UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA

Prostate Cancer Biochemical Recurrence Prediction using bpMRI Radiomics, Clinical and Histopathological Data

Mónica Leiria de Mendonça Miranda da Silva

Mestrado Integrado em Engenharia Biomédica e Biofísica
Perfil em Sinais e Imagens Médicas

Dissertação orientada por:
Dr. Nickolas Papanikolaou
Dr. Nuno Matela

2021

Acknowledgements

I would like to thank my supervisor Dr. Nickolas Papanikolaou for giving me the opportunity to develop this project at the Champalimaud Foundation, providing guidance, insight and understanding.

I would also like express my gratitude towards Professor Dr. Nuno Matela, who also supervised this project, for always showing comprehension and availability, and for motivating me throughout the years as my professor at Faculdade de Ciências da Universidade de Lisboa.

I wish to thank Dr. Jorge Fonseca, for his openness to share with me his specialised knowledge on prostate cancer clinical care and for clearing my doubts. Also, I thank Dr. Ana Gaivão, for giving me essential initial training on MRI prostate anatomy, and Dr. António Beltran, for so kindly assisting me during histopathological data collection and curation.

I also extend my sincere thanks to my colleagues from Computational Clinical Imaging group, to our daily companions from Radiopharmacology research group, and to the friends within the Champalimaud Research community. Your presence and companionship made this internship so joyful.

I wish to express my heartfelt appreciation towards special people who have supported me throughout the course of this project. To Davide, for consistently giving me crucial support and understanding, and for providing advice and guidance. To João, for his immense generosity, unconditional support, motivation and help. To Dianne, for her soul-filling friendship, but also for the image editing and English proofreading in a field far from her professional scope. To Carolina S., for the daily companionship and for being by my side when most needed. To Andreia, for her availability, essential exchange of ideas and clearing of doubts. And to Carolina D., for her kindness, helpful specialised knowledge and discussions.

Finally, I would like to acknowledge with deep gratitude the support of all my dearest friends and family, who have shown me never-ending love, understanding and courage.

Resumo

O cancro da próstata é a segunda doença oncológica mais frequente nos homens, sendo frequentemente tratado com remoção cirúrgica total do órgão, denominada prostatectomia radical. Apesar dos avanços no diagnóstico e da evolução das terapias cirúrgicas, 20–35% dos candidatos a prostatectomia radical com intuito curativo sofrem de recidiva bioquímica, uma condição que representa o insucesso do tratamento inicial e também o primeiro sinal de progressão da doença. Em particular, dois terços dos casos de recidiva bioquímica ocorrem dentro de um período de dois anos. Ocorrendo cedo, este estado implica uma maior agressividade biológica da doença e um pior prognóstico, uma vez que pode dever-se à presença de doença oculta, localmente avançada ou metastática. Apesar de o prognóstico devido ao desenvolvimento de recidiva bioquímica variar, em geral está associado a um risco acrescido de desenvolvimento de doença metastática e de mortalidade específica por cancro da próstata, representando assim uma importante preocupação clínica após terapia definitiva.

Contudo, os modelos preditivos de recidiva bioquímica actuais não só falham na explicação da variabilidade dos resultados pós-cirúrgicos, como não têm habilidade para intervir cedo no processo de decisão de tratamento, uma vez que dependem de informação provinda da avaliação histopatológica da peça cirúrgica da prostatectomia ou da biópsia.

Actualmente, o exame padrão para diagnóstico e para estadiamento do cancro da próstata é a ressonância magnética multiparamétrica, e as características provindas da avaliação dessas imagens têm mostrado potencial na caracterização do(s) tumor(es) e para predição de recidiva bioquímica. "Radiomics", a recente metodologia aplicada à análise quantitativa de imagens médicas tem mostrado ter capacidade de quantificar objectivamente a heterogeneidade macroscópica de tecidos biológicos como tumores. Esta heterogeneidade detectada tem vindo a sugerir associação a heterogeneidade genómica que, por sua vez, tem demonstrado correlação com resistência a tratamento e propensão metastática. Porém, o potencial da análise radiómica das imagens de ressonância magnética (MRI) multiparamétrica da próstata para previsão de recidiva bioquímica pós-prostatectomia radical ainda não foi totalmente aprofundado.

Esta dissertação propôs explorar o potencial da análise radiómica aplicada a imagens pré-cirúrgicas de ressonância magnética biparamétrica da próstata para previsão de recidiva bioquímica, no período de dois anos após prostatectomia radical. Este potencial foi avaliado através de modelos predictivos com base em dados radiómicos e parâmetros clinicohistopatológicos comummente adquiridos em três fases clínicas: pré-biópsia, pré- e pós-cirúrgica.

93 pacientes, de um total de 250, foram eleitos para este estudo retrospectivo, dos quais 20 verificaram recidiva bioquímica. 33 parâmetros clínico-histopatológicos foram recolhidos e 2715 variáveis radiómicas baseadas em intensidade, forma e textura, foram extraídas de todo o volume da próstata caracterizado em imagens originais e filtradas de ressonância magnética biparamétrica, nomeadamente, ponderadas em T2, ponderadas em Difusão, e mapas de coeficiente

de difusão aparente (ADC).

Embora os pacientes elegíveis tenham sido examinados na mesma instituição, as características do conjunto de imagens eram heterogéneas, sendo necessário aplicar vários passos de processamento para possibilitar uma comparação mais justa. Foi feita correcção do campo tendencial (do inglês, “bias”) e segmentação manual das imagens T2, registo tanto para transposição das delineações do volume de interesse entre as várias modalidades imagiológicas como para correcção de movimento, cálculo de mapas ADC, regularização do campo de visão, quantização personalizada em tons cinza e reamostragem.

Tendo os dados recolhidos uma alta dimensionalidade (número de variáveis maior que o número de observações), foi escolhida a regressão logística com penalização L_1 (LASSO) para resolver o problema de classificação. O uso da penalização aliada à regressão logística, um método simples e commumente usado em estudos de classificação, permite impedir o sobreajuste provável neste cenário de alta dimensionalidade. Além do popular LASSO, recorremos também ao algoritmo Priority-LASSO, um método recente para lidar com dados “ómicos” e desenvolvido com base no LASSO. O Priority-LASSO tem como princípio a definição da hierarquia ou prioridade das variáveis de estudo, através do agrupamento dessas mesmas variáveis em blocos sequenciais. Neste trabalho explorámos duas maneiras de agrupar as variáveis (Clínico-histopatológicas vs. Radiómicas e Ciínico-histopatológicas vs. T2 vs. Difusão vs. ADC). Além disso, quisemos perceber qual o impacto da ordem destes mesmos blocos no desempenho do modelo. Para tal, testámos todas as permutações de blocos possíveis (2 e 24, respectivamente) em cada um dos casos.

Assim, uma estrutura de aprendizagem automática, composta por métodos de classificação, validação-cruzada *k-fold* estratificada e repetida, e análises estatísticas, foi desenvolvida para identificar os melhores classificadores, dentro um conjunto de configurações testado para cada um dos três cenários clínicos simulados. Os algoritmos de regressão logística penalizada com LASSO e o Priority-LASSO efectuaram conjuntamente a selecção de características e o ajuste de modelos. Os modelos foram desenvolvidos de forma a optimizar o número de casos positivos de recidiva bioquímica através da maximização das métricas área sob a curva (AUC) e medida-F (*Fmax*), derivadas da análise de curva característica de operação do receptor (ROC).

Além da comparação das implementações Priority-LASSO com o caso em que não houve agrupamento de variáveis (isto é, LASSO), foram também comparados dois métodos de normalização de imagens com base no desempenho dos modelos (avaliado por *Fmax*). Um dos métodos tinha em conta o sinal de intensidade proveniente da próstata e de tecidos imediatamente circundantes, e outro apenas da próstata. Paralelamente, também o efeito do método de amostragem SMOTE, que permite equilibrar o número de casos positivos e negativos durante o processo de aprendizagem do algoritmo, foi avaliado no desempenho dos modelos. Com este método, gerámos casos sintéticos para a classe positiva (classe minoritária) para recidiva bioquímica, a partir dos casos já existentes.

O modelo de regressão logística com Priority-LASSO com a sequência de blocos de variáveis *Clínico-histopatológicas, T2, Difusão, ADC* e com restrição de esparsidade de cada bloco com o parâmetro *pmax* = (1,7,0,1), foi seleccionada como a melhor configuração em cada um dos cenários clínicos testados, superando os modelos de regressão logística LASSO.

Durante o desenvolvimento dos modelos, e em todos os cenários clínicos, os modelos com melhor desempenho obtiveram bons valor médios de *Fmax* (mínimo–máximo: 0.702–0.754 e 0.910–0.925 para classe positiva e negativa de recidiva bioquímica, respectivamente). Contudo,

na validação final com um conjunto de dados independentes, os modelos obtiveram valores *Fmax* muito baixos para a classe positiva (0.297–0.400), revelando um sobreajuste, apesar do uso de métodos de penalização. Também se verificou grande instabilidade nos atributos seleccionados. Contudo, os modelos obtiveram razoáveis valores de medida-F (0.779–0.833) e de Precisão (0.821–0.873) para a classe de recidiva bioquímica negativa durante as fases de treino e de validação, pelo que estes modelos poderão ter valor a ser explorado.

Os modelos pré-biopsia tiveram desempenho inferior no treino, mas sofreram menos de sobreajuste. Os classificadores pré-operatórios foram excessivamente optimistas, e os modelos pós-operatórios foram os melhores a detectar correctamente casos negativos de recidiva bioquímica.

Outros resultados observados incluem a superioridade no desempenho dos modelos baseados em imagens que usaram o método de normalização realizado apenas com o volume da próstata, e o inesperado resultado de que o uso método de amostragem SMOTE não ter trazido melhoria na classificação de casos positivos de recorrência bioquímica, nem nos casos negativos, durante a validação dos modelos.

Tendo em contas às variáveis seleccionadas e a sequência de prioridade dos melhores modelos Priority-LASSO, concluímos que os atributos radiómicos provindos da análise de textura de imagens MRI ponderadas em T2 poderão ter potencial para distinguir pacientes que não irão sofrer recidiva bioquímica inicial, conjuntamente com níveis iniciais de antigénio específico da próstata, num cenário pré-biópsia. A inclusão de parâmetros pré- ou pós-operatórios não adicionou valor substancial para a classificação de casos positivos de recidiva bioquímica em conjunto com variáveis radiómicas de MRI biparamétrica. Estudos com alto poder estatístico serão necessários para elucidar acerca do papel de atributos de radiómica baseados em imagens de bpMRI como predictores de recidiva bioquímica.

Palavras-Chave: Recidiva Bioquímica; Cancro da Próstata; *Radiomics*; bpMRI; LASSO.

Abstract

Primary prostate cancer is often treated with radical prostatectomy (RP). Yet, 20–35% of males undergoing RP with curative intent will experience biochemical recurrence (BCR). Of those, two-thirds happen within two years, implying a more aggressive disease and poorer prognosis. Current BCR risk stratification tools are bounded to biopsy- or to surgery-derived histopathological evaluation, having limited ability for early treatment decision-making. Magnetic resonance imaging (MRI) is acquired as part of the diagnostic procedure and imaging-derived features have shown promise in tumour characterisation and BCR prediction. We investigated the value of imaging features extracted from preoperative biparametric MRI (bpMRI) combined with clinicohistopathological data to develop models to predict two-year post-prostatectomy BCR in three simulated clinical scenarios: pre-biopsy, pre- and postoperative.

In a cohort of 20 BCR positive and 73 BCR negative RP-treated patients examined in the same institution, 33 clinicohistopathological variables were retrospectively collected, and 2715 radiomic features (based on intensity, shape and texture) were extracted from the whole-prostate volume imaged in original and filtered T2- and Diffusion-weighted MRI and ADC maps scans. A systematic machine-learning framework comprised of classification, stratified *k*-fold cross-validation and statistical analyses was developed to identify the top performing BCR classifiers' configurations within three clinical scenarios. LASSO and Priority-LASSO logistic regression algorithms were used for feature selection and model fitting, optimising the amount of correctly classified BCR positive cases through AUC and F-score maximisation (*Fmax*) derived from ROC curve analysis. We also investigated the impact of two image normalisation methods and SMOTE-based minority oversampling on model performance.

Priority-LASSO logistic regression with four-block priority sequence *Clinical, T2w, DWI, ADC*, with block sparsity restriction *pmax* = (1,7,0,1) was selected as the best performing model configuration across all clinical scenarios, outperforming LASSO logistic regression models. During development and across the simulated clinical scenarios, top models achieved good median *Fmax* values (range: 0.702–0.754 and 0.910–0.925 for BCR positive and negative classes, respectively); yet, during validation with an independent set, the models obtained very low *Fmax* for the target BCR positive class (0.297–0.400), revealing model overfitting. We also observed instability in the selected features. However, models attained reasonably good F-score (0.779–0.833) and Precision (0.821–0.873) for BCR negative class during training and validation phases, making these models worth exploring. Pre-biopsy models had lower performances in training but suffered less from overfitting. Preoperative classifiers were overoptimistic, and postoperative models were the most successful in detecting BCR negative cases.

T2w-MRI textured-based radiomic features may have potential to distinguish negative BCR patients together with baseline prostate-specific antigen (PSA) levels in a pre-biopsy scenario. The inclusion of pre- or postoperative variables did not substantially add value to BCR positive cases classification with bpMRI radiomic features. Highly powered studies with curated imaging data are needed to elucidate the role of bpMRI radiomic features as predictors of BCR.

Keywords: Prostate Cancer; Biochemical Recurrence; Radiomics, bpMRI; LASSO.

Contents

List of Figures

List of Tables

List of Abbreviations

AUC	Area Under the Curve
BCR	Biochemical Recurrence
bpMRI	Biparametric Magnetic Resonance Imaging
CT	Computed Tomography
CV	Cross-Validation
CZ	Central Zone
DRE	Digital Rectal Examination
DW	Diffusion-Weighted
EAU	European Association of Urology
EPE	Extraprostatic Extension
ESUR	European Society of Urogenital Radiology
FOV	Field of View
GS	Gleason Score
IQR	Interquartile Range
ISUP	International Society of Urological Pathology
ITN	Index Tumour Nodule
(k-)NN	(k-)Nearest Neighbour
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Logistic Regression
ML	Machine Learning
mpMRI	Multiparametric Magnetic Resonance Imaging
MRI	Magnetic Resonance Imaging
PCa	Prostate Cancer

PET	Positron Emission Tomography
PI-RADS	Prostate Imaging-Report and Data System
P-LASSO	Priority-LASSO
PLND	Pelvic Lymph Node Dissection
PLR	Penalised Logistic Regression
PSA	Prostate-Specific Antigen
PZ	Peripheral Zone
RARP	Robot-Assisted Radical Prostatectomy
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RP	Radical Prostatectomy
RT	Radiotherapy
SI	Signal Intensity
SMOTE	Synthetic Minority Over-sampling Technique
T2w	T2-Weighted
TZ	Transitional Zone
VOI	Volume of Interest

Motivation and Project Outline

Prostate cancer (PCa) is a major health concern in the male population. With incidence mainly dependent on age, the disease affects more than 40% of men over the age of 60, and it is the most frequently diagnosed type of cancer among men in over half of the countries in the world [1, 2]. Although often having an indolent course, PCa remains the second-leading cause of cancer death in men.

Advances in the diagnosis through improved risk classification methods, imaging techniques and biomarkers, and the development of PCa therapeutics have enhanced the ability to stratify patients by risk and allowed clinicians to recommend therapy based on cancer prognosis and patient preference. Particularly, prostate surgery removal techniques continue to evolve as treatment-related adverse effects are better defined. However, candidates for initial treatment with radical prostate surgery removal (radical prostatectomy, RP) often suffer from biochemical recurrence (BCR), a condition that represents treatment failure and the first sign of disease progression. Although failed treatment outcomes may vary, BCR is associated with increased risk for metastatic disease and PCa-specific mortality. If present in early stages, BCR usually implies a biologically more aggressive disease and poor prognosis, possibly due to the presence of occult disease, whether locally advanced or metastatic.

The current clinical parameters fail to explain some of the outcome variability. Decision-making regarding treatment is challenging due to the current PCa clinical difficulties in evaluating the aggressiveness and progression of the disease.

Thus, it is necessary to understand why primary treatment for localised disease assigned with curative intent leads to unsuccessful outcomes, such as BCR occurrence, in a significant proportion of patients. Performing BCR risk-prediction in an early-stage of the diagnosis (before radical treatment) is reasonably the criteria of a successful and clinically useful prognostic tool.

Radiomics have been showing potential in characterising tumour image heterogeneity. Findings suggest association of image heterogeneity with genomic heterogeneity, and correlation with increased treatment resistance and metastatic propensity [3, 4]. We believe that the potential of radiomics analysis of preoperative standard-of-care magnetic resonance imaging (MRI) for early BCR prediction after radical surgery has not been fully explored yet. Moreover, radiomic characterisation of the whole prostate pre-surgical condition for prognostic purposes after surgery is still unexplored. Identifying patients with high risk of BCR prior to primary RP would allow for effective decision-making throughout PCa management. It would also enhance patient counselling, enabling rational application of treatment intensification and, consequently, improve patients' prognosis.

The different sections of this dissertation focus on the clinical problem of prostate cancer biochemical recurrence after radical prostatectomy treatment, where the role of biparametric magnetic resonance imaging (bpMRI) was explored with radiomics analysis for the development of a model for early-BCR prediction. Clinical and histopathological data routinely assessed in different stages of PCa diagnosis were also included and evaluated in this project.

The content of this dissertation is organised as follows:

Chapter 1 includes a description of prostate cancer and its standard of care, where we state how biochemical recurrence is defined, detected and managed.

The main concepts of the methodology approach are introduced in **Chapter 2**, describing the process of Radiomics analysis. An overview of supervised machine learning is given, focusing on penalised logistic regression modeling.

In **Chapter 3** we describe the clinicohistopathological and image data used and image processing methodology applied. The radiomics analysis and machine-learning framework is announced.

Chapter 4 presents the main findings for each proposed model and **Chapter 5** discusses the results, the models used to achieve them, their characteristics and limits. Considerations of the techniques applied are also provided, as well as limitations of this study and avenues for future work.

Finally, in **Chapter 6**, we discuss the results obtained in the proposed study.

1 INTRODUCTION

1.1 Prostate Anatomy and Carcinoma

The prostate gland is an exocrine gland of the male reproductive system, with an inverted pyramidal shape. It is located in the pelvic region, directly inferior to the bladder and in front of the rectum, as depicted in Figure 1.1. Superiorly to the prostate are the seminal vesicles, two paired glands producers of seminal fluid, that become linked to the ejaculatory duct at the base of the prostate. Within the prostate sits the (prostatic) urethra coming from the bladder. Exteriorly, the gland is enclosed with a fibrous “capsule”, with nerves and vascular plexus. Veins around the prostate drain into internal iliac veins, and lymphatic drainage is made both to the internal iliac and the sacral lymph nodes.

Measuring approximately 3 cm in height by 2.5 cm in depth, prostate estimated weight is from 7 to 16 g for an adult [5]. The gland size increases at two distinct stages during physical development: initially at puberty to reach its normal size, then again after 60 years of age leading to benign prostatic hyperplasia [5]. This gland contributes with 20–30% of the fluid emitted from the male reproductive tract, facilitating a successful fertilisation [6].

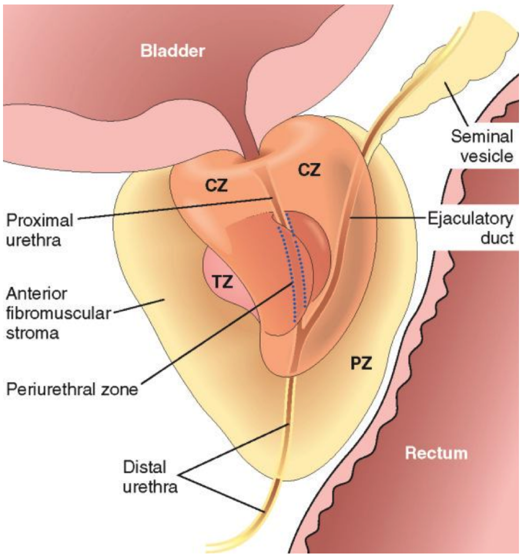
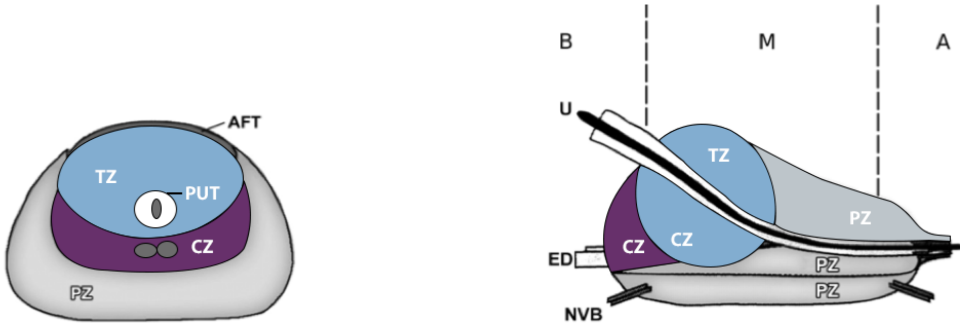


Figure 1.1: Prostate gland localisation within the male reproductive system (sagittal anatomy scheme from [7]).

The prostate gland is typically separated into three zones: peripheral (PZ), central (CZ) and transitional (TZ) zones (illustrated in Figures 1.1, 1.2 and 1.3). These zones, respectively,



(a) Transverse anatomy of the prostate.

(b) Sagittal anatomy of the prostate.

Figure 1.2: Prostate anatomy zonal division in transverse (a) and sagittal (b) planes. AFT: anterior fibromuscular tissue, CZ: central zone, ED: ejaculatory duct, NVB: neurovascular bundle, PUT: periurethral tissue, PZ: peripheral zone, U: urethra, TZ: transitional zone, B: base, M: median gland, A: apex (adapted from [9]).

constitute 70%, 25% and 5% of the prostate gland [8]. The prostate is divided into three longitudinal portions depicted in Fig. 1.2b: (i) Base, (ii) Median gland, and (iii) Apex.

The prostate gland undergoes vast changes during a man's lifetime. Along with prostatitis and benign prostatic hyperplasia, prostate cancer (PCa) is a common pathological condition at later life stages.

PCa development predispositions varies between prostate's zones. 70%–80% of PCa disease originate in PZ, 10%–20% in TZ [11], and only 5% in CZ [12]. However, cases developing in CZ are typically more aggressive and more likely to invade other organs, due to their location [12].

This disease can grow with slow- or rapid-growing tumours. Slow disease, accounting for up to 85% of PCas cases, progresses slowly and usually stays confined to the gland. In contrast, rapid-growing tumours metastasise from the prostate to other organs, primarily the bones, significantly affecting the morbidity and mortality rate [13].

PCa is the most frequently diagnosed cancer among men in over half of the countries in the world, affecting more than 40% of men over the age of 60 [2]. Despite often having an indolent course, it is the second leading cause of oncological death in men worldwide, and first in 46 countries, counting in 2018, 1.3 million new cases and causing 359000 deaths [1]. The disease incidence mainly depends on age, with a prevalence ranging from 5% at age < 30 years, increasing by an odds ratio of 1.7 per decade, to a prevalence of 59% by age > 79 years [2].

Considering that the global population aged 60 or more is expected to nearly double by 2050 [14], the management of age-related diseases such as PCa, with incidence greatly increasing with age, presents a health challenge worldwide.

Despite PCa being a common disease, little is known about its etiology. A variety of exogenous and environmental factors may have an impact on PCa incidence and risk of progression (e.g. family history, genetics, obesity and diet) [15]. Thus, neither there is yet clinical applicability to screen for genetic PCa susceptibility, nor specific preventive protocols to reduce the risk of developing the disease.

Currently, only screening for PCa enables early detection. Yet, improvement of diagnosis may enable optimum choice of therapy, reducing the increasing PCa economic burden that comes especially from interventional treatment [16].

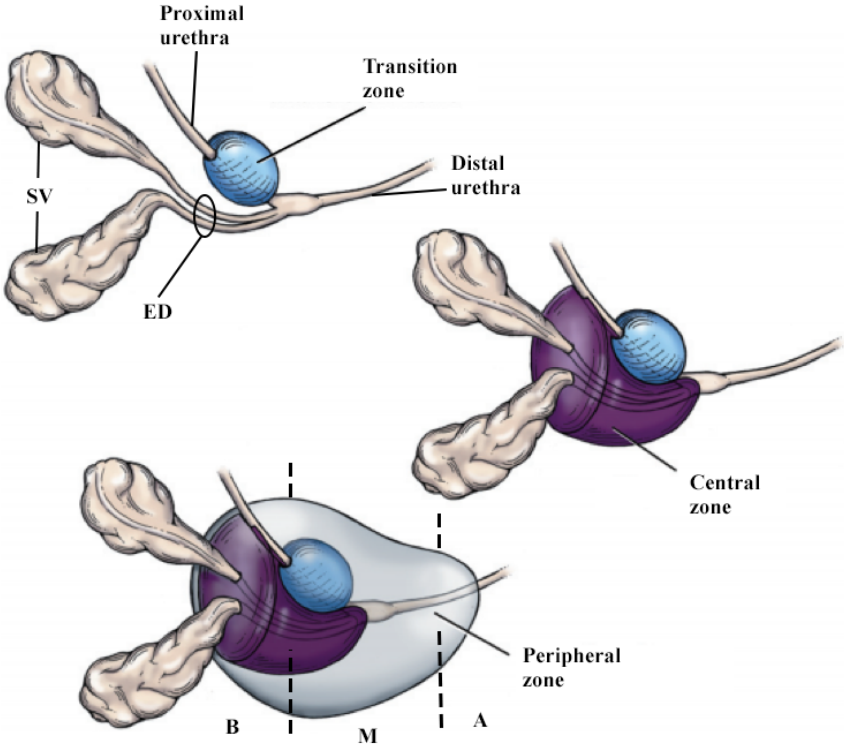


Figure 1.3: Adult prostate gland: zonal anatomy and relationships (adapted from [10]).

