostate`.

[6] Kryvenko ON, Epstein JI. Prostate cancer grading: a decade after the 2005 modified Gleason grading system. Archives of pathology & laboratory medicine. 2016;140(10):1140-52.

[7] Cookson MS, Aus G, Burnett AL, Canby-Hagino ED, D'Amico AV, Dmochowski RR, et al. Variation in the definition of biochemical recurrence in patients treated for localized prostate cancer: the American Urological

Association Prostate Guidelines for Localized Prostate Cancer Update Panel report and recommendations for a standard in the reporting of surgical outcomes. The Journal of urology. 2007;177(2):540-5.

[8] Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. The lancet oncology. 2011;12(3):245-55.

[9] O'Donnell A, Wolsztynski E, Cronin M, Moghaddam S. Improving the Post-Operative Prediction of BCR-Free Survival Time with mRNA Variables and Machine Learning. Cancers. 2023;15(4):1276.

[10] Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PloS one. 2013;8(6):e66855.

[11] Chen Y, Marotti JD, Jenson EG, Onega TL, Johnson KC, Christensen BC. Concordance of DNA methylation profiles between breast core biopsy and surgical excision specimens containing ductal carcinoma in situ (DCIS). Experimental and molecular pathology. 2017;103(1):78-83.

[12] Lee J, Lee EH, Park HY, Kim WW, Lee RK, Chae YS, et al. Efficacy of an RNA-based multigene assay with core needle biopsy samples for risk evaluation in hormone-positive early breast cancer. BMC cancer. 2019;19(1):1-8.

[13] Zanetti-Dällenbach R, Vuaroqueaux V, Wight E, Labuhn M, Singer G, Urban P, et al. Comparison of gene expression profiles in core biopsies and

corresponding surgical breast cancer samples. Breast Cancer Research. 2006;8(4):1-8.

[14] Cooperberg MR, Simko JP, Cowan JE, Reid JE, Djalilvand A, Bhatnagar S, et al. Validation of a cell-cycle progression gene panel to improve risk stratification in a contemporary prostatectomy cohort. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2013;31(11):1428-34.

[15] Knezevic D, Goddard AD, Natraj N, Cherbavaz DB, Clark-Langone KM, Snable J, et al. Analytical validation of the Oncotype DX prostate cancer assay–a clinical RT-PCR assay optimized for prostate needle biopsies. BMC genomics. 2013;14(1):1-12.

[16] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer cell. 2010;18(1):11-22.

[17] Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187-202.

[18] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of statistical software. 2011;39(5):1.

[19] Ridgeway G. The state of boosting. Computing science and statistics. 1999:172-81.

[20] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The annals of applied statistics. 2008;2(3):841-60.

[21] Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, et al. Improving breast cancer survival analysis through competition-based multidimensional modeling. PLoS computational biology. 2013;9(5):e1003047.

[22] Gilhodes J, Dalenc F, Gal J, Zemmour C, Leconte E, Boher JM, et al. Comparison of variable selection methods for time-to-event data in high-dimensional settings. Computational and Mathematical Methods in Medicine. 2020;2020(1):6795392.

[23] Wang H, Li G. Extreme learning machine Cox model for high-dimensional survival analysis. Statistics in medicine. 2019;38(12):2139-56.

[24] Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise reduction in speech processing. Springer; 2009. p. 1-4.

[25] Goh L, Song Q, Kasabov N. A novel feature selection method to improve classification of gene expression data. In: Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29; 2004. p. 161-6.

[26] Mason CH, Perreault Jr WD. Collinearity, power, and interpretation of multiple regression analysis. Journal of marketing research. 1991;28(3):268-80.

[27] Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature medicine. 2002;8(8):816-24.

[28] Van Wieringen WN, Kun D, Hampel R, Boulesteix AL. Survival prediction using gene expression data: a review and comparison. Computational statistics & data analysis. 2009;53(5):1590-603.

[29] Ishwaran H, Kogalur UB. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). manual; 2021. R package version 2.13.0. Available from: `https://cran.r-project.org/package=randomForestSRC`.

[30] Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B. mboost: Model-Based Boosting; 2022. R package version 2.9-7. Available from: `https://CRAN.R-project.org/package=mboost`.

