Application of machine learning techniques yields improvements in the predictive ability of urine biomarker panels for prostate cancer; analysis of the Movember GAP1 Urine Biomarker project

By

Shea P. O'Connell

Norwich Medical School
University of East Anglia

A thesis submitted to Norwich Medical School at the University of East Anglia in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

JANUARY 2021

Abstract

Prostate cancer is a considerable clinical problem worldwide, with large amounts of variation seen in the clinical outcome of patients with apparently similar disease. The diagnostic and prognostic tool-sets currently available to clinicians lack both sensitivity and specificity, not taking into account the molecular variability of the disease. The successful development of non-invasive prognostic biomarker tests has the potential to impact the large numbers of patients with a clinical suspicion of prostate cancer but that ultimately do not require invasive investigation and stressful follow-up.

The Movember Global Action Plan 1 (GAP1) Urine Biomarker Consortium had the aim of developing of a muti-modal urine test for the accurate discrimination of disease status. The consortium of 12 collaborating institutes collected 1,258 urine samples that were subsequently assayed by a range of biochemical techniques. The main aim of this thesis was to apply statistical learning techniques to these data in order to robustly develop prognostic models for prostate cancer.

The Prostate Urine Risk (PUR) model was developed using solely NanoString data from cell-free RNA samples, and reported strong utility for predicting the outcome of an initial prostate biopsy (AUCs > 0.70 for Gleason $\geq$ 3+4 and $\geq$ 4+3). Additionally displaying remarkable use in an active surveillance sub-cohort, PUR identified patients at a higher apparent risk of disease progression, reporting a hazard ratio = 8.23 (95% CI: 3.26 - 20.81).

The effects of altering the statistical methodology applied to the data were quantified, where ensemble algorithms presented the best solution to capturing the most amount of information. Using this information a machine learning framework was designed to produce multivariable risk prediction models incorporating strong internal validation compliant with the TRIPOD reporting guidelines.

This framework was used to construct three risk models, each integrating information from different fractions of urine. All showed strong potential for clinical utility, reporting AUCs in excess of 0.8 for predicting Gleason $\geq$ 3+4, and approaching AUC = 0.9 for ruling out the presence of any cancer on biopsy. The net benefit of adopting these risk models was determined via simulation of a population-level cohort, where each model has the potential to result in large reductions to the numbers of unnecessary biopsies currently undertaken.

In conclusion, the analyses presented here demonstrate the large amount of information that can be captured within urine. If these models are validated in future studies by the proposed clinical trial designs they could dramatically change the treatment pathway for prostate cancer, reducing costs to healthcare systems and ultimately unnecessary stress to patients.

Acknowledgements

I would like to thank my supervision team, Dr. Daniel Brewer, Dr. Jeremy Clark, Professor Colin Cooper and Mr Robert Mills, without whom I would not have been able to complete the work in this thesis.

My particular thanks go to Dan, who was always supportive in the face of seemingly-endless negative results and difficult data conversations, as well as providing countless hours of feedback on manuscripts and thesis chapters. Jeremy was very helpful in providing a huge amount of insight into the biology of prostate cancer and the clinical pathway of patients with cancer.

I am indebted to the 700+ donors to the Andy Ripley Memorial fund, without which I would not have had the financial support to undertake this research. Finally, I am forever grateful to my partner Lisa Gibson, who has always been there for me throughout the bumpy journey of a PhD and in navigating the world of academia.

Contributorship statements

Shea P. Connell, Marcel Hanna, Frank McCarthy, Rachel Hurst, Martyn Webb, Helen Curley, Helen Walker, Rob Mills, Richard Y. Ball, Martin G. Sanda, Kathryn L. Pellegrini, Dattatraya Patil, Antoinette S. Perry, Jack Schalken, Hardev Pandha, Hayley Whitaker, Nening Dennis, Christine Stuttle, Ian G. Mills, Ingrid Guldvik, Chris Parker, Daniel S. Brewer, Colin S. Cooper, Jeremy Clark (2020) *A Four-Group Urine Risk Classifier for Predicting Outcome in Prostate Cancer Patients*, BJUI, May 20;124(4):609-620. doi: 10.1111/bju.14811

SPC, JC, CSC & DSB drafted the manuscript. SPC, and DSB performed the statistical analysis . MH, FC, RM, and CP setup clinical collection and developed clinical methodology. MGS, ASP, JS, HP, HW & IGM all conceived gene-probes for NanoString interrogation. RH, MW, HC, HW, KLP, ASP, ND, MGS, CP & CS were involved in sample collection, extraction and preparation at their respective institutes. RWB oversaw histopathological analysis of biopsies. CSC, JC, JS, FC, and MH, IG, SPC, MH, FC, ND, CS, CP, DSB conceived and designed the studies. SPC, HC, DP, and DSB performed NanoString data analyses. DSB, CSC & JC had joint and equal contributions to senior authorship. All authors read and approved the manuscript. All authors critiqued the manuscript for intellectual content.

Shea P. Connell, Eve O'Reilly, Alexandra Tuzova, Martyn Webb, Rachel Hurst, Robert Mills, Fang Zhao, Bharati Bapat, Colin S. Cooper, Antoinette S. Perry, Jeremy Clark, Daniel S. Brewer (2020) *Development of a multivariable risk model integrating urinary cell DNA methylation & cell-free RNA data for the detection of significant prostate cancer*, The Prostate, 2020 (1 - 12). `doi:10.1002/pros.23968`

SPC drafted the manuscript and conceived, designed, and performed the statistical analyses. ER, AT, FZ, and BB were involved in sample collection and methylation analyses at their respective institutes. MW, RH, and RM were involved in sample collection and NanoString analyses as well as development of clinical methodologies. DSB, JC, ASP, and CSC had joint and equal contributions to senior authorship and were contributors in writing the manuscript. All authors read and approved the manuscript. All authors critiqued the manuscript for intellectual content.

Shea P. Connell, Robert Mills, Movember GAP1 Urine Biomarker Consortium, Hardev Pandha, Richard Morgan, Jeremy Clark, Colin S. Cooper, Daniel S. Brewer (2020) *Integration of urinary EN2 and cell-free RNA in developing a multivariable risk model for the detection of prostate cancer in biopsy naieve patients*, TBD

SPC drafted the manuscript and conceived, designed, and performed the statistical analyses. HP and RM were involved in sample collection and ELISA analyses at their respective institutes. JC and RM were involved in sample collection and NanoString analyses as well as development of clinical methodologies. DSB, JC, RM, HP, and CSC had joint and equal contributions to senior authorship and were contributors in writing the manuscript. CSC, JC, HP and DSB provided the original idea for this study. All authors read and approved the manuscript. All authors critiqued the manuscript for intellectual content.

Shea P. Connell, Maria Frantzi, Agnieszka Latosinska, Martyn Webb, William Mullen, Martin Pejchinovski, Harald Mischak, Mark Salji, Colin S. Cooper, Jeremy Clark, Daniel S. Brewer (2020) *Development of a risk model integrating cell-free RNA & proteomic data for the pre-biopsy detection of prostate cancer from urine*, TBD

CSC, JC and DSB proposed and implemented the original concept for this multi-omics study as part of GAP1 Movember study. SPC conceived and designed the statistical analyses; HM, WM, AL have developed the background methodology for proteomics data acquisition and post-analytical processing; MF, MP acquired the proteomics data by CE-MS and MW, MS were involved in sample collection and laboratory analyses of the RNA samples; SPC performed the statistical analyses; MF and SPC drafted the manuscript. HM, CSC, JC and DSB had joint and equal contributions to senior authorship and were contributors in writing the manuscript. All authors read and approved the manuscript. All authors critiqued the manuscript for intellectual content.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other university or other institute of learning.

Shea Peter Connell
September 2020

Wordcount: 43,663

Acronyms

AS - Active Surveillance
AUA - American Urological Association
AUC - Area Under the ROC Curve
BPH - Benign prostatic hyperplasia
CAPRA - The UCSF cancer of the prostate risk assessment
CE-MS - Capillary electrophoresis-mass spectrometry
CI - Confidence interval
CRO - Clinical research organisation
DCA - Decision Curve Analysis
DRE - Digital rectal examination
EAU - European Association of Urology
ELISA - Enzyme-linked immunosorbent assay
EPI - ExoDX Prostate (IntelliScore)
ERSPC - European randomized study of screening for prostate cancer
GAP1 - Global Action Plan 1
GBM - Gradient boosting machine
GP - General practitioner
Gs - Gleason Score
HR - Hazard Ratio
KM - Kaplan-Meier estimator
LASSO - Least absolute shrinkage and selection operator
LUTS - Lower urinary tract symptoms
MZSF - Maximum Z score amongst the shadow features
NEC - No Evidence of Cancer
NHST - Null-hypothesis significance testing
NICE - National Institute of Clinical Excellence
NIM - Normalised index of methylation
NNUH - Norfolk and Norwich University Hospital
NPV - Negative predictive value
OOB - Out-of-bag
OR - Odds Ratio
PCPTRC - Prostate Cancer Preventional Trial Risk Calculator
PH - Proportional Hazards
PHI - Prostate Health Index
PI-RADS - Prostate imaging-reporting and data system
PPV - Positive predictive value
PSA - Prostate specific antigen

PUR - Prostate Urine Risk
RCT - Randomised Control Trial
ROC - Receiver Operator characteristic
RP - Radical prostatectomy
SJH - St. James's Hospital
TNM - TNM Classification of Malignant Tumors
TPM - Trans-perineal mapping
TRIPOD - Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
TRUS - Trans-rectal ultrasound
UCL - University College London
UTI - Urinary tract infection

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cancers form a group of diseases characterised by abnormal cell growth with the potential to invade or spread to other parts of the body. Cancer arises from multiple acquired heritable genetic mutations that drive disease progression. Typically, cancer is a disease of old age, though some cancers such a leukaemia or brain cancers are particularly prevalent in young children[1]. Cancer requires multiple genetic and epigenetic alterations to be acquired before a cell can escape growth regulation and proliferate uncontrollably, invading surrounding and distant tissues, disrupting the body's basic functions, and potentially causing death.

Prostate cancer is remarkably common in Western society, accounting for 26.3% of all male cancers diagnosed in the United Kingdom in 2015, more than any other single cancer[1]. It is so common in fact, that autopsy studies have shown detectable prostate cancer is present in 24 - 40% of men at the time of their death[2,3]. We are still unsure as to why prostate cancer is so common and what causes it to appear in so many men before their deaths, the prostate has even been described as an inherently precancerous organ, predetermined to develop dysplasia and cancer as we age[4]. Survival rates following diagnosis are very good, with current 10-year survival reaching approximately 84% in the UK[1], making prostate cancer a disease that men more commonly die with rather than from. However, due to such a high prevalence and subsequent rates of diagnosis, prostate cancer still accounts for 13% of all UK male cancer deaths[1].

Considering the application of significant scientific effort over past decades, the clinical appraisal of patients suspected of having prostate cancer still primarily relies on prostate-specific antigen (PSA) levels, a single broad and error-prone blood biomarker. Taken in isolation, 75% of men in the PSA "grey zone" (4 - 10 ng/mL) have been found to not have prostate cancer on biopsy[5]. Confirmation of disease status is via invasive needle biopsy that in and of itself suffers from sampling problems, leading to both over- and underestimation of disease status. Current biopsy techniques can result in more negative results than cancer findings, although this can vary from centre to centre[6–8]. There is a clear need for clinically implementable tools able to selectively identify those men that can be safely removed from clinical pathways and adequately stratify those men harbouring disease that requires intervention.

An opportune point for the triage and risk assessment of patients suspected to harbour prostate cancer would be prior to an initial biopsy. This would allow the lowest risk patients to forgo invasive biopsy whilst simultaneously identifying high-risk patients in need of fast-tracking through to more aggressive treatment options. Liquid biopsy techniques that are minimally- or non-invasive have gained huge traction in biomarker discovery for a multitude

of malignancies[9,10]. Both the ductal nature of the prostate and interconnected nature of the male urological system make urine an ideal means for holistically sampling the prostate non-invasively. Sloughed-off cells, secreted proteins, nucleic acids and extracellular vesicles from normal and cancerous prostate tissue can all find their way into the urine through prostatic ducts that drain into the urethra[8,10–15].

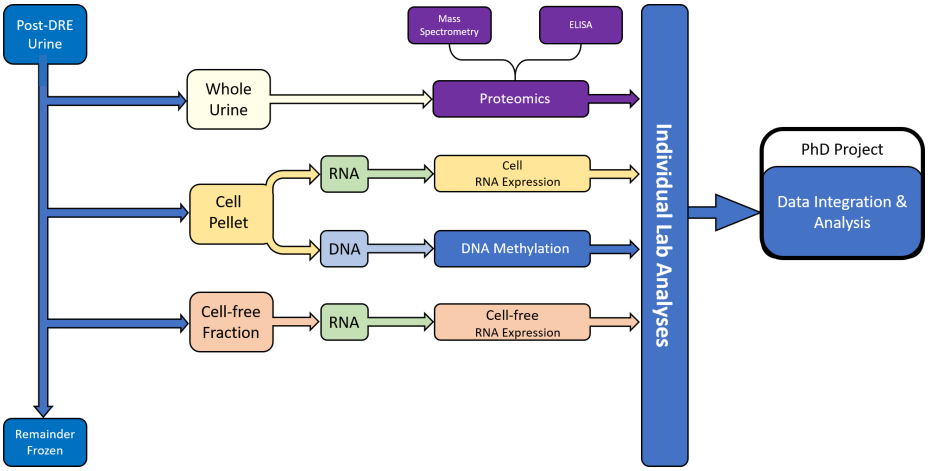


Figure 1.1: Workflow illustration of data sources from the Movember GAP1 Urine Biomarker project. Individual laboratories analysed samples, generating data and analysing indivudal datasets where possible.

In 2012, the Movember Global Action Plan 1 (GAP1) initiative was launched; a global collaboration between multiple institutes focusing on developing new biomarker candidates for prostate cancer in urine, plasma, serum and extracellular vesicles. A prime aim of the GAP1 urine biomarker initiative was to develop a multi-modal urine test for the discrimination of disease state. The consortium of 12 collaborating institutes across seven countries analysed a total of 1,258 samples across a range of analytical methods including transcriptomic, proteomic, methylation and ELISA assays (Figure 1.1. Due to limitations in the amount of material obtained from samples, not all analyses could be performed on every sample, however sufficient overlap was present in a number of key assays that will be explored later.

The focus of this thesis explores how machine learning can be utilised to optimally and robustly harness the information contained across multiple fractions of urine, in order to develop a non-invasive test for prostate cancer. As a first step I explore how transcriptomic data available from a single urinary fraction can be used to accurately discriminate disease status in patients. I then progress to improve upon this method using more robust statistical methodology and data processing before describing an encompassing framework for the rapid prototyping of predictive models. I apply this framework as part of data integration studies across differing methods of assaying urine, to successfully develop three multimodal & multivariable models which demonstrate the potential for clinical utility, named ExoMeth, ExoGrail and ExoSpec. The ExoMeth model is a multivariable risk prediction model that incorporates information from clinically available parameters, cellular methylation targets, and cell-free RNA gene information. Able to predict biopsy outcome with clinically useful precision, ExoMeth displays the potential to reduce biopsy rates by >65%, if externally validated and implemented. With this considered, in my final chapter I describe the design

of a clinical validation trial for the developed predictive models and use RNA sequencing data to suggest new targets for future biodiscovery trials.

1.1 My guiding philosophy - robust, reproducible and relevant analyses

The vast majority of cancer biomarkers fail to translate to the clinic; with only 1% of published discoveries entering clinical practice[16,17]. The lack of uptake could be attributed to a variety of issues, including a lack of robustness by identifying dataset-specific features or by sub-optimal statistical practice. Additionally, some biomarkers may not answer a clinically relevant question, or are cost prohibitive for the predicted effect sizes. Such is the case with the PCA3 test, a urine test for predicting biopsy outcome in patients with a previously negative biopsy, that has been recently recommended against by the National Institute for Clinical Excellence (NICE) due to being uneconomical for the reported clinical utility[5].

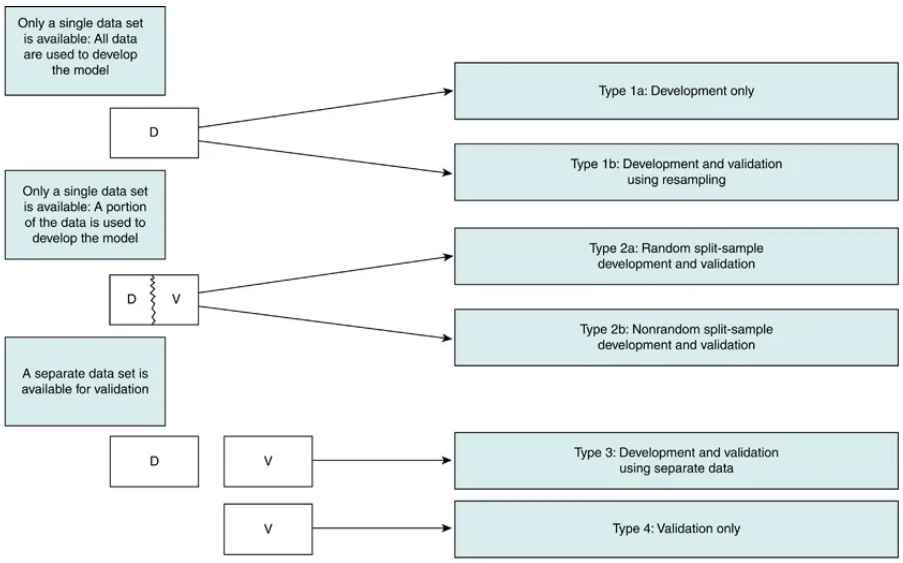
In this thesis I try to avoid these pitfalls by holding statistical robustness, quantitative reproducibility and clinical relevance as key tenets throughout my analyses:

- Robustness is achieved by the extensive use of bootstrap resampling, simulation, and the avoidance of point-estimates or over-reliance on P values for interpretation of results.
- Through documented structuring of analysis scripts and the adoption of statistical programming best practice, all results can be quantitatively reproduced if required.
- The clinical utility and translational potential of models are assessed by the use of clinically relevant endpoints and the quantification of effect sizes if models were to be applied at a population level.

1.1.1 TRIPOD guidelines

In 2008, Glasziou and colleagues assessed the reporting quality of 80 trials and systematic reviews in health research, finding over half of them to be inadequate[18]. There are several reasons that make inadequate reporting problematic. Insufficient details concerning the design and implementation of a trial leave readers without the ability to critically appraise the reliability of published results and their interpretations. Additional ethical and moral reasons highlight the need for adequate reporting[19].

As a result of this the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) Network was established as an international initiative to improve the quality of healthcare research through the promotion of transparent and accurate reporting[20]. Through the EQUATOR Network several guidelines have now been developed, encompassing recommendations for clinical trials, prognostic markers, genetic risk prediction, and most importantly for these works, guidelines for the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)[21].



Analysis Type	Description
Type 1a	Development of a prediction model where predictive performance is then directly evaluated using exactly the same data (apparent performance).
Type 1b	Development of a prediction model using the entire data set, but then using resampling (e.g., bootstrapping or cross-validation) techniques to evaluate the performance and optimism of the developed model. Resampling techniques, generally referred to as 'internal validation', are recommended as a prerequisite for prediction model development, particularly if data are limited (6, 14, 15).
Type 2a	The data are randomly split into two groups: one to develop the prediction model, and one to evaluate its predictive performance. This design is generally not recommended or better than type 1b, particularly in case of limited data, because it leads to lack of power during model development and validation (14, 15, 16).
Type 2b	The data are nonrandomly split (e.g., by location or time) into two groups: one to develop the prediction model and one to evaluate its predictive performance. Type 2b is a stronger design for evaluating model performance than type 2a, because it allows for nonrandom variation between the 2 data sets (6, 13, 17).
Type 3	Development of a prediction model using one data set and an evaluation of its performance on separate data (e.g., from a different study).
Type 4	The evaluation of the predictive performance of an existing (published) prediction model on separate data (13).

Types 3 and 4 are commonly referred to as 'external validation studies.' Arguably, type 2b is as well, although it may be considered an intermediary between internal and external validation.

Figure 1.2: Types of prediction model studies covered by the TRIPOD statement. D = development data; V = validation data. Adapted from the TRIPOD Statement

The TRIPOD guidelines are a set of recommendations for the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes, most of which form the focus of this thesis. Additionally TRIPOD describes the types of prediction model study designs and provides an evidence level hierarchy for assessing the development and validation of such studies (Figure 1.2).

With this considered, the analyses reported in Chapter 6 and Chapter 7 fully adhere to these reporting guidelines. Similarly, the validation clinical trials in Chapter 8 are designed to ensure that evidence generated is of the highest quality and few, if any, additional trials are required before clinical implementation becomes a tangible goal.

1.2 Aims and objectives of this thesis

1.2.1 Aims

To utilise the data available from the Movember GAP1 Urine Biomarker project to its fullest extent by developing robust, reproducible models for risk prediction that assess multiple streams of data in a urine sample.

1.2.2 Objectives

- To describe the current clinical pathway for prostate cancer patients and identify the need for more precise tests.
- To develop a risk prediction model using NanoString data, with validation in a sub-cohort.
- To assess whether different machine learning algorithms and training labels can be used to improve model performance and provide more clinical utility.
- To investigate the use of resampling methods and the bootstrap to maximally utilise all available data in small, high-dimensionality datasets without compromising robustness or reproducibility.
- To apply this framework to all available overlaps of urine data of an appropriate sample size ($n > 200$ typically) from the GAP1 study to develop predictive risk models.
- To design clinical validation trials for these models to expediate their adoption in clinical practice
- To identify targets and modalities of interest for future biodiscovery studies.

1.3 Chapter overview

- **Chapter** 2**:** Background information is given on prostate cancer, the current toolset available to clinicians and the need for more advanced biomarkers.
- **Chapter** 3**:** Detailed descriptions of all methods applied in this thesis are given, serving as a reference for analytical and statistical methods.
- **Chapter** 4**:** This chapter describes the development and validation of a four-group risk prediction model for prostate cancer and the application of this model to an active surveillance sub-cohort.
- **Chapter** 5**:** Improvements to this model are investigated in this chapter by exploring different combinations of machine learning algorithms, training labels, and resampling strategies.
- **Chapter** 6**:** This chapter describes the structure of a semi-automated machine learning framework for biodiscovery in overlapping datasets. Using the findings from the previous chapter, the Framework implements Random Forests for both feature selection and model creation, whilst creating comparator models from individual datasets.
- **Chapter** 7**:** In this chapter the Framework is applied to a number of overlapping datasets, developing both promising models and negative results.

- **Chapter 8:** The results from all previous chapters are considered as a whole, discussing the strengths and weaknesses of the analyses within this thesis. A clinical trial design to validate the analyses presented is proposed.

1.4 Thesis output

This thesis has produced: peer reviewed papers; talks at academic conferences; intellectual property; and preprints. Full details of these outputs are described below.

1.4.1 Peer reviewed papers

- Shea P. Connell *et al.* (2020) *Development of a multivariable risk model integrating urinary cell DNA methylation & cell-free RNA data for the detection of significant prostate cancer*, The Prostate, 2020 (1 - 12). `doi:10.1002/pros.23968`
- Shea P. Connell *et al.* (2019). *A four-group urine risk classifier for predicting outcomes in patients with prostate cancer.* BJU International, 124(4). doi: 10.1111/bju.14811

1.4.2 Papers under review

- Shea P. Connell *et al.* (2020) *Development of a risk model integrating cell-free RNA & proteomic data for the pre-biopsy detection of prostate cancer from urine*, TBD
- Shea P. Connell *et al.* (2020) *Integration of urinary EN2 and cell-free RNA in developing a multivariable risk model for the detection of prostate cancer in biopsy naieve patients*, TBD

1.4.3 Invited talks & accepted posters

- **Detecting clinically significant prostate cancer with urine: A multivariable risk model integrating urinary proteomic and cell-free RNA data** - Poster, talk European Society for Urological Research 2019, Porto, Portugal.
- **Using urine to diagnose prostate cancer: developing two multimodal diagnostic models reproducibly within R** - Invited Talk, R/Medicine 2019, Boston, USA.
- **Predicting outcome in prostate cancer patients using a multi-signature risk classifier, derived from urinary extracellular vesicles** - Poster, European Association of Cancer Researchers Tracking Cancer 2019. Awarded Clinical and Metastasis Poster Prize.

Chapter 2

Background

2.1 Summary

In this chapter I provide an overview of the main biological and medical concepts and approaches relevant to both the treatment of prostate cancer and the projects explored in this thesis. I provide a brief anatomical description of the prostate, the epidemiology and clinical presentation of prostate cancer and the current toolsets used by clinicians in the appraisal of prostate cancer. Some of the leading existing non-invasive methods for the detection of prostate cancer are discussed and framed in the context of clinical utility. Finally, the applications of machine learning for biomarker discovery and development are discussed, considering the benefits and limitations to medical settings and prostate cancer specifically.

2.2 Cancer and the prostate

A polygenic disease, cancer requires multiple genetic and epigenetic alterations to be acquired before a cell can escape growth regulation and proliferate uncontrollably. Typically a disease of old age, acquired genetic mutations are compounded over time to drive cancer progression, though some cancers such as leukaemia or brain cancers are particularly prevalent in young children[1]. The hallmarks of cancer comprise eight biological capabilities, detailed fully by Hanahan *et al.*[22,23], but briefly, are; sustaining proliferative signalling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis resisting cell death, reprogramming of energy metabolism and evading immune destruction. Basic research aims to underpin the specific mutations and alterations to biological pathways that drive prostate cancer, however this is outside of the scope of the current work and readers are directed to excellent articles by Gundem *et al.*[24] and Schlomm *et al.*[4] for further information.

2.2.1 The Prostate

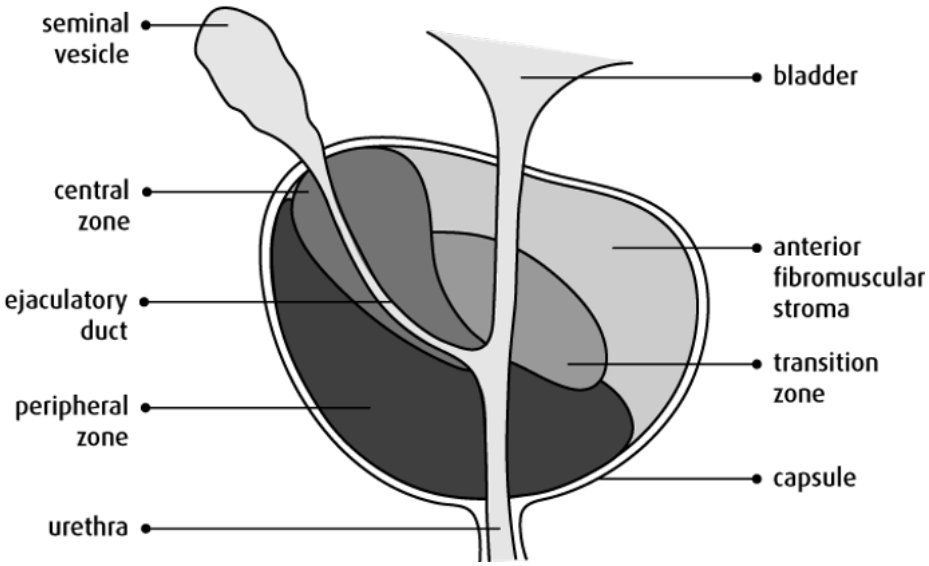


Figure 2.1: The zonal anatomy of the prostate gland (adapted from The Canadian Cancer Society)

The prostate is a fibromuscular secretory gland of the male reproductive system approximately the size of a walnut, forming the sexual accessory tissue with the Cowper and Littre glands, seminal vesicles and ampullae[25]. The composition of the prostate is approximately 70% glandular, with the remaining 30% composed of fibromuscular stroma. The tissue of the prostate is formed of many branching ducts, surrounded by the stroma, itself formed of connective tissue and muscle fibres (Figure 2.1). The cells lining these ducts produce prostatic fluid, an alkaline liquid with high levels of zinc, polyamines and citric acid[26]. The prostatic fluid also contains some secretory proteins and proteolytic enzymes including prostate specific antigen (PSA), a key biomarker used in the diagnosis of prostate cancer that is usually involved in liquefying semen immobilised within the seminal coagulum[27].

The human prostate is anatomically defined in terms of zones, as described by the works of McNeal[28–30] and detailed in Figure 2.1. The prostate is split into:

1. Peripheral Zone

   The peripheral zone comprises up to 65% of the mass of a healthy prostate, but accounts for the origin of 70 to 80% of prostatic carcinomas[31,32]. This high proportion of reported diagnoses may be influenced by the proximity to the rectum, where it the most readily sampled area for biopsy and digital rectal examination[31,32].

2. Anterior fibromuscular stroma

   This region of the prostate overlies the urethra anteromedially, meeting the smooth muscle of the bladder neck and external urethra sphincter. As opposed to the oversampling of the peripheral zone, the anterior zone represents a historically underrepresented zone by diagnosis, despite being reported to account for 20% of all prostate cancer in radical prostatectomy samples and cancers in this region are associated with higher prostate cancer-specific mortality[33–35].

3. Central Zone

   The central zone forms approximately 30% of the total glandular mass, with ducts fanning perpendicular to the ejaculatory duct. Tumours originating here have been reported to be associated with more aggressive disease, but only account for approximately 2.5% of reported cancers[36,37]

4. Transition Zone

   Two lobes of glandular tissue bordering the urethra form the transition zone, where benign prostatic hyperplasia (BPH, a non-malignant enlargement of the prostate) originates. Accounting for only 5% of the total prostate volume, incidence of cancer is reported to vary from 4% to over 20% and has been associated with more favourable outcomes[38,39].

2.2.2 Prostate Cancer

> "The prostate is a precancerous organ that inevitably develops dysplasia and cancer over time" — Schlomm *et al.* (2015)[4]

Prostate cancer is diagnosed remarkably frequently in high-income countries, accounting for 26% of all cancer diagnoses in the UK in 2017[1]. Indeed, autopsy studies have shown that prostate cancer is present in 25 to 40% of men of all ages at the time of their death, increasing with age and approaching 85% in 81 to 95 year-old patients[2,40,41], leading some to believe that the prostate is an inherently abnormal tissue[4]. Despite this exceptional frequency, only a small percentage of cancers progress to become clinically apparent[5]. Less than 15% of detected cancers progress to kill the patient within 10 years of diagnosis, which when coupled with the high incidence, makes prostate cancer responsible for the largest number of male cancer deaths in the UK[1].

Some tumours are undeniably aggressive, progressing rapidly and requiring immediate clinical intervention with curative intent, whilst a large proportion show very slow growth and are indolent in nature. Determining which patients will require treatment and those that do not need any intervention is a key clinical issue for many healthcare systems. This issue is non-trivial and is one of the main focuses of this thesis.

2.3 The diagnostic and prognostic toolsets for prostate cancer

Diagnosis of prostate cancer in the United Kingdom typically follows published guidelines from the National Institute of Health and Care Excellence (NICE), whilst in the EU it is the European Association of Urology (EAU) and USA guidelines by the American Urological Association (AUA). All regularly review available literature and update their guidelines to recommend or rescind treatments, procedures and medicines in light of emerging evidence-based research, though the most recent update from NICE will be the focus of this section[5]. Broadly speaking, there are four key tools used to provide clinicians with adequate information to make initial treatments about decision at the time of diagnosis: serum PSA levels, digital rectal examination (DRE), multi-parametric magnetic resonance imaging (mpMRI) and needle biopsy.

2.3.1 PSA

Prostate specific antigen (PSA) is a glycoprotein encoded by the *KLK3* gene. Secreted by the epithelial cells of the prostate, and not produced by any other organ in the body, PSA is theoretically the ideal marker for prostate-specific diseases. PSA can be detected in the serum of healthy patients as well as those with prostate cancer where elevated levels are associated with increasing disease severity. However, elevated PSA can also be caused by more benign conditions including BPH and prostatitis, or be influenced by external factors such as physical and sexual activity and even by temperature[42–44]. Despite the observed variability, PSA remains the most widely used biomarker for the early detection of prostate cancer.

Since clinical adoption of PSA levels in the 1980s, the reported incidence of prostate cancer has steadily risen, whilst the proportion of patients dying prostate cancer-specific deaths has decreased[45]. Historically, widespread PSA screening was a strategy employed by healthcare systems to aid the early detection of prostate cancer, though this resulted in the over-diagnosis and subsequent over-treatment of patients with indolent disease. As such, NICE, the AUA and the EAU all advocate against the use of routine PSA screening. Instead, PSA testing is typically triggered by one of the following circumstances, or symptoms that can be associated with prostate cancer:

- Patients older than 50 years of age who request a PSA test.
- Lower urinary tract symptoms (LUTS), such as nocturia, urinary frequency, hesitancy, urgency or retention.
- Erectile dysfunction
- Visible haematuria
- Unexplained symptoms that could be due to metastatic prostate cancer, such as bone pain, weight loss or lower back pain.

The clinical application of PSA from its use for triggering a biopsy as a result of elevated PSA through to assessing relapse following treatment is discussed below.

2.3.2 Digital Rectal Examination

The digital rectal examination (DRE) is a clinical technique used to assess the prostate by palpation for size, firmness, discernible nodules or lumps that may indicate the presence of prostate cancer, or need for further investigations. A DRE is performed often as one of the first clinical lines of investigation when lower urinary tract symptoms are reported by the patient, or some other clinical suspicion exists. Performed by a general practitioner (GP) or urologist a DRE is performed by inserting a finger into the rectum of a patient and palpating the prostate through the wall of the colon. However, as only the posterior of the prostate can be felt by DRE, anteriorly located cancers or cancers not impacting the prostatic capsule cannot be felt. Efficacy of the DRE is questionable, as the results can vary according to clinician, position of the prostate or body mass of the patient, whilst nodules can not be apparent upon a repeat examination. With a reported sensitivity of 51% and specificity of 59%, the DRE in isolation is of little diagnostic use[46]

2.3.3 Needle Biopsy

Currently the only method to definitively diagnose prostate cancer is via needle biopsy of the prostate and subsequent histopathological identification of tumour tissue[5]. Tissue is

commonly obtained through one of two methods, a trans-rectal ultrasound-guided (TRUS) biopsy or trans-perineal template prostate mapping (TPM) biopsy. TRUS biopsies usually collect 10 to 12 tissue cores through the wall of the rectum and into the prostate under a local anaesthetic. TPM biopsy can collect upwards of 24 cores through the perineum using TRUS or mpMRI information to map a template of the prostate to a grid and avoiding the urethra and under a general anaesthetic by NICE guidelines[5]. Both methods are associated with some degree of sampling error both in over- and under-estimation of disease status in the case of a cancer finding, estimated for TRUS-biopsy to be 29% and 14%, respectively[47]. Where a negative biopsy outcome is recorded, TRUS biopsy is associated with far higher rates of false negatives, reported to reach 20 - 30% when compared to TPM biopsy[48].

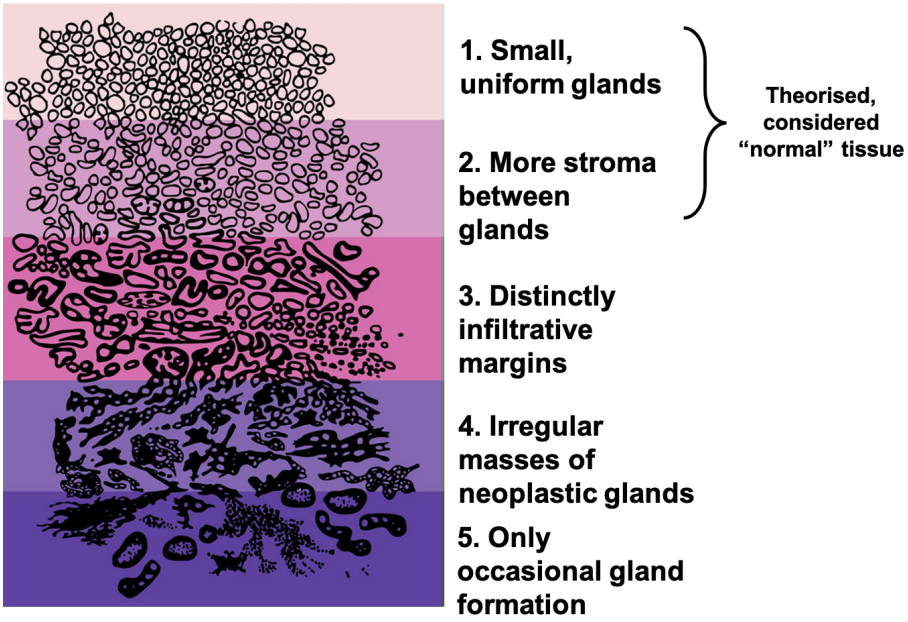


Figure 2.2: Representation of idealised Gleason patterns in tissue samples. Gleason patterns 1 and 2 are theorised and are not observed in prostate tissue. Adapted from The National Institute for Health

Tissue obtained from needle biopsy is appraised for the presence and advancement of cancer by means of the Gleason score. The Gleason score is a grading system for identifying the histological morphology of prostate tissue as a measure of cellular differentiation[49,50]. Since its inception almost half a century ago, the Gleason score has been gradually and repeatedly refined by pathologists to remain as one of the most important single markers available to clinicians[50–52]. Calculated as the sum of two patterns, the Gleason score is the sum of the most prevalent and second most common patterns when reporting biopsy outcomes (Figure 2.2). In radical prostatectomy samples, Gleason score may also be reported with a tertiary pattern[51].

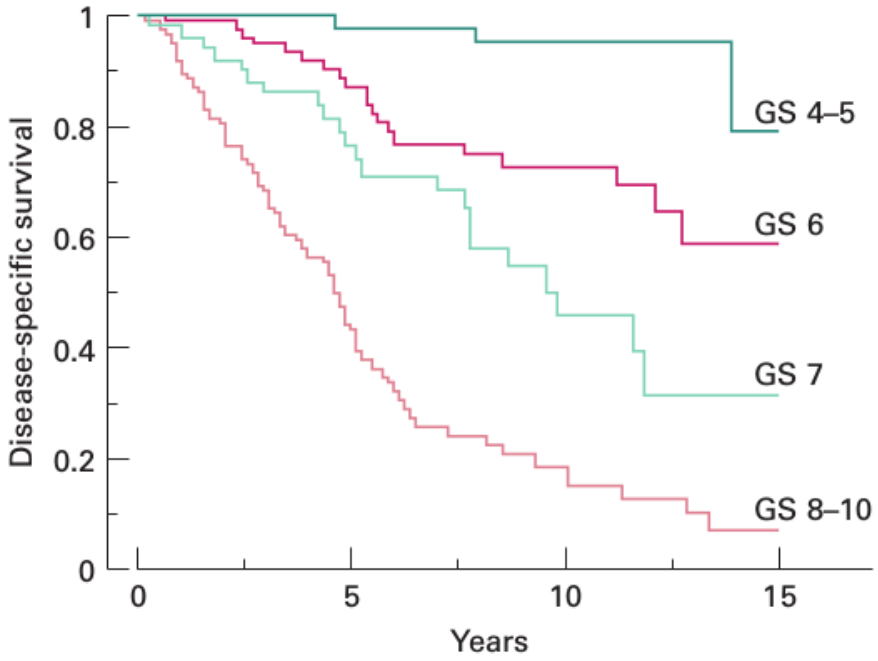


Figure 2.3: Prostate cancer-specific survival for patients diagnosed by TRUS biopsy with tumours of differing total Gleason scores. Adapted from Egevad *et al.* (2002), reported $P < 0.001$.

The Gleason score is still the most useful prognostic tool currently available (Figure 2.3), with no superior tool identified in the 50 years since its inception. The grading of tissue biopsy requires significant time and skilled pathologists and, even then, changes in assigned Gleason scores have been observed due to operator and location biases[53]. As will be discussed below, the nature of needle biopsy itself has drawbacks in sampling error and infection risks. Compounded by the multifocal nature of prostate cancer, finding a new and objective method of holistically assessing prostate health is key to improving patient care.

2.3.4 MRI

The latest addition to the NICE guidelines in 2019 was the adoption of multiparametric magnetic resonance imaging (mpMRI) as a first-line triage device prior to biopsy for patients with a clinical suspicion of prostate cancer (from PSA, LUTS, DRE or combination of the three) that are eligible for radical treatment[5]. Previously mpMRI was only formally recommended to patients following a negative initial biopsy, where a clinical suspicion remained[54]. As a non-invasive imaging technique, mpMRI makes a good first-line investigation to filter patients with benign or indolent conditions from the clinical pathway without the need for needle biopsy. Reported on the 5-point Likert scale prostate imaging-reporting and data system (PI-RADS), imaging data is scored by trained radiologists according to the likelihood of finding clinically significant prostate cancer on biopsy, defined as any one of: Gleason score ≥ 7, tumour volume >0.5 mL, or extraprostatic extension[55].

The 2019 NICE guidelines for the clinical use of mpMRI as a first-line option prior to biopsy are based on three studies; Porpiglia *et al.*,[56] the PRECISION trial from Kasivisvanathan *et al.*[57] and the PROMIS trial, reported in Ahmed *et al.*[58]. Both Porpiglia *et*

al. and Kasivisvanathan *et al.* used a PI-RADS score of 3 as a threshold for biopsy and showed that up to twice as many people with clinically significant cancers were likely to be identified through the use of mpMRI-influence biopsy rather than prostate biopsy alone. The PROMIS trial provided evidence that there is still a chance of missing clinically significant disease at that threshold and so NICE has still maintained that biopsy should not be definitively ruled out for patients with a PI-RADS score of 1 or 2[5,58]

2.3.5 TNM Staging

Upon confirmation of localised prostate cancer in a tissue biopsy the tumour is staged according to the tumour, node, metastasis (TNM) standardised system for malignant tumours. As described and maintained by the Union for International Cancer Control, it is comprised of three main parts with subsections for further description:

- T – size and extent of the primary tumour:
  - TX – Primary tumour could not be assessed
  - T0 – No evidence of primary tumour
  - T1 – Clinically insignificant tumour neither palpable or visible in imaging
    - T1a – Incidental histological finding in less than 5% of tissue
    - T1b – Incidental histological finding in more than 5% of tissue
    - T1c – Tumour identified via needle biopsy
  - T2 – Tumour confined within the prostate
    - T2a – Tumour in half or less of one lobe
    - T2b – Tumour in more than one half of one lobe
    - T2c – Tumour in both lobes, but still confined to prostate gland
  - T3 – Tumour exhibits extra-capsular extension beyond the prostate
    - T3a – Extra-capsular extension
    - T3b – Invasion of the seminal vesicles by the tumour
  - T4 – Fixed tumour or invasion of surrounding organs other than the seminal vesicles.
- N – presence/absence and extent of regional lymph node metastasis:
  - NX – Regional lymph nodes cannot be assessed
  - N0 – No regional lymph node metastasis
  - N1 – regional lymph node metastasis
- M – presence/absence of distant metastasis:
  - M0 – No distant metastasis
  - M1 – Distant metastasis
    - M1a – Non-regional lymph nodes
    - M1b – Bones
    - M1c – Other sites with or without bone disease

The above describes clinical staging, the extent of cancer at time of diagnosis. Pathological staging describes tumour extent following radical prostatectomy (RP), often providing a more accurate classification due to biopsy sampling error and other factors. Pathological staging criteria are almost identical to clinical staging, though there is no T1 classification for pathological TNM.

2.4 The current clinical pathway for patients

The current clinical journey from clinical suspicion through to diagnosis, prognosis and treatment of prostate cancer, as described by NICE, the EAU or AUA, is a complex one. There is no single absolute path for any given man from presenting with symptoms, being diagnosed with a prostatic adenocarcinoma, and being offered treatment with curative intent or placed onto a protocol of active surveillance, or, unfortunately in some cases, provided with palliative care.

Current NICE guidelines for the treatment of patients suspected to have prostate cancer can be broadly categorised into three distinct sections; detection and diagnosis, active surveillance, and management. The focus of the work in this thesis is primarily on the prognosis of prostate cancer biopsy outcomes with a secondary focus on prediction in active surveillance, therefore the curative treatment of prostate cancer will not be covered in detail. It's additionally important to state that while NICE and other organisations provide the official guidelines, clinical practice can, and does, diverge dramatically from them. For example, the Movember GAP3 Active Surveillance project supplies hugely varying protocols for the management of lower risk patients and from personal communications, some clinicians are moving away from TRUS biopsy towards MRI-guided and template biopsies.

2.4.1 Diagnosis

Patients presenting at primary care with a clinical suspicion of prostate cancer are offered a PSA test and a DRE. Causes for suspicion include lower urinary tract symptoms (LUTS; increased frequency or urgency of urination, incontinence, painful urination or excessive nocturia for example), erectile dysfunction, haematuria, or other unexplained symptoms that could be the results of metastatic disease such as bone pain and weight loss. Patients over the age of 50, or with familial history can also request a PSA test at primary care. However, careful consideration and discussion concerning the potential benefits and limitations of the PSA test are required before performing the test in all cases.

As previously discussed, an elevated PSA alone is not definitive evidence of prostate cancer, but the results of a DRE can supplement an elevated PSA result to aid clinical decision making. With most detected cancers located in the peripheral zone, it is reported that a tumour >0.2 mL can be detected by clinician DRE[59] and in up to 18% of cases a tumour is detected by DRE alone, in the absence of elevated PSA levels[59].

If a clinical suspicion of prostate cancer remains following the results of a DRE and PSA test, patients are next offered mpMRI or prostate biopsy, dependent on local resources and consideration of the patient's eligibility for radical treatment. Where mpMRI is undertaken it is common practice to omit prostate biopsy for patients with a PI-RADS score of 1 or 2, instead opting for PSA surveillance at 3 to 6 month intervals unless patient-reported symptoms indicate the need for further investigation. Biopsy is usually offered to patients with a PI-RADS score of 3 or more and NICE recommend that the decision to refer patients for confirmatory needle biopsy be made on the combined results of the DRE, PSA test, associated risk factors (race, familial links etc) and mpMRI. Consideration towards the overall health and co-morbidities of the patient, where a discussion concerning the potential of living with the diagnosis of clinically insignificant cancer are also undertaken. Risk calculators and nomograms for risk estimation at this point prior to biopsy exist, and can be used by clinicians to estimate the risks of many different endpoints such as seminal vesicle invasion or predicted Gleason on biopsy[60,61]. However, the calculators require strong

external validation and calibration, and the Predict tool is the only one currently endorsed by NICE guidelines[5,62].

The current NICE guidance for needle biopsy of the prostate is MRI-influenced TRUS biopsy, with NICE recommending against TPM biopsy as a first-line choice[5]. This recommendation is based on the intensive resource requirements to undertake a TPM biopsy; with general anaesthetic and extensive histological analysis of the at least 24 cores taken.

2.4.2 Risk Stratification and Prognosis

Several systems exist for stratifying patients into categories based on the severity of their disease following a confirmed diagnosis of prostate cancer. The most common system, used by NICE and EAU is the D'Amico Risk Classification for prostate cancer, designed to assess the five-year risk of biochemical recurrence following radical therapy, D'Amico Risk uses a combination of PSA, Gleason score and tumour staging attained prior to treatment[63] (Table 2.1). This risk categorisation forms the backbone of the clinical pathway and is utilised to inform the decision of putting prostate cancer patients forward for radical treatment, active surveillance or watchful waiting.

Table 2.1: D'Amico risk stratification parameters for patients with localised prostate cancer

Risk Level	PSA		Gleason Score		Clinical Staging
Low	<10 ng/mL	and	6 or below	and	T2a
Intermediate	10 - 20 ng/mL	or	7	or	T2b/c
High	>20 ng/mL	or	8 or above	or	T3

2.4.3 Treatment or Active Surveillance?

Methods for the treatment for prostate cancer with curative intent are broadly grouped into surgical, radiotherapy and chemotherapy-based approaches or some combination of the above. The side-effects of any treatment are not insignificant, with urinary incontinence and erectile dysfunction widely reported following radical prostatectomy (RP) in 47% and 36% of patients, respectively[64], whilst the risk of adverse cardiovascular events are significantly increased under a regimen of androgen deprivation therapy[65]. These life-altering side-effects are the prime reason for reducing the overall rates of treatment for patients that do not strictly require it.

Using the D'Amico system, most patients currently diagnosed with Low risk disease in the UK generally forgo immediate treatment and instead are enrolled onto a program of Active Surveillance (AS). The aim of AS is to delay or avoid altogether the treatment of patients until it is clear that intervention is essential, with the goal to avoid over-treatment and the side-effects of treatment without adversely influencing prostate cancer-specific mortality.[5,66,67] A typical active surveillance regime by NICE standards include repeated PSA measurement at 6 month intervals, with a repeat biopsy at 2 years following initial enrolment[5]. The clinical trigger points for intervention with curative intent are highly variable and depend on the attending clinician[68], ranging from threshold levels and doubling time of PSA to adverse histology or a volume increase on mpMRI[8,68]

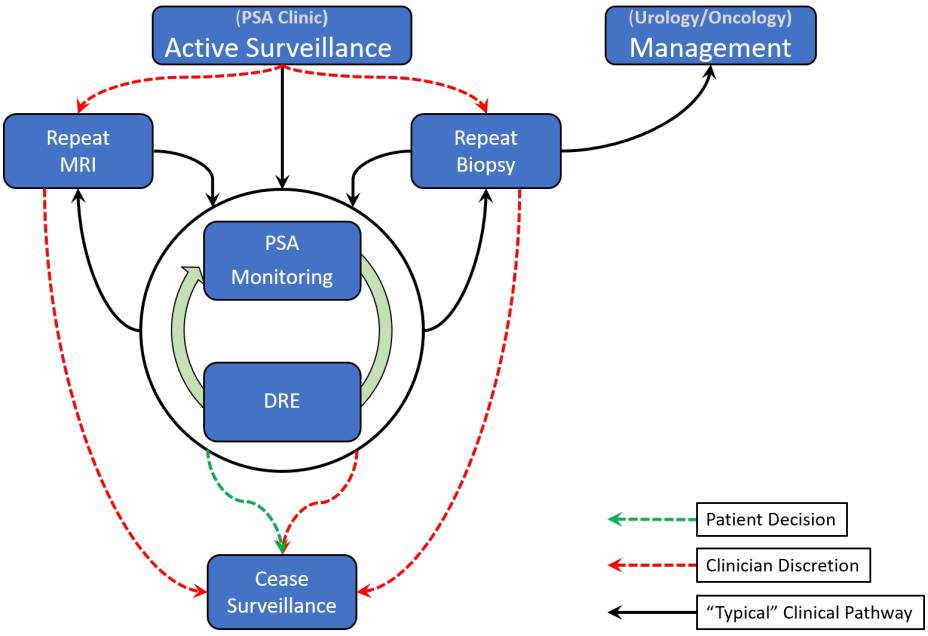


Figure 2.4: The generalised clinical pathway for prostate cancer patients enrolled onto active surveillance programs.A patient can elect for, or refuse treatment at any point within this pathway.

The generalised clinical pathway for patients on active surveillance is relatively well-defined by NICE, although it is important to note that patients can decide at any point to forgo any treatment or biopsy, or to elect for treatment with curative intent (Figure 2.4). Rates of self-election for treatment can be exceptionally high in some cohorts, with a large study finding that 32.8% of patients on an AS protocol received radical therapy without meeting the criteria for progression within 10 years of enrolment[69].

An issue with AS is the lack of a formal mechanism for ceasing active surveillance for those that fail to show any signs of progression, other than by self-election by the patient, the clinician's discretion or, unfortunately, by death. In practice, this means that once patients are placed onto AS, they are monitored until their disease spreads and requires treatment, or they die from other causes. Nevertheless, active surveillance has proven to be one of the best contemporary methods for protecting patients with favourable prognoses from unnecessary treatment, where rates of metastasis in well-managed cohorts can be as low as 0.1 - 2.8%[70].

Even with high rates of treatment, both for-cause and self-election, the disease-specific mortality rates at 10 years following diagnosis are not drastically different between those receiving treatment and those that do not[69]. Therefore, for patients with life expectancies of less than 10 years or with significant co-morbidities that effect their eligibility for radical treatment, watchful waiting is generally chosen. Watchful waiting involves similar surveillance as AS protocols, however the treatment on detection of disease progression is palliative rather than curative[5].

2.5 The clinical problem

Prostate cancer is not only biologically heterogeneous, it is also clinically diverse, where no two patients with the same clinical presentation will share identical outcomes; not all patients diagnosed will need treatment, whilst others will critically require intervention to save their life. Deciphering the best methods to adequately stratify these patients is where the clinical problem arises. Currently neither researchers nor clinicians can achieve this to satisfactorily avoid over-diagnosis and over-treatment. Reported rates of over-diagnosis vary widely, from 1.7% to 67%, dependent on population sampled and the criteria used to send patients forward for biopsy[40]. Due to the uncertainty and current inability to accurately prognosticate patients, many with disease thought not to be of immediate concern will self-elect for radical treatment, itself not free of consequences, whilst others may refuse treatment altogether through a perceived lack of benefit.

Prostate cancer has historically been over-diagnosed; patients have received a diagnosis of an indolent form of the disease, one that would not have otherwise become clinically apparent or significant enough to be life-altering in the absence of a diagnosis, where Gleason 3+4 is considered to be the key threshold. Over-diagnosis has both immediate and long-term impacts for patients and healthcare systems alike. In the short-term patients must face the knowledge that they have cancer, with the anxiety this brings causing some to seek life-altering treatment, regardless of a prognosis[71,72].

Due to the ageing population of the United Kingdom, the incidence of prostate cancer is projected to grow by upwards of 65%[73]. As incidence increases, so too will the numbers of patients inappropriately biopsied, diagnosed with indolent disease and potentially receive radical therapy which has large ramifications for quality of life. The life-altering results of radical treatment are pronounced; erectile dysfunction and urinary incontinence are reported in a large proportion of patients following radical prostatectomy (RP) (47% and 36%, respectively), whilst the risk of an adverse cardiovascular event is significantly elevated for patients receiving androgen deprivation therapy[65]. Even the side-effects of receiving a needle biopsy are not insignificant; physical and mental distress, haematuria, painful urination and in some cases, life threatening sepsis[58] have all been reported[5]. This illustrates the clear and immediate need for clinically implementable tools able to precisely and non-invasively identify patients that can either be safely removed from treatment pathways, or those requiring further follow up.

2.5.1 PSA reliability, or lack thereof

Serum PSA measured in isolation is a poor predictor of prostate health, as reported by NICE 75% of patients with an elevated PSA have been found to not have prostate cancer on biopsy[5]. PSA levels are not specific to cancer and are influenced by many factors, lacking the specificity to discriminate between benign conditions such as BPH or infection and indolent cancer from aggressive. Additionally, due to a lack of standardisation in commercial PSA assays, significant discordance between PSA readings has been reported between different commercial suppliers[74,75]. This means that if different assay kits are used between care centres, the reported PSA levels may be systematically higher or lower than initially reported at the patient's primary care centre.

The European Randomised Study of Screening for Prostate Cancer (ERSPC) showed that PSA-based screen resulted in a 20% reduction of prostate cancer-specific mortality, though at the cost of 40% of all patients diagnosed in the study posessing a clinically

insignificant level of disease[76]. These findings overall meant a reported 24% reduction in overall quality of life years gained due to PSA screening[77] and an important limitation of the ERSPC study was the choice of a sole PSA threshold of 3 ng/mL to trigger biopsy, as opposed to a more reasoned approach considering the other risk factors detailed above.

Similarly, the Cluster Randomised Trial of PSA Testing for Prostate Cancer (CAP) trial observed no benefits in overall mortality rates as the result of a low-intensity PSA screening intervention in approximately 500,000 UK men[6]. Even within the control arm of the CAP trial, most biopsies performed according to NICE guidelines were negative for cancer on biopsy (personal communication with Richard Martin, lead author). Indeed, within the datasets explored throughout this thesis, patients were found to not have cancer with serum PSA levels ranging from 4 - 30 ng/mL, detailing the reliability issues serum PSA measurement faces as a prognostic tool in isolation.

2.5.2 Sampling error of biopsy

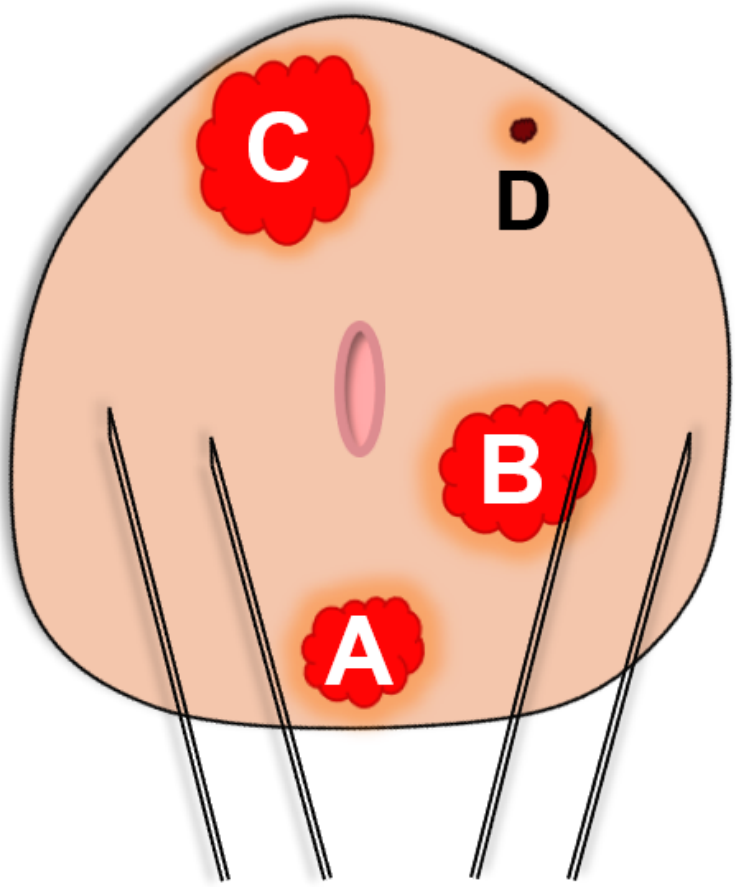


Figure 2.5: Examples of sampling errors in detecting clinically significant tumours for both tissue biopsy (A to D) and in MRI (D). A – random error, missing tumour in needle biopsy. B – Random error, attributing lower risk due to biopsy location and volume of tumour present. C – Systematic error due to lesion that is located anteriorly in the prostate, an area that is under sampled by biopsy. D – Technical "multifocality" error, small volume high-grade tumour missed due to the resolution limit of MRI and sampling size and position in tissue biopsy

Any method of sampling the prostate will be associated with some degree of error, the only solution to this is exhaustive survey of the whole organ, defeating the purpose of sampling. However, for a sampling strategy to be considered fit for purpose, the measurements taken and results reported must be representative of the whole. TRUS, and to a less extent TPM, biopsy do not meet this criteria particularly well. TRUS, whilst the only method actively recommended by NICE, suffers from four important sampling errors, both random and systematic, that drastically impact its clinical utility as a tool for diagnosis of patients (Figure 2.5). Indeed, two cross-sectional studies examined the utility of TRUS biopsy compared to mpMRI[58] and systematic template biopsy[78], led to NICE concluding that

whilst a cancer-positive TRUS biopsy leads to a very large increase (Likelihood Ratio >10) in the probability of significant (Gleason $\geq$ 7) disease being present, the opposite is not true; a negative TRUS biopsy does not meaningfully alter the likelihood of disease being present[79].

Biopsy itself is not without risk, possible complications associated with TRUS biopsy include bacterial infections, haematuria, painful urination and in some cases, life threatening sepsis. Infectious complications affect 1 - 4% of patients undergoing TRUS biopsy[80], whilst severe sepsis has been described in 0.1 - 3.5% of patients following a TRUS biopsy[81]. One study from Ontario, Canada reported that the hospital admission rate for infection-related complications within 30 days of the procedure increased from 1.0% in 1996 to 4.1% in 2005.[81] The reported incidence of urinary tract infections (UTI) after TRUS biopsy typically ranges between 2% and 6% with approximately 30%-50% of these patients having accompanying bacteremia[80].

2.5.3 Variability and costs of MRI

Whilst mpMRI has now been fully integrated into current clinical pathways, some questions and uncertainty still remains concerning the accuracy found in tightly controlled studies that pre-date NICE recommendation. It was observed in the PROMIS trial that 27% of biopsy-naïve patients with elevated PSA and non-suspicious mpMRI results could avoid a biopsy[58]. However, this performance was recorded under the strict controls of a clinical trial, where MRI scanners were carefully calibrated and systematically monitored by an external clinical research organisation (CRO)[58]. Indeed, Walz reports that several centres were unable to participate in the PROMIS trial due to the quality of their MRI scanners, despite having expert radiologists on staff[82].

Even in the case of expert radiologists, it has been reported that there is considerable variability in PI-RADS score assignment between operators and significant changes in cancer detection rates where 13-60% of patients with a PI-RADS score <3 harboured clinically significant prostate cancer.[83] Coupled with the fact that mpMRI is not a cheap technique, with costs ranging from approximately £700 to £1332 dependent on modelled scenario[84], there is clearly some room for improvements to be made in the non-invasive assessment of biopsy-naïve men.

2.5.4 Risk stratification is not fit for purpose

Due to the aforementioned heterogeneity and complexity of prostate cancer, we currently lack a consistent system for accurately sub-typing tumours differing in prognosis or treatment response based upon the expression of a few key genes, similar to the *ERBB2* over-expressing, basal and subliminal subtypes of breast cancers[85]. In its current guise, D'Amico risk stratification does not accurately predict the outcome of an individual patient with definitive certainty, instead categorising patients into one of three broad Risk groups based upon clinically available information (Table 2.1). If all tumours presenting with identical clinical symptoms behaved identically this would not be an issue, however as discussed, these clinically identical tumours can, and are, genetically disparate. Similarly to PSA and Gleason scoring no superior tools are currently available, and the D'Amico Risk classification system has been repeatedly shown to possess clinical utility above anything else available[86,87].

A key compromise employed by the D'Amico system for clinical simplicity is the broad

categorisation of PSA, a continuous measure that retains more information if kept continuous. By D'Amico a patient presenting with a PSA of 19.9 mg/mL has substantially improved survival odds compared to a man with a PSA of 20 mg/mL. Simultaneously this same patient is expected to be at similar risk as someone with half the PSA. Of course, this can be overcome with some leeway in categorisation and experienced clinicians are highly unlikely to treat it as a hard-line. Nevertheless, this does mean that more clinical expertise and nuance is required rather than a standardised system or model that can appropriately appraise patients.

Other risk stratification models do exist and are implemented in other healthcare systems, such as the cancer of the prostate risk assessment (CAPRA) score[88]. CAPRA utilises a 0 - 10 score and, similarly to D'Amico, stratifies patients based upon the predicted risk of recurrence following radical prostatectomy[63,88]. Whilst CAPRA has been demonstrated to discriminate prostate cancer better than D'Amico through multiple studies[88–90], there is no mention of CAPRA in NICE or EAU guidance on diagnosis, likely due to the requirements of detailed histopathological information at the time of scoring that, in my experience is not always available.

2.6 Biomarker discovery and development

The methods described above assess patients on average, providing benefit across a population or cohort, but lacking specificity or confidence in individual patients. This non-specificity is not aided by the treatment of patients based on the clinical appearance of their cancer, with little consideration for the underlying biology and intracellular environment of individual tumours. Personalised medicine and more specific biomarkers could be used to provide an alternative approach to this, guiding diagnosis and treatment based upon precise disease-specific markers collected from the prostatic transcriptome, proteome, genome, epigenome, or from any combination of these.