1.2 Prostate Cancer Clinical Management[1]

1.2.1 Screening

Prostate cancer screening is usually performed through measurement of blood prostate-specific antigen (PSA) levels. Yet, as PSA is organ- but not cancer-specific, its usage for screening remains controversial. Although there is evidence that PSA-based screening leads to a reduction in PCa-specific mortality, it does not seem to have an effect on overall mortality rates [18]. For this reason, there are ongoing debates and clinically unsolved challenges on the prevention of PCa overdiagnosis and overtreatment of clinically insignificant cases.

1.2.2 Clinical Diagnosis

Most PCa diagnoses are made in symptomatic men. The disease is usually suspected in men over 50 years old presenting with lower urinary tract symptoms, visible haematuria (presence of blood in the urine) or erectile dysfunction. However, the same symptoms are present in benign conditions affecting the prostate, making early-stage diagnosis challenging.

The current clinical tools and modalities for the diagnosis of PCa include digital rectal examination (DRE), PSA measurement, biopsy and imaging in the form of multiparametric magnetic resonance imaging (mpMRI) and transrectal ultrasound-guided scan.

[1] This section is written in accordance with the European Association of Urology Guidelines on Prostate Cancer, 2019 edition [17], a document that aims to assist medical professionals with curated evidence on the best current PCa clinical management practices.

1.2.2.1 Digital Rectal Examination

While DRE can detect tumour volumes greater than 0.2 mL, and an abnormal examination alone can detect of 18% of PCa cases [19], DRE abnormal findings usually lead to biopsy indication and require more diagnostic information from PSA levels and imaging evaluation.

1.2.2.2 Prostate-Specific Antigen

PSA is a multipurpose biomarker used in prostate screening, diagnosis and staging, as well as in monitoring of disease progression and treatment efficacy.

The usage of PSA serum concentration levels as a biomarker has revolutionised the diagnosis of PCa [20]. Despite not existing agreed standards for measuring PSA [21], higher serum concentration levels indicate greater likelihood of PCa, making PSA a better independent predictor of cancer than DRE or imaging with transrectal ultrasound [22].

Yet, PSA level assessment suffers from limitations that make the diagnosis of PCa more challenging. PSA is organ- but not cancer-specific, as it may be elevated in other non-malignant conditions of the prostate, such as benign prostatic hypertrophy or prostatitis. Also, many men may harbour PCa despite having low serum PSA, precluding definitions of optimal PSA thresholds for detecting clinically significant PCa.

Aiming to improve the predicting value of PSA and PCa detection, several variations of serum PSA have been described, taking into account the levels kinetics, density and age-specific reference ranges. However, their clinical value is limited, usually not providing additional information compared with PSA alone [23, 24].

Thus, PCa is clinically suspected following a positive DRE and elevated PSA levels after screening. Yet, definitive diagnosis depends on histopathological verification of adenocarcinoma in prostate biopsy cores specimens or, when undergone radical surgery, whole-prostate specimen.

1.2.2.3 Baseline Biopsy

The need for prostate biopsy is based on PSA level, suspicious DRE, or imaging (further discussed in section 1.2.2.4), where imaging can be acquired before or after biopsy.

On baseline biopsies, i.e. with no prior imaging with mpMRI, or when mpMRI findings were not suspicious, the biopsy procedure is standardly guided by ultrasound. A minimum of six different samples are taken from the prostate zones, from its apex, median, and base, covering both the left and right side.

Biopsy cores are examined individually with Gleason score system (GS) [25], assessing the aggressiveness of the tumour(s) according only to its cancerous architecture. An increased GS score implies a more abnormal glandular structure, that is associated with worse prognosis.

The overall biopsy GS is reported as a summation of the GS of the most prevalent and the GS of the worst Gleason patterns found (e.g. 4 + 3). The Gleason score has been adapted by the International Society of Urological Pathology (ISUP), leading to the ISUP grading system [26–28] (shown in Table 1.1). ISUP grading main characteristic is that it splits-up GS 7 cancers into ISUP grade 2 (primary Gleason grade 3) and ISUP grade 3 (primary Gleason grade 4) because of their distinct prognostic impact strengthens. Thus, it separates the intermediate-risk group into a low-intermediate (ISUP grade 2) and high intermediate-risk (ISUP grade 3) group.

Moreover, ISUP prognostic grade groups have been validated as predictive of BCR, response to therapy, and cancer-related mortality in several large-scale studies [26, 27].

Table 1.1: International Society of Urological Pathology 2014 grades relationship with Gleason score [26, 27].

Gleason score	**ISUP grade**
2–6	1
7 (3+4)	2
7 (4+3)	3
8 (4+4 or 3+5 or 5+3)	4
9–10	5

Table 1.2: PCa detection rates (%) by mpMRI for tumour volume and ISUP grade group in RP specimen [30]. mpMRI shows good sensitivity to detect and localise ISUP grade ≥ 2, but shows worst performance for lower ISUP grades and small tumour volumes.

ISUP grade group	**Tumour volume (mL)**		
	$< \mathbf{0.5}$	**0.5–2**	$> \mathbf{2}$
ISUP grade 1	21–29%	43–54%	67–75%
ISUP grade 2–3	63%	82–88%	97%
ISUP grade > 4	80%	93%	100%

1.2.2.4 Imaging of Primary Prostate Cancer

Multiparametric MRI refers to a MRI protocol typically consisting of high-resolution T2-weighted (T2w) anatomical scans, as well as two functional sequences: diffusion-weighted imaging (DWI) with derived apparent diffusion coefficient (ADC) maps, and dynamic contrast-enhanced (DCE) imaging [29]. When performed before biopsy, it is a sensitive tool to detect clinically relevant tumour and to localise sites for targeted biopsies [17].

The most relevant anatomical sequence for detection of primary PCa is a T2w scan. This sequence offers a good depiction of the gland anatomy, where a tumour region can be identified as hypointense. However, the tissues of the CZ and TZ are difficult to distinguish and are usually merged into a common region, denominated as central gland.

In PCa, the two main tissue properties probed with functional MRI are diffusion and perfusion. Diffusion-weighted imaging probes the random translational motion of water molecules. As tumour tissue is characterised by increased cellular density, these densely packed regions restrict the extracellular movement of water molecules, resulting in hyperintense regions in the DWI sequence. An apparent diffusion coefficient map can be derived from DWI images acquired at different gradient strengths. ADC map provides a quantitative measure of restricted diffusion for each voxel and is usually low in tumour regions. Dynamic contrast-enhanced MRI is used to quantify tissue perfusion by dynamically monitoring the distribution of injected contrast agent in the region of interest. The high energetic demand of growing tumour cells is accompanied by the formation of a new vascular bed which is often disorganised, inefficient and leaky.

mpMRI has progressively been improving and playing a crucial role on PCa clinical diagnosis. Association of T2-weighted imaging with at least one functional imaging technique (DWI, DCE) has shown good sensitivity to detect and localise PCa with ISUP grade ≥ 2 [30–32]. Yet, imaging assessments revealed less capable to identify ISUP grade 1 PCa [33], especially in case of small tumour volumes (detection rates in Table 1.2) [30].

mpMRI exams are also imperative in determining the biopsy approach or technique that should optimise diagnosis, in function of the patient setting (biopsy-naive or with prior negative-biopsy) [17]. This imaging technique has been increasingly used in biopsy procedures to localise suspicious areas [33].

1.2.2.5 Pathology of Prostate Needle Biopsies

Definitive pre-treatment diagnosis of PCa depends on histopathologic verification of biopsy specimens, where each core is processed separately. The diagnostic criteria include features characteristic for cancer, and major and minor features favouring and/or against carcinoma. Reports are individually made for each biopsy core, with inclusion of the location and histopathological findings – histological type and ISUP grade [26, 27] (Table 1.1). Finally, the global ISUP grade is reported, taking into account all biopsies positive for carcinoma, and, if present, intraductal carcinoma pattern, lymphovascular invasion, perineural invasion and extraprostatic extension (EPE) state.

1.2.2.6 Histopathology of Radical Prostatectomy Specimen

In the case of choice of local treatment with RP (discussed in Section 1.2.4.1), the definitive diagnosis of PCa is made by histopathology assessment of the RP specimen. Evaluation of the whole-gland specimen enables evaluation of cancer location, multifocality and heterogeneity.

The pathology report provides essential prognostic characteristics, relevant for further clinical decision-making. It describes PCa in terms of: (i) pathological stage (pTNM staging, see Appendix A), (ii) histopathological type (acinar, ductal or mixed), (iii) histological ISUP grade, and (iv) surgical margins status: location, extent and laterality of EPE, presence of bladder neck or seminal vesicle invasion, location and extent of positive surgical margins.

These pathologic findings comprise highly-valuable prognostic characteristics, essential for clinical decision-making and prognosis estimation [26].

1.2.3 Clinical Staging and the Role of Imaging

The extent of PCa is evaluated by DRE and PSA, and may be supplemented with mpMRI, bone scanning and computed tomography (CT), enabling Tumour, Nodal and Metastatic (TNM) staging. This staging can be based on clinical (cTNM) and pathological (pTNM) findings and, if indicated, on additional imaging findings, as the TNM system largely parallels across the different diagnostic modalities. The detailed clinical description of each TNM stage can be found in Appendix A.

1.2.3.1 Tumour (or Local) Staging

mpMRI also plays an important role in local tumour staging (T staging), capable of identifying EPE and seminal vesicle invasion.

The most relevant anatomical sequence for local staging on mpMRI is T2-weighted. At 1.5 Tesla (medium MR field strength), mpMRI has good specificity but low sensitivity for detecting tumour extension through the prostatic capsule (T3 stages) [34]. One drawback is that mpMRI cannot detect microscopic EPE. However, its sensitivity increases with the radius of extension within periprostatic fat [35, 36]. The use of high field strength (3 Tesla) or functional imaging in addition to T2-weighted imaging improves sensitivity for EPE or seminal vesicle invasion detection [34], i.e. allowing for full T3 stages (T3a, T3b) evaluation.

Table 1.3: EAU risk groups for biochemical recurrence of localised and locally advanced prostate cancer [43].

Definition			
Low-risk	**Intermediate-risk**	**High-risk**	
PSA <10 ng/mL	PSA 10–20 ng/mL	PSA >20 ng/mL	any PSA
and GS <7 (ISUP grade 1)	or GS 7 (ISUP grade 2/3)	or GS >7 (ISUP grade 4/5)	any GS
and cT1-2a	or cT2b	or cT2c	cT3-4 or cN+
Localised			**Locally advanced**

Given mpMRI low sensitivity for focal (microscopic) EPE, it is not recommended for local staging in low-risk patients [37–39]. However, mpMRI can still be useful for treatment planning.

1.2.3.2 Nodal and Metastatic Staging

The detection of nodal spread (N staging) can be performed with anatomical MRI, abdominal CT or positron emission tomography (PET). Both MRI and CT methods indirectly assess nodal invasion by allowing for measuring nodal diameter and morphology, but there are no threshold metrics to discriminate non-metastatic from metastatic lymph nodes [40]. Although DW MRI is able to detect metastases in normal-sized nodes, a negative DW MRI cannot rule out the presence of LN metastases [41, 42]. Thus, MRI is not recommended for this staging evaluation. Instead, prostate-specific membrane antigen-based PET/CT nuclear imaging modality is increasingly used for N-staging, given its higher sensitivity for lymph node involvement.

For the assessment of metastatic spread (M staging), which in PCa disease usually takes the form of bone metastasis, bone scintigraphy has been the most widely used method, despite the increasing popularity of PET scans.

1.2.3.3 PCa Biochemical Recurrence Risk Groups

Patients are classified with EAU risk group according to risk of BCR after RP, into low-, intermediate- and high-BCR risk groups. The EAU risk group uses cTNM, Gleason grading system and PSA values, with criteria shown in Table 1.3. This system is reference in PCa standard of care that enables both prognosis estimation, and adequate treatment decision-making [43].

1.2.4 Treatment

The main treatment options for localised PCa are active surveillance or watchful waiting, radical prostatectomy (RP) and radiotherapy (RT).

The field of RP has seen technical improvement with the introduction of robot-assisted surgery, allowing more precise dissections potentially leading to better preservation of functional structures [44–47]. Similarly, RT has also seen great technological advances in the past decades, with enhanced dosage delivery. These modalities have shown great oncologic control, none showing superiority over another, in terms of overall- and prostate-cancer-specific survival [48].

1.2.4.1 Radical Prostatectomy

Alongside RT, radical prostatectomy, i.e. surgical removal of the prostate gland, is one of the main active treatment options for primary PCa. Its goal is the eradication of cancer, while whenever possible, preserving continence and potency. RP is considered a standard treatment for clinically localised cancer and, thus, with primary indication to be performed with curative intent. The ideal candidate for RP is the person with disease that is pathologically confined to the prostate and who, if untreated, would suffer morbidity or mortality from the malignancy.

Prostatectomy can be performed by open-, laparoscopic- or robot-assisted (RARP) approaches, where none has clearly shown superiority in terms of functional or oncological results [44]. Yet, RARP it is minimally invasive and has shown to reduce blood loss [44]. RP involves the removal of the entire prostate and seminal vesicles, and it may also be performed with pelvic lymph node dissection (PLND), when the stage of the disease demands so.

RP can be chosen as primary treatment for the three EAU risk groups of localised PCa: low-, intermediate- and high-risk groups (as described in Table 1.3) [17]. In patients with low-risk disease, RP is weakly recommended. However, it can be offered as an alternative to active surveillance, if the patient accepts the trade-off between toxicity and prevention of disease progression. Instead, RP treatment is strongly recommended for patients with intermediate-risk disease with life expectancy greater than ten years. Nerve-sparing RP surgery is suitable for patients with low risk of EPE and, in the case of estimated risk of positive lymph nodes, the surgery should involve also extended PLND. Concerning patients with high-risk localised disease, RP is not firmly recommended, as focal treatment is no longer advised. Yet, RP can be offered as part of a multi-modal therapy, in very specific staging conditions and implying PLND. For such patients, radiotherapeutic treatments are usually prioritised.

1.2.5 Follow-up after Local Treatment

Follow-up procedures are routinely used to detect PCa progression or residual disease. Measurement of PSA constitutes a cornerstone in follow-up after local treatment, as PSA is an extremely sensitive biomarker in this setting.

During the initial-post treatment period, PSA levels are routinely monitored to evaluate treatment success. As the risk of treatment failure is highest during the initial post-treatment period [49, 50], PSA measurement, disease-specific history and DRE (if considered) are recommended at three, six and twelve months post-operatively, reducing to every six months thereafter until three years, and then annually.

Normal PSA values after active treatment with RP or RT differ [51, 52], but recurrence of PSA levels almost always precedes clinical recurrence. No recent consensus exists regarding the best definition of PSA relapse after local treatment, since not all PSA increases have the same clinical value [53]. Reaching the state of BCR does not necessarily indicate that an individual will develop clinically relevant recurrence and/or die of his disease, with studies reporting that 23–43% of patients with BCR after primary surgery develop clinical recurrence [54], with only 16.4% dying from their disease [55].

1.2.5.1 Postprostatectomy PSA Monitoring and Biochemical Recurrence

In a postprostatectomy setting, PSA levels are expected to be undetectable within six weeks, given the complete removal of the prostate tissue and potential affected structures (e.g. seminal vesicles and lymph nodes) [56]. However, when rising PSA levels are detected, this dictates RP treatment failure, and the patient is said to have developed biochemical recurrence.

Post-radical prostatectomy biochemical recurrence is most commonly defined as a value of 0.2 ng/mL (nanograms per millilitre) with subsequent rising [57]. However, no recent consensus exists regarding the best definition of PSA relapse or BCR after local treatment. It is currently argued that a higher threshold of 0.4 ng/mL and rising has better prognostic value for prediction of further metastases [51, 58, 59].

About 20–35% of patients who undergo radical treatment (RP or RT) develop BCR, where most patients do so within 7–10 years [54, 60–62].

BCR occurrence most likely represents the first sign of progression after surgery [63]. This disease stage is thought to be indicative of residual cancer, either locally recurrent, with micrometastases present in the prostatic bed, or due to distant metastases, or both. Regarding local relapse, the absence of a complete prostatic capsule at the apex and the need to preserve the pelvic structures (essentially urethral sphincter and neurovascular bundles) are the principal reasons of this high incidence of local BCR [64]. However, elevated PSA serum levels could be also due to residual glandular healthy tissue in the post-prostatectomy fossa [65].

Patients experiencing BCR are at an increased risk of development of distant metastases, and of PCa-specific and overall mortality [53]. However, the effect size of BCR as a risk factor for mortality is highly variable, as only certain patient subgroups with BCR might be at an increased risk of mortality.

As BCR occurrence is associated with worst prognosis, it requires substantial changes in treatment decisions. Once this disease stage is diagnosed, it is determined whether BCR has developed at local or distant sites. For local recurrences, most patients undergo salvage RT based on BCR diagnosis, without histological proof or imaging of local recurrence. Although imaging has the potential to play a role in detecting both local recurrence and distant metastases, it is mostly reserved for assessment of metastatic disease [17].

Early-BCR Measurement of time from RP surgery to BCR seems to convey relevant prognostic value, where it has been related with disease progression. The shorter the time to BCR, the higher the risk of developing distant metastases, PCa-specific and overall mortality [53, 66]. An early BCR usually implies a biologically more aggressive disease and poor prognosis [66], which can be due to the presence of occult metastases or locally advanced disease [49, 54, 67]. The impact of having early BCR is immense, antedating metastatic disease progression and PCa specific mortality by an average of 7 and 15 years, respectively [54, 67].
The time-point to define ”early BCR” is still under discussion. Yet, more than two-thirds of BCR cases occur within the first two years of treatment, and it has been demonstrated that patients developing BCR within two years after RP have a 20% higher rate of metastatic progression at 5 years compared to those with more delayed BCR [54].

Relevance of BCR Prediction The current management of BCR is still a dilemma for doctors, as there are no prospective random control trials available to allow for firm recommendations. It is still necessary to understand the true impact of BCR on oncological outcomes, given that this disease stage frequently affects PCa treated patients, therefore posing a main clinical concern after definitive therapy. As most severe BCR cases occur shortly after treatment, identifying patients at high-risk of early-BCR may help clinical decision-making process, which can consequently improve patients’ prognosis.

1.3 Literature Review on Biochemical Recurrence Prediction

1.3.1 The Role of Clinicohistopathology information, MRI and Radiomics

Both pre- and postoperative clinical and histopathology data have been explored to give solution to BCR prediction after RP, and more recently computational and quantitative methods of imaging have been used for the development of such prognostic tools.

1.3.1.1 Clinically Accepted BCR Predictive Tools

The probability of BCR varies according to baseline risk characteristics, such as preoperative PSA level, clinical and biopsy staging. Since these parameters vary considerably among patients and exert different effects on the probability of BCR, calculating the individual BCR risk is a better approach than the assignment of average BCR risk [68].

To address this heterogeneity, nomograms and probability graphs were designed to generate individualised probabilities of BCR after RP. Examples of widely accepted models are the 1998 D'Amico et al. risk stratification scheme [69], the 2005 Cancer of the Prostate Risk Assessment (CAPRA) score [43] and the 2006 Stephenson et al. nomogram [70], an enhanced version of Kattan's 1998 nomogram [71]. These models, widely used within the urologic community, rely on commonly available clinical and histopathological variables. Although these models have been validated and clinically accepted, their accuracy values of around 70% leave room for improvement.

Developing stratification tools to explain variability seen in patient outcome after RP can potentially be obtained by including information from pre-treatment mpMR imaging.

1.3.1.2 Imaging in PCa Management: from MRI to Bi- and Multi-parametric MRI

mpMRI scans are well established in PCa clinical management. Since 2010, consensus guidelines on mpMRI reading and interpretation by radiologists of the different mpMRI sequences have been defined [72]. In 2012, the European Society of Urogenital Radiology published MR guidelines and PI-RADS v1 [29], which raised and directed clinical awareness of the potential of this imaging modality. mpMRI has since been rapidly accepted into widespread clinical practice. A second iteration of the guidelines was published in 2015 [73], that widen the acceptance of mpMRI beyond Europe. In response to ambiguities and limitations of the guidelines, these were updated into v2.1 in 2019 [74]. These efforts from the scientific and urological community aimed at promoting a high-quality standardised mpMRI in the work-up of men with suspect PCa.

Radiographic assessment of the prostate through mpMRI allows the retrieval of information for tumour diagnosis, staging and treatment planning, combining anatomical and functional data. mpMRI imaging modality has been used in numerous studies to develop diagnostic tools, for detection and differentiation of prostate cancerous and noncancerous tissue [75–81], and for disease staging, identifying PCa grade [78, 82] and detecting clinically significant lesions [83].

Adding to the aforementioned most relevant problems in the disease domain of PCa is treatment outcome prediction. Before mpMRI was established as the clinical imaging method for PCa management by ESUR [29], PSA failure or BCR prediction after RP started to be addressed with pelvic and endorectal MRI [84, 85]. Later, with the combination of T2w and DWI MRI modalities [86, 87], the first steps of usage of what we currently designate as biparametric

MRI were taken, and led to the conclusion that functional imaging information was valuable for BCR prediction.

Following the first published ESUR guidelines for PCa [29], a plethora of studies were designed to validate and illustrate the value of mpMRI in PCa clinical management in a range of diverse above-mentioned PCa problems [88–91].

Treatment Outcome Prediction with mpMRI Regarding RP treatment outcome prediction for localised PCa disease, different methodologies for exploration of mpMRI data were studied. These ranged from visual evaluation of exams and creating derived semantic attributes (such as state of disease invasion of surrounding structures) [92–95], to proposal of new scoring schemes [92, 96], to prostate capsule 3D shape characteristics assessment [97]. Some studies explored the combination of preoperative patient characteristics and MR findings with post-operative variables coming from histopathology evaluation of RP specimens [95, 98, 99], while others focused on high-risk BCR groups imaging characteristics [100].
Many of these studies compared their results with conventional clinically accepted nomograms, concluding that their performance could be improved by the addition of MRI findings [93, 95, 99, 101].

1.3.2 Quantitative Imaging and Radiomics using bp- and mpMRI

mpMRI has become increasingly important for the clinical assessment of PCa, but its interpretation is generally variable, owing to its relatively subjective nature [102], and to its dependence on the study at hand and reader experience [102–105]. To make mpMR imaging more objective and reliable, researchers focused on identifying quantitative imaging parameters that could be extracted from the different sequences acquired on a mpMRI protocol [88–91, 106].

The field of Radiomics emerged, dealing with the extraction of quantifiable features such as texture, size and shape from clinical images [4, 107–109]. The underlying assumption is that images collected during routine clinical care contain latent information regarding tumour behavior that can be computed using a wide variety of quantitative image characterisation algorithms. The extraction of these radiomic features enables the conversion of collections of digital clinical images into structured quantitative data that can help model tumour behaviour, ultimately eliminating subjectivity in clinical image assessment.

The performance of Radiomics in PCa imaging has been studied to take advantage of the abundant data available with prostate bp- or mpMRI exams, to tackle the most relevant problems in PCa management. Thus, studies for radiomics-driven tumour lesions localisation and detection have been developed [77, 79, 110, 111], as well as lesion classification as benign or malignant [112–114], or as clinically significant or insignificant [115, 116]. Radiomics has also been used for development of non-invasive biomarkers for tumour staging prediction [112, 117, 118].

1.3.2.1 BCR Prediction with Radiomics

So far, Radiomics has not been explored in depth in the context of predicting biochemical recurrence (BCR) of PCa.

On heterogenous cohorts, Shiradkar et al. identified a radiomic signature derived from annotated cancerous lesions on pretreatment bpMRI (T2w and ADC maps) that was predictive of PCa BCR after treatment, whether it was radical prostatectomy, radiotherapy or hormone therapy [119]. Also, Gnep et al. computed a specific category of radiomic features, based on textural analysis of a prostate's peripheral zone, and found them associated with recurrent disease after radiotherapy [120].

Focusing on high-risk PCa patients, Bourbonne et al. found prognostic value on radiomic features derived from prostatic tumours delineated on T2w and ADC for prediction of BCR and BCR-free-survival after surgery [121, 122], while Fernandes et al. found value on anatomical T2w radiomic features for the prediction of radiotherapeutic treatment outcome [123].

Thus, on heterogeneous cohorts, tumour-extracted imaging features from T2w MRI and ADC maps were associated with recurrent disease after different active treatment modalities. This information might be used to determine in which patients conventional treatment is likely to succeed and those where treatment intensification or adding secondary therapy would be beneficial. Early identification of increased recurrence risk can potentially impact clinical management and subsequent follow-up.

1.3.3 Tumour Index in Histopathology and MRI Assessment: Association with BCR prediction

Prostate cancer often presents as a multifocal disease, with two or more tumour nodules, with heterogeneity in Gleason score characterising the disease [124]. Among tumour nodules, the dominant/index tumor nodule (ITN) refers to a single tumour lesion likely to harbor the most aggressive biological behavior [125], possibly dictating the overall disease's prognosis.

In histopathologic studies, the grade of ITN has been strongly associated with BCR occurrence after RP [126], whereas in studies using T2w MR imaging, the assessment of the volume of ITN contributed to the prediction of adverse RP outcomes [127]. However, detection of ITN may not always be possible by radiologists on MR imaging, as it mostly depends on its size [128, 129].

2 APPROACH: RADIOMICS AND LASSO LOGISTIC REGRESSION

2.1 Radiomics

During the last century, physicists have focused on innovating imaging techniques assisting radiologists to improve cancer detection and diagnosis. However, human diagnosis still suffers from low repeatability, erroneous detection or interpretation of abnormalities throughout clinical decision. Errors in diagnosis can be driven by observer limitations (e.g. constrained human visual perception, fatigue, distraction) and by the complexity of the clinical cases.