[31] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1-22. Available from: `https://www.jstatsoft.org/v33/i01/`.

[32] Hastie T, R T, J F. The Elements of Statistical Learning. Springer; 2009.

[33] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2021. Available from: `https://www.R-project.org/`.

[34] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543-6.

[35] Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996;15(4):361-87.

[36] Royston P. Tools for checking calibration of a Cox model in external validation: prediction of population-averaged survival curves based on risk groups. The Stata Journal. 2015;15(1):275-91.

[37] Bewick V, Cheek L, Ball J. Statistics review 13: receiver operating characteristic curves. Critical care. 2004;8(6):1-5.

[38] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making. 2006;26(6):565-74.

[39] Holmes DT. Statistical methods in laboratory medicine. In: Contemporary Practice in Clinical Chemistry. Elsevier; 2020. p. 15-35.

[40] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289-300.

[41] Freedland SJ, Sutter ME, Dorey F, Aronson WJ. Defining the ideal cutpoint for determining PSA recurrence after radical prostatectomy. Urology. 2003;61(2):365-9.

[42] Wang Y, Ledet RJ, Imberg-Kazdan K, Logan SK, Garabedian MJ. Dynein axonemal heavy chain 8 promotes androgen receptor activity and associates with prostate cancer progression. Oncotarget. 2016;7(31):49268.

[43] Krizkova V, Dubova M, Susova S, Vycital O, Bruha J, Skala M, et al. Protein expression of ATP-binding cassette transporters ABCC10 and ABCC11 associates with survival of colorectal cancer patients. Cancer chemotherapy and pharmacology. 2016;78(3):595-603.

[44] Pan KF, Lee WJ, Chou CC, Yang YC, Chang YC, Chien MH, et al. Direct interaction of β-catenin with nuclear ESM1 supports stemness of metastatic prostate cancer. The EMBO journal. 2021;40(4):e105450.

[45] Zhao S, Geybels MS, Leonardson A, Rubicz R, Kolb S, Yan Q, et al. Epigenome-Wide Tumor DNA Methylation Profiling Identifies Novel Prognostic Biomarkers of Metastatic-Lethal Progression in Men Diagnosed with Clinically Localized Prostate CancerDNA Methylation Biomarkers and Prostate Cancer Prognosis. Clinical Cancer Research. 2017;23(1):311-9.

[46] Mottet N, van den Bergh RC, Briers E, Van den Broeck T, Cumberbatch MG, De Santis M, et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer—2020 update. Part 1: screening, diagnosis, and local treatment with curative intent. European urology. 2021;79(2):243-62.

[47] Kalfazade N, Kuskucu AM, Karadag S, Sahin S, Aras B, Midilli K, et al. Quantification of PSA mRNA levels in peripheral blood of patients with localized prostate adenocarcinoma before, during, and after radical prostatectomy by quantitative real-time PCR (qRT-PCR). International urology and nephrology. 2009;41(2):273-9.

[48] Chen J, Wang S, Tsai C, Lin C. Selection of differentially expressed genes in microarray data analysis. The pharmacogenomics journal. 2007;7(3):212-20.

Chapter 5

Discussion & Conclusion

This chapter provides a brief recapitulative summary of the overall research aims and contributions of this thesis. This is followed by a discussion on the nature and limitations of the work carried out in each of the three papers (i.e., Chapters 2, 3 and 4). We then provide an overview of further research carried out by the group at the periphery of these three chapters, and identify directions of future work. The chapter closes with an overall conclusion for this thesis.

5.1 Research Aims & Contributions

As outlined in the introduction of this thesis, our overarching goal was to determine the potential of machine learning methods for survival analysis in prostate cancer when mRNA variables are included, with the following specific aims.

1. Identify appropriate machine learning methods for survival analysis with the potential for improved performance compared to traditional statistical methods.

2. Determine whether the inclusion of mRNA expression information can improve predictive performance of biochemical recurrence in the **post-operative** setting.

3. Determine whether the inclusion of mRNA expression information has the potential to improve predictive performance of biochemical recurrence in the **pre-operative** setting.

The research aimed to contribute to the improvement of clinical decision-making in prostate cancer prognosis with timely precision medicine, as well as providing a concise and complete summary of the use of machine learning methods for survival analysis in cancer, their performance, and common pitfalls.