The concept of personalised medicine has brought about changes in how researchers consider pathogenesis, taking a more holistic approach considering altered biological pathways and processes as a whole rather than the historical search for individual biomarkers of disease state. Complex diseases, such as cancer, have complex causes and so, this change has been of particular benefit to such polygenic diseases that typically involve large numbers of genes, molecular processes and environmental factors all acting simultaneously to invoke the functional changes observed at the tissue and cellular level[91]. With the advent of high-throughput technologies a multitude of genomic and proteomic biomarkers have already been identified as potential predictors for prostate cancer. The aforementioned complexity of prostate cancer means that there is a very low likelihood of a single biomarker existing that is capable of explaining a large amount of variance and possessing clinical utility. Instead, it is more likely that clinical benefit can be derived in from panels of already known biomarkers rather than searching for and testing of, novel targets identified through basic research. Such a "targeted" biodiscovery trial was the focus of the Movember GAP1 Urine Biomarker project, and subsequently, this thesis. Through the application of machine learning methods and robust analyses, I hypothesise that a multiplexed panel of already known biomarkers can be generated with potential for strong clinical utility for a urine-based biomarker test.

It is my opinion that the most societal, clinical and economical benefit would come from such a multiplexed biomarker panel that could be administered non-invasively to triage

patients prior to any invasive and costly needle biopsy. With suitable predictive accuracy such a panel would enable repeated monitoring for those considered at risk, reducing the burden of stress and uncertainty for patients. This of course, is unlikely to be stumbled upon and so, directed studies aiming to incorporate such a test into current clinical pathways would need to be carefully designed and considered, rather than attempting to implement wholesale change to how patients are treated. Indeed, before NICE or another clinical body would consider advocating adoption of a test, substantial scientific efforts are required in clinical trials, epidemiological studies and cost-benefit analyses, no small feat. A good example of this is the developers of the PCA3 urine test detailing the approximately 12-year long journey they took from basic discovery of the *DD3* gene through to approval by the US Food and Drug Administration of the PCA3 urine biomarker test, predicting the likelihood of a cancer finding on re-biopsy of a patient[92].

2.6.1 Why urine?

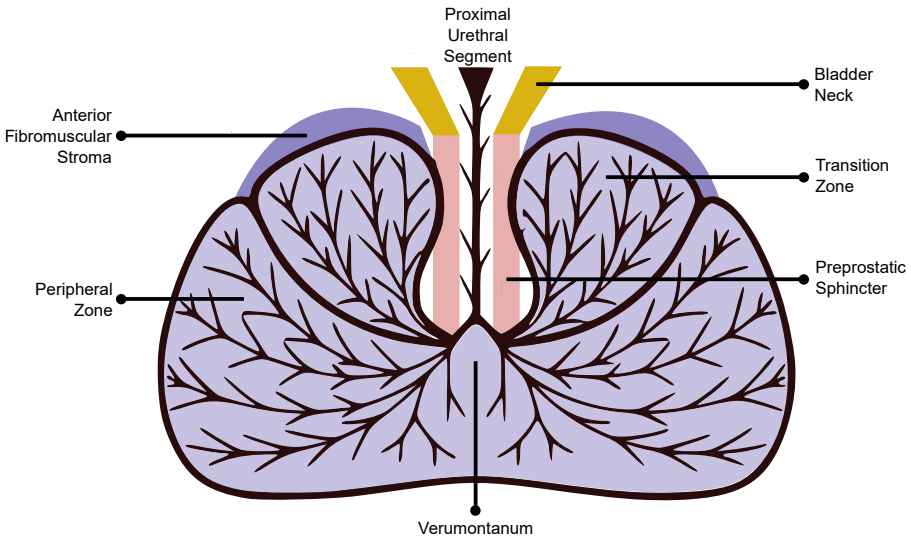


Figure 2.6: Oblique coronal section of the prostate showing the branching pattern of the prostatic ducts, where medial transition zone ducts penetrate the sphincter. Adapted from Abdominal Key

Liquid biopsy is a minimally invasive technique that has gained large scale adoption in prospecting for novel biomarkers of urologic malignancies in recent years, with blood, urine and semen all explored[10]. The average age range of those presenting with prostate cancer somewhat precludes the use of semen as a universally convenient source of biomarkers. Whilst venepuncture is simple, the proteolytic activity of serum is higher than in urine and the large volume of blood in humans dilutes the concentration of biomarkers dramatically[93]. It is the ductal nature of the prostate and interconnectedness of the male urological system that makes urine potentially the optimal means for convenient sampling of almost the entire prostate directly (Figure 2.6). Sloughed off cells, secreted proteins, nucleic acids and extracellular vesicles from both normal and cancerous tissue can all find their way into the urine through these prostatic ducts that drain into the urethra[12–15].

Following a DRE it has been well-documented that cells from the prostate, proteins, and

markers strongly associated with prostate cancer such as *PCA3* and the *TMPRSS2:ERG* gene-fusion can be detected within the urine[12–15]. Indeed, several urine-based tests for use in diagnosing prostate cancer have now been developed and are in various stages of validation (see Section 2.6.2), showing that urine-based liquid biopsy may very well have the potential to augment the wide-spread use of invasive tissue biopsy. There are of course several technical complications that need to be surmounted for urine sampling to be widely adopted; urine samples exhibit large variability between samples, dependent on factors such as sample volume, protein concentration, pH, RNA yield and, if a DRE is performed, DRE efficiency. Much of this variation can be accounted for with adoption of strict collection protocol, and to some extent, mitigated with normalisation and careful selection of markers robust to such variation.

2.6.2 Existing urine biomarker tests

Several prognostic biomarkers and clinical tests have been developed for a range of uses throughout the clinical pathway of prostate cancer. Each has their own specific use, many filling the niches where PSA fails to perform, such as when to repeat a negative biopsy where clinical suspicion remains, or predicting cancer specific mortality from a biopsy sample. Some have been established and FDA-approved for a number of years, whilst others are relatively new and are still being evaluated for clinical efficacy. Despite the demonstrable clinical utility of these tests, they have yet to reach widespread clinical adoption. Furthermore, most of the currently approved or marketed biomarkers for prostate cancer to-date are tissue-based assays relying on tumour samples from a needle biopsy which, as previously discussed, is less than ideal in its currently implemented form. Regardless of needle biopsy accuracy, the requirement for tissue precludes the use of tissue-based tests as triage devices and so are not within the scope of this thesis.

The PCA3 test

Prostate cancer antigen gene 3 (*PCA3*) is a prostate-specific non-coding RNA that was originally described as the *DD3* gene in 1999 by Bussemakers *et al.*[94]. *PCA3* is expressed at rates 60 to 100 times higher in prostate cancer tissue than in normal tissue and transcripts can be detected in the urine sediment[13]. Twelve years after the initial discovery, the PCA3 test gained FDA approval, and uses quantitative amplification of *PCA3* and PSA transcripts in the urinary cell pellet from post-DRE samples, scaled to *KLK3* levels. Licensed by Gen-Probe, the clinical application of the test is to predict the likelihood of a cancer finding on a repeat biopsy, following an initial negative one. This is a grey area in the clinical pathway where the uncertainty surrounding biopsy accuracy can mean a clinical suspicion remains if PSA levels and the DRE indicated the presence of disease.

The PCA3 test has been shown to have lower sensitivity than serum PSA, but significantly higher specificity, positive predictive value (PPV) and negative predictive value (NPV)[95]. Modifications to the original test by including further biomarkers, including the *TMPRSS2-ERG* gene fusion, has been shown to markedly increase predictive utility for predicting outcome of subsequent biopsies[96,97], showing the potential for multiple markers to increase clinical performance.

Whilst analytically useful, the PCA3 test is currently recommended against by NICE, as they concluded that it does not provide an overall net benefit in the clinic above currently implemented standards of care[5]. Most patients with an initial negative biopsy are likely to

have lower volume, lower risk disease that does not require immediate intervention and in the absence of a PCA3 test, would not be detected. Subjecting these patients to further biopsy does not significantly improve their outcome, and only does so at high costs to healthcare providers, where it averages £178.70 per test[5,98].

SelectMDx

SelectMDx is a urine test providing two likelihoods for interpretation by clinicians; the probability of any cancer being present on biopsy, and the probability for high-grade versus low-grade disease. The test quantifies the mRNA expression levels of three genes, *DLX1*, *HOXC6* and *KLK3* and incorporates information from clinically available risk factors to produce the risk score. It is designed to decrease the number of unnecessary biopsies; at cut-offs with an NPV of 98% for Gleason ≥ 7 cancer the decrease in total biopsies performed is estimated to be 42%.[99]

Since initial development in 2016, the SelectMDx test has now been calibrated and validated in a large, multicentre study of 1,955 patients. This trial reported AUCs around 0.8, but the assay suffers from low specificity (53%)[100]. This has led it to have a restricted use-case in biopsy naïve patients with a PSA <10 ng/mL. Drawbacks considered, a health economics analysis found that SelectMDx improved health outcomes and lowered costs for American patients at risk of prostate cancer[101]. As the US healthcare landscape is starkly different to European and UK systems, this benefit remains to be quantified in a single-payer or nationalised healthcare system such as the NHS and so is currently not endorsed by NICE.

ExoDx Prostate (IntelliScore)

Granted FDA Breakthrough Designation in 2019, the ExoDX Prostate (IntelliScore) (EPI) is another fully-realised urine-based test reportedly based on the exosomal expression signature of three gene transcripts; *ERG*, *SPDEF* and *PCA3*. The EPI test is intended for patients 50 years of age or older with a PSA between 2 - 10 ng/mL and determines the patient's risk of clinically significant (Gleason 7) prostate cancer upon biopsy. Derived from qPCR-derived values input to a simple regression formula to generate an EPI output that is transformed to between 0 and 30.

EPI has been validated in two prospective trials, at first in a more general trial to define the intended use-case population and thresholds ($n = 499$)[102], followed by a registered prospective adaptive clinical trial to validate the performance in a large external cohort ($n = 503$)[103]. The EPI test shows good clinical utility for pre-biopsy prediction, outperforming the standards of care within their respective studies, though only marginally, and integration of EPI and standards of care did not yield significant uplifts in predictive ability. Interestingly, despite a continuous EPI score, the test result is dichotomised at a threshold of 15.6, a suboptimal approach that is likely to discard useful information[104].

2.7 The applications of machine learning for biodiscovery & prostate cancer

Much of oncology can be reduced to a problem of prediction and probability estimation. Indeed, the initial decision to biopsy is based on a prediction of the likelihood of a tumour being found, and in advanced disease treatment, where decisions about life extending versus

palliative care inherently involve the use of survival analyses[105]. With regards to prostate cancer, the decision to initiate a search for a tumour is dictated by the elevated likelihood of significant cancer, often as interpreted by the primary care physician, using some heuristic weighted combination of risk factors such as PSA levels, age, family history and DRE findings. In ideal settings these predictors would allow for the design of robust models that could objectively quantify this likelihood rather than the qualitative and conditional assessment from clinicians and patients in their absence.

Treatment decisions are similarly affected, where the outcome of a biopsy is coupled with numerous shared decisions from patients, urologists, oncologists, radiologists and other care providers to designate the best course of action in light of a prostate cancer diagnosis. Again, no quantitative and objective methods exist within the clinical pathway to better guide these decisions, appropriately appraising patient risk.

The biology of prostate cancer is highly complex, and cancer researchers must generally deal with high dimensional, noisy data with innumerable confounding factors from epidemiological and societal levels, through to biases in data collection and processing. Machine learning has recently come to the forefront in attempts to cut through the noise and find the signal, extracting useful insights about diseases that can be used to improve patient care, either directly through predictive clinical models, or indirectly by deconvoluting the biology driving a disease.

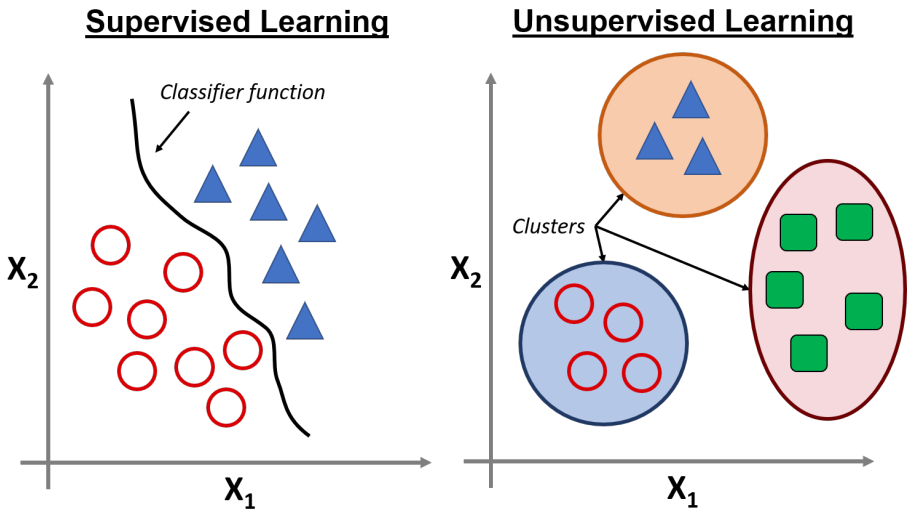


Figure 2.7: Simplified examples of supervised and unsupervised learning methods with two variables, X_1 and X_2

Machine learning is a sub-field of statistics and computer science, concerned with pattern recognition at its base level[106]. A machine learning algorithm is anything that uses previously generated data to "learn" the parameters and coefficients of a statistical model that can describe the system without being explicitly programmed to do so. This statistical model can then be applied to new, previously unseen data to generate predictions about the likely class or value of a given sample based on the previous observations. Due to the dramatic reduction in the costs of computation over recent decades, machine learning has become the go-to technique in numerous fields including natural language processing, computer vision, autonomous driving, computational biology and, where this thesis is most

interested, medical diagnostics and prognostics.

Machine learning methods typically fall under one of two umbrella terms; supervised and unsupervised (semi-supervised reinforcement approaches do exist, but I shall not mention them further in this work[107]). Supervised methods use labelled training data such as biopsy outcome, where the outcome is known *a priori*, to learn a function that can adequately describe the system and be generalised to unlabelled data, correctly determining its value[108]. Unsupervised methods rely on structure within the data to cluster objects with similar attributes together[108]. Where unsupervised methods attempt to define the class to which data should belong without input, supervised methods attempt to define what is different between predetermined groups (Figure 2.7). Whilst simple to visualise and model with only two variables, real-world problems can have many hundreds of dimensions, making subsequent decision spaces humanly impossible to visualise or perceive. There are many debates and strong opinions about what constitutes a statistical or machine learning method, in this thesis I consider there to be no difference between a statistical model and a machine learning one bar complexity, with the two terms used interchangeably.

2.7.1 The "Black Box" issue

Machine learning techniques have been at the core of many recent advances in cancer research, including prognostic models for prostate cancer[8,99,109–112]. However, one issue commonly overlooked is the role of humans and trust in interpretation of prediction model outputs. If a model is to be used, it must be trusted in both its predictions and behaviour, to do no harm. This is trivial for cases of simple regressions where model coefficients can be related directly to covariates, outcomes and predictions for unseen ranges of data. However, as the complexity of a model builds, not only does its explanatory power grow but also its opacity[113]. When complex machine learning methods such as artificial neural networks and gradient boosting are employed, this opacity can be pushed to make models impossible to be interpreted by human brains, or so called "black boxes". Even simple neural networks can have thousands of intermediate parameters hidden in each layer, obscured from the input layers, with changes of each affecting not only the model output, but other downstream parameters too[114]. Attempts have been made to reduce the opacity of certain complex machine learning algorithms, including specialised neural network structures that show intermediate predictions or simple linear models that can locally model the decision space of another, more complex model to explain what led to a certain predicted value[115,116].

This thesis will aim to ameliorate these issues by carefully considering whether more complex algorithms yield sufficient improvements in predictive ability to warrant their opacity. The inputs to models will also be investigated in more detail, exploring distributions and expression patterns across differing severities of prostate cancer and interactions with other clinical features of the disease.

2.8 Discussion

In this chapter the primary biological and medical background of prostate cancer relevant to this thesis has been discussed. The diagnostic and prognostic challenges that clinicians face with current guidelines highlights the critical need for more precise testing methods for patients suspected to harbour prostate cancer, ideally in a non-invasive triage setting. Urine can provide a solution for this; due to the interconnected nature of the male urinary tract the prostate is well placed for sampling via liquid biopsy. Indeed, there are existing

biomarker panels that utilise urine for disease appraisal, including the now defunct PCA3 test and the newly validated SelectMDx test that shows great promise if approved by the FDA, NICE and other regulatory bodies.

The path to clinical adoption of any prognostic test is long and full of pitfalls as demonstrated by the PCA3 test that took 12 years to approval only to be recommended against soon after[92]. With this considered, the careful design and development of prognostic models in collaboration with practising clinicians is key to ensuring a smoother journey to the clinic and improving patient care. The more targeted experimental approaches for biomarker discovery implemented in here pose less unknowns when compared to whole 'omic approaches and the *a priori* reduction of variables for consideration makes robust test development less complex overall. As will be explored in later chapters, the application of machine learning techniques to such datasets needs to be carefully considered so as not to produce overly optimistic results that then cannot be replicated.

Chapter 3

Methods

3.1 The Movember GAP1 Urine Biomarker Cohort

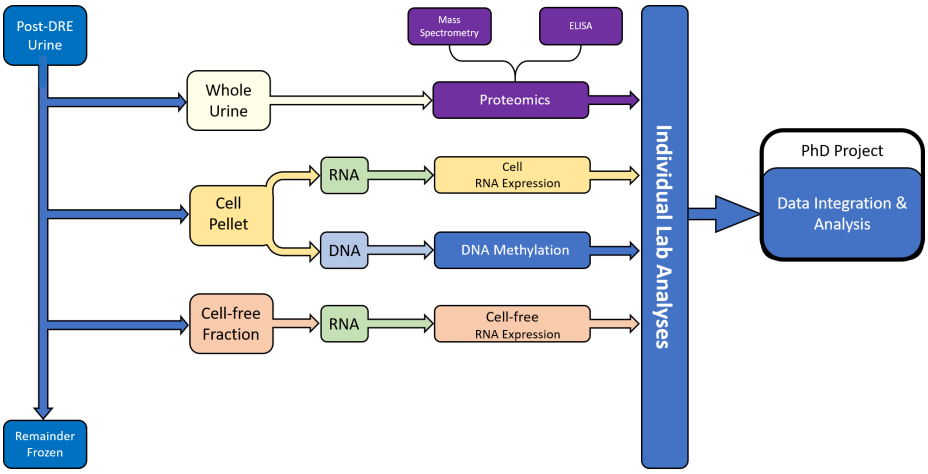


Figure 3.1: The workflow of samples within the Movember GAP1 urine biomarker project. Samples were fractionated and distributed to collaborating laboratories for individual analyses. Not all samples received all analyses.

The Movember GAP1 Urine Biomarker Cohort comprised of first-catch post-digital rectal examination (DRE) urine samples collected at diagnosis between 2009 and 2015 from urology clinics at the Norfolk and Norwich University Hospital (NNUH, Norwich, UK), Royal Marsden Hospital (RMH, London, UK), St. James's Hospital (Dublin, Republic of Ireland), urology clinics within the University Health Network (UHN, Toronto, Canada), and from primary care and urology clinics of Emory Healthcare (Atlanta, USA). Samples were processed as three fractions; whole-urine aliquots, cell-pellet and cell-free, isolated according to the Movember GAP1 Protocol described by Connell *et al.*[8] (Figure 3.1).

Sample collections and processing were ethically approved in their country of origin: NNUH samples by the East of England REC, Dublin samples by St. James's Hospital. iii) RMH by the local ethics committee, iv) Emory Healthcare samples by the Institutional Review board of Emory University, and v) UHN samples by the research ethics boards of all centres and Sinai Health System, Toronto, Canada. Trans-rectal ultrasound (TRUS) guided biopsy was used to provide biopsy information.

Within the Movember GAP1 cohort were 87 patients enrolled on an Active Surveillance (AS) programme at the RMH[117], subsequently known as the Movember GAP1 AS Cohort. Eligibility criteria for this AS programme included histologically proven prostate cancer, age 50–80, clinical stage T1/T2, PSA < 15 ng/mL, Gleason $\leq 3+3$ (Gleason $\leq 3+4$ if age > 65), and $< 50\%$ percent positive biopsy cores. Progression was defined as the detection of disease by clinical criteria that typically triggers the requirement for therapeutic intervention. Clinical criteria of progression were either: PSA velocity >1 ng/mL per year or adverse histology on repeat biopsy, defined as primary Gleason ≥ 4 or $\geq 50\%$ biopsy cores positive for cancer. mpMRI criteria for progression were either: detection of >1 cm3 prostate tumour, an increase in volume >100% for lesions between 0.5 - 1 cm^3, or T 3/4 disease[117].

Figure 3.2: Available sample numbers within each of the Movember GAP1 datasets analysed. Numbers shown are unprocessed sample sizes prior to filtering or preprocessing of datasets.

The original intention of the GAP1 project was to assay all samples by all analytical techniques. Practical constraints limited this, resulting in only a very small number of samples receiving all analyses ($n = 13$, data not shown). The overlap of available samples between assay methods was primarily limited to pairs, with the largest commonly with NanoString data and one other source. Five datasets in total were considered for analyses, with the individual datasets and overlaps fully described where used (Figure 3.2).

3.1.1 NanoString

Figure 3.3: Representation of the NanoString nCounter hybridisation system using reporter and capture probes, including a selection of fluorescently labelled beads on the reporter probe, assigned to a specific detection target for digital quanitification. Adapted from NanoString's marketing materials and protocols.

NanoString is a method of direct digital quantification of gene transcripts utilising microscopy and bar-coded hybridisation of specific probes. The NanoString nCounter system uses two probes; a capture and reporter probe. The probes are designed to have a complementary sequence to specific transcripts corresponding to genes of interest. Each gene-probe has a distinct string of fluorescently labelled beads that can be observed as a colourimetric barcode under oil-immersion microscopy and automatically registered with NanoString's software, with the possibility to multiplex up to 800 possible gene-probes in a single assay. Capture probes are electrophoretically pulled down and immobilised onto a capture surface, with unbound sequences removed. The gene reporter probes are then hybridised to complementary nucleic acid sequences of the capture probe to form a conjugate of capture, reporter, and target sequences (Figure 3.3).

NanoString expression analysis of the Movember GAP1 cohort samples consisted of 167 probes representing 164 genes (Table 3.1). Quantification was performed at the Human Dendritic Cell Laboratory, Newcastle University using 100 ng of cDNA that was produced by amplification of extracted RNA from samples. 137 of the gene-probes were selected based on previously proposed controls alongside diagnostic and prognostic prostate cancer biomarkers within tissue and control probes. 30 additional probes were selected as overexpressed in prostate cancer samples when next generation sequence data generated from 20 urine derived cell-free RNA (cf-RNA) samples were analysed (data not shown). Target gene sequences were provided to NanoString, who designed the probes according to their protocols[118].

Table 3.1: Gene-probes included on the NanoString assay in the Movember GAP1 cohort

Gene	Full Name	*Gene*	Full Name
AATF	apoptosis antagonizing transcription factor	*MEX3A*	mex-3 RNA binding family member A
ABCB9	ATP binding cassette subfamily B member 9	*MFSD2A*	major facilitator superfamily domain containing 2A

Table 3.1: Gene-probes included on the NanoString assay in the Movember GAP1 cohort *(continued)*

Gene	Full Name	*Gene*	Full Name
ACTR5	ARP5 actin-related protein 5 homolog	*MGAT5B*	mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase, isozyme B
AGR2	anterior gradient 2, protein disulphide isomerase family member	*MIR146A*	microRNA 146a
ALAS1	5'-aminolevulinate synthase 1	*MIR4435-2HG*	MIR4435-2 host gene
AMACR	alpha-methylacyl-CoA racemase	*MKI67*	marker of proliferation Ki-67
AMH	anti-Mullerian hormone	*MME*	membrane metalloendopeptidase
ANKRD34B	ankyrin repeat domain 34B	*MMP11*	matrix metallopeptidase 11
ANPEP	alanyl aminopeptidase, membrane	*MMP25*	matrix metallopeptidase 25
APOC1	apolipoprotein C1	*MMP26*	matrix metallopeptidase 26
AR ex 9	Androgen Receptor splice variant	*MNX1*	motor neuron and pancreas homeobox 1
AR ex 4-8	Androgen Receptor	*MSMB*	microseminoprotein beta
ARHGEF25	Rho guanine nucleotide exchange factor 25	*MXI1*	MAX interactor 1, dimerization protein
AURKA	aurora kinase A	*MYOF*	myoferlin
B2M	beta-2-microglobulin	*NAALADL2*	N-acetylated alpha-linked acidic dipeptidase like 2
B4GALNT4	beta-1,4-N-acetyl-galactosaminyltransferase 4	*NEAT1*	nuclear paraspeckle assembly transcript 1 (non-protein coding)
BRAF	B-Raf proto-oncogene, serine/threonine kinase	*NKAIN1*	Na+/K+ transporting ATPase interacting 1
BTG2	BTG anti-proliferation factor 2	*NLRP3*	NLR family pyrin domain containing 3
CACNA1D	calcium voltage-gated channel subunit alpha1 D	*OGT*	O-linked N-acetylglucosamine (GlcNAc) transferase
CADPS	calcium dependent secretion activator	*OR51E2*	olfactory receptor family 51 subfamily E member 2
CAMK2N2	calcium/calmodulin dependent protein kinase II inhibitor 2	*PALM3*	paralemmin 3

Table 3.1: Gene-probes included on the NanoString assay in the Movember GAP1 cohort *(continued)*

Gene	Full Name	*Gene*	Full Name
CAMKK2	calcium/calmodulin dependent protein kinase kinase 2	*PCA3*	prostate cancer associated 3 (non-protein coding)
CASKIN1	CASK interacting protein 1	*PCSK6*	proprotein convertase subtilisin/kexin type 6
CCDC88B	coiled-coil domain containing 88B	*PDLIM5*	PDZ and LIM domain 5
CDC20	cell division cycle 20	*PLPP1*	phospholipid phosphatase 1
CDC37L1	cell division cycle 37 like 1	*PPFIA2*	PTPRF interacting protein alpha 2
CDKN3	cyclin dependent kinase inhibitor 3	*PPP1R12B*	protein phosphatase 1 regulatory subunit 12B
CERS1	ceramide synthase 1	*PSTPIP1*	proline-serine-threonine phosphatase interacting protein 1
CKAP2L	cytoskeleton associated protein 2 like	*PTN*	pleiotrophin
CLIC2	chloride intracellular channel 2	*PTPRC*	protein tyrosine phosphatase, receptor type C
CLU	clusterin	*PVT1*	Pvt1 oncogene (non-protein coding)
COL10A1	collagen type X alpha 1 chain	*RAB17*	RAB17, member RAS oncogene family
COL9A2	collagen type IX alpha 2 chain	*RIOK3*	RIO kinase 3
CP	ceruloplasmin	*RNF157*	ring finger protein 157
MIATNB	MIAT neighbour	*MRPL46*	mitochondrial ribosomal protein L46
DLX1	distal-less homeobox 1	*RPL18A*	ribosomal protein L18a
DNAH5	dynein axonemal heavy chain 5	*RPL23AP53*	ribosomal protein L23a pseudogene 53
DPP4	dipeptidyl peptidase 4	*RPLP2*	ribosomal protein lateral stalk subunit P2
ECI2	enoyl-CoA delta isomerase 2	*RPS10*	ribosomal protein S10
EIF2D	eukaryotic translation initiation factor 2D	*RPS11*	ribosomal protein S11
EN2	engrailed homeobox 2	*SACM1L*	SAC1 suppressor of actin mutations 1-like (yeast)

Table 3.1: Gene-probes included on the NanoString assay in the Movember GAP1 cohort *(continued)*

Gene	Full Name	*Gene*	Full Name
TMPRSS2/ERG	transmembrane protease, serine 2/ERG fusion	*SCHLAP1*	SWI/SNF complex antagonist associated with prostate cancer 1 (non-protein coding)
ERG	ERG, ETS transcription factor	*SEC61A1*	Sec61 translocon alpha 1 subunit
ERG 3 ex 4-5	ERG, ETS transcription factor	*SERPINB5*	serpin family B member 5
ERG3 ex 6-7	ERG, ETS transcription factor	*SFRP4*	secreted frizzled related protein 4
FDPS	farnesyl diphosphate synthase	*SIM2*	single-minded family bHLH transcription factor 2
FOLH1	folate hydrolase 1	*SIM2*	single-minded family bHLH transcription factor 2
GABARAPL2	GABA type A receptor associated protein like 2	*SIRT1*	sirtuin 1
GAPDH	glyceraldehyde-3-phosphate dehydrogenase	*SLC12A1*	solute carrier family 12 member 1
GCNT1	glucosaminyl (N-acetyl) transferase 1, core 2	*SLC43A1*	solute carrier family 43 member 1
GDF15	growth differentiation factor 15	*SLC4A1*	solute carrier family 4 member 1
GJB1	gap junction protein beta 1	*SMAP1*	small ArfGAP 1
GOLM1	golgi membrane protein 1	*SMIM1*	small integral membrane protein 1 (Vel blood group)
HIST1H1C	histone cluster 1 H1 family member c	*SNCA*	synuclein alpha
HIST1H1E	histone cluster 1 H1 family member e	*SNORA20*	Small nucleolar RNA SNORA20
HIST1H2BF	histone cluster 1 H2B family member f	*SPINK1*	serine peptidase inhibitor, Kazal type 1
HIST1H2BG	histone cluster 1 H2B family member g	*SPON2*	spondin 2
HIST3H2A	histone cluster 3 H2A	*SRSF3*	serine and arginine rich splicing factor 3
HMBS	hydroxymethylbilane synthase	*SSPO*	SCO-spondin
HOXC4	homeobox C4	*SSTR1*	somatostatin receptor 1

Table 3.1: Gene-probes included on the NanoString assay in the Movember GAP1 cohort *(continued)*

Gene	Full Name	*Gene*	Full Name
HOXC6	homeobox C6	*ST6GALNAC1*	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 1
HPN	hepsin	*STEAP2*	STEAP2 metalloreductase
HPRT1	hypoxanthine phosphoribosyltransferase 1	*STEAP4*	STEAP4 metalloreductase
IFT57	intraflagellar transport 57	*STOM*	stomatin
IGFBP3	insulin like growth factor binding protein 3	*SULF2*	sulfatase 2
IMPDH2	inosine monophosphate dehydrogenase 2	*SULT1A1*	sulfotransferase family 1A member 1
ISX	intestine specific homeobox	*SYNM*	synemin
ITGBL1	integrin subunit beta like 1	*TBP*	TATA-box binding protein
ITPR1	inositol 1,4,5-trisphosphate receptor type 1	*TDRD1*	Tudor domain containing 1
KLK2	kallikrein related peptidase 2	*TERF2IP*	TERF2 interacting protein
KLK3 ex 1-2	kallikrein related peptidase 3	*TERT*	telomerase reverse transcriptase
KLK3 ex 2-3	kallikrein related peptidase 3	*TFDP1*	transcription factor Dp-1
KLK4	kallikrein related peptidase 4	*TIMP4*	TIMP metallopeptidase inhibitor 4
LBH	limb bud and heart development	*TMCC2*	transmembrane and coiled-coil domain family 2
POTEH-AS1	POTEH antisense RNA 1 (POTEH-AS1), long non-coding RNA. prostate-specific P712P mRNA	*TMEM45B*	transmembrane protein 45B
MAK	male germ cell associated kinase	*TMEM47*	transmembrane protein 47
MAPK8IP2	mitogen-activated protein kinase 8 interacting protein 2	*TMEM86A*	transmembrane protein 86A
Mar-05	membrane associated ring-CH-type finger 5	*TRPM4*	transient receptor potential cation channel subfamily M member 4

Table 3.1: Gene-probes included on the NanoString assay in the Movember GAP1 cohort *(continued)*

Gene	Full Name	*Gene*	Full Name
MCM7	minichromosome maintenance complex component 7	*TWIST1*	twist family bHLH transcription factor 1
MCTP1	multiple C2 and transmembrane domain containing 1	*UPK2*	uroplakin 2
MDK	midkine (neurite growth-promoting factor 2)	*VAX2*	ventral anterior homeobox 2
MED4	mediator complex subunit 4	*VPS13A*	vacuolar protein sorting 13 homolog A
MEMO1	mediator of cell motility 1	*ZNF577*	zinc finger protein 577
MET	MET proto-oncogene, receptor tyrosine kinase		

Counts quantified from NanoString platforms require normalisation to account for the amount of sample, variations in assay efficiency and other factors that influence non-biological variability. Positive control sequences of known concentrations are included by NanoString for assessing quality control, along with negative probes that do not align with any part of the human transcriptome to assess non-specific binding of sample material.

Unless otherwise specified normalisation of NanoString data was performed using the *NanoStringNorm* R package for preprocessing and followed the recommended protocols from NanoString[118]. This consisted of confirming binding density was in a suitable range (0.1 - 2.2), with failing samples removed from further analysis ($n = 14$). Positive control normalisation, considered the most fundamental normalisation step, used correction factors calculated from NanoString-supplied positive control probes for each sample. The correction factor CF was calculated for a given sample i by using the geometric mean G of the positive controls across n samples to divide the arithmetic mean of G:

$$CF_i^{Pos} = \frac{\sum_n^i (G_{Pos_i})}{n(G_{Pos_i})}$$

This positive control normalisation attempts to correct for technical variance introduced between NanoString cartridges and within nCounter runs or between differing nCounter machines. Due to the inclusion of known quantities of controls, the ground-truth is very well known and so, positive control normalisation represents the least error prone (both technical and assumption-wise) of all the methods. Samples with $0.3 > CF > 3$ were removed ($n = 13$).

Following positive control normalisation estimation of non-specific binding was estimated by use of the negative control probes. Counts quantified from gene-probes were thresholded according to the calculated background, with any counts less than $\mu_{Background} + 2SD$ from the background set to 0.

3.1.2 Methylation

The Epigenetic Cancer of the Prostate Test in Urine (epiCaPture) is a multibiomarker panel developed to quantitatively measure DNA hypermethylation at the 5′-regulatory regions of six genes previously associated with prostate cancer (GSTP1, SFRP2, IGFBP3, IGFBP7, APC, and PTGS2), in the urinary cell pellet fraction[119–122]. These data were generated by collaborators in University College Dublin and Trinity College Dublin by assaying urinary cell pellet samples within the Movember GAP1 cohort, and were previously described by O'Reilly *et al.*. The Infinium HumanMethylation450 BeadChip (HM450k) assay kit was used to assay samples and return methylation values for each gene.

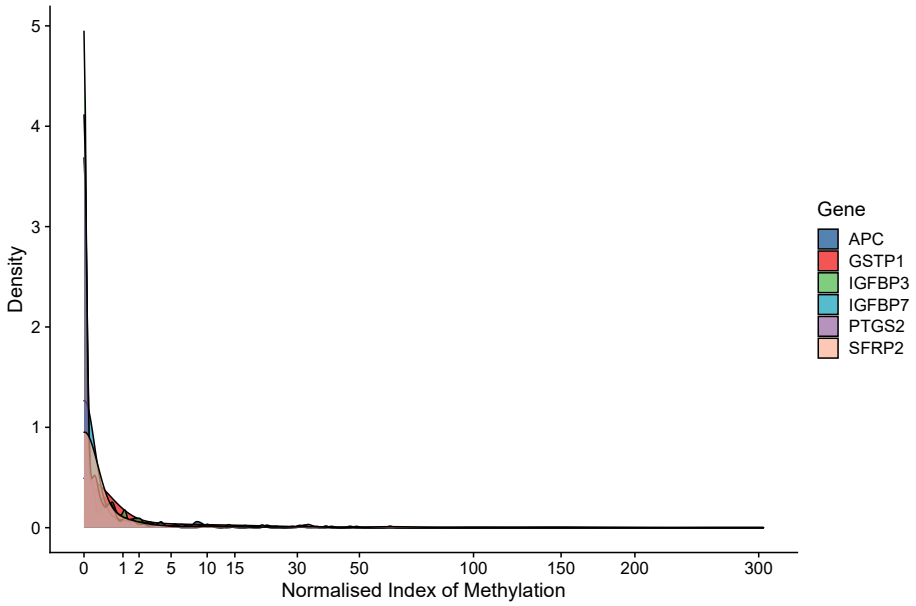


Figure 3.4: The extremely skewed distribution of the normalised index of methylation values for each quantified gene within the methylation dataset

Methylation data ($n = 406$) are presented as the "normalised index of methylation" (NIM), calculated relative to a 100% DNA methylation control and to a control qPCR reaction that only amplifies bisulfite-modified DNA. NIM represents the methylation percentage of a gene relative to this standard, with 0 representing no methylation quantified. The data are highly skewed for all genes, with no methylation quantified for most samples (Figure 3.4). No further alterations were made to the data prior to use.

3.1.3 ELISA and EN2

Urinary levels of 10 proteins were quantified by enzyme linked immunosorbent assay (ELISA) performed by three different collaborating laboratories. Proteins previously associated with prostate cancer MSMB, GDF15 and CD10[123,124] were quantified by collaborators at University College London (UCL), along with urinary creatinine levels, hypothesised to normalise urine concentration based on kidney function. Five proteins from the kallikrein family were assayed by collaborators at the University of Toronto. Two KLKs highly specific to prostate tissue (KLK2 and KLK4)[125,126] were quantified, with the remaining three (KLK6, KLK7 and KLK11) not regularly overexpressed in prostate tissue

but previously identified as prognostic markers of disease status[127]. Engrailed-2 (EN2) was assayed by collaborators at the University of Surrey, and has been established as a biomarker of prostate and bladder cancers previously[128].

The available ELISA dataset was large, encompassing 471 samples. All proteins were reported as concentration (ng/mL), with no preprocessing or normalisation undertaken.

3.1.4 Mass Spectrometry

Capillary electrophoresis mass spectrometry (CE-MS) was undertaken by collaborators at Mosaique Diagnostics, and followed previously established protocols for sample preparation and data acquisition[129]. Briefly, whole urine aliquots were digested with a 2M urea solution, and fractionated by ultracentrifugation to retain proteins and polypeptides <20 kDa. The fractionated extracts were desalted and lyophilised, then re-suspended in high-performance-liquid-chromatography grade water for CE-MS detection. The peak list of detected peptides was deconvoulted using the proprietary MosaiquesVisu software[130,131]. All detected peptides were deposited, matched, and annotated to a human urinary peptide database maintained by Mosaique Diagnostics and data were presented as raw counts for each peptide, normalised to 29 collagen fragments that are considered invariant and not affected by disease status[132].

The mass spectrometry dataset is sparse, comprising a total of 18,035 peptides quantified across 340 samples. Most peptides were not quantified in more than one sample, and for robust model development required extensive *a priori* filtering before analysis. As recommended by collaborators, peptides quantified in <30% of cancer or non-cancer samples were excluded *a priori*. Preprocessing of mass spectrometry data left 643 possible peptides that were expressed in log_2 units for further analysis.

3.2 Statistical and Machine Learning Methods

This section describes the methods employed within this thesis for producing machine learning models, and statistical analytical techniques used to quantify and test differences. All statistical analysis was undertaken in R 3.5.3, and unless otherwise specified used default parameters and two-tailed tests of significance, with $P < 0.05$ accepted as the threshold for "significance".

3.2.1 Regression modelling

Regression analysis encompasses some of the most widely used and easily grasped statistical techniques for modelling relationships between input and output variables. Linear regression modelling was largely developed in the pre-computer era and is likely the most widely used statistical prediction methods used in both scientific research and commercial applications today. Even with the advent of cutting edge machine learning techniques, linear regression still has a multitude of uses, even being used to describe and interpret complex "black-box" models[116].

Linear regression models a continuous numeric variable Y as some linear product of the input predictors $(X_1, X_2, ..., X_n)$ and their coefficients $(\beta_1, \beta_2, ..., \beta_n)$ according to:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

Interpretation of linear regression models is relatively trivial; the coefficient of an input β_i represents the expected change in Y for each unit change of β_i, when all other inputs are held static. With standard linear regression, unsurprisingly only linear relationships can be modelled due to single coefficients for variables. Polynomial regression introduces higher order powers for coefficients, allowing for curvilinear relationships to be better represented but can quickly introduce overfit with the additional of higher orders. Further alterations to linear regression can be made through link functions to model different relationships of interest and outcomes. Binary events can be modelled using the *logit* link function, that bounds the continous output of a linear regression model into [0,1]. Whilst ordinal event modelling effectively applies thresholds to the output of a linear model, discretising it into the number of categorical events to be modelled.

3.2.2 Overfitting

Overfit is the term ascribed to the event where a function fit to limited data models a relationship too precisely and thus fits to some measured noise, resulting in poor predictive ability with new data. The essence of overfitting is to have unwittingly extracted some of this inherent variance or irreducible error as if it were truly representative of the underlying data structure[133].

Overfit can be introduced through multiple means, including insufficient observations or irreproducible data, though a common method is through the inclusion of too many parameters in a given predictive model. Each additional parameter in a model decreases the error associated with each observation, a key goal for machine learning algorithms, but also increases model complexity and the potential for overfit.

This error term that all statistical models attempt to minimise can essentially be decomposed into three terms; bias, variance and an irreducible error, the noise term inherent to the data that cannot be reduced, such as uncertainty surrounding a measurement:

$$Error_x = Bias^2 + Variance + Error_{Irreducible}$$

Given infinite data and a true model it is possible to reduce bias and variance to 0 simultaneously. In practice however, with both finite data and models that can only approximate relationships, there is a trade-off between bias and variance. Due to the square relationship of error and bias, an overfit model is often completely free of bias, but exhibits exceptionally large variances[133]. Overfit is one of, if not the most common issues in developing machine learning algorithms and models, and so many strategies have been developed within algorithm and model development to ameliorate this. Arguably the most effective method one can undertake to overcome overfit is the adoption of the Principle of Parsimony and minimising the included parameters of a model to reduce its complexity, where large coefficients and unregulated inclusion of parameters in a model can exert undue leverage on its output, reducing accuracy[106]. Of course, producing a simple model for a complex problem is not always feasible, such as in prediction of cancer presence or progression, where many interacting variables meaningful impact on outcome. In this case it is often more appropriate to use computational methods such as regularisation and penalisation, resampling methods or bagging.

3.2.3 Regularisation and the LASSO

As overfit is a prevalent problem, much of machine learning can be framed as an optimisation problem. Regularisation, or shrinkage, methods penalise overly complex models by

restricting or eliminating coefficients to reduce their absolute magnitude and influence on model output. The least absolute shrinkage and selection operator (LASSO) is a regression technique originally described by Tibrashani[134]. LASSO imposes strong penalisation on coefficients, where all but the most informative variables are shrunk to 0. This is achieved by the introduction of an additional error parameter to be minimised during fitting operations and was originally described with the sum of squares error function as defined by

$$\hat{\beta}_{LASSO} = \arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\}$$

Where the scalar λ is a complexity parameter ≥ 0 that controls shrinkage. When $\lambda = 0$ no coefficients are forced to zero, and as the value of λ is increased more coefficients are set to zero. λ must be determined separately and pre-specified before fitting a model to the data, commonly achieved by evaluating the k-fold cross-validated error. LASSO-based regression models in this work are fit using the *ordinalNet* package, selecting the cross-validated λ value returning the minimum error[135].

LASSO penalisation does possess several limitations beyond being a linear error function. One of note is the "large p, small n" case, with more predictors than observations, where LASSO selects at most n variables before it saturates[136]. Additionally, the LASSO struggles to deal with collinearity of variables, selecting a single variable at the cost of all other correlated ones, as opposed to an "all-relevant" approach such as the Random Forest[136,137].

3.2.4 Cross-validation

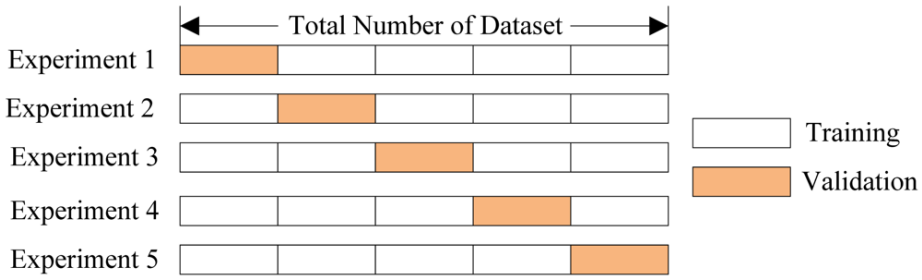


Figure 3.5: Example of the model building process with cross-validation. The full dataset is split into five folds; four are used to train the model, with the remaining fold used to validate the model and assess its fit.

Cross-validation is a method of data splitting for fitting and assessing how well a predictive model can generalise to unseen data, whilst still using all available data to fit a final model[138]. In cross-validated workflows the model fitting process is undertaken on different subsets of the available training data, that are partitioned equally into k folds. A single fold is retained as the validation data for the fitted model, and the remaining $k-1$ folds are used as training data. This process is then repeated k times, with each one of the k folds used exactly once as a validation dataset (Figure 3.5).

Predictions are then produced for each validation fold and the process repeated for another fold. This results in all available data being used to train the final model, whilst still being able to assess the model's accuracy using the results generated when each sample is "out-of-fold". Cross-validation gives a pessimistically biased estimate of performance

because most statistical models will improve if the training set is made larger[106]. Typically, results from models trained on a full dataset perform slightly better than the initial results that tuning via cross-validation on the same data would have suggested. Cross-validation gives a more accurate representation of model quality, and results in a more generalisable model, but requires that k models be trained which can be computationally intensive, and unstable dependent on the dataset[106].

3.2.5 Resampling and the bootstrap

Resampling is a simple concept that consists of drawing repeated samples from an original dataset, either with or without replacing samples. A common method of non-parametric statistical inference, resampling does not involve a reliance on generic distribution tables such as the normal distribution in order to compute approximate P values. Bootstrapping is a specific type of resampling with replacement to the same size as the original dataset, and repeated many times, where it can be used to estimate sampling distributions of an estimate, most often with the purpose of deriving robust estimates of uncertainties around a point estimate.

A key advantage of the bootstrap is in its simplicity, relying on the data itself rather than typical statistical assumptions. Whilst it is impossible to know the true confidence interval or error of a measurement for most problems, the bootstrap-derived estimates have been shown to be more accurate than standard methods using sample variances and distribution-based assumptions[139].

3.2.6 Random Forests

Originally conceived by Tin Kam Ho, the Random Forest algorithm is an ensemble method (one that aggregates the results from more than one model) for both classification and regression problems[140,141]. Further refined and subsequently trademarked by Breiman in 2001, the concept of "bagging" was added and remains today a frequently used algorithm for machine learning applications[142]. The following section is based on the works of Breiman[142], and Hastie, Tibshirani and Friedman[106]. Ensemble learning methods combine multiple individually trained classifiers synergistically to obtain better prediction results than any of the constituent methods alone could achieve[106].

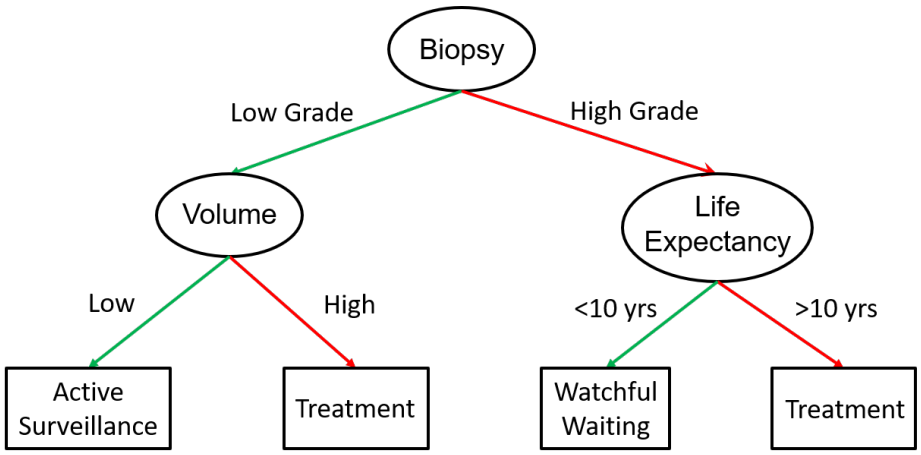


Figure 3.6: Theoretical example of a decision tree for prostate cancer treatment decision. Ellipses represent a tested attribute with squares the decision made following an outcome.

Random Forests function through the generation of an ensemble, or forest, of decision trees. Decision trees are a relatively simple concept and application of supervised learning not dissimilar from a conventional flow chart (Figure 3.6). Decision tree learning is based on a number of base algorithms outside the scope of this work but most often uses the C4.5 algorithm[143]. Decision trees are easily assembled, show low bias and, if small enough, produce directly interpretable models[106]. However, trees in isolation are inaccurate and seldom provide the best prediction accuracy achievable with a given dataset[106]. Trees grown deep, with many branches and decision nodes across multiple variables are able to learn exceptionally non-linear patterns with no underlying assumptions on distributions or linearity, but lead to strong overfitting.

In order to overcome the high variance of single trees, the Random Forest algorithm employs bootstrap-aggregation of predictions, named bagging. Informally for a regression task, bagging functions by taking the output decisions from each decision tree in the forest and taking the average, and in the case of classification tasks, the modal "vote" of the trees. The key goal in bagging is to average many noisy, but unbiased models, to reduce overall variance.

More formally, given a training set of size $Z = [(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)]$, bagging B times selects a random sample with replacement of size n from Z and fits a decision tree to each sample f_{bag}. Predictions can then be made on previously unseen samples $\hat{x}$ by selecting the majority vote for classification, or averaging outputs from all individual fit trees by

$$\hat{f}(\hat{x}) = \frac{1}{B} \sum_{b=1}^{B} f_{bag}(\hat{x})$$

In addition to bagging, Random Forests utilise random feature selection, known as the random subspace method, to further minimise variance in trained models. For M features in a given dataset $m << M$ features are selected at random without replacement. Typically, $m = \sqrt{M}$ for classification or $M/3$ for regression forests, although these are often treated as tunable parameters to optimise model fit.

A key feature of random forests in producing unbiased predictions is the use of strong internal validation through out-of-bag (OOB) samples. Due to the random sampling with

replacement described above, approximately 30% of samples are not selected when growing any given tree, though this is also a tunable parameter. OOB samples are used to validate the trained model so that, for any observation in the dataset Z_i, its prediction is generated using only those trees where Z_i did not appear. This means predictions generated from Random Forests are never actually produced using the same data that portion of the forest was trained on. This produces error metrics similar to other methods such as k-fold cross validation, but as it is intrinsic to the model fitting process, requires no data to be withheld from training at any stage, and results in a single model that can be interrogated, stored and used to produce predictions for unseen data. It is these features that make Random Forests an optimal method for producing predictive models where sample numbers are relatively limited, precluding the use of traditional train/test/validation data splitting techniques, maximising information extraction without risks of overfitting.

Random Forests fit throughout this thesis used the *randomForest* package with default parameters ($m = \sqrt{M}$ for classification and $M/3$ for regression), unless otherwise described[144].

3.2.7 Gradient Boosting Machines

Gradient Boosting Machines (GBMs) are a forward-learning ensemble method, similar in practice to Random Forests, and most commonly implemented using the same decision tree algorithms as the basis for a "weak learner". The driving heuristic behind boosting is that multiple weaker learners can be combined over iterative improvements to become a single, strong learner[145]. Boosting is conceptually similar to bagging described above, however there are key differences in both fitting of individual learners, and in the final aggregation of results.

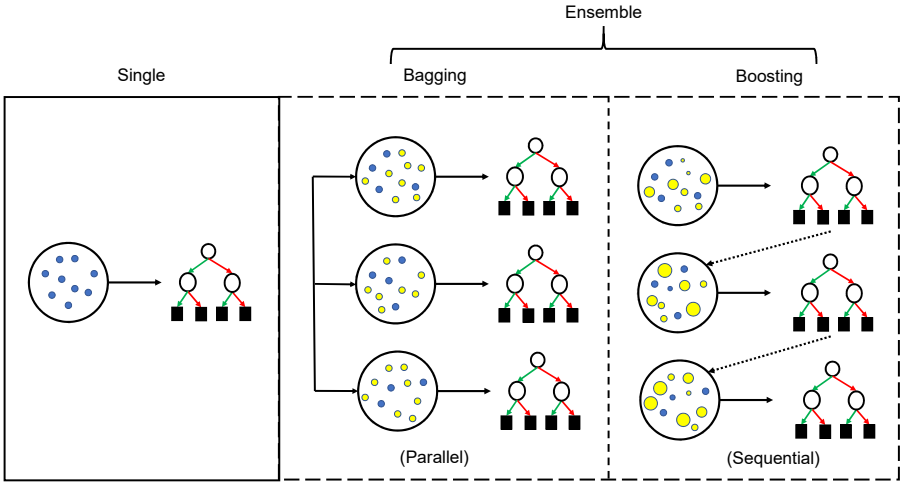


Figure 3.7: Differences in learner construction between singular and ensemble methods. In single learner methods all data (blue circles) are used to produce a single estimate, in this illustration, a decision tree. Bagging randomly samples data, with replacement (yellow circles) many times and estimates many decision trees in parallel. Boosting randomly samples weighted data (shown as size of circles) with replacement. The weight assigned to samples is dictated by the magnitude of misclassification by the previous (or initial) learner. After each boosting round, the weights are recalculated, and a new resampling is used to grow an improved decision tree. The ensemble classification steps follow once the desired number of learners have been constructed.

Where bagging involves random sampling with replacement, with any singular datum being equally likely to be selected, boosting weights data according to previous rates of misclassification; with greater weight given to those data more commonly misclassified. Once a learner is constructed, the weights for each sample are calculated according to a chosen error metric for the next round of boosting. Over each boosting round the learner is improved by essentially concentrating upon the incorrectly classified, hard to learn training data (Figure 3.7). Most specific implementations of boosting include an extra condition to stop sequentially improving a learner, to avoid overfit. This is commonly the point at which the single learner has improved to the point of being able to predict the target slightly greater than random chance. The resulting "weak learner" may not be useful in isolation, but aggregated over an ensemble of many such learners, results in vastly greater performance, similar to the bagging of Random Forests. The key differences and similarities between boosting and bagging approaches are shown in Table 3.2.

Table 3.2: Similarities and differenced between boosting and bagging methods for ensemble learners

Similarities	Differences
Both are ensemble methods to derive N learners from a single dataset	Whilst the learners are built independently for bagging, boosting adds new models that improve upon previous failures.
Both generate training data by random sampling with replacement	Boosting also determines weights to prioritise selection of more difficult cases to classify.
Both make final predictions by the average, or majority decision of the N learners	Boosting additionally weights the average, with more weight to stronger predictors.
Both reduce variance and increase model stability	Boosting reduces bias, at the cost of overfitting. Bagging reduces overfitting but does not reduce bias.

As the construction of each learner in bagging methods is completely independent and non-iterative, it is significantly faster than boosting methods, easily parallelised and distributed for computational efficiency. Several libraries exist that modify the specifics of boosting, depending on the application for improved computational efficiency[145–147], with the *XGBoost* library by far the most popular implementation for high performance use in large datasets. With this considered, GBMs implemented within this thesis use the *xgboost* R package and its C++ backend implementation of the XGBoost libraries[146].

3.2.8 Meta-ensembles or Stacking

Model stacking is a commonly employed approach for marginal gains in machine learning competitions, aiming to improve predictive accuracy by combining the predictions made from multiple models, often with disparate underlying algorithms. Model stacking *usually*, but not always, leads to improved predictive ability, but at the cost of interpretability. Most winning models on the popular data science citizen-science website Kaggle are stacked models[148].

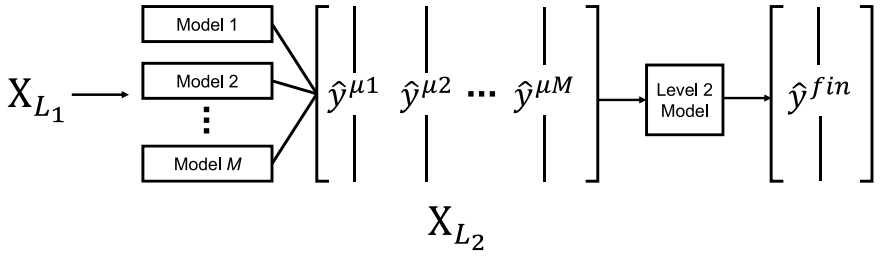


Figure 3.8: A generalised example of constructing a stacked model. M different models are built with the original training data, X_{L_1} where the predictions from these models form a second level dataset X_{L_2} formed of M features and the original number of observations. A second level model(s) can then be trained on this data to produce the final outcomes used for prediction.

Model stacking can range from simple averaging of predicted outcomes, to fitting entirely new models using the estimates from previous models as inputs along with new variables (Figure 3.8). In this thesis, the simple unweighted average will be used to produce meta-ensembles from multiple machine learning models. Preliminary work showed that there was no significant benefit derived from tuning weights used to average outputs (data not shown), with the requirement for more data for tuning outweighing small gains.

3.2.9 Boruta

Boruta is an all-relevant feature selection algorithm designed as a wrapper around the previously described Random Forest classification algorithm. Described by Kursa & Rudnicki in 2010[149], Boruta iteratively compares feature importance against a random predictor generated by permutating real data, deem "shadow features". The importance measure of a feature is determined as the loss of accuracy of classification caused by the random permutation of the feature across all trees in the forest which use that feature for classification. Variables that perform significantly worse compared to the maximally performing shadow feature at each permutation, calculated by Z score difference in mean accuracy decrease are consecutively dropped until only confirmed, stable features remain.

The Boruta algorithm calculates feature importances as follows:

1. Extend the dataset by copying of all variables and permute them to remove their association with the response, creating the shadow features.
2. Fit a Random Forest classifier on the extended dataset, and compute all Z scores.
3. Search for the maximum Z score amongst the shadow features (MZSF), and assign a "hit" to every feature with a higher Z score than MZSF.
4. For each feature with undetermined importance, perform a two-sided test of equality with the MZSF.
5. For each feature with importance significantly lower than MZSF, mark them as "unimportant" and remove from the extended dataset.
6. For each feature with importance significantly higher than MZSF, deem them "important".
7. Remove all shadow features.

8. Repeat the process until importance is assigned to all variables, or the predetermined number of Random Forests have been fitted.

A key aspect of Boruta important to the current work is that it is an *all-relevant* method, as opposed to a *minimal-optimal* selection approach such as LASSO. As features are compared to the random permuted shadow features, correlated and co-linear variables do not impact the importance ranking of one another. The non-zero importance of a shadow feature can only be attributed to randomness, and so can be used to as a reference to for confirming all truly important variables. All Boruta analysis in this thesis uses the *Boruta* package[149] to calculate feature importance, with external bootstrap resampling of 1,000 samples implemented to assess feature-set stability in resampled datasets.

3.2.10 Survival Analysis

Survival analysis is a set of statistical techniques concerned with the modelling of time-to-event data. Within medical research this would commonly be the time between the beginning of an observation period and an outcome, or "event" such as death, disease recurrence or recovery. Survival analysis commonly uses censored data; data where a subject does not have the event during the observation time and so nothing is known about their status after the observation period. This is referred to as right-censoring. Left-censored data are less common, where it is possible for the subject to have previously experienced an event unknown to the observer. An important assumption of working with censored data is that censoring is a random effect, not correlated with the outcome of interest. All survival analyses performed in this thesis use the *survival* and *rms* packages[150–152]

Kaplan-Meier (KM) curves

A common method for representing survival data graphically is through the use of Kaplan-Meier (KM) curves, that represent survival probability (S_t) as a function of time. S_t denotes the probability that a participant does not experience an event in the time t, where t can range from 0 - ∞. In practice, time is never infinite, and so the function may never equal zero across the observation period. The probability of surviving past $t = 0$ is always 1.

The KM survival distribution is a discrete-stepped survivorship curve, gaining information as each event occurs. Two variables define the KM curve at any given time t_j, the number of events $d(t_j)$ and those still at-risk $r(t_j)$. The probability of surviving longer than t or the estimator of the survival function S_{KM} is given by:

$$S_{KM} = \prod_{j:t_j \leq t} (1 - \frac{d(t_j)}{r(t_j)})$$

The survival probability past time t_j, $\hat{S_{t_j}}$ can be calculate as the probability of surviving past the previous time t_{j-1} multiplied by the probability of surviving past time t_j, given survival to at least time t_j:

$$\hat{S_{t_j}} = \hat{S}_{t_j-1} \times P(T > t_j | T \geq t_j)$$

KM curves are most useful when predictor values are categorical, and do not work easily with many categories or continuous values such as a bounded risk score, age, or gene expression.

Cox proportional hazards (PH) model

Modelling of survival influenced by more than one variable can be achieved though regression analysis and the Cox proportional hazards (PH) model. Cox PH modelling is one of the most popular statistical techniques for performing multivariable survival analysis, designed to simultaneously investigate the effects of several explanatory variables on survival time. The Cox PH model has the key assumption that the hazard to any individual over time is proportional to the hazard for any other individual; that the explanatory variables are independent of time.

Cox PH models are primarily used for one of three goals: to test if a variable has an effect on survival, to provide the hazard ratio (HR) for a variable, a point estimate of the effects on survival if one variable is altered, and to provide a confidence interval around the hazard ratio. The hazard function is central to the the Cox PH model and is defined as:

$$h(t, X) = h_0(t) \exp\Big\{ \sum_{i=1}^{p} \beta_i X_i \Big\}$$

Where $X = (X_1, X_2, ..., X_p)$ is a set of p variables, and $[\beta_1, \beta_2, ..., \beta_p]$ are a set of p coefficients corresponding to the variables. The baseline hazard function $h_0(t)$ explains how the hazard changes as a function of time only, prior to consideration of any input variables. The second function is the exponential of a linear combination of the explanatory variables. The HR describes the ratio of hazards rates between two levels of the explanatory variable. For example, if the chances of prostate-cancer specific mortality double for each distant node involved with metastases, then the HR = 2 per distant node.

3.2.11 Metrics for assessing model accuracy

ROC curves and the AUC

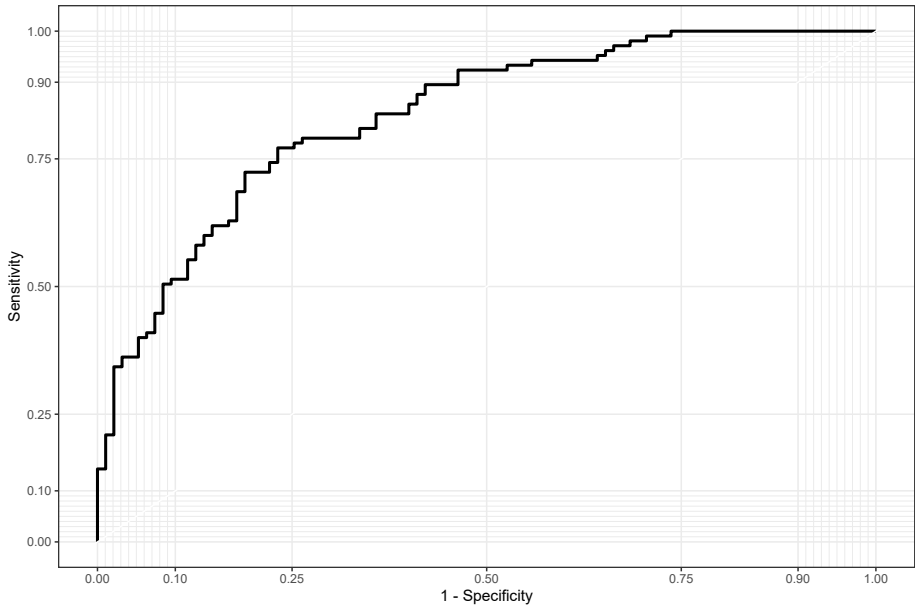


Figure 3.9: Example of a typical ROC plot using sensitivity (y-axis) and specificity (x-axis) to evaluate discriminatory ability over a range of thresholds (not shown).

One of the most commonly used tools when predicting the binary outcomes over a range of probabilities is the Receiver Operating Characteristic (ROC) curve. The ROC curve is a graphical plot of the false-positive rate plotted against the true-positive rate across all candidate thresholds. The true-positive rate, also referred to as sensitivity, is calculated as the number of true positives (TP) divided by the sum of the number of true positives and the number of false negatives (FN):

$$Sensitivity = \frac{TP}{TP + FN}$$

The false positive rate is calculated as the number of false positives divided by the sum of the number of false positives (FP) and the number of true negatives (TN), and is the inverted specificity:

$$1 - Specificity = \frac{FP}{FP + TN}$$

The Area Under the Curve (AUC) of the ROC curve is a summary metric for evaluating model discrimination. The AUC quantifies the probability that the risk scores from a randomly selected pair of samples with and without the binary test condition are correctly ordered. Confidence intervals surrounding the AUC are calculated via bootstrap resampling of results with 1,000 resamples, as specified in the *pROC* package.[153] Similarly, differences between ROC curves and AUC can be significance tested using the same bootstrapping process.