With high-throughput computing, it is possible to extract innumerable quantitative features from medical images. The conversion of digital medical images into mineable high-dimensional data, a process known as Radiomics, is motivated by the concept that biomedical images contain information that reflects underlying pathophysiology and that these relationships can be revealed via quantitative image analyses [4]. This approach has gained special strength in oncology studies since digital radiologic images are obtained for almost every patient with cancer.

Radiomics quantitative image features are based on intensity, shape and texture, offering information on tumour phenotype and microenvironment that is distinct from that provided by clinical reports, laboratory test results, and genomic or proteomic assays. These features, in conjunction with other information, can be correlated with clinical outcome data and used for evidence-based clinical decision support [108].

The practice of Radiomics involves discrete steps, each with its own challenges [4, 130, 131]. These steps include: (i) acquiring the images, (ii) perform image pre-processing, (iii) identifying the volumes of interest (VOIs) (i.e. those that may contain prognostic value), (iv) segmenting the volumes (i.e. delineating the borders of the volume with computer-assisted contouring), (v) extracting and qualifying descriptive features from the VOIs, (vi) using these features to populate a searchable database, and (vii) mining these data to develop classifier models to predict outcomes either alone or in combination with additional information (e.g. demographic, clinical, genomic, pathology data). Finally, (viii) performing model validation, preferably with an independent dataset.

Radiomics Analysis and Model Development

The heart of Radiomics lies on the extraction of high-dimension feature data to quantitatively describe attributes of VOIs, forming wide datasets where the number of features (p) is much larger than the sample size of patients studied (n), i.e. $p \gg n$. This high-dimensional setting is ruled by a set of phenomena constraints denominated by the curse of dimensionality [132], where the usual statistical techniques for estimation of the model parameters cannot be applied.

Optimal approaches to analyse Radiomics data are yet to be established. One approach can be viewing Radiomics analysis as a Machine Learning (ML) problem, through the creation of a system able to automatically learn and improve from experience without being explicitly programmed. ML algorithms can be coupled with dimensionality reduction techniques to reduce data objects to a lower dimensional feature space, thus tackling the dimensionality problem.

In recent years, machine learning has revolutionised the fields of computer vision and medical image analysis, being increasingly used for clinical applications [133, 134]. ML algorithms are able to detect patterns that are beyond human perception, to learn and even master tasks that were thought to be too complex for machines [135], as well as finding useful biomarkers. Examples of successful clinical application of ML algorithms are expert-level skin lesions classification [136], and lung and breast cancer detection [137, 138].

2.2 Supervised Machine Learning[1]

ML problems can be categorised as supervised or unsupervised. In supervised learning, the goal is to predict the value of an outcome measure based on input measures; whereas in unsupervised learning, there is no outcome measure to be used and the goal is to describe associations and patterns among a set of input measures. The breadth of techniques available is remarkable and spans statistics, data mining and machine learning. Approaches can go from neural networks, linear and logistic regressions, to deep learning.

In the statistical literature, inputs are often called the predictors or independent variables. In the more modern language of ML, the term features is preferred. The outputs are called the responses or dependent variables. Outputs vary in nature, where the output can be qualitative (also referred to categorical or discrete variables or factors), or quantitative (where some measurements are bigger than others, and measurements close in value are close in nature).

The distinction in output type leads to distinct prediction tasks names: classification, when we predict qualitative outputs, and regression, when we predict quantitative outputs. Both tasks share common traits, and both can be viewed function approximation tasks.

Inputs also vary in measurement type. We can have quantitative and qualitative input variables. A third variable type of input is ordered categorical, where there is an ordering between the values, but no metric notion appropriate to measure them.

Qualitative variables are typically represented numerically. In a case of only two classes or categories (such as BCR positive/negative), classes are often represented by a single binary digit, 0 or 1. When there are more than two categories, the most useful and commonly used coding is via ”one-hot encoding”, where a k-level qualitative variable is represented by a vector of K binary variables, where only one is “on” at a time.

Typically, one denotes an input variable by the symbol X. If X is a vector, its components can be accessed by subscripts X_j. Qualitative outputs are denoted by G (for group), while quantitative outputs are denoted by Y. The uppercase letters (X, Y, or G) refer to the generic aspects of a variables, whereas observed values are written in lowercase. Hence, the i^{th} observed value of X is written as x_i.

[1] This section was written using the book *The Elements of Statistical Learning* (Hastie, Tibshirani and Friedman, 2nd edition 2009) as the main reference [139].

Thus, the learning task can be stated as follows: given the value of an input vector X, make a good prediction of the output G or output Y, denoted by $\hat{G}$ or $\hat{Y}$. For a two-class G, one approach is to denote the binary coded output as Y, and then treat it as a quantitative output. Typically, the predictions $\hat{Y}$ will lie in $[0,1]$, and we assign to $\hat{G}$ one of the two-class label according to a cutpoint that separates the classes, for instance, $\hat{y} = 0.5$ (the median value).

The goal is to find a useful approximation $\hat{f}(\mathrm{x})$ to the function $f(x)$ that underlies the predictive relationship between the inputs and outputs.

From the whole dataset, we assemble a training (or development) set of observations $\mathcal{T} = (x_i, y_i)$ or $(x_i, g_i), i \in \{1, \dots, N\}$ with which to construct prediction rules. Being ε an additive error for the model $Y = f(X) + \varepsilon$, supervised learning attempts to learn f by minimising the error ε or an associated metric.

The observed input values x_i are fed to a learning algorithm (usually a computer program), and the algorithm produces outputs $\hat{f}(x_i)$ in response to the inputs. Throughout training phase, the algorithm learns by example, modifying its input/output relationship f in response to differences $y_i - f(x_i)$ between the original and generated outputs.

Upon completion of the learning process, the difference between the artificial and real outputs $y_i - f(x_i)$ should be as small as possible. The aim is to produce a model that is not overfitted to the training data and that it is capable to generalise, i.e., that it is useful for all sets of inputs likely to be encountered in practice.

2.2.1 Linear Logistic Regression

Here we focus on linear logistic regression (LR), a method used for classification problems. In this case, our predict $G(x)$ takes values in a discrete set G, and the input space can be divided into a collection of regions labeled according to the classification. The boundaries of these regions can be rough or smooth, depending on the complexity of the prediction function that defines them. The simplest case is when the separation of the regions of datapoints is linear, as it is done with the linear logistic regression method.

One possible way of finding linear decision boundaries is by modelling the posterior probabilities $P(G = k \mid X = x)$, i.e. the probability of $G = k$ given the evidence $X = x$. The logistic regression model arises infers the posterior probabilities of the K classes via linear functions in x, while at the same time ensuring that they sum to one, remaining in $[0,1]$.

Considering a binary classification problem (K = 2 classes), let $\mathbf{y}_i \in \{0,1\}$ be a vector of size $n \times 1$ of patients' 2-yr BCR status (negative: 0 or positive: 1), N the total number of patients, and let $\mathbf{x}_i$ be a $p \times 1$ vector of features. Let $\pi_i = P(y_i = 1 \mid \mathbf{x}_i)$ and, by total law of probability, $1 - \pi_i = P(y_i = 0 \mid \mathbf{x}_i)$.

LR makes the central assumption that $P(Y \mid X)$ can be approximated as a sigmoid function σ applied to a linear combination of input features. For a single training datapoint $(\mathbf{x}_i, y_i)$, the model assumes:

$$\pi_i(z) = P(Y = 1 \mid \mathbf{X} = \mathbf{x}_i) = \sigma(z) \tag{2.1}$$

$$\text{where} \quad z = \beta_0 + \sum_{i=1}^{p} \beta x_i \quad \text{and} \quad \sigma(z) = \frac{1}{1 + e^{-z}} \quad \text{(sigmoid function).}$$

Using the logit transformation log $[\pi/(1–\pi)]$ — a mathematical transformation that preserves the given order —, we obtain only one linear expression (linear log-odds or logits):

$$\log\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \sum_{j=1}^{p} \mathbf{x}_{ij}^T \beta_j, \quad i \in \{1,2,\ldots,n\}, \tag{2.2}$$

where β_0 is the intercept and β_j is a $p \times 1$ vector of unknown feature coefficients.

Logistic regression offers the advantage of simultaneously estimating the probabilities π_i and $1-\pi_i$ for each class and classifying subjects. The probability of classifying the ith sample in class 1 is estimated by

$$\hat{\pi}_i = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \tag{2.3}$$

The predicted class is obtained by $I\{\hat{\pi}_i > \text{cutpoint}\}$, where $I(.)$ is an indicator function of class membership.

The vector of parameters β is obtained by maximising the log-likelihood of Equation 2.2, over n observations, that is defined as:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \{y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\}. \tag{2.4}$$

The parameters are then estimated through a learning stage using a training set, where the resulting model will allow to understand the role of the input variables in explaining the outcome.

2.2.2 Penalised Logistic Regression

The high-dimensionality $p \gg n$ scenario demands feature selection techniques to eliminate redundant and non-informative features.

Penalisation methods are a type of embedded feature selection techniques that are integrated as part of the learning algorithm. These methods are used to optimise the objective function or performance of a model, slightly altering it through the addition of a penalty term.

Particularly, L_1 Penalised Logistic Regression adds a nonnegative penalty term to the equation used to estimate the models' parameters (Equation 2.4), such that the size of the feature coefficients in high-dimension classification tasks can be controlled, preventing model overfitting.

2.2.2.1 LASSO

Several penalty terms have been discussed in the literature, where LASSO (Least Absolute Shrinkage and Selection Operator) proposed by Tibshirani et al. [140] is one of the popular penalty terms. This algorithm uses L_1-penalty and performs feature selection and estimation simultaneously by constraining the log-likelihood function of the model's coefficients. The penalised method for the logistic regression (PLR) can be obtained by adding the penalty term to the negative log-likelihood function:

$$\text{PLR} = -\sum_{i=1}^{n} \{y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\} + \lambda P(\beta) \tag{2.5}$$

The estimation of the vector β is obtained by minimising Equation 2.5:

$$\hat{\beta}_{PLR} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \{ y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i) \} + \lambda P(\beta) \right], \tag{2.6}$$

where $\lambda P(\beta)$ is the penalty term that penalises the estimates. The penalty term depends on the positive tuning parameter, λ, which controls the tradeoff between fitting the data to the model and the effect of the penalty, i.e., it controls the amount of shrinkage.

Without loss of generality, it is assumed that the features are standardised $\sum_{i=1}^{n} x_{ij} = 0$ and $(n^{-1}) \sum_{i=1}^{n} x_{ij}^2 = 1, \quad \forall j \in \{1, 2, \ldots, p\}$. As a result, the intercept β_0 is not penalised. The estimation of the vector β using LASSO (L_1-penalty) is defined as:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \{ y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i) \} + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{2.7}$$

where λ is a tuning parameter and $\sum_1^p |\beta_j|$ is the L_1 LASSO penalty.

LASSO translates each coefficient by a constant factor λ. For $\lambda = 0$, we obtain the Maximum Likelihood Estimation (MLE) solution, while for large values of λ the influence of the penalty term on the coefficient estimates increases. Making λ sufficiently large will cause some of the coefficients $\hat{\beta}_j$ to be exactly zero, providing a (severe) form of feature selection.

The penalty term makes the PLR solutions nonlinear and thus efficient algorithms are available for computing the entire path of solutions as λ is varied (providing a kind of continuous subset selection of features). The choice of λ is an important part of the model fitting, where this tuning parameter is often chosen by cross-validation (CV) procedure so as to minimise an estimate of the expected prediction error or another costumed performance metric.

2.2.2.2 Priority-LASSO

Motivated by the high-dimensionality of ”omics” models and their need for sparsity and transportability, that preferably should select variables easy to collect or expected to yield good prediction accuracy, Priority-LASSO was developed by Klau et al. in 2018 [141].

Priority-LASSO (P-LASSO) [141] is a LASSO-based method designed for the incorporation of different groups of variables, called “blocks”. The principle of P-LASSO is to define a priority order for the groups of variables. Priority-LASSO approach is based on a hierarchical regression method, that successively fits Lasso regression models using the features in the order of their group’s priority, until all groups have been considered. The resulting linear predictor of each step is then used as an offset for the regression model fit to the features of the group with the next highest priority. Thus, the features of a group with lower priority only explain the part of variation that has not been explained by features of higher priority. In a standard linear regression context this means fitting the residuals of the preceding step.

Formally speaking, let G now be the number of groups under investigation. Let $\pi = (\pi_1, \ldots, \pi_G)$ be a permutation of $(1, \ldots, G)$ indicating the priority order. $\beta_j^{(\pi_g)}$ indicates the coefficient of feature j of group π_g and p_{π_g} the number of features from group π_g. The coefficients of the first step are then estimated by applying standard LASSO on the features of the group with highest priority order. Thus, minimising the mean squared error estimator (the average squared difference between the estimated values and the actual value), and including the penalty term:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p_{m_1}} x_{ij}^{(\pi_1)}\beta_j^{(\pi_1)}\right)^2 + \lambda^{(\pi_1)}\sum_{j=1}^{p_{\pi_1}}\left|\beta_j^{(\pi_1)}\right| \tag{2.8}$$

leads to the linear predictor $\hat{\eta}_{1,i}(\pi) = \hat{\beta}_1^{(\pi_1)} x_{i1}^{(\pi_1)} + \ldots + \hat{\beta}_{p_{n1}}^{(\pi_1)} x_{ip_{\pi_1}}^{(\pi_1)}$. This is used as an offset for the LASSO model fit to the next group in the following step [141]. This procedure is iterated until all groups have been considered, using different offsets $\hat{\eta}_{\pi_g,i}(\pi)$ in each step. Klau et al. [141] emphasise that the information used to produce the model of a subsequent step has already been used to compute the offsets of the previous steps. Therefore, they recommend using cross-validation to estimate the offsets. Otherwise, variability that could be explained by lower-priority groups might be removed, although it is not explained by previous groups.

2.2.3 Model Selection and Assessment

When developing ML models in a data-rich situation, the dataset can be split into three parts: training, validation and test sets. The training set is used to fit the models; the validation set serves to estimate prediction error for model selection; and the test set is used for assessment of the generalisation error of the final chosen model. A well-performing model should reveal good generalisation performance, that relates to its prediction capability on independent data. This independent data is ideally kept aside and used only at the end of the data analysis.

The evaluation of model performance for the different data splits is extremely relevant during model development, as it guides both (i) model selection, i.e the process of estimating the performance of different models in order to choose the best one; and, having chosen a final model, it allows to perform (ii) model assessment, providing a measure of the quality of the ultimately chosen model in terms of its prediction error (generalisation error) on new data.

The dataset splitting is necessary because training error by itself does not give a good estimate of the test error, that is used to evaluate the final model. In fact, training error consistently decreases with model complexity, dropping to zero if the usage of the training data continues to increase. This can happen because the model adapts to complicated underlying structures, leading to an overfit model uncapable to generalise in test set. Yet, typically a model will have a tuning parameter(s) that varies the complexity of a model, where the aim is to find the value and correspondent intermediate model complexity that minimises test error.

There is no general rule on how to choose the number of observations in each of the three parts, as this depends on the signal-to-noise ratio in the data and the training sample size. However, a typical split might be 50% for training, and 25% each for validation and testing.

Yet, it is very common to be have insufficient data to split it into three parts, and thus there are methods designed for such situations, such as cross-validation, to provide reasonable estimates of the expected prediction error.

2.2.3.1 Cross-Validation

Cross-validation (CV) method was designed to approximate the validation step by efficient sample re-use. It is one of the simplest and most widely used methods for estimating prediction error, enabling model selection and a reliable estimate of test error of the final chosen method.

***K*-fold CV** K-fold cross-validation uses part of the available data to fit the model, and a different part to test it. The data is split into K roughly equal-sized parts. For the kth part, the model is fit to the other $K-1$ parts of the data, and we can calculate the prediction error of the fitted model when predicting the kth part of the data. This is repeated to $k = 1, 2, \ldots, K$ and then we combine the K estimates of prediction error. Common K values are 3, 5 and 10.

Stratified *K*-fold CV As some classification problems can exhibit a large imbalance in the distribution of the target classes, performing stratified data sampling is widely suggested. In such cases, stratified k-fold cross validation ensures that the relative class frequencies are approximately preserved in each train and validation fold.

Repeated *K*-fold CV To improve the estimated performance of a ML model, we can use repeated k-fold CV, where the mean result is expected to be a more accurate estimate of the true unknown underlying mean performance of the model on the dataset. We repeat the CV procedure multiple times and report the mean results across all folds from all runs. The data sample is shuffled prior to each repetition, resulting in different splits of the sample.

2.2.4 Discrimination Performance with ROC Analysis

When the results of a diagnostic test or algorithm are binary, discrimination performance is typically measured in terms of sensitivity (the proportion of test-positive subjects out of all disease-positive subjects) and specificity (the proportion of test-negative subjects out of all disease-negative subjects). Before the classification algorithm presents the final results in a dichotomous manner, it first calculates a probability-like continuous output and then collapses the continuous output into categorical results by applying a threshold. Therefore, applying different threshold levels to the continuous output, multiple pairs of sensitivity and specificity values will be obtained. As the threshold to predict the output is lowered, the sensitivity increases while the specificity decreases, and vice versa. To deal with these multiple pairs of sensitivity and specificity values, one can draw a graph by using sensitivity as the y coordinate and specificity (i.e. the false-positive rate) as the x coordinate. Each discrete point on the graph, created by using different threshold levels for a positive test result, is called an operating point, and the Receiver Operating Characteristic (ROC) curve is estimated by connecting these operating points, as depicted in Figure 2.1. ROC curve analysis [142] is an effective method for determining the discrimination performance of a binary classification model.

The most used summary measure of a ROC curve is the area under the ROC curve (AUC). The AUC can be interpreted as the average value of sensitivity for all possible values of specificity or the average value of specificity for all possible values of sensitivity. It can take on any value between 0 and 1, and the closer the AUC is to 1, the better the discrimination performance of the diagnostic test. However, the practical lower limit for the AUC is 0.5, because if we were to rely on pure chance to discriminate binary conditions, the ROC curve would fall along the diagonal line from points (0,0) to (1,1) (see Fig. 2.1). For this study, it is noteworthy that AUC, sensitivity, specificity are independent of disease or condition prevalence [143].

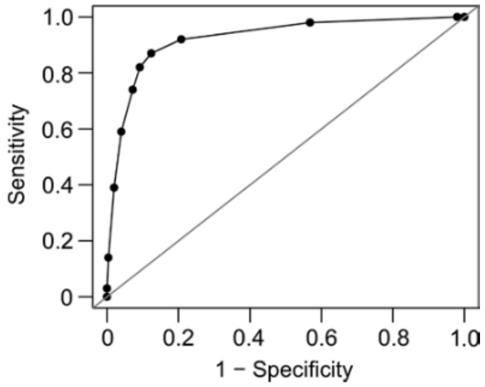


Figure 2.1: Graph of a receiver operating characteristic (ROC) curve of a binary classification model [143]. The datapoints represent various thresholds to distinguish a positive from a negative test result, resulting in different pairs of model's sensitivity and specificity.

Table 2.1: Confusion matrix, classifying all four possible types of outcomes from a binary classifier.

	Condition positive	**Condition negative**
Test outcome positive	True positive (TP)	False positive (FP)
Test outcome negative	False negative (FN)	True negative (TN)

2.2.4.1 Optimal Cutpoint Determination

ROC analysis can also be used to find the optimal cut-off value for turning the continuous output from the ML algorithm into dichotomous results [143]. Determination of the optimal cut-off value for classification should take into account various factors such as the sensitivity and specificity, disease prevalence, costs of different decisions, and clinical setting. For example, a cut-off value with a higher sensitivity is preferable in a screening setting, whereas a cut-off value with a high specificity is more appropriate in a confirmatory setting. Consequently, various methods can be used to make this determination.

2.2.4.2 Performance Evaluation Criteria

Appropriate evaluation criteria are crucial for assessing the binary classification performance of the methods. In the bi-class scenario, one class with very few training samples but high identification importance is referred to as the positive class; the other as the negative class.

A binary classifier classifies all data instances as positive or negative, producing four types of outcomes: two types of correct (or true) classification, true positives (TP) and true negatives (TN), and two types of incorrect (or false) classification, false positives (FP) and false negatives (FN). A 2×2 table formulated with these four outcomes is called a confusion matrix (Table 2.1).

Several measures can be derived using the confusion matrix, and common evaluation criteria include accuracy, recall, precision and specificity. However, as the minority class may bias the decision boundary and has little impact on accuracy [144], we focus on performance evaluation metrics recall, precision, F-score, and specificity, as defined below.

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad \text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$F\text{-score} = 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \qquad \text{Specificity} = \frac{\text{TN}}{\text{FN}+\text{TN}}$$

In information retrieval, **Recall** denotes the percentage of retrieved objects that are relevant; in the context of imbalanced classification with the minority class being relevant, that is the percentage of correctly classified minority instances. **Precision** denotes the percentage of relevant objects that are identified for retrieval, e.g. the percentage of the retrieved instances belonging to the minority class. **F-score** represents a harmonic mean between recall and precision. The harmonic mean of two numbers tends to be closer to the smaller of the two. Hence, a high F-score value ensures that both recall and precision are reasonably high. **Specificity** denotes the percentage of correctly classified majority instances.

2.2.5 Classification on Imbalanced Data

A major challenge to effective healthcare data analytics is highly skewed data class distribution, which is referred to as the imbalanced classification problem. This leads to disadvantages when building a model, as the classification output tends to be biased, being more sensitive to detecting the majority class and less sensitive to the minority class. Data imbalance also affects the choice of proper model evaluation metrics, where model accuracy may not be indicative of true performance. This is crucial when solving a healthcare classification problem, where there is an increased risk associated with false-negative or false-positive predictions. Health-related datasets are often imbalanced because of real world factors, namely, the minority class is simple a rare event when compared to controls, which makes finding data that would balance the class distribution of the dataset difficult. As most standard classifiers, such as logistic regression, implicitly assume that both classes are equally common, being designed for maximising overall classification accuracy, it can be necessary to handle imbalance datasets with techniques such as up-sampling of the minority class.

Minority oversampling with SMOTE Simple replication of the minority class cases can lead to overfitting, as it makes the classifier algorithm learning more and more specific regions of the minority class, thus not spreading the classifier decision boundary as intended. Instead, SMOTE (synthetic minority over-sampling technique) [145], has been widely used to balance class distribution during model development. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors (samples from the minority class with the smallest Euclidean distance from the original sample). Depending upon the amount of over-sampling required, neighbours from the k nearest neighbours are randomly chosen. For instance, if the amount of over-sampling needed is 200%, only two neighbours are chosen and one sample is generated in the direction of each. Synthetic samples are generated by: (1) taking the difference between the feature vector (sample) under consideration and its nearest neighbor; (2) multiplying this difference by a random number between 0 and 1, and (3) adding it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach forces the decision region of the minority class to become more general, potentially solving the class imbalance problem.

3 MATERIALS AND METHODS

This project was conducted in collaboration between the Urology Unit and Radiology Clinical Service from Champalimaud Clinical Centre and the Computational Clinical Imaging Group from Champalimaud Research, Lisbon, Portugal. The project received ethical approval from the Institutional Review Board, where the written documents submitted for ethics evaluation can be found in Appendix B.

Image data handling and processing procedures were automatised through Python scripting, enabling the reproduction of this work. This project's GitHub repository can be found in [146].

3.1 Study Population

In a single centre, 250 patients with localised PCa were retrospectively selected from a consecutive cohort treated with laparoscopic radical prostatectomy, robotic and non-robotic, between January 2014 and December 2017. Patient demographics are presented in Table 4.1.

Inclusion criteria were (i) histopathologically-confirmed diagnosis of PCa, (ii) pelvic bpMRI examination for PCa performed at this study's institution, and (iii) a minimum follow-up of two years with PSA level assessment.

We excluded patients that had (i) received pre- or postoperative adjuvant therapy (hormone or radiation); (ii) incomplete follow-up information; (iii) RP surgery type different that laparoscopy; (iv) external MRI exams; or (v) unavailable or incomplete MRI exams.

The final study population comprised 93 patients, 23 BCR-positive and 70 BCR-negative for two years following RP treatment. 157 out of 250 patients were excluded, with detailed criteria that can be found in a patient population flowchart in Appendix C.

Biochemical recurrence was diagnosed to patients with a single postoperative PSA value equal or greater than 0.2 ng/mL [57, 147].

3.2 Patient Data

Collection of retrospective clinical, histopathological and imaging data was performed by the author of this project, and reviewed by institutional urologists, pathologists and radiologists.

3.2.1 MR Image Data

Patients' radiographic images were retrieved from the institution Picture Archiving and Communication System. Medical imaging data contained both image and image-related data and metadata (e.g. scanner and patient information, acquisition settings) in DICOM format.

3.2.1.1 Image Data Characterisation

Diagnostic and preoperative axial T2-weighted and DW MR imaging was performed with institutional 1.5-Tesla Achieva (n = 32), or 1.5- (n = 52) and 3.0-Tesla Ingenia (n = 9) MRI systems (Philips Healthcare, The Netherlands). The sequences followed a standard-of-care mpMRI protocol compliant with PI-RADS v1 or v2 [29, 73] (depending on the year of acquisition). Other sequences (e.g. anatomical T1w, and coronal and sagittal T2w scans, and functional DCE MRI) acquired as part of the protocol were not included. Characterisation of axial T2w and DWI scans is available in Table D.1.

3.2.1.2 De-identification

Following the principle of providing the minimum amount of confidential information necessary to accommodate downstream analysis of imaging data, we de-identified our raw DICOM images [148]. More descriptive details of this task are given in Appendix D.1.

3.2.1.3 Quality Assessment

Horos [149], a dedicated, open source software for visualisation of medical images, was used to visually perform image quality assessment. This allowed us to remove inadequate image series (faulty and localiser images, duplicated exams, scanner post-processing series).