Chapter 2 presented a systematic review of the literature focused on addressing the first research aim, to identify machine learning methods used commonly for survival analysis in cancer studies and, from this, garner insight as to whether machine learning yielded the ability to outperform traditional models. A second objective of this review was to identify which methods, if any, showed superior performance. Despite the lack of consistency in reports across all studies reviewed, it was found that machine learning methods could outperform the Cox proportional hazard model in most cancer types, although with varied predictive performance. A sub-analysis of eleven studies which reported on prostate cancer yielded mixed findings (the CPH either outperformed or compared to ML methods in two of these papers (both SVM), but ML models outperformed CPH for all other comparisons). This low number of publications prevented further conclusions on the use of ML for survival analysis in PCa. The ML models in our own analyses were ultimately selected for their prevalence in the literature, their performance in previous studies, and the possibility to evaluate and interpret variable contribution in these models. They also all allow modelling of censored patient follow-up information.

The analysis of post-operative BCR prediction models in Chapter 3 was focused on addressing the second research aim while also considering the

first. The ML models considered (LASSO Cox, Boosted Cox and RSF) were compared to the CPH model and a state-of-the-art nomogram, the MSKCC model. This study demonstrated that the inclusion of mRNA variables increased predictive performance, with respect to the C-index, compared to models using only clinical variables, although the level of improved performance varied across modelling types. Two mRNA variables in particular, namely Dynein axonemal heavy chain 8 (DNAH8) and Endothelial Cell Specific Molecule 1 (ESM1), appeared consistently across the different models. Both of these have been noted in other reports for their association with prostate cancer. In our benchmark, CPH and RSF models using only clinical variables outperformed the reference MSKCC nomogram in terms of model discrimination and calibration. While the traditional CPH model with prefiltered mRNA variables demonstrated the best discrimination performance, its poorer calibration at specific endpoints and overall calibration may be suggestive overfitting to some of the sub-cohorts. Conversely, the best-performing RSF-based pipeline outperformed other modelling strategies with respect to discrimination and calibration for 2-year, 5-year, overall BCR-free survival prediction, and AUC at both 2-year and 5-year horizons. RSF also gave rise to the most consistent net benefit at these time-points in the decision curve analysis. Not all mRNA-inclusive pipelines yielded an increase in the predictive performance over their clinical-only counterpart. These results demonstrated that the model-building process is non-trivial and that not all ML pipelines can improve on traditional methods.

The analysis of Chapter 4 focused to pre-operative BCR prediction models, addressing the third research aim while again considering the first. This work demonstrated the potential of mRNA information (including in particular DNAH8, ABCC11, ESM1 and PI15) for improved pre-operative prediction

of time-to-BCR in PCa and is, to the best of our knowledge, the first publication to propose such pre-operative transcriptomics models. This approach relies upon the assumption that the mRNA information used in the models is unlikely to have changed significantly between biopsy and RP time-points. Our benchmark indicated that RSF and boosted Cox models, in particular, were suitable candidates for the design of dedicated nomograms for this task. RSF yielded the best discriminatory performance overall with good calibration, a continuous net benefit across all threshold probabilities, and a more stable selection of features. The implemented pipelines showed promising performance on a relatively small cohort in terms of model discrimination, calibration, and predictive performance of overall survival as well as 3- and 5-year survival, which are all common current clinical endpoints. Following validation of these findings, these models could be implemented into a clinical setting using biopsy samples.

5.2 Discussion

5.2.1 Systematic Review

The systematic review in Chapter 2 showed that although there was a lack of consistency in reporting and in study designs, machine learning methods could outperform the Cox proportional hazard model in most cancer types. However, in the specific analysis of prostate cancer studies, comparable studies found contradicting results. Thus, the question of the viability of machine learning methods for survival analysis in prostate cancer remains open.

Undertaking a systematic review of the application of machine learning for survival analysis in cancer contributes to the field beyond its initially outlined aims. First, given the broad scope of this study and a particular surge in the

topic in the last few years, the review provides a structured and comprehensive understanding of the current uses and impact of ML-based survival analysis in cancer. This is particularly novel in statistics, where systematic reviews are not commonly conducted.