AUC and ROC plots are arguably the most widely used metrics for reporting model accuracy in a diagnostic or prognostic setting. ROC plots have several key disadvantages, namely the obfuscation of risk thresholds and the equal weighting given to the costs of false positives and false negatives, which are often radically different in a clinical setting[154]. There is an argument that a simple binary threshold robustly estimated and validated has more use in specific clinical settings, with similarly binary outcomes. However, where risk stratification is concerned, risk prediction models require good calibration across all thresholds to be considered clinically useful and robust to new patient populations[21].

Decision Curve Analysis

In decision curve analysis (DCA)[154], a clinical judgement of the relative benefits (treating a true-positive) and harms (treating a false-positive) associated with prediction models is made across a range of threshold probabilities. Net benefit is computed by subtraction of the proportion of all patients who are false-positives from the proportion who are true positives at a certain threshold, weighted by the relative harm of a false positive and a false negative result. Net benefit depends on the cost and benefit of intervention, the prevalence P of the outcome of interest in the population and the model's ability to accurately assign risk to the correct outcomes. A model's classification accuracy is measured by the true-positive rate (TPR_R), the proportion of cases with risk above risk threshold R; the false-positive rate (FPR_R) is the proportion of controls with risk above risk threshold R. The net benefit NB to the population of using the risk model at the specified risk threshold R is:

$$NB_R = TPR_R\ P - \frac{R}{1 - R}\ FPR_R\ (1 - P)$$

There are a few important observations about this expression, as it requires that the risk threshold R has been chosen rationally, implicitly including the costs and benefits

of intervention. In DCA plots, R is commonly plotted along a range, for the reader to interpret at their own acceptable thresholds. A challenge in interpreting decision curves stems from the challenge in interpreting NB itself. A specific difficulty is that the NB is in units of "Benefit". Mathematically, the maximum possible value of NB is achieved when the $TPR = 1$ and $FPR = 0$; meaning we can never do better than intervening on all cases and no controls and the maximum possible $NB = P$. Instead standardised NB sNB can be used as a metric slightly easier to interpret as $sNB = NB/P$[155]. One reason is that sNB always has a maximum value of 1.0, providing a sense of large and small on a percent scale.

Net benefit can also be readily used to calculate the reduction in unneccessary interventions, typically where the routine intervention is to treat all (such as in prostate cancer where a suspicious clinical examination typically results in a biopsy). Reduction in this case is calculated as:

$$Reduction = \frac{(NB_{Model} - NB_{All}) \times 100}{R/(100 - R)}$$

Vickers, Calster & Steyerberg provide a very intuitive guide to interpreting DCA and readers are directed there for further information[156].

Estimation plots

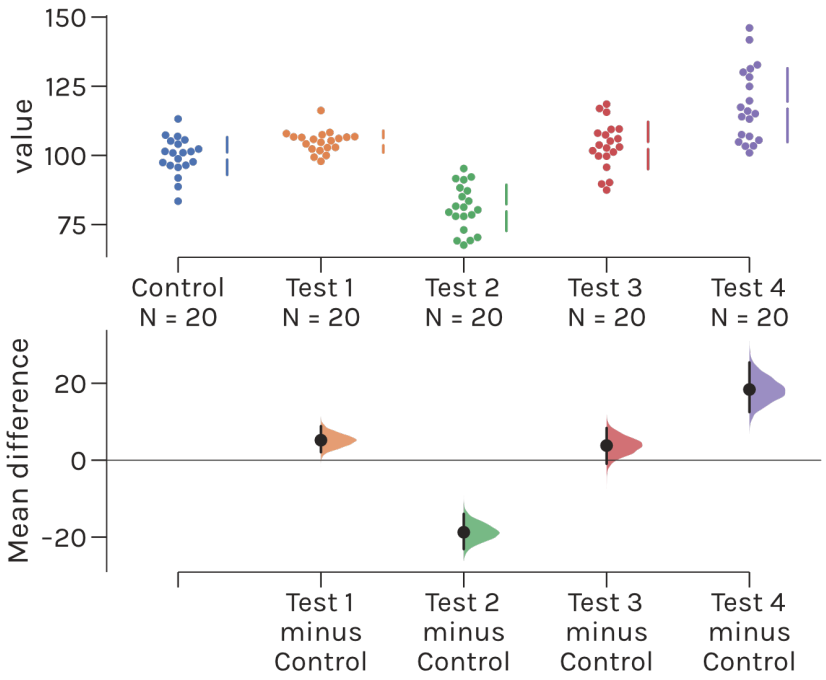


Figure 3.10: A generalised example of estimation plots. Individual raw data for each group of interest are shown as points, with confidence intervals shown as gapped bars in the upper panel. The lower panels shows bootstrap estimated effect sizes relative to the control group. Adapted from github.com/ACCLAB/DABEST-python

Estimation graphics are a means to avoid reliance on null-hypothesis significance testing (NHST) and the assumptions connected to statistical distributions. Proposed by Ho *et al.* in 2019, estimation statistics is a simple framework based upon the effect size of differences through bootstrap resampling of the available data, and displaying the results as familiar concepts of means, mean differences and error bars[157]. Estimation plots have two key features that define them: all data points are shown in a swarm plot, which orders each point to show the distribution, and the effect size of differences are presented as the bootstrapped 95% confidence intervals on separate, but aligned axes to the raw data (Figure 3.10).

Estimation plots and calculations were produced using the *dabestr* package[157] and 1,000 bootstrap resamples were used to visualise a robust effect size estimate of model predictions between risk groups where used.

Chapter 4

Development of the Prostate Urine Risk Scores

4.1 Summary

This chapter describes the development and internal validation of the Prostate Urine Risk (PUR) model, produced by Dr. Daniel Brewer using NanoString data from the Movember GAP1 cohort. Using a conventional training/test data split, PUR was trained on D'Amico status of patients using a LASSO-penalised ordinal regression. PUR showed good clinical utility for biopsy prediction, reporting good accuracy for Gs $\geq$ 3+4 (AUC = 0.76 (95% CI: 0.69 - 0.83)) and Gs $\geq$ 4+3 (AUCs = 0.72 (95% CI: 0.63 - 0.81)). This is a promising result that if externally validated, could result in substantial improvements to patient care.

In an active surveillance sub-cohort, PUR appeared to predict disease progression up to five years in advance, (HR = 8.23; 95% CI: 3.26 – 20.81). However, this AS performance was hampered by cohort effects, where D'Amico status alone returned similar predictive ability (HR = 6.51; 95% CI: 2.57 - 16.43). This is discussed in detail, with potential explanations and solutions provided.

This work is adapted from the original publication "A four-group urine risk classifier for predicting outcomes in patients with prostate cancer" by Connell *et al.* in BJU International, published 20th May 2019. Where work was completed by someone other than myself, this has been clearly stated.

4.2 Background

The progression of prostate cancer is highly heterogeneous[158], and risk assessment at the time of diagnosis is a critical step in the management of the disease. Based on the information obtained prior to treatment, key decisions are made about the likelihood of disease progression and the best course of treatment for localised disease. D'Amico stratification[63], which classifies patients as Low-, Intermediate-, or High-risk of PSA-failure post-radical therapy, is based on Gleason score (Gs)[159], PSA and clinical stage, and has been used as a framework for guidelines issued in the UK, Europe and USA[5,66,160]. Low-, and some favourable Intermediate-risk, patients are generally offered active surveillance[117,160] (AS) while unfavourable Intermediate-, and High-risk patients are considered for radical therapy[117]. Other classification systems, such as CAPRA score[88], use additional clinical information, assigning simple numeric values based on age, pre-treatment PSA, Gleason

score, percentage of biopsy cores positive for cancer and clinical stage for an overall 0-10 CAPRA score. The CAPRA score has shown favourable prediction of PSA-free survival, development of metastasis and prostate cancer-specific survival[161].

Prostate cancer is often multifocal[162], with disease state often underestimated by TRUS biopsy alone[163] and overestimated by multi-parametric-MRI (mpMRI), most often in the case of Prostate Imaging Reporting and Data System (PI-RADS) 3 lesions[58]. Sampling issues associated with needle biopsy of the prostate have prompted the development of non-invasive urine tests for aggressive disease, which examine prostate-derived material, harvested within urine[102,110,164]. Recent successes in this field are illustrated by three studies carried out on whole urine for predicting the presence of Gs ≥ 7 on initial biopsy: Tomlins *et al.* (2016), and McKiernan *et al.* (2016) used PCA3 and TMPRSS2-ERG transcript expression levels, whilst Van Neste *et al.* (2016) used HOXC6 and DLX1 in combination with traditional clinical markers[99,102,164].

The objectives of this study were to develop a classification model that could predict D'Amico & CAPRA risk group pre-biopsy from a single urine sample, and to additionally test the classifier's utility as a predictor of disease progression in a sub-cohort of AS patients with five years of clinical follow-up.

4.3 Materials & Methods

4.3.1 Patient samples and clinical criteria

Samples collected within the Movember GAP1 cohort that had been processed to harvest extracellular vesicles and interrogated by NanoString were used here, as described in Chapter 3.

D'Amico classification used Gleason and PSA criteria as per D'Amico *et al.*[63]. CAPRA classification used the criteria as described by Cooperberg *et al.*[88]. Where multiple biopsies were taken the results from the closest biopsy to initial urine sample collection were used. Men were defined to have no evidence of cancer (NEC) with a PSA normal for their age or lower[165] and as such, were not subjected to biopsy. Metastatic disease, defined by a PSA >100 ng/mL, were excluded from analyses.

4.3.2 Expression analyses

NanoString data were adjusted relative to internal positive control probes as per Chapter 3, with the following changes. The ComBat algorithm was used to adjust for inter-batch and inter-cohort bias[166]. Data were further adjusted by means of a correction factor (CF) for input amount by normalisation to two invariant and highly expressed housekeeping gene-probes, GAPDH and RPLP2. The CF for a given sample i, was calculated as the total mean of GAPDH and RPLP2 expression, divided by the sample-specific mean of GAPDH and RPLP2:

$$CF_i = \frac{\sum_j \bar{x}_{GAPDH,RPLP2}}{n(\bar{x}_{GAPDH_i,RPLP2_i})}$$

All data were expressed relative to KLK2 as follows: samples with low KLK2 (counts <100) were removed, and data log2 transformed. Data were further normalised by adjusting the median of each probe across all samples to 1, with the interquartile range adjusted to that of KLK2. More formally, for each sample i and gene-probe j, the KLK2 normalised

value, $\hat{y}_{i,j}$ was calculated as:

$$\hat{y}_{i,j} = \frac{\frac{y_{i,j} - median_j}{IQR_j} \times IQR_{KLK2}}{y_{i,j}}$$

No correlation was seen with respect to patient's drugs, cohort site, urine pH, colour or sample volume (all $P > 0.05$; Chi-square and Spearman's Rank tests, data not shown). The work in this section was completed by Helen Curley.

4.3.3 Model production and statistical analysis

All statistical analyses and model construction presented here were undertaken in R version 3.4.1[167], and unless otherwise stated utilised base R and default parameters. The Prostate Urine Risk (PUR) signatures were constructed from the training dataset as follows: for each probe, a univariate cumulative link model was fitted using the R package *clm* with risk group as the outcome and NanoString expression as inputs. Each probe that had a significant association with risk group ($P < 0.05$) was used as input to the final multivariate model. A constrained continuation ratio model with an L_1 penalisation was fitted to the training dataset using the *glmnetcr* library[168], an adaptation of the least absolute shrinkage and selection operator (LASSO) method[169]. Default parameters were applied using the LASSO penalty and values from all probes selected by the univariate analysis used as input. The final multivariable model was selected according to the minimum Akaike information criterion and incorporated all probes not removed by the LASSO penalty. Model construction was performed by Daniel Brewer. Ordinal logistic regression was undertaken using the *ordinal* library[170].

Bootstrap resampling of ROC analyses used the *pROC* library[153] for calculation, statistical tests and production of figures, with 2,000 resamples used. Random predictors were generated by random sampling from a uniform distribution between 0 and 1.

Survival analyses were undertaken where follow-up of AS patients allowed and used progression as an endpoint, where progression criteria were either: PSA velocity >1 ng/ml per year or Gs $\geq$ 4+3 or $\geq$ 50% cores positive for cancer on repeat biopsy. Cox proportional hazards models utilised risk signatures as a continuous variable. Kaplan-Meier (KM) estimators were calculated based on the median optimal threshold to minimise the Log-rank test P-value from 10,000 resamples of the cohort with replacement to ensure robustness. As the clinical costs of missing significant cancer are far higher than an unnecessary biopsy or investigation, where multiple samples were analysed from the same AS patient the sample with the highest PUR-4 signature was used in survival analyses and KM estimators. Where multiple samples were available from a patient, only a single sample was used.

Decision curve analysis (DCA)[154] examined the potential net benefit of adopting PUR signatures in clinical settings and was reported as standardised net benefit as per Chapter 3 as it is more interpretable when compared to net benefit[155].

In order to ensure DCA was representative of a more general population, the prevalence of Gleason grades within the Movember cohort was adjusted via bootstrap resampling to match that observed in a population of 219,439 men that were in the control arm of the Cluster Randomised Trial of PSA Testing for Prostate Cancer (CAP) Trial[6]. Briefly, for the biopsied men within this CAP cohort, 23.6% were Gs 6, 8.7% Gs 7 and 7.1% Gs 8 or greater, with 60.6% of biopsies being prostate cancer negative. These proportions were used to perform stratified random sampling with replacement of the Movember cohort to

produce a synthetic dataset of 300 samples. Standardised net benefit was calculated on the resampled dataset, and the process repeated for a total of 1,000 resamples. The mean standardised net benefit for PUR-4 and the "treat-all" options over all iterations was used to produce the presented figures to account for variance in sampling.

Stability of temporally-spaced samples from the same patient were assessed by simulation against a null model. This null model was generated by random sampling of two non-related samples from the whole Movember GAP1 cohort and measuring the Euclidean distance between samples using their PUR signatures. This was repeated to produce a simulated population the same size as the real paired samples. The mean distance was calculated and the resampling with replacement process was repeated 100,000 times and the real distances from paired samples compared to this synthetic distribution.

4.4 Results

4.4.1 The Clinical Cohort

The Movember cohort comprised of 535 post-DRE urine samples collected from four centres (NNUH, $n = 312$; RMH, $n = 87$; Atlanta, $n = 85$; Dublin, $n = 17$). Multiple, longitudinal samples within the Movember cohort were provided by 20 of the 87 men enrolled on an AS program at the RMH (Figure 4.11. The median time between collection of multiple samples was 185 days (IQR: 122-252 days) and were treated independently from one another. Samples originated from men categorised as having either No Evidence of Cancer (NEC, $n = 92$) or localised prostate cancer at time of urine collection, as detected by TRUS biopsy ($n = 443$). Patients with cancer were further subdivided into three risk categories using D'Amico criteria: Low (L), $n = 134$; Intermediate (I), $n = 208$; and High-risk (H), $n = 101$. Patients with metastatic cancer at collection were excluded from analyses ($n = 35$). Further characteristics of the Movember cohort are available in Table 4.1.

Table 4.1: Characteristics of the Training and Test datasets

Characteristic	Training	Test
Total, n (%)	358 (67.0)	177 (33.0)
Collection Centre:		
NNUH	203	109
RMH	83	38
Dublin	9	8
Atlanta	63	22
PSA, ng/ml, mean (median; IQR)	10.6 (6.9, 6.4)	10.9 (6.9, 7)
Age, yr, mean (median; IQR)	65.8 (67, 11)	67.2 (67, 11)
Family history, %; no, yes, NA	3.0, 6.1, 90.8	0.6, 6.2, 93.3
First biopsy, n (%)	298 (82.78)	145 (81.46)
Prostate volume, ml; mean (median; IQR)	59.2 (49.8, 30.4)	61.1 (49.2, 32.8)
PSAD, ng/ml; mean (median; IQR)	0.29 (0.19, 0.16)	0.29 (0.18, 0.17)
Suspicious DRE, n	107	52
Diagnosis:		
NEC, n (%)	62 (17.3)	30 (17.0)
D'Amico Low n (%)	89 (24.9)	45 (25.4)
D'Amico Intermediate n (%)	139 (38.8)	69 (39.0)
D'Amico High n (%)	61 (17.0)	27 (15.3)
Metastatic (bone scan) n (%)*	7 (2.0)	6 (3.3)
CAPRA:		
Low (0-2) n (%)	97 (33.7)	49 (33.7)
Intermediate (3-5) n (%)	108 (37.5)	53 (36.6)
High (< 7) n (%)	83 (28.8)	43 (29.7)
Gleason:		
Gleason, n:	292	144
Gs = 6, n (%)	119 (40.8)	64 (44.4)
Gs = 7, n (%)	131 (44.9)	56 (38.9)
Gs > 7 n (%)	42 (14.4)	24 (16.7)

* Data from patients with metastatic disease confirmed by bone scan after sample collection were used, and classified as D'Amico High-risk.

4.4.2 Selection of cell-free fractions

Based on earlier analyses and previously published results by Pellegrini *et al.*[171], the cell-free and extracellular vesicle fraction in urine samples were selected for this study.

4.4.3 Development of the Prostate Urine Risk Signatures

Samples in D'Amico categories Low, Intermediate and High-risk, together with NEC samples were divided into the Movember Training dataset (two-thirds of samples; $n = 358$) and the Movember Test dataset (one-third of samples; $n = 177$) by random assignment, stratified by risk category (Table 4.1).

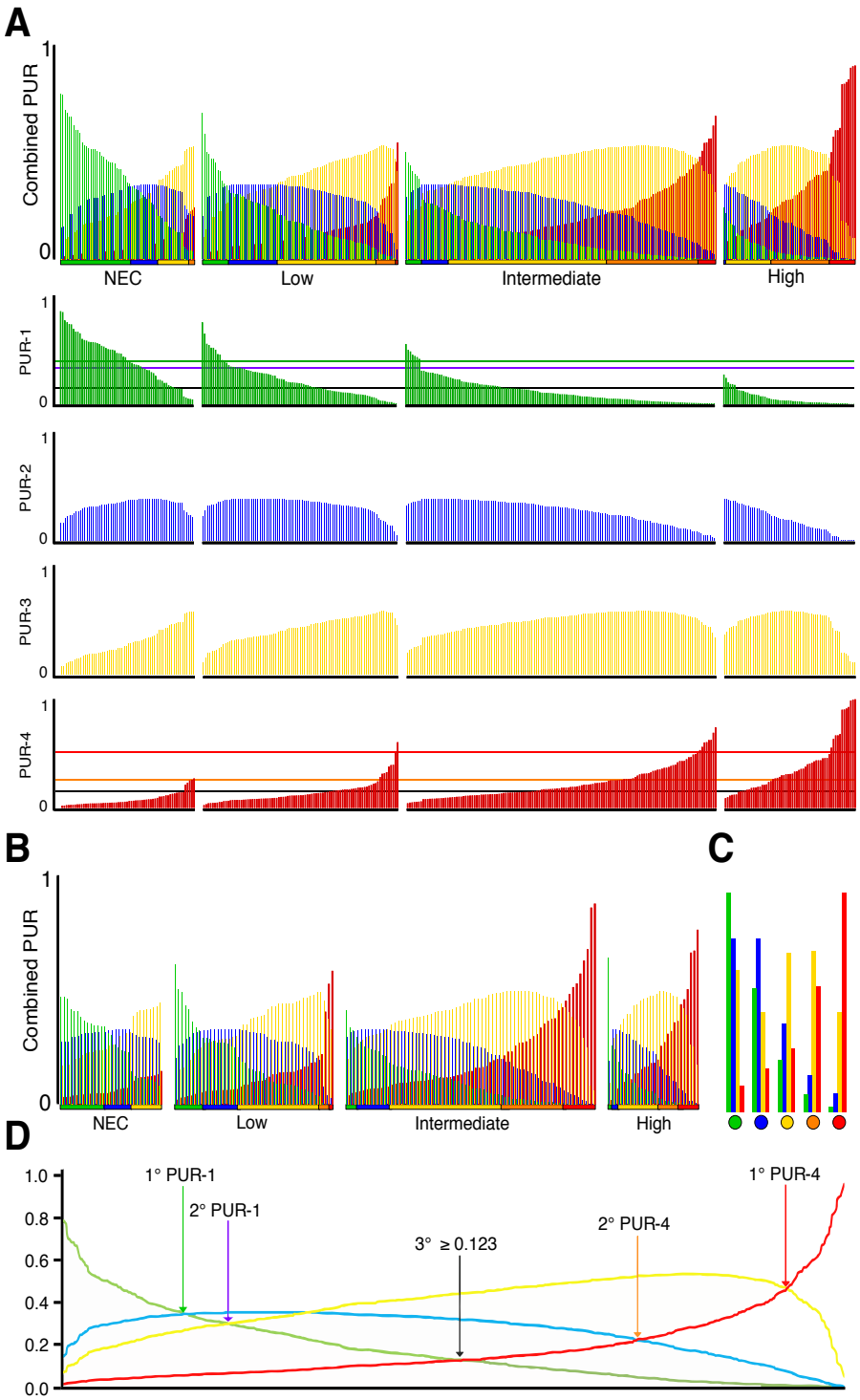


Figure 4.1: A) PUR profiles (PUR-1 – green, PUR-2 – blue, PUR-3 – yellow, PUR-4 – red) for the Training dataset, grouped by D'Amico risk group and ordered by ascending PUR-4 score. Horizontal lines indicate where the PUR thresholds lie as shown in D). B) PUR profiles in the Test dataset. C) Examples of samples with primary PUR signatures, where coloured circles indicate the primary PUR signal for that sample; 1° PUR-1 (green), 1° PUR-2 (blue), 1° PUR-3 (yellow), 2° PUR-4 (orange) and 1° PUR-4 (red). D) The outline of the four PUR signatures for all samples ordered in ascending PUR-4 (red) to illustrate where 1°, 2° and the 3° crossover point of PUR-1 and PUR-4 lie.

Table 4.2: Coefficients of the 36 gene probes included as variables in the final PUR model and the intercepts.

PUR variable:	Coefficient
Intercept	-2.1781570
AMACR	0.6829973
AMH	0.3363184
ANKRD34B	0.1673693
APOC1	0.3712274
AR (exons 4-8)	-0.4771042
DPP4	-1.3364905
ERG (exons 4-5)	0.0256132
GABARAPL2	0.5138853
GAPDH	-0.9188083
GDF15	0.2792761
HOXC6	0.6543025
HPN	-0.4625853
IGFBP3	-1.2101205
IMPDH2	0.4543117
ITGBL1	-0.1094984
KLK4	-1.5051707
MARCH5	-1.4391403
MED4	-1.0766399
MEMO1	-1.9473755
MEX3A	0.2318072
MME	-0.9433935
MMP11	0.9918169
MMP26	0.3549589
NKAIN1	0.0352952
PALM3	0.1954966
PCA3	2.7549211
PPFIA2	-0.7369071
SIM2 (short)	0.9031434
SMIM1	-0.2209302
SSPO	0.9231364
SULT1A1	1.7614731
TDRD1	0.2666629
TMPRSS2/ERG fusion	0.4792269
TRPM4	0.0594701
TWIST1	-0.2593533
UPK2	0.6382611
Cp 1	2.4258354
Cp 2	1.4855935
Cp 3	-0.4792212

NanoString data for 167 gene-probes were obtained for each sample. The data was processed and normalised by Helen Curley (see Methods above). The optimal model, produced by Daniel Brewer, for prediction of D'Amico status (NEC, Low-, Intermediate-, High-Risk) as defined by the LASSO criteria in a constrained continuation ratio model, (see Methods for full details) incorporated information from 36 probes (Table 4.2) and was applied to both training and test datasets (Figure 4.1A, B). For each sample the 4-signature PUR-model was interpreted as the probability of containing NEC (PUR-1), L (PUR-2), I (PUR-3) and H (PUR-4) material within samples (Figure 4.1A, B). The sum of all four PUR-signatures in any individual sample was 1 (PUR1 + PUR2 + PUR3 + PUR4 = 1). The strongest PUR-signature for a sample was termed the primary (1°) signature while the second highest was called the secondary (2°) signature (Figure 4.1C, D).

4.4.4 Pre-biopsy Prediction of D'Amico risk, CAPRA score and Gleason:

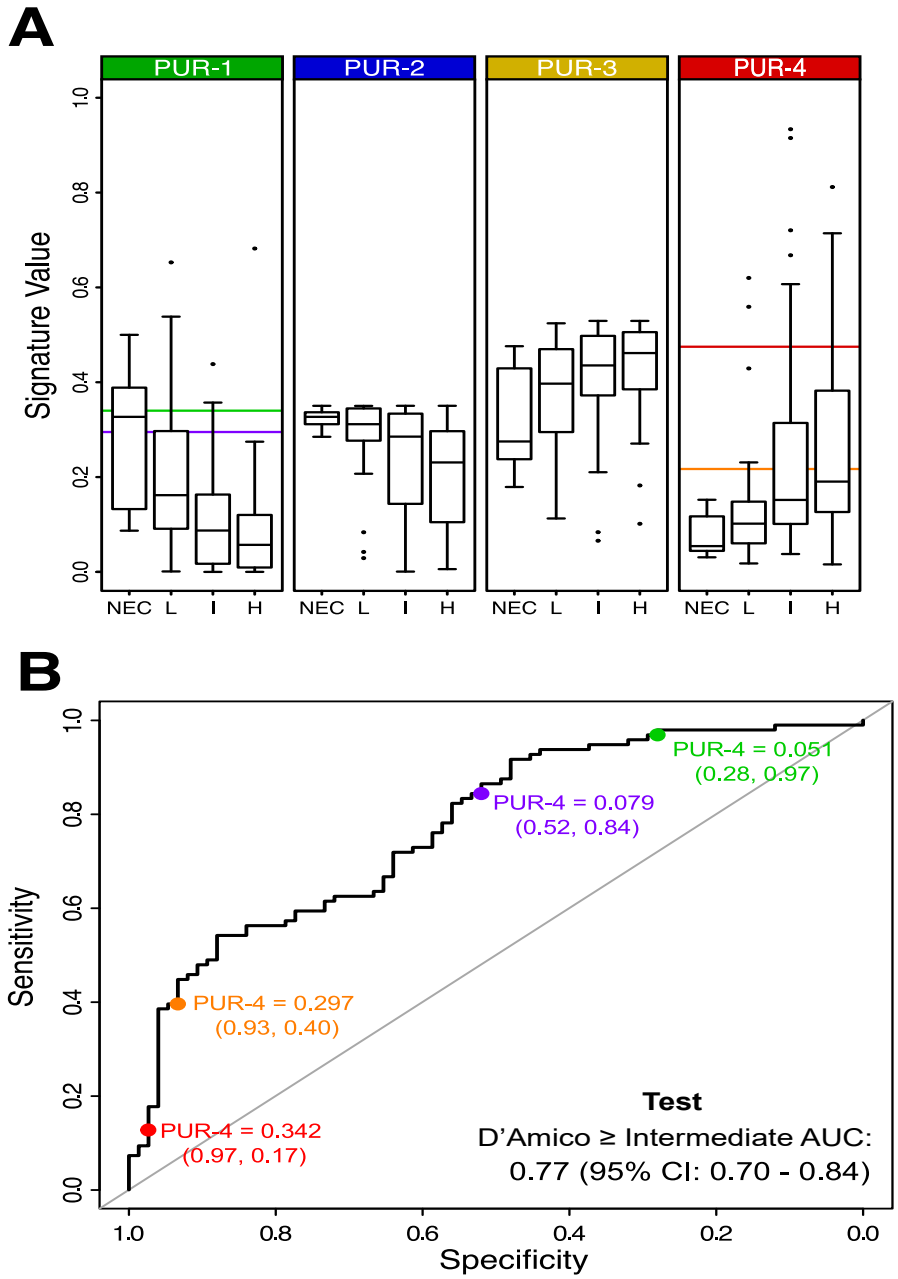


Figure 4.2: A) Boxplots of PUR signatures in samples categorised as no evidence of cancer (NEC, $n = 30$) and D'Amico risk categories; (L – Low, $n = 45$, I – Intermediate,$n = 69$ and H – High risk, $n = 27$) in the Test dataset. Horizontal lines indicate where the PUR thresholds lie for: 1° PUR-1 (Green), 2° PUR-1 (Purple), 1° PUR-4 (Red), 2° PUR-4 (Orange). B) ROC curve of PUR-4 predicting the presence of significant (D'Amico Intermediate or High risk) prostate cancer prior to initial biopsy in the Test dataset. Coloured circles indicate the specificity and sensitivity

Primary PUR-signatures (PUR-1 to 4) were found to significantly associate with each clinical category (NEC, L, I, H respectively) in both training and test sets ($P < 0.001$, Wald test for ordinal logistic regression in both Training and Test datasets, Figure 4.2A, B, Table 4.3).

Table 4.3: Assignment matrix of samples based on their primary PUR signature and actual D'Amico Risk category in the Training and Test datasets

PUR Assignment	NEC	Low Risk	Intermediate Risk	High Risk
Training				
1° PUR-1	**63**	24	13	0
1° PUR-2	26	**47**	26	2
1° PUR-3	7	23	**47**	22
1° PUR-4	0	4	38	**58**
Test:				
1° PUR-1	**48**	28	10	4
1° PUR-2	24	**38**	31	7
1° PUR-3	9	22	45	23
1° PUR-4	0	12	**53**	**35**

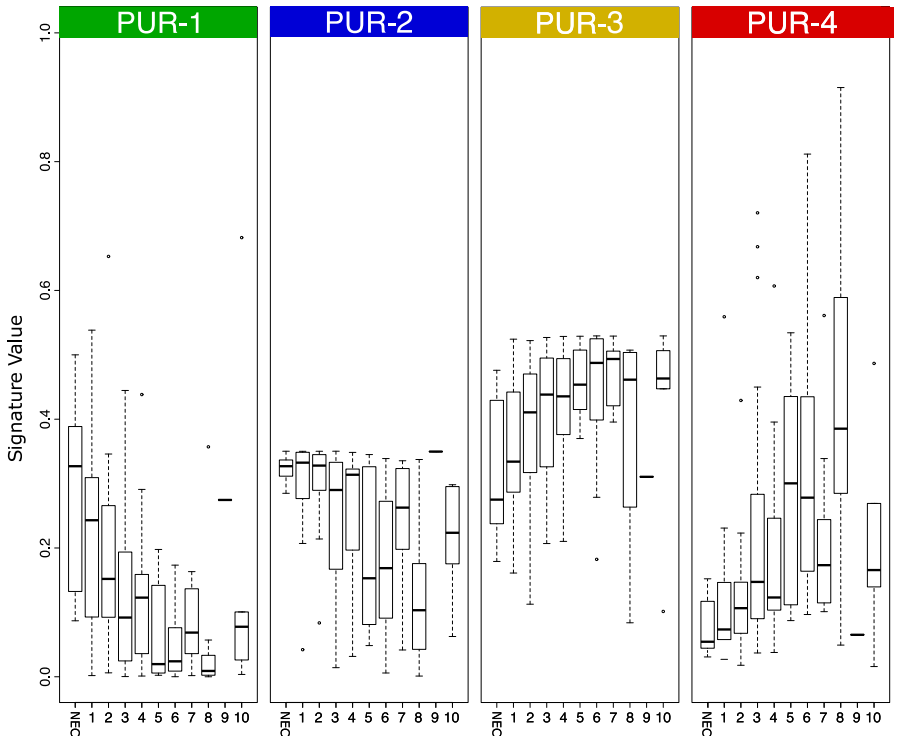


Figure 4.3: Boxplots of PUR signatures relative to no evidence of cancer (NEC) and CAPRA scores 1 – 10 in the Test dataset. Numbers of samples within each group are as detailed in the table above.

A similar association was observed with CAPRA score ($P < 0.001$, Wald test for ordinal logistic regression; Figure 4.3.

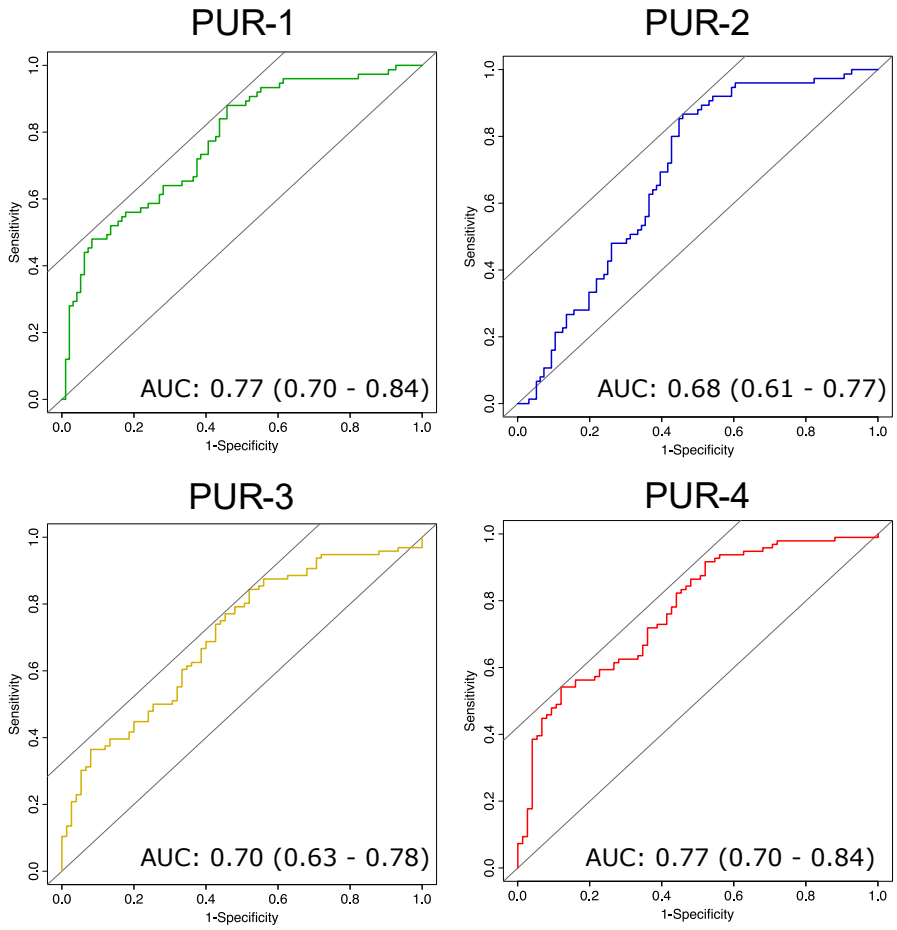


Figure 4.4: ROC curves for each of the four PUR signatures (Green – PUR-1, Blue – PUR-2, Yellow – PUR-3, Red – PUR-4) predicting presence of D'Amico Intermediate- or High-risk cancers on initial biopsy in the test dataset.

Based on recommended guidelines[54,66,160], the distinction between D'Amico low and intermediate-risk is considered critical because radical therapy is commonly recommended for patients with high and intermediate-risk cancer. We therefore initially tested the ability of the PUR-model to discriminate the presence of H or I disease from L or NEC upon initial biopsy. Each of the four PUR-signatures alone were able to predict the presence of significant disease (Risk category $\geq$ Intermediate, Area Under the Curve (AUC) ≥ 0.68 for each PUR signature, Test dataset; Figure 4.4, and were significantly better than a random predictor ($P < 0.001$, bootstrap test, 2,000 resamples). However, PUR-1 and PUR-4 were equally best at discerning significant disease; AUCs for both PUR-4 and for PUR-1 in the Test dataset were 0.77 (95% CI: 0.70 - 0.84), (Figure 4.2B).

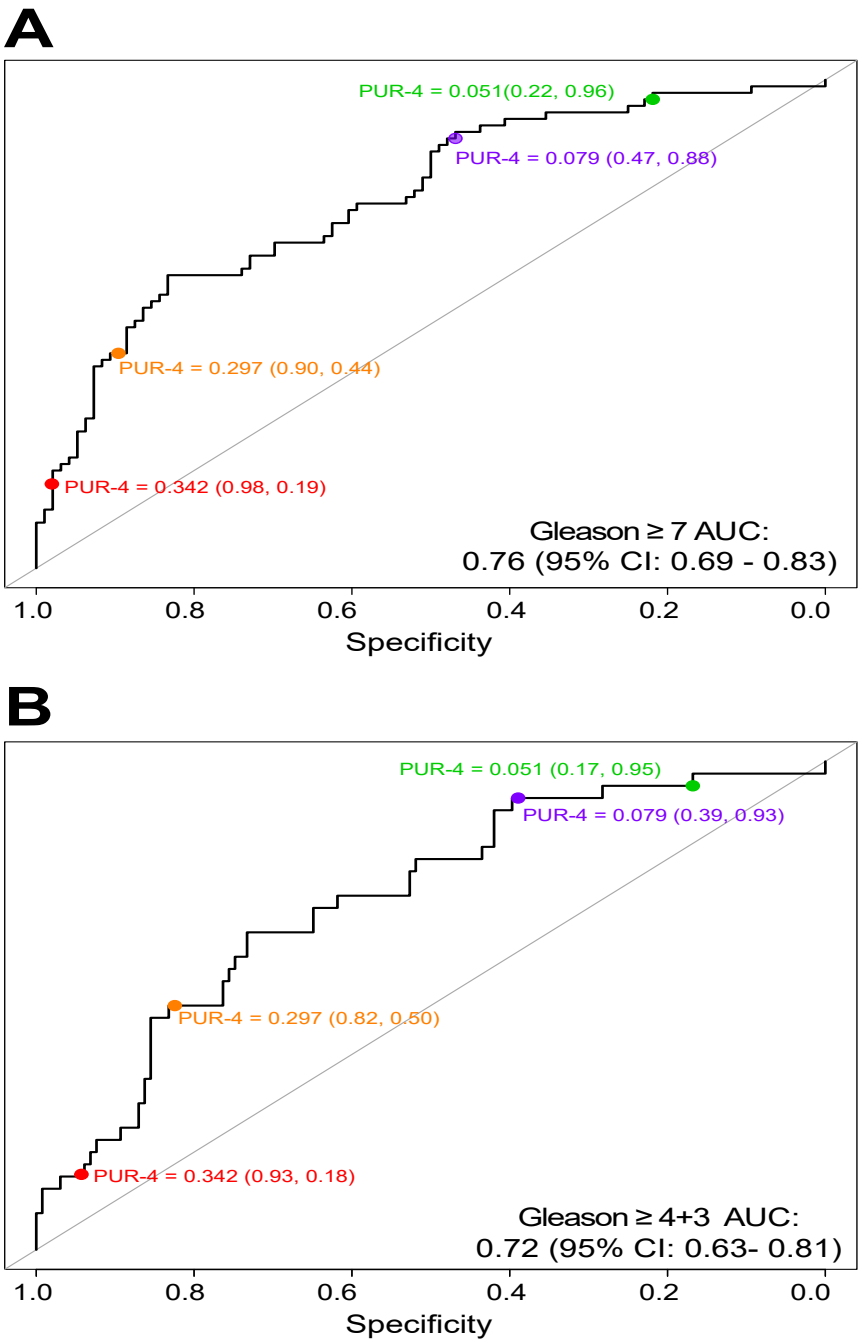


Figure 4.5: ROC plots for PUR-4 predicting the presence/absence of: A) Gleason $\geq$ 7 on initial biopsy in the Test dataset or B) Gleason $\geq$ 4+3 in the Test dataset. Coloured circles indicate the specificity and sensitivity, respectively, of thresholds along the ROC curve that correspond to the indicated PUR-4 thresholds

When Gleason score alone was considered we found that PUR-4 predicted Gs $\geq$ 3+4 with AUCs of 0.78 (95% CI: 0.73 - 0.82) (Training) and 0.76 (95% CI: 0.69 - 0.83) (Test) and Gs $\geq$ 4+3 with AUCs of 0.76 (95% CI: 0.70 - 0.81) (Training) and 0.72 (95% CI: 0.63 -

0.81) (Test) (Figure 4.5). The ability to predict Gs $\geq$ 3+4 was particularly relevant because this was previously chosen as an endpoint for aggressive disease in other urine biomarker studies, where AUCs of 0.77, 0.78 and 0.74 were reported by McKiernan *et al.*[102], Tomlins *et al.*[164] and Van Neste *et al.*[99], respectively.

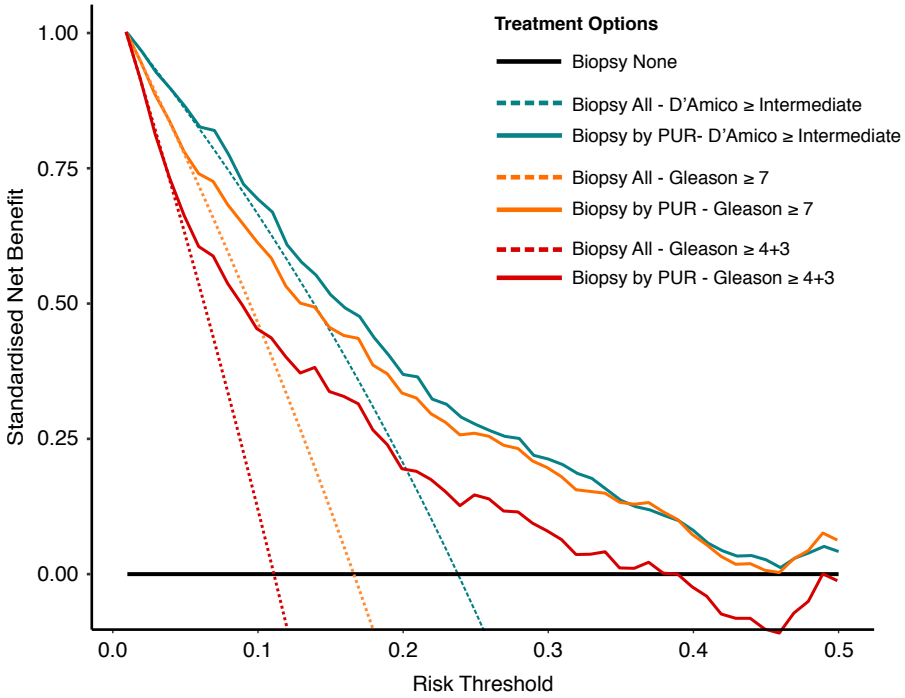


Figure 4.6: DCA plot depicting the standardised net benefit of adopting PUR-4 as a continuous predictor for detecting significant cancer on initial biopsy, when significant is defined as: D'Amico risk group of Intermediate or greater (teal), Gs $\geq$ 3+4 (orange) or Gs $\geq$ 4+3 (red). To assess benefit in the context of cancer arising in a non-PSA screened population of men we used data from the control arm of the CAP study(30). Bootstrap analysis with 100,000 resamples was used to adjust the distribution of Gleason grades in the Movember cohort to match that of the CAP population. For full details see Methods.

Decision curve analysis (DCA, section 3.2)[154] examined the potential net benefit of using PUR-signatures in a non-PSA screened population. Biopsy of men based upon their PUR-4 score provided a net benefit over biopsy of men based on current clinical practice across all thresholds (Figure 4.6).

4.4.5 Active surveillance cohort:

Active Surveillance cohort characteristics

Table 4.4: Active surveillance cohort characteristics.

Characteristic:	
Patients, n	87
Age, year, mean (median; IQR)	64 (66, 7)
PSA, ng/ml, mean (median; IQR)	7.8 (7.5, 3.3)
D'Amico:	
Low n (%)	55 (63)
Intermediate n (%)	32 (37)
CAPRA:	
Low (0-2) n (%)	59 (68)
Intermediate (3-5) n (%)	27 (31)
High (>5) n (%)	1 (1)
Gleason Score:	
Gs < 7, n	79
Gs = 3+4, n	7
Gs = 4+3, n	1
Number of biopsies:	
1	14
2	28
>2	35
NA:	10
Number of negative biopsies following a positive:	
1	26
2	3
NA:	58
Progressed to treatment due to:	
PSA increase	17
Adverse histopathology	6
mpMRI criteria only	9
Non-progressed to treatment due to:	
Any criteria	49
Self-elected for treatment:	3
Died of other causes:	3

Gs = Gleason score; IQR = interquartile range; PSA = prostate-specific antigen; NA = not available; mpMRI = multiparametric magnetic resonance imaging

Within the Movember cohort were 87 men enrolled in AS at the Royal Marsden Hospital, UK. The median follow-up time from initial urine sample collection was 5.7 years (range 5.1 – 7.0 years) (Table 2). The median time from initial urine sample collection to progression or final follow up was 503 days (range 0.1 – 7.4 years). Full AS cohort characteristics are available in Table 4.4.

PUR model performance in AS cohort

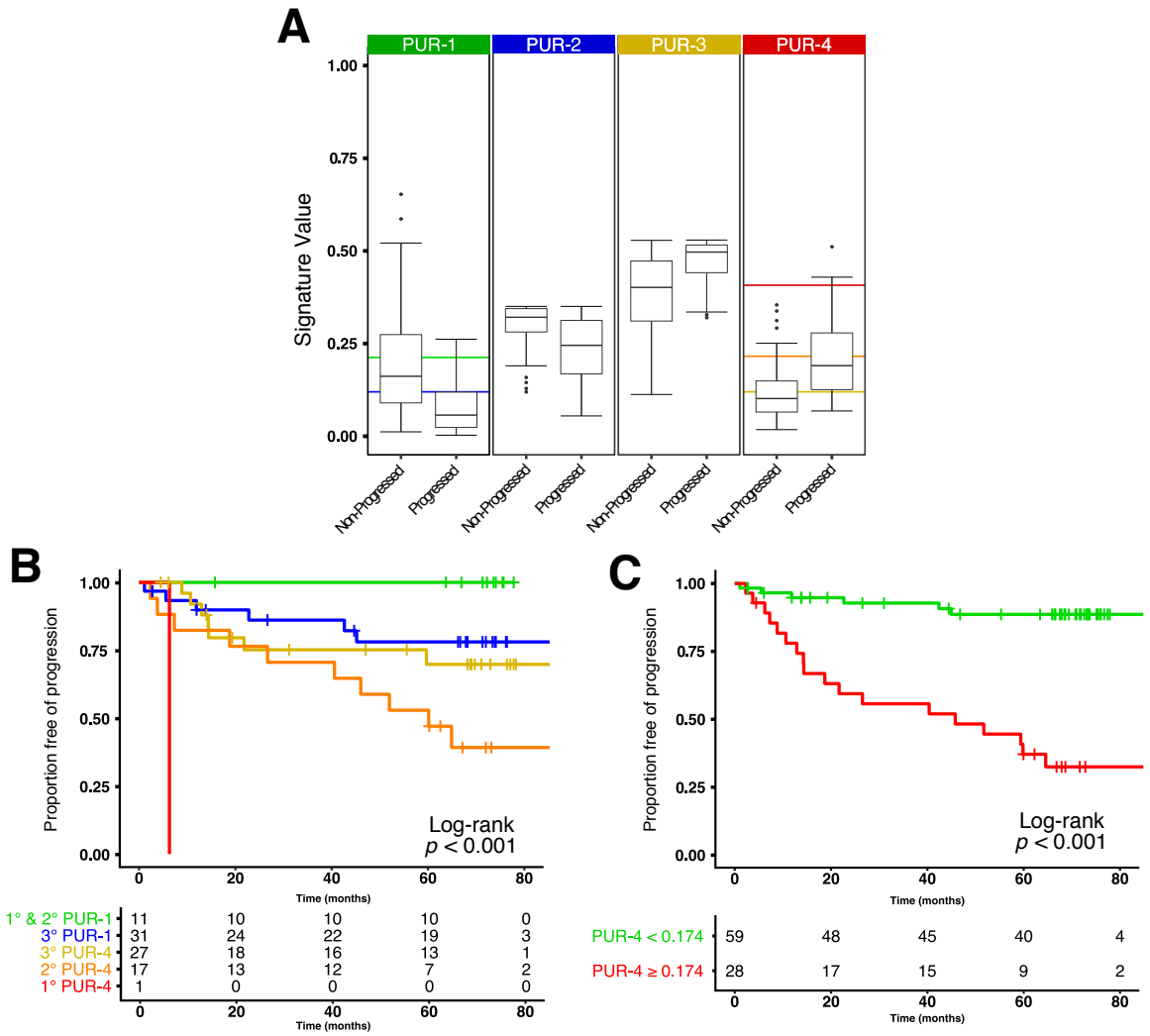


Figure 4.7: A) PUR profiles of patients on active surveillance that had met the clinical criteria, not including mpMRI criteria, for progression ($n = 23$) or not ($n = 49$) at five years post urine sample collection. Progression criteria were either: PSA velocity > 1 ng/ml per year or Gs $\geq$ 4+3 or $\geq$ 50% cores positive for cancer on repeat biopsy. PUR signatures for progressed vs non-progressed samples were significantly different for all PUR signature ($P < 0.001$, Wilcoxon rank sum test). Horizontal line colour indicates the thresholds for PUR categories described in: B) Kaplan-Meier plot of progression in active surveillance patients with respect to PUR categories described by the corresponding colours; Green - 1° and 2° PUR-1, Blue - 3° PUR-1, Yellow - 3° PUR-4, Orange - 2° PUR-4, Red - 1° PUR-4 and the number of patients within each PUR category at the given time intervals in months from urine collection. C) Kaplan-Meier plot of progression with respect to the dichotomised PUR thresholds described by the corresponding colours Green – PUR-4 < 0.174, Red – PUR-4 ≥ 0.174 and the number of patients within each group at the given time intervals in months from urine collection.

Calculation of Kaplan-Meier estimators with samples divided on the basis of 1°, 2° and 3° PUR-1 and PUR-4 signatures showed significant differences in clinical outcome ($P < 0.001$, log-rank test, Figure 4.7B) and was robust (log-rank test $P < 0.05$ in 93.6% of 100,000 resamples with replacement, see section 3.2 for full details). Proportion of PUR-4, a continuous variable, had a significant association with clinical outcome ($P < 0.001$; IQR HR = 5.87, 95% CI: 1.68 – 20.46); Cox Proportional hazards model).

A robust optimal threshold of PUR-4 was generated through bootstrap resampling of the AS cohort with replacement. At each resample, the PUR-4 threshold that minimised the *p*-value reported from the Log-rank test was recorded, therefore maxmising the discriminatory ability of a dichotomised PUR-4 for predicting survival. This was repeated over 10,000 resamples with replacement to ensure robustness and avoid overfitting to specific samples. The median PUR-4 threshold over all resamples was selected (PUR-4 = 0.174) to dichotomise patients into poor prognosis and good prognosis groups. The two groups were found to have a large difference in time to progression: 60% progression within 5 years of urine sample collection in the poor prognosis group compared to 10% in the good prognosis group ($P < 0.001$, log-rank test, 4.7C, HR = 8.23; 95% CI: 3.26 – 20.81). This result is robust ($P < 0.05$ in 99.8% of 100,000 resamples with replacement, see section 4.3.3 for full details).

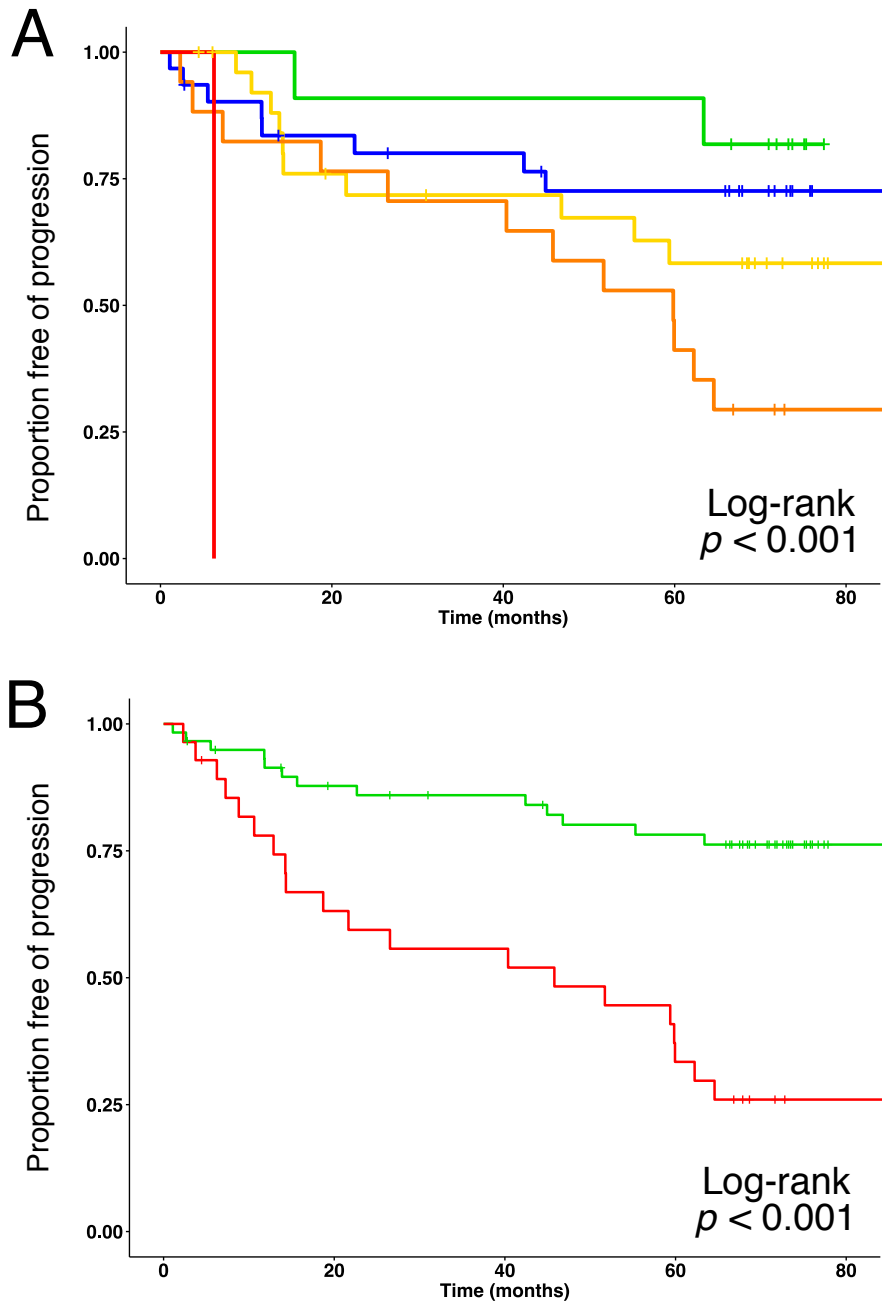


Figure 4.8: A) Kaplan-Meier plot of AS progression, including mpMRI criteria over time in days with respect to PUR thresholds described by the corresponding colours Green - 1° and 2° PUR-1, Blue - 3° PUR-1, Yellow - 3° PUR-4, Orange - 2° PUR-4, Red - 1° PUR-4. B) Kaplan-Meier plot of progression, including mpMRI criteria, with respect to the dichotomised PUR thresholds described by the corresponding colours Green – PUR-4 $<$ 0.174, Red – PUR-4 $=$ 0.174 and the number of patients within each group at the given time intervals in months from urine collection.

When mpMRI criteria for progression were also included, both primary PUR-status and

dichotomised PUR threshold remained a significant predictor of progression ($P < 0.001$ log–rank test, Figure 4.8).

Potential confounding in survival outcomes affects the intepretation of model performance:

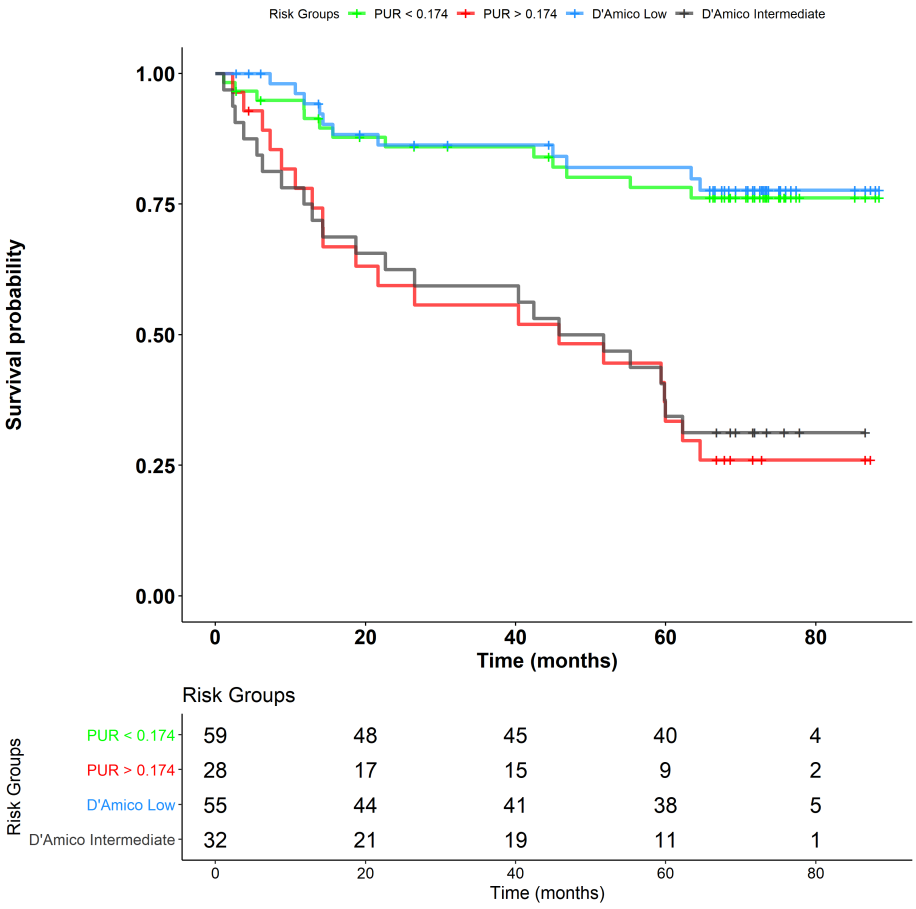


Figure 4.9: Kaplan-Meier plot and risk tables of AS progression with either D'Amico category alone (Dashed darker lines), or dichotomised PUR (Solid brighter lines) defining the risk groups. The table underneath the main figure details the number of patients still at risk of progression within each group at a given time on the x-axis.

When D'Amico Risk category was considered as the sole predictor variable for progression into a Cox proportional hazards regression model, it was found to be a significant predictor of progression in AS, returning similar hazard ratios and numbers at risk when compared to the dichotomised PUR thresholds ($P < 0.001$ log-rank test, HR = 6.51; 95% CI: 2.57 - 16.43, Figure 4.9).

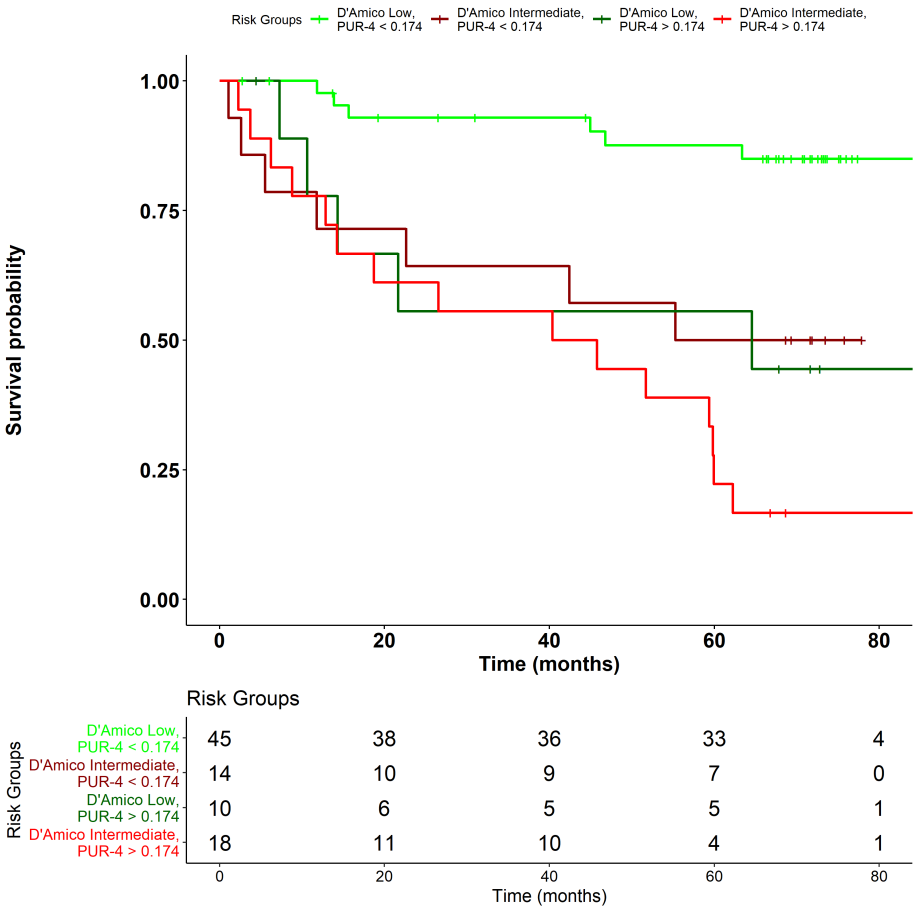


Figure 4.10: Kaplan-Meier plot and risk tables of AS progression considering both DÁmico category and PUR-4 status to define the risk groups. The table underneath the main figure details the number of patients still at risk of progression within each group at a given time on the x-axis.

When both D'Amico Risk category was considered alongside the dichotomised PUR-4 threshold the lowest risk group (D'Amico Low-risk and PUR-4 < 0.174) were found to have a very low rate of progression whilst the highest risk group (D'Amico Intermediate and PUR-4 >0.174) had significantly worse rates of progression ($p < 0.001$, Log-rank test, Figure 4.10). However, this introduced larger uncertainty for the two groups in between where PUR-4 status and D'Amico disagreed, with no significant differences in progression from either the highest or lowest risk groups.

4.4.6 Longitudinal stability of the PUR model in urine samples

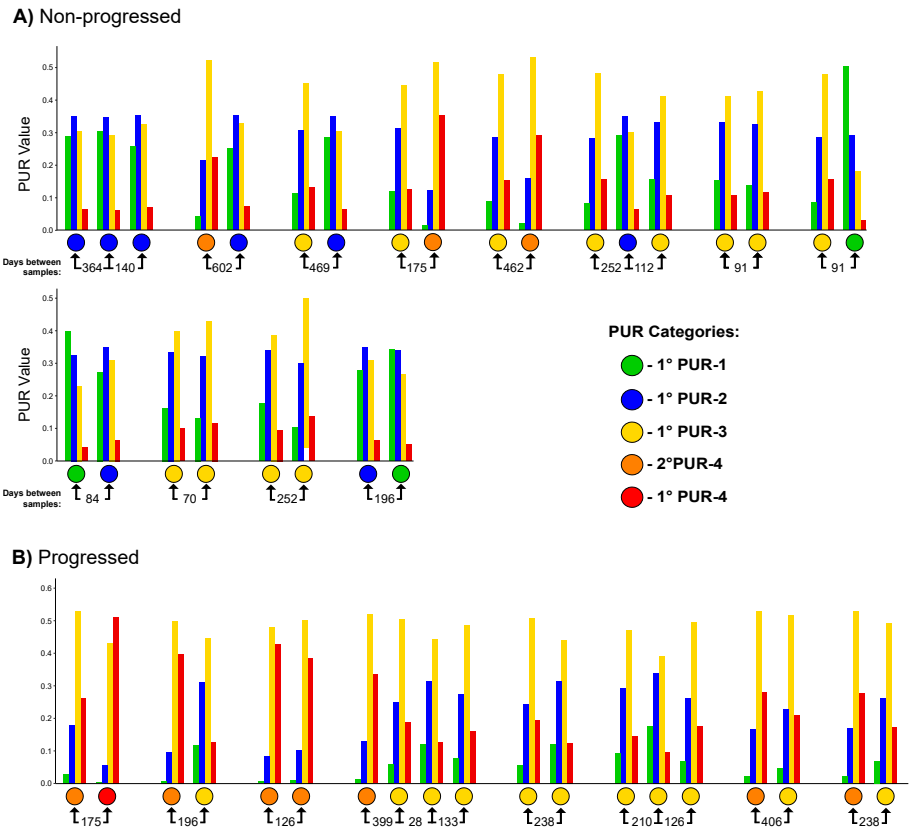


Figure 4.11: PUR signatures from Active Surveillance longitudinal samples: 1° PUR-1 (Green), 2° PUR-1 (Purple), 1° PUR-2 (Blue), 1° PUR-3 (Yellow), 2° PUR-4 (Orange), 1° PUR-4 (Red). Samples within each numbered box are from a single patient with coloured circles underneath indicating primary PUR signature. A) patients that did not reach clinical progression criteria. B) patients that reached clinical progression criteria. Arrows and numbers under coloured circles detail the number of days between consecutive samples from a patient.