T2w image series were also visually examined for detection of distortion and motion-artifacts, and presence of unwanted objects (e.g. catheters). Other commonly susceptible MRI artifacts such as bias field were addressed at a later stage.

We also filtered DW MRI exams according to the most commonly acquired low and high b-value parameters within the dataset: b = 0 and b = 1000 s/mm^2, respectively, compliant with PI-RADS v2.1 technical recommendations [74].

3.2.2 Clinical and Histopathologic Data

Clinical and histopathological information was collected from the institution's clinical information system, retrospectively reviewing electronic medical records and histopathology reports, for a total of 33 variables. All histopathological data (biopsy and specimen) followed International Society of Urological Pathology and World Health Organisation criteria, 2014 modified Gleason scoring system and grade groups [27], and were updated to the American Joint Committee on Cancer 2017 eighth edition [150].

Pre-biopsy Data The clinical pre-biopsy variables collected from medical records included baseline PSA level (also stratified into '< 10', '$10 - 20$', and '> 20' ng/mL, according to EAU-ESUR-ESTRO-SIOG Guidelines on Prostate Cancer [11]).

Preoperative Data From the histopathology biopsy report, the overall biopsy Gleason score was retrieved (and coded as ISUP grade), including worst biopsy-core Gleason score found, pathological clinical stage and age.

Treatment Data Radical prostatectomy information was retrieved from medical records and categorised regarding technique (robotic/non-robotic) and type (with/without lymphadenectomy). Date of treatment was also registered, allowing the calculation of time span between MRI exam and treatment.

Postoperative Data From histopathology radical prostatectomy specimen report, we collected Gleason score (specimen and tumour index), grade group at RP and pathological stage category. Other analysed parameters included lymphovascular or perineural invasion, extraprostatic extension, seminal vesicle invasion, and surgical margin status. Patients were classified with an adapted version of the EAU risk group for BCR of localised PCa (grouping of patients with similar risk of BCR after local treatment), baseline PSA level and specimen Gleason score, but without typical consideration of the clinical staging (not available for all patients).

Follow-Up Data Patients were followed after surgery, with serum PSA levels being typically checked at 3-month intervals for the first year and every 6 months for the subsequent 2 years. The endpoint of our study was BCR occurrence within two years after radical prostatectomy, defined as positive if PSA levels ≥ 0.2 ng/mL [57, 147]. Time to BCR was measured from the date of RP treatment to the date of the first documented PSA ≥ 0.2 ng/mL.

3.2.3 Clinicohistopathological Characteristics Descriptive Analysis

Clinicohistopathological variables were split by BCR group (positive or negative within 2 years of surgery) and summarised accordingly (see Table 4.1). Fisher's Exact Test or Wilcoxon Rank Sum Test were applied to categorical or continuous factors, respectively, to compare the distribution of the independent variables across the two levels of the BCR grouping variable.

3.3 Image Processing and Radiomics

Images were converted from DICOM to NIfTI files to ease the following image processing steps. Image preparation details can be found in Appendix D.3.

3.3.1 Image Preprocessing

Despite quality of MRI data acquisition being constantly improving, MRI is still significantly affected by acquisition artifacts. The main artifacts are bias-field related intensity non-uniformity, inter-acquisition non-standardness and Gaussian noise [151], which correction facilitates the application of computarised analysis techniques [152–154]. Bias field correction was performed on T2w images with *N4ITK* algorithm [155]. Image intensity non-standardness was addressed only after field-of-view regularisation (Section D.8.1). Gaussian noise was minimised during MR image reconstruction process. A description of these artifacts categories and respective correction implementations are provided in Appendix D.4.

3.3.2 Image Processing

3.3.2.1 Segmentation

Image segmentation stands for separation of structures of interest from the background and from each other, constituting regions- or volumes-of interest (ROI, VOI, respectively) used for image analysis. T2w axial plane image series, providing the best resolution and contrast to reveal the anatomy of the prostate [29, 72, 110], was used to test different types of segmentation methods (automatic, semi-automatic and manual). Given the poor results of automatised segmentation, we chose to perform manual delineation of the whole-prostate gland.

Training was provided by an experienced urologic radiologist on how to localise the prostate within the pelvic area. *Slicer* (v. 4.10.1) [156], an open-source software platform for medical imaging processing and three-dimensional visualisation, was used to perform segmentation. A more detailed description of the segmentation procedures can be found in Appendix D.5.

3.3.2.2 Registration

A critical assumption of this image processing framework was that the prostate volume segmented on T2w imaging could be automatically transferred to the remaining image sequences we wished to extract information from: b-1000 DWI and ADC maps (that derive from DWI sequences). This transference could be achieved through image registration, an iterative process of aligning an unregistered image into a template image via a geometric transformation [157].

Through alignment of T2w into DWI-b0 intrasubject series, we derived and applied the resultant geometrical transform necessary to register T2w VOIs into low b-value DWI space. Afterwards, we registered high b-value into low b-value DWI images (performing motion correction). This way, the previously computed DWI VOIs were valid for high-b-value DWI images.

We utilised *SimpleElastix* medical image registration library for registration tasks [158]. The resultant final images and VOIs were visually inspected and validated using *Slicer* [156]. Specifications of the registration procedure are available in Appendix D.6.

3.3.2.3 ADC Map Calculation

We calculated the quantitative Apparent Diffusion Coefficient maps from DWI 0 and 1000 s/mm^2 (motion corrected) image series, on a voxel-by-voxel basis, using an in-house Python script. Mathematical details of this implementation are given in Appendix D.7.

3.3.3 Post-Processing

Given the high inter-patient heterogeneity in bpMRI exams, namely in acquisition parameters, image dimensional characteristics, FOV and used scanners (see Table D.1), we implemented additional image processing procedures to diminish the variability of image properties encountered and thus enable fairer VOI comparison. This required the usage of *SimpleITK* interface [159] and *pyRadiomics* package [160] for Python language.

3.3.3.1 Cropping and Normalisation

Image dimensions were regularised prior to image signal intensity normalisation, as normalisation takes into account all grey levels composing the image. For such, we cropped the images with an individually calculated, close-fitting bounding-box that delimited the segmented prostate volume, in T2w, DWI b1000 and ADC maps. This generated cubic-shaped volumes with prostate and immediately surrounding tissues, creating the "bounding-box" image dataset. Similarly, we created a second group of volumes containing the segmented prostate but without surrounding tissues, the "prostate-only" image dataset.

With image intensity normalisation, we transformed axial T2w and DWI b1000 original intensities for both "bounding-box" and "prostate-only" image datasets, and later compared them on developed models performance (Section 4.2). ADC maps signal intensities were not normalised due to their image intensity signal quantitative nature.

3.3.3.2 Binning and Resampling

We discretised images' grey values with costumised binning with fixed bin-width following literature recommendations for textural feature analysis [160, 161]. We also homogenised image resolution for consistent calculation of quantitative features, downsampling T2w images to a grid of 0.66 x 0.66 x 4.5 mm voxels, and DWI-b1000 and ADC maps to a grid of 1.95 x 1.95 x 7 mm voxels. All aforementioned steps are described in detail in Appendix D.8.

Again, these image processing steps led to the creation of two different image datasets of T2w, b1000 DWI and ADC maps: "bounding-box" (with prostate and surrounding tissues) and "prostate-only" image datasets. For simplicity of the description of following steps we will refer only to a single image dataset, even though these steps were run for both datasets.

3.3.4 Radiomic Feature Extraction

2D Radiomic analysis was conducted on the segmented volume labels of the axial plane of T2w, high b-value DWI and ADC maps. We used *PyRadiomics* 3.0 [160], an open-source Python package that extracts region-wise engineered features from 2D images (from a 3D volume). The extraction computes single values per feature for a ROI. A total of 839 for T2w, 839 for high-b-value DWI and 1037 for ADC maps region-level features were extracted for each patient, in a total of 2715 radiomic features. Details of the extracted features are given in Appendix D.9.

3.4 Model Development and Validation

2715 Radiomic features from T2w, b1000 DWI and ADC map of 93 patients, together with 33 clinicohistopathological features, were used to develop a ML framework to discriminate between BCR positive and negative cases, within two years after radical prostatectomy.

The dataset was transformed into three datasets: (i) pre-biopsy, (ii) preoperative and (iii) postoperative, where the inclusion of clinical and histopathological variables varied according to their clinical availability at each corresponding scenario (see Tables D.4 and D.5). Radiomic features were present in all three clinical settings, as PCa imaging examination is performed before biopsy.

The following steps were performed to each dataset and normalisation method ("prostate-only" and "bounding-box"). On the final stage of this workflow, we used SMOTE minority-sampling technique to deal with class unbalance, generating a randomly-oversampled and balanced dataset.

3.4.1 Data Preparation

Observations were randomly shuffled and then randomly split into train and test sets, in the ratio of 70:30. Train set had 66 patients (15 BCR positive, 51 BCR negative) and test set comprised 27 patients (5 BCR positive, 22 BCR negative). The training dataset was used for parameter tuning, feature selection and error estimation using cross-validation, whereas the test dataset was only used at the end, to access models' generalisation capabilities. More details on data preparation are given in Appendix D.10.1.

3.4.2 Classification

We chose the logistic regression (LR) algorithm to solve this problem through supervised machine learning, for its simplicity and widespread use in classification studies.

3.4.3 Minority Cases Oversampling

For performance comparison between non-oversampled and oversampled datasets, SMOTE [145] was applied, generating a class-balanced training dataset of 51 BCR positive and 51 BCR negative cases. We used $k = 5$ as the number of nearest neighbours to create the new examples of the minority class, with R software *DMwR* package [162].

3.4.4 Cross-Validation

A 4-fold stratified cross-validation procedure was applied to the development set to train and to estimate the performance of LR classifiers using embedded penalisation methods (further described in Section 3.4.6). The procedure involved stratification to ensure the generated folds preserved the percentage of samples for each class, with respect to the original sample.

In each CV round, candidate classifiers were trained using 75% of the development set (3 folds) and evaluated on the remaining 25% (1 fold). Repeating the process over all the folds and collecting the resulting fitted probabilities generated a vector of the same length as the size of the development set, with values between 0 and 1. This vector, after being dichotomised into two classes with criteria that will be further detailed on Section 3.4.8, was then compared with true labels of the patients, using appropriate evaluation measures to assess the classification performance of the trained algorithm being considered. The complete procedure of generating four subsets to perform 4-fold cross validation was repeated 14 times.

3.4.5 Statistical Analyses of Classifier Performance

To determine the best performing classification algorithm among those developed, whether in terms of their hyperparameters, structures or between top candidates, a statistical analysis of the performance of all the tested classification models was conducted using the Friedman-Nemenyi or Friedman-Wilcoxon tests [163]. These tests, which also account for multiple or pairwise hypothesis testing, were used to assess the statistical significance of the relative difference of performance of the penalisation algorithms and their resultant LR models, in terms of their relative ranks across the 14 CV runs. We summarised the models' performance with a single summary measure, F-score, as it combines both the concerns of precision and recall in one number. Specifically, we used *Fmax*, calculated from ROC curve, as it will be defined in Section 3.4.8.

Statistical analyses were conducted with the R statistical program (version 3.6.1) [164], using *stats* (statistical calculations) and *PMCMR* (multiple comparison tests [165]) packages. A significance level of 0.05 was established as significant, and Shapiro-Wilk's test [166] was used to assess variable's normality.

3.4.6 Feature Selection with Regularisation

For this high-dimensional logistic regression problem, we used LASSO [140] and Priority-LASSO [141] regularisation techniques, integrated as part of the LR learning algorithm, with R packages *glmnet* and *prioritylasso* [167, 168].

3.4.6.1 Regularisation Configuration

The magnitude of L_1-penalisation achieved with LASSO and P-LASSO was controlled by a constant penalty parameter *lambda*. This parameter was tuned via a 4-fold stratified CV, such that the mean cross-validated AUC across folds was maximal. For such, the models were fitted at 100 sequential values of *lambda* automatically chosen.

For both LASSO and Priority-LASSO, we imposed different limits to the maximum number of features in each resulting model. Models contained, at most, 9 variables (following the "one variable to ten observations" rule of thumb for our dataset of N = 93).

LASSO LASSO logistic regression model fitting was implemented with restriction to the number of selected variables, with a parameter $pmax = \{5, 6, 7, 8, 9\}$.

Priority-LASSO Priority-LASSO method required the definition of block of variables with a priority order to successively fit LASSO models. As there was no prior knowledge about the influence of variables on BCR occurrence, we altered the P-LASSO method to allow for the algorithm to automatically run over all possible block priority sequences of the defined blocks.
We implemented P-LASSO in two ways: P-LASSO-2-blocks and P-LASSO-4-blocks. For P-LASSO-2-blocks, features were segregated into two blocks: *Clinical* (i.e. all variables not derived from imaging) and *Radiomics*. For this variant, two block priority sequences were possible, as shown in Table 3.1. For P-LASSO-4-blocks implementation, features were divided into four blocks: *Clinical*, *T2w*, *DWI-b1000* and *ADC*. This analysis strategy led to 24 different block priority permutations, as defined in Table 3.2.
We constricted P-LASSO model fitting regarding the maximum amount of non-zero coefficients produced. As the algorithm implementation in R only allows to specify the maximal number of coefficients for each block, we defined several *pmax* sequential vectors with integers which specified the number of maximal coefficients for each block.
For Priority-Lasso-2-blocks, we ran the CV scheme with $pmax = \{(1,5), (1,7), (1,8), (2,7), (3,6), (5,4)\}$. For Priority-Lasso-4-blocks, we restricted the fitting of each of the four constitutive blocks with $pmax = \{(1,6,0,1), (1,7,0,1), (2,1,2,1), (2,2,2,2), (3,2,1,0), (3,2,2,0), (3,3,0,0), (3,3,2,0), (3,3,3,0)\}$. This way, P-LASSO models also included, at most, 9 variables.

Table 3.1: Priority-LASSO-2-blocks priority sequences implemented, according to the variables' categories.

Priority-LASSO-2-blocks		
Priority Sequence	Block 1 (higher priority)	Block 2 (lower priority)
Radiomics > Clinical	Radiomics	Clinical
Clinical > Radiomics	Clinical	Radiomics

Table 3.2: Priority-LASSO-4-blocks priority sequences implemented, according to the variables' categories.

Priority-LASSO-4-blocks				
Priority Sequence	Block 1 (highest priority)	Block 2	Block 3	Block 4 (lowest priority)
1	T2	b1000	ADC	Clinical
2	T2	b1000	Clinical	ADC
3	T2	ADC	b1000	Clinical
4	T2	ADC	Clinical	b1000
5	T2	Clinical	b1000	ADC
6	T2	Clinical	ADC	b1000
7	b1000	T2	ADC	Clinical
8	b1000	T2	Clinical	ADC
9	b1000	ADC	T2	Clinical
10	b1000	ADC	Clinical	T2
11	b1000	Clinical	T2	ADC
12	b1000	Clinical	ADC	T2
13	ADC	T2	b1000	Clinical
14	ADC	T2	Clinical	b1000
15	ADC	b1000	T2	Clinical
16	ADC	b1000	Clinical	T2
17	ADC	Clinical	T2	b1000
18	ADC	Clinical	b1000	T2
19	Clinical	T2	b1000	ADC
20	Clinical	T2	ADC	b1000
21	Clinical	b1000	T2	ADC
22	Clinical	b1000	ADC	T2
23	Clinical	ADC	T2	b1000
24	Clinical	ADC	b1000	T2

3.4.7 Reproducibility and Stochastic Procedures

To ensure all models were comparable and that reproducibility of results was guaranteed, we used a fixed set of 14 randomly generated numbers (seeds) that underlid all random components throughout the different trials.

3.4.8 Assessment of Classifier Performance

Classifier performance can be evaluated in terms of a variety of measures. Although the Area under the ROC Curve, AUC, is the most commonly used performance measures in radiomics studies, it is not reliable in cases of unbalanced classes [169, 170], which is the case in this study. Thus, in addition to AUC, class-specific Precision, Recall and F-measure evaluation measures, which are more suited for unbalanced class situations [170, 171], were also used. Like AUC, these measures range from 0 to 1, with higher values indicating better classification performance.

The ROC curve can be derived by varying the threshold that is applied to discern the classification scores into the two classes. For this study, the maximum value of F-measure for the BCR-positive (minority) class achieved across all these thresholds, also termed *Fmax*, as well as the associated values of Precision and Recall, termed *Pmax* and *Rmax* respectively, were used to evaluate the candidate classifiers tested in the framework. The corresponding classification score threshold that yielded this value of *Fmax* was also recorded for each of the classifiers.

The threshold for the final classifier was obtained by averaging the threshold that yielded the highest *Fmax* value for the corresponding classification algorithm in each of the fourteen cross-validation rounds. The final classifier was applied in combination with this threshold to the independent validation set to obtain binary predicted labels for the constituent patients, which were then evaluated in terms of AUC, F-measure, Precision and Recall. Similarly to the development set, the validation of the model using the left-out test set was repeated 14 times to investigate the stability of the model's solutions.

ROC curve calculation and determination of optimal cutpoints analyses were performed with *pROC* and *cutpointr* R packages [172, 173].

3.4.9 Validation on a Left-out Patient Cohort

For each simulated clinical scenario, the best performing classifier identified by our framework was applied to the independent validation set of 27 PCa patients to assess the classifier's generalisability to new patient populations. In parallel, we also applied this test set to the corresponding classifiers using oversampling of the minority class during model development, to explore the effect of class balancing in classification. Performance assessments were carried using the classifier evaluation measures discussed previously.

3.4.10 Other Analyses

Image Normalisation Method Influence on Classifier Performance

The two normalisation methods, bounding-box and prostate-only, were compared based on the *Fmax* performance metric of the models built with Pre-Biopsy datasets, as these datasets contain the highest portion of radiomics variables.

A broad comparison of models' performances was made for each type of model (LASSO, Priority-LASSO-2- and -4-blocks) *Fmax*'s distributions regarding the normalisation method implemented, using a pair-wise Wilcoxon signed-rank test. Here, all models' tested configurations (different hyperparameters and block priority sequences) were considered together.

Then, we fine-tuned this analysis for Priority-LASSO's specific implementations. As Priority-Lasso (-2 and -4-blocks) were implemented with different block priority sequences, representing different models, we summarised the performance statistics for each arrangement of variable blocks and normalisation method. After selecting the top sequences that yielded the highest median *Fmax* values, we carried out pairwise Wilcoxon signed-rank tests to evaluate the effect of the normalisation method on the given top sequences' performances.

4 RESULTS

Here we describe the study cohort in terms of clinicohistopathologic features. We show the results of comparison of the image normalisation methods regarding models' discriminatory performance. The remaining analysis is carried out with one selected normalisation method. We present the results of the application of the ML framework for the different simulated clinical scenarios (pre-biopsy, preoperative and postoperative), which included model selection, evaluation and validation.

4.1 Patient Clinicohistopathological Characteristics

Characteristics of 93 patients in the study are summarised in Table 4.1, as well as respective parameter association with 2-year biochemical recurrence status. The mean age at diagnosis was 60.7 years, and mean preoperative PSA was 7.94 ng/mL. The patients were treated with radical prostatectomy alone (60.2%) or radical prostatectomy with extended lymph node dissection (39.8%). The minimum follow-up for the entire cohort of patients was 2 years. During follow-up, 20 patients (21.5%) experienced BCR. The mean age at time of biochemical recurrence was 61.4 years (range: 54.3–75.9) with a time to recurrence of 9.75 ± 7.87 months (range: 2–24).

Localised disease was seen in 63% of patients, and locally advanced disease was observed in 37%. Lymph node metastasis was seen in 4 of the 37 patients who underwent lymphadenectomy. Pure acinar adenocarcinoma was the most frequent histologic subtype (71%), but mixed acinar/non-acinar histology was seen in 29% of patients. Twenty-three percent of patients had positive resection margins. Perineural invasion and lymphovascular invasion were seen in 83% and 4% of patients, respectively.

Preoperative PSA and its level, treatment type, grade group classification of overall specimen and of index tumor nodule, histologic subtype, surgical resection margin status and linear extension, extraprostatic linear extension, perineural and lymphovascular invasion, the adapted EAU-ESUR 2018 group risk and risk stratification by group grade were associated with biochemical recurrence in our cohort series (Table 4.1).

The most common grade group was 2 (GG2), both for the index tumor nodule (57%) and overall prostate cancer (66.7%). In general, the higher-grade group was seen in index tumor nodule and not in overall prostate cancer: GG3, 30.1% versus 16.1%; GG4, 2.2% versus 0%; and GG5, 3.2% versus 2.2%.

Table 4.1: Clinical and pathologic features of 93 patients who underwent laparoscopic radical prostatectomy, and parameters associated with biochemical recurrence. Continuous variables are present as mean ± SD (range) and categorical variables as frequencies (percentage). Legend: SD = standard deviation; PSA = prostate-specific antigen (in ng/mL), GS = Gleason score, cc = cubic centimeters; AJCC = American Joint Committee on Cancer.

	Biochemical Recurrence			
	No (N = 73)	Yes (N = 20)	Overall (N = 93)	*p* value
	78.5%	21.5%		
Age at diagnosis	60.9 ± 5.5	59.9 ± 5.4	60.7 ± 5.4	0.263[2]
	(46.0 – 72.0)	(53.0 – 75.0)	(46.0 – 75.0)	
Preoperative PSA, ng/mL	7.09 ± 4.66	11.06 ± 8.76	7.94 ± 5.96	**0.017**[2]
	(0.38 – 23.50)	(4.30 – 39.00)	(0.38 – 39.00)	
Preoperative PSA level, ng/mL				**0.034**[1]
< 10	61 (83.6%)	14 (70.0%)	75 (80.6%)	
10 - 20	10 (13.7%)	2 (10.0%)	12 (12.9%)	
> 20	2 (2.7%)	4 (20.0%)	6 (6.5%)	
Biopsy grade group				0.123[1]
(GS) (highest core)				
Grade group 1 (GS ≤ 6)	8 (11.0%)	1 (5.0%)	9 (9.7%)	
Grade group 2 (GS 3 + 4 = 7)	43 (58.9%)	8 (40.0%)	51 (54.8%)	
Grade group 3 (GS 4 + 3 = 7)	18 (24.7%)	9 (45.0%)	27 (29.0%)	
Grade group 4 (GS 8)	4 (5.5%)	1 (5.0%)	5 (5.4%)	
Grade group 5 (GS 9–10)	0 (0.0%)	1 (5.0%)	1 (1.1%)	
Surgery type				0.147[1]
Non-robotic	57 (78.1%)	12 (60.0%)	69 (74.2%)	
Robotic	16 (21.9%)	8 (40.0%)	24 (25.8%)	
Treatment type				**0.043**[1]
Radical Prostatectomy	48 (65.8%)	8 (40.0%)	56 (60.2%)	
Radical Prostatectomy w/ lymphadenectomy	25 (34.2%)	12 (60.0%)	37 (39.8%)	
Prognostic Grade group				**0.006**[1]
(GS) overall prostate				
Grade group 1 (GS ≤ 6)	14 (19.2%)	0 (0.0%)	14 (15.1%)	
Grade group 2 (GS 3 + 4 = 7)	49 (67.1%)	13 (65.0%)	62 (66.7%)	
Grade group 3 (GS 4 + 3 = 7)	10 (13.7%)	5 (25.0%)	15 (16.1%)	
Grade group 4 (GS 4 + 4 = 8)	0 (0%)	0 (0%)	0 (0%)	
Grade group 5 (GS 9–10)	0 (0.0%)	2 (10.0%)	2 (2.2%)	
Prognostic grade group				**0.005**[1]
(GS) index tumor				
Grade group 1 (GS ≤ 6)	7 (9.6%)	0 (0.0%)	7 (7.5%)	
Grade group 2 (GS 3 + 4 = 7)	45 (61.6%)	8 (40.0%)	53 (57.0%)	
Grade group 3 (GS 4 + 3 = 7)	19 (26.0%)	9 (45.0%)	28 (30.1%)	
Grade group 4 (GS 8)	2 (2.7%)	0 (0.0%)	2 (2.2%)	
Grade group 5 (GS 9–10)	0 (0.0%)	3 (15.0%)	3 (3.2%)	
Index tumor nodule	1.9 ± 1.3	2.0 ± 0.8	1.9 ± 1.2	0.148[2]
linear extension, cm	(0.4 – 11.0)	(0.4 – 3.4)	(0.4 – 11.0)	
Index tumour volume, cc	16 ± 10	18 ± 10	16 ± 10	0.555[2]
	(1 – 39)	(5 – 40)	(1 – 40)	
pT status (AJCC 2017)				0.177[1]
pT2	49 (67.1%)	10 (50.0%)	59 (63.4%)	
pT3a	20 (27.4%)	7 (35.0%)	27 (29.0%)	
pT3b	4 (5.5%)	3 (15.0%)	7 (7.5%)	

pN status				0.066[1]
N0	23 (31.5%)	10 (50.0%)	33 (35.5%)	
N1	2 (2.7%)	2 (10.0%)	4 (4.3%)	
Nx	48 (65.8%)	8 (40.0%)	56 (60.2%)	
Histological subtype				**0.027**[1]
Mixed	17 (23.3%)	10 (50.0%)	27 (29.0%)	
Acinar	56 (76.7%)	10 (50.0%)	66 (71.0%)	
Surgical resection margin status				**0.013**[1]
R0	61 (83.6%)	11 (55.0%)	72 (77.4%)	
R1	12 (16.4%)	9 (45.0%)	21 (22.6%)	
Surgical resection margin extension, mm	0.3 ± 1.0 (0.0 – 7.0)	2.3 ± 3.4 (0.0 – 13.0)	0.7 ± 1.9 (0.0 – 13.0)	**0.001**[2]
Extraprostatic extension				0.121[1]
No	48 (65.8%)	9 (45.0%)	57 (61.3%)	
Yes	25 (34.2%)	11 (55.0%)	36 (38.7%)	
Extraprostatic linear extension, mm	1 ± 2 (0 – 16)	2 ± 2 (0 – 8)	1 ± 2 (0 – 16)	**0.022**[2]
Perineural invasion				**0.019**[1]
No	16 (21.9%)	0 (0.0%)	16 (17.2%)	
Yes	57 (78.1%)	20 (100.0%)	77 (82.8%)	
Lymphovascular invasion				**0.030**[1]
No	72 (98.6%)	17 (85.0%)	89 (95.7%)	
Yes	1 (1.4%)	3 (15.0%)	4 (4.3%)	
Seminal vesicle invasion				0.606[1]
No	69 (94.5%)	18 (90.0%)	87 (93.5%)	
Yes	4 (5.5%)	2 (10.0%)	6 (6.5%)	
EAU-ESUR 2018 Group Risk (adapted)				$<$ **0.001**[1]
Low (PSA $<$ 10 , GS $<$ 7)	13 (17.8%)	0 (0.0%)	13 (14.0%)	
Intermediate (10 $<$ PSA $<$ 20 or GS $=$ 7)	58 (79.5%)	14 (70.0%)	72 (77.4%)	
High (PSA $>$ 20 or GS $>$ 7)	2 (2.7%)	6 (30.0%)	8 (8.6%)	
Overall prostate group grade risk stratification				**0.013**[1]
Low-risk (Group grade 1)	14 (19.2%)	0 (0.0%)	14 (15.1%)	
Medium-risk (Group grade 2)	49 (67.1%)	13 (65.0%)	62 (66.7%)	
High-risk (Group grade $\geq$ 3)	10 (13.7%)	7 (35.0%)	17 (18.3%)	

[1]: Fisher’s Exact Test; [2]: Wilcoxon Rank Sum test.