A systematic review differs significantly from a traditional literature review's approach, rigour, and outcomes. While a literature review typically offers a broad overview of existing research on a topic, summarising and synthesising findings based on the author's selection of sources, a systematic review follows a predefined, transparent methodology designed to minimise bias and provide a comprehensive and objective synthesis of all relevant studies. This involves meticulously defining inclusion and exclusion criteria, systematically searching multiple databases, and critically appraising each study for quality and relevance. The benefits of a systematic review include greater reliability and reproducibility of findings, as it aims to provide a clear, evidence-based summary that can inform practice and guide future research. However, the systematic review process is more time-consuming and requires multiple reviewers, often making it less feasible for such topics with rapidly evolving research.

As the electronic databases were searched on the 24th of February, 2023, studies published after this date are not included. Additionally, its strict methodology can potentially exclude relevant studies that do not meet predefined criteria, again potentially overlooking emerging methods. Common to systematic reviews is the opportunity to extend the work with timely updates. Also, alternative search strategies, with changes to the search string structure, can be considered in future reviews that could help to expand on this and increase the scope of the findings.

This review is especially valuable because it identifies and addresses the

significant discrepancies in methodologies, reporting practices, and findings across the literature. These inconsistencies create challenges in comparing studies and drawing reliable conclusions about the efficacy and applicability of different machine learning techniques. By highlighting these discrepancies, the systematic review underscores the need for improved reporting standards and advocates for more rigorous and standardised comparative studies. This can drive future research toward more consistent and reliable outcomes, ultimately advancing the field. Unlike in randomised control trials, conducting a study on modelling methods is not as rigorously controlled, and until recently, there were no specific guidelines to follow. The ChAMAI checklist, developed in 2020, is a hugely beneficial tool not only for conducting quality reviews of machine learning methods in a medical setting but can also be used by researchers to ensure that their research is rigorous, well-reported and reproducible. However, as with all checklists, it cannot cover all aspects that should be considered. For example, although a model's stability could be regarded as an essential requirement for clinical usability, it is not a criterion on the checklist.

Moreover, in line with the study's primary aim, the systematic review showed that machine learning methods have the potential to outperform conventional statistical methods by analysing the small sub-cohort of studies that could be directly compared to one another despite broader discrepancies in the literature. This comparative analysis provides valuable insights into which machine learning approaches might be more effective in survival analysis for cancer, offering a foundation for future research and application. However, these findings must be considered in a broader context and are subject to potential bias. Though deep learning and multi-task models were found to have superior performance, conducting a formal meta-analysis to determine if the

perceived improvement was statistically significant across the studies was not possible. Thus, its output can be considered an exploratory analysis rather than definitive results. In summary, this systematic review is a crucial step toward consolidating knowledge, improving research practices, and guiding future work in the context of machine learning for survival analysis in cancer, but the question of best-performing methods remains undetermined.

5.2.2 Post-operative Prediction

The study of Chapter 3 showed that the ML methodologies considered can facilitate improved predictive performance post-prostatectomy for PCa patients, over current clinical nomograms. One of the primary strengths of the post-operative model is its potential for integration into routine clinical practice with relative ease. Since histological examination of surgically removed tumour samples is already routine, incorporating mRNA assays could be a feasible enhancement to clinical workflows. However, the feasibility of this addition is contingent on the balance between patient benefit and monetary and time costs. Although the cost of mRNA assaying has decreased over time, a thorough cost-benefit analysis is necessary to determine if it can be implemented routinely in clinical settings without overburdening healthcare resources. The increased net benefit shown in the decision curve analysis gives an insight into the clinical benefit to patients, however further assessment of the clinical relevance, following external validation would be necessary prior to implementation.