Multiple urine samples, collected at varying intervals, were available for 20 of the patients in the AS cohort, allowing for an assessment of stability of the PUR profiles over time (Figure 4.11.

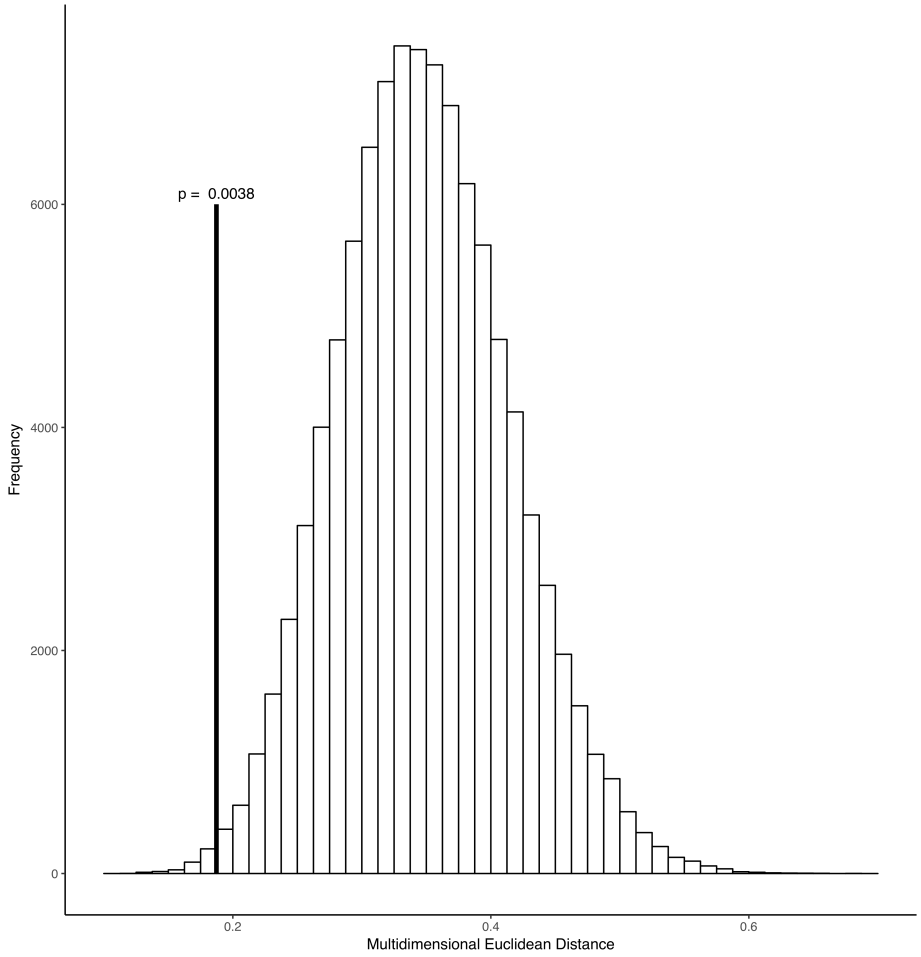


Figure 4.12: Distribution of the mean Euclidean distances recorded by comparing two randomly selected samples from the Movember GAP1 cohort with replacement to generate 20 pairs of random samples. This was repeated 100,000 times to generate the distribution shown. The vertical line details the mean Euclidean distance of the non-progressed samples in the AS cohort. The P value is calculated as the proportion of simulated results more stable than the real results.

Stability was assessed via simulation against a null model of purposefully non-stable urine samples. This null model was generated by random selection with replacement of two samples from the Movember cohort, and the Euclidean distance between them recorded. This was repeated 20 times to form a synthetic null model dataset the same size as the real set of paired samples, and the mean distance recorded over 100,000 iterations of this process. The mean distance of the samples from patients who did not progress ($n = 12$) was found to be significantly more stable than those generated from this simulated process ($P = 0.0038$, Figure 4.12), whilst the samples from patients deemed to have progressed did not pass this stability test ($n = 8$, $P = 0.059$).

4.5 Discussion

The clinical outcome of patients with prostate cancer is well-established to be highly variable, even within risk stratified groups such as D'Amico. Attempts to address this have been made, including further categorisation into favourable- and unfavourable-Intermediate Risk disease groups[172] and the development of the CAPRA classification system[88]. Molecular medicine-based approaches have also been used to develop tissue-based assays and genomic tests for aggressive disease[173–175].

A more holistic assessment of cancer status prior to invasive tissue biopsy would be clinically useful. Urine biomarkers present such a propsect, and may be used to supplement the current clinical standards for stratification of prostate cancer patients. Previous risk models developed using urine biomarkers have been designed specifically for singular purposes, such as the detection of prostate cancer following a negative biopsy (PCA3 test), or to detect Gs $\geq$ 3+4 upon an initial biopsy[12,99,102,164]. Here the PUR risk model was developed to provide a non-invasive and simultaneous assessment of the proportions of non-cancerous, "normal" tissue and D'Amico Low-, Intermediate- and High-risk prostate cancer harboured by a patient. The use of D'Amico risk types as an ordinal outcome, as opposed to a more binary biopsy-based one, is unique and could aid the deconvolution of complex cancerous states into more clinically translatable forms for monitoring the development of disease over time. For example, in men diagnosed with lower grade Gleason 6 disease and displaying a primary PUR-1, they may be enrolled onto a very low-frequency active surveillance program and monitored non-invasively over time.

For prediction the presence of significant prostate cancer at an initial biopsy, PUR compared favourably to other published biomarkers that use less complex transcript expression systems, and involve fewer probes[12,99,102,164]. The PUR classifer, based on the cf-RNA expression of 36 gene-probes relative to *KLK2*, *RPLP2* and *GAPDH*, can be used as a versatile predictor of clinical risk. Of note were the inclusion of well-described genes *PCA3*, *TMPRSS2-ERG* and *HOXC6* within the optimal PUR model defined by the LASSO criteria, while *DLX1* was not. The ability of PUR-4 status to predict TRUS detected Gs $\geq$ 3+4 was similar (AUC = 0.76; 95% CI = 0.69 – 0.83, Test) to these published models using *PCA3*/*TMPRSS2-ERG* (AUC = 0.74 - 0.78)[102,164] and *HOXC6*/*DLX1* (AUC = 0.77)[99].

Current clinical practice assesses patient's disease using PSA, needle biopsy of the prostate and mpMRI. However, up to 75% of men with a raised PSA ($\geq$3 ng/ml) are negative for prostate cancer on biopsy[7,54], whilst in absence of a raised PSA, 15% of men are found to have prostate cancer, with a further 15% of these cancers being high-grade[176]. This illustrates the considerable need for additional biomarkers that can make pre-biopsy assessment of prostate cancer more accurate. In this respect we show that both PUR-4 and PUR-1 are each equally good at predicting the presence of Intermediate or High-risk prostate cancer as defined by D'Amico criteria or by CAPRA status, while in DCA analysis PUR provided a net benefit in a a representative population of UK patients. With the increased adoption of mpMRI it would be useful in future studies to correlate PUR, and other urine-based markers, with MRI findings and radical prostatectomy outcomes.

Variation in clinical outcomes are also well recognised for patients entered onto AS.[177] We found that PUR worked well when applied to men on AS monitored by PSA and biopsy, and also in patients monitored by mpMRI. A potential limitation of this study is that we have not been able to test the PUR stratification in an independent and more conservatively managed active surveillance cohort. However, based on our observations approximately 13% of the RMH AS cohort could have been safely removed from AS monitoring for a minimum

of five years. An interesting feature is that in some patients the PUR urine signature predicted disease progression up to five years before it was detected by standard clinical methods. This prognostic information could potentially also aid the reduction of patient-elected radical intervention in active surveillance men which in some cohorts can be as high as 75% within three years of enrolment[177]. Indeed, we would view the use of PUR within the context of active surveillance as its major potential clinical application. Repeated longitudinal measurements of PUR status could help correctly assess and track a patient's risk over time in a non-invasive manner. The stability investigations while promising, were carried out in small numbers and *in silico*, it would be beneficial in future studies to collect multiple temporally-spaced samples from more patients to more completely assess stability.

Unfortunately, the predictive performance of PUR in AS usage appears to be confounded, or at least diminished by the prognostic ability of D'Amico risk categories of patients alone in the RMH AS cohort presented here. Risk stratification of AS patients based solely on D'Amico risk showed that Intermediate-risk patients were approximately 6.5 times more likely to meet the clinical criteria for progression than those Low-risk patients, similar to the performance of PUR alone. A possible explanation for this is that as the PUR classifier is trained on D'Amico risk, it is entirely possible that the AS performance is simply a reflection of accurate classification to the D'Amico labels, and not a deeper insight into a patient's disease status. Alternatively, it is possible that Intermediate risk patients should not be placed on AS programmes and the observations here are a reflection of that. Regardless, this is a question that cannot be answered with this cohort and requires more data, ideally in a larger future study across multiple centres with differing AS enrolment criteria. The design of such a study is considered in Chapter 8. PUR did show some additional utility when combined with D'Amico information, where those deemed most, and least at-risk of progression would be given more certainty, though at the cost of increased variance for low-PUR Intermediate and high PUR Low-risk patients (Figure 4.10). However, more data are required to be able to draw sound conclusions about predictive ability in active surveillance with the small numbers and high variance presented here.

Regardless of the limitations and shortfalls described above, PUR represents a promising new & versatile urine risk model capable of detecting aggressive prostate cancer, albeit with a need for external validation. The differences in cf-RNA profiles across the spectrum of patients in the Movember GAP1 Urine Biomarker cohort leaves no doubt that the presence of prostate cancer substantially influences the transcripts found in urine samples.

There is certainly scope for improvement within this study; whilst using D'Amico as the training label obviously shows good clinical utility for PUR, it may be suboptimal for biopsy prediction as D'Amico was initially developed for predicting treatment failure following radical therapy[63]. Similarly, the LASSO-penalised ordinal regression used here is useful for relatively direct interpretation of model coefficients, but is univariate and reliant on linear patterns in univariate gene expression across disease statuses. A good example of where the regression implemented here may lose information in this regard is the fact that only 40 to 50% of prostate tumours are *TMPRSS2/ERG* fusion positive[178]. This results in higher-grade, *TMPRSS2/ERG* negative tumours being predicted a lower risk score than might be achieved using a method that can account for this non-linearity. As discussed in 3, different machine learning algorithms can learn different decision spaces, accounting for interactions across multiple genes and lend themselves well to robust analysis methods through integration of bootstrap resampling. In the next chapter, I will explore the potential for increased performance by using different combinations of training labels and machine

learning algorithms.

Chapter 5

An empirical exploration of supervised machine learning algorithms and validation strategies

5.1 Summary

Exploring the potential gain from the application of disparate machine learning algorithms is important to fully utilising datasets, even more so where data are limited and expensive to gather. Optimising the combination of algorithm and training label for predictive modelling can improve the utility of an clinically implementable test based upon machine learning, with little cost. In this chapter I present the results generated from exploring a number of machine learning algorithms to the NanoString data in the Movember GAP1 cohort and explore the consequences.

The main aim was to experimentally assess whether different algorithms do in fact perform differently across a range of clinical outcomes, and if so, does one perform consistently best? Secondly the data themselves are investigated to ascertain whether certain clinical outcomes are more predictable than others with the available data. Finally, the impact of incorporating clinically available variables and feature importances are explored to see whether the predictive ability of models can be increased by using data readily collected by clinicians as part of the normal clinical pathway.

Results showed that whilst the LASSO regression employed by PUR in the previous chapter did indeed perform well, it was never the optimal algorithm for any outcome to be predicted. Instead, Random Forests consistently produced the most discriminatory models for predicting biopsy outcome, and are well suited to non-linear gene expression patterns. The predictive performance of fitted models was highly dependent on the random split of data chosen for training, leading to the conclusion that the NanoString dataset itself is highly variable, and more robust methods employing internal validation should be considered for training models in future.

5.2 Background

The No Free Lunch Theorem suggests that for any given algorithm or predictive model, superior performance in one class of problem comes at the cost of reduced performance in another[179]. Put simply, there is no universally "best" machine learning algorithm with which to predict all possible clinical outcomes for a patient with any given training label; even the most accurate of models will perform sub-optimally in *some* test case. Extracting maximal utility from a dataset is important in any application of statistical learning, however when patient treatment and outcomes are of concern it is critical to ensure data is used optimally, maximising the chances of successful results being robust, and able to be validated in further patient data[180,181].

The previous chapter showed that the PUR model possesses good clinical utility for predicting a range of scenarios prior to an initial biopsy, and an apparent ability to predict long-term outcomes in active surveillance use (See Chapter 4 for full details). PUR is a LASSO-based ordinal logistic regression model trained on D'Amico Risk categories, and was able to predict both D'Amico risk and biopsy outcome with a clinically acceptable degree of accuracy (AUC $>$ 0.75 in a test dataset). Given its relative simplicity, a LASSO ordinal regression will only discriminate linear relationships between variables and outcomes. Prostate cancer is highly heterogeneous; clinically, spatially and molecularly[182], so it is unlikely that all important predictors of prostate cancer outcome are linearly and monotonically related to disease severity.

Typical machine learning algorithms, such as the LASSO-penalised ordinal regression of PUR, represent "narrow" or "weak" artificial intelligence; able to predict exactly what trained on, but performing poorly in other scenarios[183,184]. As discussed in Chapter 4, this was not found to be the case for PUR, displaying utility for a range of endpoints. This may be caused by one, or both, the existence of latent information in the NanoString dataset where genes provide information towards more than one outcome as they are fundamentally linked to the pathobiology of the disease, or an overlap or confounding in clinical outcomes exists; where an increase in D'Amico Risk is in part defined for a majority of patients by an increase in Gleason pattern. This leads to the conclusion that it should be tested whether PUR represents the optimal means for predicting multiple clinical outcomes by exploring the implementation of different algorithms and training labels within the Movember GAP1 NanoString dataset. The most parsimonious approach for any attempt to access potential latent information would be to apply different machine learning algorithms, capable of modelling and representing disparate solutions from identical component search spaces. Additionally, altering the training label used for model fitting may yield changes in clinical utility.

A number of known-prognostic markers such as PSA levels and patient age are recorded as part of the treatment pathway for all patients, and would be available at the time of a urine sample being collected prior to biopsy. The most commonly employed approach is usually to only consider the added benefit of including clinically available information after initial model development[96,97,99,102]. This approach allows for simpler evaluation by isolating components, and a greater insight into cancer biology, if the study design allows for causal inference. However, it assumes that none of the underlying biomarkers interact with, or are dependent on, the clinically available parameters. Consideration of clinically available features such as age and PSA levels for predictive ability alongside NanoString or other biochemical markers adds no additional cost or complexity but may lead to higher predictive ability.

In this chapter, the primary aim was to undertake a pilot study searching through a number of algorithm - training label - outcome prediction permutations in an attempt to answer three key questions:

1. Does model performance vary when the training label is altered?
2. Do different machine learning algorithms perform differently?
3. Are certain clinical questions easier to answer than others, within the confines of the Movember GAP1 NanoString dataset?

Following this, the impact of integrating clinically available parameters is quantified and their relative importance for predicting a range of outcomes measured against the NanoString gene-probes. The final aims of this chapter are to decide upon a final modelling strategy that can be used to produce robust, interpretable models that are capable of predicting clinically relevant outcomes better than current standards of care.

5.3 Methods

Three main algorithms were applied within this chapter to fit statistical models; LASSO-penalised ordinal regression, Random Forests, and Gradient Boosting Machines (GBMs). Models were additionally produced as a meta-ensemble from the output of these three algorithms. For full details see Section 3.2 of Chapter 3.

5.3.1 NanoString data

The NanoString dataset described in Chapter 4 was used here with the exception that samples from raised PSA, negative biopsy patients ($n = 129$) were included. As these patients were actively biopsied, their disease status is arguably more known than those patients without any biopsy information. Of course, there is likely to be missed disease within both groups, given the prevalence of prostate cancer in general, and the inaccuracies of TRUS-biopsy (see Chapter 2 for further details).

5.3.2 Curation of Training and Test datasets

In order to provide an unbiased assessment of model effectiveness, it is necessary at the very least to create a hold-out validation dataset that is not actively used for model fitting. As in Chapter 4, this is often achieved through a stratified random sampling approach, though usually with an additional, externally collected dataset for true validation[21].

Training and Test datasets were created as follows: data were sampled 1,000 times, randomly selecting 67% of data for training and 33% for validation at each resample without replacement. The proportions of D'Amico clinical categories were held constant at each resampling iteration. Median expression was calculated for each gene-probe across all samples in both training and test datasets and the Euclidean distance between the training and test datasets measured at each iteration. The iteration with the minimum variable distance between the two datasets was used to select the final training ($n = 347$) and test ($n = 225$) datasets, whilst samples collected from high PSA-negative biopsy and metastatic patients were additionally included in the final test dataset ($n = 354$) dataset. The impact of randomly chosen training and test splits is explored in section 5.4.3.

5.3.3 Model training labels and variables

Models were fitted to one of six different training labels based upon clinical endpoints. Two labels were ordinal, multi-class variables, with the remainder simple binary classification (Table 5.1).

Table 5.1: Training labels used as targets for model construction. > indicate the direction of a continuing ordinal variable, where only forward direction is considered possible.

Label Name	Outcome Type	Label Levels	Clinical Outcome
D'Amico	Ordinal	NEC > L > I > H	NEC and D'Amico Risk categories
TriSig	Ordinal	NEC > LC > HC	NEC, Predominantly Gleason Pattern 3 (3+3 or 3+4), Predominantly Gleason 4 or greater ($\geq$ 4+3)
Cancer vs No Cancer	Binary	1,0	NEC vs any D'Amico outcome
Cancer vs High-Risk Cancer	Binary	1,0	D'Amico High Risk vs all other outcomes
Extremes	Binary	1,0	Subsampling NEC and High-Risk samples only for training
Gleason $\geq$ 4+3	Binary	1,0	Gleason $\geq$ 4+3
Gleason $\geq$ 3+4	Binary	1,0	Gleason $\geq$ 3+4

TriSig is a three-level ordinal outcome, categorised according to the dominant Gleason pattern in a sample and is based on an assumption that the strongest molecular signal will be derived from the most common cellular morphology in the absence of highly detailed histopathology data.

5.3.4 Model construction and selection of user-tunable parameters

Each of the three algorithms used have user-tunable parameters to control for a variety of conditions or outcomes during the model fitting process. For example, the elastic net penalty α, controls for the severity of penalisation between LASSO ($\alpha = 1$) and Ridge ($\alpha = 0$) regression penalisation. The number of tunable parameters varies considerably between algorithms, where α is the only meaningful parameter within the elastic net framework, XGBoost employs five complex, interacting parameters requiring careful tuning (Chapter 3.2). The parameters used here are seen in Table 5.2.

Table 5.2: The tunable parameters of the machine learning algorithms implemented, their possible ranges and the values used in practice.

Algorithm	Tunable Parameters	Possible Range	Values Used
LASSO (glmnet)	α	0 - 1	1
Random Forest	p subsample	0 - p	0 - p
	Trees grown	1 - ∞	801
XGBoost	Tree depth	1 - ∞	3 - 12
	Child Weight	1 - n	1 - 20
	p subsample ratio	0 - 1	0.5 - 1
	p subsample	0 - 1	0.15 - 0.7
	η	0 - 1	0.001 - 0.3

α = elastic net penalty; p = number of input variables
n number of samples η = learning rate

LASSO

Constrained continuation ratio LASSO ($\alpha = 1$) ordinal logistic regression models were created using the *ordinalNet* package, with default values, selecting the minimum lambda value that returned the best performance metric over 20-fold cross-validation. For a full description see Chapter 3.2. Chosen performance metrics were dependent on classification type; binary classification training labels maximised the AUC, whilst the two ordinal models were optimised to minimise the multiclass log-loss function:

$$-\sum_{C=1}^{M} y_{j,Label} \log(p_{j,Label})$$

Where M = number of classes, y = binary indicator (0 or 1) if class label *Label* is the correct classification for observation j and p - predicted probability observation j is of class C.

Random Forests

The number of features to be subsampled in Random Forest models p were tuned within 10-fold cross-validation loops, repeated 5 times in the training dataset, with the final model fit to the whole training dataset. Parameter tuning was carried out using the *caret* package, sequentially searching through all possible values for the variable subsample number as detailed in Table 5.2. The number of trees grown was fixed at 801, as it was empirically shown to be adequate, and an odd number allows for the rare case of a tie in individual tree votes to be solved. For a full description see Chapter 3.2.

Gradient Boosting Machines

Gradient boosting machines used the *xgboost* R package, implementing the XGBoost libraries and algorithm[146], for a full description see Chapter 3.2. The range of parameters to be tuned was established prior to optimisation and expanded to included every permutation, resulting in >10,000 combinations to be searched through.

For each term of the parameter grid a decision tree-based XGBoost GBM was fitted within a 5-fold cross-validation, stratified on clinical category, with boosting stopped once the out-of-fold error metric has not improved in 20 boosting rounds. The final model was constructed over the entire training dataset with the parameters that maximised the cross-validated performance, with one less boosting round to avoid overfitting.

Meta-ensemble

Meta-ensemble predictors were constructed by calculating the simple mean of the LASSO, Random Forest, and GBM model outputs for a given sample, where all model outputs were in the range 0 - 1.

5.3.5 Evaluation of model performance

Constructed models were evaluated for clinical utility over different available outcomes ranging from clinically relevant biopsy outcomes and diagnoses following the result of a biopsy (Gleason $\geq$ 3+4, Gleason $\geq$ 4+3 and D'Amico $\geq$ Intermediate Risk), through to more clinically irrelevant outcomes that give more insight into the how difficult certain outcomes are to model with the available data (D'Amico $\geq$ High Risk and any diagnosis of prostate cancer following a biopsy). Area under the ROC curve (AUC, see Section 3.2) was used to quantify predictivity of trained models over these outcomes. Kruskal-Wallis tests were used to test for any difference in AUC between models, with pairwise comparisons made using Wilcox rank-sum tests and Benjamin-Hochberg adjustment for multiple comparisons.

5.3.6 Assessment of dataset variability

The effects of specific training-test split were examined in the most performant set of models by resampling the NanoString dataset. Data were randomly split 67/33 into Training and Validation datasets, stratified by D'Amico category. The selected models were then fit to these splits of data and AUCs from the models recorded in the validation dataset across outcomes. This process was repeated 1,000 times, at each iteration creating a new training/test split without replacement. All statistical tests are reported as the two-sided probability.

5.3.7 Inclusion of clinically available parameters

A total of nine clinically available or technical variables were integrated into the NanoString dataset and used for modelling to evaluate their clinical utility (Table 5.3. Where nominal categorical variables were considered they were dummy coded (one-hot encoding) to avoid misinterpretation of order by algorithms that re-encode categorical variables as numeric[185].

Table 5.3: Available non-NanoString parameters for use in predictive models and feature selection

Variable	Description	Levels
Age	Age in years	Continuous
PSA	Serum PSA (ng/mL)	Continuous
RNA Amount	Quantity of input RNA (ng)	Continuous
pH	pH of urine at collection	0 - 14
DRE	Size of prostate as estimated by DRE	Small, Medium, Large, Unknown
Family History	Previous family history of prostate cancer	Yes, No, Unknown
Smoking	Smoking history	Yes, No, Unknown
Urine Vol	Volume of urine collected (mL)	Continuous
Alcohol	Consumption of alcohol	Yes, No, Unknown

5.3.8 Feature Selection

The importance of features within each modelled outcome was investigated, allowing for relative importances of clinically available parameters and NanoString gene-probes to be compared. Feature selection was undertaken by application of the Boruta algorithm[149], fully described in Chapter 3, Section 3.2. Variables were positively retained so long as they were not significantly worse than the maximally performing Shadow feature, that is variables deemed "Tentative" were selected. Boruta was applied for each training label using a maximum of 100 repeats, accepting or rejecting variables at a significance level of $P < 0.01$.

5.4 Results

5.4.1 Choice of training labels, clinical outcomes and machine learning algorithm

All permutations of training label and algorithm were used to fit machine learning models to the NanoString data from 347 samples across the 167 gene-probes. This resulted in a total of 28 multivariable risk models. Each model was assessed for its predictive accuracy in the validation split of 225 samples, according to the AUC measured against the range of available clinical outcomes.

Prediction accuracy by clinical outcome

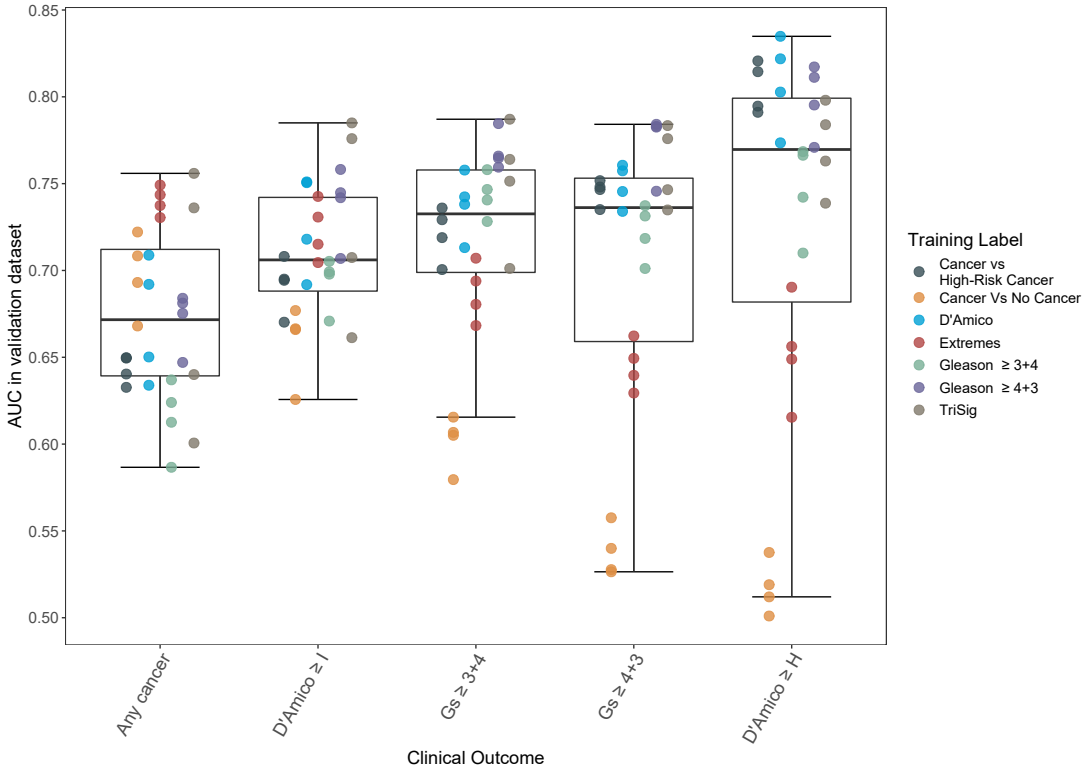


Figure 5.1: Average AUC returned from models predicting each clinical outcome (x-axis), algorithms and training labels are grouped, with coloured points detailing the specific training label. The algorithm used to train each model is not shown here.

Table 5.4: P values derived from pairwise comparisons of AUCs with respect to clinical outcome using Wilcoxon rank sum test and Benjamin-Hochberg adjustment.

	Any cancer	D'Amico ≥ I	D'Amico ≥ H	Gs ≥ 3+4
D'Amico ≥ I	0.026			
D'Amico ≥ H	0.0093	0.0428		
Gs ≥ 3+4	0.0131	0.2573	0.0575	
Gs ≥ 4+3	0.0321	0.2933	0.0605	0.9805

Significant differences were observed in the accuracy with which clinical outcomes could be predicted, when considered across all training labels ($P < 0.001$, Kruskal-Wallis rank sum test, Figure 5.1). The average ability of fitted models to discriminate the presence of any cancer regardless of training label was significantly lower than for all other outcomes (all $P < 0.05$, pairwise Wilcoxon rank sum test, Figure 5.1 & Table 5.4). Prediction of D'Amico ≥ Intermediate risk also returned significantly lower AUCs than when predicting D'Amico ≥ H ($P < 0.05$, Wilcoxon rank sum test, Figure 5.1 & Table 5.4). No other significant

differences were observed at this grouped level (Table 5.4).

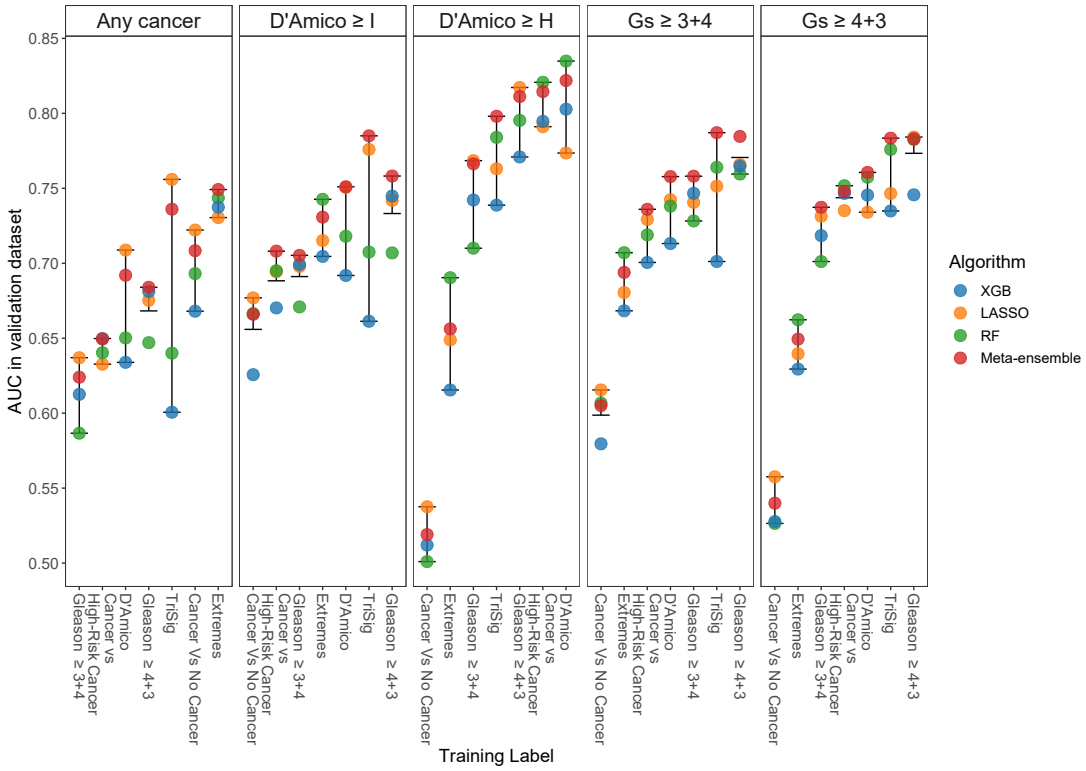


Figure 5.2: AUC performance of trained models (x-axis) in the validation dataset. Facets detail the clinical outcome being predicted, with each coloured points detailing the specific algorithm used for to generate the model.

When considering the ability to predict each clinical outcome in turn, pairwise comparisons were made between the models and statistically significant differences in AUC were observed only for the prediction of D'Amico High Risk ($P < 0.05$, Table 5.5). These differences were predominantly between the models trained on the Cancer vs No Cancer training label compared to all others (Table 5.5, $P < 0.05$, pairwise Wilcox rank sum test).

Table 5.5: P values derived from pairwise comparisons of AUCs with respect to training label used when predicting an outcome of D'Amico ≥ H. Calculated using Wilcoxon rank sum test and Benjamin-Hochberg adjustment.

Training Label	D'Amico	Extreme	HC v C	Gs ≥ 3+4	Gs ≥ 4+3	TriSig
C v NC	$P = 0.043$	$P = 0.043$	$P = 0.043$	$P = 0.043$	$P = 0.043$	$P = 0.043$
D'Amico		$P = 0.043$	$P = 0.72$	$P = 0.043$	$P = 0.537$	$P = 0.15$
Extreme			$P = 0.043$	$P = 0.043$	$P = 0.043$	$P = 0.043$
HC v C				$P = 0.043$	$P = 0.886$	$P = 0.15$
Gs ≥ 3+4					$P = 0.043$	$P = 0.537$
Gs ≥ 4+3						$P = 0.247$

Prediction accuracy by training label

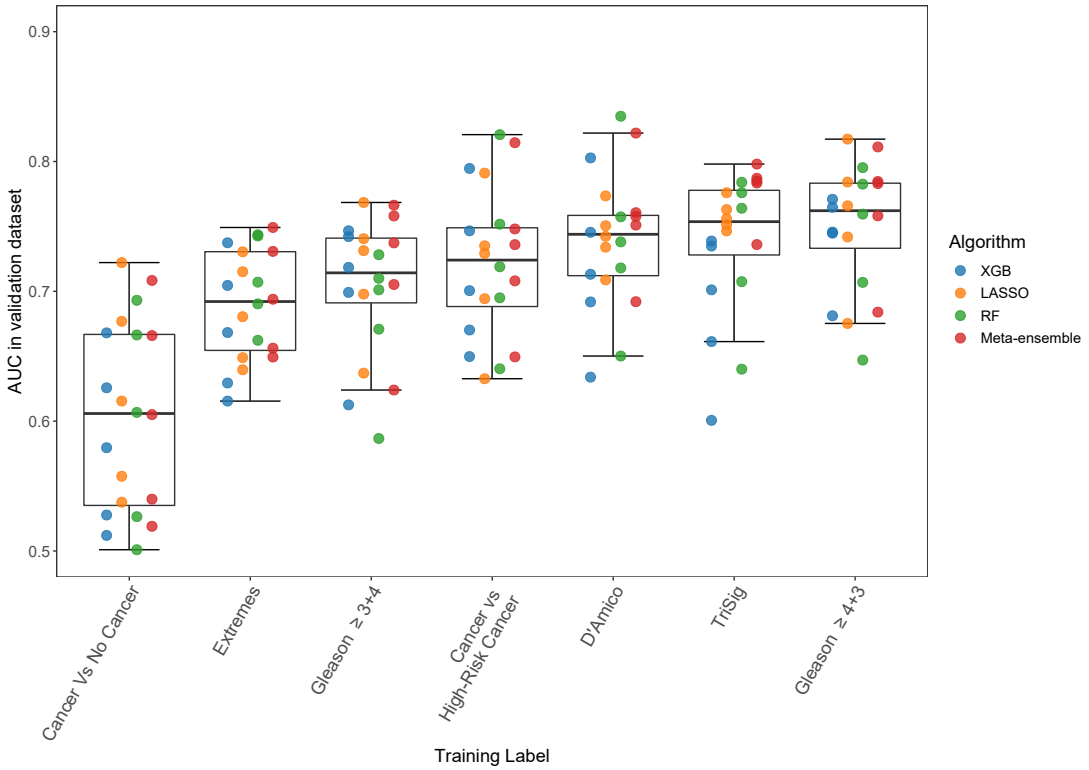


Figure 5.3: Average AUC returned from models according to the training label used (x-axis) to fit the model, averaged over both the outcome being predicted (not shown) and algorithm used to fit the model (colour)

Differing the training label used to fit models had a significant effect on the AUCs reported across all algorithms and clinical outcomes assessed ($P < 0.001$, Kruskal-Wallis test, Figure 5.3). Training of models on the binary Cancer vs No Cancer label resulted in significantly lower AUCs (median AUC = 0.61; IQR = 0.13, $P < 0.001$, Wilcox rank sum) than any other

training label, whilst the binary Gleason $\geq$ 4+3 label returned models with significantly higher accuracy (median AUC = 0.76; IQR = 0.05) than the Extreme and Gleason $\geq$ 3+4 labels (median AUC = 0.692; IQR = 0.076 and AUC = 0.714; IQR = 0.05, respectively. $P < 0.01$, pairwise Wilcoxon rank-sum tests, 5.6)

Table 5.6: P values from pairwise comparisons of AUCs between different training labels using Wilcox rank sum test and Benjamin-Hochberg adjustment.

	Cancer vs High-Risk cancer	Cancer vs No cancer	D'Amico	Extremes	Gleason $\geq$ 3+4	Gleason $\geq$ 4+3
Cancer Vs No Cancer	0					
D'Amico	0.318	0				
Extremes	0.1159	0.0006	0.003			
Gleason $\geq$ 3+4	0.5468	0.0001	0.0978	0.3271		
Gleason $\geq$ 4+3	0.1366	0	0.3369	0.0004	0.0067	
TriSig	0.2817	0	0.5468	0.002	0.0232	0.5468

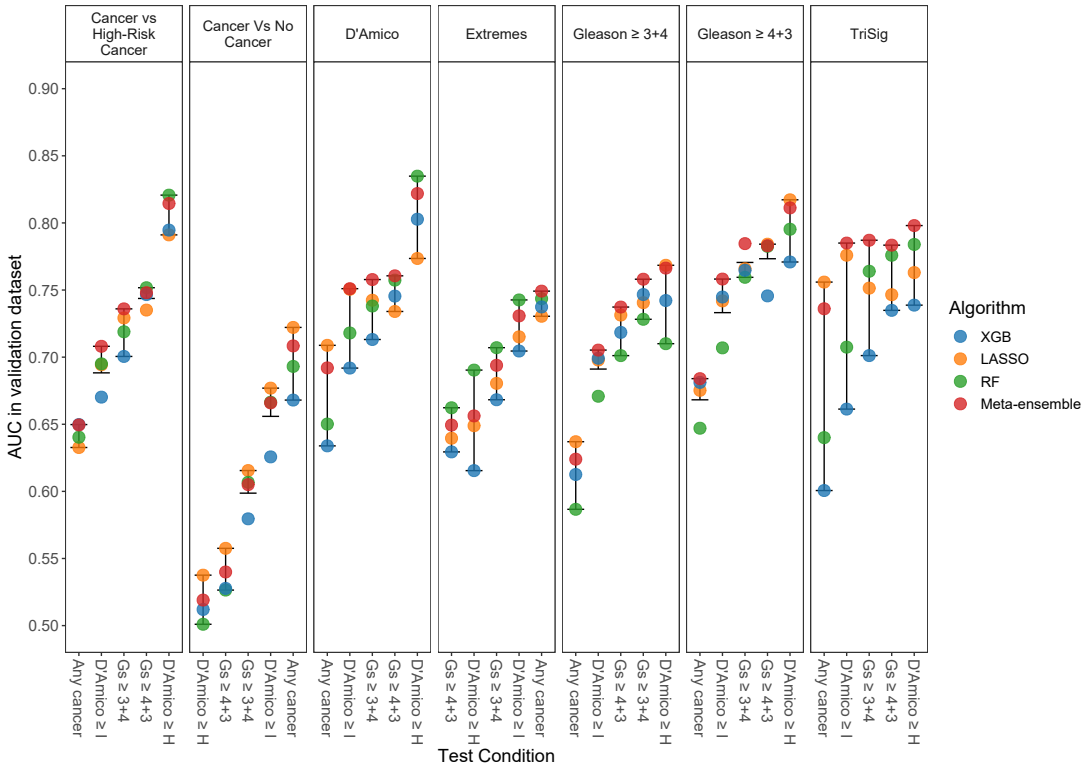


Figure 5.4: Detailed AUCs from models in the validation dataset according to the training label used to specify the model (panels), across different clinical outcomes (x-axis). Coloured points show the machine learning algorithm used to fit the model in the training data

The performance of models predicting each clinical outcome within a training label

varied greatly, with a mean AUC difference of 0.24 between the minimally and maximally performing model(Figure 5.4). The Cancer Vs No Cancer training label resulted in the worst performing models regardless of algorithm, with models predicting D'Amico $\geq$ H and Gs $\geq$ 4+3 returning AUCs close to that of a random predictor (AUC = 0.52 and 0.54, respectively, Figure 5.4). Pairwise comparisons of AUC showed that TriSig was the only label to not display significant differences in AUCs ($P > 0.05$ Wilcox rank sum test, Figure 5.4 TriSig panel).

Prediction accuracy by machine learning algorithm

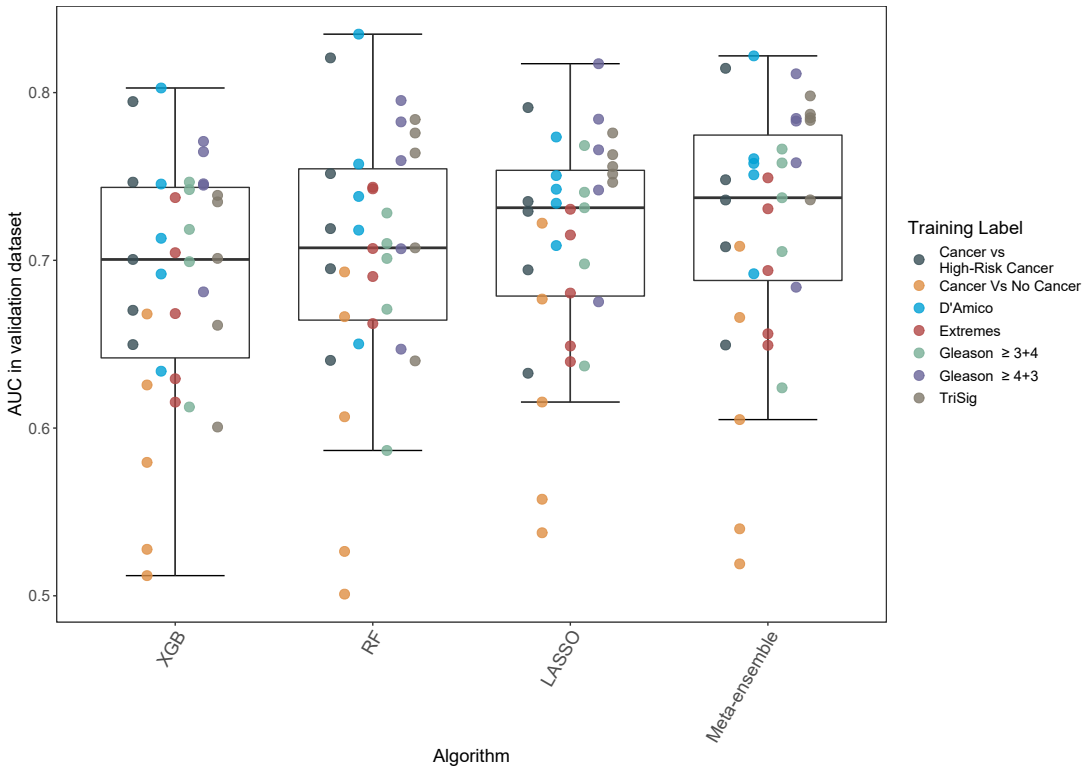


Figure 5.5: Average AUC performance of models in the validation dataset according to the machine learning algorithm used to define them. Point colour details the specific training label used for model fit. The clinical outcomes being predicted are not indicated here.

Choice of machine learning algorithm had no significant effect on model predictivity when all training labels and outcomes were considered together ($P > 0.05$ Kruskal Wallis test, Figure 5.5). All algorithms displayed a large variance in predictive accuracy depending on the training label used (mean AUC range within algorithm = 0.302). Models fitted using the Cancer Vs No Cancer label were particularly of note, returning AUCs low enough to constitute outliers within the results from the Random Forest, LASSO and Meta-ensemble algorithm-derived models (Figure 5.5).

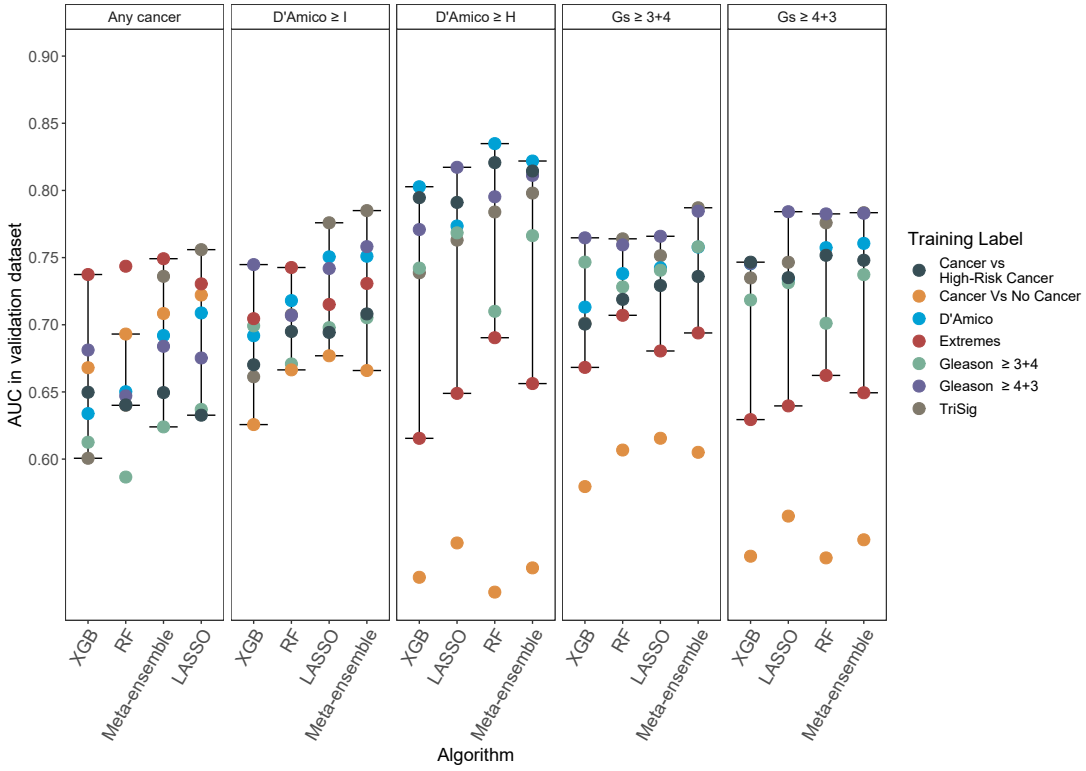


Figure 5.6: AUCs returned from models trained using different machine learning algorithms (x-axis), with panels detailing the specific clinical outcome being predicted. Coloured points detail the training label used to fit models.

Algorithm-dependent differences in AUC accuracy of models across different clinical questions showed no significant differences in AUC (all $P > 0.05$, pairwise Wilcox rank-sum tests with Benjamin-Hochberg corrections for multiple comparisons, Figure 5.6). Random Forest-based models or the meta-ensemble predictions always appeared as one of the top two algorithms, as determined by ranking the AUCs for each outcome and training label used.

Conclusions

Following the initial modelling approaches presented above the decision was made to reduce the number of models and training labels used moving forwards to reduce complexity. Only the most performant combinations were retained for resampling approaches and integration of clinically available parameters. The XGBoost algorithm was not selected for further study, as whilst it did not perform significantly worse than the others, the opaqueness and requirement for extensive parameter tuning did not result in any detectable performance increase. Similarly, the Cancer Vs No Cancer and Extreme training labels were discarded as they consistently resulted in models less accurate than others.

Models positively retained for further investigations were; LASSO-based regression models trained on the TriSig and Gleason $\geq$ 4+3 labels, Random Forest-based models trained on D'Amico and Gleason $\geq$ 4+3 labels, and meta-ensembles combining the results from these two algorithms. No changes were made in the clinical outcomes used to assess clinical

utility of trained models.

5.4.2 Integration of clinical and non-NanoString biochemical parameters

A total of nine additional variables were available for integration into the NanoString dataset, two urine sample-derived variables (RNA amount and pH) with the remainder clinically available parameters (Table 5.3). Categorical variables were dummy encoded (see Methods above) and integrated into the dataset for both modelling and later feature selection using the Boruta algorithm.

5.4.3 The effects of clinical variables and resampling training/test splits

Following integration of the additional clinical variables, models were fit to each one of 1,000 random training/test splits of the datasets, and assessed for clinical utility. Due to serum PSA levels partially defining the D'Amico Risk categories, predictions of D'Amico Risk category were not included in these assessments. Both the choice of input variables and training labels had significant effects on the predictive accuracy of trained models. The training labels retained for this portion of the study were the ordinal D'Amico and TriSig labels in addition to the binary Gleason $\geq$ 4+3 label.

Differences in AUC dependent solely on the specific training and test split of the dataset were considerable (mean 95% CI of AUC = 0.168). Large differences in AUC were also observed according to the outcome being predicted (Figures 5.7, 5.8 and 5.9). Differences in AUC were assessed statistically by counting the number of times a greater AUC was returned in each of the 1,000 resamples when comparing differing sets of input variables. There were no significant differences in the AUCs returned between models using NanoString genes only and models using clinical variables only, regardless of algorithm, training label, or outcome being predicted (all $P > 0.05$ by simulation analysis of 1,000 paired resamples).

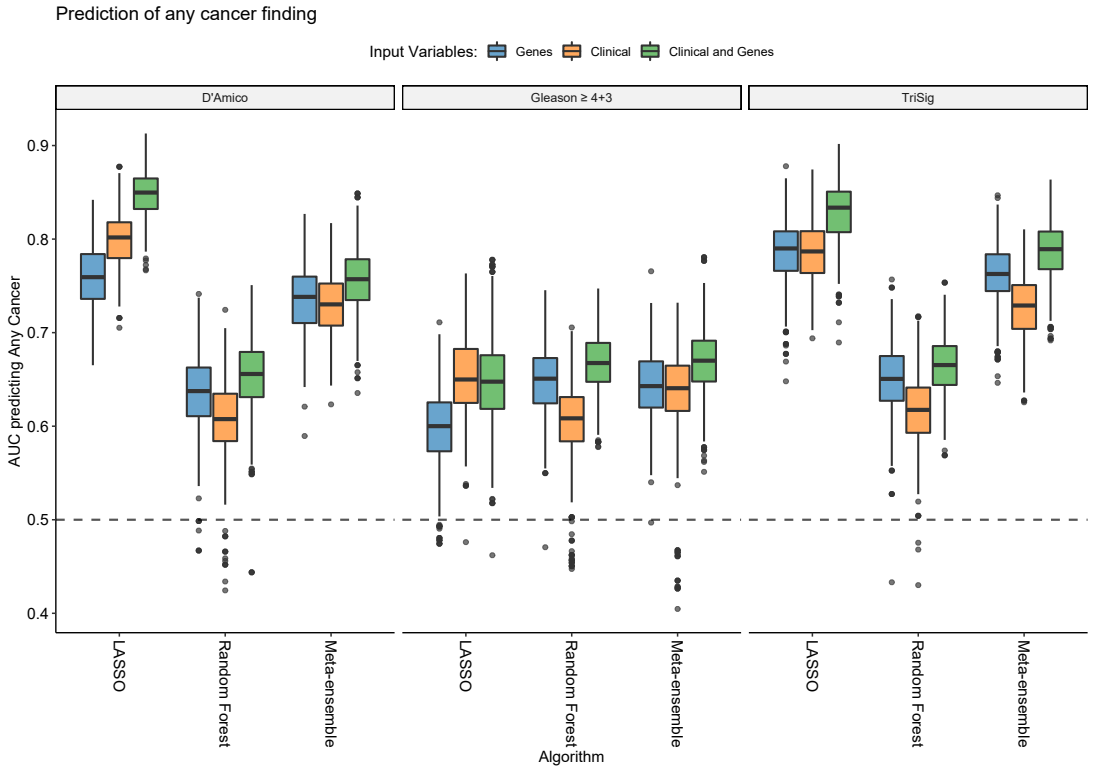


Figure 5.7: AUCs returned when predicting any cancer outcome on biopsy, from models fit the training labels: D'Amico category, binary Gleason = 4+3 outcome, or TriSig (facets). x-axis in each facet details the different algorithms used. Models were fit to 1,000 random training and test splits of the data, with different subsets of input variables used at each split, dictated by colour Blue - NanoString gene-probes only; Orange - Clinically available parameters only; Green - both NanoString and clinical variables.

The prediction of any cancer on biopsy showed highly variable results across the 1,000 resamples, in some cases resulting in models with predictive accuracy below that of a random predictor (AUC = 0.5) depending on the training label used (Figure 5.7). The combination of NanoString genes and clinically available parameters performed significantly better than consideration of only NanoString genes for the LASSO-based models trained on D'Amico category ($P = 0$, simulation in 1,000 resamples) and binary Gleason $\geq$ 4+3 ($P = 0.028$, simulation in 1,000 resamples). The same combination of genes and clinical variables returned higher AUCs than models only using clinical variables when using D'Amico-fitted LASSO models ($P = 0.041$, simulation in 1,000 resamples), and meta-ensemble TriSig-trained models ($P = 0.041$, simulation in 1,000 resamples).

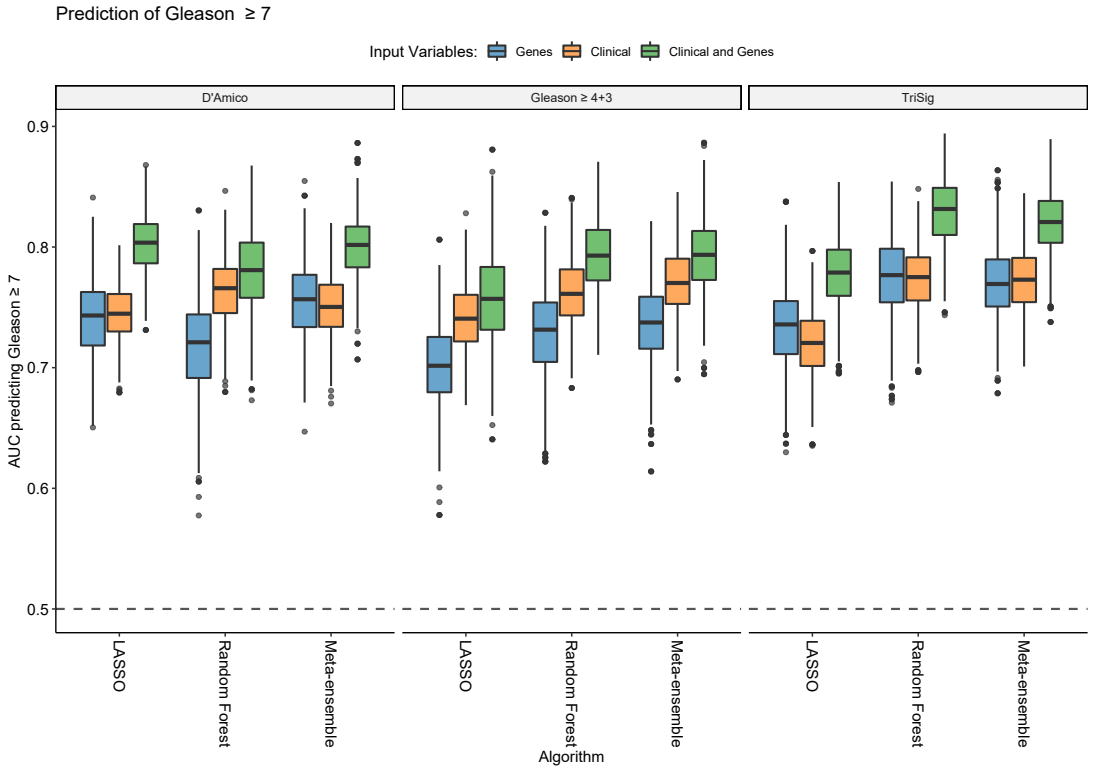


Figure 5.8: AUCs returned when predicting a biopsy outcome of Gleason = 7, from models fit to training labels: D'Amico category, binary Gleason outcome, or TriSig (facets). Models were fit to 1,000 random training and test splits of the data, with different subsets of input variables used at each split, dictated by colour Blue - NanoString gene-probes only; Orange - Clinically available parameters only; Green - both NanoString and clinical variables.

Prediction of Gleason ≥ 7 on biopsy was less variable and more accurate than when predicting any cancer outcome (Figure 5.8). The mean AUC for predicting Gleason ≥ 7 was greater than 0.72 for all models across algorithm and training labels(Table 5.7).

Once again the inclusion of clinically available parameters resulted in significantly higher accuracy from models than when using NanoString variables alone as input (all $P < 0.05$ by simulation, Figure 5.8). Clinical variables alone returned lower AUCs than when both NanoString and clinical variables were considered for the TriSig-trained LASSO and Random Forest models ($P = 0.047$ and 0.041, respectively, , simulation in 1,000 resamples) and the D'Amico trained LASSO and meta-ensemble models ($P = 0.004$ and 0.035, respectively, simulation in 1,000 resamples).

Table 5.7: Summary statistics of the AUCs returned from models predicting a biopsy result of Gleason $\geq$ 7.

Input Variables	Algorithm	Training Label	Median AUC	IQR AUC
Genes	LASSO	D'Amico	0.743	0.044
		Gleason $\geq$ 4+3	0.702	0.046
		TriSig	0.736	0.044
	Random Forest	D'Amico	0.721	0.053
		Gleason $\geq$ 4+3	0.732	0.049
		TriSig	0.777	0.044
	Meta-ensemble	D'Amico	0.757	0.043
		Gleason $\geq$ 4+3	0.737	0.043
		TriSig	0.769	0.039
Clinical	LASSO	D'Amico	0.745	0.031
		Gleason $\geq$ 4+3	0.741	0.039
		TriSig	0.720	0.037
	Random Forest	D'Amico	0.766	0.037
		Gleason $\geq$ 4+3	0.761	0.038
		TriSig	0.775	0.036
	Meta-ensemble	D'Amico	0.750	0.035
		Gleason $\geq$ 4+3	0.770	0.037
		TriSig	0.773	0.037
Clinical and Genes	LASSO	D'Amico	0.804	0.033
		Gleason $\geq$ 4+3	0.757	0.052
		TriSig	0.779	0.038
	Random Forest	D'Amico	0.781	0.046
		Gleason $\geq$ 4+3	0.793	0.042
		TriSig	0.832	0.039
	Meta-ensemble	D'Amico	0.802	0.034
		Gleason $\geq$ 4+3	0.794	0.041
		TriSig	0.821	0.035

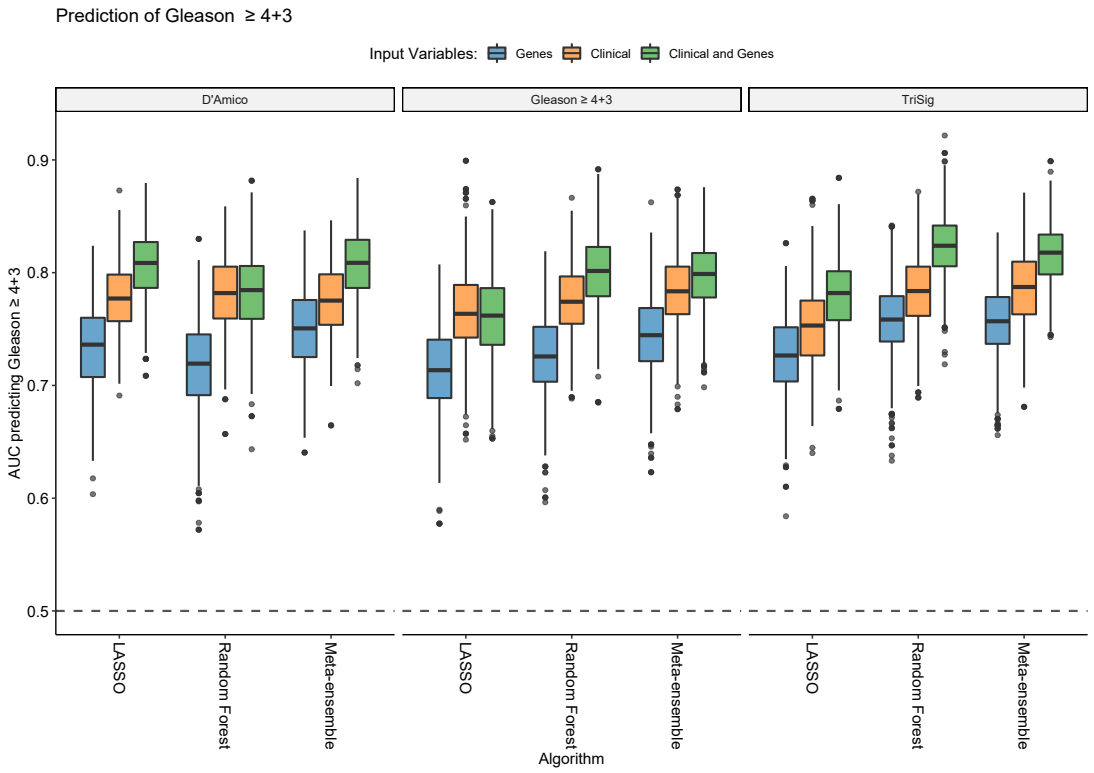


Figure 5.9: AUCs returned when predicting a biopsy outcome of Gleason = 4+3, from models fit to training labels: D'Amico category, binary Gleason outcome, or TriSig (facets). Models were fit to 1,000 random training and test splits of the data, with different subsets of input variables used at each split, dictated by colour Blue - NanoString gene-probes only; Orange - Clinically available parameters only; Green - both NanoString and clinical variables.

Where the detection of Gleason $\geq$ 4+3 disease prior to biopsy was considered, on average clinical variables alone outperformed the NanoString variables across each of the three training labels and algorithms apart from LASSO-based models trained on the Gleason $\geq$ 4+3 label (Figure 5.9).

Paired analysis of the differences in AUC at each of the 1,000 iterations showed that using both NanoString gene-probes and clinical variables results in models returning significantly higher AUCs than consideration of NanoString genes alone (all $P < 0.001$, simulation in 1,000 resamples), but not when compared to using clinical variables as the sole input variables (all $P > 0.05$, simulation in 1,000 resamples).