4.2 Comparison of Image Normalisation Methods

We aimed to verify if model performance would differ according to the normalisation method used for image pre-processing, since it may affect radiomic features' calculation.

4.2.1 Broad Algorithm Performance Comparison

Overall *Fmax* performance distribution of the three implemented algorithms is depicted in Figure 4.1, encompassing all tested hyperparameters and block priority sequences.

A Wilcoxon signed-rank test revealed that normalisation method induced statistically significant differences between LASSO's algorithms' *Fmax* performance (V = 1791, $p = 0.00016$, Wilcoxon signed-rank test), where prostate-only normalisation models led to solutions with improved *Fmax*, whereas bounding-box normalisation models consistently provided solutions that led to lower but more stable *Fmax* values.

As for both Priority-LASSO algorithms (2- and 4-blocks), *Fmax* distribution of all models' configurations tested did not statistically differ across normalisation methods.

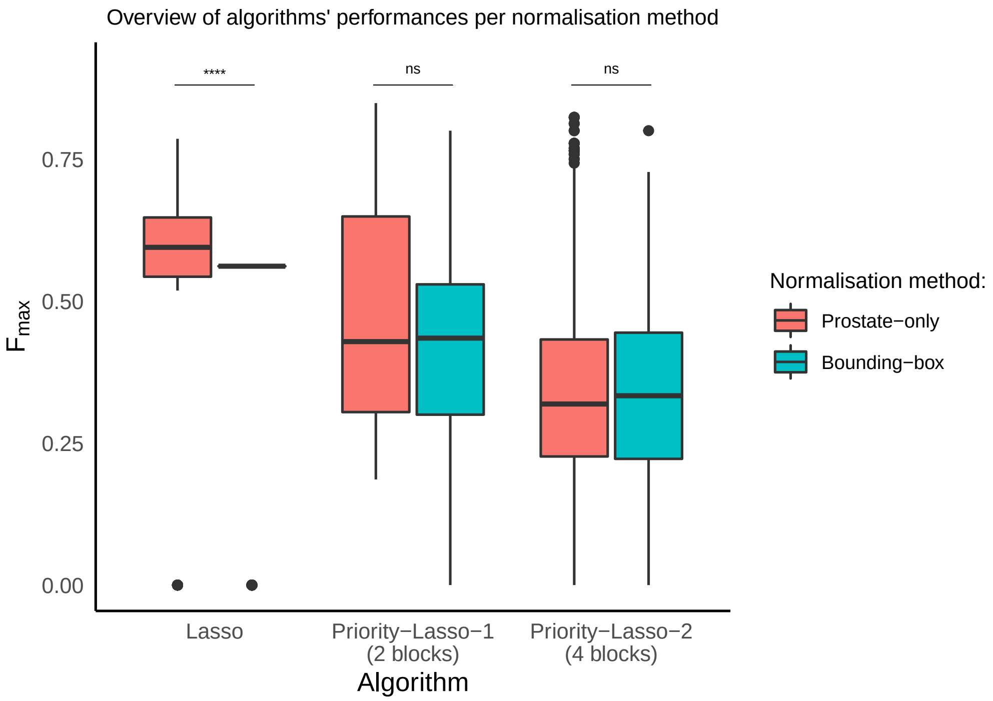


Figure 4.1: Boxplots and statistical analysis of overall *Fmax* performance for the both normalisation methods applied, across the different penalised LR models. Results of a Wilcoxon signed-ranked test for both prostate-only and bounding-box normalisation *Fmax* performances, encompassing all tested models' hyperparameters. Significant differences in *Fmax*, according to each normalisation method, was found for LASSO algorithm. Priority-LASSO algorithms' overall *Fmax* performance did not show differences between normalisation methods. Legend: 'ns' $p > 0.05$; $****$ $p < 0.0001$.

4.2.2 Model Performance Comparison: Priority-LASSO Implementations

Priority-LASSO algorithms were implemented with different priority sequences, as we aimed to understand which types of data could have higher relevance to solve this classification problem. In Figures 4.2 and 4.3, we compare *Fmax* performances of each priority sequence implemented in P-LASSO-2-blocks and -4-blocks models, respectively, for each normalisation method.

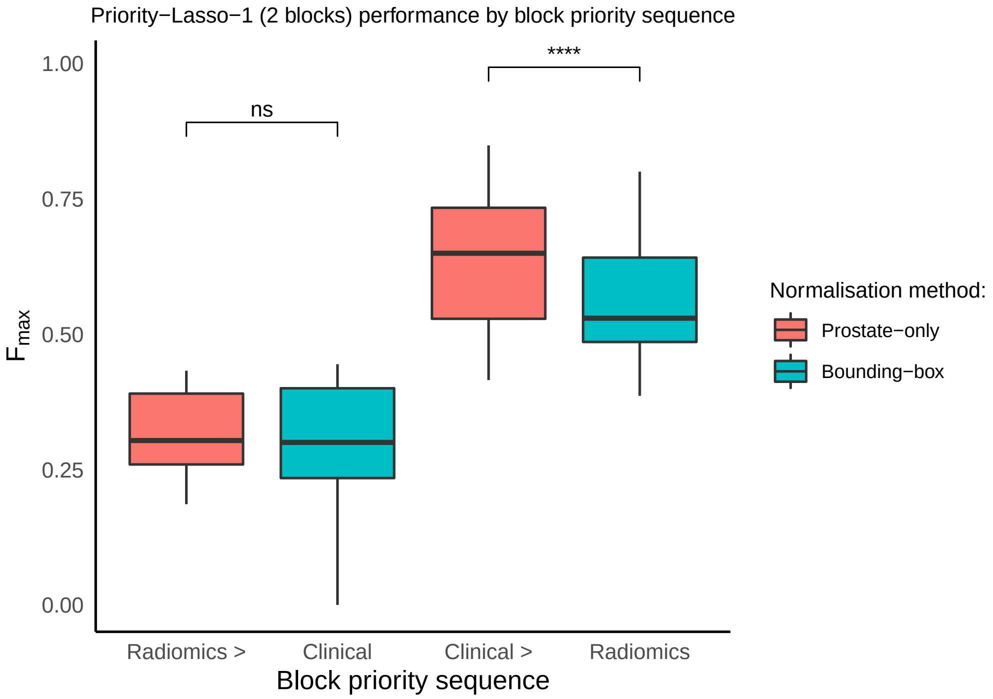


Figure 4.2: Boxplots and statistical analysis of Priority-LASSO-1 (2 blocks) *Fmax* performances for each normalisation method and block priority sequence implemented. *Fmax* distributions' means were compared using Wilcoxon signed-rank test. Significant differences in performance were found for *Clinical* > *Radiomics* priority sequence. Legend: 'ns' $p > 0.05$, '$****$' $p < 0.0001$.

4.2.2.1 Priority-LASSO-2-blocks

Regarding *Fmax* performances of each of the two block priority sequences implemented with P-LASSO-2-blocks, shown in Figure 4.2, we found a statistical difference between normalisation methods when comparing *Fmax* distributions of P-LASSO-2-blocks ran with *Clinical* > *Radiomics* block priority sequence, where prostate-only normalisation led to greater *Fmax* values than bounding-box normalisation (V = 2672, $p < 0.0001$, Wilcoxon signed-rank test).

No statistical difference between *Fmax* distributions for models developed with Radiomics features having a higher priority than Clinical features was found when comparing the two normalisation methods (V = 1578, $p > 0.05$, Wilcoxon signed-rank test).

These results with pre-biopsy data also allowed to conclude that block priority sequence *Radiomics* > *Clinical* consistently performed worse than the alternative sequence. As such, we will not consider the priority sequence *Radiomics* > *Clinical* for any further Priority-LASSO-1 (2-blocks) analysis of any clinical scenario, only *Clinical* > *Radiomics*.

4.2.2.2 Priority-LASSO-4-blocks

All 24 possible block priority sequences derived from 4 blocks (*Clinical*, *T2*, *b1000* and *ADC*) were tested for comparison of *Fmax* performances for prostate-only and bounding-box normalisation methods. In Figure 4.3 are the results of the statistical analysis comparing the effect of normalisation methods on Priority-LASSO-4-blocks priority sequences' performances. Using pair-wise Wilcoxon signed-rank tests, 20 out of 24 block priority sequences revealed significant differences in *Fmax*. However, there was not a predominant superiority of one normalisation method over the other, as the frequency that prostate-only normalisation led to superior or inferior values of Fmax compared to bounding-box normalisation was the same (10 out of 20).

Thus, prostate-only normalisation did not consistently perform better than bounding-box normalisation, or vice-versa, where *Fmax* performance superiority, or inferiority, was dependent on the priority sequence used to develop the model.

However, the results depicted in Figure 4.3 reveal superior *Fmax* performances achieved through the implementation of 19 and 20 block priority sequences (bps), i.e. *Clinical, T2, b1000, ADC* and *Clinical, T2, ADC, b1000*, respectively. Here, prostate-normalisation led to improved *Fmax* for both bps19 (V = 8320, $p < 0.0001$, Wilcoxon signed rank test) and bps20 (V = 8704, $p < 0.0001$, Wilcoxon signed rank test), when comparing with the alternative method.

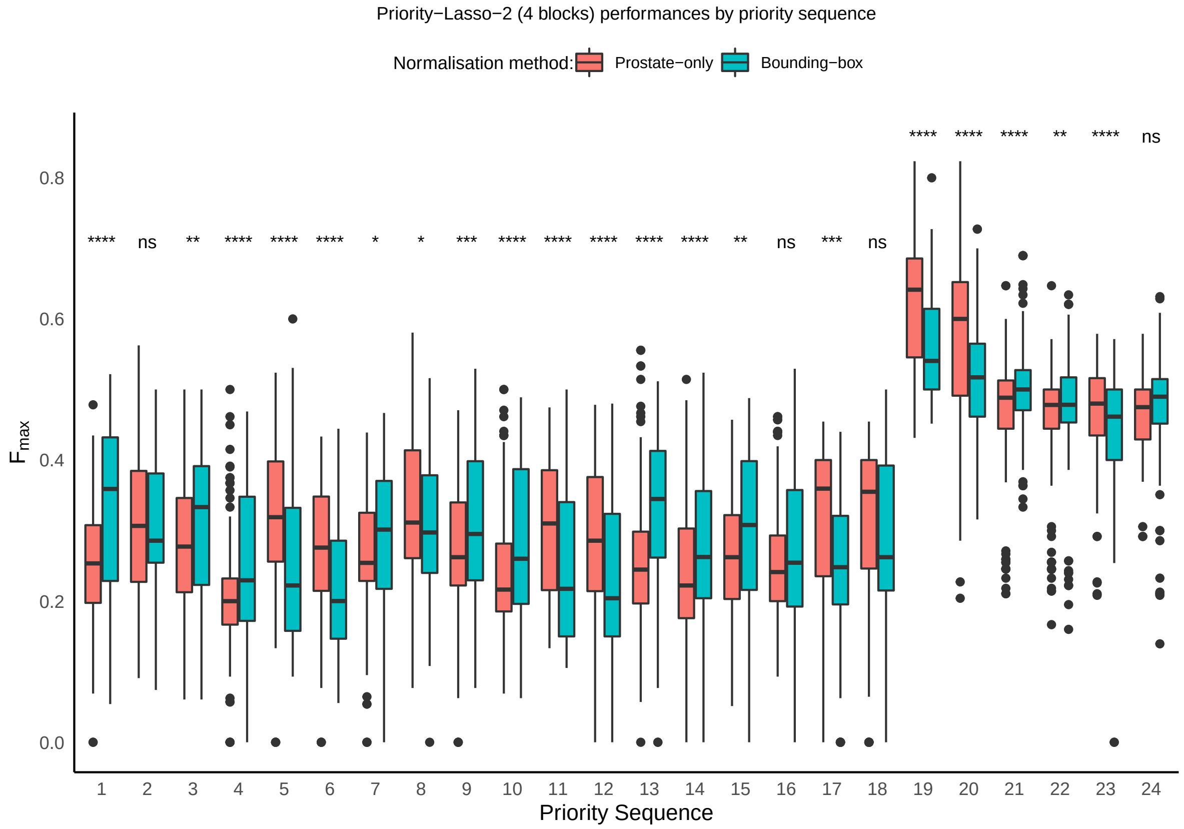


Figure 4.3: Boxplots and statistical analysis of the effects of normalisation method on the *Fmax* performance of the different priority sequences implemented for Priority-LASSO-4-blocks models. Significance level results of a two-sample non-parametric Wilcoxon signed-rank test comparing each of the 24 priority sequences' *Fmax* performance according to each normalisation method are depicted. Only 4 priority sequences' performances revealed to not differ for normalisation method. These results also highlight clearly superior performances of priority sequences 19 and 20. Legend: 'ns' $p > 0.05$; '$*$' $p \leq 0.05$; '$**$' $p \leq 0.01$; '$***$' $p \leq 0.001$; '$****$' $p \leq 0.0001$.

Conclusion Overall, and across all three different regularisation methods applied within the pre-biopsy clinical scenario, models fitted with prostate-only normalised image data performed better than the corresponding models fitted with image data using bounding-box normalisation. For Priority-LASSO-4-blocks, this performance difference was highlighted by the top performing algorithm configurations, 19 and 20. Given these findings of persistent higher performance using prostate-only normalisation, only results derived from it will be presented in the following sections.

4.3 Model Selection

Here we analyse all candidate models for BCR classification modelling problem, aiming to choose one among them. The model selection process results are presented for each simulated clinical scenario: pre-biopsy, preoperative and postoperative (colour coded in blue, green and orange).

4.3.1 Pre-Biopsy

4.3.1.1 LASSO

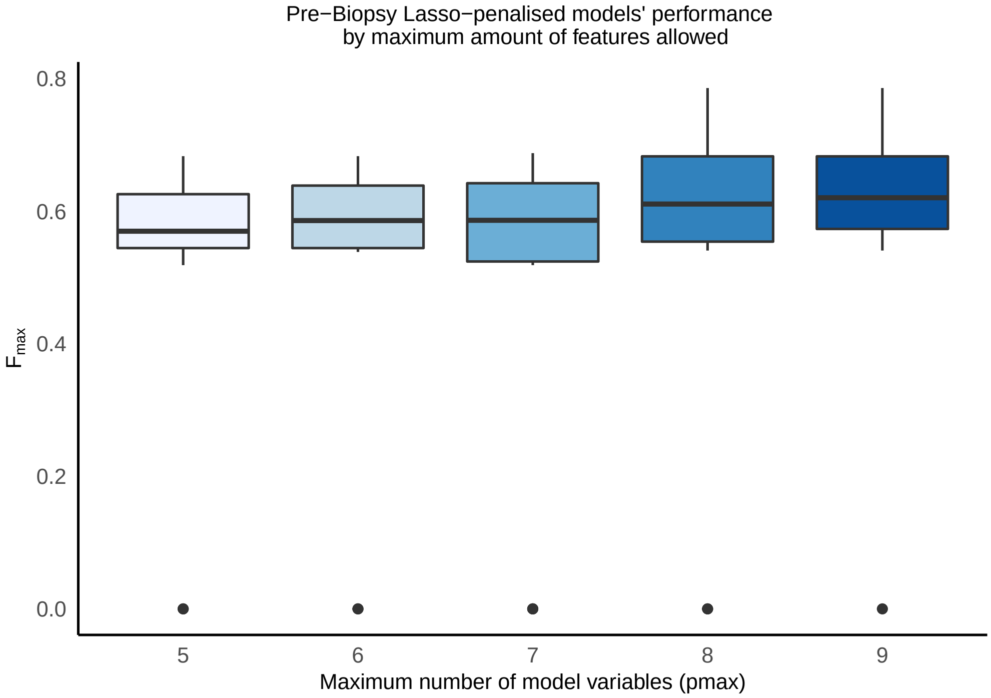


Figure 4.4: Boxplots and statistical analysis of the effects of the maximum number of selected variables (*pmax*) on LASSO models's performance, measured by *Fmax*. Statistical results from multiple pairwise comparisons with Nemenyi test showed no significant differences in measured *Fmax* for any of the maximum limit of allowed variables, highlighting a considerable stability of pre-biopsy LASSO models' solutions.

Figure 4.4 shows Pre-biopsy LASSO models' performance when varying the maximal number of variables to be included in it (*pmax*). Model solutions were stable through the different model configurations. Median (IQR) *Fmax* values for *pmax* = 5, 6 and 7 were 0.570 (0.544 to 0.623), 0.586 (0.544 to 0.639) and 0.586 (0.524 to 0.642). For *pmax* = 8, 9, the median values were 0.6107 (0.554 to 0.683) and 0.620 (0.573 to 0.683), respectively. Even though Friedman test indicated statistical significance ($\chi^2(4) = 17.193$, $p = 0.002$), Nemenyi post-hoc test for LASSO models performance based on *Fmax* did not reveal any pairwise comparison with statistical significance.

Given an overall lower *Fmax* performance for LASSO models developed with *pmax* = 5, 6, 7, and a wider distribution amplitude of *Fmax* values for *pmax* = 8, we chose as the best performing Pre-Biosy LASSO model the one developed with *pmax* = 9, which attained higher median *Fmax* value (0.620).

4.3.1.2 Priority-LASSO-2-blocks

As previously mentioned at the end of the section 4.2.2.1, we discarded Priority-Lasso-2-blocks *Radiomics* > *Clinical* block priority sequence for its considerable inferior performance (Figure 4.2) when comparing with the alternative *Clinical* > *Radiomics* sequence (V = 0, $p < 0.0001$, Wilcoxon signed rank test), regardless of parameter configurations.

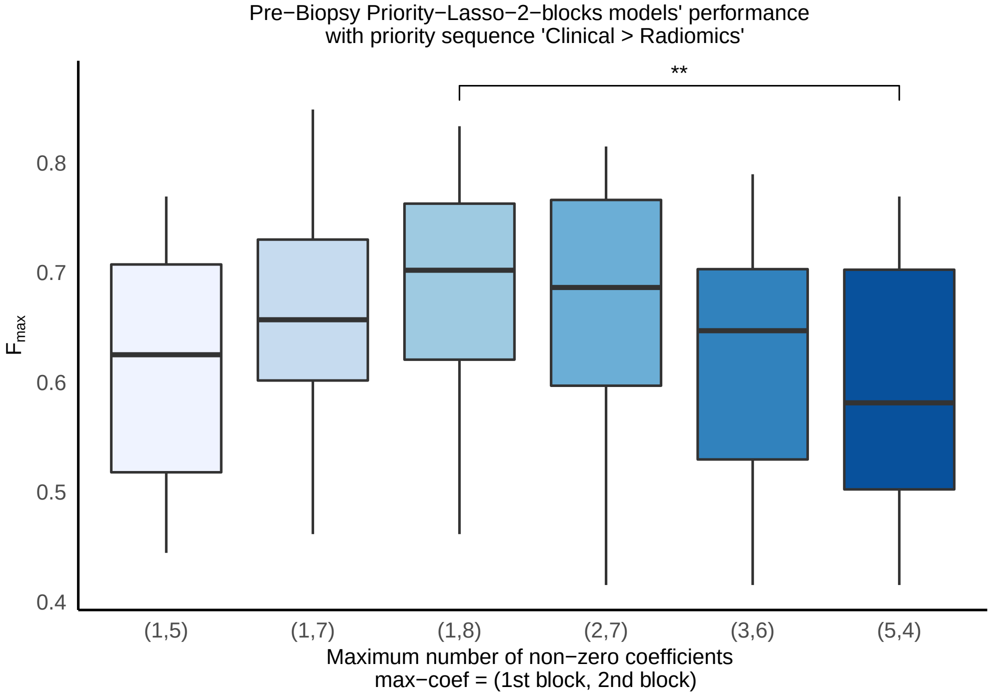


Figure 4.5: Boxplot and statistical analysis comparing Pre-biopsy Priority-LASSO-2-blocks models' performances developed with *Clinical* > *Radiomics* block priority, according to the maximum number of non-zero coefficients of each constitutive block. Significant difference in *Fmax* score was found for *pmax* = (1,8) vs. (5,4), through multiple pairwise comparisons Nemenyi test. No other statistical differences were found for other model configurations. Legend: '∗∗' $p < 0.01$.

The *Fmax* distributions and the results of the statistical analysis performed to compare the performances of Pre-biopsy Priority-LASSO-2-blocks models developed with priority sequence *Clinical* > *Radiomics* and different *pmax* hyperparameter values are shown in Figure 4.5.

Friedman test indicated significance for differences in *Fmax*, according to the maximum allowed parameters for each of the two blocks ($\chi^2(5) = 18.135$, $p < 0.01$).

Median (IQR) *Fmax* values for *pmax* = (1,5), (1,7), (1,8) were 0.625 (0.518 to 0.707), 0.657 (0.602 to 0.730) and 0.702 (0.621 to 0.763), respectively. Then, for *pmax* = (2,7), (3,6), (5,4), the median values were 0.686 (0.597 to 0.766), 0.647 (0.530 to 0.703) and 0.581 (0.502 to 0.703). Post-hoc multiple pairwise comparisons with Nemenyi test revealed a significant difference for *Fmax* scores for models developed with *pmax*= (5,4) and (1,8) ($p < 0.01$). There were no significant differences between any other pairwise comparisons ($p > 0.05$).

Given such, we chose as the best performing configuration *pmax* = (1,8) for Priority-LASSO-2-blocks, with block priority sequence *Clinical* > *Radiomics*, as it attained the highest median *Fmax* value (0.702) during model development.

4.3.1.3 Priority-LASSO-4-blocks

Fmax performance distributions resulting from the implementation of Pre-Biopsy Priority-LASSO-4-blocks with 24 priority sequences, where each encompasses all tested model configurations in it, are presented in Figure 4.6.

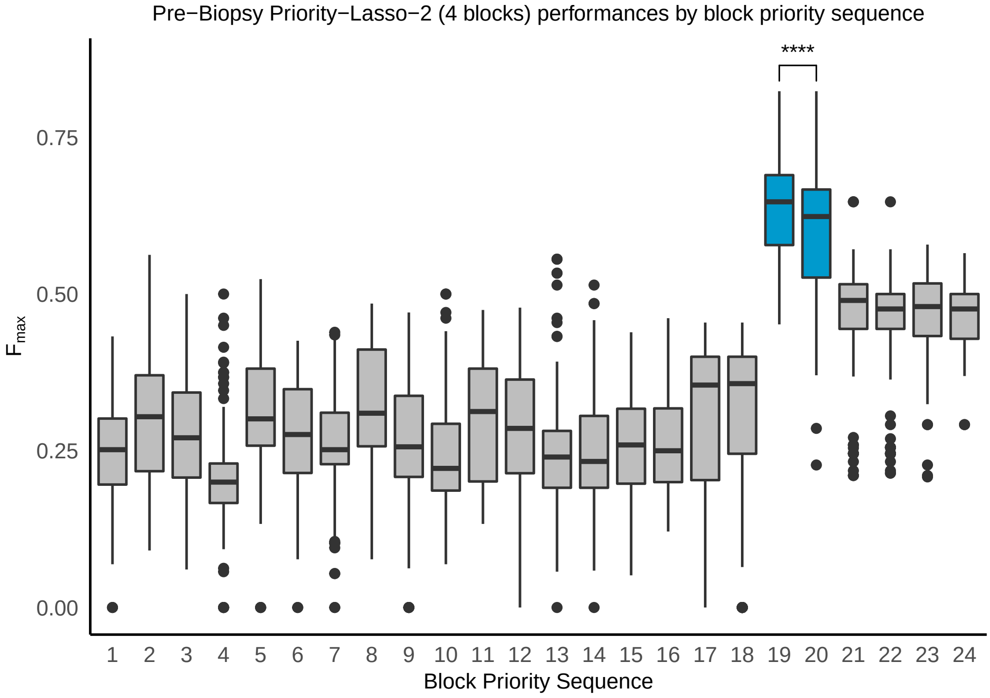


Figure 4.6: Boxplots of *Fmax* performance distributions for each block priority sequence tested in Priority-LASSO-4-blocks, and statistical analysis of the highest performance sequences 19 and 20. Results from a Wilcoxon test revealed a highly significant difference between *Fmax* scores using priority sequence 19 and sequence 20, where the former had a better performance. Legend: '∗∗∗∗' $p < 0.0001$.