There are also notable opportunities for improving future research into the model-building pipelines. The study utilised correlation and univariate Cox proportional hazards models for pre-filtering and feature selection, which are widely recognised in the literature. However, alternative approaches could

offer improved performance. For instance, clustering or principal component analysis (PCA) could be employed for dimensionality reduction, potentially improving model performance by capturing more complex interactions between variables. The study opted not to use Principal Component Analysis (PCA) due to the loss in interpretability induced by the introduction of composite variables in the modelling. This could hinder the acceptance and implementation of the model in a clinical setting, where a clear understanding and communication of risk factors are essential. The decision to prioritise interpretability is motivated by the end goal but has the potential to limit the model's predictive power. Another consideration for future work in this area is the use of transformation for the clinical variables that are non-linear. PSA for instance is known to vary non-linearly with hazard and so in state-of-the-art models such as the MSKCC nomogram splines were used to adjust for this. There is thus potential for this or other transformations to improve the utility of PSA in a traditional clinical model and improve its predictive ability. Due to the nature of the follow-up occurring at regular intervals i.e. 3, 6, 12 months, there is potential for bias due to the true, unmeasured time-to-BCR being shorter than the recorded time. As the PSA measurement required for BCR evaluation requires a blood test it would likely not be feasible in a clinical setting to increase the frequency of testing both from a financial and patient welfare perspective. With regard to the need for external validation it is also noted that due to the relatively small sample size not all possible patient profiles are contained in the training data and though resampling techniques have need implemented to mitigate this to a point, a larger, more diverse sample will provide better evaluation of the entire patient population and thus more accurate evaluation of prediction performance.

Furthermore, another important consideration is that not all patients in

the original study had mRNA analysis. As described, by the authors [2] this was due to tumour size, i.e. some of the tumours were too small to allow for all four analyses in their study (array comparative genomic hybridization, mRNA, microRNA and sequencing) to be undertaken. Thus, patients with smaller tumour sizes may not be well represented in the model. A comparison of the distribution of the clinical variables was considered to determine potential bias in the missingness. The Mann-Whitney U test (Table A in appendix) was used for continuous variables, and χ^2 or Fisher's Exact Test, as appropriate, were used for categorical variables (Table B in appendix). As can be seen in Table B, only two of the variables, clinical and pathological stage, showed a statistically significant association with mRNA analysis. As the staging variables represent the size of the tumour this may seem to intuitively indicate a bias present based on the tumour size, as described by the original authors, i.e. patients with smaller tumours (or lower stage) are less likely to have had mRNA analysis. However, the distribution indicates that those without mRNA analysis where more likely to have a higher stage (larger tumour) which is not in keeping with the expected results. This may potentially indicate that other factors affected the decision of the authors to analyse the mRNA of some patients. This could imply some selection bias in the models produced in this work, again emphasising the need for external validation.

Lastly, there remains a need for external validation, particularly given the single-centre nature of the original dataset. External validation is necessary for the generalisability of the findings, as the model may perform differently in varied clinical settings or populations. Multi-centre studies and external datasets are necessary to ensure the robustness and applicability of the predictive model across diverse patient groups. However, it is worth noting that

the use of a single-centred study mitigates potential issues, such as differences in clinical practice influencing the results. Also, as mRNA assaying methods can vary, the combination of multiple datasets for the initial generation of the models would require that all use the same techniques and include the same variables or the study's findings would be further limited.

5.2.3 Pre-operative Prediction

Although the consideration of the time-point of evaluation is not novel, as evidenced by studies from Kattan and colleagues published in the late nineties and early two-thousands, it is not so common in the literature, despite its importance for clinical application. The post-operative approach considered in Chapter 3 provides a promising predictive methodology for patients who must undergo prostatectomy, with clear improvement over current clinical nomograms. The pre-operative approach considered in Chapter 4 allows for prediction before prostatectomy, which may ultimately provide timely information for earlier decision-making in clinical settings. This aligns with the clinical need for early and accurate individualised treatment of prostate cancer.

This study offers a significant contribution to the research field in that sense and the potential to significantly impact the treatment decision-making process. The ability to predict biochemical recurrence before surgery could help identify patients who may not require radical prostatectomy, thereby reducing the number of unnecessary surgeries. This could not only enhance survival rates by optimising treatment strategies but also improve patients' overall quality of life by minimising the risk of surgical complications and preserving normal physiological functions.

One of the key strengths of this study, like the post-operative study, is its

feasibility. Since biopsy is already a routine procedure in prostate cancer diagnosis, and histological assessments are commonly performed, incorporating mRNA analysis at this stage is practical and could be integrated into existing clinical workflows. However, the same cost-benefit analysis would be necessary for its incorporation. The ability to apply such models pre-operatively ensures that critical decisions are made earlier in the disease course, potentially leading to more effective and targeted therapies; thus, the cost-benefit may be more effective at this time-point than in the post-operative setting. Again, the DCA gives an insight into the clinical benefit of the models but further validation would be required prior to implementation.