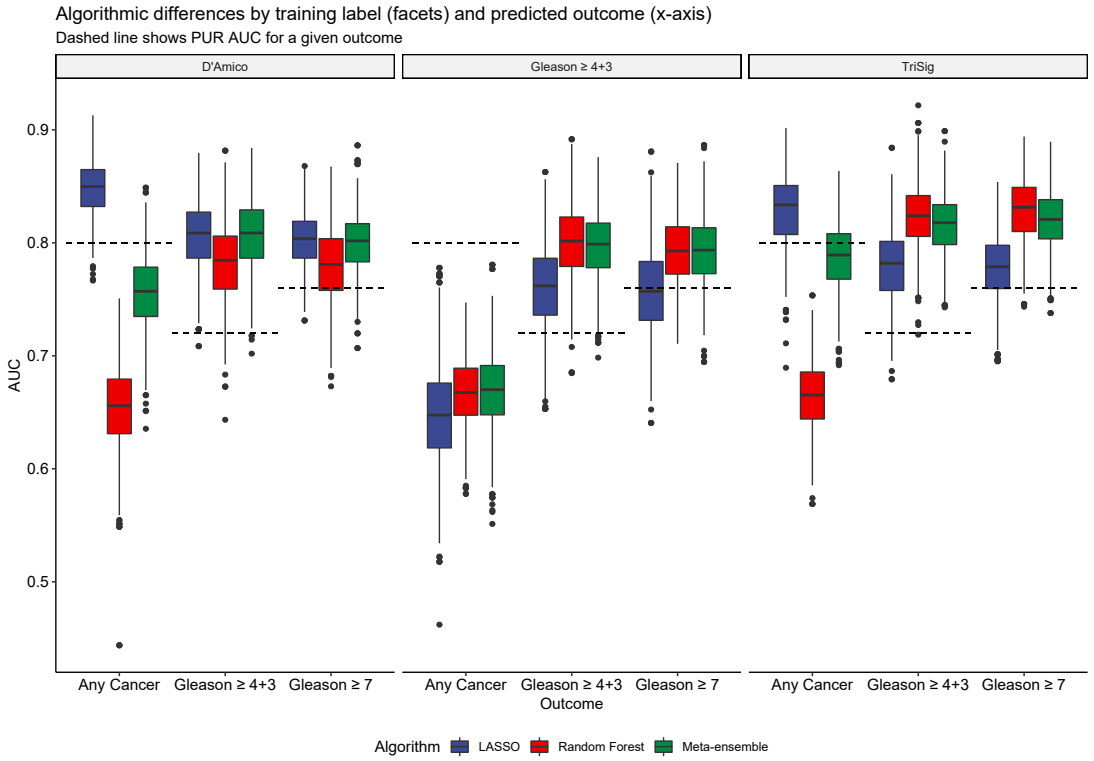


Figure 5.10: Predictive accuracy of models trained using both NanoString and clinical variables as inputs. AUCs were calculated by training models over 1,000 random training/test splits of the data and are presented on the y-axis. Differing clinical outcomes are shown on the x-axis, whilst fill colour denotes the algorithm used. Panels separate the results from the three different training labels

When models were trained the D'Amico risk categories, predicting any cancer as an outcome returned significantly different AUCs between algorithms (all $P < 0.001$ by simulation analysis of 1,000 resamples, Figure 5.10 D'Amico panel). When predicting any Gleason pattern on biopsy, no significant differences were observed between algorithms (all $P > 0.1$, simulation in 1,000 paired resamples).

Different algorithms did not result in significant changes to AUC when models were trained using the binary Gleason $\geq$ 4+3 label (all $P > 0.5$, simulation in 1,000 paired resamples, Figure 5.10 Gleason $\geq$ 4+3 panel). When models were trained using the TriSig training label, again only prediction of any cancer resulted in significantly different AUCs between algorithms, with LASSO ordinal regression models posessing higher predictive accuracy than Random Forests or meta-ensemble models, which were also different from one another (all $P < 0.01$, simulation in 1,000 paired resamples, Figure 5.10 TriSig panel).

5.4.4 Feature selection

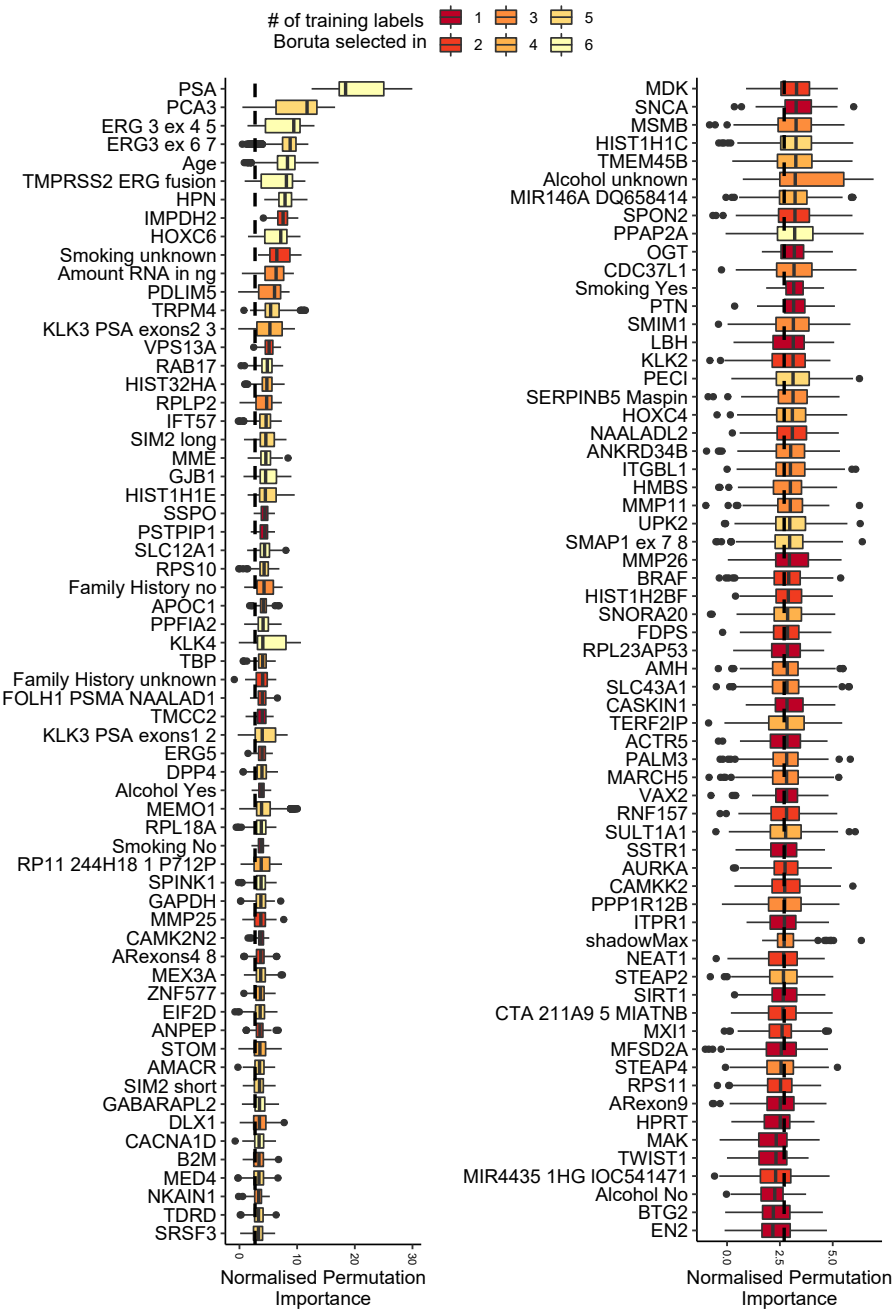


Figure 5.11: Normalised permutation importance for each variable, averaged across the six training labels (detailed in Table 5.1). All 167 gene-probes and clinically available parameters were supplied as inputs. Colours indicate the number of times each variable was confirmed over the training labels. For example, serum PSA is confirmed in every single training label, and on average, is the most important single feature. Dashed line indicates the median Shadow Max importance. Only features selected for at least one of the training labels are shown here.

The Boruta feature selection algorithm was applied using six training labels (TriSig, Gleason $\geq$ 4+3, Gleason $\geq$ 3+4, D'Amico, Any Cancer, Cancer vs High Risk Cancer) using the full NanoString and clinical data as inputs, assessing how well each feature predicts the training label of choice. A total of 185 variables were selected as either "Confirmed" or "Tentative" in at least one of the training labels examined (Figure 5.11). "Tentative" variables were retained, as the aims of this study are not to be as robust as possible, but an attempt to remove some of the unusable variance from the data.

Serum PSA was positively selected across all training labels, with a median permutation importance almost double that of the next most important variable, *PCA3* as quantified by NanoString. *PCA3* averaged as the most important NanoString derived variable, confirmed in five of the six training labels, only rejected when the Cancer vs High-Risk Cancer training label was considered (Appendix Table A.1). In comparison to the PUR model described in Chapter 4, all but two of the 36 genes were retained across the different training labels, with *MIC1* and *IGFBP3* rejected across all training labels (Appendix Table A.1).

The "Smoking Unknown" clinical parameter appeared to be an important variable, where it was selected in the D'Amico and TriSig models. However this can only be an artefact of data collection; it's not feasible for an unknown smoking status to hold information about the outcome of a patient. This was confirmed as, when examined the "Smoking Unknown" level accounts for 41% of all patients, highlighting the importance of thorough data collection and curation.

5.5 Discussion

5.5.1 The relative ease of predicting different prostate cancer outcomes

In this chapter we have shown that some clinical outcomes can be predicted with much higher accuracy than others, such as the specific prediction of High-Risk disease following a biopsy (Figure 5.1). However, prediction of such an outcome is not clinically useful, for example the identification of High-Risk patients in isolation can be well-predicted by markers already collected; it is the detection of patients with Intermediate-risk disease that is seen as a key clinical threshold[5] and the identification of High-Risk patients in isolation is not considered overly difficult. As expected, the clinically important outcomes such as D'Amico $\geq$ Intermediate Risk and Gleason $\geq$ 3+4 represent a more difficult problem, which may be due to underlying tumour heterogeneity of prostate cancer[182].

Currently used diagnostic frameworks such as D'Amico and the Gleason score roughly capture the clinical behaviour of prostate cancer, but may not be a good representation of the disease if molecular subtypes or heterogeneous groups within clinical categories exist. An example of this is seen when models using the No Cancer vs Any Cancer label were fit, resulting in poorly performing predictive models. When all clinical risk categories are combined in such a manner, clearly a large amount of variance in the "cancer" category will be present. This makes it difficult for any algorithm to appropriately model and represent the variance, the results of which are seen in the poor performance of models trained using the"No Cancer vs Any Cancer" label Figures 5.3 and 5.4).

5.5.2 Algorithmic choices

Smaller differences were observed in algorithmic performance than were initially hypothesised, as each of the algorithms used typically learn very different decision spaces. However,

whilst the differences observed were small and non-significant, it was decided that the complexity and opaqueness of XGBoost did not justify the performance returned from models fitted using it (Figure 5.5). Random Forests and the meta-ensemble based models did, however rank amongst the most performant algorithms. Given the relative ease of implementing Random Forests, they would be well suited for use in future work.

It may be that molecular variance within levels of a given training label makes boosting techniques suboptimal. If two or more molecular subtypes do exist within the same level of a label, when weights are attributed in a given boosting round to poorly classified samples from before, XGBoost cannot produce appropriate weak learners for all subtypes. This is a problem that may be improved with advances in semi-supervised machine learning methods applied to large prostate cancer datasets, such as variational autoencoders or using specific molecular subtypes as predicted by other unsupervised frameworks such as the DESNT classifier[112].

5.5.3 The importance of data splitting strategy

The specific split of data chosen for training and validating a model showed a very large effect on the apparent predictive utility of models. This resample-specific effect was so large that some models even returned predictions with an accuracy below that of a random predictor (Figure 5.7). This is a downfall of the more conventional machine learning methodology of a data splitting strategy, and inherent variability in the data. As not all information is used to learn from, it's very possible for small subsets of samples containing important patterns to not be represented in a training dataset. This results in models with poor generalisability; appearing accurate in the training set but performing poorly in previously unseen data.

Of course, the training/test/validation strategy is good for guarding against overfit and model tuning before true external validation, and works where the full population variance can be captured in a training set of an appropriate size. The limitations of this are that it requires at least three relatively large datasets, two of which cannot be used for training a model. This is not ideal in studies such as this one, with limited observations or scope to collect further samples. Instead it may be more appropriate to use a resampling or cross-validation based approach to derive a model that is protected from extensive overfitting and poor generalisation, whilst still utilising all available information for specifying the final model, as is recommended by TRIPOD guidelines[21]. Of course, given a larger dataset or multiple cohorts, implementation of internal validation methods within a traditional training/validation split would also be advised.

5.5.4 Solutions and conclusions

A potential solution to the variance in the observed predictive accuracy is to employ some form of resampling and internal validation methods during the model fitting stages, such as the bootstrap or cross-validation. Arguably a resampling-based approach should also be applied to any feature selection; the application of the Boruta algorithm to the whole dataset here, without any means of guarding against overfit, was sub-optimal and so this is something that should be improved upon in future works. If these experiments were to be repeated it might be considered wise to include a feature reduction step in advance of predictive modelling to remove as much variance from the data as possible. The use of a modelling approach that incorporates bootstrap resampling, such as the application of the Random Forest algorithm would be beneficial, as out-of-bag (OOB) predictions can be

obtained with relative ease, whilst still training a single model that uses all available data. An added benefit of OOB predictions is that they are as accurate as a validation dataset of equal size to the training dataset[186].

Scope-reduction, lowering one's expectations of what should be achieved with the data, represents a useful non-analytical method to improve power and produce stronger models. Rather than attempting to produce a fully validated model from limited, highly variable data, concentrating solely on the development of a **robust** model, leaving external validation to a future study with a pre-specified model. This approach would arguably be of more benefit, and where the data within the Movember GAP1 study are considered, it is not feasible to produce a truly validated model that is TRIPOD compliant[21] as all the data were collected in one study and often assayed together from a single collection site with no temporal separation, the most that could be achieved is an internally validated model[180,181].

Within the dataset used here, the Random Forest algorithm and a meta-ensemble frequently returned the highest predictive accuracies. However different algorithm and training labels work well in different scenarios, depending on the outcome being predicted. The work shown here represents a detailed empirical evaluation of how varying training labels and algorithms can cause a large change in the accuracy of developed models, which also varies considerably according to the clinical outcome being predicted in a single training and test split of the NanoString dataset. To robustly produce a model in these data, the data splitting strategy needs to be replaced by resampling, as internal validation done well is far better than an unrepresentative split of the data[186,187]. The addition of clinically available variables dramatically improves performance, though reduces the potential for biological interpretation if a study was designed with this in mind[187,188].

Considering the large effect the inclusion of simple clinical variables had on model performance, future studies would benefit from including these easily available parameters at the model fitting stage, and investigating the added benefit of integrating data from multiple sources. As discussed, multiple urine fractions were examined within the Movember GAP1 study, and where sample overlap allows, such an investigation of integrated analyses would be of great interest. This is explored in the next chapter. Combining the previously discussed strategy of scope reduction and employing strong internal validation methods using bootstrap resampling of datasets has the potential to develop prognostic models that, in theory, would require fewer and more targeted validation clinical trials. This would reduce research costs and expedite the journey of such a model to clinical adoption. In the next chapter I will describe the development and deployment of a machine learning frame that is designed to develop such models robustly and according to TRIPOD guidelines.

Chapter 6

Development of a machine learning biodiscovery framework based on bootstrap resampling and Random Forests

6.1 Summary

In this chapter I describe the development of the FrameWork, a machine learning pipeline for the robust development of TRIPOD-compliant multivariable risk prediction models and its application for integrated analyses of overlapping datasets within the Movember GAP1 study. The FrameWork employs a bootstrap resampling-based feature selection process coupled with Random Forests to produce a single pre-specified model that can be interrogated for clinical utility and aims to recognise the uncertainty inherent to prostate sampling.

The motivation and methodology for the FrameWork are discussed, followed by a flagship example of applying the FrameWork in the Movember GAP1 datasets, ExoMeth; a multivariable risk prediction model integrating data from urinary cell-free RNA expression, hypermethylation within the urinary cell-pellet and clinically available parameters. On an initial TRUS biopsy, ExoMeth accurately predicted the presence of Gleason score $\geq$ 3+4, AUC = 0.89 (95% CI: 0.84 - 0.93) and was additionally capable of detecting any cancer on biopsy, AUC = 0.91 (95% CI: 0.87 - 0.95). As ExoMeth Risk Score (range 0-1) increased, the likelihood of high-grade disease being detected on biopsy was significantly greater (OR = 2.04 per 0.1 ExoMeth increase, 95% CI: 1.78 - 2.35). Application of ExoMeth provided a net benefit over current standards of care and has the potential to reduce unnecessary biopsies by 66% when a risk threshold of 0.25 is accepted.

This work is adapted from the original publication "Development of a multivariable risk model integrating urinary cell DNA methylation & cell-free RNA data for the detection of significant prostate cancer" by Connell *et al.* published 9th March 2020 in *The Prostate*. All analysis presented in this chapter was completed by myself.

6.2 Background

The primary aim of the GAP1 initiative was to produce a multi-modal urine biomarker panel for the discrimination of disease state. However, the development of a multivariable risk model is a complex task with the validation of such a model even more so, requiring careful planning and stringent data controls to ensure validity and avoid bias or overfit. As explored in Chapters 4 & 5, achieving TRIPOD-compliant external validation of any model developed in the Movember GAP1 trial is not feasible[21]. With this considered and following the recommendations in Chapter 5, I have reduced the scope for analyses to more soundly develop robust multivariable risk prediction models that have the best possibility of later validation.

Results from two aspects of the GAP1 studies were published prior to this work, assaying differing urinary fractions; epiCaPture assessed hypermethylation of urinary cell DNA[11], and PUR quantified transcript levels in cell-free extracellular vesicle mRNA (cf-RNA) using NanoString (See Chapter 4)[8]. Both of these tests were able to discriminate some level of clinically significant disease and exhibited differing predictive characteristics; where epiCaPture was well suited to detecting the highest grade disease (Gleason score $\geq$ 8, AUC = 0.86), PUR was better matched to the deconvolution of lower risk and indolent disease, as detailed by its apparent prognostic ability in active surveillance use and detection of disease with a Gleason $\geq$ 3+4 (HR in active surveillance = 8.23, AUC = 0.76, see Chapter 4 for full details). PUR utilised LASSO-penalised cumulative link regression models, whilst EpiCaPture used a simple addition of all methylation marker values to generate risk scores, with both using conventional training/test splits of the available data as a validation strategy. Explored in-depth in Chapter 5, this is a suboptimal strategy for maximal utilisation of the data, both in terms of algorithm choice and data splitting. Instead, application of methods with strong internal validation and a resampling approach would likely show improved clinical utility for both sources of urinary biomarkers, increase robustness and avoid the variability previously observed with specific splits of data.

Where many variables are considered, such as the NanoString dataset of 167 gene-probes, or in 'omics-based assays, feature selection is a key step to isolate feature-sets that are actually of use for the outcome of interest, rather than simply random noise. However, feature selection performed poorly can have serious consequences for model performance and generalisability, resulting in overly optimistic predictions and dataset-specific feature-sets being selected[189]. Considering this, and building on the results presented in Chapter 5, in this chapter I explore the application of more robust methods through bootstrap resampling and the Boruta algorithm for feature selection, followed by model fitting using the Random Forest algorithm. Wrapped into a semi-automated pipeline for rapid prototyping of multivariable prognostic models compliant with TRIPOD guidelines[21], this process was named the FrameWork. The FrameWork is designed to robustly produce models that can be interrogated for clinical utility and recognises the uncertainty inherent to the results reported from TRUS biopsy of the prostate. The benefits of using bootstrap resampling, in both feature selection and modelling are two-fold. For feature selection, bootstrap resampling avoids the selection of features based solely on a strong link in a small subset of samples, specific to the dataset[189]. Similarly, the bagging employed by Random Forests (see Section 3.2), and the use of out-of-bag predictions, taken from decision trees that do not feature the sample in question, results in an effective validation dataset equal to the size of the training dataset[186].

With a suitable overlap in the numbers of patient samples analysed by both methods, it

was hypothesised that the original methods applied in O'Reilly *et al.* (2019) and Connell *et al.* (2019, Chapter 4) could be improved upon whilst simultaneously investigating whether data from these two assay methods could be complementary. The integration of both datasets could result in a more holistic model with predictive ability greater than the sum of its parts, able to encapsulate the clinical heterogeneity of prostate cancer and reach the levels of prognostic accuracy and clinical utility required for widespread adoption. The diagnostic accuracy of the developed models are determined by the ability to predict the presence of Gleason ≥ 7 and Gleason $\geq 4+3$ disease on biopsy, both critical distinctions, where patients with Gleason ≥ 7 are recommended radical therapy[5], whilst patients with Gleason 4+3 have significantly worse outcomes than Gleason 3+4 patients[53]. Mindful that many cancer biomarkers fail to translate to the clinic, the development of the presented model has been carried out adhering to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines[21].

6.3 Methods

6.3.1 Patient population and characteristics

Samples within the Movember GAP1 cohort (see Sections 9.2 & 3) that were analysed for both cell-pellet methylation and cf-RNA were eligible for selection for model development in this chapter ($n = 207$).

Samples from patients with metastatic disease (confirmed by a positive bone-scan or PSA >100 ng/mL, $n = 10$) were excluded, resulting in a dataset of 197 samples used for model development and named the ExoMeth cohort. The samples analysed in the ExoMeth cohort were collected from the Norfolk and Norwich University Hospital (NNUH, Norwich, UK) and St. James's Hospital (SJH, Dublin, Republic of Ireland).

6.3.2 Sample Processing and analysis

Hypermethylation at the 5'-regulatory regions of six genes (*GSTP1*, *SFRP2*, *IGFBP3*, *IGFBP7*, *APC* and *PTSG2*) in urinary cell-pellet DNA was assessed using quantitative methylation-specific PCR as described by O'Reilly *et al.* (2019).

Cell-free RNA (cf-RNA) was isolated and quantified from urinary extracellular vesicles using NanoString technology as described in Section 3.1.1. NanoString data used here were normalised according to NanoString guidelines using NanoString internal positive controls, and log_2 transformed as described in Section 3.1.1. Clinical variables considered were serum PSA, age at sample collection, DRE impression and urine volume collected.

6.3.3 Statistical Analysis

All analyses, model construction and data preparation were undertaken in R version 3.5.3[167], and unless otherwise stated, utilised base R and default parameters. All data and code required to reproduce these analyses can be found at `https://github.com/UEA-Cancer-Genetics-Lab/ExoMeth`.

Feature Selection

In total 177 variables were available for consideration (cf-RNA ($n = 167$), methylation ($n = 6$) and clinical variables ($n = 4$). For full list see Appendix Table B.1). To avoid

dataset-specific features being positively selected[189] (as may have occurred in Chapter 5), a robust feature selection workflow was developed utilising the Boruta algorithm[149] with a bootstrap resampling loop. Fully described in Chapter 3, Boruta is a Random Forest-based algorithm that iteratively compares feature importance against random predictors, named "shadow features". Features that perform significantly worse than these shadow features are consecutively dropped until only a stable feature-set remains.

The available data were resampled 1,000 times with replacement, and Boruta was applied to each resample, using the TriSig training label described in Section 5.3.3, briefly samples are categorised according to Gleason pattern into either no cancer, predominantly Gleason pattern 3 (Gleason 6 and 3+4) or predominantly Gleason pattern 4 or higher (Gleason $\geq$ 4+3). Importance measures and final decisions from Boruta were recorded at each iteration and aggregated over all 1,000 resamples. Features were only positively retained for model fitting if they were confirmed as stable and important by Boruta in $\geq 90\%$ of the resampled datasets.

Comparator Models

As shown in the previous chapter, it is entirely possible for clinical features alone to perform well for predicting a biopsy outcome, even out-performing NanoString data in many scenarios. Therefore, in order to objectively evaluate the potential clinical utility of a fully integrated model featuring methylation, cf-RNA and clinical features, additional models using subsets of the available features were trained as comparators:

- A clinical standard of care (SoC) model was trained by incorporating age, PSA, urine volume and clinician DRE impression:
- A model using only the available DNA methylation probes (Methylation, $n = 6$);
- A model only using NanoString gene-probe information (NanoString, $n = 167$).

The fully integrated ExoMeth model was trained by considering information from all of the above variables ($n = 177$). Each set of variables for comparator models were independently selected via the bootstrapped Boruta feature selection process described above to select the most optimal subset of variables possible for each predictive model. The bootstrap resamples used for each feature-set were identical and used the same random seed.

Model Construction

After feature selection, all models were trained via the random forest algorithm[142], using the *randomForest* package[144] with default parameters except for resampling without replacement and 401 trees being grown per model. Risk scores from trained models are presented as the out-of-bag predictions; the aggregated outputs from decision trees within the forest where the sample in question has not been included within the resampled dataset[142].

All models were trained using the TriSig label, modified to be treated as a continuous variable by the Random Forest algorithm. Treating this label continuously attempts to recognise that two patients with the same Gleason-scored TRUS biopsy-detected cancer may not share the exact same proportions of tumour pattern, or overall disease burden within their prostate. With larger, template-biopsied cohorts with exhaustive data collection in future, it may be possible to fit models directly to the proportion of Gleason pattern recorded. This modified continuous version of the TriSig label is solely used for model fitting and is not represented in the reporting of any clinically relevant endpoint measurements, or for determining predictive ability and clinical utility.

Statistical evaluation of model predictivity

Area Under the Receiver-Operator Characteristic curve (AUC) metrics were produced using the *pROC* package[153], with confidence intervals calculated via 1,000 stratified bootstrap resamples (See Section 3.2 for full details). Estimation plots and calculations were produced using the *dabestr* package[157] and 1,000 bootstrap resamples were used to visualise a robust effect size estimate of model predictions between risk groups.

Decision curve analysis (DCA, Section 3.2)[154] examined the potential net benefit of using ExoMeth in a clinical setting with a suitable population and is presented as standardised net benefit (sNB) calculated with the *rmda* package[190]. As presented in Chapter 4, in order to ensure the DCA results presented were representative of a more general population, the prevalence of Gleason scores within the ExoMeth cohort were adjusted via bootstrap resampling to match those observed in the control arm of the Cluster Randomised Trial of PSA Testing for Prostate Cancer (CAP) Trial[6] (as described in Connell *et al.* (2019), Chapter 4 and Chapter 3).

6.4 Results

6.4.1 The ExoMeth development cohort

Cell-pellet methylation and cell-free NanoString data were available for a total of 197 patients within the Movember GAP1 cohort, with the majority originating from the NNUH and forming the ExoMeth development cohort (Table 6.1). The proportion of significant (Gleason $\geq$ 7) disease in the ExoMeth cohort was well-balanced for modelling (49%).

Table 6.1: Characteristics of the ExoMeth development cohort

	Cancer: Cancer finding (N = 120)	Cancer: No cancer finding (N = 77)
Collection Centre:		
NNUH, n (%)	113 (94)	68 (88)
SJH, n (%)	7 (6)	9 (12)
Age:		
minimum	53.00	42.00
median (IQR)	69.50 (65.00, 76.00)	66.00 (59.00, 71.00)
mean (sd)	69.97 ± 7.44	65.70 ± 8.53
maximum	86.00	82.00
PSA:		
minimum	3.60	0.20
median (IQR)	10.05 (6.90, 18.20)	6.70 (4.20, 8.80)
mean (sd)	17.50 ± 18.82	7.44 ± 5.59
maximum	95.90	30.30
Prostate Size (DRE Estimate):		
Small, n (%)	12 (10)	14 (18)
Medium, n (%)	56 (47)	29 (38)
Large, n (%)	37 (31)	22 (29)
Unknown, n (%)	15 (12)	12 (16)
Gleason Score:		
0, n (%)	0 (0)	77 (100)
6, n (%)	24 (20)	0 (0)
3+4, n (%)	42 (35)	0 (0)
4+3, n (%)	23 (19)	0 (0)
$\geq$ 8, n (%)	31 (26)	0 (0)
Biopsy Result:		
Biopsy Positive, n (%)	120 (100)	0 (0)
Biopsy Negative, n (%)	0 (0)	53 (69)
No Biopsy, n (%)	0 (0)	24 (31)

6.4.2 Feature selection and model development

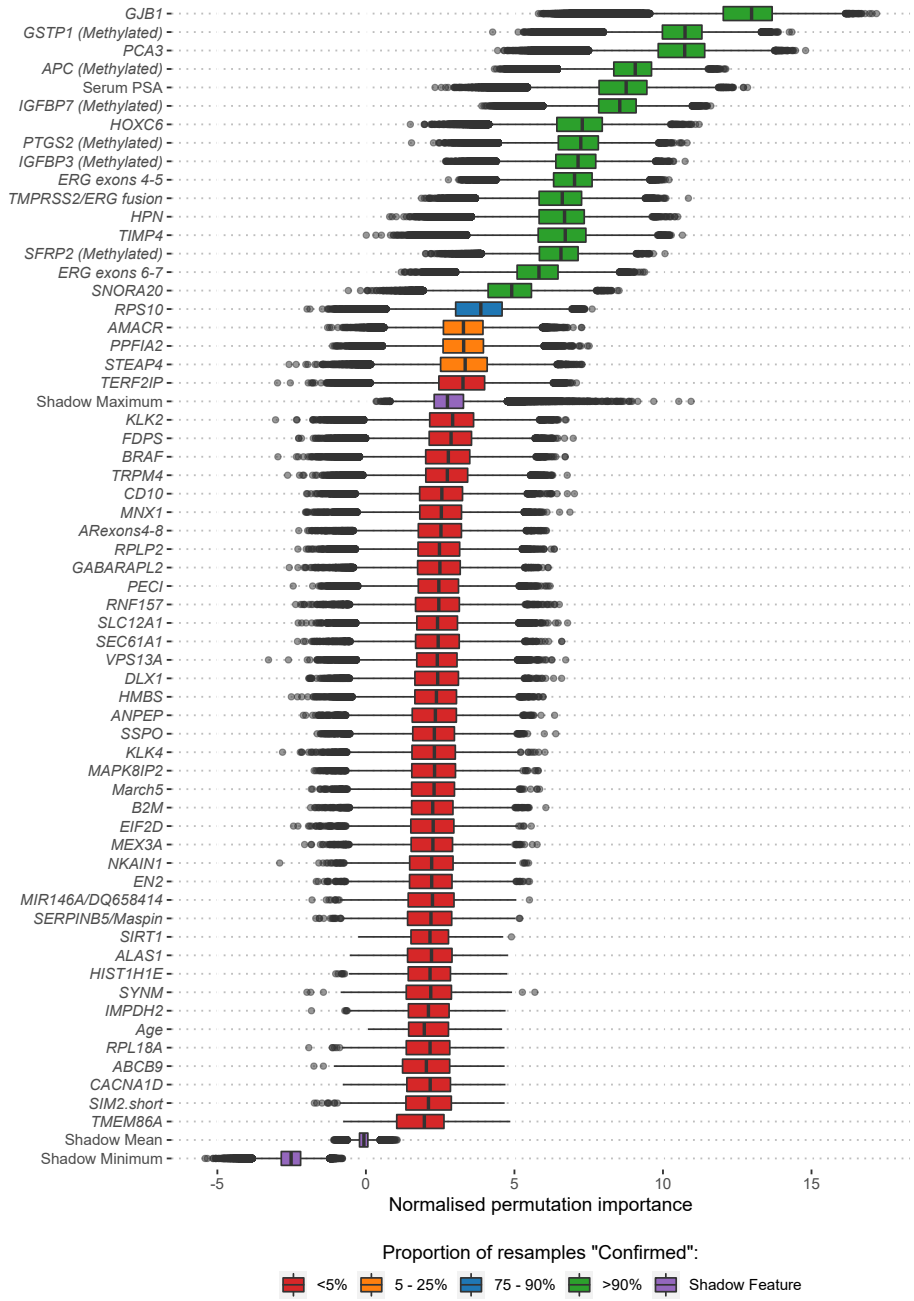


Figure 6.1: Boruta analysis of variables available for the training of the ExoMeth model. Variable importance was determined over 1,000 bootstrap resamples of the available data and the decision reached recorded at each resample. Colour indicates the proportion of the 1,000 resamples a variable was confirmed to be important in. Variables confirmed in at least 90% of resamples were selected for predictive modelling (Green). Those variables rejected in every single resample are not shown here.

Application of the bootstrap resampled Boruta portion of the FrameWork selected four feature-sets, one for each of the groups of input variables considered (Table 6.2). These feature-sets were then used as inputs to four different Random Forest comparator models trained on the continuous TriSig label based on Gleason score; a standard of care (SoC) model using only clinical information (age and PSA), a model using only methylation data (Methylation, 6 genes), a model using only cf-RNA information (ExoRNA, 12 gene-probes) and the integrated model, named ExoMeth (16 variables) (Table 6.2). The ExoMeth model is a multivariable risk prediction model incorporating clinical, methylation and cf-RNA variables. Each of the variables retained for the ExoMeth model were confirmed in every resample and notably included variables from clinical, methylation and cf-RNA sources (Figure 6.1). Full resample-derived Boruta variable importances for the SoC, Methylation and ExoRNA comparator models can be seen in Appendix Figures B.1, B.2 and B.3, respectively.

In the SoC comparator model only PSA and age were selected as important predictors, with urine volume and DRE impression not selected. All methylation probes were selected as important in both the independent Methylation model and integrated ExoMeth models (Table 6.2). 12 cf-RNA gene-probes were selected for the ExoRNA model, notably containing both variants of the *ERG* gene-probe and *TMPRSS2/ERG* fusion gene-probe, alongside *PCA3*. All variables of the ExoMeth model were also selected in each one of the comparator models.

Table 6.2: Boruta-derived features positively selected for each model. Features are selected for each model by being confirmed as important for predicting biopsy outcome, categorised as a modified ordinal variable (see Methods) by Boruta in $\geq 90\%$ of bootstrap resamples

	SoC	Methylation	ExoRNA	ExoMeth
Clinical Parameters	Serum PSA	-	-	Serum PSA
	Age	-	-	-
Methylation Targets	-	*GSTP1*	-	*GSTP1*
	-	*APC*	-	*APC*
	-	*SFRP2*	-	*SFRP2*
	-	*IGFBP3*	-	*IGFBP3*
	-	*IGFBP7*	-	*IGFBP7*
	-	*PTGS2*	-	*PTGS2*
cf-RNA Targets	-	-	*AMACR*	-
	-	-	*ERG* exons 4-5	*ERG* exons 4-5
	-	-	*ERG* exons 6-7	*ERG* exons 6-7
	-	-	*GJB1*	*GJB1*
	-	-	*HOXC6*	*HOXC6*
	-	-	*HPN*	*HPN*
	-	-	*PCA3*	*PCA3*
	-	-	*PPFIA2*	-
	-	-	*RPS10*	-
	-	-	*SNORA20*	*SNORA20*
	-	-	*TIMP4*	*TIMP4*
	-	-	*TMPRSS2/ERG* fusion	*TMPRSS2/ERG* fusion

6.4.3 ExoMeth predictive ability

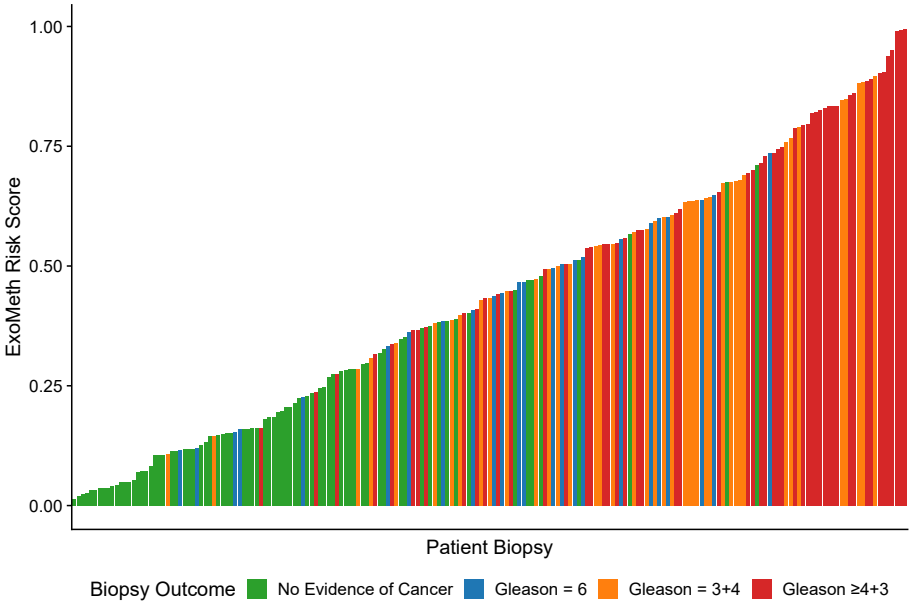


Figure 6.2: Waterfall plot of the ExoMeth risk score for each patient. Each coloured bar represents an individual patient's calculated risk score and their true biopsy outcome, coloured according to Gleason score (Gleason) . Green - No evidence of cancer, Blue – Gleason 6, Orange - Gleason 3+4, Red - Gleason $\geq$ 4+3

As ExoMeth Risk Score (range 0-1) increased, the likelihood of detecting disease with a higher Gleason score on biopsy was significantly greater (Proportional odds ratio = 2.04 per 0.1 ExoMeth increase, 95% CI: 1.78 - 2.35; ordinal logistic regression, Figure 6.2). Metastatic patients, not used in any stage of model fitting, were used as a means of assessing ExoMeth's calibration (n = 10). The median ExoMeth risk score for metastatic patients was 0.83 (IQR = 0.279), placing it above the 90th percentile of ExoMeth risk scores. One metastatic sample had a lower than expected ExoMeth score of 0.55 where no methylation target was quantified at all for this sample, perhaps reflecting a technical failure of the sample.

Table 6.3: AUC of random forest models for detecting differing outcomes on initial biopsy. Brackets show 95% confidence intervals of the AUC, calculated from 2,000 bootstrap resamples.

Initial biopsy outcome:	SoC	Methylation	ExoRNA	ExoMeth
Gleason $\geq$ 4+3:	0.75 (0.68 - 0.82)	0.77 (0.69 - 0.85)	0.74 (0.67 - 0.82)	0.81 (0.74 - 0.87)
Gleason $\geq$ 3+4:	0.73 (0.66 - 0.80)	0.78 (0.71 - 0.84)	0.81 (0.75 - 0.87)	0.89 (0.84 - 0.92)
Any Cancer	0.70 (0.62 - 0.77)	0.73 (0.65 - 0.80)	0.86 (0.81 - 0.91)	0.91 (0.87 - 0.95)

ExoMeth was superior to all other models, returning an AUC for the prediction of

Gleason $\geq$ 3+4 = 0.89 (95% CI: 0.84 - 0.93) and 0.81 (95% CI: 0.75 - 0.87) when prediction of Gleason $\geq$ 4+3 was considered (Table 6.3).

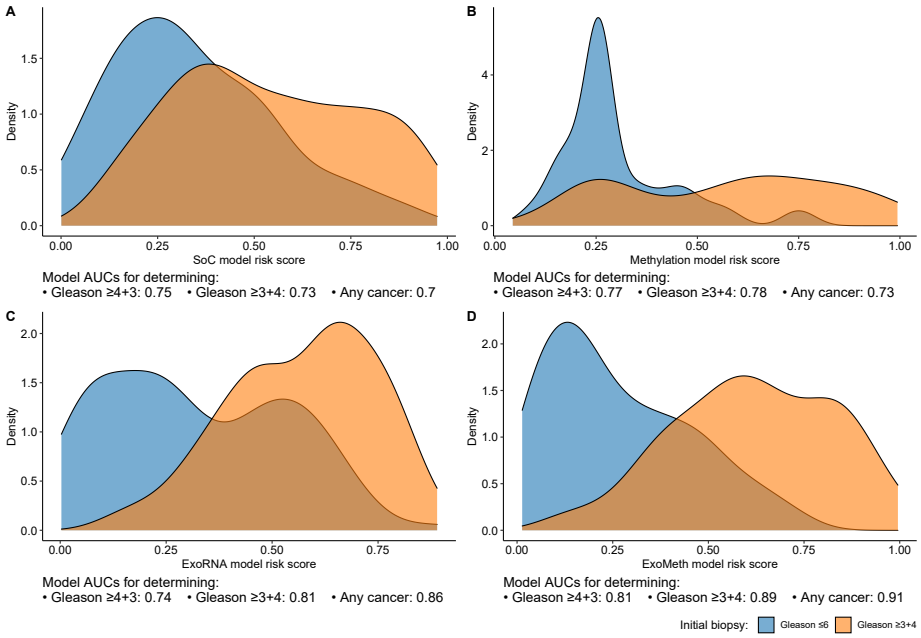


Figure 6.3: Density plots detailing risk score distributions generated from four trained models. Models A to D were trained with different input variables; A - SoC clinical risk model, including Age and PSA, B - Methylation model, C -ExoRNA model and D - ExoMeth model, combining the predictors from all three previous models. The full list of variables in each model is available in Table 6.1. Fill colour shows the risk score distribution of patients with a significant biopsy outcome of Gleason $\geq$ 3+4 (Orange) or Gleason $\leq$ 6 (Blue).

Density plots were used to explore the distribution of risk scores for models in more detail than the AUC alone. These plots showed that ExoMeth achieved a better discrimination of Gleason $\geq$ 3+4 disease from less clinically significant outcomes when compared to any of the other models (ExoMeth AUC all $P < 0.01$ bootstrap test, 1,000 resamples, Figure 6.3). The SoC model, whilst returning respectable AUCs, would misclassify more patients with indolent disease as warranting further investigation than all other models (Figure 6.4A), broadly representative of what is observed clinically. For example, to classify 90% of patients with Gleason = 7 disease correctly would require an SoC risk score of 0.237, which would misclassify 65% of patients with less significant disease.

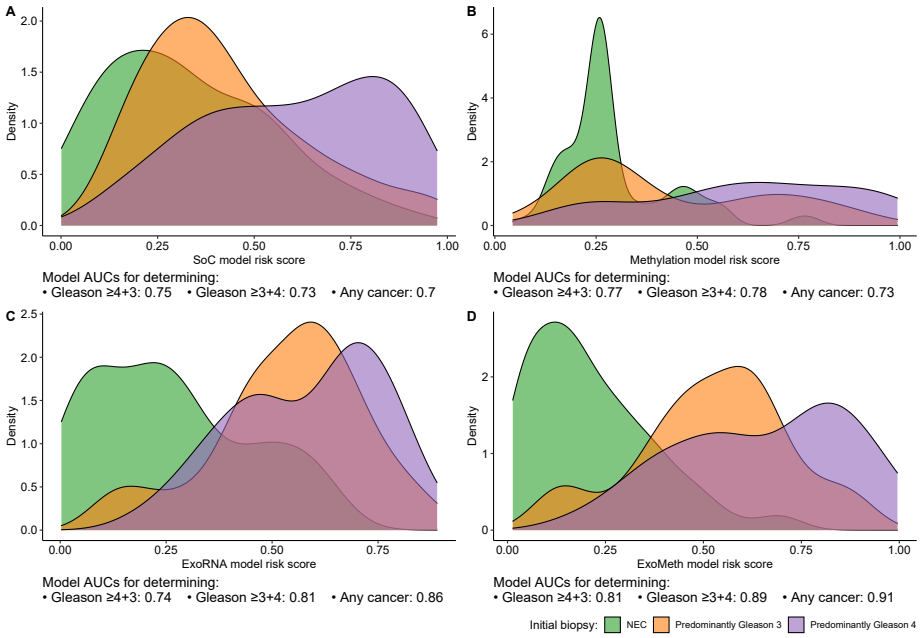


Figure 6.4: Density plots detailing risk score distributions generated from four trained models. Models A to D were trained with different input variables; A - SoC clinical risk model, including Age and PSA, B - Methylation model, C -ExoRNA model and D - ExoMeth model, combining the predictors from all three previous models. The full list of variables in each model is available in Table 6.1. Fill colour shows the risk score distribution of patients with with respect to biopsy outcome: No evidence of cancer (Blue), Gleason = 6 or 3+4 (Orange), Gleason $\geq$ 4+3 (Green)

This discriminatory ability of the ExoMeth model over all comparators was highlighted further when biopsy outcomes are considered as biopsy negative, Gleason 6 or 3+4, or Gleason $\geq$ 4+3 (Figure 6.4). The Methylation comparator model improved upon SoC, by drawing the risk distribution of Gleason $\leq$ 6 patients into a far more pronounced peak but displaying poor discrimination of higher risk patients, with a bimodal distribution of risk score where almost 50% of patients with Gleason $\geq$ 3+4 have risk scores equal to benign patients (Figure 6.4B). The opposite occurred in the NanoString comparator model, which exhibited a broad bimodal distribution for lower-risk patients that would see them unnecessarily subjected to biopsy (Figure 6.4C).

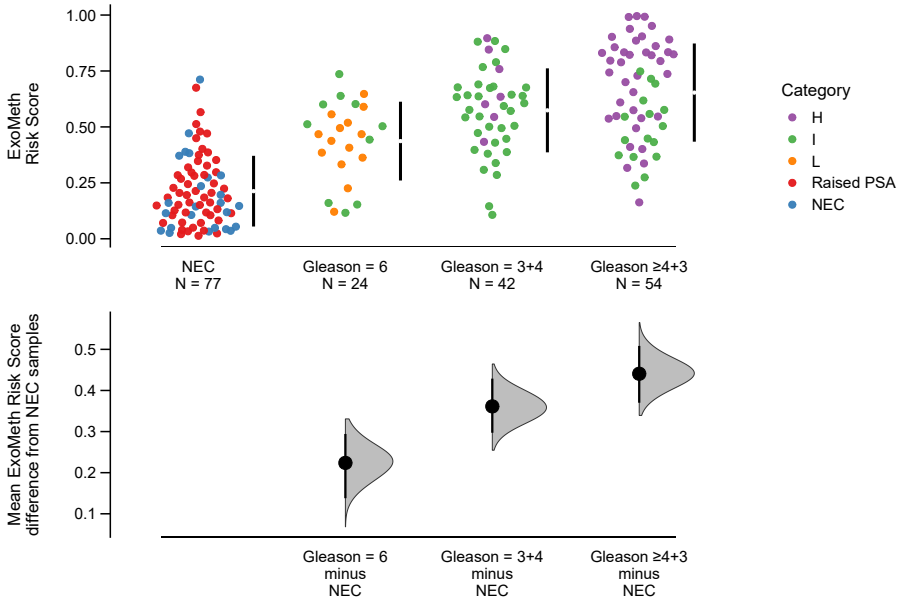


Figure 6.5: Estimation plot of the ExoMeth risk score The top row details individual patients as points, separated according to Gleason score on the x-axis and risk score on the y-axis. Points are coloured according to clinical risk category; NEC - No evidence of cancer, Raised PSA - Raised PSA with negative biopsy, L -D'Amico Low-Risk, I - D'Amico Intermediate Risk, H - D'Amico High-Risk. Gapped vertical lines detail the mean and standard deviation of risk scores for each group. The lower panel shows the mean differences in risk score of each group, as compared to the NEC samples. Mean differences and 95% confidence interval are displayed as a point estimate and vertical bar respectively, using the sample density distributions calculated from a bias-corrected and accelerated bootstrap analysis from 1,000 resamples.

Resampling of ExoMeth predictions via estimation plots allowed for comparisons of mean ExoMeth differences between clinical groups (1,000 bias-corrected and accelerated bootstrap resamples, Figure 6.5). Mean ExoMeth differences relative to patients with no evidence of cancer and different grades of cancer finding on biopsy were: Gleason 6 = 0.22 (95% CI: 0.14 – 0.30), Gleason 3+4 = 0.36 (95% CI: 0.28 – 0.42) and Gleason $\geq$ 4+3 = 0.44 (95% CI: 0.37 – 0.51) (Figure 6.5).

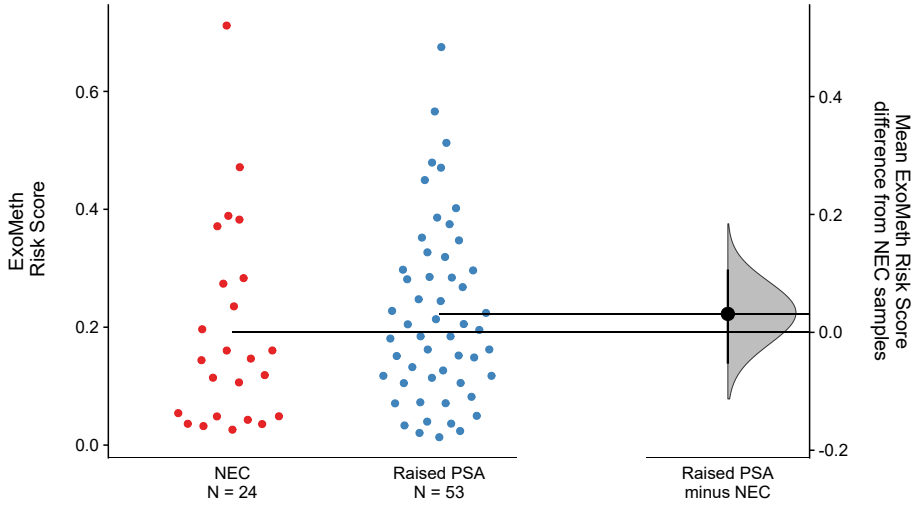


Figure 6.6: Estimation plot of the ExoMeth risk score in No evidence of cancer (NEC) and raised PSA, negative biopsy samples. The left panel details individual patients as points with ExoMeth risk score on the y-axis. Points are coloured according to clinical risk category; NEC - No evidence of cancer, Raised PSA - Raised PSA with negative biopsy. The right panel shows the distribution of the mean bootstrapped differences in risk score between NEC and Raised PSA samples. The horizontal lines show the mean difference of ExoMeth risk score relative to the NEC category. Mean difference and 95% confidence interval are displayed as a point estimate and vertical bar respectively, using the sample density distributions calculated from a bias-corrected and accelerated bootstrap analysis from 1,000 resamples.

Notably no significant differences were observed in ExoMeth risk score between patients with a raised PSA but negative for cancer on biopsy and patients with no evidence of cancer (mean difference = 0.03 (95% CI: 0.05 – -0.10), Figure 6.6).

6.4.4 Net Benefit of ExoMeth

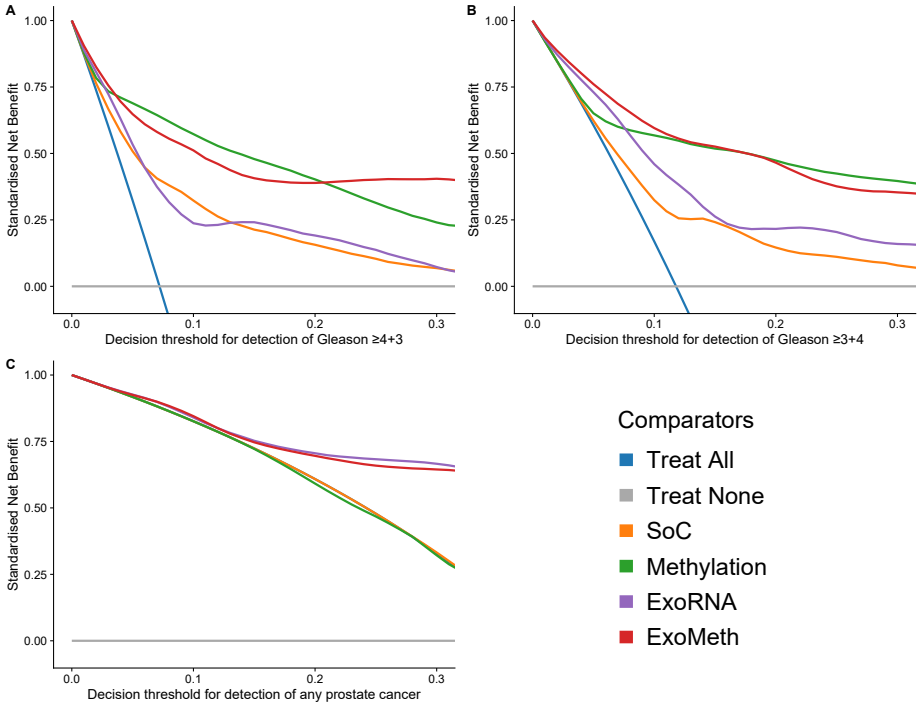


Figure 6.7: Decision curve analysis (DCA) plots detailing the standardised net benefit (sNB) of adopting different risk models for aiding the decision to biopsy patients who present with a PSA $\geq$ 4 ng/mL. The x-axis details the range of risk a clinician or patient may accept before deciding to biopsy. Panels show the sNB based upon the detection of varying levels of disease severity: A - detection of Gleason $\geq$ 4+3, B - detection of Gleason $\geq$ 3+4, C - any cancer; Blue- biopsy all patients with a PSA >4 ng/mL, Orange - biopsy patients according to the SoC model, Green - biopsy patients based on the methylation model, Purple - biopsy patients based on the NanoString model, Red - biopsy patients based on a the ExoMeth model. To assess the benefit of adopting these risk models in a non-PSA screened population we used data available from the control arm of the CAP study. DCA curves were calculated from 1,000 bootstrap resamples of the available data to match the distribution of disease reported in the CAP trial population. Mean sNB from these resampled DCA results are plotted here. See Methods for full details.

Decision curve analysis was used to examine the net benefit of adopting ExoMeth within a population of patients suspected to have prostate cancer, and with a PSA level suitable to trigger a diagnostic biopsy by NICE guidelines (PSA $\geq$ 4 ng/mL). The biopsy of patients based upon their ExoMeth risk score consistently provided a net benefit over current standards of care (represented by the SoC model) across all decision thresholds examined and was the most consistent amongst all comparator models across a range of biopsy endpoints considered clinically relevant (Figure 6.7). Of the patients with Gleason $\geq$ 7 disease, 95% had an ExoMeth risk score $\geq$ 0.283.

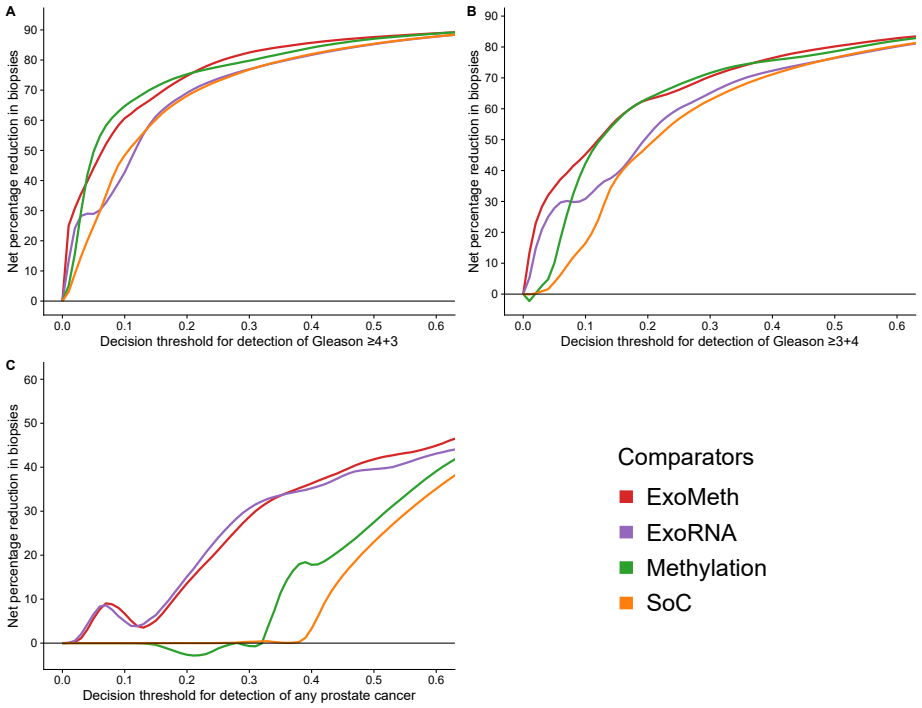


Figure 6.8: Net percentage reduction in biopsies, as calculated by DCA measuring the benfit of adopting different risk models for aiding the decision to biopsy patients who who would otherwise undergo biopsy by current clinical guidelines. The x-axis details the range of accepted risk a clinician or patient may accept before deciding to biopsy. Panels show the reduction in biopsies per 100 patients based upon the detection of varying levels of disease severity: A - detection of Gleason $\geq$ 4+3, B - detection of Gleason $\geq$ 3+4 and C - any cancer. Coloured lines show differing comparator models; Orange - biopsy patients according the to the SoC model, Green - biopsy patients based on the methylation model, Purple - biopsy patients based on the ExoRNA model, Red - biopsy patients based on a the ExoMeth model. To assess the benefit of adopting these risk models in a non-PSA screened population we used data available from the control arm of the CAP study. DCA curves were calculated from 1,000 bootstrap resamples of the available data to match the distribution of disease reported in the CAP trial population. Mean sNB from these resampled DCA results are used to calculate the potentially reductions in biopsy rates here. See Methods for full details.

At a decision threshold of 0.25 (accepting a 1 in 4 chance of an outcome before accepting biopsy), ExoMeth could result in up to 66% fewer unnecessary biopsies of patients presenting with a suspicion of prostate cancer, without missing substantial numbers of patients with aggressive disease, whilst if Gleason $\geq$ 4+3 were considered the threshold of clinical significance, the same decision threshold of 0.25 could save 79% of patients from receiving an unnecessary biopsy (Figure 6.8).

6.5 Discussion

The accurate discrimination of disease state in men prior to a confirmatory diagnostic biopsy would mark a significant development and impact large numbers of men suspected of harbouring prostate cancer. Currently up to 75% of men with a raised PSA ($\geq$ 4 ng/mL) are negative for prostate cancer on biopsy[5–7], which has led to a concentration of research efforts to address this problem using non-invasive methods. Several such biomarkers capable of detecting Gleason $\geq$ 3+4 disease using urine samples have been reported with accuracy superior to current clinical methods, including PUR in Chapter 4[8,12,99,102]. However, in each of these examples, only a single assay method or aspect of prostate cancer pathobiology is considered. With the molecular heterogeneity of prostate cancer considered[191], a more holistic approach to appraise disease status is necessary.

Published data has shown that urine can contain a wealth of useful cancer biomarkers within RNA, DNA, cell-free DNA, DNA methylation and proteins[8,11,192–194]. However, the analyses presented here were at the time of publication, the first attempt to integrate information from multiple biomarkers collected within the same sample for the detection of prostate cancer prior to biopsy. A combination of miRNA and methylation markers in tissue samples has been recently reported to predict biochemical recurrence following radical prostatectomy (HR = 1.35, 95% CI: 1.06–1.73)[195]. I have shown here that an improved prognostic marker can be produced by harnessing the information derived from different molecular entities in urine and clinically available parameters for patients suspected to have prostate cancer. The ExoMeth model integrates NanoString quantified cf-RNA data with hypermethylation data from six previously identified genes[11] and serum PSA levels.

A practical consideration in the case of deploying a urine test based on ExoMeth is the requirement of more than one assay, raising material costs and complexity. The added complexity increases the likelihood of technical failures that cause false positives or negatives, as observed with one of the assayed metastatic samples. Such a problem could potentially be solved by a repeated assay, but would require stringent quality control procedures. Of course, due to the non-invasive nature of liquid biopsy, repeated urine samples are not a health concern like a repeated biopsy may be. Cost increases can be mitigated if the predictive utility of ExoMeth is upheld in validation and health economic studies, where the potential reduction in biopsies and/or mpMRI scans would likely fair outweigh the cost of a urine test.

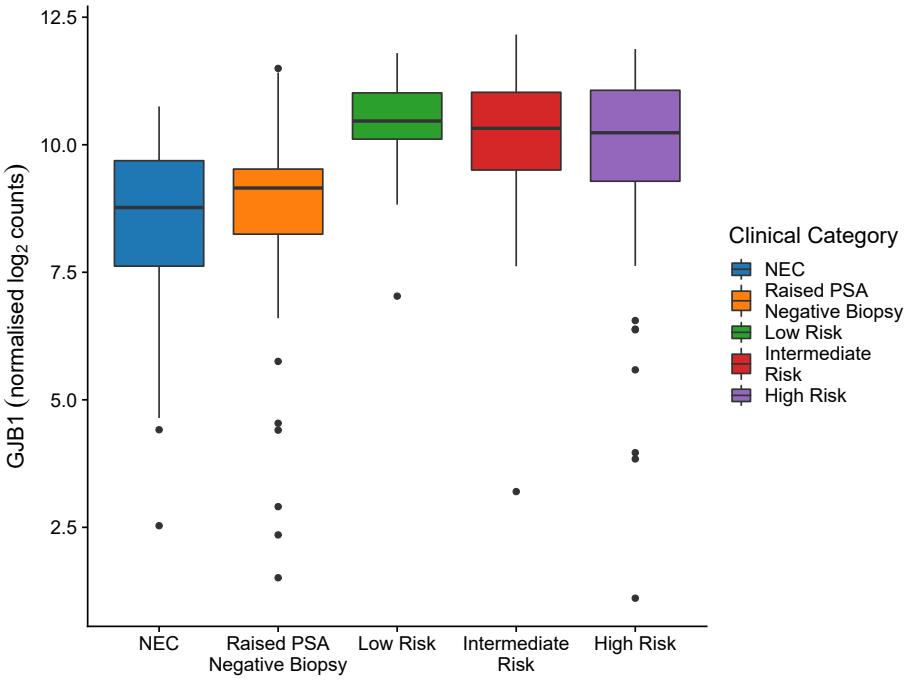


Figure 6.9: Expression *GJB1* cf-RNA levels in the ExoMeth cohort, relative to clinical risk category.

Features selected for ExoMeth included gene-probes well known to be associated with prostate cancer and proven in other diagnostic tests, such as *HOXC6*[99], *PCA3*[12,164] and the *TMPRSS2/ERG* gene fusion[164]. ExoMeth additionally incorporated the *GJB1* cf-RNA gene-probe as the most important variable for predicting biopsy outcome (Figure 6.1). Whilst *GJB1* has been reported to be a prognostic marker for favourable outcomes in renal cancer, there is no evidence of its use as a prognostic biomarker in prostate cancer to date[196,197]. Interestingly, *GJB1* was not selected in the PUR modelling process, where it's expression appears to be binary as a cancer versus no cancer marker rather than monotonically increasing with risk category (Figure 6.9).

ExoMeth was able to correctly predict the presence of significant prostate cancer on biopsy with an AUC of 0.89, representing a significant uplift when compared to other published tests (AUCs for Gleason ≥ 7 : PUR = 0.77[8], ProCUrE = 0.73[192], ExoDX Prostate IntelliScore = 0.77[102], SelectMDX = 0.78[99], epiCaPture Gleason $\geq$ 4+3 AUC = 0.73[11]). Furthermore, ExoMeth resulted in accurate predictions even where serum PSA levels alone were inaccurate: these are patients biopsied likely due to their raised PSA, but ultimately no prostate cancer was found. These patients ExoMeth scores were largely no different to NEC patients (Figure 6.6), meaning they could have avoided this invasive procedure, whilst those diagnosed with cancer were still accurately stratified by ExoMeth (Figure 6.4). Of the three patients with no evidence of cancer on biopsy with an ExoMeth risk score >0.55, two were positive for the *TMPRSS2/ERG* fusion transcript in NanoString analyses (data not shown), implying that a tumour may have been missed and a re-biopsy of these patients may be necessary, as *TMPRSS2/ERG* is not seen in benign tissue[198].

Adoption of ExoMeth as a triage-style tool into clinical pathways could see a large amount of patients removed from the clinical pathway much earlier than current standards allow for. Not only would this save healthcare systems money, it would save patients from

undue stress and worry of investigations for prostate cancer. Putting 66% fewer patients forward for biopsy, or mpMRI scans if ExoMeth is validated against mpMRI, would free huge amounts of resources for clinicians and tertiary care systems alike. The ease of sampling of patients has an additional benefit in that repeated sampling and collection outside of a hospital environment is feasible, such as in a primary care setting. ExoMeth, if successfully validated in larger external cohorts presents a very real opportunity for considerable changes to be made in how patients suspected to have prostate cancer are initially risk assessed.

The results presented here show that through careful consideration of statistical methodology, a predictive risk model can be successfully and robustly developed to minimise the potential for overfitting and bias, even within a relatively small dataset. Extensive application of bootstrap resampling and out-of-bag predictions ensures that whilst the ExoMeth development cohort comprises of only 197 samples, the effective dataset size is doubled[186]. ExoMeth nonetheless requires validation in an independent external cohort, compliant to TRIPOD guidelines[21] before its use a clinical risk model can be considered. The design of such a study is presented in Chapter 8, including considerations of updated clinical guidelines to evaluate the clinical utility of supplementing mpMRI with ExoMeth. For many men harbouring indolent prostate cancer, ExoMeth could greatly impact their experience of prostate cancer care when compared to current clinical pathways.

The FrameWork shows great promise in the ability to rapidly develop models that can be pre-specified and used in further studies with minimal further work. In the next chapter I shall discuss further applications of the FrameWork to additional overlapping datasets within the Movember GAP1 study, showcasing the successful development of an additional two multivariable risk models. It is important to highlight limitations as well as successes, and with this considered, examples of unsuccessful attempts at model development are also presented.