Visual inspection of these distributions allows to pinpoint block priority sequences 19 and 20 as having considerable higher performance. For sequences 1 to 18, where the highest priority is given to features derived from one image modality (T2, b1000 or ADC, see Table 3.2), the models' performances were almost always below *Fmax* = 0.5, whereas for sequences 21 to 24, this metric's median value was approximately 0.5. Sequences 19 and 20 are *Clinical* > *T2* > *b1000* > *ADC* and *Clinical* > *T2* > *ADC* > *b1000*, respectively.

There was a statistically significant difference in *Fmax* metric according to the implemented priority sequence ($\chi^2(23) = 1531.2$, $p < 0.0001$, Friedman test). Post-hoc analysis with Wilcoxon signed-rank sum test was conducted only for priority sequences 19 and 20, with median (IQR) Fmax values of 0.6471 (0.578 to 0.690) and 0.624 (0.526 to 0.667), respectively, where a highly significant difference between both was found (V = 2675.5, $p < 0.0001$, Wilcoxon test).

Thus, we concluded that the *Fmax* performance achieved with Pre-biopsy Priority-Lasso-4-blocks priority sequence 19 was significantly different from sequence 20, where it achieved higher results throughout the tested models' configurations (see Table 3.2).

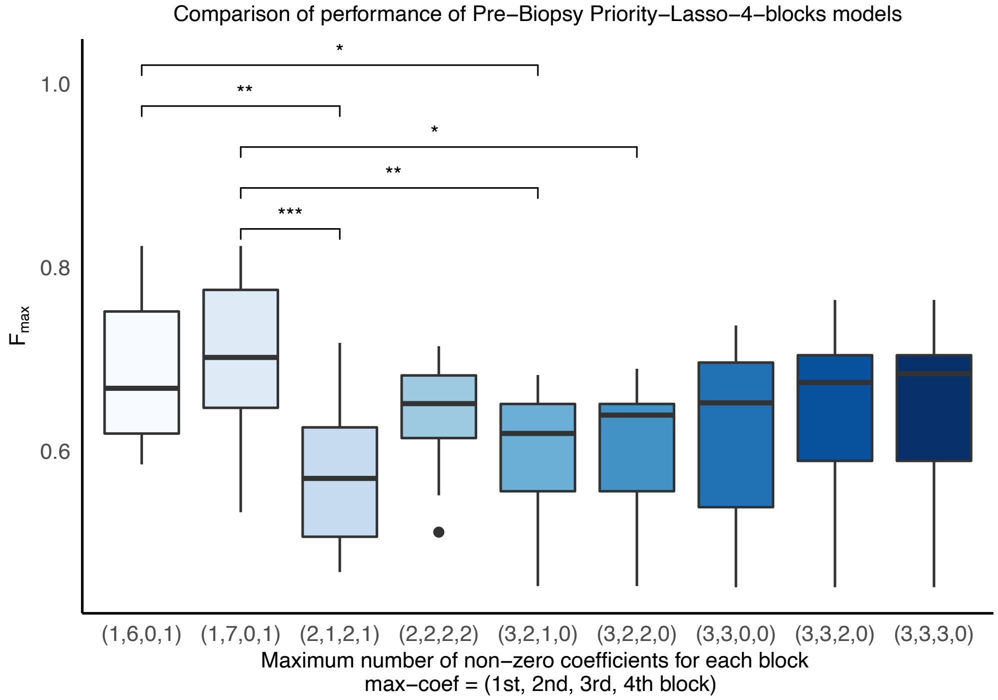


Figure 4.7: Boxplots and statistical analysis of Fmax scores with Pre-biopsy Priority-LASSO-4-blocks, priority sequence 19: *Clinical* > *T2* > *b1000* > *ADC* for different block configurations. Results of statistically significant differences according to Nemenyi pairwise multiple comparison test are shown. Legend: '∗' $p \leq 0.05$; '∗∗' $p \leq 0.01$; '∗∗∗' $p \leq 0.001$.

Parameter Tuning of Best Priority Sequence Figure 4.7 shows how *Fmax* performances of Pre-biopsy Priority-LASSO-4-blocks implemented with sequence 19 (*Clinical* > *T2* > *b1000* > *ADC)* varied with limitations of maximal number of features per block, as specificied by *pmax*. *Fmax* scores obtained for the referred model significantly varied with tested model configuration ($\chi^2(10) = 55.978$, $p < 0.0001$, Friedman test). The top performing configurations were *pmax* = (1,7,0,1) and (1,6,0,1) (Figure 4.7), with median (IQR) Fmax scores of 0.702 (0.647 to 0.776) and 0.668 (0.619 to 0.752), respectively. No statistical difference was found between them (p = 1, Nemenyi test).

However, post-hoc analysis with Nemenyi test revealed significant difference for *Fmax* scores of models developed with *pmax* = (1,7,0,1) and (2,1,2,1) ($p < 0.001$), as well as with *pmax* = (3,2,1,0) ($p < 0.01$) and (3,2,2,0) ($p < 0.05$). Also, we found evidence of significant differences between *Fmax* scores for *pmax* = (1,6,0,1) and (2,1,2,1) ($p < 0.01$), as well as (3,2,1,0) ($p < 0.05$).

Based on this, we chose the best performing Pre-biopsy Priority-LASSO 4-blocks model configuration, in terms of Fmax score, as the one with *pmax* set to (1,7,0,1), using priority sequence 19: *Clinical* > *T2* > *b1000* > *ADC*.

4.3.2 Preoperative

4.3.2.1 LASSO

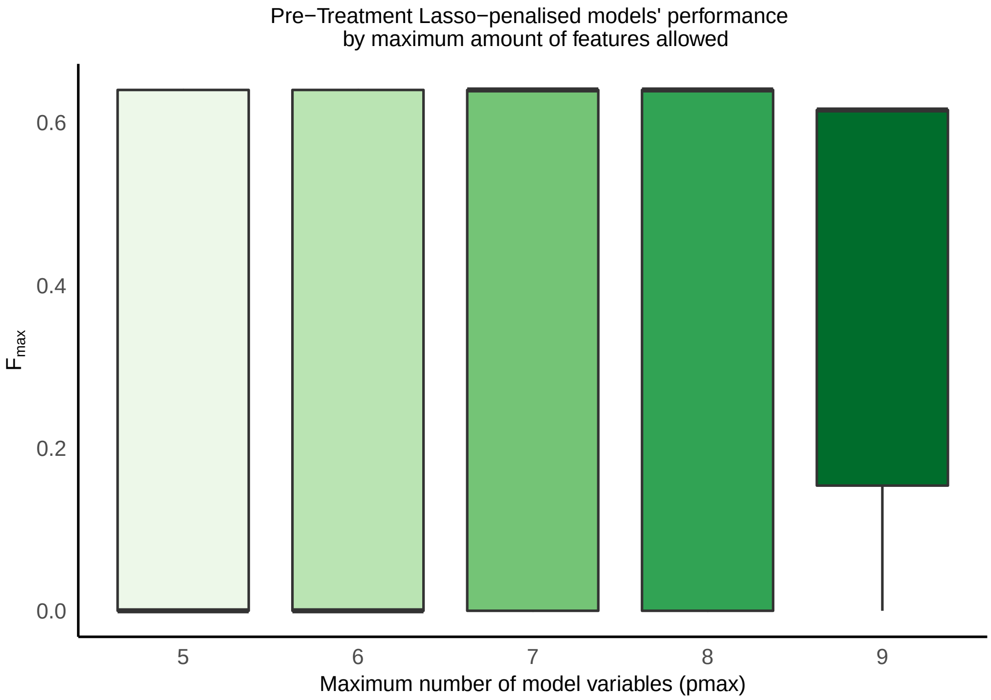


Figure 4.8: Boxplots of Preoperative LASSO models' performance (*Fmax* scores) according to the maximum allowed number of model variables (pmax). Statistical results showed no significant differences in Fmax scores (Friedman test).

Preoperative LASSO logistic regression models' performance according to different maximal number of model variables are depicted in Figure 4.8. These models' solutions rendered very poor *Fmax* scoring for $pmax = 5$ and 6, with median (IQR) Fmax scores of 0 (0 to 0.64). For $pmax = 7$ and 8, the results were very unstable, with median (IQR) being 0.640 (0 to 0.640). Comparatively, LASSO performance with $pmax = 9$ was less unstable, with *Fmax* scores with a median value (IQR) of 0.6154 (0.154 to 0.6154).

The presented model variants performed similarly, as there was no statistically significant difference in *Fmax* scores' distributions according to Friedman's test ($\chi^2(4) = 3.302$, $p > 0.05$).

Given the slightly more stable performance of LASSO solutions with $pmax = 9$, we chose this model configuration as the best preoperative LASSO model for subsequent analysis.

4.3.2.2 Priority-LASSO-2-blocks

We discarded Priority-LASSO-2-blocks *Radiomics > Clinical* block priority sequence for its considerable inferior performance when comparing with the alternative *Clinical > Radiomics* sequence (V = 11, $p < 0.0001$, Wilcoxon signed rank test), regardless of parameter configurations (as referred in Section 4.2.2.1). Therefore, we tackle our analysis to the alternative sequence *Clinical > Radiomics*.

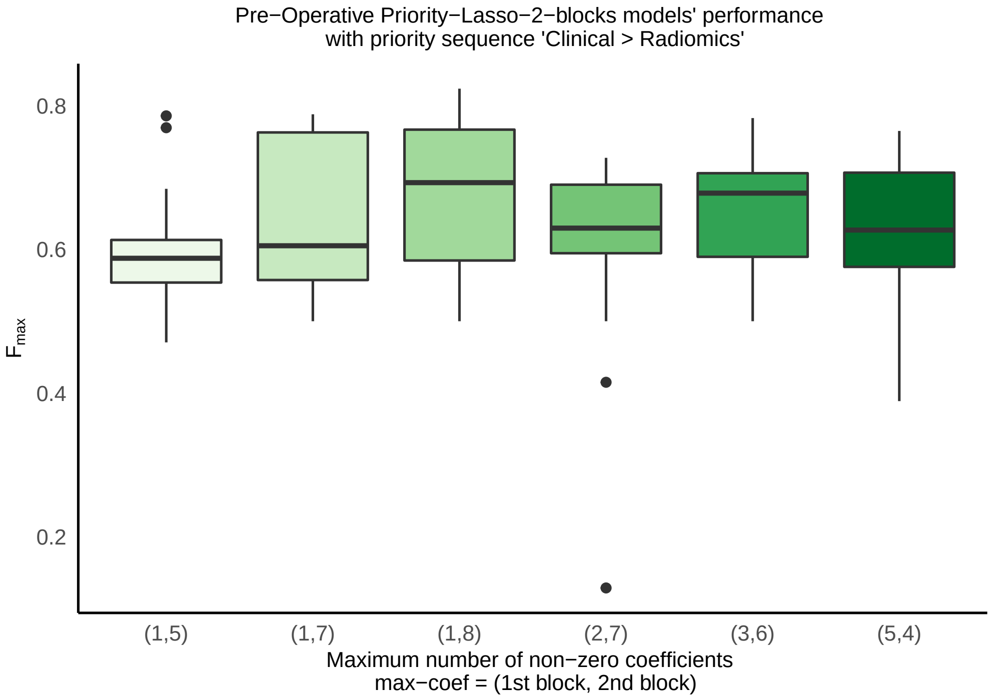


Figure 4.9: Boxplot of Preoperative Priority-LASSO-2-blocks models' performance (*Fmax*) developed with *Clinical* > *Radiomics* block priority, according to the maximum number of non-zero coefficients for each constitutive block. No statistical differences were found between models' implementations (Friedman test).

In Figure 4.9 we have Preoperative Priority-LASSO-2-blocks implemented with priority sequence *Clinical* > *Radiomics* and different values for the hyperparameter *pmax*, which limits the amount of variables in each constitutive block.

There was no statistically significant difference in *Fmax* scores resulting from Preoperative Priority-LASSO-2-blocks penalised models, depending on the amount of maximum allowed variables ($\chi^2(5) = 8.7255$, $p > 0.05$, Friedman test).

Median (IQR) Fmax values for $pmax = \{(1,5), (1,7), (1,8)\}$ were 0.588 (0.554 to 0.613), 0.605 (0.557 to 0.763) and 0.693 (0.585 to 0.767), respectively. As for $pmax = (2,7), (3,6), (5,4)$, the median (IQR) values were 0.630 (0.595 to 0.690), 0.678 (0.590 to 0.706) and 0.627 (0.576 to 0.707).

We chose as the best performing configuration $pmax = (1,8)$ that attained the highest median *Fmax* score (0.693) during model development.

4.3.2.3 Priority-LASSO-4-blocks

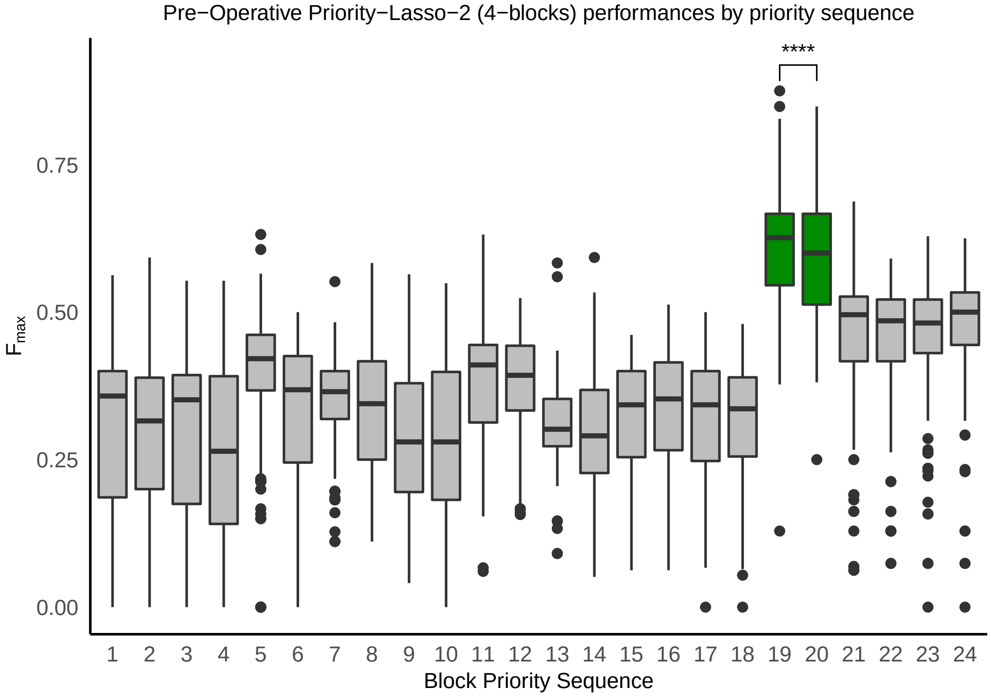


Figure 4.10: Boxplots of *Fmax* performance for implemented block priority sequences in Preoperative Priority-LASSO-4-blocks, and statistical analysis of the highest performance sequences 19 and 20. Results from a Wilcoxon test revealed a highly significant difference between *Fmax* scores of priority sequences 19 and 20. Legend: '$****$' $p < 0.0001$.

Figure 4.10 presents the performance of Preoperative Priority-LASSO-4-blocks, of the different block priority sequences implemented in terms of *Fmax*.

As in Pre-biopsy scenario analysis, the visualisation of the performance of the 24 implemented block priority sequences with Priority-LASSO-4-blocks lets us pinpoint the sequences 19 and 20 as those having achieved considerable higher performance than the remaining – which median *Fmax* values were approximately equal or less than 0.5.

There was a statistically significant difference in *Fmax* scores depending on the block priority sequence used to fit Priority-LASSO-4-blocks models ($\chi^2(23) = 1227.7$, $p < 0.0001$, Friedman test). Post-hoc analysis with Wilcoxon signed-rank sum test was conducted for the top-performing pair of sequences 19 and 20, with median (IQR) Fmax values of 0.626 (0.546 to 0.667) and 0.600 (0.513 to 0.667), that indicated a highly statistically significant difference between them ($V = 1807.5$, $p < 0.0001$).

Thus, we conclude that *Fmax* score distributions achieved with the implementation of sequence 19th was significantly different than 20th, with higher results when considering all tested models' configurations.

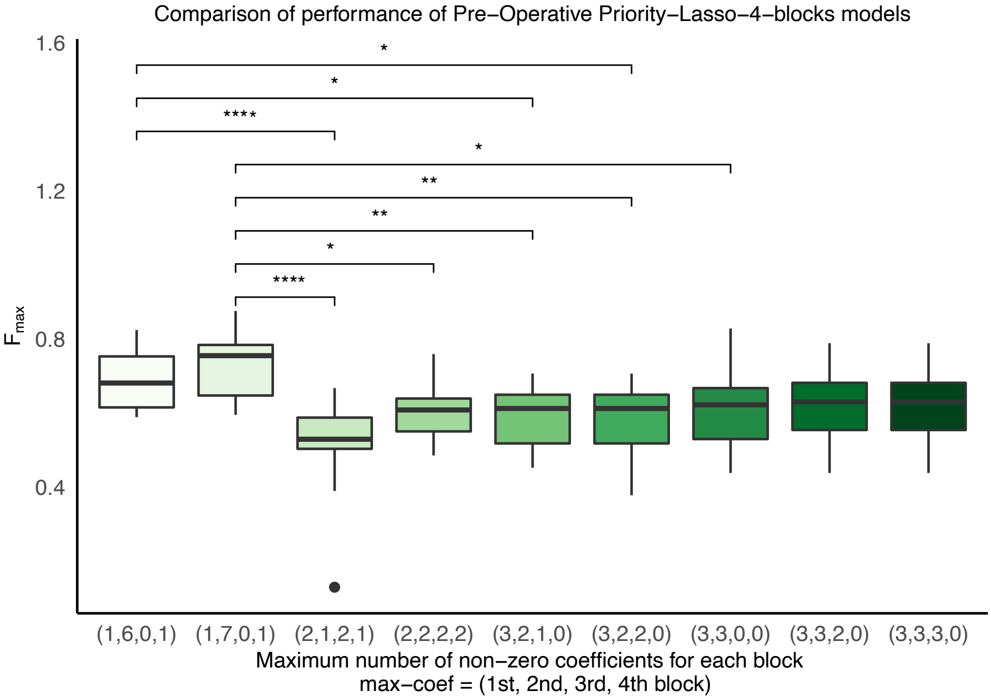


Figure 4.11: Boxplots and statistical analysis of *Fmax* scores with Preoperative Priority-LASSO-4-blocks for different block configurations, for block sequence 19. Results of statistically significant differences according to Nemenyi pairwise comparison test are shown. Legend: '$*$' $p \leq 0.05$; '$**$' $p \leq 0.01$; '$****$' $p \leq 0.0001$.

Parameter Tuning of Best Priority Sequence In Figure 4.11 we see the different resulting performances of Preoperative Priority-LASSO-4-blocks models implemented with top sequence 19 (*Clinical* $>$ *T2* $>$ *b1000* $>$ *ADC*), when varying the hyperparameter *pmax* that delimited the maximal number of variables per block.

There was a statistically significant difference in *Fmax* scores according to different limitations on the number of variables per block ($\chi^2(8) = 46.119$, $p < 0.0001$, Friedman rank sum test). For priority sequence 19, the top performing configurations were *pmax* = (1,7,0,1) and (1,6,0,1) (Figure 4.11, with median (IQR) *Fmax* scores of 0.754 (0.647 to 0.783) and 0.680 (0.615 to 0.752), respectively. Overall post-hoc analysis with Nemenyi test did not reveal a significant difference between these two configurations ($p < 0.05$).

Based on this, we chose as the best performing Preoperative Priority-LASSO-4-blocks model configuration, in terms of *Fmax*'s score, as the one with *pmax* set to (1,7,0,1), for block priority sequence 19.

4.3.3 Postoperative

4.3.3.1 LASSO

All Postoperative LASSO models yielded *Fmax* scores of 0.667, across *pmax* = {5, 6, 7, 8, 9} and in all 14 CV runs. Besides providing very stable solutions, Postoperative LASSO developed models which consistently selected 3 variables.

Given the equality of the results, no statistical analysis to evaluate differences in *Fmax* distributions was necessary to be performed. Any of the considered model configurations could be considered for further analysis and comparison. However, LASSO implementation with *pmax* = 9 was selected to allow for a better comparison with the previous clinical scenarios' results.

4.3.3.2 Priority-LASSO-2-blocks

As previously mentioned at the end of the Section 4.2.2.1, we discarded the implemented priority sequence *Radiomics* > *Clinical* of Priority-LASSO-2-blocks for its evident inferior (*Fmax*) performance compared with the alternative *Clinical* > *Radiomics* (V = 0, $p < 0.0001$, Wilcoxon signed rank test), irrespectively of parameter configurations.

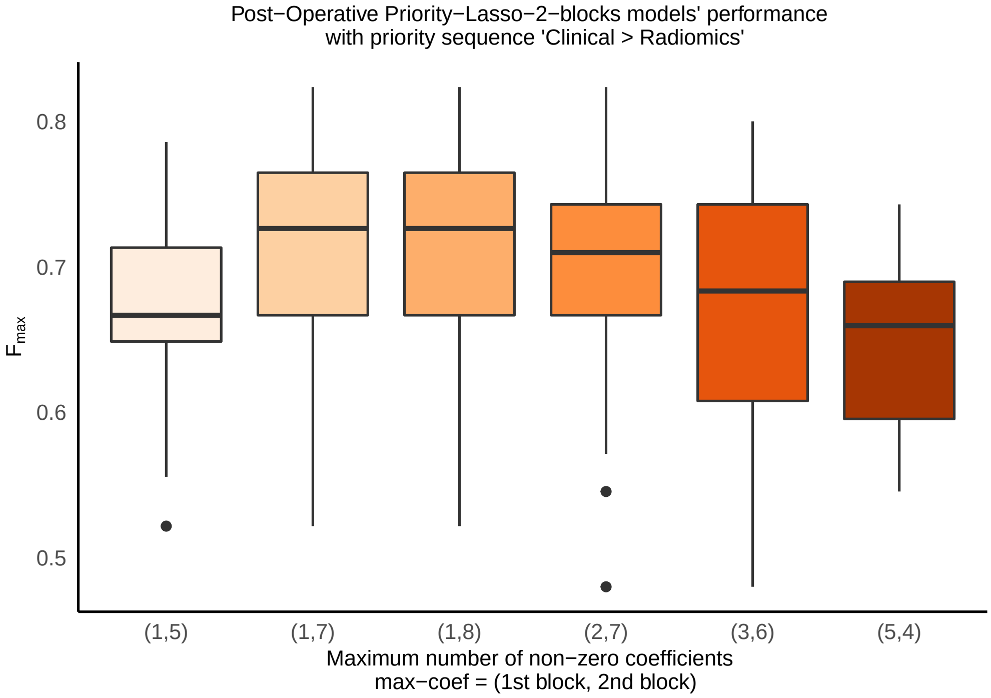


Figure 4.12: Boxplot of Postoperative Priority-LASSO-2-blocks models' performances (*Fmax* score) using priority block sequence *Clinical* > *Radiomics*, according to different restrictions of maximum allowed number of variables in each constitutive block. Both *pmax* = (1,7) and (1,8) configurations rendered the same *Fmax* scores' distributions, of median value 0.726. No significant differences in *Fmax* scores' distributions with different priority block restrictions were found with Friedman test.

In Figure 4.12, we show how the different *pmax* hyperparameters with which Postoperative Priority-LASSO-2-blocks' models were developed affected *Fmax* performance.

Friedman test did not reveal significance for differences in *Fmax* median scores, according to the maximum allowed parameters for each of the two blocks ($\chi^2(5) = 8.7115$, $p > 0.05$). In particular, the assessed model performance metrics for *pmax* = (1,7) and (1,8) did not differ, meaning that restricting the model in one unit on the second priority block did not interfere with the solutions created.

Given such, we will choose Priority-LASSO-2-blocks configuration with *pmax* = (1,8) with median *Fmax* = 0.7263, to allow for a better comparison in-between clinical scenarios.

4.3.3.3 Priority-LASSO-4-blocks

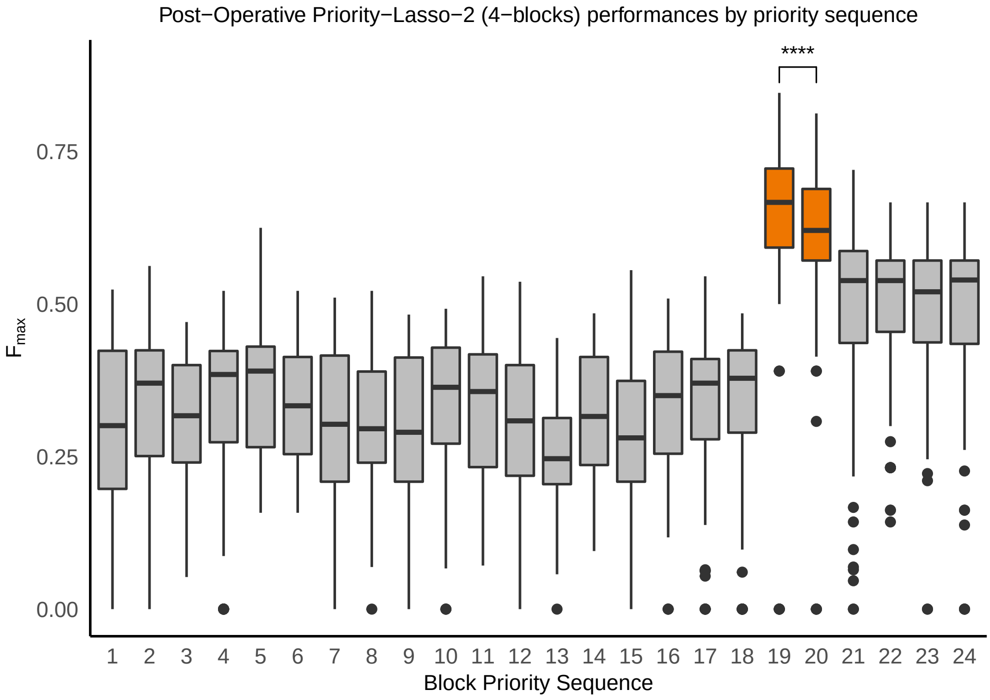


Figure 4.13: Boxplot of *Fmax* performance distributions for each block priority sequence implemented with Postoperative Priority-LASSO-4-blocks, and statistical analysis of the highest performance sequences 19 and 20 (highlighted in orange). A pairwise Wilcoxon test revealed a highly significant difference between sequences 19's and 20's *Fmax* scores (V = 2840, $p < 0.0001$). Legend: '$****$' p $\leq$ 0.0001.