Of note, any pre-operative prediction model for BCR following prostatectomy may only be trained on patients who actually underwent RP, and is thus subject to survivorship bias as such models are inherently blind to patients who, though eligible for RP, do not undergo surgical intervention for one reason or another (reason which may associate with the level of risk for such patients). This is true not only for the models produced in this study but for current state-of-the-art methods as well. This thus limits the use of the pre-operative model to such patients and would be inappropriate for treatment decision-making in other cohorts for whom prostatectomy is not being considered. Additionally, such models are inherently blind to patients, though eligible for RP, do not undergo surgical intervention for one reason or another in model building.

However, the conclusions of this study rest on the critical assumption that mRNA expression levels in trans-operatively acquired tumour tissue samples are representative of those in pre-operative biopsy samples. Since mRNA data from biopsy samples were unavailable for the cohort studied, this assumption could not be directly validated, posing a limitation. Although post-hoc anal-

ysis suggested that mRNA expression in patients who received neoadjuvant treatment was comparable to those who did not, this observation can only suggest, rather than conclusively prove, that mRNA expression is unaffected by the treatment. Moreover, even if statistical comparisons indicated similarities within this subset, it would not definitively confirm that pre-operative mRNA expression levels are consistent with post-operative ones across all patients. Consideration should also be given to the possible evolution of the tumour between the time of biopsy and the time of surgical intervention. The longer this period is, the greater a significant difference in the mRNA expression is likely to be. This was not possible to evaluate with the dataset used and thus consolidates the need for further investigation. Validation of the results of the pre-operative model with mRNA data acquired pre-operatively is necessary. Therefore, while the study provides promising results, further research is necessary to validate the assumption underpinning its predictive model to ensure its reliability and generalisability in clinical practice.

5.3 Further & Future Research

To inform this future work, some preliminary research undertaken during this PhD is described here, which has not been published in any previously discussed work.

5.3.1 Pre-operative Genetic Biomarkers

Incorporating mRNA measurements in pre-operative assessment would require an extraction from tumour-biopsy samples. This research group explored opportunities to access data of this type for this purpose during this PhD project. Unfortunately, biopsy mRNA data from most prostate cancer

studies tend to be held in privately owned databases, and our efforts thus far have not been successful. As such, validation on pre-operative mRNA data remains a direction of future work.

As mentioned in the previous section, tumour-biopsy samples are already required for routine clinical diagnosis. Thus, the additional workload would relate solely to the mRNA assay, which has become more cost-reasonable and feasible in recent years. However, another alternative would be to use a liquid-biopsy analysis of the blood-based measurements of these mRNA variables. This would, however, require substantial additional research into whether these variables are present in the blood and if their levels are relatable to those in a tumour sample.

During this PhD project, research was undertaken into the feasibility of the liquid-biopsy data[1] acquired from a collaboration with University College Dublin. This Irish cohort consisted of 271 patients with both clinical variables (including biopsy Gleason Score, DRE and PSA) and six serum biomarkers (CD14, IGFBP3, APOAII, GC Globulin, ZAG and PEDF). In an effort to predict BCR-free survival, a number of modelling frameworks similar to those used in Chapters 3 and 4 were implemented, with the additional inclusion of two other regularisation techniques (namely, ridge and elastic net penalisations of the CPH model). RSF was found to be comparable to CPH at baseline (p-value = 0.5964), while all other models showed statistically significant improvement (p-values < 0.001). Models with biomarkers alone

[1]The cohort was collected with funding from the Irish Cancer Society (PCI11WAT) for the Prostate Cancer Research Consortium biocollection and biomarkers measured with funding of Science Foundation Ireland (SFI) under Grant Number 15/IA/3104. Ethical approval for sample collection and processing were from the IRB at the Mater Misericordiae University Hospital (1/378/660) and samples analysis by University College Dublin ethics committee (LS-E-11-39).

were not able to improve on models using only clinical variables. Models with clinical variables and the PEDF biomarker performed better than when all six biomarkers were included, in keeping with previous findings by Moghaddam et al. [1]. Ridge and boosted Cox models showed the best performance when biomarkers were included. These findings again led to the conclusion that machine learning models leveraging genetic information have the potential to improve on traditional CPH methods, this time in a low-dimensional setting.