Chapter 7

Successes and Failures of the FrameWork

7.1 Summary

The work in this chapter describes analysis of additional datasets in the GAP1 cohort and application of the FrameWork, described in the previous chapter. Two additions to the "Exo-X" series of multivariable risk prediction models are developed, in the form of ExoGrail and ExoSpec. Negative results from an ELISA panel completed on a large portion of the GAP1 cohort is described, showing that not all previously identified biomarker results can be replicated successfully. The main results of this chapter are split into self-contained sections for each application of the FrameWork that encompass the methods, results, and conclusions, with a final discussion considering overarching implications for validation and further development.

ELISA data quantifying twelve different proteins from multiple collection sites were available for a large number of samples in the GAP1 cohort ($n = 471$). Initial analyses of these data revealed no gain in predictive ability over clinical standards of care. Integrated analyses with NanoString data ($n = 237$) were additionally attempted and analysis of variables by Boruta showed only the Engrailed-2 (EN2) protein to be of predictive use. It was decided to disregard the other ELISA proteins, and concentrate on only EN2 and NanoString data moving forwards.

ExoGrail represents another successful outputs from the application of the FrameWork, developed by the integration of NanoString data with the singular EN2 protein biomarker. ExoGrail was able to predict a biopsy outcome with accuracy exceeding standards of care, with an AUC = 0.89 (95% CI: 0.85 – 0.94) when predicting the presence of any cancer and discriminating more aggressive Gleason $\geq$ 3+4 disease returning an AUC = 0.84 (95% CI: 0.78 – 0.89). Development of the ExoGrail model showed that improvements in how patients are risk assessed can be found from the relatively simple integration of data from a single protein marker during the model fitting process.

ExoSpec was developed using NanoString data and the high-dimensionality proteomic dataset, featuring over 30,000 variables and requiring modifications to be made to the FrameWork to function. The resampled Boruta feature selection portion is replaced with a LASSO-penalisation to reduce computational intensity without too much of a reduction in the robustness of the entire pipeline. The final ExoSpec model incorporated 4 cf-RNA transcripts, 6 peptides and 2 clinically available parameters, and accurately predicted an

outcome of Gleason $\geq$ 3+4 with an AUC = 0.83 (95% CI: 0.77 - 0.88). ExoSpec requires an additional, more targeted study prior to consideration of a full validation study.

The ExoGrail and ExoSpec results are adapted from two original publications, under review at the time of submission; "Development of a multivariable risk model integrating urinary proteomic and cell-free RNA data to detect significant prostate cancer" by Connell *et al.* under review at the *British Journal of Cancer* and "Integration of urinary EN2 protein & cell-free RNA data in the development of a multivariable risk model for the detection of prostate cancer prior to biopsy" by Connell *et al.*, submitted to *eBioMed*. All analysis presented in this chapter was completed by myself.

7.2 Background

Given the successful application of the FrameWork in Chapter 6 during the development of the ExoMeth model[111], it was hypothesised that similar prognostic potential may be uncovered in other available overlaps of GAP1 data. There are additional intersectional datasets of an appropriate size for developing prognostic models ($n \approx 200$), with multiple experiments performed on the same samples. The sources for these data were: cf-RNA quantified by NanoString, urinary levels of 12 proteins from ELISA data, and proteomics captured by capillary-electrophoresis mass spectrometry (CE-MS). Notably, the analyses of overlapping data were always between NanoString and one other data source: NanoString and CE-MS data, NanoString and EN2 ELISA data, and NanoString and ELISA-panel datasets. These three intersectional datasets represent differing approaches to biodiscovery and biomarker development studies, despite all being from the same, much larger GAP1 study.

Targeted analyses of previously identified biomarkers make an attractive prospect, they are a known quantity with a history of successful results relating to pathobiology or prognostication, such as the Engrailed-2 protein. EN2 is a transcription factor with an essential function in early development and has been shown to be secreted from cells, and taken up by others[199,200]. EN2 is a proven biomarker of its own, having shown to have diagnostic utility for both predicting biopsy outcome and tumour volume in radical prostatectomy patients[128,201]. Quantified by a simple ELISA, integration of EN2 and cf-RNA NanoString data are trivial, with subsequent application of the FrameWork completed with relative ease given the highly targeted nature of the data. In the wider ELISA dataset were a further 11 proteins, including the Kalikrein (KLK) family that are highly expressed in prostate tissues[125,202], and proteins thought to be useful for normalising urine volume and concentration, such as creatinine, typically used for assessing kidney function and muscle damage[203]. These ELISA data represent a semi-focused approach, with pre-existing hypotheses of biological function but lacking in strong prior data for their use as prognostic biomarkers.

At the other end of study complexity in biodiscovery are whole 'omics discovery studies, such as the CE-MS peptidomics dataset presented here. Representing the entirety of a singular biological aspect, they are hugely multidimensional, often in the tens of thousands of eligible predictors. The untargeted nature of these data make integration much less trivial, requiring careful handling to extract useful information. To reduce false discovery rates and overfitting by normal feature selection methods, *a priori* filtering is one of the strongest tools[204]; removing the less informative features with simple rules and conditions rather than statistical tests. Nonetheless, even following *a priori* filtering, a large featureset

can still remain and require large amounts of compute power for statistical analysis.

The main purpose of these analyses was to investigate whether the FrameWork truly represents a "one-size fits all" solution, regardless of the approach to biodiscovery. Or, if more tailored approaches are required for maximal utilisation of data, assessing whether the FrameWork be adapted appropriately. Three self-contained sections are presented below, each dealing with the featuresets described above to avoid confusion of methodological alterations and consideration of multiple outputs at once.

7.3 Analysis of ELISA data reveals little clinical utility

7.3.1 Methods

Patient cohort and characteristics

Samples within the Movember GAP1 cohort (see Sections 9.2 & 3) that were analysed for whole-urine protein levels by ELISA ($n = 471$) were eligible for the ELISA cohort (Table 7.1). Samples with both ELISA and cf-RNA transcript level data available were eligible for the ExoLISA development cohort and used for model development ($n = 237$, Table 7.3).

Sample Processing

Cell-free RNA (cf-RNA) was isolated and quantified from urinary extracellular vesicles using NanoString technology as described in Section 3.1.1. All NanoString data presented here were normalised according to NanoString guidelines using NanoString internal positive controls, and log_2 transformed as described in Section 3.1.1.

Urinary protein levels were quantified by sandwich ELISA using monoclonal antibodies to; MSMB, GDF15, CD10, Creatinine, KLK2, KLK4, KLK7, KLK11, and EN2. Clinical variables considered were serum PSA and age at sample collection.

Feature Selection

Only 14 variables in total were considered in the ELISA cohort and feature selection was not deemed necessary.

A total of 181 variables were available within the ExoLISA development cohort for prediction (cf-RNA ($n = 167$), the clinical variables of patient age and serum PSA ($n = 2$), and ELISA quantified proteins ($n = 12$). Feature selection during the analysis of the ExoLISA cohort followed the same FrameWork process as described in Chapter 6. Briefly, the Boruta algorithm was applied to 1,000 resamples of the ExoLISA development cohort with replacement. Features were only positively retained for model fitting if selected by Boruta in $\geq 90\%$ of resampled Boruta runs.

ELISA Model Construction and evaluation:

For the ELISA cohort analysis, three Random Forest-based models were trained using subsets of the available variables across the patient population, similarly to those models described in Chapter 6. A clinical standard of care (SoC) model was trained by incorporating only patient age and PSA; a model using only the ELISA values (ELISA, $n = 12$), and an ELSoC model was trained by incorporating information from both ELISA-derived variables and clinically available parameters ($n = 14$).

All models were trained via the random forest algorithm[142], using the *randomForest* package[144] with the same parameters and process described in Chapter 6. Risk scores from trained models are presented as the out-of-bag predictions, as described in Section 3.2.

Random Forest models were fit using the TriSig label, modified to be treated continuously as in Chapter 6. Area Under the Receiver-Operator Characteristic curve (AUC) metrics were produced using the package[153], with confidence intervals calculated via 1,000 stratified bootstrap resamples to initially assess potential predictive utility.

7.3.2 Results

ELISA cohort and preliminary analysis:

A total of 12 proteins were quantified by ELISA, with data available for a total of 471 samples, making it the second largest dataset within the Movember GAP1 study after the NanoString dataset (Table 7.1).

Table 7.1: Characteristics of the ELISA cohort

	Cancer finding	No cancer finding
Collection Centre:		
Atlanta, n (%)	33 (12)	9 (4)
NNUH, n (%)	156 (58)	154 (75)
Toronto, n (%)	78 (29)	41 (20)
Age:		
minimum	47.00	37.00
median (IQR)	67.00 (61.00, 72.00)	66.00 (60.00, 71.00)
mean (sd)	67.03 ± 8.06	65.34 ± 8.15
maximum	91.00	85.00
PSA:		
minimum	0.80	0.01
median (IQR)	8.20 (5.69, 13.25)	5.93 (3.69, 8.28)
mean (sd)	12.90 ± 14.29	6.58 ± 4.64
maximum	95.90	30.30
Prostate Size\(DRE Estimate):		
Small, n (%)	21 (8)	27 (13)
Medium, n (%)	87 (33)	65 (32)
Large, n (%)	65 (24)	54 (27)
Unknown, n (%)	93 (35)	57 (28)
Gleason Score:		
0, n (%)	2 (1)	203 (100)
6, n (%)	90 (34)	0 (0)
3+4, n (%)	81 (30)	0 (0)
4+3, n (%)	46 (17)	1 (0)
$\geq$ 8, n (%)	48 (18)	0 (0)
Biopsy Result:		
Biopsy Positive, n (%)	267 (100)	0 (0)
Biopsy Negative, n (%)	0 (0)	136 (67)
No Biopsy, n (%)	0 (0)	68 (33)

Random Forest based models trained using the available ELISA data showed that ELISA model had very little additional predictive utility above the clinical standard of care model that uses only serum PSA levels and patient age information to predict outcome (all $P > 0.05$ by bootstrap test with 1,000 resamples, Table 7.2). The ELSoC model performed significantly better than both the SoC and ELISA models for predicting all biopsy outcomes (all $P < 0.001$ by bootstrap test with 1,000 resamples), though returning AUCs below those of the other established urine tests discussed in section 2.6.2. With this considered, further development of an ELISA-based model was ceased, and the potential utility of integrating cf-RNA data was investigated through the ExoLISA cohort.

Table 7.2: AUC from Random Forest models trained using: only clinical variables (SoC), peptide data (ELISA), or both ELISA and clinical data (ELSoC) for detecting different biopsy outcomes. Brackets show 95% confidence intervals of the AUC, calculated over 1,000 bootstrap resamples

Initial biopsy outcome:	SoC	ELISA	ELSoC
Gleason $\geq$ 4+3:	0.70 (0.63 - 0.76)	0.71 (0.65 - 0.77)	0.78 (0.73 - 0.83)
Gleason $\geq$ 3+4:	0.67 (0.62 - 0.72)	0.69 (0.64 - 0.74)	0.74 (0.69 - 0.79)
Any Cancer	0.67 (0.62 - 0.72)	0.68 (0.63 - 0.73)	0.71 (0.66 - 0.76)

The ExoLISA cohort and important features

A total of 204 samples with ELISA and cf-RNA data were available, collected from the NNUH ($n = 173$) and urology clinics in Atlanta, USA ($n = 31$). These samples formed the ExoLISA development cohort, and used for integrated analyses, as described above (Table 7.3.

Table 7.3: Characteristics of the ExoLISA cohort

	Cancer finding	No cancer finding
Collection Centre:		
Atlanta, n (%)	27 (20)	4 (6)
NNUH, n (%)	110 (80)	63 (94)
Age:		
minimum	47.00	48.00
median (IQR)	68.00 (62.00, 75.00)	66.00 (59.00, 71.50)
mean (sd)	68.34 ± 8.14	65.33 ± 8.30
maximum	91.00	82.00
PSA:		
minimum	0.80	0.30
median (IQR)	8.90 (6.20, 14.00)	5.50 (2.65, 8.15)
mean (sd)	14.18 ± 16.01	6.58 ± 5.77
maximum	95.90	30.30
Gleason Score:		
0, n (%)	0 (0)	67 (100)
6, n (%)	36 (26)	0 (0)
3+4, n (%)	51 (37)	0 (0)
4+3, n (%)	24 (18)	0 (0)
$\geq$ 8, n (%)	26 (19)	0 (0)
Biopsy Result:		
Biopsy Positive, n (%)	137 (100)	0 (0)
Biopsy Negative, n (%)	0 (0)	40 (60)
No Biopsy, n (%)	0 (0)	27 (40)

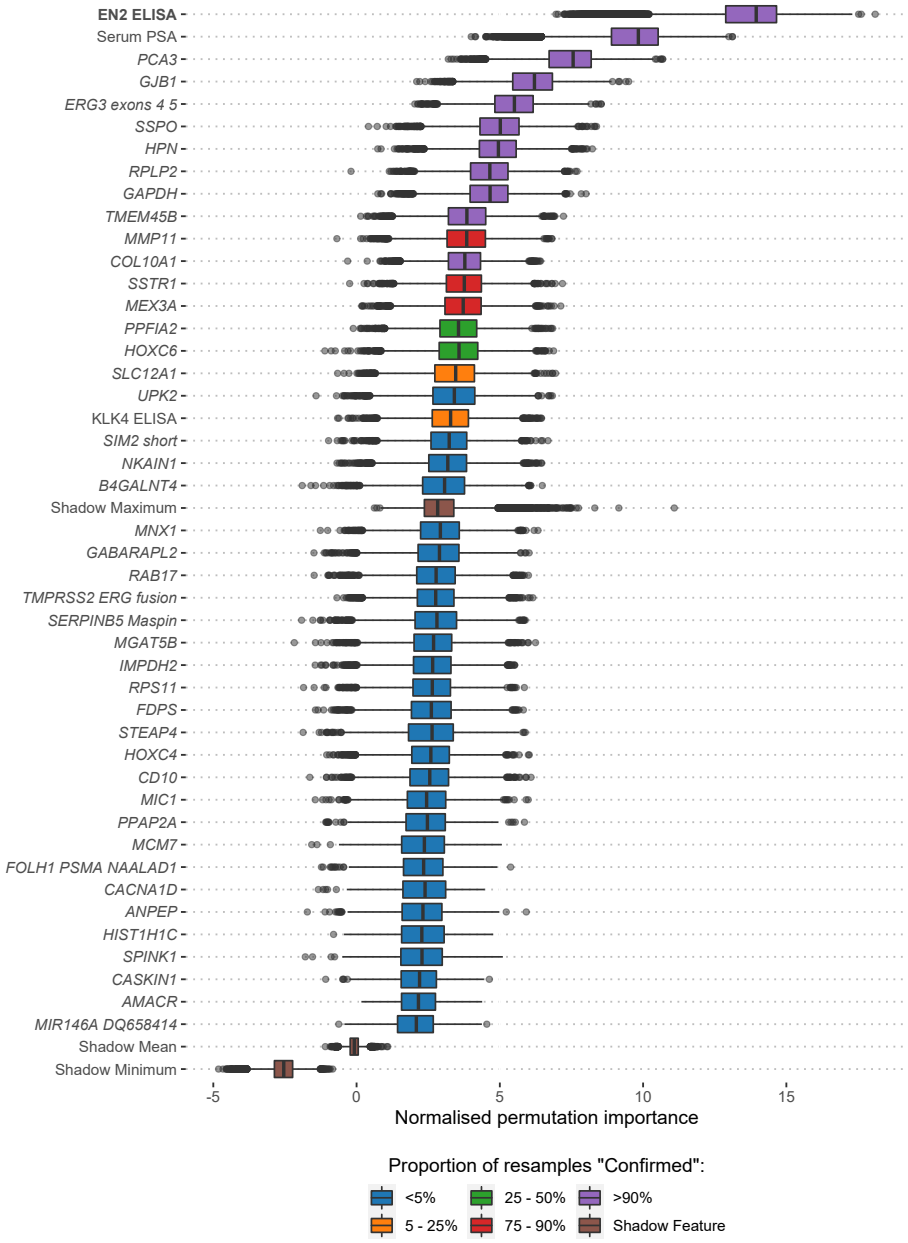


Figure 7.1: Boruta analysis of the ExoLISA cohort, using all available variables. 1,000 resamples with replacement of the available data were made, with the normalised permutation importance of each variable recorded at each iteration, along with the decision of Boruta within that resample. Fill colour shows the proportion of resamples that a feature was positively retained by Boruta. Those features selected in $\geq$ 90% of resamples were selected for fitting predictive models. Variables rejected in all of the 1,000 resamples are not shown here

Boruta analysis of the available predictors showed that the Engrailed-2 (EN2) protein was the only protein to be confirmed in more than 25% of resampled Boruta runs (Figure 7.1). KLK4 was the only other protein marker to not be rejected in every single resample,

showing the low predictive usefulness of the ELISA-derived data. EN2 was additionally selected as the singular most important variable for predicting biopsy outcome with almost triple the normalised permutation importance of the *PCA3* cf-RNA gene-probe (Figure 7.1). Considering this information, it was decided again to cease model development, and instead concentrate on integrated analysis of only available EN2 and cf-RNA data, in an effort to reduce the number of variables to consider, increase the sample size, and the power of any subsequent models developed.

7.3.3 Conclusions

Initial analyses of the large ELISA cohort revealed little predictive utility was to be found in the available protein biomarkers above clinical standards of care. When the kalikrein (KLK) proteins are considered, this may have been somewhat expected. Whilst KLK2 and KLK4 are highly specific to prostate tissue[125,126], KLK6 has been associated primarily with neuroplasticity in the central nervous system[205], and KLK7 is highly expressed in the epidermis of the skin[206]. KLK6, KLK7 and KLK11 have previously identified as prognostic markers of disease status, but this remains to be validated in larger cohorts with the explicit goal of disease prediction[127].

When the available data were reduced to include all samples with additional cf-RNA data, only EN2, was identified as posessing predictive utility greater than the randomly permuted shadow features of Boruta (Figure 7.1). The sizeable increase in predictive utility of protein data may be due to an artefact of the dataset itself, when EN2 initially appears to discriminate disease status more clearly in the smaller ExoLISA cohort than in the larger ELISA cohort (Figure 7.2). However, questions about data quality of the ELISA cohort data have been raised (personal communication), where not all samples appear to have been collected to the agreed upon Movember standard procedures.

The decision to concentrate solely on the overlap of EN2 and cf-RNA samples therefore has a number of benefits. In addition to reducing the variables to consider there is a reduction in variance of the data, as the all of the samples quantified for both EN2 and cf-RNA originate the NNUH and were quantified by a single laboratory. Whilst the reduced variance is welcome, it likely comes at the cost of increased bias and reduced generalisability for developed models, as the variance in sample quality is something likely to be encountered in real world deployment of a urine test. This is something to be considered in the design of a validation trial, where extra samples should be collected where a developed model may require updating to account for this.

Collection of further samples is realistically the only way to find out whether observed increases in predictive ability in smaller amounts of data are caused by random splits of data, similar to that seen in Chapter 5, or whether the reduced variance is more indicative of a more carefully curated patient cohort. Regardless, neither can be answered by application of the FrameWork, and instead the rest of this chapter will concentrate on solvable analytical problems, including integrated analysis of EN2 and cf-RNA data.

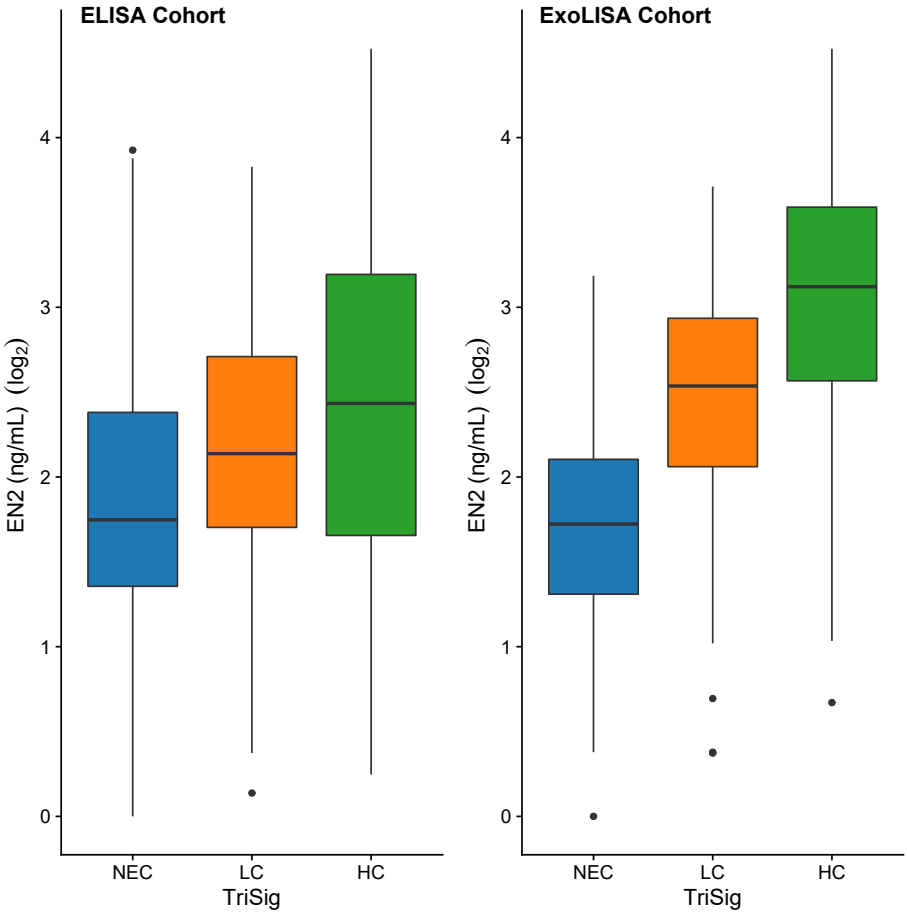


Figure 7.2: Quantified levels of EN2 in the ELISA cohort ($n = 471$) and the ExoLISA cohort ($n = 237$), shown according to TriSig level - No Evidence of Cancer (NEC), Gleason 3+3 or 3+4 (LC) or Gleason $\geq$ 4+3 (HC).

7.4 ExoGrail: an ideal scenario of few predictors and previously identified biomarkers

7.4.1 Methods

Patient cohort and characteristics Samples within the Movember GAP1 cohort (see Sections 9.2 & 3) that were analysed for both cf-RNA by NanoString and whole-urine EN2 protein levels by ELISA were eligible for the ExoGrail development cohort and used for model development ($n = 207$). All samples analysed in the ExoGrail cohort were collected from the Norfolk and Norwich University Hospital (NNUH, Norwich, UK).

Sample Processing

Cell-free RNA (cf-RNA) was isolated and quantified from urinary extracellular vesicles using NanoString technology as described in Section 3.1.1. All NanoString data presented here were normalised according to NanoString guidelines using NanoString internal positive controls, and log_2 transformed as described in Section 3.1.1.

Urinary EN2 protein concentration was quantified by ELISA using a monoclonal anti-mouse EN2 antibody, as described by Morgan et al. (2011)[128]. Clinical variables considered were serum PSA, age at sample collection, DRE impression and urine volume collected.

Statistical analysis

All data and scripts required to reproduce the ExoSpec analyses can be found at `https://github.com/UEA-Cancer-Genetics-Lab/ExoGrail`.

Feature Selection

A total of 172 variables were available within the ExoGrail development cohort for prediction (cf-RNA ($n = 167$), clinical variables ($n = 4$) and urinary EN2 ($n = 1$). Feature selection during the development of the ExoGrail cohort followed the same FrameWork process as described in Chapter 6. Briefly, the Boruta algorithm was applied to 1,000 resamples of the ExoGrail development cohort with replacement. Features were only positively retained for model fitting if selected by Boruta in $\geq 90\%$ of resampled Boruta runs.

Comparator Models

To evaluate potential clinical utility, additional models were trained as comparators using subsets of the available variables across the patient population, similarly to those models described in Chapter 6. A clinical standard of care (SoC) model was trained by incorporating age, PSA, T-staging and clinician DRE impression; a model using only the EN2 ELISA values (EN2, $n = 1$); and a model only using NanoString gene-probe information (NanoString, $n = 167$). The fully integrated ExoGrail model was trained by incorporating information from all of the above variables ($n = 177$). Each set of variables for comparator models were independently selected via the bootstrapped Boruta feature selection process described above to select the most optimal subset of variables possible for each predictive model.

Model Construction

All models were trained via the random forest algorithm[142], using the *randomForest* package[144] with the same parameters and process described in Chapter 6. Risk scores from trained models are presented as the out-of-bag predictions, as described in Section 3.2.

Models were trained on the TriSig label, again modified as a continuous label as in Chapter 6.

Statistical evaluation of models

Area Under the Receiver-Operator Characteristic curve (AUC) metrics were produced using the package[153], with confidence intervals calculated via 1,000 stratified bootstrap resamples. Density plots of model risk scores, and all other plots were created using the *ggplot2* package[207]. Partial dependency plots were calculated using the *pdp* package[208]. Estimation plots and calculations were produced using the *dabestr* package[157] and 1,000 bootstrap resamples were used to visualise robust effect size estimates of model predictions. Decision curve analysis (DCA)[154] examined the potential net benefit of using ExoGrail in the

clinic. Standardised net benefit (sNB) was calculated with the *rmda* package[190] and presented throughout our decision curve analyses as it is a more directly interpretable metric compared to net benefit[155].

7.4.2 Results

The ExoGrail Development cohort

Both urinary protein and transcriptomic data were available for 207 patients within the Movember GAP1 cohort, with all samples originating from the NNUH to form the ExoGrail development cohort (Table 7.4). The proportion of Gleason ≥7 disease in the ExoGrail cohort was 48%.

Table 7.4: Characteristics of the ExoGrail development cohort

	Cancer finding	No cancer finding
Collection Centre:		
NNUH, n (%)	130 (100)	77 (100)
Age:		
minimum	53.00	45.00
median (IQR)	68.50 (65.00, 76.00)	65.00 (59.00, 71.00)
mean (sd)	69.71 ± 7.67	65.22 ± 8.10
maximum	91.00	82.00
PSA:		
minimum	4.10	0.30
median (IQR)	10.35 (6.82, 16.48)	6.10 (3.70, 8.80)
mean (sd)	17.08 ± 18.33	7.89 ± 8.72
maximum	95.90	63.80
Prostate Size\(DRE Estimate):		
Small, n (%)	13 (10)	13 (17)
Medium, n (%)	64 (49)	34 (44)
Large, n (%)	38 (29)	21 (27)
Unknown, n (%)	15 (12)	9 (12)
Gleason Score:		
0, n (%)	0 (0)	77 (100)
6, n (%)	30 (23)	0 (0)
3+4, n (%)	48 (37)	0 (0)
4+3, n (%)	24 (18)	0 (0)
≥ 8, n (%)	28 (22)	0 (0)
Biopsy Result:		
Biopsy Positive, n (%)	130 (100)	0 (0)
Biopsy Negative, n (%)	0 (0)	52 (68)
No Biopsy, n (%)	0 (0)	25 (32)

Feature selection and model development

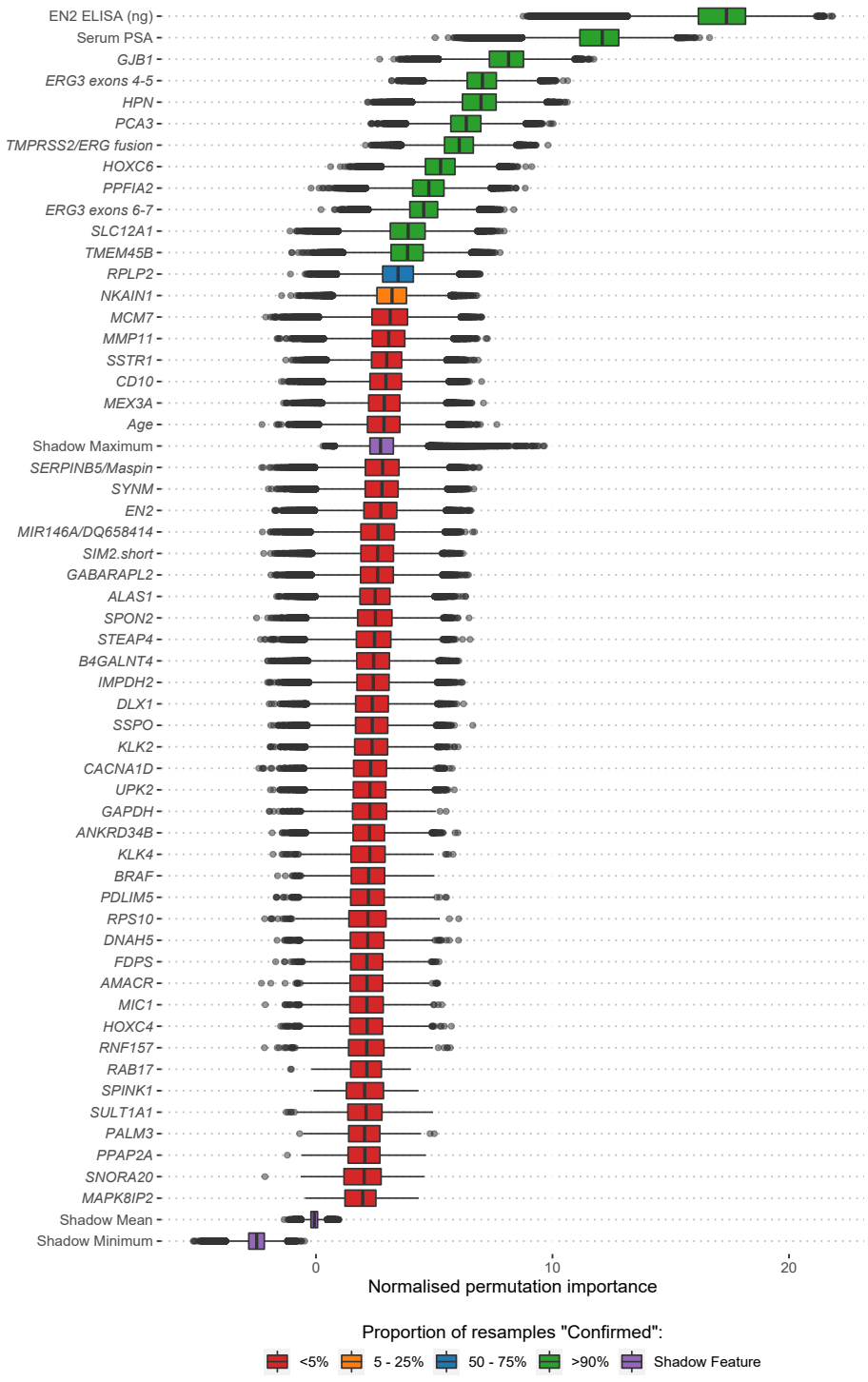


Figure 7.3: Analysis of variables available for the training of the ExoGrail model through the application of the Boruta algorithm via bootstrap resampling. 1,000 resamples with replacement of the available data were made, with the normalised permutation importance of each variable recorded at each iteration, along with the decision of Boruta within that resample. Fill colour shows the proportion of resamples that a feature was positively retained by Boruta.

Using the robust feature selection framework described above four models were produced in total; a standard of care (SoC) model incorporating only clinically available parameters (age and PSA), a model using urinary EN2 protein levels as the sole predictor variable (Engrailed), a model using only cf-RNA information (ExoRNA, 11 gene-probes) and the integrated model, named ExoGrail that incorporated variables from all three sources (12 variables) (Table 7.5). The ExoGrail model is a multivariable risk prediction model incorporating clinical parameters, urinary EN2 protein levels and cf-RNA expression information. When the resampling strategy was applied for feature reduction using Boruta, 12 variables were selected for the ExoGrail model. Each of the retained variables were positively selected in every resample and notably included information from clinical and cf-RNA variables, as well as urinary EN2 (Figure 7.3).

Table 7.5: Boruta-derived features positively selected for each model. Features are selected for each model by being confirmed as important for predicting biopsy outcome, categorised as a modified ordinal variable (see Methods) by Boruta in $\geq 90\%$ of bootstrap resamples

	SoC	Engrailed	ExoRNA	ExoGrail
Clinical Parameters	Serum PSA	-	-	Serum PSA
	Age	-	-	-
Methylation Targets	-	EN2 (ELISA)	-	EN2 (ELISA)
cf-RNA Targets	-	-	*AMACR*	-
	-	-	*ERG* exons 4-5	*ERG* exons 4-5
	-	-	*ERG* exons 6-7	*ERG* exons 6-7
	-	-	*GJB1*	*GJB1*
	-	-	*HOXC6*	*HOXC6*
	-	-	*HPN*	*HPN*
	-	-	*NKAIN1*	*NKAIN1*
	-	-	*PCA3*	*PCA3*
	-	-	*PPFIA2*	*PPFIA2*
	-	-	*RPLP2*	-
	-	-	-	*SLC12A1*
	-	-	*TMEM45B*	*TMEM45B*
	-	-	*TMPRSS2/ERG* fusion	*TMPRSS2/ERG* fusion

In the SoC comparator model only PSA and age were selected as important predictors. Urinary EN2 levels were confirmed as important in the independent Engrailed model as the sole variable, and also within the ExoGrail model (Table 2). For the cf-RNA model, 11 transcripts were selected, notably including both variants of the *ERG* gene-probe and *TM-*

PRSS2/ERG fusion gene-probe. ExoGrail incorporated an additional cf-RNA transcript, *SLC12A1*, which was not previously selected in the ExoRNA comparator model. When this was examined by partial dependency plots, an additive non-linear interaction effect was observed between quantified levels of urinary EN2 and counts of *SLC12A1* on the predicted ExoGrail risk signature output (Figure 7.4).

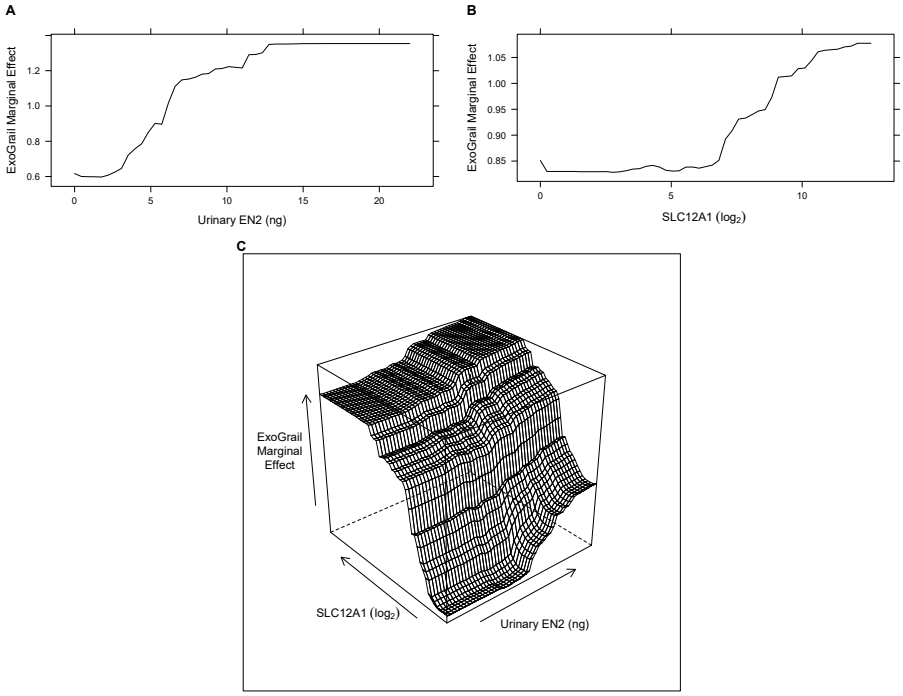


Figure 7.4: Partial dependency plots detailing the marginal effects and interactions of *SLC12A1* and urinary EN2 on predicted ExoGrail Risk Score. A - Partial dependency of ExoGrail on urinary EN2, B - Partial dependency of ExoGrail on *SLC12A1*, C - Partial dependency of ExoGrail on both *SLC12A1* and urinary EN2

ExoGrail predictive ability

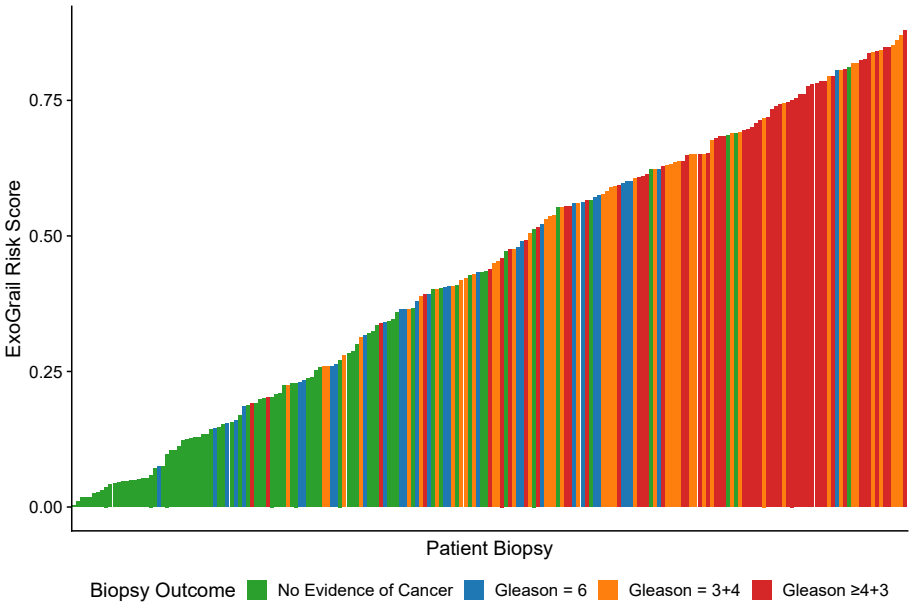


Figure 7.5: Waterfall plot of the ExoGrail risk score for each patient. Each coloured bar represents an individual patient's calculated risk score and their true biopsy outcome, coloured according to Gleason score . Green - No evidence of cancer, Blue – Gleason = 6, Orange - Gleason = 3+4, Red - Gs ≥ 4+3

As ExoGrail Risk Score (range 0-1) increased, the likelihood of high-grade disease detection on TRUS-biopsy was significantly greater (Proportional odds ratio = 2.21 per 0.1 ExoGrail increase, 95% CI: 1.91 - 2.59; ordinal logistic regression, Figure 7.5). The median ExoGrail risk score for metastatic patients was 0.76 (n = 11). These patients were not included in the ExoGrail development cohort and can be considered as a positive control for model calibration.

Table 7.6: AUC of random forest models for detecting differing outcomes on initial biopsy. Brackets show 95% confidence intervals of the AUC, calculated from 2,000 bootstrap resamples.

Biopsy outcome:	SoC	Engrailed	ExoRNA	ExoGrail
Gleason ≥ 4+3:	0.77 (0.69 - 0.84)	0.81 (0.73 - 0.87)	0.67 (0.59 - 0.75)	0.84 (0.78 - 0.90)
Gleason ≥ 3+4:	0.72 (0.65 - 0.79)	0.83 (0.77 - 0.88)	0.77 (0.70 - 0.83)	0.90 (0.85 - 0.94)
Any Cancer	0.75 (0.68 - 0.82)	0.81 (0.74 - 0.87)	0.81 (0.75 - 0.87)	0.89 (0.85 - 0.94)

ExoGrail was superior to all other models for the detection of Gleason ≥ 3+4 (AUC = 0.90 (95% CI: 0.85 - 0.94), P <0.001, bootstrap test with 1,000 resamples) and for any cancer (AUC = 0.89 (95% CI: 0.85 - 0.94), P <0.001, bootstrap test with 1,000 resamples) (Table 7.6). When Gleason ≥ 4+3 was considered, ExoGrail returned an AUC = 0.84 (95% CI: 0.78 - 0.90), outperforming the SoC and cf-RNA models (P <0.001, bootstrap test with

1000 resamples), whilst the Engrailed model displayed similar performance by AUC metrics (Table 7.6).

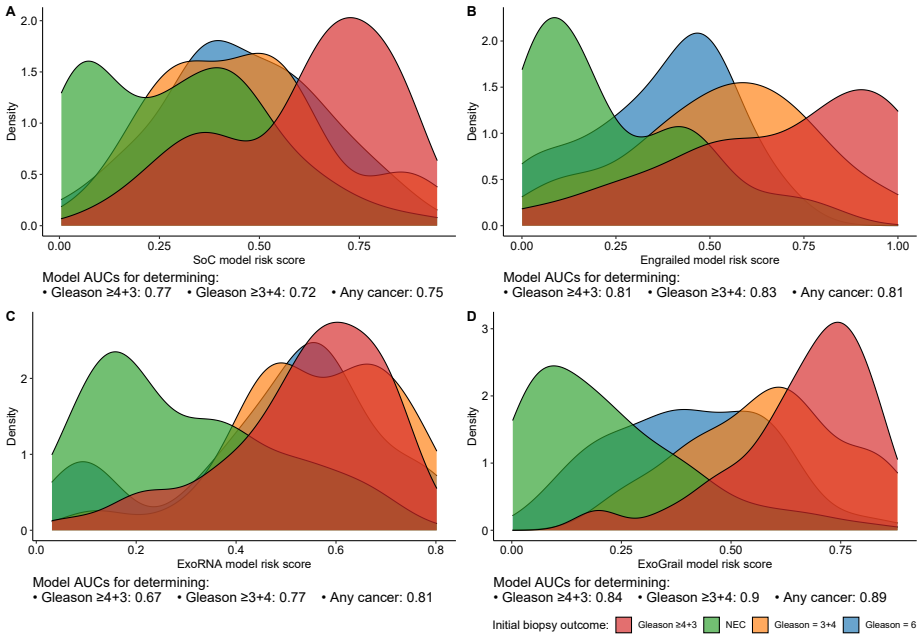


Figure 7.6: Risk score distributions of the four trained models, calculated as the out-of-bag predictions and represented as density plots. AUCs for each model's predictive ability for clinically relevant outcomes are detailed underneath each panel. Each random forest model was fit using differing input variables; A - SoC clinical risk model, including Age and PSA, B - Engrailed model, C -ExoRNA model and D - ExoGrail model, combining predictors from all three modes of analysis. The full list of variables in each model is available in Table 1. Fill colour shows the risk score distribution of patients with respect to biopsy outcome: No evidence of cancer (Green), Gleason 6 (Blue), Gleason 3+4 (Orange), Gleason $\geq$ 4+3 (Red).

As revealed by the distributions of risk scores and AUC, ExoGrail achieved a clearer discrimination of disease status Gleason $\geq$ 3+4 disease from other outcomes when compared to any of the other models (ExoGrail all $P < 0.01$ bootstrap test, 1,000 resamples, Figure 7.6).

Investigation of risk score distributions found that whilst the SoC model returned respectable AUCs and detection of the higher grade disease (Gleason $\geq$ 3+4), it displayed a relative inability to clearly stratify intermediate disease states. This uncertainty would cause large numbers of patients to be inappropriately selected for further investigation (Figure 7.6A). For example, to classify 90% of patients with Gleason 7 disease correctly, an SoC risk score of 0.251 would misclassify 64.5% of men with less significant, or no disease. The Engrailed model detailed clearer discrimination, though featured a bimodal distribution of patients without prostate cancer (Figure 7.6B, green density plot), falsely identifying 51.4% of patients with low grade disease as similar to those with more clinically significant disease (Figure 7.6B). Whilst the AUCs returned for the ExoRNA model were lower, the distribution of risk scores shows that ExoRNA could more accurately discriminate cancer from

non-cancer than either the SoC or EN2 models, a key clinical step in the triage of patients prior to biopsy (Figure 7.6C).

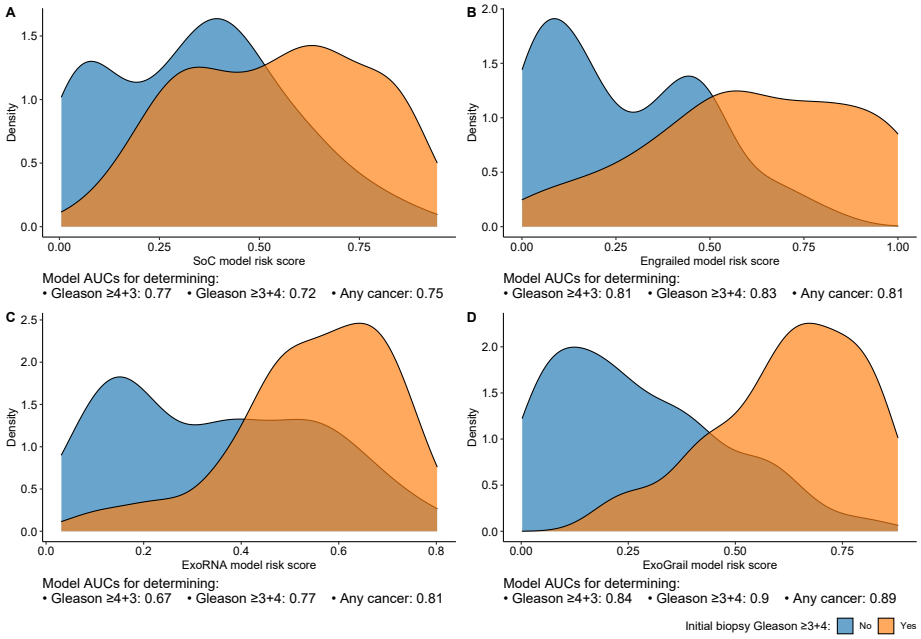


Figure 7.7: Density plots detailing risk score distributions generated from four trained models. Models A to D were trained with different input variables; A - SoC clinical risk model, including Age and PSA, B - Engrailed model, C -ExoRNA model and D - ExoGrail model, combining the predictors from all three previous models. The full list of variables in each model is available in Table 1. Fill colour shows the risk score distribution of patients with a significant biopsy outcome of Gs $\geq$ 3+4 (Orange) or Gs $\geq$ 6 (Blue)

Examination of ExoGrail scores displayed similar distributions for NEC patients as the ExoRNA model whilst also being able to more accurately separate different cancer outcomes from biopsy, resulting in fewer misclassifications no-cancer patients if binary detection of 95% of Gleason $\geq$ 3+4 were considered (28% of NEC patients misclassified). The greater discriminatory ability of the ExoGrail model when biopsy outcomes are considered as a binary Gleason $\geq$3+4 threshold can also be seen in Figure 7.7.

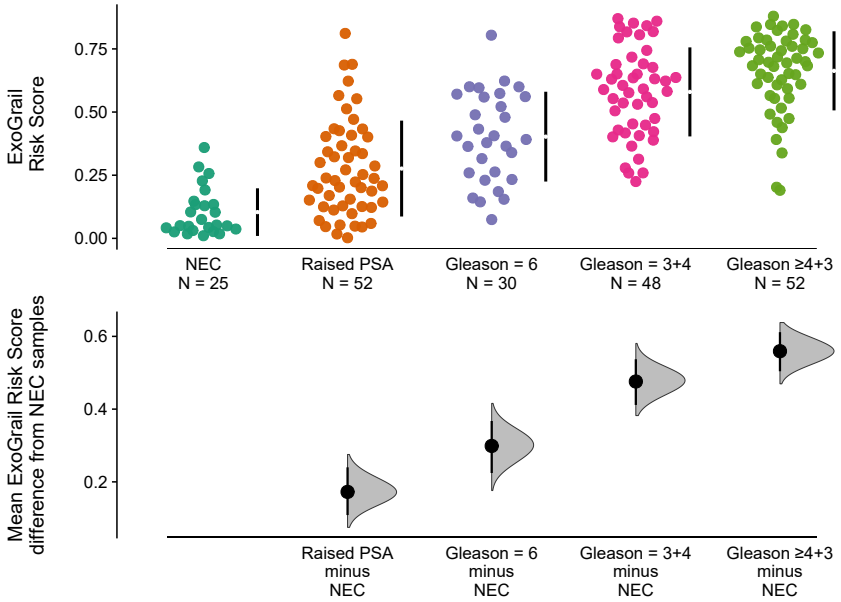


Figure 7.8: Mean ExoGrail risk score differences between biopsy outcomes, as represented by Estimation plots. Individual patient risk scores (y-axis) are presented as points in the top panel, separated according to Gleason score (x-axis) with gapped vertical lines detailing the mean and standard deviation of each clinical group's ExoGrail risk score. Mean ExoGrail risk score differences relative to the no evidence of cancer (NEC) group are shown in the bottom panel. Mean difference and 95% confidence intervals are shown as a point estimate and vertical bar, respectively, with density plots generated from 1,000 bias-corrected and accelerated bootstrap resamples.

Comparisons of ExoGrail mean ExoGrail signatures between groups was performed with resampling and estimation plots (1,000 bias-corrected and accelerated bootstrap resamples, Figure 7.8). The mean ExoGrail differences between patients with no evidence of cancer on biopsy were: Gleason 6 = 0.3 (95% CI: 0.22 - 0.37), Gleason 3+4 = 0.48 (95% CI: 0.41 - 0.54) and Gleason $\geq$ 4+3 = 0.56 (95% CI: 0.5 - 0.61). Of note, patients with no evidence of cancer had a lower ExoGrail risk score (mean difference = 0.17 (95% CI: 0.11 - 0.24)) than those with a raised PSA but no findings of cancer on biopsy (Figure 7.8).

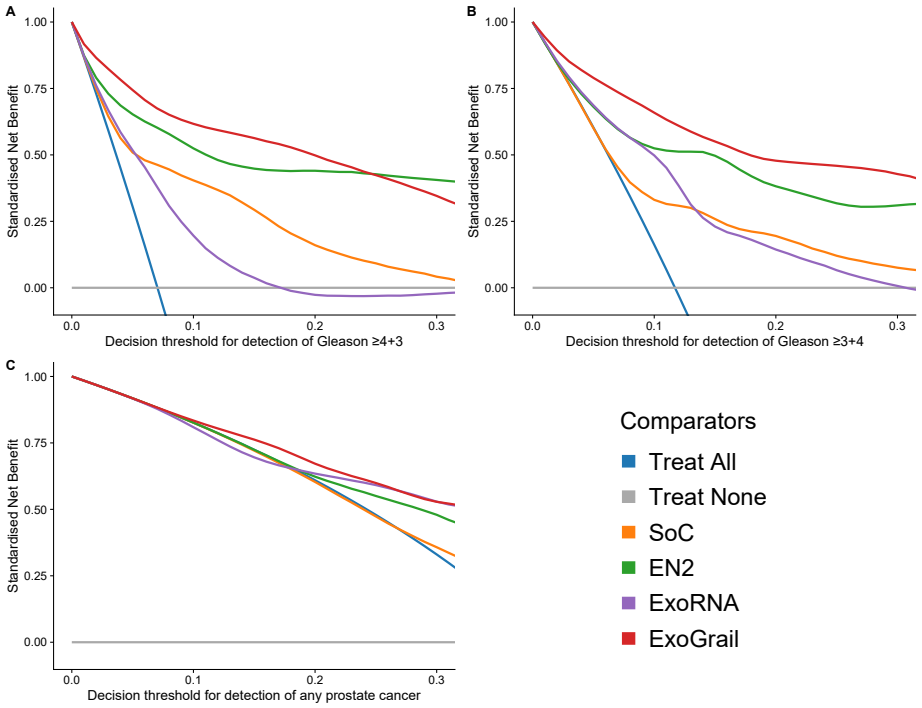


Figure 7.9: Exploration of the standardised net benefit (sNB) by decision curve analysis (DCA) for adopting risk models to aid the decision to undertake an initial biopsy for patients presenting with a serum PSA $\geq$ 4 ng/mL, where current clinical practice is to biopsy all patients. The accepted patient/clinician risk threshold for accepting biopsy is detailed on the x-axis. Different biopsy outcomes are shown in each of the three panels; A - detection of Gleason $\geq$ 4+3, B - detection of Gleason $\geq$ 3+4, C - any cancer; Blue- biopsy all patients with a PSA $>$ 4 ng/mL, Orange - biopsy patients according to the SoC model, Green - biopsy patients based on the Engrailed model, Purple - biopsy patients based on the exoRNA model, Red - biopsy patients based on the ExoGrail model. To assess the benefit of adopting these risk models in a clinically relevant population we used data available from the control arm of the CAP study for proportionally resampling the ExoGrail cohort. DCA curves were calculated from 1,000 bootstrap resamples of the available data to match the distribution of disease reported in the CAP trial population. Mean sNB from these resampled DCA results are plotted here.

Decision curve analyses examined the net benefit of ExoGrail adoption in a population of patients with a clinical suspicion of prostate cancer and a PSA level suitable to trigger biopsy ($\geq$ 4 ng/mL). The biopsy of men based upon their ExoGrail risk score provided a net benefit over current standards of care across all decision thresholds examined and was the most consistent amongst all comparator models across a range of clinically relevant endpoints for biopsy (Figure 7.9).

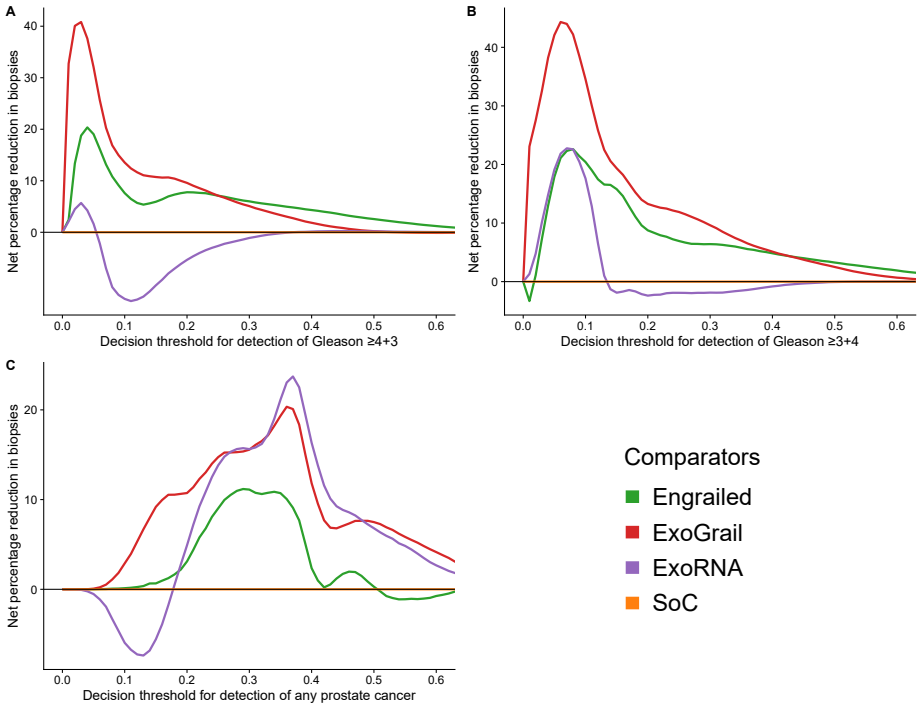


Figure 7.10: . Estimation of biopsy reduction, as calculated by comparing the DCA-calculated net benefit of each risk model to the net benefit of the standard of care (SoC) model. The accepted patient/clinician risk threshold for accepting biopsy is detailed on the x-axis. Different biopsy outcomes are shown in each of the three panels; A - detection of Gleason $\geq$ 4+3, B - detection of Gleason $\geq$ 3+4 and C - any cancer. Coloured lines show differing comparator models; Blue- biopsy all patients with a PSA $>$ 3 ng/mL, Orange - biopsy patients by according the to the SoC model, Green - biopsy patients based on the Engrailed model, Purple - biopsy patients based on the ExoRNA model, Red - biopsy patients based on a the ExoGrail model. To assess the benefit of adopting these risk models in a clinically relevant population we used data available from the control arm of the CAP study for proportionally resampling the ExoGrail cohort. DCA curves were calculated from 1,000 bootstrap resamples of the available data to match the distribution of disease reported in the CAP trial population. Net benefit, averaged over all resamples are used to calculate the potentially reductions in biopsy rates here.

Using the SoC model as the baseline with which to compare the potential for biopsy reduction of each model, we found that ExoGrail could reduce unnecessary biopsy rates by upwards of 40%, depending on accepted patient-clinician risk. For example, if a decision threshold of 0.1 were accepted, representing a perceived risk of 1 in 10 for Gleason $\geq$ 3+4 on biopsy, ExoGrail could result in up to a 35% reduction in unnecessary biopsies of men presenting with a suspicion of prostate cancer, whilst also correctly identifying patients with more aggressive disease. If Gleason $\geq$ 4+3 were considered the threshold of clinical significance, a more conservative decision threshold of 0.05 could save 32% of men from receiving an unnecessary biopsy (Figure 7.10).

7.4.3 Discussion

The results here represent the potential that can be harnessed both from the FrameWork, and the data within the Movember GAP1 project, showing that an improved multivariable risk prediction model can be robustly developed from the information derived from multiple urine fractions in conjunction with clinically based measurements. The final model named ExoGrail incorporated markers from genes well known to be associated with prostate cancer, and proven in other urinary tests such as *PCA3*, *HOXC6*, and the *TMPRSS2/ERG* gene fusion. An interesting, non-linear interaction between the urinary protein levels of EN2 and quantified levels of the *SLC12A1* transcript shows both the benefit of considering information from multiple biological sources, and using statistical approaches capable of handling non-linear terms readily (Figure 7.4).

As shown with ExoMeth in Chapter 6, application of the FrameWork for the integration of datasets from sources with previously proven prognostic utility is able to produce favourable results. Where multiple assays could be performed at the same time on urine samples in future there may be more of a benefit to considering their results simultaneously within a single risk model or applying the FrameWork, as opposed to considering each in isolation separately or at best, additively. However, the same artefacts of the smaller EN2 dataset discussed above in section 7.3.3 cannot be ruled out without a larger dataset to examine.

7.5 ExoSpec: high-dimensionality data require alterations to the FrameWork

7.5.1 Methods

Patient cohort and characteristics

Samples within the Movember GAP1 cohort (see Sections 9.2 & 3) with both cf-RNA NanoString data and capillary electrophoresis mass spectrometry (CE-MS) analysis completed were used for the ExoSpec development cohort ($n = 192$). All samples analysed in the ExoSpec cohort were collected from the Norfolk and Norwich University Hospital (NNUH, Norwich, UK).

Sample Processing

Cell-free RNA (cf-RNA) was isolated and quantified from urinary extracellular vesicles using NanoString technology as described in Section 3.1.1. All NanoString data presented here were normalised according to NanoString guidelines using NanoString internal positive controls, and log_2 transformed as described in Section 3.1.1.

Clinical variables considered were serum PSA, age at sample collection, DRE impression and urine volume collected. Peptidomic analysis by CE-MS was performed by Mosaique Diagnostics in Germany using protocols previously established by Metzger et al. (2013)[209], and described in Section 3.1.1.

Statistical analysis

Peptide data were filtered *a priori* (see Section 3.1.4. All data and scripts required to reproduce the ExoSpec analyses can be found at `https://github.com/UEA-Cancer-Genetics-`

`Lab/ExoSpec`.

Feature Selection

Following *a priori* filtering, the ExoSpec cohort comprised a total of 814 possible variables for predictive modelling including cf-RNA ($n = 167$), peptides ($n = 643$) and clinical variables ($n = 4$) was derived. The large featureset, even post-filtering, made feature selection by resampled-Boruta infeasible. Alterations were made to the FrameWork described in Chapter 6 at the feature selection stage, with cross-validation and penalisation used to assess variable importance as follow: variables significantly associated to TriSig level were robustly identified by the application of a 20-fold cross validated LASSO-penalised generalised linear model, fit using the *glmnet* package[210]. Only those variables with coefficients were not decreased to zero by the LASSO penalisation were considered further, and positively selected as input for model fitting.

Comparator Models

To evaluate potential clinical utility, additional models were trained as comparators using subsets of the available variables across the patient population, similarly to those models described in Chapter 6. A clinical standard of care (SoC) model was trained by incorporating age, PSA, T-staging and clinician DRE impression; a model using only the pre-filtered peptides (MassSpec, $n = 643$); and a model only using NanoString gene-probe information (NanoString, $n = 167$). The fully integrated ExoSpec model was trained by incorporating information from all of the above variables ($n = 814$). Each set of variables for comparator models were independently selected via the cross-validated LASSO feature selection process described above to select the optimal subset of variables possible for each predictive model.

Model Construction

See section 7.4.1 above.

Statistical evaluation of models

See section 7.4.1 above.

7.5.2 Results

The ExoSpec Development cohort

Paired cf-RNA and proteomic data were available for 192 patients within the Movember GAP1 cohort, all originating from the NNUH and forming the ExoSpec development cohort (Table 7.7). The proportion of Gleason ≥ 7 disease in the ExoSpec cohort was 53%.

Table 7.7: Characteristics of the ExoSpec development cohort

	Cancer finding	No cancer finding
Collection Centre:		
NNUH, n (%)	133 (100)	59 (100)
Age:		
minimum	53.00	45.00
median (IQR)	70.00 (65.00, 76.00)	67.00 (59.50, 71.00)
mean (sd)	70.23 ± 7.81	66.15 ± 8.30
maximum	91.00	82.00
PSA:		
minimum	4.10	0.30
median (IQR)	10.40 (6.90, 16.60)	5.30 (2.30, 7.95)
mean (sd)	16.81 ± 17.36	6.44 ± 5.96
maximum	95.90	30.30
Prostate Size (DRE Estimate):		
Small, n (%)	12 (9)	16 (27)
Medium, n (%)	67 (50)	25 (42)
Large, n (%)	38 (29)	14 (24)
Unknown, n (%)	16 (12)	4 (7)
Gleason Score:		
0, n (%)	0 (0)	59 (100)
6, n (%)	31 (23)	0 (0)
3+4, n (%)	48 (36)	0 (0)
4+3, n (%)	25 (19)	0 (0)
≥ 8, n (%)	29 (22)	0 (0)
Biopsy Result:		
Biopsy Positive, n (%)	133 (100)	0 (0)
Biopsy Negative, n (%)	0 (0)	36 (61)
No Biopsy, n (%)	0 (0)	23 (39)

Feature selection and development of models

LASSO-penalised general linear models were applied through 20-fold cross-validation to the individual feature sets; the cf-RNA variables, the urinary mass spectrometry counts, and the clinically available parameters, as described above. Following *a priori* filtering,643 peptides were considered for feature selection through regression analysis, with a 13 peptides selected as predictive for biopsy outcome. Among the 13 significant predictive peptides were fragments of matrix metalloproteinase-2 (MMP2), three peptide fragments of fibrinogen alpha chain (FGA), NAD kinase (NADK) and Histone H1.4 (HIST1H1E) were all identified with increased urinary abundance in prostate cancer patients (Table 7.8). Substrates of MMP2, such as fibrillar type I, collagen 1 alpha 1 (COL1A1) and basal collagen type IV (COL4A3, COL4A5) were detected in decreased abundance, as well as glutamate dehydrogenase 1 (GLUD1) (Wilcox Rank Sum Test, Table 7.8).

Application of the same LASSO regression to the cf-RNA feature-set resulted in six significant transcripts being positively selected. The *ERG* exons 4 - 5 gene-probe along with

PCA3, *TMEM45B* and *SLC12A1* transcripts was identified in increased urinary abundance, whilst *SERPINB5* and *SNORA20* were expressed at decrease levels in prostate cancer patients(Table 7.8). Patient age and PSA level were the only clinically available parameters to be selected by the LASSO regression (Table 7.8).

Consideration of the fully integrated featureset identified 12 variables as important towards predicting outcome across all available clinical, cf-RNA and peptide variables (Table 7.8). Of the 13 peptides selected for input to the MassSpec model and the 11 cf-RNA geneprobes, only four peptides and three cf-RNA probes were retained for the final ExoSpec model. Interestingly the ExoSpec model further incorporated an additional two peptide fragments of FGA not deemed to be useful when only peptides were considered (Table 7.8).

Table 7.8: Features selected by the cross-validated LASSO to be used as input variables for each Random Forest comparator model.