On Figure 4.13 we show the distributions of *Fmax* scores derived from Priority-LASSO-4-blocks models implemented with 24 different block priority sequences and statistical analysis results.

The visual representation of *Fmax* metric distributions on Figure 4.13 allows to notice a considerable higher performance of sequences 19 and 20. Sequences from 1 to 18 performed very poorly, where an improvement was seen from sequences 19 to 24 (block priority sequences specified in Table 3.2).

Friedman test indicated a statistically significant difference in *Fmax* metric values according to the priority sequence ($\chi^2(23) = 1163.4$, $p < 0.0001$). Post-hoc analysis with Wilcoxon signed-rank sum test was conducted only for top performing sequences, 19 and 20, with median (IQR) *Fmax* values of 0.667 (0.593 to 0.722) and 0.621 (0.571 to 0.689), respectively. A highly significant difference between these two sequences *Fmax* performances was found (V = 2840, $p < 0.0001$).

Thus, we concluded that priority sequence 19 was significantly different than 20, where it achieved higher results throughout the tested models' configurations.

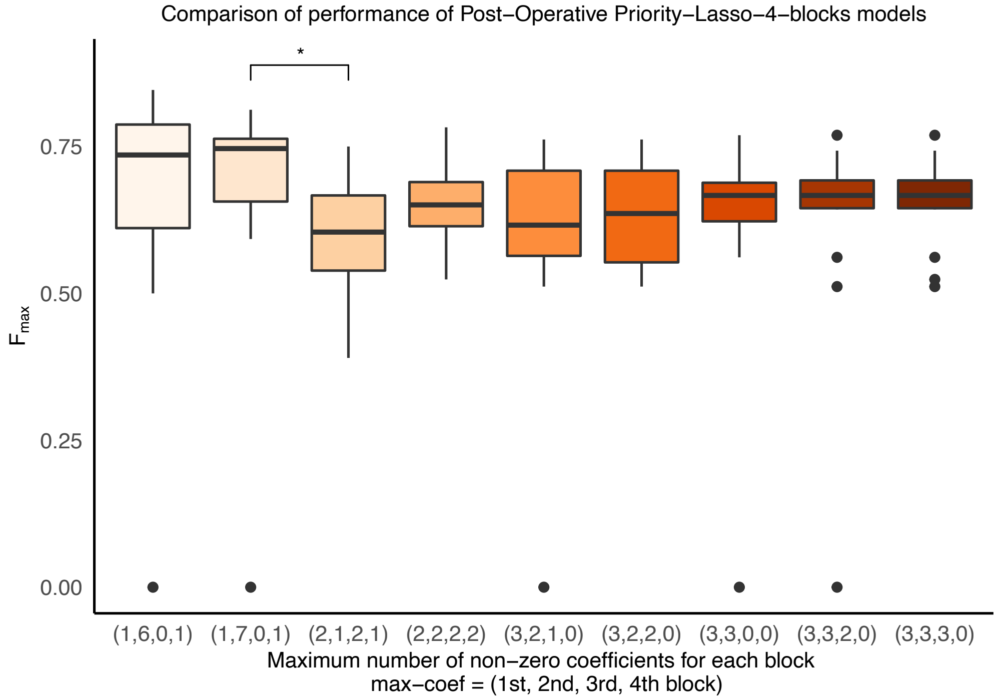


Figure 4.14: Boxplots and statistical analysis of *Fmax* scores with Postoperative Priority-LASSO-4-blocks for different block configurations, with priority sequence *Clinical* > *T2* > *b1000* > *ADC*. *Fmax* scores' distributions with *pmax* = (1,7,0,1) and (2,1,2,1) were statistically significantly different (Nemenyi post-hoc test). Legend: '∗' $p \leq 0.05$.

Parameter Tuning of Best Priority Sequence The priority sequence 19 (*Clinical* > *T2* > *b1000* > *ADC*), identified as the top performing priority sequence, was implemented with different limitations on the number of variables per block (*pmax*). On Figure 4.14, the distributions of *Fmax* scores and statistical analysis results are shown.

Postoperative Priority-LASSO-4-blocks models implemented with priority sequence 19 achieved *Fmax* performances that statistically significant differed according to the maximum number of variables allowed for each block ($\chi^2(8) = 20.163$, $p < 0.01$, Friedman test).

Particularly, a statistically significant difference on *Fmax* scores of models developed with *pmax* = (1,7,0,1) and (2,1,2,1) was found ($p < 0.05$, Nemenyi post-hoc test). This seems to correspond to the difference between higher and lower performing model configurations, with median (IQR) *Fmax* scores of 0.746 (0.656 to 0.763) and 0.604 (0.539 to 0.667), respectively.

Based on this, we chose Priority-LASSO-4-blocks' configuration with *pmax* = (1,7,0,1) as the best Postoperative model for this algorithm type.

4.3.4 Conclusions

The best performing models for each clinical scenario, within each algorithm type, regardless of clinical scenario, were LASSO with *pmax* = 9, P-LASSO-2-blocks with *pmax* = (1,8) and P-LASSO-4-blocks with *pmax* = (1,7,0,1).

4.4 Top Performing Models Comparison per Clinical Scenario

The results of the classification algorithms applied with different penalisation methods, evaluated in a repeated CV setting, for different clinical scenarios, are shown in Fig. 4.15.

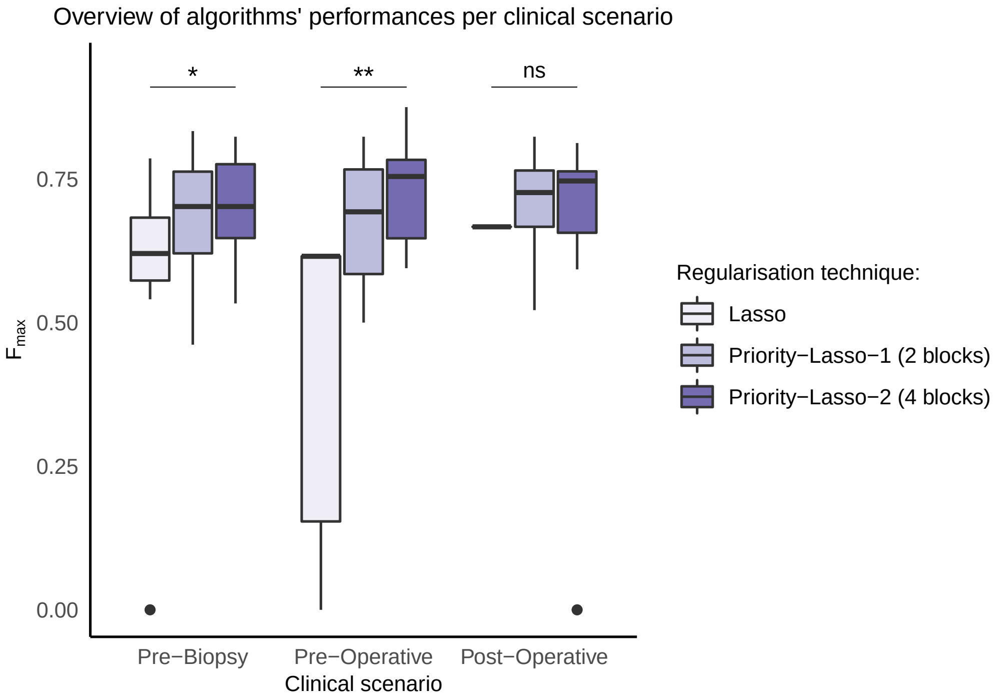


Figure 4.15: Overview of the performance of the best algorithms, per clinical scenario and model type (regularisation method applied). Statistical analysis of the different algorithms' *Fmax* performances (Friedman test) for the three clinical scenarios considered. Legend: 'ns' $p > 0.05$, '$*$' $p \leq 0.05$, '$**$' $p \leq 0.01$.

For Pre-Biopsy scenario, there was a statistically significant difference in *Fmax* scores distribution for the different implemented algorithms ($\chi^2(2) = 6.8727$, $p < 0.05$, Friedman test). Post-hoc analysis with Nemenyi test revealed a significant difference between LASSO and P-LASSO-4-blocks ($p < 0.05$). However, no statistically significant difference was found between LASSO and P-LASSO-2-blocks nor between the two implementations of P-LASSO ($p > 0.05$).

The same held true for the Preoperative setting, where *Fmax* distributions revealed to be statistically significant different ($\chi^2(2) = 10.429$, $p < 0.01$, Friedman test) and Nemenyi post-hoc test revealed significant difference between *Fmax* distributions of LASSO and Priority-LASSO-4-blocks (p < 0.01), but no differences between Priority-LASSO-2-blocks and -4-blocks nor between LASSO and Priority-LASSO-2-blocks ($p > 0.05$).

Lastly, Friedman test yielded no statistically significant difference between *Fmax* distributions of any of Postoperative models ($\chi^2(2) = 5.4815$, $p > 0.05$).

Overall, Lasso performances were very unstable, across clinical scenarios, and Priority-LASSO-4-blocks achieved higher median *Fmax* values across clinical scenarios (0.702, 0.754, 0.746 respectively for Pre-Biopsy, Preoperative and Postoperative).

4.4.1 Best Models Selection

Having chosen the model parameters in the most favourable sense of *Fmax* metric in each model and each clinical scenario, and combining the statistical results depicted in Figure 4.15, the best performing model across all clinical scenarios was Priority-LASSO-4-blocks implemented with priority sequence 19: *Clinical* > *T2* > *b1000* > *ADC*, with *pmax* = (1,7,0,1). Thus, it was chosen as the algorithm to be the final radiomics-and-clinicohistopathologic-based classifier of each clinical scenario.

The associated thresholds for binarising the probabilistic predictions generated by the classifiers into discrete BCR negative/positive labels were 0.27779, 0.28451 and 0.26213, the average of the thresholds found to maximise *Fmax* for Pre-biopsy, Preoperative and Postoperative Priority-LASSO-4-blocks in the fourteen cross-validation runs, respectively.

Overall, the best performing model of all the models tested in this ML-framework, in terms of median *Fmax* achieved during development was Preoperative Priority-LASSO-4-blocks with *pmax* = (1,7,0,1), with median (IQR) *Fmax* of 0.7543 (0.647 to 0.783).

4.5 Model Performance Evaluation with Oversampling

The performance of the best classification algorithms selected within this framework was evaluated with and without SMOTE random over-sampling, to address the substantial class imbalance in the development set, i.e. a much higher number of BCR-negative patients (n = 51) than BCR-positive ones (n = 15). For all the algorithms and for each clinical scenario, random over-sampling led to significantly improved performance across all the evaluation measures as compared to no oversampling (pairwise Wilcoxon signed rank-sum tests, $p < 0.0004$).

For oversampled models, the correspondent thresholds that allowed the dichotomisation of the probabilistic predictions into the two classes, optimising *Fmax* metric, were found to be 0.52840, 0.54989 and 0.54854, as the mean values of the thresholds for Pre-Biopsy, Preoperative and Postoperative models, respectively.

In Figure 4.16 we show *Fmax* distributions and the referred difference in performance is visible. And thus, we can conclude that SMOTE improved the performance of the best performing classifiers considered, on the development set.

Overall, the algorithm that performed the best was Priority-LASSO. The implementation of this algorithm with 4 blocks performed the best for models where no oversampling technique was used, whereas Priority-LASSO with 2 blocks performed better for oversampled development sets. The latter, achieved median *Fmax* values of 0.913, 0.924 and 0.929 for Pre-Biopsy, Preoperative and Postoperative scenarios.

Regardless of the usage of oversampling techniques, the best performing Priority-LASSO models were found for Preoperative scenario.

Given the considerable higher performance of models where SMOTE oversampling was conducted, we will use them for parallel evaluation of the final models.

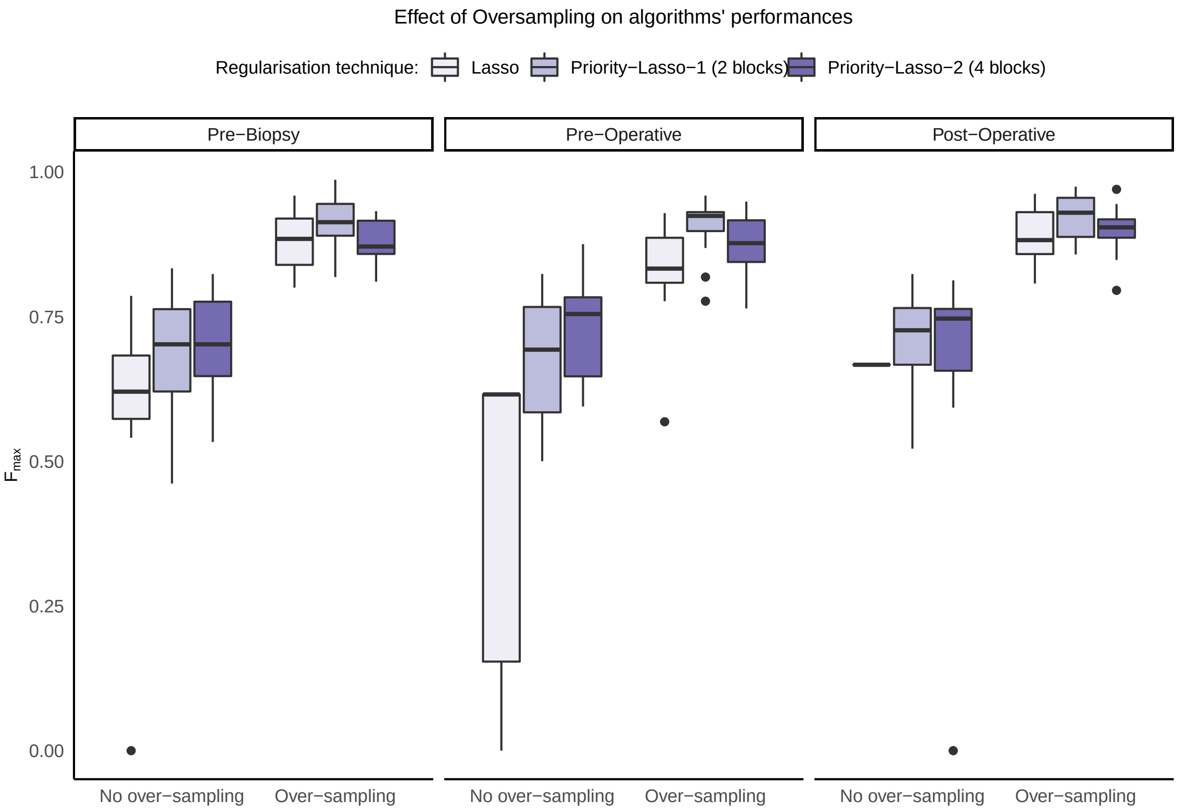


Figure 4.16: Boxplots of the performance of the selected algorithms (*Fmax*) for each clinical scenario, according to the usage or non-usage of SMOTE over-sampling technique in the development set.

4.6 Evaluation of Final Classifier

The main goal of this study was to compare 2-yr BCR classification methods using biparametric MRI radiomics data together with clinicohistopathological information. In this section we first give an overview of the classification performance of the selected models, for each clinical scenario.

The final classifiers of each clinical scenario, i.e. Priority-LASSO-4-blocks with block priority sequence 19: *Clinical* > *T2* > *b1000* > *ADC* and $pmax$ = (1,7,0,1), developed with and without random over-sampling of the minority class, were applied to the independent cohort to assess the generalisability of the models. The independent cohort can also be referred to as the validation or test set. Models' performances were measured in terms of AUC (threshold-free metric), F-measure (*Fmax*), Precision (*Pmax*) and Recall (*Rmax*) (Table 4.2), for each class (BCR positive as the minority class, and BCR negative as the majority class). The final values were taken as the median from the metrics' distributions of each classifier, calculated over 14 CV runs.

We recall that the original (non-oversampled) dataset was divided into train and test sets with a ratio of 70:30, in a stratified way, and models were developed and evaluated with a repeated stratified cross-validation scheme, with 14 runs. The development set was composed of 66 patients: 15 BCR positive (22.7%), 51 BCR negative (77.3%). The validation set comprised 27 cases: 5 BCR positive (18.5%) and 22 BCR negative (81.5%).

The SMOTE oversampled dataset followed the same premises regarding the pipeline of evalu-

Table 4.2: Evaluation of the final Priority-LASSO-based classifiers, developed with and without SMOTE oversampling technique, for different clinical scenarios, on the independent validation set of 27 PCa patients (5 positive, 22 negative). Evaluation metrics are AUC from ROC curve, F-measure (*Fmax*), Precision (*Pmax*) and Recall (*Rmax*), where the final values were taken as the median from the metrics' distributions of each learner calculated over 14 CV iterations.

	No Oversampling							
Clinical Scenario	**Set**	**AUC**	**BCR positive (minority) class**			**BCR negative (majority) class**		
			Fmax	***Pmax***	***Rmax***	***Fmax***	***Pmax***	***Rmax***
Pre-Biopsy	Train	0.883	0.702	0.656	0.733	0.913	0.927	0.875
	Test	0.739	0.400	0.382	0.450	0.829	0.873	0.799
Pre-Operative	Train	0.882	0.754	0.682	0.828	0.925	0.949	0.900
	Test	0.638	0.297	0.333	0.292	0.779	0.821	0.783
Post-Operative	Train	0.895	0.746	0.659	0.823	0.910	0.947	0.900
	Test	0.691	0.382	0.333	0.400	0.833	0.845	0.823
	Oversampling							
Pre-Biopsy	Train	0.923	0.871	0.872	0.886	0.878	0.884	0.868
	Test	0.640	0.312	0.250	0.450	0.754	0.826	0.655
Pre-Operative	Train	0.933	0.877	0.882	0.904	0.869	0.905	0.886
	Test	0.583	0.258	0.200	0.464	0.725	0.805	0.614
Post-Operative	Train	0.958	0.904	0.935	0.869	0.909	0.881	0.935
	Test	0.701	0.369	0.303	0.536	0.823	0.869	0.804

ation, with the exception that the models were developed with a balanced cohort: 51 randomly generated BCR positive cases and 51 negative (original) BCR cases. The validation of the oversampled models was made with the same test set of 27 patients as specified above, since we aimed at understanding the impact of minority class balance during model development on performance.

Table 4.2 allows to compare models' performances when classifying the train and test sets, in each clinical setting. Also, we can compare the models' performances with and without the usage of SMOTE oversampling techniques.

We recall that our models were developed with the goal of optimising the number of correctly classified BCR positive cases (through F-measure maximisation), and not with the aim to maximise overall accuracy. For this reason, this commonly used metric is not reported.

The most evident aspect of our final models' evaluation is the performance discrepancy between training and testing sets, that is the typical depiction of model overfitting during development. Training results are too optimistic and testing results demonstrate that models were not able to generalise to unseen data.

In any case, all classifiers achieved good performance during training phase to classify BCR-positive cases, especially in terms of *Fmax* and *Rmax*, with values ranging from 0.702 to 0.754 (non-oversampled) and from 0.871 to 0.904 (oversampled). In fact, the machine-learning framework was designed to optimise *Fmax* metric so that the algorithms were learning with the specific goal of correctly identify BCR positive cases. However, despite this implemented maximisation procedure during model training, the models performed poorly during model validation with the test set, with minority class-specific performance metrics yielding values lower than 0.450.

Yet, models revealed great performance on the classification of the BCR-negative class, both during training and validation phases.

SMOTE oversampling was implemented to artificially balance the ratio of positive and negative cases showcased to the models during training phase, with expectations that it could improve BCR positive cases' classification. Comparing the performances of models that used SMOTE sampling technique to the performance of models that did not, we see that the SMOTE models generally provided superior performances during training. However, for validation phase, models developed with oversampling technique performed worst for almost all class-specific metrics, except minority-class *Rmax* (Recall). We conclude that this technique was overoptimistic during validation and performed worse than non-oversampling models during test, and thus not providing improvements of neither the minority nor the majority classes classification. Therefore, only results based on non-oversampled models are discussed from here on.

Despite Pre-biopsy models having yielded the lowest performance during training phase, for both minority and majority classes, in validation phase it achieved the highest median values of AUC (0.739), *Fmax* (0.400), *Pmax* (0.382) and *Rmax* (0.450) for the BCR positive class and *Pmax* (0.873) for the BCR negative cases.

The Preoperative classifier produced the most accurate predictions of both classes during training, except for the threshold-free measure of AUC. However, it was an overoptimistic performance, given that Preoperative models showed the worst generalisability among all, with a huge discrepancy between Train and Test sets' classification performances of the two classes. Regarding the majority class classification, i.e. BCR negative cases, Post-operative model was the best performer in terms of *Fmax* (0.833) and *Rmax* (0.823).

As the chosen model configuration was the same for all clinical scenario's final model (Priority-LASSO-4-blocks, sequence 19 with $pmax$ = (1,7,0,1)), we conclude that the addition of clinico-histopathological data, that successively becomes available through the normal course of PCa clinical management, did not improve the classification of 2-year BCR positive cases.

4.6.1 Selected Variables

On Table 4.3 we categorise the variables selected for the final models, over the 14 CV runs. Regression coefficients are not reported as they were too variable within the considered CV runs. Yet, features were organised by their level of importance for the calculated solution, i.e. by descending order of their coefficients, indicating the relative weight of each category of feature. Then, we grouped the chosen variables by their type (clinical, histopathological or image) and, in the case of image, by image modality (T2, DWI b1000 or ADC map) and radiomic group type (first order, texture, shape). Finally, variables were counted accordingly.

First, the variety of the results in terms of selected features over 14 cross-validation runs shows model instability and variability possibly coming from dataset partitioning during development, that is highly undesirable.

Yet, given the exploratory nature of this study, we grouped variables in these broad categories shown in Table 4.3 to better understand which of the many computed radiomic variable types partaking our dataset were possibly the most promising ones for prognostic prediction or characterisation of the whole prostate gland preoperative conditions.

Table 4.3: Categorisation of the final models' selected variables, across 14 CV runs, for each clinical scenario. Variables were grouped according to their importance (relative weight for the calculated solution), type, image modality and radiomic group, and counted accordingly.

Pre-Biopsy

Importance	Variable Type	Image type	Class	Counts
1	Clinical*	-	-	14
2	T2	log	Texture	8
		Wavelet	Texture	5
		Original	Shape	1
	ADC	log	Texture	1
3	T2	Wavelet	Texture	9
		log		4
4	T2	Wavelet	Texture	11
		log		1
5	T2	Wavelet	Texture	10
	ADC	log		1
6	T2	Wavelet	Texture	10
7	T2	Wavelet	Texture	8
	ADC		First order	1
8	T2	Wavelet	Texture	5
	ADC			1

Preoperative

Importance	Variable Type	Image type	Group Type	Counts
1	Clinical*	-	-	13
	T2	log	Texture	1
2	T2	log	Texture	9
		Wavelet	Texture	3
		Original	Shape	2
3	T2	Wavelet	Texture	7
		log	Texture	6
		Original	Shape	1
4	T2	Wavelet	Texture	10
	T2	log	Texture	3
	ADC	Wavelet	First-order	1
5	T2	Wavelet	Texture	12
6	T2	Wavelet	Texture	12
7	T2	Wavelet	Texture	10
		log		2
8	T2	Wavelet	Texture	7
	ADC	log		1
9	ADC	log	Texture	4

Postoperative

Importance	Variable Type	Image type	Group Type	Counts
1	Clinical**	-	-	13
	Clinical*	-	-	1
2	T2	log	Texture	7
		wavelet	Texture	4
		original	Shape	1
3	T2	log		6
		wavelet	Texture	6
4	T2	Wavelet	Texture	8
		log		4
5	T2	Wavelet	Texture	10
	ADC		First-order	1
6	T2	Wavelet	Texture	8
	ADC			1
7	T2	Wavelet	Texture	7
	ADC	log		1
8	T2	Wavelet	Texture	7
9	ADC	log	Texture	4
		Wavelet		1

*: Baseline PSA; **: Surgical resection margin extension

For these model configurations, Clinical features were almost always favoured and given higher weight in model solutions. Only one was selected, given the configuration of the chosen top model. The clinical variable *Baseline PSA* was very frequently included, except for Postoperative clinical scenario. No biopsy-derived variables were selected when available. Regarding image features, T2w-derived variables are predominantly found of relevance, where ADC ones are only chosen as of higher importance by one out of 14 models or found more represented in lower levels of importance. Clearly, DWI features were not found to be informative for these models' solutions.

PyRadiomics allows preprocessing of (applying filtering to) the original image before feature extraction. We see that the variables that were shown to play a role in this analysis were mostly computed from Laplacian of Gaussian (LoG) and Wavelet images, and only a few on the Original. None of the features computed from Gradient nor Local Binary Pattern filtered images were chosen by any model.

Regarding the class of selected features, texture-based ones were only informative when calculated from LoG or Wavelet images. Textural features subclasses GLCM, GLRM, GLSZM and GLDM were considered in all models, leaving aside only NGTDM class. Shape-based variables were only selected by a maximum of 2 models per clinical scenario, and always based on Original images. Finally, first-order features, that were only extracted for ADC maps given its imaging characteristics, were scarce and derived only from Wavelet images.