5.3.2 Dimensionality Reduction

High dimensionality in the mRNA expression environment is one of the main reasons for the consideration of ML in prognostic modelling. Consideration of alternative dimensionality-reduction techniques, such as principal component analysis (PCA) prior to the implementation of a traditional statistical CPH model, could alleviate limitations relating to data size without the need for ML to model survival. As mentioned in the previous section, such dimensionality reduction techniques were not implemented in this work as a choice based on their lack of transparency, which renders their implementation in a clinical setting very challenging. Another way to reduce the number of features being considered could be to include features previously found to be associated with poor prognosis or known to be involved in the cancer tumour progression process. This type of hand-selection has several potential issues, particularly concerning selection bias, but particularly limits the possibility of finding novel biomarkers.

During this PhD project, research was undertaken to explore the potential benefit of mRNA variables deemed independently to be predictors of poor prognosis of prostate cancer in a multivariate context. Twenty mRNA variables were investigated based on a literature review of seventeen studies of

a single genetic attribute for its prognostic potential, along with the three variables found by Taylor et al., 2010 in the original investigation of this dataset [2]. There was a lack of consistency in the definition of 'poor prognosis' across these studies, with some not providing any definition. For the purpose of consistency in evaluation, the definition of poor prognosis was considered as recurrence in the two years post-RP in this investigation. Thus, logistic regression was used as the primary modelling approach. Models were created using a 5-fold cross-validation framework. A clinical model was built to distinguish a baseline analysis including PSA, clinical stage and biopsy Gleason score.

A model that included each mRNA variable univariately added to this clinical model was then produced to show their added benefit. A multivariable model was built, including all mRNA variables with a correlation threshold of 0.9 (to alleviate multiple collinearity issues). Finally, unsupervised PCA was performed to account for the potential polygenicity, and another model was built from these new scores. The clinical variable model produced 88% test accuracy. The mRNA variables were then added individually; however, none were found to improve model performance statistically significantly. The multivariate model with all 20 variables had lower cross-validated accuracy than the clinical model (0.79 vs. 0.88, respectively). A model with clinical variables and the first three principal components yielded 92% test accuracy. It was clear from this research that the inclusion of genetic attributes such as mRNA variables can improve the prediction of poor prognosis prostate cancer when compared to clinical variables alone. However, the consideration of single variables in isolation does not add benefit to predictive capabilities. Multivariable models can improve on this, but collinearity may dampen the effects, thus showing the benefit of PCA. PCA has been shown here to help

to negate this and has been shown to have the highest test accuracy. The success of this approach also indicates the polygenicity of prostate cancer, thus compounding the inaccuracy of single gene investigation or the off-hand use of pre-determined 'poor prognosis' predictors without proper investigation.

5.3.3 Beyond Biochemical Recurrence

A final element for future consideration revolves around the concept that a single endpoint such as BCR-free survival may not be the most appropriate, or at least not the only outcome of interest. The win-ratio [3], initially developed by Pocock et al. in 2012, is a valuable metric in the context of clinical trials, especially when multiple endpoints of interest can be hierarchically arranged to assess overall effect of the treatment. Clinical trials often involve the evaluation of diverse endpoints, such as survival rates, disease progression, and treatment response, each representing distinct facets of a treatment's efficacy. By calculating the win-ratio, researchers can systematically weight the relative importance of these endpoints, effectively capturing the comprehensive impact of a drug on patient outcomes. This approach enables a nuanced evaluation that aligns with the hierarchical structure of endpoints, ensuring that the primary objective, typically a critical clinical endpoint like overall survival, retains its prominence while still acknowledging the relevance of secondary endpoints. In doing so, the win-ratio approach provides a holistic perspective on the treatment's performance, aiding in informed decision-making and enhancing the interpretability of clinical trial results.