	SoC	MassSpec	ExoRNA	ExoSpec	Fold Change (No Cancer vs Cancer)
Clinical Parameters	Serum PSA	-	-	Serum PSA	10.37*
	Age	-	-	Age	4.08*
Peptide Targets	-	-	-	FGA	2.05†
	-	-	-	FGA	1.66†
	-	COL1A1	-	-	-0.45†
	-	COL1A1	-	-	-0.84†
	-	MMP2	-	-	2.26†
	-	COL2A1	-	COL2A1	2.80†
	-	COL4A4	-	-	2.42†
	-	GLUD1	-	GLUD1	-1.12†
	-	COL1A1	-	-	2.67†
	-	COL4A3	-	-	-0.71†
	-	HIST1H1	-	HIST1H1	2.82†
	-	COL4A5	-	-	-0.81†
	-	NADK	-	-	0.34†
	-	FGA	-	FGA	2.49†
cf-RNA targets	-	-	*ERG* exons 4-5	*ERG* exons 4-5	2,25†
	-	-	*PCA3*	*PCA3*	2.08†
	-	-	*SERPINB5/Maspin*	-	-0.27†
	-	-	*SLC12A1*	*SLC12A1*	1.81†
	-	-	*SNORA20*	-	-0.26†
	-	-	*TMEM45B*	*TMEM45B*	0.94†

* Absolute fold change † log_2 fold change

The features above, identified as possessing a significant association to biopsy outcome, were subsequently used to train four Random Forest-based models, utilising differing subsets of the available featuresets:

1) A clinical standard of care (SoC) model, using only clinically available information (n = 2).
2) A model using only peptide data from CE-MS (MassSpec, n = 13).
3) A cf-RNA model, using only NanoString-derived cf-RNA counts (ExoRNA, n = 6).

4) The fully integrated model, using information from clinical, peptide and cf-RNA data (ExoSpec $n = 12$).

Predictive utility of ExoSpec and comparator models

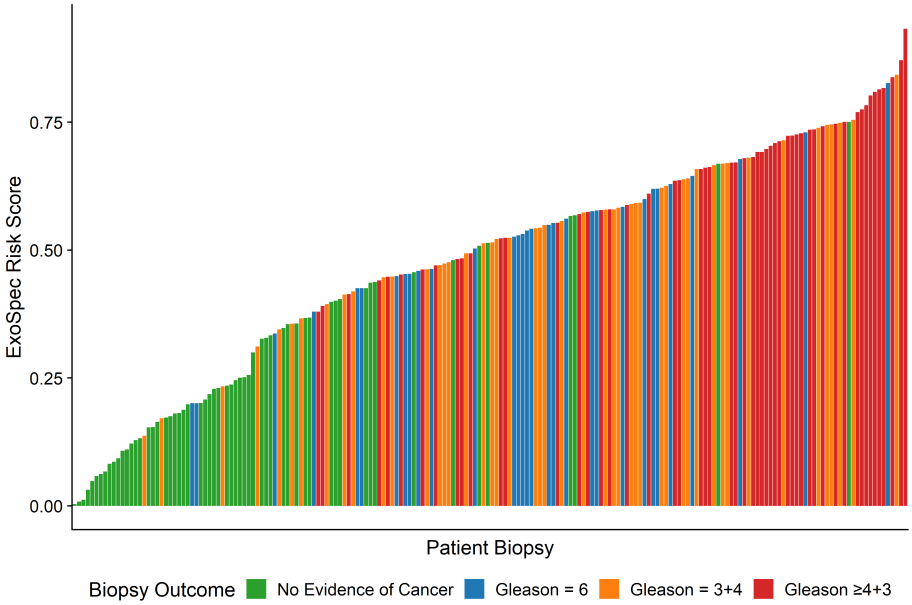


Figure 7.11: ExoSpec risk score for each patient, presented as a waterfall plot. Each individual biopsy is represented as a coloured bar, where the height represents the predicted risk score, and filled according to the Gleason score (Gs). In a perfectly calibrated model the colours would be ordered with no overlap. Green - No evidence of cancer, Blue – Gs 6, Orange - Gs 3+4, Red - Gs $\geq$ 4+3.

ExoSpec appeared to be well calibrated, ordering patients appropriately in ascending risk, with an increased ExoSpec score (range 0 - 1) resulting in a significantly higher likelihood of more aggressive disease being detected upon an initial biopsy (Proportional odds ratio = 2.26 per 0.1 ExoSpec increase, 95% CI: 1.91 - 2.71; ordinal logistic regression, Figure 7.11)

The AUC returned by ExoSpec for predicting the presence of Gleason $\geq$ 3+4 = 0.83 (95% CI: 0.77 - 0.88), superior to the SoC model, as well as both molecular based models (MassSpec and ExoRNA, all $P < 0.001$, bootstrap test, 1,000 resamples, Table 7.9). When the detection (or exclusion) of any cancer on initial biopsy was considered, ExoSpec showed a remarkable predictive ability, with an AUC of 0.91 (95% CI: 0.86 - 0.96, Table 7.9).

Table 7.9: AUC of random forest models for detecting differing outcomes on initial biopsy. Brackets show 95% confidence intervals of the AUC, calculated from 1,000 bootstrap resamples.

Biopsy outcome:	SoC	MassSpec	ExoRNA	ExoSpec
Gleason =4+3:	0.76 (0.69 - 0.83)	0.70 (0.62 - 0.77)	0.67 (0.58 - 0.74)	0.82 (0.75 - 0.88)
Gleason =3+4:	0.71 (0.64 - 0.78)	0.69 (0.60 - 0.76)	0.75 (0.67 - 0.81)	0.83 (0.76 - 0.88)
Any Cancer	0.78 (0.70 - 0.85)	0.76 (0.68 - 0.83)	0.84 (0.78 - 0.90)	0.91 (0.86 - 0.95)

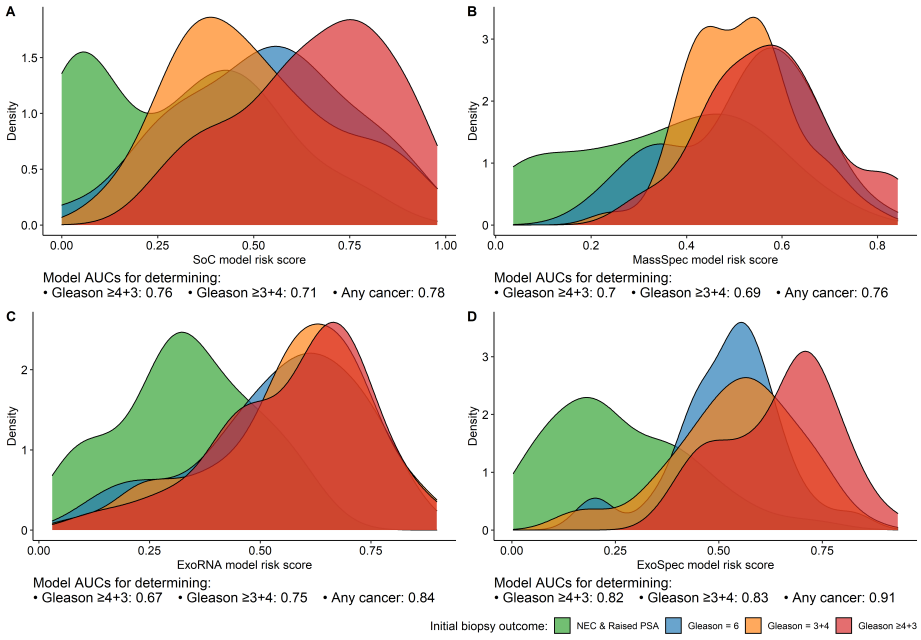


Figure 7.12: Risk score distributions generated by the four comparator models fit to the data, where each comparator was fit with different input variables. A - SoC clinical risk model, including Age and PSA, B - MassSpec model incorporating peptide data, C -ExoRNA model, utilising only cf-RNA data D - ExoSpec model, integrating clinical parameters, peptide data and cf-RNA data. Biopsy outcomes are indicated according to fill colour, where a clinically significant biopsy outcome (Gs $\geq$ 3+4) is orange and Gs $\leq$ 6 on biopsy is blue.

When risk score distributions were explored, it was confirmed that the SoC model broadly reflected the problems currently exhibited in the clinic. The SoC model was able to discriminate the lowest and highest risk patients with good accuracy, but not accurately separate clinically significant Gleason 3+4 disease from Gleason 6, with the latter possessing a higher mean SoC risk score than the more indolent disease group (Figure 7.12A). Both the MassSpec and the ExoRNA comparator models performed similarly, with no significant improvement of the one AUC estimate over the other for predicting biopsy outcome (Figure 7.12B & 2C). The integrated ExoSpec model displayed clear improvements and showed substantial value in excluding a false cancer finding on initial biopsy (Figure 7.12D).

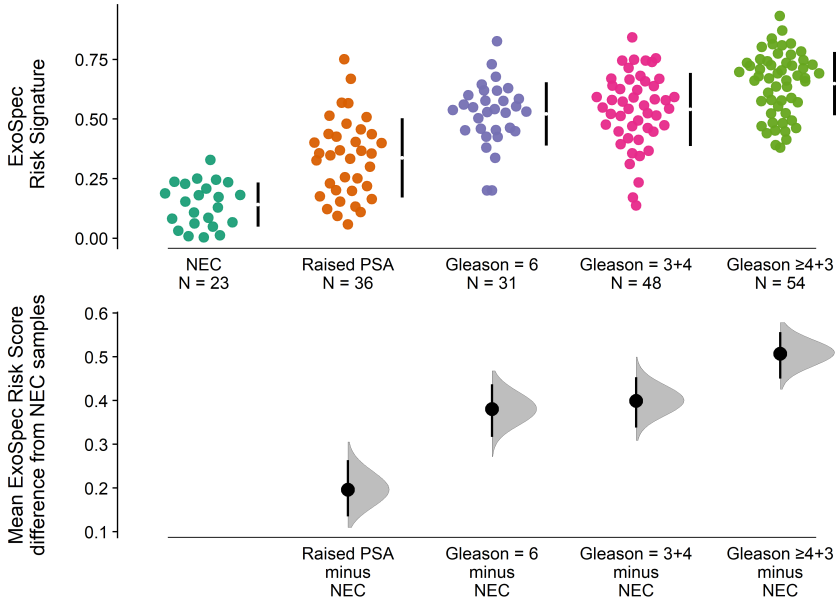


Figure 7.13: Estimation plots for the ExoSpec risk signature, where the top row details each patient biopsy as a point, stratified by Gleason score across the x-axis and ExoSpec risk signature on the y-axis. Each patient sample point is coloured according to their D'Amico clinical risk category; NEC - No evidence of cancer, Raised PSA - Raised PSA with negative biopsy, L -D'Amico Low-Risk, I - D'Amico Intermediate Risk, H - D'Amico High-Risk. Mean and standard deviation ExoSpec risk signatures for each group is shown by the gapped vertical lines. The bottom panel shows mean differences in ExoSpec signatures relative to NEC patient samples. Calculated from bias-corrected and accelerate bootstrap resampling (1,000 resamples with replacement), sample density distributions are presented with a point estimate and vertical bar to show mean difference and 95% confidence intervals, respectively.

Resampling of ExoSpec predictions via estimation plots allowed for comparisons of mean ExoSpec scores between groups by 1,000 bias-corrected and accelerated bootstrap resamples (Figure 7.13). The mean ExoSpec score differences between patients with no evidence of cancer on biopsy were: Gleason 6 = 0.38 (95% CI: 0.32 - 0.44), Gleason 3+4 = 0.4 (95% CI: 0.34 - 0.45) and Gleason ≥4+3 = 0.51 (95% CI: 0.45 - 0.56). Notably, patients with a raised PSA but negative for cancer on biopsy had a higher ExoSpec risk score than those with no evidence of cancer (mean difference = 0.2 (95% CI: 0.13 - 0.26)). Patients negative for cancer findings and with elevated PSA levels also exhibited a wider ExoSpec score distribution than other clinical categories, suggesting these patients may not form a homogeneous molecular or biological group (Figure 7.13).

Net benefit of adopting ExoSpec

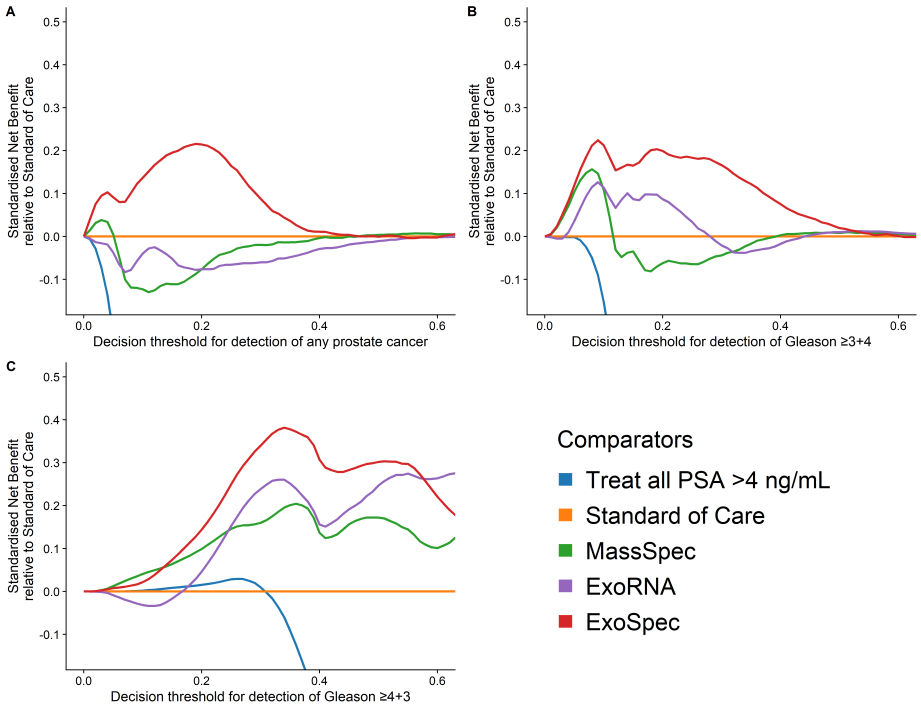


Figure 7.14: Standardised net benefit (sNB) of adopting each comparator model into clinical practice, displayed as decision curves, relative to standards of care. Accepted risk thresholds for the interpreter before agreeing to biopsy are shown on the x-axis. Each panel shows the relative sNB of a different biopsy outcome result when compared to standards of care: A-detection of any prostate cancer, regardless of Gleason, B - detection of Gleason = 3+4, C - detection of Gleason = 4+3. Coloured lines in each panel detail the comparator: Orange – biopsy of patients according to current standards of care, Green - biopsy patients based on the MassSpec model, Purple - biopsy patients based on the ExoRNA model, Red - biopsy patients based on the ExoSpec model. Data presented here were calculated from 1,000 stratified bootstrap resamples of the available data to match the disease proportions reported from the control arm of the CAP study. The mean sNB from these resamples were calculated and presented here.

Decision curve analysis examined the net benefit of adopting ExoSpec in a population of patients suspected to harbour prostate cancer, with a PSA threshold of 4 ng/mL, suitable to trigger biopsy by current clinical guidelines[5]. Using the SoC model as the baseline with which to compare ExoSpec to, the biopsy of patients based upon their ExoSpec risk score consistently provided a net benefit across all decision thresholds and endpoints examined and was the only comparator model not apparently harmful at some threshold when compared to the SoC model (Figure 7.14). The ExoSpec model again showed a synergistic ability to rule out disease on an initial biopsy, greater than each of the comparator models in isolation (Figure 7.14.

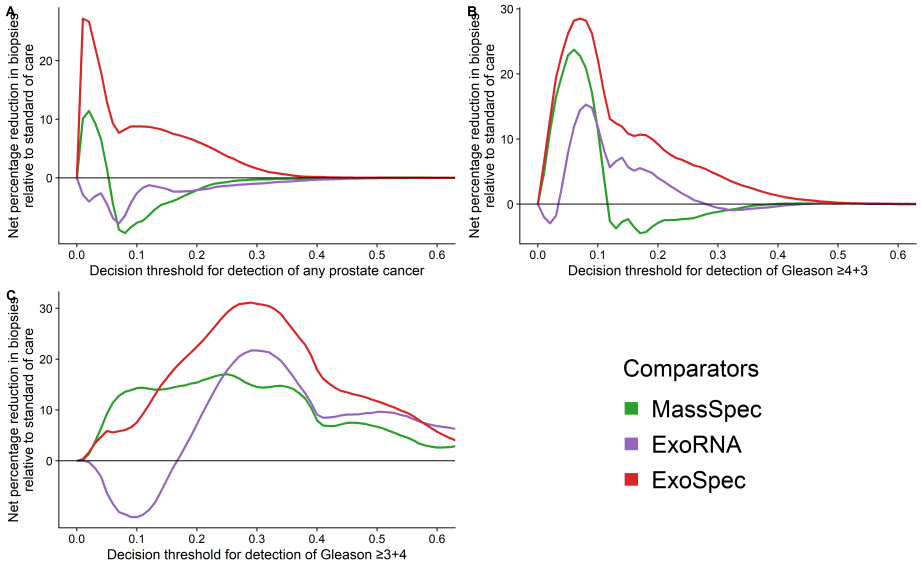


Figure 7.15: Potential reductions in unnecessary biopsies when considering different biopsy outcomes, calculated by measuring net benefit that the adoption of difference comparator risk models could bring compared to standards of care. Accepted risk thresholds for the interpreter before agreeing to biopsy are shown on the x-axis. Each panel details the percentage reduction in biopsies for a differing biopsy outcome. Each panel shows the relative sNB of a different biopsy outcome result when compared to standards of care: A- detection of any prostate cancer, regardless of Gleason, B - detection of Gleason = 3+4, C - detection of Gleason = 4+3. Coloured lines in each panel detail the comparator: Orange – biopsy of patients according to current standards of care, Green - biopsy patients based on the MassSpec model, Purple - biopsy patients based on the ExoRNA model, Red - biopsy patients based on the ExoSpec model. Data presented here were calculated from 1,000 stratified bootstrap resamples of the available data to match the disease proportions reported from the control arm of the CAP study. The mean change in biopsies performed were calculated across all resamples and presented here as a percentage, for full details see Methods.

Once more, when compared to the SoC model, ExoSpec could result in a reduction in unnecessary biopsies by more than 30% for detecting clinically significant (Gleason $\geq$ 7) disease across a range of reasonable decision thresholds (0.1 – 0.3, Figure 7.15).

7.5.3 Discussion

ExoSpec, whilst showing promising results, represents some of the limitations in the design of the FrameWork through the use of such highly-dimensional data seen in the CE-MS dataset. The Boruta algorithm as a feature selection tool is very good as an all-relevant feature selection approach robust to collinear predictors and interactions[149]. However, this comes at the cost of increasing computational requirements, where Boruta scales with $O(P \times N)$, where P and N are the numbers of attributes and samples, respectively[149]. With the resampling employed within the FrameWork this computational requirement is then multiplied further by both the number of resamples taken and the number of comparator

models, making it an untenable prospect. With the computational limitations considered, a more appropriate initial approach was applied here, using cross-validation and LASSO penalisation to identify variables important to predicting biopsy outcome.

Connell *et al.* have previously reported cf-RNA derived features as predictive biomarkers for significant prostate cancer[8,111] (See Chapters 4 and 6), including *ERG* exons 4 - 5, *PCA3*, *SERPINB5*/Maspin, *SLC12A1* and *TMEM45B* selected here. Similarly, several collagen and fibrinogen fragments selected in ExoSpec have also been previously reported as CE-MS biomarkers for discrimination of prostate cancer patients from those without malignancy[129] and also for detecting significant prostate cancer[211]. Importantly, not all significant features from the individual comparator models added value to the integrated ExoSpec model. These observed differences could be attributed in part to redundant information shared between the multiple methods of appraising disease status of the prostate.

Application of LASSO-penalised general linear model for feature selection is arguably suboptimal for avoiding overfitting through feature selection, even within 20-fold cross-validation as the variability seen in Chapter 5 between data splits is still likely to have an influence. I would recommend a further, more targeted study examining a much smaller selection of peptides quantified by CE-MS before embarking on a costly validation study of ExoSpec.

7.6 Conclusions

Developed in this chapter, ExoGrail joins ExoMeth as a promising example of what is achievable by the FrameWork using data from more targeted analyses. ExoGrail outperformed both clinical standards and the individual sources of biomarkers in isolation, with predictive accuracy that could result in sizeable changes to the patient journey for those suspected to have prostate cancer. The interactions observed between EN2 protein levels and the cf-RNA *SLC12A1* gene-probe show that by integrating information from multiple sources can result in risk models with increased utility (Figure 7.4). Furthermore, the use of a machine learning algorithm such as Random Forests that can natively account for both the non-linearity of each biomarker and their interaction means no additional data or prior knowledge is required to model those relationships.

However, the FrameWork as conceived does not represent a singular solution for all types of biodiscovery datasets, where it was unable to handle very large featuresets without unreasonable amounts of computation. With alterations to the feature selection process by using cross-validated and penalised generalised linear models it was still possible to perform robust feature selection in a dataset of over 800 parameters. With only 20-folds for cross-validation, this process will not be as robust to overfitting as Random Forest based methods that employ resampling, nor able to consider as many features. A key limitation of the LASSO when dealing with more predictors than observations ($p > n$) is that the LASSO, at most selects n features, and may not be stable[212]. This does not appear to be an issue encountered here, where strong *a priori* filtering was employed to further reduce the number of peptides considered from >30,000 down to 643, with the final ExoSpec model retaining 12 variables for final modelling. This result would benefit from further experimental work, considering a more targeted set of peptides, perhaps by generating a list with less strong regularisation.

Both ExoGrail and ExoSpec models require validation studies before clinical use can be considered in any meaningful way, with ExoSpec arguably requiring more careful validation

to further ascertain whether the features selected here are indeed generalisable to a larger cohort whilst examining fewer peptides. Implementation of a test based on data integration across multiple assays poses a technical challenge for implementation. Requiring additional, potentially complex and disparate assay methods increasing the risk of technical failures (false positive or negatives from a single assay), logistical complexity and increased material costs. This is something that requires consideration when moving towards validation and commercial implementation. A study of the health economics and feasibility of implementation would be wise, though this falls outside the scope of this thesis, that concentrates on the analytical aspects of prognostic biomarker development. In the next chapter the design of a large, multicentre prospective collected clinical trial will be considered, with the explicit aim of easing the journey to clinical use for the developed risk models presented in this thesis.

Chapter 8

Discussion

I have demonstrated in this thesis that clinically useful information is available from urine where prostate cancer is concerned. Coupled with the ease of non-invasive sampling, urine biomarkers could be responsible for a considerable change to both the risk stratification of patients suspected to have prostate cancer, and the long-term monitoring of those with low risk disease. Consideration of multiple biological aspects within urine, from whole urine through to the cell-free and cell pellet fractions, can result in considerably more predictive information being uncovered when analysed carefully. Investigations such as these are only really possible with large, collaborative studies such as the Movember GAP1 project, where the specific expertise of collaborators allows a wide biodiscovery net to be cast. Unfortunately in this case due to practical constraints, not all samples were assayed by all methods, making overlaps between datasets somewhat limited. Regardless, considerable value was shown to be within the GAP1 datasets through the development of several promising prognostic risk models, research publications, and the generation of two pieces of intellectual property.

8.1 Results from this thesis

Chapter 4 described the initial development of one such risk model called PUR, integrating information from NanoString data and using a training/test data splitting strategy to fit a LASSO-penalised constrained continuation link ratio regression model. PUR compared favourably to other published urine tests, with additional utility in the apparent prediction of long-term outcomes in a small AS sub-cohort. More detailed investigation showed that this ability may be driven by cohort-specific effects rather than predictive utility of the model itself, requiring further study to definitely answer which is the case. Prognostication of active surveillance patients via non-invasive sampling could meaningfully change how patients with lower risk prostate cancer are monitored, and hopefully reduce the rates of self-election for treatment that can be reported reach over 30%[69]. Of course, with better triage tools that can save an unnecessary biopsy, less indolent disease overall would be diagnosed and therefore, fewer patients being enrolled onto AS.

Biological data are commonly highly variable, derived from complex systems with many diverse interactions and the data are characterised by multiple dependencies and interactions[213]. This complexity is increased even more so in observational cohort-based biodiscovery studies such as the GAP1 project, where experimental design and manipulation of variables cannot be altered to compensate for variance. The reduction of variance

in the data is absolutely key where predictive analytics and the development of clinical risk models are concerned. Careful consideration of modelling strategies can, and does, have a large effect on a model's predictive ability. Explored in Chapter 5, the effects of altering the machine learning algorithm and training labels were quantified, where ensemble algorithms presented the best solution to capturing the most amount of variance. The Random Forest algorithm consistently outperformed both LASSO-based linear regression analyses and gradient boosting machines using decision trees. The benefits of the Random Forest algorithm are many, including the inherent use of bootstrap resampling and aggregation of results across many weak learners to provide a single strongly internally-validated model without any additional user input[142]. This makes the algorithm ideal for biodiscovery applications, where there are often no strong prior hypotheses or assumptions concerning expression patterns or variable distribution.

Incorporating the findings from Chapter 5, further analyses were undertaken in Chapters 6 and 7, developing a robust feature selection and machine learning framework and applying it to the NanoString dataset alongside methylation, ELISA, and proteomic datasets. The developed mulitmodal risk prediction models revealed even greater clinical utility for biopsy reduction and risk stratification than PUR. Where prediction of disease status is the priority, rather than understanding the pathobiology of prostate cancer, consideration of multiple modalities within urine samples alongside clinically available parameters results in better models. Of course, this somewhat removes biological interpretability, but this is often more of a concern for studies concentrating on more basic research rather than the translational outcomes explored here. Regardless of this limitation, it was shown that the *GJB1* cf-RNA gene-probe was frequently the most important variable for predicting biopsy outcome, a novel finding not reported before for prostate cancer. Multiple markers previously associated with prostate cancer were additionally selected with widespread agreement across each of the models, including the *TMPRSS2/ERG* gene fusion and *PCA3* cf-RNA probes from NanoString data. The clinical relevance of serum PSA for disease identification was also confirmed and improved upon when considered with multiple other variables to account for the non-specificity of PSA alone that is widely documented[5,160,214].

Each of these models compared favourably the results reported by existing urine tests, including the now validated SelectMDx and ExoDx Prostate Intelliscore tests. Where PUR equalled their reported discriminatory ability determined by AUC for predicting Gleason $\geq$ 7 (SelectMDx = 0.81, $n = 715$ with a PSA < 10 ng/mL[100], ExoDx = 0.71, $n = 519$ with PSA 2 - 10 ng/mL[103]), it proved to have potentially novel predictive use in active surveillance that as of yet, has not been matched. The ExoMeth and ExoGrail models exceeded both aforementioned tests, as the integration of multiple modalities had a synergistic effect on biopsy prediction. To the best of my knowledge there are currently no other urine-based tests that assay multiple aspects of the prostate for identifying disease status, though a combination of methylation and miRNA in tissue has been shown to add value for predicting biochemical recurrence[195]. If implementation of these models as a clinical test is deemed practical and validated, they could have a sizeable impact on patient care in the UK.

8.2 Potential clinical impacts

Almost 50,000 cases of prostate cancer are diagnosed each year in the UK, with up to 80% of these PSA-detected cancers being clinically irrelevant[1]. Without treatment, they would never have caused symptoms or endangered the life of the patient. Treatment with curative

intent, such as surgery, chemotherapy or radiotherapy, should be focused on patients with clinically significant disease, not those with indolent forms of prostate cancer in order to save them from the potentially life-altering side-effects of treatment. Concentrating on removing these patients from the clinical pathway earlier on, rather than identification of the most at-risk would have an impact on larger numbers of men, and greater savings in healthcare systems.

The main impacts of the results from this thesis will be for those patients with a clinical suspicion of prostate cancer (raised PSA, lower urinary tract symptoms etc.). Implementation of a urine test for triaging these patients prior to any invasive biopsy would remove a sizeable amount of stress and worry for the patient. In the medium term, and with further causal research the personalised molecular-level information for patients has the potential to more accurately inform clinician decisions about why the disease may require treatment. With suitable validation and clinical adoption of tests, the long-term impacts of this work would be evidenced in policy and guideline changes, particularly those of NICE and the European Association of Urology. With updated guidelines and subsequent reduction in the number of biopsies performed, economic impacts could be seen in reduced material and clinician costs, especially within expensive tertiary care settings.

8.3 Requirements to realise this impact

In order to realise the potential impacts, all of the models reported in this thesis require robust external validation in order to generate suitable levels of evidence to allow prospective treatment of patients based on their urinary molecular profiles. Strong internal validation will always best weak-external validation, as evidenced by the TRIPOD guideline's preferences for type 1b analyses over type 2a[21]. Whilst the combination of internal validation methods such as cross-validation and resampling with external validation in one study design is the ideal setting, this isn't often feasible and instead compromises have to be made. In the case of the work presented in this thesis, the decision to use resampling and not use weak validation methods was made to avoid caveats being placed on models and producing potentially overly optimistic results.

There is no single solution to guarantee validation and subsequent adoption of any clinical risk model, as Rittenhouse *et al.* described in their journey for the first breakthrough urine test to receive FDA approval, it takes years and a constant awareness of the challenges that must be faced[92]. Planning for future regulatory and practical challenges at each stage of study design should be prioritised above many other immediate goals to avoid endless iterations of studies. Moving carefully and with purpose when study design is considered greatly increases the chances, not only of successful validation, but of the evidence generated be acceptable to clinical bodies like NICE and the EAU. Even in the case of a failed validation, careful study design allows for the collected data to have secondary and tertiary uses in new model development and updating, or for more practical purposes such as assay optimisation. In the next chapter I will describe the design of a validation study that also formed part of a successful grant application to Prostate Cancer UK that has already begun sample collection.

Chapter 9

Future Work

9.1 Summary

The design of future studies with the express aim to validate the models developed during this thesis, including options to update or calibrate models if needed are described in this chapter. The evidence levels presented by the TRIPOD guidelines are considered again to provide the rationale for a multi-centre cohort study including both prospective and retrospective curation of cohorts. A carefully curated retrospective cohort of patients from the Norfolk and Norwich University Hospital will be used as an initial validation and exploration cohort. The data from here can be used to assess the calibration of multivariable risk prediction models against very well characterised patients and if required, recalibrate models without the risk of losing validation potential.

Following the internal calibration and validation of models, multiple external cohorts will be prospectively collected from urology clinics around the UK. Samples will be collected from patients as they present at clinics in order to realistically capture a "snapshot" of disease proportions in a real, local population. Each external cohort can then be used as a validation cohort of their own, or if samples numbers do not allow this, pooled and assessed as one larger cohort with blocking for collection site. The collection of samples from an active surveillance (AS) cohort is also described, with the aim to more definitively answer how predictive of outcome the PUR model truly is. The curation of this AS dataset will be very finely controlled, with patients very well characterised and followed for at minimum five years. Following this observation period the predictive utility of the PUR model or D'Amico status can be assessed. If neither is deemed to be prognostic at five years, the dataset can be used to train a time-dependent survival model incorporating NanoString data, specifically for use in AS cohorts. This of course would then require further validation in a new study over multiple years.

9.2 Introduction

Fewer than 1% of published cancer biomarkers see clinical adoption[16]. This is to be expected somewhat, considering most initial publications are academic in nature, predominantly describing a discovery. Additionally the research aims and scope at the start of early-stage biodiscovery trials can be ambiguous as to the key decisions required to see any new discovery through to clinical validation. Quality control rules, patient populations to collect, inclusion/exclusion criteria, data processing, and early interruption for preliminary analyses

all have to be considered with one eye on validation in the future[16,21,215]. As Rittenhouse *et al.* describe with the PCA3 test, the time-scale over which these need to be considered can be over a decade before clinical acceptance of a biomarker[92]. Obviously it is not realistic to expect to simultaneously develop, validate, and gain regulatory approval for a clinical test for prostate cancer in a single study, instead multiple studies are required to generate strong evidence of utility and surety in individual predictions from a risk model before adoption can be considered. However with careful thought and design, it is certainly possible to reduce the number of repeated studies required to develop and validate a urine test for prostate cancer[92].

9.2.1 Compliance to TRIPOD guidelines

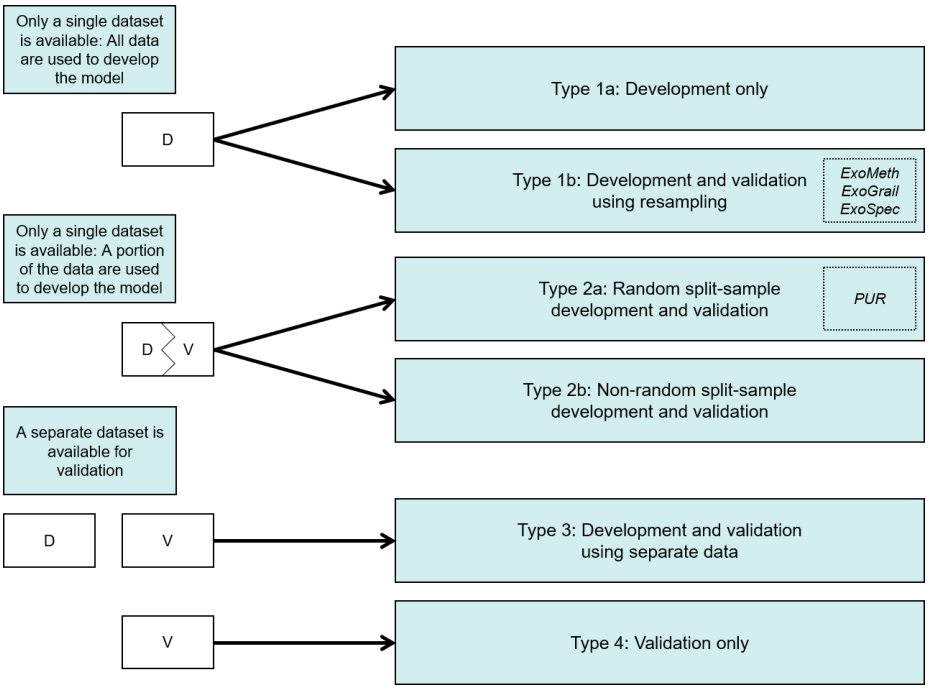


Figure 9.1: Types of prediction model studies covered by the TRIPOD statement. D = development data; V = validation data. Models described within this thesis are italicised. Adapted from the TRIPOD Statement

Considering the hierarchy of analysis types described by the TRIPOD guidance briefly covered in Chapter 1, we can examine where the models presented in this thesis fall (Figure 9.1). Where PUR meets the requirements of a Type 2a study using a randomly selected split of data to create development and validation datasets, the models developed by application of the FrameWork fall into Type 1b analyses (Figure 9.1). Type 2a analyses are generally not recommended nor necessarily better than Type 1b as they generally lead to a lack of power[181,216,217] and, as observed in Chapter 5, can result in highly unstable model performance dependent on the random split of the data. Instead, the authors of the TRIPOD guidelines recommend Type 1b analyses using internal validation methods such as resampling as a prerequisite for model development, especially where data are limited[181,217,218].

For the models developed during this thesis the next step will be to perform Type 4 validation analysis; with the sole aim being to validate each of the models. In theory

this simply requires enough samples need be collected to fairly evaluate the predictive performance of the pre-specified models in new data. However, there are certain drawbacks and inefficiencies to such a simple design. Updating of models cannot be achieved without an additional study[217], furthering the costs and time taken before potential clinical adoption. Additionally, whilst the reduced data collection requirements of a validation only study are welcome, they limit the usefulness of the data beyond validation, where more rigorous collection of data could see new models developed regardless of validation outcome.

With the above considered, in the following sections I will describe the process and rationale for a multi-centre study, with samples collected and analysed in such a way to allow for recalibration or updating if necessary. Additionally, if recalibration of models is not required, the results of the suggested study can be used to simply provide stronger evidence of clinical utility in further external cohorts. One reason for models requiring updating may be that as data from the GAP1 study were collected some time ago, analytical or clinical methods and assays have changed since initial sample collection, such as the routine use of mpMRI as a first-line triage tool. Our sample collection procedure has also been updated and improved, resulting in an at-home collection protocol that results in more stable samples from patients, improving the quality of extracted RNA and avoiding the requirement of a DRE[219]. These changes will alter both the underlying distribution of molecular patterns in the data, and the reported proportions of disease within cohorts, which will have an unknown impact on model stability and needs to be assessed before embarking on external validation.

9.2.2 Goals of future studies

> "A validation study has a specific goal: quantifying the performance of a model in other data." — TRIPOD Statement[21]

Validation of developed models is the primary aim of the designed studies described here. However, there are multiple secondary objectives that if achieved, substantially improve the likelihood of realising a urine test for prostate cancer:

- Validate models across *multiple* external centres using existing models, as published.
- Ensure models are suitably calibrated, with predicted risk scores matching disease proportions in a well-characterised population.
- Compare prognostic ability to current clinical standards including multiparametric MRI.
- Collection of additional parameters to enable calculation of popular nomograms and risk calculators.
- Prove predictive ability in active surveillance usage and/or enable development of a specific time-dependent prediction model

Patient recruitment across multiple different cohorts and incorporating differing experimental/recruitment designs are required in order to achieve these goals. For example, suitable calibration of models needs a very carefully curated and well-characterised patient cohort that would be retrospectively collected. On the other hand, any external validation should occur using prospectively collected samples that accurately represent the real patient population seen within the healthcare system being used. The underlying theory in considering such a rigorous multi-cohort design at this stage allows for an immediate transition from successful model validation to large randomised control clinical trials (RCT). Whilst a superiority RCT is several years away at the earliest, it is envisaged that the intervention

would put patients forward for biopsy, or remove them from the treatment pathway, based upon their calculated risk score from a validated model, whilst the control would be the currently implemented clinical standards.

9.3 A three cohort design

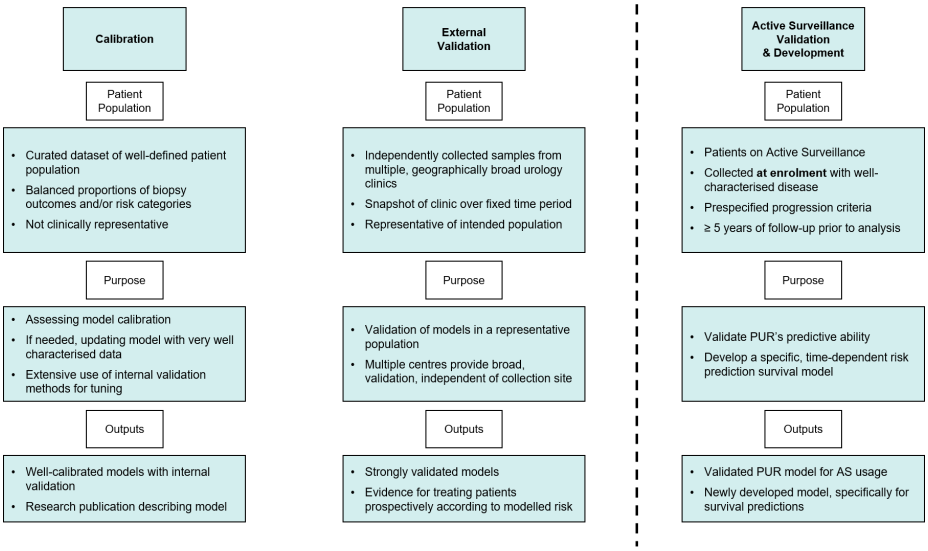


Figure 9.2: Broad overview of the three cohorts to be collected as part of a future validation study. The Active Surveillance cohort is collected and analysed entirely separately from the other two cohorts, with five years of follow-up prior to commencing analysis.

Three main cohorts are to be considered in the proposed studies; a model calibration cohort, a multi-centre external validation cohort and an active surveillance-specific validation and model development cohort (Figure 9.2). Each arm of this study is designed to achieve differing goals, again with the primary aim to produce strongly validated multivariable risk prediction models, robust to changes in study population and ready for large-scale superiority RCTs. Samples across all cohorts will be assayed by the same methods: collecting cf-RNA data using NanoString, with ideal data collection also including whole urine EN2 levels, and urinary cell-pellet methylation.

9.3.1 The Calibration Cohort

The calibration of risk model estimates is described as the agreement between estimated and observed number of events in a patient population[220]. Similar to assessing predictive utility, models can be calibrated on several levels, in the mean, weak, moderate, or strong sense as described by Van Calster *et al.*[220]. Calibration is crucial as poorly calibrated models can lead to misleading predictions such as a systematic overestimation of risk, which leads to overtreatment[221]. Undertaken at the validation phase in new data, the reporting of calibration performance is recommended the TRIPOD guidelines[21], as multiple systematic reviews have found that calibration is reported far less often than discrimination[222–226].

Van Calster *et al.* provide an excellent review on the available specific analytical techniques for assessing, avoiding, and correcting poor calibration[227].

Two main sources are thought to affect model calibration, either related to the model/algorithm in question, or to an external force such as the patient population[227]. Patient populations tend to change over time naturally as clinical practice, referral patterns or healthcare policies change[228,229]. These changes can lead to alterations in the prevalence/incidence of clinically significant disease compared to what was originally modelled, resulting in poor calibration. As mentioned, the Movember GAP1 cohort was collected several years ago, prior to the widespread adoption of mpMRI as a triage tool[5] which alone makes calibration a key task moving forwards. The second set of potential causes for mis-calibration relate to algorithmic and statistical methodology choices. Statistical overfitting is common[227], where overfitted predictions capture too much random variance in the original data. Though great care has been taken through this thesis to avoid such by widespread use of resampling, statistical overfit cannot be ruled out with the small datasets used, again making calibration absolutely necessary.

With an idealised disease with a perfect ground-truth, diagnosis or outcome is known definitively. For example, a recording of death in the UK is definitive, with no ambiguity[230]. As discussed before, this is not the case where prostate cancer is considered. Large uncertainties surround a negative biopsy outcome if using TRUS-guided needle biopsy, and even in the case of a cancer-positive finding, due to the low sampling rate specific disease burden can be under- or over-reported when compared to radical prostatectomy[47]. With this considered, the goal of the Calibration Cohort is to assess the calibration of developed models in a very well-defined patient population, carefully curated to remove as much uncertainty about patient status as is reasonably possible. Patients will be retrospectively recruited to this cohort from the local Norfolk and Norwich University Hospital (NNUH), where inclusion criteria will include:

- Extensive template-derived biopsy information, including detailed histopathological analysis
- True multiparametric MRI (comprising information from T1 and T2 weighting, diffusion-weighted imaging and dynamic contrast enhancement).
- PSA <50 ng/mL
- Eligible for treatment, no significant co-morbidities.
- Not an extreme outlier in clinical risk category:
  - For example a patient presenting with a PSA of 45 ng/mL but Gleason 3+3 on biopsy

Collection of this cohort allows for detailed investigations into where a model may perform well, or report erroneous results, including evaluating associations of clinical measures to molecular patterns. This may allow for a deeper insight into the biology of prostate cancer, and the links between the pathobiology and clinical presentation to be investigated. Such a detailed cohort has not been previously described and would be valuable for future study, regardless of how well models are calibrated.

9.3.2 The External Validation Cohort and sub-cohorts

Assuming models have been accurately calibrated and updated if required in the Calibration Cohort, true validation can then be considered in external cohorts. Of course, collection of these samples can be concurrent with the Calibration Cohort, though once analysed or

assessed for predictive utility, cannot be used again for validation without the introduction of both concious and unconscious bias[21]. Therefore the Validation Cohort can be considered a "single-use" dataset, making it wise to be sure of accurate models before approaching validatory analyses.

Validation over multiple smaller sub-cohorts provides stronger evidence of clinical utility than a single large cohort, as each provides evidence commensurate with that of a TRIPOD level 4 analysis (Figure 9.1). These sub-cohorts can form "narrow" or "broad" validation, dependent on how they are collected[231]. Narrow validation could be achieved through the successful external validation of a model in a similar setting or population to the one it was derived or updated in[231]. In this case, that would be in the form of further NNUH-based cohorts, that would be temporally separated from the Calibration Cohort. This would not be ideal for multiple reasons; it requires some unspecified length of time to pass between cohort collections, collection naturally takes longer, and any bias that may be present due to geographical or socio-economic factors, or status as a university hospital remains. Broad validation is therefore the far more reasoned option, externally validating a model in multiple varied settings and populations, where successful validation provides evidence that predictions from a model can be confidently used in future patients, regardless of setting[21,231]. Additionally, the collection of samples can be coordinated in parallel, scaled according to funding and time constraints, or add additional centres if and when required with relative ease.

The External Validation Cohort presented here comprises three or more sub-cohorts, each operating entirely independently from one another. Each sub-cohort will be collected from a designated centre as and when patients present at a clinic. This "snapshot" recruitment is designed to ensure that the collected samples from each centre are representative of disease proportions reported at that centre. A potential weakness of this is an underlying assumption of roughly equal recruitment rates and losses to follow-up at each centre. Models can still be validated if this assumption is broken, either by reporting results with a caveat, or through over-recruiting patients from affected clinics. Laboratory analysis of samples should be mixed rather than batched, to ensure any laboratory batch-effects are equally spread across sub-cohort samples.

All patients recruited will have to be deemed "eligible for biopsy"; criteria set by each centre in isolation and agreed upon prior to commencing any sample collection, fixed in place for the duration of sample collection. Allowing for local variations such as slightly different PSA thresholds or triage processes, only further strengthens the evidence generated by successful validation, showing that non-random variation between centres can be handled by the model being validated. Relatively minimal clinical criteria are required to be collected at this stage in order to evaluate model performance against clinical endpoints: mpMRI outcome, Gleason score, serum PSA and patient age are all that is needed. Again, once these samples are assessed to validate predictive performance, they can only be used for development of new models and further updating, where the variance between centres may not be ideal for stable model development, and specifying collection centre as a variable would only serve to limit options for validation.

9.3.3 The Active Surveillance Validation and Development Cohort

Entirely separate from the previous two cohorts designed primarily for the validation of models predicting biopsy outcomes, the AS Validation and Development Cohort serves two purposes. The main aim with this prospectively collected cohort is to answer the questions

that arose in Chapter 4, namely whether the predictive ability of PUR to predict outcome five years in advance of progression is driven by predicted risk from the PUR model itself, or from cohort-specific effects driving the underlying outcome, where the patients that progressed were predominantly D'Amico Intermediate Risk and therefore inherently had higher PUR4 risk scores.

Patients recruited into the AS Validation and Development cohort will be done so retrospectively, following enrolment onto an AS programme at the NNUH. Specific enrolment criteria will be set by the consultant clinician overseeing the programme, though all patients will have received extensive mpMRI scans, and have serum PSA levels assessed at regular intervals for inclusion. Progression criteria will be similarly set by the attending physician, recorded, and ideally fixed in place for the duration of the study, with self-elected treatment recorded as a loss to follow-up. The cohort will be monitored for a minimum of 36 months, preferably waiting five years before outcomes are assessed to ensure a suitable number of progression events are recorded. If feasible, urine samples could be collected at regular intervals along with PSA and mpMRI data, which would both allow for further, more detailed measurement of model stability over time, and for deriving multi-state and interval survival models[232] for outcomes other than a binary progression/non-progression.

A secondary use of this cohort is in the construction of an AS-specific, time-dependent survival risk prediction model. As discussed in Chapter 4, there are no personalised medicine tools available for making time-series predictions of survival outcomes for AS patients. Given the apparent predictive ability of PUR in AS despite not being designed to do as such, it's hypothesised that the development of a model with the explicit goal of time-dependent prediction is possible with suitable data. Suitable methods for developing a model capable of returning a predicted time interval for survival outcome are an ongoing area of research in machine learning[233]. Overall survival predictions may be produced using methods similar to those employed in this thesis, including coxnet models using the elastic net framework[234], analogous to the LASSO penalised ordinal regressions of Chapter 4 and Chapter 5, or random survival forests, adapted from the Random Forest algorithm[235].

9.4 Comparisons to clinical standards and calculators

The Calibration and Validation Cohorts share the same key endpoints for assessing model performance in the discrimination of binary biopsy outcomes of Gleason $\geq$ 3+4 and $\geq$ 4+3, as reported during their development. mpMRI marks a new addition to clinical practice since data were originally collected, and can be used in two ways; as an outcome in itself to be predicted by models, or as direct competition to predictive models for predicting biopsy outcomes. Given the cost and time intensive nature of mpMRI[58], if a simple non-invasive urine test could provide similar ability to PIRADS scores for patient triage prior to a biopsy, it could result in large savings to healthcare systems.

As was discussed in Chapters 6 & 7, predictive utility in isolation means very little without a point of reference to current standards of care. So far in this thesis this has been achieved by constructing new clinical models from available data. There is an argument that these models themselves require validation before considered true reflections of clinical practice. Instead, with the additional of a few clinical variables, comparisons to already validated nomograms and calculators can be made. For example, with the inclusion of free PSA and -2proPSA measurements (a biologically inactive precursor to PSA[236]), the Prostate Health Index (PHI) can be compared against. The PHI is primarily used to

predict the likelihood of high grade, Gleason >7 disease on biopsy[237], but has proven utility in predicting progression likelihood in Active Surveillance[238]. This would make the PHI an ideal comparator within the Calibration and Active Surveillance Cohorts suggested here. As the Validation Cohort and sub-cohorts are designed to be as minimally intrusive on local practices, it is not feasible to request extra information that is not routinely collected.

Instead, where full external validation is concerned the measures likely to be validated against would include PIRADS scores derived from mpMRI information, biopsy outcome or clinical risk category from pre-biopsy PSA levels and simpler clinical tools such as the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC)[239]. Whilst not perfect, the PCPTRC is simple, and uses regularly collected variables of race, age, PSA, family history, DRE impression and prior biopsy history to calculate a likelihood of significant (Gleason $\geq$ 7) cancer on biopsy, with proven clinical utility[239]. In the original Movember GAP1 cohort collection of these parameters was attempted, but many reported high levels of missingness (e.g. >75% for family history). Extra attention will need to be made in future studies, with quality control procedures in place at the time of data-entry.

9.5 Sample sizes

Sample size calculation for developing risk prediction models is not trivial. Datasets must be "big enough", capturing the population variance well enough to be of use when applied to new individuals. Traditional power calculations do not work as there are no estimates of effect size or statistical power to consider so instead, various blanket "rules of thumb" have been proposed, debated and debunked[240–245]. Riley *et al.* described a sound methodology in 2020 based on sound statistical practice and considering the number of variables, the study population, and the margin of acceptable error in a prediction[246]. Unfortunately, the methods focus primarily on regression-based algorithms and models, making it mostly inappropriate for our use here and we have to predominantly focus on feasibility and practical constraints.

9.5.1 The Calibration Cohort

Collection of this cohort will be the most resource intensive, but also the most crucial that it is undertaken thoroughly. Curation of balanced clinical categories will require a longer recruitment time as very low and high risk patients will present at urology clinics with lower frequency than those with intermediate disease. ExoMeth, ExoGrail and ExoSpec were all developed in approximately 200 patient samples, whilst PUR was developed with 335 patients. Consideration of how stable the feature-sets were for each of these models showed that even with extensive resampling, the selected variables were very stable (Chapters 6 & 7). Therefore the proposed size of the Calibration Cohort is 400 patients. This not only allows for assessment of calibration and updating, but if curated carefully enough, could be used for future translational and basic research studies, maximising the value of the data.

9.5.2 External Validation sub-cohorts

Designed as a "snapshot" of each clinic, recruiting patients as they present, a minimum of 100 patient samples per centre is recommended, though this would be guided by local practices as to how many samples they can collect. A minimum of three external centres would result in a total validation cohort of 300 patients at the very least, but further centres

would be desirable. As each centre operates in isolation to one another, it is also possible to add additional centres as the project continues, and if funds allow.

9.5.3 AS Validation and Development Cohort

The time-dependent nature of active surveillance makes collection of large, well-described cohorts difficult without extensive collaboratory efforts such as those seen in the Movember GAP3 AS study of over 15,000 patients on AS programs[247]. Access to this cohort would be an ideal goal following successful validation of PUR, or the development of a specific time-dependent AS model as described above.

In the earliest instance, local collection and follow-up of patients from the NNUH will suffice, as enrolment, monitoring and intervention criteria can all be tightly controlled with the help of local clinical collaboration. With this considered, and from personal communications with NNUH clinicians, it would not be unreasonable to recruit 200 patients. These would be closely monitored for a minimum of 36 months, with rolling enrolment and analysis once reaching 36 months of observation.

9.6 Discussion

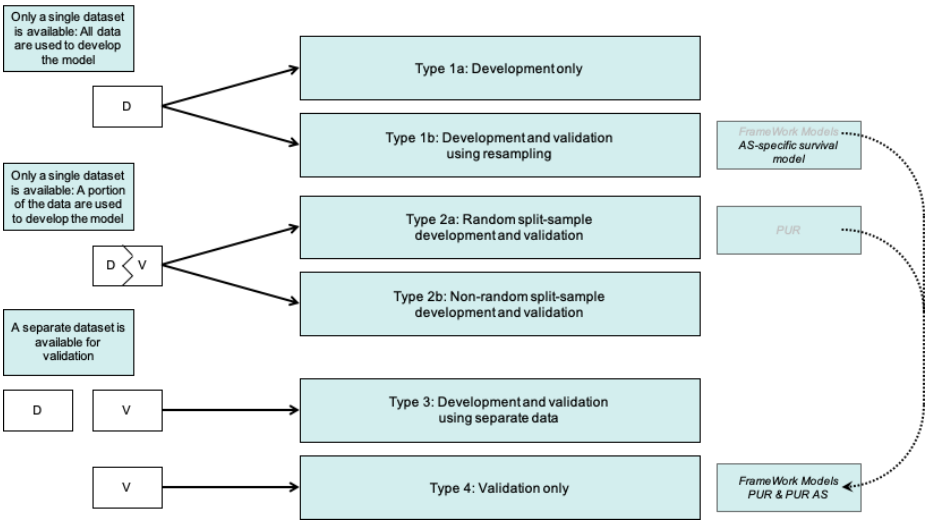


Figure 9.3: Evidence generated by successful completion of the proposed trials. Models in grey represent the current status of validation, with their updated counterparts in black. D = development data; V = validation data. Models described within this thesis are italicised. Adapted from the TRIPOD Statement

Successful completion of the proposed studies could result in numerous outputs, including research publications, generation of intellectual property and TRIPOD type 4 validated models ready for large RCTs (Figure 9.3). This design formed part of a successful grant application to Prostate Cancer UK, where sample collection has begun. Further to this, the collection of urine samples from the very large Movember GAP3 active surveillance study has also begun, with access to over 15,000 patients across the globe[68].

Adherence to the structure of the proposed trials would hopefully avoid the "over-promise, under deliver" of many cancer biomarkers, where it is easier by far to design analyses **to** TRIPOD guidelines rather than **adapting** the reporting of results to fit afterwards[21]. An important consideration is that the data generation process cannot be materially changed, for example through assay optimisation or changes to the sample collection process. Alterations to the data generation process change the underlying variance and distribution of the data being modelled, removing any certainty that future predictions will be reliable. If clinical adoption of models and improvements to patient care are truly the end-goal, then the allure of constant iteration and methodological improvement must be resisted.

Appendix A

Chapter 5:

Table A.1: Table of Boruta decisions for each variable with at least one training label decision rendered as "Tentative" (?) or "Confirmed" (✓). The total number of confirmed or tentative decisions are recorded, as well as whether the variable in question appears in the PUR model previously described. Variables rejected for individual training labels are shown with 7

Variable	Total	In PUR Model	TriSig	Gleason ≥ 4+3	Gleason ≥ 3+4	D'Amico	Any Cancer	Cancer vs High Risk Cancer
Age	6	No	✓	✓	✓	✓	✓	✓
CACNA1D	6	No	✓	?	?	✓	?	✓
ERG 3 ex 4 5	6	Yes	✓	?	✓	✓	✓	✓
GABARAPL2	6	Yes	?	?	?	✓	?	✓
GJB1	6	No	✓	✓	✓	✓	✓	?
HOXC6	6	Yes	✓	✓	✓	✓	✓	?
HPN	6	Yes	✓	✓	✓	✓	✓	✓
KLK4	6	Yes	✓	✓	?	✓	✓	✓
MME	6	Yes	✓	?	✓	✓	✓	✓
PPAP2A	6	No	✓	?	?	?	?	✓
PPFIA2	6	Yes	✓	?	✓	✓	✓	?
PSA	6	No	✓	✓	✓	✓	✓	✓
RAB17	6	No	✓	✓	✓	✓	✓	✓
RPL18A	6	No	✓	?	✓	✓	✓	?
SLC12A1	6	No	✓	✓	✓	✓	✓	✓
SPINK1	6	No	✓	?	✓	✓	✓	✓
TMPRSS2 ERG fusion	6	Yes	✓	?	✓	✓	✓	?
AMACR	5	Yes	?	✓	✓	?	✗	✓
DPP4	5	Yes	✓	✓	✓	?	✗	✓
EIF2D	5	No	?	✓	✓	?	✗	✓
ERG3 ex 6 7	5	No	✓	✗	✓	✓	✓	?
GAPDH	5	Yes	✓	?	✓	✗	✓	✓
HIST1H1C	5	No	?	✓	?	✗	?	✓
HIST1H1E	5	No	✓	?	✓	✗	✓	✓

Table A.1: Table of Boruta decisions for each variable with at least one training label decision rendered as "Tentative" (?) or "Confirmed" (✓). The total number of confirmed or tentative decisions are recorded, as well as whether the variable in question appears in the PUR model previously described. Variables rejected for individual training labels are shown with 7 *(continued)*

Variable	Total	In PUR Model	TriSig	Gleason ≥ 4+3	Gleason ≥ 3+4	D'Amico	Any Cancer	Cancer vs High Risk Cancer
IFT57	5	No	✓	✓	✓	?	✗	✓
KLK3 PSA exons1 2	5	No	✓	?	?	?	✓	✗
MED4	5	Yes	?	?	?	✓	✗	✓
MEMO1	5	Yes	✓	✓	?	?	✗	✓
MEX3A	5	Yes	✗	?	✓	✓	✓	?
PCA3	5	Yes	✓	?	✓	✓	✓	✗
PECI	5	No	✓	?	?	✗	✓	?
RPS10	5	No	✓	✓	✗	?	✓	✓
SIM2 long	5	No	✓	?	✓	✗	✓	✓
SIM2 short	5	Yes	?	✗	✓	?	?	✓
SMAP1 ex 7 8	5	No	✓	?	?	?	?	✗
SRSF3	5	No	?	✓	?	✓	✗	✓
TDRD1	5	Yes	?	?	✓	?	✓	✗
TRPM4	5	Yes	✓	✓	✓	✓	✓	✗
UPK2	5	Yes	?	?	✓	?	✓	✗
HIST32HA	4	No	✓	✓	✓	✗	✓	✗
HOXC4	4	No	?	✗	✓	?	?	✗
KLK3 PSA exons2 3	4	No	✓	?	✗	?	✓	✗
MIR146A DQ658414	4	No	?	✗	✓	✓	?	✗
RP11 244H18 1 P712P	4	No	?	✓	?	✗	✗	✓
SNORA20	4	No	?	?	✗	✗	?	?
STEAP2	4	No	?	?	✗	?	✗	?
STOM	4	No	?	✓	?	✗	✗	✓
SULT1A1	4	Yes	?	?	✗	?	?	✗
TBP	4	No	✓	?	✓	✗	✗	✓
TERF2IP	4	No	?	✗	?	?	✗	✓
TMEM45B	4	No	?	?	✗	✗	✓	✓
ZNF577	4	No	✓	?	✓	✗	✗	✓
Alcohol unknown	3	No	✓	✗	?	✗	?	✗
AMH	3	Yes	✗	?	?	✗	✗	?
Amount RNA in ng	3	No	✓	✗	✓	✗	✓	✗
ANKRD34B	3	Yes	?	?	✗	✗	✓	✗
APOC1	3	Yes	✓	✗	✓	✗	✗	✓
B2M	3	No	✓	?	✗	✗	?	✗
CDC37L1	3	No	?	✗	?	✗	✗	✓

Table A.1: Table of Boruta decisions for each variable with at least one training label decision rendered as "Tentative" (?) or "Confirmed" (✓). The total number of confirmed or tentative decisions are recorded, as well as whether the variable in question appears in the PUR model previously described. Variables rejected for individual training labels are shown with 7 *(continued)*

Variable	Total	In PUR Model	TriSig	Gleason >/ 4+3	Gleason >/ 3+4	D'Amico	Any Cancer	Cancer vs High Risk Cancer
DLX1	3	No	?	✓	✗	✗	✗	?
Family History no	3	No	✓	✗	?	✗	✓	✗
HMBS	3	No	?	?	✗	✗	✗	?
ITGBL1	3	Yes	✗	✗	✗	?	?	?
MARCH5	3	Yes	?	✗	✗	✗	?	?
MMP11	3	Yes	✗	?	✗	✗	?	?
MSMB	3	No	✗	✓	?	✗	✗	✓
NKAIN1	3	Yes	✓	✗	✗	✗	✓	?
PALM3	3	Yes	✗	✗	?	?	✗	?
PDLIM5	3	No	✓	✗	?	✗	✓	✗
PPP1R12B	3	No	✗	?	?	✗	✗	?
RPLP2	3	No	✓	✗	?	✗	✓	✗
SERPINB5 Maspin	3	No	?	✗	✗	?	✓	✗
SLC43A1	3	No	?	✗	✓	✗	?	✗
SMIM1	3	Yes	✗	?	✗	✗	?	✓
STEAP4	3	No	?	?	✗	?	✗	✗
ANPEP	2	No	✗	?	✗	✗	✗	✓
ARexons4 8	2	Yes	?	✗	✗	✗	✓	✗
AURKA	2	No	✗	?	✗	✗	✗	?
BRAF	2	No	?	✗	✗	✗	✗	?
CAMKK2	2	No	?	?	✗	✗	✗	✗
CTA 211A9 5 MIATNB	2	No	?	✗	✗	✗	✗	?
ERG5	2	No	✓	✗	✗	✗	✓	✗
Family History unknown	2	No	✓	✗	✗	✗	?	✗
FDPS	2	No	?	✗	✗	✗	✗	?
FOLH1 PSMA NAALAD1	2	No	✓	✗	✗	✗	✓	✗
HIST1H2BF	2	No	✗	?	✗	✗	✗	?
IMPDH2	2	Yes	✓	✗	✗	✗	✓	✗
KLK2	2	No	?	✗	✗	✗	✗	?
MDK	2	No	✗	?	✗	✗	✗	?
MIR4435 1HG lOC541471	2	No	✗	✗	?	✗	?	✗
MMP25	2	No	?	✓	✗	✗	✗	✗
MXI1	2	No	?	✗	✗	?	✗	✗
NAALADL2	2	No	✓	✗	?	✗	✗	✗
NEAT1	2	No	✗	✗	✗	✗	?	?

Table A.1: Table of Boruta decisions for each variable with at least one training label decision rendered as "Tentative" (?) or "Confirmed" (✓). The total number of confirmed or tentative decisions are recorded, as well as whether the variable in question appears in the PUR model previously described. Variables rejected for individual training labels are shown with 7 *(continued)*

Variable	Total	In PUR Model	TriSig	Gleason ≥ 4+3	Gleason ≥ 3+4	D'Amico	Any Cancer	Cancer vs High Risk Cancer
RNF157	2	No	?	✗	✗	✗	?	✗
RPS11	2	No	?	✗	✗	✗	?	✗
Smoking unknown	2	No	✓	✗	✗	✗	✓	✗
SPON2	2	No	?	✗	✗	✗	✓	✗
VPS13A	2	No	✗	✓	✗	✗	✗	✓
ACTR5	1	No	✗	✗	✗	✗	✗	?
Alcohol No	1	No	?	✗	✗	✗	✗	✗
Alcohol Yes	1	No	✓	✗	✗	✗	✗	✗
ARexon9	1	No	✗	✗	✗	✗	?	✗
BTG2	1	No	✗	✗	?	✗	✗	✗
CAMK2N2	1	No	✗	✗	✗	✗	✗	✓
CASKIN1	1	No	✗	✗	✗	✗	?	✗
EN2	1	No	✗	✗	✗	✗	?	✗
HPRT	1	No	✗	✗	✗	✗	✗	?
ITPR1	1	No	✗	✗	✗	✗	✗	?
LBH	1	No	✗	?	✗	✗	✗	✗
MAK	1	No	✗	✗	✗	✗	?	✗
MFSD2A	1	No	✗	✗	✗	✗	?	✗
MMP26	1	Yes	✗	?	✗	✗	✗	✗
OGT	1	No	✗	✗	✗	✗	✗	?
PSTPIP1	1	No	✗	✗	✗	✗	✗	✓
PTN	1	No	✗	✗	✗	✗	✗	?
RPL23AP53	1	No	✗	✗	✗	✗	?	✗
SIRT1	1	No	?	✗	✗	✗	✗	✗
Smoking No	1	No	✓	✗	✗	✗	✗	✗
Smoking Yes	1	No	✓	✗	✗	✗	✗	✗
SNCA	1	No	✗	✗	✗	✗	✗	?
SSPO	1	Yes	✓	✗	✗	✗	✗	✗
SSTR1	1	No	✗	✗	✗	✗	?	✗
TMCC2	1	No	✗	✗	✗	✗	✗	✓
TWIST1	1	Yes	?	✗	✗	✗	✗	✗
VAX2	1	No	✗	✗	✗	✗	?	✗

Appendix B

Chapter 6:

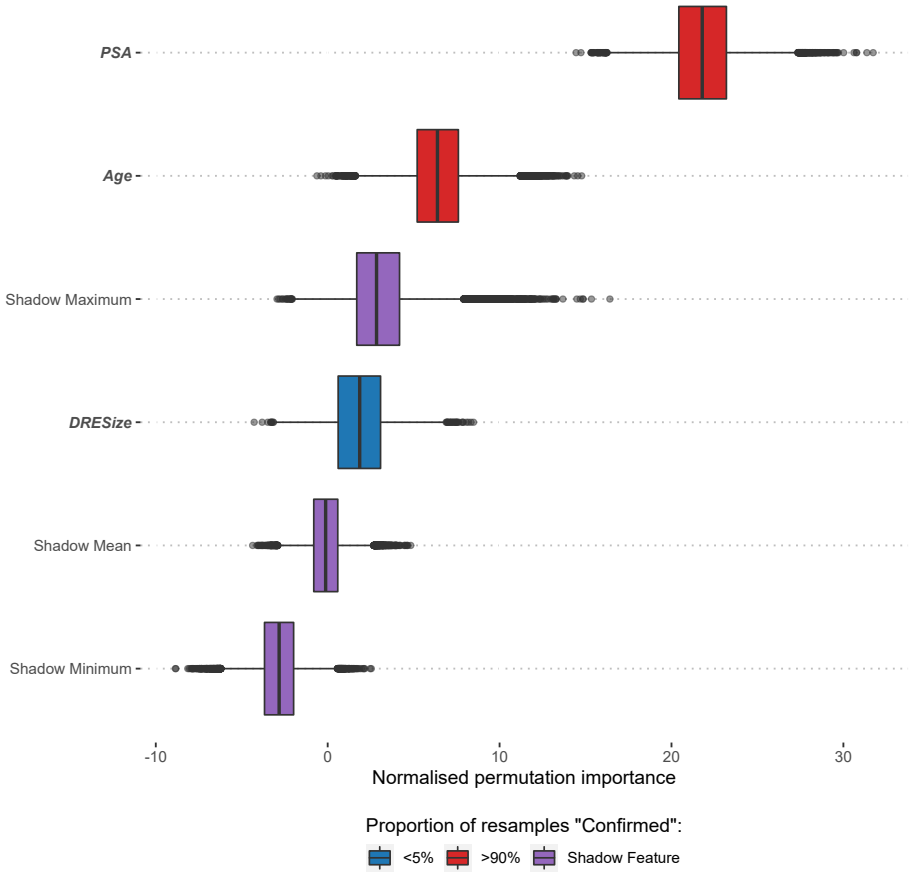


Figure B.1: Boruta analysis of variables available for the training of the SoC model. Variable importance was determined over 1,000 bootstrap resamples of the available data and the decision reached recorded at each resample. Variable origins are denoted by font; clinical variables are italicised and emboldened. Colour indicates the proportion of the 1,000 resamples a variable was confirmed to be important in. Variables confirmed in at least 90% of resamples were selected for training predictive models.

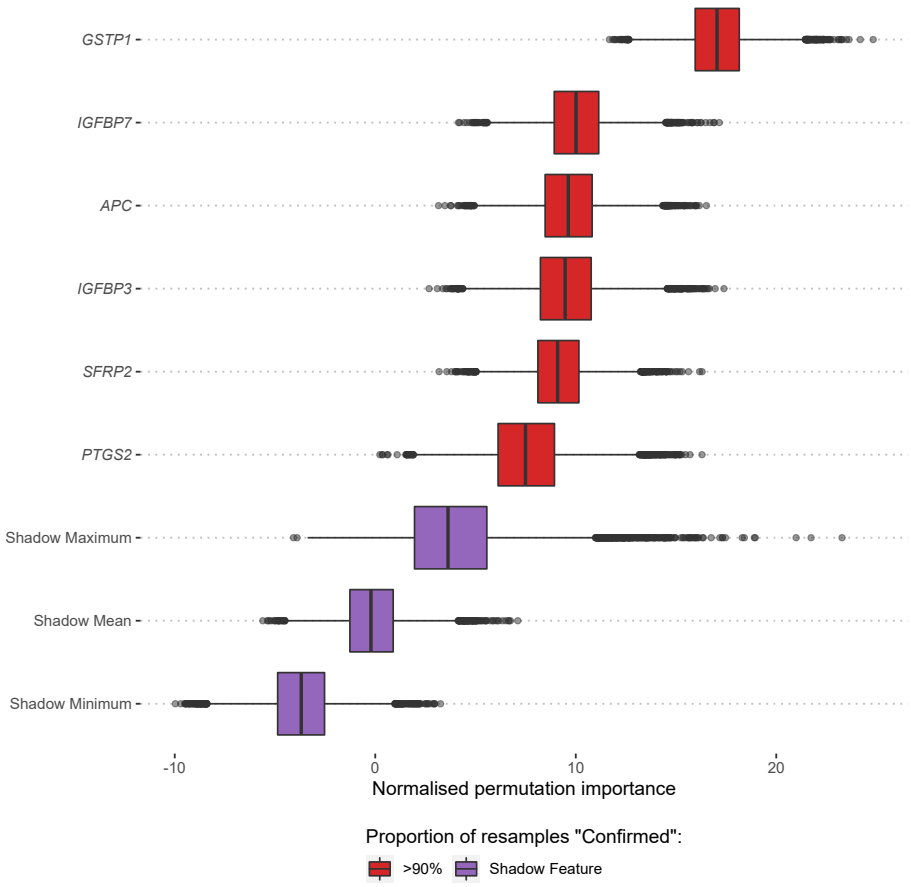


Figure B.2: Boruta analysis of variables available for the training of the Methylation model. Variable importance was determined over 1,000 bootstrap resamples of the available data and the decision reached recorded at each resample. Variable origins are denoted by font; methylation variables are italicised. Colour indicates the proportion of the 1,000 resamples a variable was confirmed to be important in. Variables confirmed in at least 90% of resamples were selected for training predictive models.

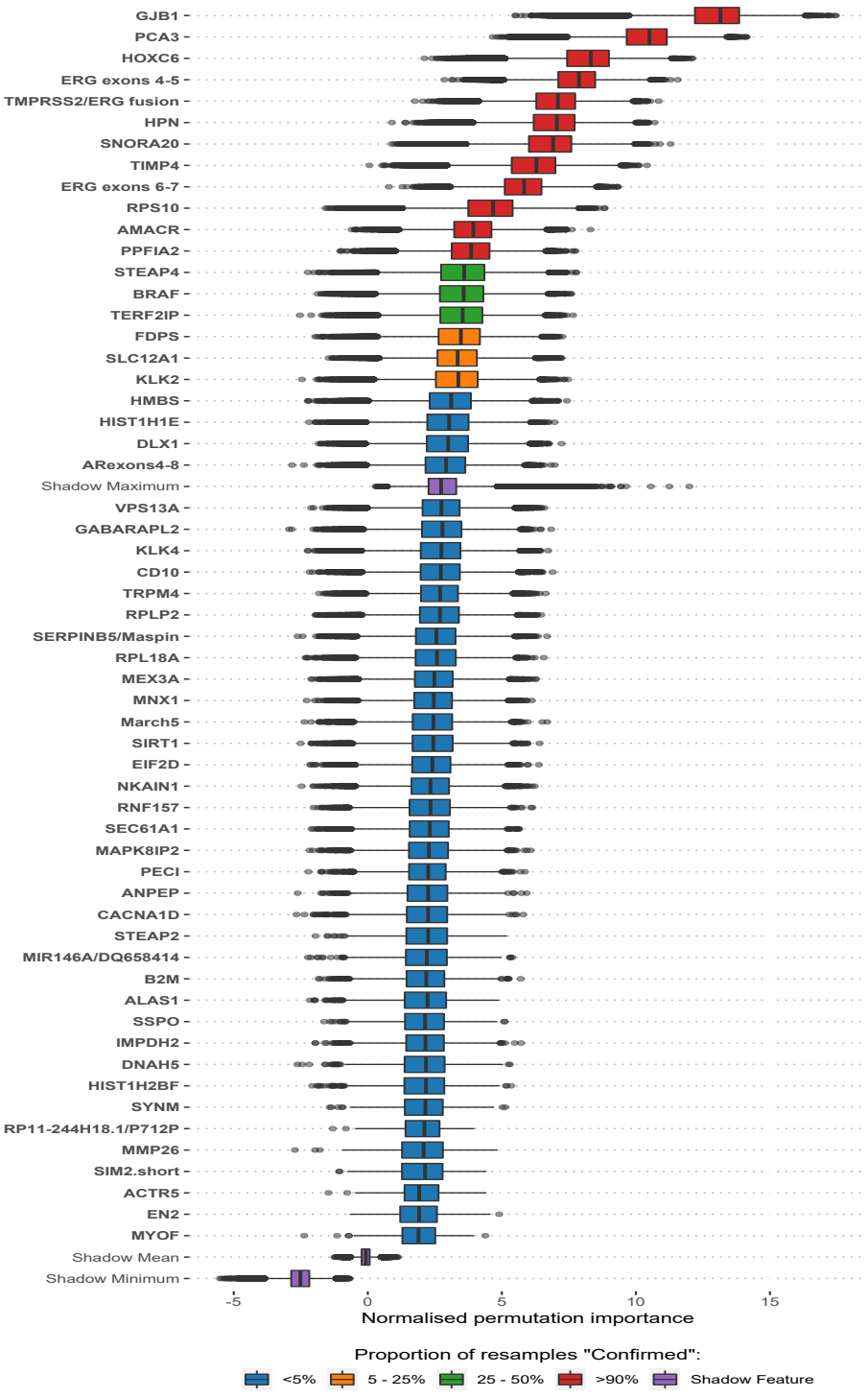


Figure B.3: Boruta analysis of variables available for the training of the ExoRNA model. Variable importance was determined over 1,000 bootstrap resamples of the available data and the decision reached recorded at each resample. Variable origins are denoted by font; clinical variables are emboldened. Colour indicates the proportion of the 1,000 resamples a variable was confirmed to be important in. Variables confirmed in at least 90% of resamples were selected for training predictive models.

Table B.1: List of all features available for selection as input variables for each model prior to bootstrapped Boruta feature selection.

SoC	*Methylation*	*ExoRNA*	ExoMeth
PSA	*GSTP1*	*Mar-05*	PSA
UrineVol	*APC*	*AATF*	UrineVol
DRESize	*SFRP2*	*ABCB9*	DRESize
Age	*IGFBP3*	*ACTR5*	Age
	IGFBP7	*AGR2*	*GSTP1*
	PTGS2	*ALAS1*	*APC*
		AMACR	*SFRP2*
		AMH	*IGFBP3*
		ANKRD34B	*IGFBP7*
		ANPEP	*PTGS2*
		APOC1	*Mar-05*
		ARexon9	*AATF*
		ARexons4-8	*ABCB9*
		ARHGEF25	*ACTR5*
		AURKA	*AGR2*
		B2M	*ALAS1*
		B4GALNT4	*AMACR*
		BRAF	*AMH*
		BTG2	*ANKRD34B*
		CACNA1D	*ANPEP*
		CADPS	*APOC1*
		CAMK2N2	*ARexon9*
		CAMKK2	*ARexons4-8*
		CASKIN1	*ARHGEF25*
		CCDC88B	*AURKA*
		CD10	*B2M*
		CDC20	*B4GALNT4*
		CDC37L1	*BRAF*
		CDKN3	*BTG2*
		CKAP2L	*CACNA1D*
		CLIC2	*CADPS*
		CLU	*CAMK2N2*
		COL10A1	*CAMKK2*
		COL9A2	*CASKIN1*
		CP	*CCDC88B*
		CTA-211A9.5/MIATNB	*CD10*
		DLX1	*CDC20*
		DNAH5	*CDC37L1*
		DPP4	*CDKN3*
		EIF2D	*CKAP2L*

EN2
ERG exons 4-5
ERG exons 6-7
ERG5
FDPS
FOLH1/PSMA/NAALAD1
GABARAPL2
GAPDH
GCNT1
GJB1
GOLM1
HIST1H1C
HIST1H1E
HIST1H2BF
HIST1H2BG
HIST32HA
HMBS
HOXC4
HOXC6
HPN
HPRT
IFT57
IGFBP3
IMPDH2
ISX
ITGBL1
ITPR1
KLK2
KLK3/PSA(exons1-2
KLK3/PSA(exons2-3
KLK4
LASS1
LBH
MAK
MAPK8IP2
MCM7
MCTP1
MDK
MED4
MEMO1
Met
MEX3A
MFSD2A
MGAT5B
CLIC2
CLU
COL10A1
COL9A2
CP
CTA-211A9.5/MIATNB
DLX1
DNAH5
DPP4
EIF2D
EN2
ERG exons 4-5
ERG exons 6-7
ERG5
FDPS
FOLH1/PSMA/NAALAD1
GABARAPL2
GAPDH
GCNT1
GJB1
GOLM1
HIST1H1C
HIST1H1E
HIST1H2BF
HIST1H2BG
HIST32HA
HMBS
HOXC4
HOXC6
HPN
HPRT
IFT57
IGFBP3
IMPDH2
ISX
ITGBL1
ITPR1
KLK2
KLK3/PSA(exons1-2
KLK3/PSA(exons2-3
KLK4
LASS1
LBH
MAK

MIC1

MIR146A/DQ658414
MIR4435-1HG/lOC541471
MKi67
MMP11
MMP25

MMP26
MNX1
MSMB
MXI1
MYOF

NAALADL2
NEAT1
NKAIN1
NLRP3
OGT

OR52A2/PSGR
PALM3
PCA3
PCSK6
PDLIM5

PECI
PPAP2A
PPFIA2
PPP1R12B
PSTPIP1

PTN
PTPRC
PVT1
RAB17
RIOK3

RNF157
RP11-244H18.1/P712P
RP11-97O12.7
RPL18A
RPL23AP53

RPLP2
RPS10
RPS11
SACM1L
SChLAP1

SEC61A1
SERPINB5/Maspin
SFRP4

MAPK8IP2

MCM7
MCTP1
MDK
MED4
MEMO1

Met
MEX3A
MFSD2A
MGAT5B
MIC1

MIR146A/DQ658414
MIR4435-1HG/lOC541471
MKi67
MMP11
MMP25

MMP26
MNX1
MSMB
MXI1
MYOF

NAALADL2
NEAT1
NKAIN1
NLRP3
OGT

OR52A2/PSGR
PALM3
PCA3
PCSK6
PDLIM5

PECI
PPAP2A
PPFIA2
PPP1R12B
PSTPIP1

PTN
PTPRC
PVT1
RAB17
RIOK3

RNF157
RP11-244H18.1/P712P
RP11-97O12.7

SIM2.long
SIM2.short

SIRT1
SLC12A1
SLC43A1
SLC4A1 S
SMAP1 ex 7-8

SMIM1
SNCA
SNORA20
SPINK1
SPON2

SRSF3
SSPO
SSTR1
ST6GALNAC1
STEAP2

STEAP4
STOM
SULF2
SULT1A1
SYNM

TBP
TDRD
TERF2IP
TERT
TFDP1

TIMP4
TMCC2
TMEM45B
TMEM47
TMEM86A

TMPRSS2/ERG fusion
TRPM4
TWIST1
UPK2
VAX2

VPS13A
ZNF577

RPL18A
RPL23AP53

RPLP2
RPS10
RPS11
SACM1L
SChLAP1

SEC61A1
SERPINB5/Maspin
SFRP4
SIM2.long
SIM2.short

SIRT1
SLC12A1
SLC43A1
SLC4A1 S
SMAP1 ex 7-8

SMIM1
SNCA
SNORA20
SPINK1
SPON2

SRSF3
SSPO
SSTR1
ST6GALNAC1
STEAP2

STEAP4
STOM
SULF2
SULT1A1
SYNM

TBP
TDRD
TERF2IP
TERT
TFDP1

TIMP4
TMCC2
TMEM45B
TMEM47
TMEM86A

TMPRSS2/ERG fusion
TRPM4

	TWIST1
	UPK2
	VAX2
	VPS13A
	ZNF577

References

1. Cancer Research UK. Prostate cancer incidence statistics. 2019. `http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer/incidence`. Accessed June 29, 2019.

2. Soos G, Tsakiris I, Szanto J, Turzo C, Haas PG, Dezso B. The prevalence of prostate carcinoma and its precursor in Hungary: An autopsy study. 2005;48(5):739-744. doi:10.1016/j.eururo.2005.08.010

3. Sánchez-Chapado M, Olmedilla G, Cabeza M, Donat E, Ruiz A. Prevalence of prostate cancer and prostatic intraepithelial neoplasia in Caucasian Mediterranean males: An autopsy study. 2003;54(3):238-247. doi:10.1002/pros.10177

4. Schlomm T, Weischenfeldt J, Korbel J, Sauter G. The Aging Prostate Is Never "Normal": Implications from the Genomic Characterization of Multifocal Prostate Cancers. 2015;68(3):348-350. doi:10.1016/j.eururo.2015.04.012

5. National Institute for Health and Care Excellence. *Prostate cancer: diagnosis and management [2019]*. NICE

6. Martin RM, Donovan JL, Turner EL, et al. Effect of a Low-Intensity PSA-Based Screening Intervention on Prostate Cancer Mortality. 2018;319(9):883. doi:10.1001/jama.2018.0154

7. Lane JA, Donovan JL, Davis M, et al. Active monitoring, radical prostatectomy, or radiotherapy for localised prostate cancer: Study design and diagnostic and baseline results of the ProtecT randomised phase 3 trial. 2014;15(10):1109-1118. doi:10.1016/S1470-2045(14)70361-4

8. Connell SP and, Hanna M, McCarthy F, et al. A Four-Group Urine Risk Classifier for Predicting Outcome in Prostate Cancer Patients. May 2019. doi:10.1111/bju.14811

9. Palmirotta R, Lovero D, Cafforio P, et al. Liquid biopsy of cancer: a multimodal diagnostic tool in clinical oncology. 2018;10:1758835918794630-1758835918794630. doi:10.1177/1758835918794630

10. Junker K, Heinzelmann J, Beckham C, Ochiya T, Jenster G. Extracellular Vesicles and Their Role in Urologic Malignancies. 2016;70:323-331. doi:10.1016/j.eururo.2016.02.046

11. O'Reilly E, Tuzova AV, Walsh AL, et al. epiCaPture: A Urine DNA Methylation Test for Early Detection of Aggressive Prostate Cancer. 2019;(3):1-18. doi:10.1200/PO.18.00134

12. Hessels D, Klein Gunnewiek JMT, Van Oort I, et al. DD3PCA3-based molecular urine analysis for the diagnosis of prostate cancer. 2003;44(1):8-16. doi:10.1016/S0302-2838(03)00201-X

13. Nilsson J, Skog J, Nordstrand A, et al. Prostate cancer-derived urine exosomes: A novel approach to biomarkers for prostate cancer. 2009;100(10):1603-1607. doi:10.1038/sj.bjc.6605058

14. Bologna M, Vicentini C, Festuccia C, et al. Early diagnosis of prostatic carcinoma based on in vitro culture of viable tumor cells harvested by prostatic massage. 1988;14(6):474-476. `http://www.ncbi.nlm.nih.gov/pubmed/3181228`.

15. Garret M, Jassie M. Cytologic examination of post prostatic massage specimens as an aid in diagnosis of carcinoma of the prostate. 20(2):126-131. `http://www.ncbi.nlm.nih.gov/pubmed/1065172`.

16. Kern SE. Why your new cancer biomarker may never work: Recurrent patterns and remarkable diversity in biomarker failures. 2012;72(23):6097-6101. doi:10.1158/0008-5472.CAN-12-3232

17. Diamandis EP. The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? 2012;10(1):87. doi:10.1186/1741-7015-10-87

18. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? 2008;336:1472. doi:10.1136/bmj.39590.732037.47

19. Moher D. Reporting research results: A moral obligation for all researchers. 2007.

20. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. 2010;7:e1000217. doi:10.1371/journal.pmed.1000217

21. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. 2015;67(6):1142-1151. doi:https://doi.org/10.1016/j.eururo.2014.11.025

22. Hanahan D, Weinberg Ra. The hallmarks of cancer. 2000;100:57-70.

23. Hanahan D, Weinberg RA, Pan KH, et al. Hallmarks of Cancer: The Next Generation. 2011;144(5):646-674. doi:10.1016/j.cell.2011.02.013

24. Gundem G, Van Loo P, Kremeyer B, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*. 2015;520(7547):353-357. doi:10.1038/nature14347

25. Franks LM. Biology of the prostate and its tumors. In: *The Treatment of Prostatic Hypertrophy and Neoplasia*. Springer Netherlands; 1974:1-26. doi:10.1007/978-94-015-7190-6_1

26. Costello LC, Franklin RB. Prostatic fluid electrolyte composition for the screening of prostate cancer: A potential solution to a major problem. 2009;12:17-24. doi:10.1038/pcan.2008.19

27. Balk SP, Ko Y-J, Bubley GJ. Biology of prostate-specific antigen. 2003;21(2):383-391. doi:10.1200/JCO.2003.02.083

28. McNeal JE. Origin and development of carcinoma in the prostate. 1969;23(1):24-34. doi:10.1002/1097-0142(196901)23:1<24::AID-CNCR2820230103>3.0.CO;2-1

29. McNeal JE. The zonal anatomy of the prostate. 1981;2(1):35-49. doi:10.1002/pros.2990020105

30. McNeal JE. Normal histology of the prostate. 1988;12(8):619-633. `http://www.ncbi.nlm.nih.gov/pubmed/2456702`.

31. Bostwick DG, Burke HB, Djakiew D, et al. Human prostate cancer risk factors. 2004;101:2371-2490. doi:10.1002/cncr.20408

32. Elbuluk O, Muradyan N, Shih J, et al. Differentiating Transition Zone Cancers from Benign Prostatic Hyperplasia by Quantitative Multiparametric Magnetic Resonance Imaging. 2016;40(2):218-224. doi:10.1097/RCT.0000000000000353

33. Abdelsayed GA, Danial T, Kaswick JA, Finley DS. Tumors of the anterior prostate: Implications for diagnosis and treatment. 2015;85:1224-1228. doi:10.1016/j.urology.2014.12.035

34. Ghai S, Haider MA. Multiparametric-MRI in diagnosis of prostate cancer. 2015;31(3):194-201. doi:10.4103/0970-1591.159606

35. Koppie TM, Bianco FJ, Kuroiwa K, et al. The clinical features of anterior prostate cancers. 2006;98(6):1167-1171. doi:10.1111/j.1464-410X.2006.06578.x

36. Vargas HA, Akin O, Franiel T, et al. Normal central zone of the prostate and central zone involvement by prostate cancer: Clinical and mr imaging implications. 2012;262(3):894-902. doi:10.1148/radiol.11110663

37. Cohen RJ, Shannon BA, Phillips M, Moorin RE, Wheeler TM, Garrett KL. Central Zone Carcinoma of the Prostate Gland: A Distinct Tumor Type With Poor Prognostic Features. 2008;179(5):1762-1767. doi:10.1016/j.juro.2008.01.017

38. Troncoso P, Babaian RJ, Ro JY, Grignon DJ, Eschenbach AC von, Ayala AG. Prostatic intraepithelial neoplasia and invasive prostatic adenocarcinoma in cystoprostatectomy specimens. 1989;34(6 Suppl):52-56. `http://www.ncbi.nlm.nih.gov/pubmed/2603286`.

39. Lee JJ, Thomas IC, Nolley R, Ferrari M, Brooks JD, Leppert JT. Biologic differences between peripheral and transition zone prostate cancer. 2015;75(2):183-190. doi:10.1002/pros.22903

40. Loeb S, Bjurlin MA, Nicholson J, et al. Overdiagnosis and overtreatment of prostate cancer. 2014;65:1046-1055. doi:10.1016/j.eururo.2013.12.062

41. Stemmermann GN, Nomura AMY, Chyou PH, Yatani R. A Prospective Comparison of Prostate Cancer at Autopsy and as a Clinical Event: The Hawaii Japanese Experience. 1992;1(3):189-193. `http://www.ncbi.nlm.nih.gov/pubmed/1306104`.

42. Boniol M, Autier P, Perrin P, Boyle P. Variation of Prostate-specific Antigen Value in Men and Risk of High-grade Prostate Cancer: Analysis of the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Study. 2015;85(5):1117-1122. doi:10.1016/j.urology.2015.02.013

43. Lujan M, Pascual C, Rodriguez N, et al. Impact of the weather on the serum levels

of prostatic specific antigen (PSA). 2006;59:247-252. `http://www.ncbi.nlm.nih.gov/pubmed/16724709`.

44. Salama G, Noirot O, Bataille V, et al. Seasonality of Serum Prostate-Specific Antigen Levels: A Population-Based Study. 2007;52(3):708-714. doi:10.1016/j.eururo.2006.11.042

45. Catalona WJ, Smith DS, Ratliff TL, Basler JW. Detection of Organ-Confined Prostate Cancer Is Increased Through Prostate-Specific Antigen—Based Screening. 1993;270(8):948-954. doi:10.1001/jama.1993.03510080052031

46. Naji L, Randhawa H, Sohani Z, et al. Digital rectal examination for prostate cancer screening in primary care: A systematic review and meta-analysis. *Annals of Family Medicine*. 2018;16(2):149-154. doi:10.1370/afm.2205

47. Moreira Leite KR, Camara-Lopes LHA, Dall'Oglio MF, et al. Upgrading the Gleason Score in Extended Prostate Biopsy: Implications for Treatment Choice. *International Journal of Radiation Oncology Biology Physics*. 2009;73(2):353-356. doi:10.1016/j.ijrobp.2008.04.039

48. Taira AV, Merrick GS, Galbreath RW, et al. Performance of transperineal template-guided mapping biopsy in detecting prostate cancer in the initial and repeat biopsy setting. 2010;13(1):71-77. doi:10.1038/pcan.2009.42

49. Mellinger GT, Gleason D, Bailar 3rd J. The histology and prognosis of prostatic cancer. 1967;97(2):331-337. doi:10.1016/S0022-5347(17)63039-8

50. Epstein JI, Allsbrook WC, Amin MB, et al. The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. In: *American Journal of Surgical Pathology*. Vol 29.; 2005:1228-1242. doi:10.1097/01.pas.0000173646.99337.b1

51. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. 2016;40(2):244-252. doi:10.1097/PAS.0000000000000530

52. Epstein JI. An Update of the Gleason Grading System. 2010;183:433-440. doi:10.1016/j.juro.2009.10.046

53. Stark JR, Perner S, Stampfer MJ, et al. Gleason score and lethal prostate cancer: Does $3 + 4 = 4 + 3$? 2009;27(21):3459-3464. doi:10.1200/JCO.2008.20.4669

54. NICE. Costing Statement: prostate cancer: diagnosis and treatment. 2014;(June):1-23.

55. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European Urology*. 2019;76(3):340-351. doi:https://doi.org/10.1016/j.eururo.2019.02.033

56. Porpiglia F, Manfredi M, Mele F, et al. Diagnostic Pathway with Multiparametric Magnetic Resonance Imaging Versus Standard Pathway: Results from a Randomized Prospective Study in Biopsy-naïve Patients with Suspected Prostate Cancer. *European Urology*. 2017;72(2):282-288. doi:10.1016/j.eururo.2016.08.041

57. Kasivisvanathan V, Rannikko AS, Borghi M, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *New England Journal of Medicine*. 2018;378(19):1767-1777. doi:10.1056/NEJMoa1801993

58. Ahmed HU, El-Shater Bosaily A, Brown LC, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. 2017;389(10071):815-822. doi:10.1016/S0140-6736(16)32401-1

59. Richie JP, Catalona WJ, Ahmann FR, et al. Effect of patient age on early detection of prostate cancer with serum prostate-specific antigen and digital rectal examination. 1993;42(4):365-374. `http://www.ncbi.nlm.nih.gov/pubmed/7692657`.

60. Chun FKH, Karakiewicz PI, Briganti A, et al. Prostate Cancer Nomograms: An Update. 2006;50:914-926. doi:10.1016/j.eururo.2006.07.042

61. Shariat SF, Kattan MW, Vickers AJ, Karakiewicz PI, Scardino PT. Critical review of prostate cancer predictive tools. 2009;5(10):1555-1584. doi:10.2217/fon.09.121

62. Thurtle DCAL David R. AND Greenberg. Individual prognosis at diagnosis in nonmetastatic prostate cancer: Development and external validation of the predict prostate multivariable model. *PLOS Medicine*. 2019;16(3):1-19. doi:10.1371/journal.pmed.1002758

63. D'Amico AV, Whittington R, Bruce Malkowicz S, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. 1998;280(11):969-974. doi:10.1001/jama.280.11.969

64. Tsikis ST, Nottingham CU, Faris SF. The Relationship Between Incontinence and Erectile Dysfunction After Robotic Prostatectomy: Are They Mutually Exclusive? *Journal of Sexual Medicine*. 2017;14(10):1241-1247. doi:10.1016/j.jsxm.2017.08.002

65. Lester JF, Mason MD. Cardiovascular effects of hormone therapy for prostate cancer. 2015;7:129-138. doi:10.2147/DHPS.S50549

66. Mottet N, Bellmunt J, Bolla M, et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. 2017;71(4):618-629. doi:10.1016/j.eururo.2016.08.003

67. Cornford P, Bellmunt J, Bolla M, et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer. 2017;71(4):630-642. doi:10.1016/j.eururo.2016.08.002

68. Bruinsma SM, Zhang L, Roobol MJ, et al. The movember foundation's gap3 cohort: A profile of the largest global prostate cancer active surveillance database to date. *BJU International*. 2018;121(5):737-744. doi:10.1111/bju.14106

69. Hamdy FC, Donovan JL, Lane JA, et al. 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. 2016;375(15):1415-1424. doi:10.1056/NEJMoa1606220

70. Tosoian JJ, Carter HB, Lepor A, Loeb S. Active surveillance for prostate cancer: Current evidence and contemporary state of practice. 2016;13:205-215. doi:10.1038/nrurol.2016.45

71. Bellardita L, Valdagni R, Van Den Bergh R, et al. How does active surveillance for prostate cancer affect quality of life? A systematic review. 2015;67:637-645. doi:10.1016/j.eururo.2014.10.028

72. Ruane-McAteer E, Porter S, O'Sullivan JM, Santin O, Prue G. Active surveillance for favorable-risk prostate cancer: Is there a greater psychological impact than previously thought? A systematic, mixed studies literature review. 2017;26(10):1411-1421. doi:10.1002/pon.4311

73. Cancer incidence and mortality projections in the UK until 2035. *British Journal of Cancer*. 2016;115(9):1147-1155. doi:10.1038/bjc.2016.304

74. Murthy V, Rishi A, Gupta S, et al. Clinical impact of prostate specific antigen (PSA) inter-assay variability on management of prostate cancer. *Clinical Biochemistry*. 2016;49(1):79-84. doi:10.1016/j.clinbiochem.2015.10.013

75. Forde JC, Marignol L, Blake O, et al. Standardization of assay methods reduces variability of total PSA measurements: An Irish study. *BJU International*. 2012;110(5):644-650. doi:10.1111/j.1464-410X.2011.10923.x

76. Schröder FH, Hugosson J, Roobol MJ, et al. Screening and Prostate-Cancer Mortality in a Randomized European Study. 2009;360(13):1320-1328. doi:10.1056/NEJMoa0810084

77. Heijnsdijk EAM, Wever EM, Auvinen A, et al. Quality-of-Life Effects of Prostate-Specific Antigen Screening. 2012;367(7):595-605. doi:10.1056/NEJMoa1201637

78. Nafie S, Mellon JK, Dormer JP, Khan MA. The role of transperineal template prostate biopsies in prostate cancer diagnosis in biopsy naïve men with PSA less than 20 ng ml-1. *Prostate Cancer and Prostatic Diseases*. 2014;17(2):170-173. doi:10.1038/pcan.2014.4

79. National Institute for Health and Care Excellence. *Prostate cancer: diagnosis and management [D] Evidence review for diagnosing and identifying clinically significant prostate cancer NICE guideline NG131 Evidence reviews*. NICE; 2019. `https://www.nice.org.uk/guidance/ng131/evidence/d-diagnosing-and-identifying-clinically-significant-prostate-cancer-pdf-6779081777`.

80. Pinkhasov GI, Lin YK, Palmerola R, et al. Complications following prostate needle biopsy requiring hospital admission or emergency department visits - Experience from 1000 consecutive cases. 2012;110:369-374. doi:10.1111/j.1464-410X.2011.10926.x

81. Nam RK, Saskin R, Lee Y, et al. Increasing Hospital Admission Rates for Urological Complications After Transrectal Ultrasound Guided Prostate Biopsy. *Journal of Urology*. 2010;183(3):963-969. doi:10.1016/j.juro.2009.11.043

82. Walz J. The "PROMIS" of Magnetic Resonance Imaging Cost Effectiveness in Prostate Cancer Diagnosis? *European Urology*. 2018;73(1):31-32. doi:10.1016/j.eururo.2017.09.015

83. Sonn GA, Fan RE, Ghanouni P, et al. Prostate Magnetic Resonance Imaging Interpretation Varies Substantially Across Radiologists. December 2018. doi:10.1016/j.euf.2017.11.010

84. Faria R, Soares MO, Spackman E, et al. Optimising the Diagnosis of Prostate Cancer in the Era of Multiparametric Magnetic Resonance Imaging: A Cost-effectiveness

Analysis Based on the Prostate MR Imaging Study (PROMIS). 2018;73(1):23-30. doi:10.1016/j.eururo.2017.08.018

85. Carey LA, Perou CM, Livasy CA, et al. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. 2006;295(21):2492-2502. doi:10.1001/jama.295.21.2492

86. Hernandez DJ, Nielsen ME, Han M, Partin AW. Contemporary Evaluation of the D'Amico Risk Classification of Prostate Cancer. 2007;70(5):931-935. doi:10.1016/j.urology.2007.08.055

87. Boorjian SA, Karnes RJ, Rangel LJ, Bergstralh EJ, Blute ML. Mayo Clinic Validation of the D'Amico Risk Group Classification for Predicting Survival Following Radical Prostatectomy. 2008;179(4):1354-1361. doi:10.1016/j.juro.2007.11.061

88. Cooperberg MR, Freedland SJ, Pasta DJ, et al. Multiinstitutional validation of the UCSF cancer of the prostate risk assessment for prediction of recurrence after radical prostatectomy. 2006;107(10):2384-2391. doi:10.1002/cncr.22262

89. Scattoni V, Lazzeri M, Lughezzani G, et al. Head-to-head comparison of prostate health index and urinary PCA3 for predicting cancer at initial or repeat biopsy. 2013;190(2):496-501. doi:10.1016/j.juro.2013.02.3184

90. Lughezzani G, Budäus L, Isbarn H, et al. Head-to-Head Comparison of the Three Most Commonly Used Preoperative Models for Prediction of Biochemical Recurrence After Radical Prostatectomy. 2010;57(4):562-568. doi:10.1016/j.eururo.2009.12.003

91. Pasic MD, Samaan S, Yousef GM. Genomic medicine: New frontiers and new challenges. 2013;59(1):158-167. doi:10.1373/clinchem.2012.184622

92. Rittenhouse H, Blase A, Shamel B, Schalken J, Groskopf J. The long and winding road to FDA approval of a novel prostate cancer test: Our story. 2013;59(1):32-34. doi:10.1373/clinchem.2012.198739

93. Minciacchi VR, Zijlstra A, Rubin MA, Di Vizio D. Extracellular vesicles for liquid biopsy in prostate cancer: Where are we and where are we headed? 2017;20(3):251-258. doi:10.1038/pcan.2017.7

94. Bussemakers MJG, Van Bokhoven A, Verhaegh GW, et al. DD3: A new prostate-specific gene, highly overexpressed in prostate cancer. 1999;59(23):5975-5979. doi:10.1038/ncb2161

95. Vlaeminck-Guillem V, Ruffion A, Andre J. Place du test urinaire PCA3 pour le diagnostic du cancer de la prostate. 2008;18:259-265. doi:10.1016/j.purol.2008.03.029

96. Salagierski M, Sosnowski M, Schalken JA. How accurate is our prediction of biopsy outcome? PCA3-based nomograms in personalized diagnosis of prostate cancer. 2012;65:110-112. doi:10.5173/ceju.2012.03.art1

97. Leyten GHJM, Hessels D, Jannink SA, et al. Prospective multicentre evaluation of PCA3 and TMPRSS2-ERG gene fusions as diagnostic and prognostic urinary biomarkers for prostate cancer. 2014;65(3):534-542. doi:10.1016/j.eururo.2012.11.014

98. National Institute for Health and Care Excellence. *Prostate cancer update Health*

economic model report Health economic model report HE.1 General HE.2 RQ8: Managing people at increased risk of prostate cancer HE.2.1 Decision problem.; 2019.

99. Van Neste L, Hendriks RJ, Dijkstra S, et al. Detection of High-grade Prostate Cancer Using a Urinary Molecular Biomarker–Based Risk Score. 2016;70(5):740-748. doi:10.1016/j.eururo.2016.04.012

100. Multicenter Optimization and Validation of a 2-Gene mRNA Urine Test for Detection of Clinically Significant Prostate Cancer before Initial Prostate Biopsy. *Journal of Urology*. 2019;202(2):256-262. doi:10.1097/JU.0000000000000293

101. Govers TM, Caba L, Resnick MJ. Cost-Effectiveness of Urinary Biomarker Panel in Prostate Cancer Risk Assessment. *Journal of Urology*. 2018;200(6):1221-1226. doi:10.1016/j.juro.2018.07.034

102. McKiernan J, Donovan MJ, O'Neill V, et al. A novel urine exosome gene expression assay to predict high-grade prostate cancer at initial biopsy. 2016;2(7):882-889. doi:10.1001/jamaoncol.2016.0097

103. A Prospective Adaptive Utility Trial to Validate Performance of a Novel Urine Exosome Gene Expression Assay to Predict High-grade Prostate Cancer in Patients with Prostate-specific Antigen 2–10 ng/ml at Initial Biopsy. *European Urology*. 2018;74(6):731-738. doi:10.1016/j.eururo.2018.08.019

104. Senn SJ. Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. *Proceedings of the International Statistical Institute, 55th Session*. 2005:1-13.

105. Vickers AJ. Prediction Models in Cancer Care Introduction: Cancer as a Prediction Problem. 2011;61:315-326. doi:10.3322/caac.20118

106. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer New York; 2009. doi:10.1007/978-0-387-84858-7

107. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*. 1996;4:237-285. doi:10.1613/jair.301

108. Mohri M. *Foundations of Machine Learning - Book*. MIT Press; 2012:414. doi:ISBN 978-0-262-01825-8

109. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. 2015;13:8-17. doi:10.1016/j.csbj.2014.11.005

110. Donovan MJ, Noerholm M, Bentink S, et al. A molecular signature of PCA3 and ERG exosomal RNA from non-DRE urine is predictive of initial prostate biopsy result. 2015;18(4):370-375. doi:10.1038/pcan.2015.40

111. Connell SP, O'Reilly E, Tuzova A, et al. Development of a multivariable risk model integrating urinary cell DNA methylation and cell-free RNA data for the detection of significant prostate cancer. *Prostate*. 2020;80(7):547-558. doi:10.1002/pros.23968

112. Luca BA, Brewer DS, Edwards DR, et al. DESNT: A Poor Prognosis Category of

Human Prostate Cancer. March 2017. doi:10.1016/j.euf.2017.01.016

113. Lipton ZC. The Mythos of Model Interpretability. In: *ICML Workshop on Human Interpretability in Machine Learning.*; 2016. `https://arxiv.org/pdf/1606.03490.pdf%20http://arxiv.org/abs/1606.03490`.

114. Haykin S. *Neural Networks: A Comprehensive Foundation.* 1st ed. USA: Prentice Hall PTR; 1994.

115. Graves A, Wayne G, Danihelka I. Neural Turing Machines. 2014. doi:10.3389/neuro.12.006.2007

116. Lei T, Barzilay R, Jaakkola T. Why Should I Trust You? Explaining the Predictions of Any Classifier. 2016. doi:10.1145/2939672.2939778

117. Selvadurai ED, Singhera M, Thomas K, et al. Medium-term outcomes of active surveillance for localised prostate cancer. 2013;64(6):981-987. doi:10.1016/j.eururo.2013.02.020

118. Geiss GK, Bumgarner RE, Birditt B, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. 2008;26(3):317-325. doi:10.1038/nbt1385

119. Perry AS, O'Hurley G, Raheem OA, et al. Gene expression and epigenetic discovery screen reveal methylation of SFRP2 in prostate cancer. *International Journal of Cancer.* 2013;132(8):1771-1780. doi:10.1002/ijc.27798

120. Sullivan L, Murphy TM, Barrett C, et al. IGFBP7 promoter methylation and gene expression analysis in prostate cancer. *Journal of Urology.* 2012;188(4):1354-1360. doi:10.1016/j.juro.2012.06.002

121. Perry AS, Loftus B, Moroose R, et al. In silico mining identifies IGFBP3 as a novel target of methylation in prostate cancer. *British Journal of Cancer.* 2007;96(10):1587-1594. doi:10.1038/sj.bjc.6603767

122. Bastian PJ, Ellinger J, Heukamp LC, Kahl P, Müller SC, Rücker A von. Prognostic Value of CpG Island Hypermethylation at PTGS2, RAR-beta, EDNRB, and Other Gene Loci in Patients Undergoing Radical Prostatectomy. *European Urology.* 2007;51(3):665-674. doi:10.1016/j.eururo.2006.08.008

123. Whitaker HC, Kote-Jarai Z, Ross-Adams H, et al. The rs10993994 risk allele for prostate cancer results in clinically relevant changes in microseminoprotein-beta expression in tissue and urine. *PLoS ONE.* 2010;5(10). doi:10.1371/journal.pone.0013363

124. Ross-Adams H, Lamb A, Dunning M, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine.* 2015;2(9):1133-1144. doi:10.1016/j.ebiom.2015.07.017

125. Stephenson SA, Verity K, Ashworth LK, Clements JA. Localization of a new prostate-specific antigen-related serine protease gene, KLK4, is evidence for an expanded human kallikrein gene family cluster on chromosome 19q13.3-13.4. *Journal of Biological Chemistry.* 1999;274(33):23210-23214. doi:10.1074/jbc.274.33.23210

126. Emami N, Diamandis EP. Utility of kallikrein-related peptidases (KLKs) as cancer biomarkers. 2008;54:1600-1607. doi:10.1373/clinchem.2008.105189

127. Paliouras M, Borgono C, Diamandis EP. Human tissue kallikreins: The cancer

biomarker family. 2007;249:61-79. doi:10.1016/j.canlet.2006.12.018

128. Morgan R, Boxall A, Bhatt A, et al. Engrailed-2 (en2): A tumor specific urinary biomarker for the early diagnosis of prostate cancer. *Clinical Cancer Research*. 2011;17(5):1090-1098. doi:10.1158/1078-0432.CCR-10-2410

129. Zürbig P, Renfrow MB, Schiffer E, et al. Biomarker discovery by ce-ms enables sequence analysis via ms/ms with platform-independent separation. *ELECTROPHORESIS*. 2006;27(11):2111-2125. doi:10.1002/elps.200500827

130. Kaiser T, Hermann A, Kielstein JT, et al. Capillary electrophoresis coupled to mass spectrometry to establish polypeptide patterns in dialysis fluids. In: *Journal of Chromatography a*. Vol 1013. Elsevier; 2003:157-171. doi:10.1016/S0021-9673(03)00712-X

131. Wittke S, Fliser D, Haubitz M, et al. Determination of peptides and proteins in human urine with capillary electrophoresis-mass spectrometry, a suitable tool for the establishment of new diagnostic markers. In: *Journal of Chromatography a*. Vol 1013. Elsevier; 2003:173-181. doi:10.1016/S0021-9673(03)00713-1

132. Siwy J, Mullen W, Golovko I, Franke J, Zürbig P. Human urinary peptide database for multiple disease biomarker discovery. 2011;5:367-374. doi:10.1002/prca.201000155

133. Burnham KP, Anderson DR, eds. Information and likelihood theory: A basis for model selection and inference. In: *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer New York; 2002:49-97. doi:10.1007/978-0-387-22456-5_2

134. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996;58(1):267-288. `http://www.jstor.org/stable/2346178`.

135. Wurm MJ, Rathouz PJ, Hanlon BM. Regularized Ordinal Regression and the ordinalNet R Package. 2017. `https://arxiv.org/pdf/1706.05003.pdf%20http://arxiv.org/abs/1706.05003`.

136. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*. 2005;67:301-320.

137. Kursa MB, Rudnicki WR. The all relevant feature selection using random forest. *CoRR*. 2011;abs/1106.5112. `http://arxiv.org/abs/1106.5112`.

138. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1974;36(2):111-133. doi:10.1111/j.2517-6161.1974.tb00994.x

139. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statist Sci*. 1996;11(3):189-228. doi:10.1214/ss/1032280214

140. Ho TK. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol 1. ICDAR '95. IEEE Computer Society; 1995:278-282. doi:10.1109/ICDAR.1995.598994

141. Ho TK. The random subspace method for constructing decision forests.

1998;20(8):832-844. doi:10.1109/34.709601

142. Breiman L. Random forests. 2001;45(1):5-32. doi:10.1023/A:1010933404324

143. Quinlan JR. *C4.5: Programs for Machine Learning*. Vol 1. Morgan Kaufmann Publishers; 1992:302. doi:10.1016/S0019-9958(62)90649-6

144. Liaw A, Wiener M. Classification and regression by randomForest. 2002;2(3):18-22. `https://CRAN.R-project.org/doc/Rnews/`.

145. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. 1997;55(1):119-139. doi:10.1006/jcss.1997.1504

146. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. 2016. doi:10.1145/2939672.2939785

147. Hastie T, Tibshirani R, Friedman J. Boosting and Additive Trees. 2003:337-388. doi:10.1007/b94608

148. Goldbloom A, Hamner B, Moser J, Cukierski M. kaggle: Your Home for Data Science. 2017. `https://www.kaggle.com/`.

149. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. 2010;36(11). doi:Vol. 36, Issue 11, Sep 2010

150. Therneau TM. *A Package for Survival Analysis in R.*; 2020. `https://CRAN.R-project.org/package=survival`.

151. Terry M. Therneau, Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. New York: Springer; 2000.

152. Harrell Jr FE. *Rms: Regression Modeling Strategies.*; 2020. `https://CRAN.R-project.org/package=rms`.

153. Robin X, Turck N, Hainard A, et al. PROC: An open-source package for r and s+ to analyze and compare roc curves. 2011;12:77.

154. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. 2006;26(6):565-574. doi:10.1177/0272989X06295361

155. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: Guidance for correct interpretation and appropriate use. 2016;34(21):2534-2540. doi:10.1200/JCO.2015.65.5654

156. Vickers AJ, Calster B van, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research*. 2019;3(1):18. doi:10.1186/s41512-019-0064-7

157. Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. Moving beyond P values: data analysis with estimation graphics. June 2019:1. doi:10.1038/s41592-019-0470-3

158. D'Amico AV, Moul J, Carroll PR, Sun L, Lubeck D, Chen MH. Cancer-specific mortality after surgery or radiation for patients with clinically localized prostate cancer managed during the prostate-specific antigen era. 2003;21(11):2163-2172. doi:10.1200/JCO.2003.01.075

159. Gleason DF MG. Prediction of prognosis for prostatic Staging, adenocarcinoma by combined histological grading and slinical. 1974;111(1):111:58-64. doi:10.1016/S0022-5347(17)59889-4

160. Sanda MG, Cadeddu JA, Kirkby E, et al. Clinically Localized Prostate Cancer: AUA/ASTRO/SUO Guideline. Part I: Risk Stratification, Shared Decision Making, and Care Options. 2018;199(3):683-690. doi:10.1016/j.juro.2017.11.095

161. Brajtbord JS, Leapman MS, Cooperberg MR. The CAPRA Score at 10 Years: Contemporary Perspectives and Analysis of Supporting Studies. 2017;71(5):705-709. doi:10.1016/j.eururo.2016.08.065

162. Andreoiu M, Cheng L. Multifocal prostate cancer: biologic, prognostic, and therapeutic implications. 2010;41:781-793. doi:10.1016/j.humpath.2010.02.011

163. Corcoran NM, Hovens CM, Hong MKH, et al. Underestimation of Gleason score at prostate biopsy reflects sampling error in lower volume tumours. 2012;109(5):660-664. doi:10.1111/j.1464-410X.2011.10543.x

164. Tomlins SA, Day JR, Lonigro RJ, et al. Urine TMPRSS2:ERG Plus PCA3 for Individualized Prostate Cancer Risk Assessment. 2016;70(1):45-53. doi:10.1016/j.eururo.2015.04.039

165. Deantoni EP, Crawford ED, Oesterling JE, et al. Age- and race-specific reference ranges for prostate-specific antigen from a large community-based study. 1996;48(2):234-239. doi:10.1016/S0090-4295(96)00091-X

166. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037

167. R Core Team. R: A Language and Environment for Statistical Computing. 2019. `https://www.r-project.org/`.

168. Archer KJ, Williams AAA. L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. 2012;31(14):1464-1474. doi:10.1002/sim.4484

169. Tibshirani R. Regression Shrinkage and Selection via the Lasso. 1996;58:267-288. doi:10.2307/2346178

170. Christensen RHB. ordinal Regression Models for Ordinal Data. 2018.

171. Pellegrini KL, Patil D, Douglas KJS, et al. Detection of prostate cancer-specific transcripts in extracellular vesicles isolated from post-DRE urine. 2017;77(9):990-999. doi:10.1002/pros.23355

172. Aghazadeh MA, Frankel J, Belanger M, et al. National Comprehensive Cancer Network® Favorable Intermediate Risk Prostate Cancer—Is Active Surveillance Appropriate? 2018;199(5):1196-1201. doi:10.1016/j.juro.2017.12.049

173. Cuzick J, Berney DM, Fisher G, et al. Prognostic value of a cell cycle progression signature for prostate cancer death in a conservatively managed needle biopsy cohort. 2012;106(6):1095-1099. doi:10.1038/bjc.2012.39

174. Knezevic D, Goddard AD, Natraj N, et al. Analytical validation of the Oncotype DX

prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies. 2013;14(1):690. doi:10.1186/1471-2164-14-690

175. Robert G, Jannink S, Smit F, et al. Rational basis for the combination of PCA3 and TMPRSS2:ERG gene fusion for prostate cancer diagnosis. 2013;73(2):113-120. doi:10.1002/pros.22546

176. Thompson IM, Pauler DK, Goodman PJ, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level ≤4.0 ng per Milliliter. 2004;350(22):2239-2246. doi:10.1056/NEJMoa031918

177. Simpkin AJ, Tilling K, Martin RM, et al. Systematic review and meta-analysis of factors determining change to radical treatment in active surveillance for localized prostate cancer. 2015;67:993-1005. doi:10.1016/j.eururo.2015.01.004

178. Tomlins SA, Bjartell A, Chinnaiyan AM, et al. ETS Gene Fusions in Prostate Cancer: From Discovery to Daily Clinical Practice. *European Urology*. 2009;56(2):275-286. doi:10.1016/j.eururo.2009.04.036

179. Wolpert DH, Macready WG. No free lunch theorems for optimization. 1997;1(1):67-82. doi:10.1109/4235.585893

180. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *Bmj*. 2009;338.

181. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: Validating a prognostic model. *Bmj*. 2009;338:b605.

182. Tolkach Y, Kristiansen G. The Heterogeneity of Prostate Cancer: A Practical Approach. *Pathobiology*. 2018;85(1-2):108-116. doi:10.1159/000477852

183. Kersting K. Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in Big Data*. 2018;1:6. doi:10.3389/fdata.2018.00006

184. Danks D. Learning. In: Frankish K, Ramsey WM, eds. *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press; 2014:151-167. doi:10.1017/CBO9781139046855.011

185. Draper NR, Smith H. "Dummy" variables. In: *Applied Regression Analysis*. John Wiley & Sons, Ltd; 2014:299-325. doi:10.1002/9781118625590.ch14

186. Breiman L. Bagging predictors. *Machine Learning*. 1996;24(2):123-140. doi:10.1023/A:1018054314350

187. 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):361-387.

188. Steyerberg EW, others. *Clinical Prediction Models*. Springer; 2019.

189. Guyon I, Elisseeff A. An introduction to variable and feature selection. 2003;3(Mar):1157-1182.

190. Brown M. rmda: Risk Model Decision Analysis. 2017. `https://cran.r-project.org/package=rmda`.

191. Ciccarese C, Massari F, Iacovelli R, et al. Prostate cancer heterogeneity: Discovering novel molecular targets for therapy. 2017;54:68-73. doi:https://doi.org/10.1016/j.ctrv.2017.02.001

192. Zhao F, Olkhov-Mitsel E, Kamdar S, et al. A urine-based DNA methylation assay, ProCUrE, to identify clinically significant prostate cancer. 2018;10(1):147. doi:10.1186/s13148-018-0575-z

193. Xia Y, Huang C-C, Dittmar R, et al. Copy number variations in urine cell free DNA as biomarkers in advanced prostate cancer. 2016;7(24):35818-35831. doi:10.18632/oncotarget.9027

194. Killick E, Morgan R, Launchbury F, et al. Role of Engrailed-2 (EN2) as a prostate cancer detection biomarker in genetically high risk men. 2013;3:2059. doi:10.1038/srep02059

195. Strand SH, Bavafaye-Haghighi E, Kristensen H, et al. A novel combined miRNA and methylation marker panel (miMe) for prediction of prostate cancer outcome after radical prostatectomy. *International Journal of Cancer*. 2019;145(12):3445-3452. doi:10.1002/ijc.32427

196. Ricketts CJ, De Cubas AA, Fan H, et al. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Reports*. 2018;23(1):313-326.e5. doi:10.1016/j.celrep.2018.03.075

197. The Human Protein Atlas. Expression of GJB1 in cancer. `https://www.proteinatlas.org/ENSG00000169562-GJB1/pathology`. Accessed May 24, 2019.

198. Tomlins SA, Laxman B, Varambally S, et al. Role of the TMPRSS2-ERG gene fusion in prostate cancer. 2008;10(2):177-188. doi:10.1593/neo.07822

199. Morgan R. Engrailed: Complexity and economy of a multi-functional transcription factor. 2006;580:2531-2533. doi:10.1016/j.febslet.2006.04.053

200. Membrane insertion and secretion of the Engrailed-2 (EN2) transcription factor by prostate cancer cells may induce antiviral activity in the stroma. *Scientific Reports*. 2019;9(1):1-9. doi:10.1038/s41598-019-41678-0

201. Pandha H, Sorensen KD, Orntoft TF, et al. Urinary engrailed-2 (EN2) levels predict tumour volume in men undergoing radical prostatectomy for prostate cancer. *BJU International*. 2012;110(6B). doi:10.1111/j.1464-410X.2012.11208.x

202. Riegman PHJ, Vlietstra RJ, Klaassen P, et al. The prostate-specific antigen gene and the human glandular kallikrein-1 gene are tandemly located on chromosome 19. *FEBS Letters*. 1989;247(1):123-126. doi:10.1016/0014-5793(89)81253-0

203. Barr DB, Wilder LC, Caudill SP, Gonzalez AJ, Needham LL, Pirkle JL. Urinary creatinine concentrations in the U.S. population: Implications for urinary biologic monitoring measurements. *Environmental Health Perspectives*. 2005;113(2):192-200. doi:10.1289/ehp.7337

204. Heinze G, Dunkler D. Five myths about variable selection. 2017;30:6-10.

doi:10.1111/tri.12895

205. Tamura H, Ishikawa Y, Hino N, et al. Neuropsin is essential for early processes of memory acquisition and Schaffer collateral long-term potentiation in adult mouse hippocampus in vivo. *Journal of Physiology.* 2006;570(3):541-551. doi:10.1113/jphysiol.2005.098715

206. Ovaere P, Lippens S, Vandenabeele P, Declercq W. The emerging roles of serine protease cascades in the epidermis. 2009;34:453-463. doi:10.1016/j.tibs.2009.08.001

207. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York; 2016. `https://ggplot2.tidyverse.org`.

208. Greenwell BM. pdp: An R package for constructing partial dependence plots. *R Journal.* 2017;9(1):421-436. doi:10.32614/rj-2017-016

209. Metzger J, Negm AA, Plentz RR, et al. Urine proteomic analysis differentiates cholangiocarcinoma from primary sclerosing cholangitis and other benign biliary disorders. *Gut.* 2013;62(1):122-130. doi:10.1136/gutjnl-2012-302047

210. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software.* 2010;33(1):1-22. `http://www.jstatsoft.org/v33/i01/`.

211. CE–MS-based urinary biomarkers to distinguish non-significant from significant prostate cancer. *British Journal of Cancer.* 2019;120(12):1120-1128. doi:10.1038/s41416-019-0472-z

212. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2005;67(2):301-320. doi:10.1111/j.1467-9868.2005.00503.x

213. Sakhanenko NA, Galas DJ. Biological Data Analysis as an Information Theory Problem: Multivariable Dependence Measures and the Shadows Algorithm. *Journal of Computational Biology.* 2015;22(11):1005-1024. doi:10.1089/cmb.2015.0051

214. Patel HD, Chalfin HJ, Carter HB. Improving Prostate Cancer Screening and Diagnosis. 2016;2(7):867-868. doi:10.1001/jamaoncol.2016.0170

215. Diamandis EP. Cancer Biomarkers: Can We Turn Recent Failures into Success? *JNCI: Journal of the National Cancer Institute.* 2010;102(19):1462-1467. doi:10.1093/jnci/djq306

216. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology.* 2001;54(8):774-781. doi:10.1016/S0895-4356(01)00341-9

217. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. 2012;98:683-690. doi:10.1136/heartjnl-2011-301246

218. Steyerberg KGMA van der W Ewout W. AND Moons. Prognosis research strategy (progress) 3: Prognostic model research. *PLOS Medicine.* 2013;10(2):1-9.

doi:10.1371/journal.pmed.1001381

219. Webb M, Manley K, Olivan M, et al. Methodology for the at-home collection of urine samples for prostate cancer detection. *BioTechniques*. 2020;68(2):65-73. doi:10.2144/btn-2019-0092

220. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology*. 2016;74:167-176. doi:10.1016/j.jclinepi.2015.12.005

221. Van Calster B, Vickers AJ. Calibration of risk prediction models: Impact on decision-analytic performance. *Medical Decision Making*. 2015;35(2):162-169. doi:10.1177/0272989X14547233

222. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: A systematic review. *PLoS Medicine*. 2012;9(5):1-12. doi:10.1371/journal.pmed.1001221

223. Wessler BS, Paulus J, Lundquist CM, et al. Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. *Diagnostic and Prognostic Research*. 2017;1(1):10. doi:10.1186/s41512-017-0021-2

224. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004

225. Kleinrouweler CE, Cheong-See FM, Collins GS, et al. Prognostic models in obstetrics: Available, but far from applicable. *American Journal of Obstetrics and Gynecology*. 2016;214(1):79-90.e36. doi:10.1016/j.ajog.2015.06.013

226. Collins GS, De Groot JA, Dutton S, et al. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. 2014;14:40. doi:10.1186/1471-2288-14-40

227. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. *BMC Medicine*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7

228. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*. 2017;24(6):1052-1061. doi:10.1093/jamia/ocx030

229. Thai TN, Ebell MH. Prospective validation of the Good Outcome Following Attempted Resuscitation (GO-FAR)score for in-hospital cardiac arrest prognosis. *Resuscitation*. 2019;140:2-8. doi:10.1016/j.resuscitation.2019.05.002

230. Office for National Statistics. User guide to mortality statistics - Office for National Statistics. 2019:1-45. `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017%20https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017#certification-of-cause-of-death`.

231. Reilly BM, Evans AT. Translating clinical research into clinical practice: Impact of

using prediction rules to make decisions. 2006;144:201-209. doi:10.7326/0003-4819-144-3-200602070-00009

232. Andersen PK. Multistate models in survival analysis: A study of nephropathy and mortality in diabetes. *Statistics in Medicine*. 1988;7(6):661-670. doi:10.1002/sim.4780070605

233. Kvamme H, Borgan O, Scheel I. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*. 2019;20:1-30. `http://jmlr.org/papers/v20/18-424.html`.

234. 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):1-13. doi:10.18637/jss.v039.i05

235. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Annals of Applied Statistics*. 2008;2(3):841-860. doi:10.1214/08-AOAS169

236. Heidegger I, Klocker H, Steiner E, et al. ProPSA is an early marker for prostate cancer aggressiveness. *Prostate Cancer and Prostatic Diseases*. 2014;17(1):70-74. doi:10.1038/pcan.2013.50

237. Catalona WJ, Partin AW, Sanda MG, et al. A multicenter study of [-2]pro-prostate specific antigen combined with prostate specific antigen and free prostate specific antigen for prostate cancer detection in the 2.0 to 10.0 ng/ml prostate specific antigen range. *Journal of Urology*. 2011;185(5):1650-1655. doi:10.1016/j.juro.2010.12.032

238. Loeb S, Catalona WJ. The Prostate Health Index: A new test for the detection of prostate cancer. 2014;6:74-77. doi:10.1177/1756287213513488

239. Vickers AJ, Sjoberg DD, Ankerst DP, Tangen CM, Goodman PJ, Thompson IM. The Prostate Cancer Prevention Trial risk calculator and the relationship between prostate-specific antigen and biopsy outcome. 2013;119(16):3007-3011. doi:10.1002/cncr.28114

240. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. *Statistics in Medicine*. 2019;38(7):1262-1275. doi:10.1002/sim.7993

241. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*. 2019;38(7):1276-1296. doi:10.1002/sim.7992

242. Jong VMT de, Eijkemans MJC, Calster B van, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. *Statistics in Medicine*. 2019;38(9):1601-1619. doi:10.1002/sim.8063

243. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: Beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology*. 2011;64(9):993-1000. doi:10.1016/j.jclinepi.2010.11.012

244. Van Smeden M, De Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*. 2016;16(1):1-12. doi:10.1186/s12874-016-0267-3

245. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*. 2019;28(8):2455-2474. doi:10.1177/0962280218784726

246. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *The BMJ*. 2020;368. doi:10.1136/bmj.m441

247. Bruinsma SM, Zhang L, Roobol MJ, et al. The Movember Foundation's GAP3 cohort: a profile of the largest global prostate cancer active surveillance database to date. *BJU International*. 2018;121(5):737-744. doi:10.1111/bju.14106

248. Canadian Cancer Society. The prostate. `http://www.cancer.ca/en/cancer-information/cancer-type/prostate/prostate-cancer/the-prostate/?region=on`. Accessed January 22, 2018.

249. Egevad L, Granfors T, Karlberg L, Bergh A, Stattin P. Prognostic value of the Gleason score in prostate cancer. *BJU International*. 2002;89(6):538-542. doi:10.1046/j.1464-410X.2002.02669.x

250. National Intutute for Health. Morphology & Grade | SEER Training. `https://training.seer.cancer.gov/prostate/abstract-code-stage/morphology.html`. Accessed June 22, 2020.

251. Abdominal Key. Nonneoplastic diseases of the prostate | Abdominal Key. 2016. `https://abdominalkey.com/nonneoplastic-diseases-of-the-prostate/`. Accessed July 20, 2020.