Current applications of the win-ratio in the literature are somewhat limited, as this is still a relatively novel method in clinical trial design. The results of a Cochrane Library search for win-ratio in randomised controlled trials (RCT) showed that of a total of 33 studies deemed eligible for inclusion,

25 were in a cardiovascular disease setting, five of which stemmed from one larger trial (EMPULSE trial). As the cardiovascular setting was the original application of the win-ratio, it was expected to predominate. Six studies investigated COVID-19, and the remaining two trials focused on stroke and kidney transplants, respectively. Of interest is the lack of studies in a cancer setting, as the types of endpoints of interest in cancer trials are well suited to the win-ratio methodology. During this PhD project, work was undertaken to create an R package for sample size and power calculation for win-ratio analyses. The developed `WRestimates` package provides easy-to-use functions that can be used in the required sample size determination and power of studies implementing the win-ratio. These allow for the calculation of sample size and power based on estimands, negating the need for complex simulation-based methods, which require many assumptions to be made of the data, thus enabling use of the win-ratio in a wider range of studies in the future. Also, during this PhD project, the authors of several recent clinical trials in prostate cancer were contacted to request access to their data for evaluation of the potential of the win-ratio in their assessment, but no response was received. Thus, further investigation is yet to be carried out.

5.4 Conclusion

This thesis demonstrates that integrating mRNA expression data through the application of machine learning methods for survival analysis improves the prediction of biochemical recurrence (BCR) in prostate cancer. The systematic review underscored the potential of machine learning in survival analysis, highlighting the need for further validation and consistency across studies. Both post-operative and pre-operative analyses showed enhanced prediction

accuracy using mRNA variables and machine learning, which could lead to more personalised and timely treatment decisions. Future research should focus on external validation, alternative dimensionality reduction techniques, and developing more interpretable models. Overall, this work advances the use of machine learning in prostate cancer prognosis and outlines critical areas for further investigation.

5.5 References

[1] Moghaddam S, Jalali A, O'Neill A, Murphy L, Gorman L, Reilly AM, et al. Integrating serum biomarkers into prediction models for biochemical recurrence following radical prostatectomy. Cancers. 2021;13(16):4162.

[2] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer cell. 2010;18(1):11-22.

[3] Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. European heart journal. 2012;33(2):176-82.

Appendix A

Table A: Comparison of the distribution of continuous clinical variables between patients with and without mRNA analysis.

Variable	**Mann-Whitney U**	**p-value**
PSA	3206	0.359
Age	3356	0.642

Table B: Comparison of the distribution of categorical clinical variables between patients with and without mRNA analysis. The total number of patients without mRNA analysis = 52, total number of patients with mRNA analysis = 135.

Variable	**# No mRNA (%)**	**# mRNA (%)**	χ^2 **Test/ Fisher's Exact Test p-value**
Biopsy Gleason Grade			0.372
6	25 (48.1%)	76 (56.3%)	
7	18 (34.6%)	46 (34.1%)	
8	7 (13.5%)	8 (5.9%)	
9	2 (3.8%)	5 (3.7%)	
Clinical Stage			0.004
1	19 (36.5%)	77 (57.0%)	
2	25 (48.1%)	53 (39.3%)	
3	8 (15.4%)	5 (3.7%)	
Pathological Gleason Grade			0.276
6	12 (23.1%)	40 (29.6%)	
7	27 (51.9%)	76 (56.3%)	
8	6 (11.5%)	11 (8.1%)	
9	7 (13.5%)	8 (5.9%)	
Pathological Stage			0.004
2	23 (44.2%)	85 (63.0%)	
3	19 (36.5%)	43 (31.9%)	
4	10 (19.2%)	7 (5.2%)	
Surgical Marginal Status (SMS)			0.143
Negative	34 (65.4%)	103 (76.3%)	
Positive	18 (34.6%)	32 (23.7%)	
Extracapsular Extension (ECE)			0.274
Absent	11 (21.2%)	41 (30.4%)	
Present	41 (78.8%)	94 (69.6%)	
Seminal Vesicle Invasion (SVI)			0.077
Negative	39 (75.0%)	117 (86.7%)	
Positive	13 (25.0%)	18 (13.3%)	
Lymph Node Involvement (LNI)			0.081
Abnormal	9 (17.3%)	10 (7.4%)	
Normal	38 (73.1%)	102 (75.6%)	
Not Assessed	5 (9.6%)	23 (17.0%)	