From these results, we can also look into model sparsity, that has two aspects: sparsity on feature level and sparsity on group or block level. The former refers to overall sparsity, i.e. the total amount of selected features in the resulting models. Instead, the sparsity on group level refers to whether features of only some groups, and not of all groups, are selected. We recall that all model configurations were limited to an overall sparsity of 9 features, and constrictions on group level varied but several possibilities were tested.

Regarding overall sparsity, we see that all 14 pre-biopsy models required a minimum of 2 variables. For preoperative models the minimum number of variables increases to 4 and for postoperative scenario it decreases to 1. Pre-biopsy models were the sparsest, where only 6 models used 8 features, followed by preoperative and postoperative models.

5 DISCUSSION

Prediction of clinical outcomes, such as prostate cancer biochemical recurrence, remains a challenging problem. The current stratification of patients regarding the risk of BCR has limited ability for treatment decision-making, since a homogenous treatment with curative intent, such as the case of radical prostatectomy, can lead to very heterogeneous oncologic outcomes. This observation raises the need of identifying the unseen sub-group of patients who are at higher risk of relapsing, whom would potentially benefit from treatment intensification, or a different clinical management strategy.

Given the significant incidence of cancer recurrence in PCa patients treated with curative surgery, numerous predictive and prognostic tools have been developed throughout the years. Although the inclusion of postoperative clinicohistopathological data has allowed for achieving higher precision in BCR prediction, it has inherent shortcomings: it is bounded to biopsy or to surgery specimen histopathology assessment. Instead, performing BCR prediction in early-stages of the clinical workup, such as pre-biopsy or preoperative, is of greater clinical value.

We investigated the predictive power of preoperative bi-parametric MRI radiomic features combined with clinicohistopathological information in foreseeing the occurrence of biochemical recurrence in PCa patients treated with radical prostatectomy.

mpMRI examination is the standard-of-care imaging for PCa diagnosis, combining anatomical and functional imaging information. These imaging exams are readily available in the early stages of PCa clinical workup, characterising the preoperative condition of the disease. Thus, mpMR imaging information was thought to be valuable for foreseeing treatment outcome such as BCR. This imaging technique is the most sensitive and specific image modality to detect and characterise the clinical aggressiveness of prostate tumours. Despite image acquisition and reporting protocols for PCa being worldwide-standardised with PI-RADS v2.1 guidelines [74], image evaluation is inherently subjective, as it depends on human interpretation. Moreover, mpMRI assessment by radiologists is known not to be 100% sensitive for clinically important lesions, that can be missed or having their size underestimated [32]. Consequently, the final diagnosis and prescribed treatment can be compromised.

Thus, we aimed at objectively quantifying the bpMR images of the whole-prostate gland with Radiomics analysis. The field of Radiomics investigates the macroscopic heterogeneity of quantitative features computed from medical images with the goal to probe the characteristics of the underlying tissue. This field, specially developed in oncology, has shown potential in characterising the heterogeneity of tumour appearance (usually correlated with increased malignancy and, thus, linked with treatment resistance and metastatic propensity [3, 4]).

A major application of the quantitative radiomic information carried by bpMR images, together with clinical variables, is their usage for the development of trainable classifier models with machine learning techniques. Radiomics, in combination with ML, can help achieve a more objective classification of medical images, thus providing a valuable tool to aid clinicians in decision-making at different phases of PCa clinical management. This study developed a systematic and ML-based framework for deriving a trainable classifier in a high-dimensionality scenario, by fine-tuning and optimising embedded regularisation techniques operating on thousands of radiomic features. This framework was used to develop a regularised logistic regression classifier operating on bpMR images of the prostate gland. It was set up in three clinical scenarios; pre-biopsy, preoperative and postoperative, where the linkage with clinicohistopathological information varied according to its respective availability at each clinical stage. We aimed to explore the potential of these combinations for distinguishing subjects that were BCR positive from those who were BCR negative within 2 years of PCa treatment with radical prostatectomy.

5.1 BCR Prediction with LASSO Logistic Regression Models

Our top classifiers across pre-biopsy, pre- and postoperative clinical settings were built on top of 2715 radiomic features and 3, 11, and 33 clinicohistopathological variables, respectively and interpreted via logistic regression using Priority-LASSO regularisation technique, with priority sequence *Clinical, T2, DWI b1000, ADC* and block variable restriction *pmax* = (1,7,0,1). The models performed with reasonably high F-measure (range: 0.702 – 0.754 and 0.910 – 0.925 for the BCR positive and negative classes, respectively) when classifying BCR positive PCa patients in the development set (Table 4.2). However, performance assessed with an independent validation set revealed a great discrepancy between train and test sets results, especially in the BCR positive class-specific metrics. This led us to conclude that training results were overoptimistic, and that— despite the regularisation —our models suffered from overfitting. For validation, we obtained very low F-measure for the target BCR positive class (0.297 – 0.400). Yet, we attained reasonably good F-measure and Precision for BCR negative class (0.779 – 0.833 and 0.821 – 0.873), which, in our opinion, make these models worth exploring.

The Area Under the ROC Curve (AUC) is routinely used to assess the performance of classification models [171], including in radiomics studies [174]. However, AUC weighs classification errors in the two classes being evaluated equally in a cumulative manner, which can lead to misleading results in situations where the dataset is substantially imbalanced. Since this study's cohort was similarly imbalanced, i.e. many more BCR-negative patients as compared BCR-positive ones, the performance of all the classifiers tested in the framework, both candidates and final, was also assessed in terms of class-specific measures, namely recall (sensitivity), precision (predictive value), and F-measure, in addition to AUC. These measures enabled us to specifically evaluate the classifiers' performance on the minority class in this study's cohorts, namely BCR-positive, which can be dominated by the majority class during testing and evaluation.

5.1.1 Regularisation Methods and Model Configurations

We implemented a workflow of model development and analysis suitable for classification problems that also aimed to tackle the curse of dimensionality, common in radiomic studies. In our study, this high-dimensionality problem was aggravated as we used multi-image modality radiomics data, extracting around 1000 features for each image (see Table D.4).

We applied embedded regularisation methods in our models to achieve model sparsity. Sparse models have higher practical utility in clinical settings, being easier to interpret and to communicate. We focused on LASSO-based regularisation of logistic regression models, where three variants (standard LASSO, two-block Priority-LASSO and four-block Priority-LASSO) were compared based on their performance and level of sparsity on three different clinical scenarios. These methods differ in their accessibility to variables. Priority-LASSO makes use of variable groups or block structures, that progressively become available to the classifier according to a priority sequence, for the search of a model solution. On the contrary, LASSO was implemented as a naive classifier, as it does not include variable aggregation criteria.

As we were aware of our limited cohort size, we restricted the maximum number of variables selected in each model, so as to prevent overfitting. For LASSO, this translates into controlling the model overall sparsity. However, given the restrictions of the Priority-LASSO algorithm implementation on R software [168], it was only possible to control model sparsity through block-by-block restriction of total number of variables. This can be viewed as a limitation for our search of the best block-by-block induced sparsity, or as a convenience when constructing models for clinical applications cointaining variables of acknowledged different preferences (e.g. different clinical availability or acquisition costs).

With the model selection workflow, we observed that, overall, models' performances depend not only on the used method but also on sparsity restrictions and clinical scenario.

First, the results affirm that using the naive strategy of treating radiomic and clinical feature groups equally, with LASSO, leads to a worse performance. This became clear throughout the model selection for each clinical scenario, and evident through the comparison of the best models per model type and clinical scenario (see Figure 4.15). Thus, our results are in agreement with the validation of Priority-LASSO algorithm [141], where the method achieves an equivalent or similar performance compared to standard LASSO.

Also, through our search of the impact of block priority sequences with the implementation of all the possible block permutations for both P-LASSO-2-blocks and P-LASSO-4-blocks, we observed that different configurations of a method with respect to the usage of group information affect the potential of using radiomics and other types of data. This diverges from the expectation of P-LASSO's authors [141], i.e. that the chosen priority sequence would have limited impact on the prediction error. Accordingly, "if a block A with strong predictive power is attributed a low priority, its predictive power will nevertheless be exploited in the prediction rule". They add that "the proportion of the variability of the outcome variable that is only explainable by block A will still be unexplained before block A is considered as a covariate block in the iterative procedure". As our study was exploratory, we could not determine, *a priori*, which block should be given higher priority, therefore justifying our more computationally intensive approach of testing all possible block permutations deriving from a primary approach for data segregation. This way, we avoided suboptimal models that may have resulted in cases in which the priority sequence does not attribute high priority to blocks with high predictive power.

In the third place, we observed that the usage of variable group information combined with prioritisation of clinical variables led, overall, to better model performances. In fact, during model selection for P-LASSO-2-blocks, we discarded completely the priority sequence *Radiomics* > *Clinical*, as the alternative achieved consistently better performance. Whereas for P-LASSO-

4-blocks, 6 out of 24 block priority sequences with Clinical variables having higher priority were consistently better in all clinical scenarios (as portrayed in Figures 4.6, 4.10 and 4.13). Indeed, according to selected top models' performance block priority sequences and respective selected features (Table 4.3), clinical variables were preferential to primarily segregate the two classes.

Finally, as we also aimed to understand the impact of the level of categorisation of variables on Priority-LASSO algorithm, through the implementation of P-LASSO-2- and -4-blocks, we found that higher segregation of radiomics data (in this case, according to their image modality with P-LASSO-4-blocks) provided better results than non-distinguishing radiomic features (P-LASSO-2-blocks). This emerged during model selection procedure and became evident when comparing these two algorithms' performances side by side — with clinical features being favoured, see Figure 4.15. Thus, according to these findings, further distinguishing the radiomics data into image modalities led to an improved exploitation of these variables for this study's problem.

A possible explanation for the poorer performance coming from LASSO models could be derived from the fact of incorporating low- (clinical) and high-dimensional (radiomics) features together. Thus, it may be possible that the few clinical features *got lost* within the huge amount of radiomic features and, therefore, their potential was not exploited as if they were considered by themselves. P-LASSO-4-blocks results support this, as forcing clinical variables to be considered firstly and separately, and only then radiomic features, led to higher performances in this study.

Also, the priority sequence *Clinical, T2, DWI, ADC* was the top priority sequence for all clinical scenarios, as depicted in Figures 4.3, 4.6, 4.10 and 4.13 (sequence 19). Clinical variables, even if not collected for prognostic purposes, are extensively and continuously proved to be of value for urologic practice, as they allow for portrayal of the disease state. Then, T2w MRI was the imaging modality with higher relevance, which might be due to the fact that this is the modality with highest spatial (resampled) resolution in this study: almost 3 times higher in-plane- and 1.5 higher out-of-plane resolutions than the remaining modalities (see Table D.3). Thus, the level of anatomical detail seems to be the most important feature. Finally, DWI and ADC maps provide functional information, but with lower level of spatial information. Particularly, the top performing *pmax* block sparsity configurations null out the DWI information, and value ADC variables instead. This could be because ADC maps combine (in this study) two DWI sequences, being richer than just one, i.e. high-b value DWI. In fact, this is directly compared through the performances of priority sequences 19 and 20 (i.e. *Clinical, T2, DWI, ADC* and *Clinical, T2, ADC, DWI*, respectively), where we observed that sequence 19 consistently performed better than sequence 20 (Figures 4.3, 4.6, 4.10 and 4.13).

5.1.2 Radiomics Model Performance and Feature Selection Instabilities

Radiomics analysis used in the context of medical imaging generates a huge volume of data to make possible the analysis of complex systems and diseases. Yet, usually a radiomic dataset is composed of features that can reach several thousands, whereas the sample size is usually less than one hundred.

Finding a good radiomics-based model in a high-dimensional setting can be challenging. Radiomics data is characterised by its heterogeneity, as confusing factors can interfere with the information of interest. Also, radiomic features are inherently susceptible to several variations, ranging from usage of different scanners, image acquisition protocols, segmentation and feature extraction procedures [175, 176]. Thus, a notorious difficulty in large-scale data analysis such as in radiomics is the handling of these confounding factors, which may induce bias, cause unreliable feature selection and high error rates.

In this already challenging scenario, not only is it important that a model has high performance and uses few features, but also that selection of such features is stable. That is, the sets of chosen features should be similar for similar datasets, as an unstable feature selection questions the reliability of the results.

Thus, it is crucial to observe that this study did not allow for the development of stable models, seen through the inconsistency on performance metrics' values and selected features along the 14 CV runs. Yet, this variability was expected, as this was an ill-posed classification problem (with many features and few samples).

Through repeated 4-fold cross-validation procedure, we encompassed the two common sources of variance when creating a model: the noise in the training data, and the randomness harnessed during the ML procedure: data partitioning, k-fold CV and initialisation.

When repeating 14 times a 4-fold cross-validation, (14 × 4) 56 different held-out sets were used to estimate the model efficacy. With one single seed, we fitted a total of 155520 models for this ML-framework (2 × (datasets with different normalisation methodologies) × 2 (oversampling / non-oversampling) × 3 (types of models: Lasso, Priority-Lasso-2-blocks, Priority-Lasso-4-blocks) × 5 (configurations of Lasso) × 2 (PL-I priority sequences) × 6 (configurations of PL-I) × 24 (PL-II priority sequences) × 9 (configurations of PL-II) = 155520 models). Thus, with fourteen seeds, we fitted 2177280 models in total.

Yet, increasing the CV runs to 25 would have fit 100 models for each configuration, allowing for fairer estimations of the performance metrics' distributions.

LASSO makes strong assumptions on the data properties. Particularly, it assumes that correlations between variables are weak [177]. Consequently, a major and still open question concerns the application of the procedure while coping with large correlations between variables. OSCAR [178] is an algorithm that was developed as a way to tackle this problem and should be further explored in future studies. Given the nature of the biparametric imaging data used in this study, we believe that our dataset could have many correlated variables.

In fact, correlation filters are widely used in Radiomics [179]. However, correlation filter methods for feature selection lack clear criteria for removal or selection of variables and, thus, using them can be seen as implemented authority into the process of modelling, ideally supposed to be data-driven. For this reason, we did not use them.

Other options to deal with a number of predictor variables that by far exceeds the number of observations could be using other techniques for dimensionality reduction: principal component analysis (PCA) [180], UMAP [181] or t-SNE [182]. These techniques could be combined with equivariant strategies such as Mapper [183], a topological data analysis algorithm that extracts global features from high-dimensional data, representing it in a compact and global form.

5.1.3 Oversampling of Minority Class

Even though logistic regression algorithms have shown higher robustness when dealing with imbalanced datasets [184], a sampling method, SMOTE, was applied to generate a class-balanced dataset for training purposes. SMOTE has been used as a data-based approach for imbalanced data in a number of prostate radiomics studies [82, 116, 117].

In this study, SMOTE produced new minority class samples by interpolating between the 15 existing training minority samples, leading to an optimally 50/50 balanced development dataset of 51 BCR positive and 51 negative cases. We aimed to explore how the class balance with SMOTE could impact the ability of the logistic regression classifier to correctly detect minority cases in the test set. As depicted in Figure 4.16, SMOTE improved the performance during training but worsened it during testing, contrary to our expectations. Therefore, the selected models of this study did not benefit from SMOTE-based augmentation of the training data.

A possible explanation for this phenomenon can be due to the fact that SMOTE was set to take into consideration $k = 5$ neighboring examples from the minority class while generating synthetic instances. This parameter might be too high, possibly giving rise to the problem of overgeneralisation that is known to be associated with SMOTE [185]. This issue is caused when the new synthetic examples are generated in overlapping areas, thus increasing the overlapping of classes. There is also the possibility of SMOTE augmenting noisy regions within the BCR positive class. Another potential reason to justify the low performance of the model when classifying the 5 BCR cases belonging to the test set could be that the 20 BCR positive cases belonging to training data were not representative of the (heterogeneous) BCR positive population.

Even with the aforementioned susceptibilities of SMOTE, this method's ability to generate larger decision boundaries is still considered a major strength [185]. In order to avoid and explore the problems mentioned above, while keeping in mind the lack of flexibility of SMOTE, we should run this method for different combinations of k and other parameters. Alternative methodologies for data imbalance corrections include usage of undersampling of the majority class. However, we did not address the class imbalance with this type of technique as it would result in a great loss of information portrayed in our already limited cohort size.

5.1.4 On Feature Selection Results

5.1.4.1 Clinical Features

In the study, the top learners favoring clinical features (P-LASSO-2-blocks and P-LASSO-4-blocks) were, overall, the best performing of all methods under investigation. Whether or not the image modality radiomic feature types are distinguished, favoring clinical features led to better prediction performances.

Baseline PSA, in combination with T2w and ADC variables, was selected for our pre-biopsy and preoperative models, which is in agreement with the current literature and other BCR prognostic studies [61, 126, 186–188]. Against our expectations, variables derived from biopsy histopathologic assessment, such as Gleason Score were not selected by any preoperative model. Yet, in fact, as shown in Table 4.1 that displays patients' clinicohistopathological characteristics for BCR positive and negative cases, we found that Biopsy Grade Group, that combines Gleason score assessment, did not statistically differ for BCR negative and positive cases (p = 0.123, Fisher's Exact Test).

Out of 14, 13 postoperative models chose surgical resection margin extension (SRME) as the most valuable (postoperative) clinical variable. In fact, patients with positive surgical margin (PSM), i.e. "tumor that extends to the surface of the prostate wherein the surgeon has cut across the tissue plane [189]" have increased risk of BCR [50, 186, 190–192]. However, PSM presence does not dictate the development of BCR. Instead, the histopathologic characteristics of the

PSM may influence the risk of BCR. Multiple investigators have sought to characterise PSM, including its length, with the rationale that a greater amount of PSM is associated with greater quantity of tumor that remained after surgery, and a greater potential for growth, biochemical recurrence, and metastases [190].

However, one cannot attribute the incomplete cancer excision to the technique of the surgeon or even the surgery type, as the incidence of PSM depends also on the characteristics of the cancer, such as its aggressiveness and location [190, 191]. Moreover, attaining a negative surgical margin at the time of RP is the primary goal of the surgeon, but it is not an isolated goal. Preserving the neurovascular tissue and maintaining maximal urethral length are crucial for maintaining erectile functionality and continence. Balancing oncologic and functional goals, which are at odds with one another, is fundamental to successfully perform RP regardless of surgical approach.

Yet, as surgical resection with microscopically negative margins remains the main curative option for prostatic cancer, some clinicians view PSM as a trigger for adjuvant therapy. Thus, it would be interesting to use PSM or SRME as a covariate, i.e. a possible predictive or explanatory variable of the dependent variable, BCR occurrence.

Generally speaking, favoring clinical features over radiomic features could be preferable for several reasons. One reason is that clinical variables have often proved to be of value, even if not for BCR prediction, but for overall PCa diagnosis and assessment, as practitioners include them in their diagnosis/therapy/follow-up routine. Moreover, a model including variables that are already considered in routine diagnostics, or variables that can be easily assessed (e.g. age, common clinical variables), are more likely to be accepted by physicians than a model including variables measured with new and/or expensive technologies, maybe even at the expense of a slightly lower prediction accuracy.

However, favouring clinical features would not be enough, as the patients' clinical characteristics in Table 4.1 portrayed the difficulty in discriminating patients solely based on clinical information. Thus, in fact, it is not surprising that clinical nomograms are not exceptional at this task, even if they do have much greater statistical power/sample size.

5.1.4.2 Radiomic Features

The great majority of selected radiomic variables were texture-based (96–97%), followed by shape-based (1.3–3%) and first-order ($< 1\%$) (available for ADC only) features.

Top performing models selected texture features derived only from Laplacian of Gaussian (LoG) or Wavelet filtered images. This is in agreement with reports affirming textural analysis value for PCa disease characterisation [120, 121, 176, 193–195], that were followed when personalising this study's radiomic feature extraction. However, Gradient and Local Binary Pattern images were not selected, contrary to other successful PCa detection studies [174, 196], where the focus was instead on cancer detection through object recognition.

The usefulness of textural analysis of tissue micro-architecture in prostate cancer aggressiveness assessment and classification has been reported in previous studies [112, 117, 194, 195, 197, 198]. PCa tumour agressiveness is associated with BCR, thus providing insight on recurrent disease. Histologically, aggressive prostate cancers are characterised by poor differentiation, glandular structure deformation, and loss of cellular integrity of the prostate gland. This disrupts the tissue cell-architectural patterns, potentially leading to decreased homogeneity and high disorder. Particularly, the association between benign and malign tumour histopathology and texture features of multi-parametric MRI of the prostate has been studied [199].

Association between textural features and clinical outcome have also been reported, but instead for PCa treated with radiotherapy. Studies have shown strong association between T2w texture features [120, 121, 123], in detriment of ADC, shape-based and first-order features [120], as it was the case in this study.

92–96% of selected texture features were based on T2w and the remaining on ADC. In fact, the T2w modality offers the highest resolution, and is therefore richest in texture information. Moreover, the low resolution of the ADC and DW images compared to the T2w images may have impacted the textural information contained in the former.

Besides, compared to DW and DCE imaging, T2w imaging is generally regarded as the most stable sequence in terms of scanner variations and gradient artifacts, a factor specially important in this study setup with heterogeneous image database.

5.1.5 Whole-prostate Approach

We can say that this exploratory study's overall findings not only support the idea that relevant information can be found on a whole-prostate level but also that this region can have potential for BCR prediction. Being multifocality a characteristic of PCa disease, the evaluation of whole-prostate imaging features may be capable of conveying prominent characteristics of the organ micro- and macro-environment, avoiding single-lesion delineation uncertainty, curation and restriction of the analysis to "humanly"-visible tumoural area.

This organ-based approach for prognostic studies is supported by previous findings describing an association between prostate micro-environment and disease relapse or progression [200, 201], whole-gland shape differences between patients who do and do not undergo BCR [97], and usage of a similar successful approach on high-BCR risk patients treated with radiotherapy [123].

However, as studies have observed that recurrence occurs mainly at the site of the primary largest and/or highest-grade index lesion[126], it is reasonable to assume that the most crucial information would lie within the pre-treatment visible single tumour region. In this case, the usage of features from the whole prostatic area might be thought to be diluting the information provided by single tumour-derived features. This could explain the higher performance obtained in BCR predictions studies using tumour-derived information for BCR prediction [120, 126, 127]. With the same train of thought, recent studies are focusing instead on prostate region-based analysis [202–204], given that tissue appearance and PCa predisposition varies with prostate zones.

5.1.6 Comparison with Clinical BCR Predictors

On Table 5.1, the most relevant clinical nomograms that were developed and externally validated for BCR prediction are compared with the clinicohistopathological-radiomics-based models developed in this study (denominated us ProBCR models). Here, models are separated by their clinical time-point applicability: pre-biopsy, preoperative or postoperative. We can compare their discrimination power according to AUC (achieved in training and testing phases) and reported externally-validated AUC of the reference clinical nomograms. We also show the characteristics of these studies' cohorts, in terms of size and BCR prevalence.

Table 5.1: Comparison of pre-biopsy, preoperative and postoperative referential clinical models that predict BCR in men treated with RP with this study's ProBCR models, made in terms of their discrimination, number of patients and BCR prevalence. No pre-biopsy clinical tools available for BCR prediction were found for comparison, and radiomic features may not be of value for BCR prediction at preoperative and postoperative stages. We found a substantial difference in cohort size between our study and others. For our study, AUC was assessed in train and test phases, whereas for the remaining studies, AUC refers to external validations. AUC = area under the ROC curve.

Reference	AUC	No. of patients	No. of BCR patients (%)
Pre-biopsy prediction of BCR			
NA	-	-	-
ProBCR model	0.883 / 0.739*	93	20 (21.5)
Preoperative prediction of BCR			
Stephenson et al. [70]	0.79**	1978	220 (11.1)
Cooperberg et al. [43]	0.66**	1439	210 (14.6)
ProBCR model	0.882 / 0.638*	93	20 (21.5)
Postoperative prediction of BCR			
Walz et al. [50]	0.82**	2875	494 (17.2)
ProBCR model	0.895 / 0.691*	93	20 (21.5)

*: Train/Test ; **: External Validation

First, there is a substantial difference in cohort size between our study and others, where ours was ten to twenty times smaller than the reference models, although with higher BCR prevalence. Referential clinical nomograms might possibly be underrepresenting the incidence of BCR stage in PCa patients treated with RP. Nevertheless, all study cohorts were highly unbalanced in terms of BCR positive and negatives cases.

5.1.6.1 Pre-Biopsy Prediction of BCR

To our knowledge, no clinical tools are available for prediction of BCR occurrence before biopsy. This absence can be due to the fact that, at pre-biopsy time-point, the only commonly assessed variables are baseline PSA and clinical stage, being possibly insufficient for patient discrimination. Our models did not include clinical stage, as it was unavailable, but instead baseline PSA and radiomics features from T2, DWI b1000 and ADC maps, achieving AUC of 0.833 and 0.739 in development and validation phases, respectively.

5.1.6.2 Preoperative Prediction of BCR

Preoperative prediction of BCR can be accomplished with the Stephenson et al. nomogram (n = 1978; discrimination: 79.0%) [70] or as modeled by Cooperberg et al. (n = 1439; discrimination: 66.0%) [43]. Both models rely on commonly available variables, such as PSA, clinical stage, and biopsy Gleason sum, and have been externally validated, thus, their usage is encouraged in the literature.

It is noteworthy that Preoperative ProBCR model showed better discrimination abilities than the Cooperberg et al. and Stephenson et al. risk stratification schemes during training (AUC 0.882 vs. 0.66 and 0.79, respectively). Also, it showed similar power to Cooperberg's nomograms during validation (AUC 0.638 vs. 0.66).

Despite this, Stephenson's model might be superior to our radiomics-based model and, in this clinical setting, radiomic feature extraction and analysis may not be of added value.

5.1.6.3 Postoperative Prediction of BCR

Prediction of BCR after RP represented the focus of several previously reported prognostic tools, yet we picked one that was specific for early BCR prediction and extensively externally validated for comparison. Walz et al. has devised a highly accurate tool for prediction of BCR 2-yr after RP (n = 2875; discrimination: 82%) [50], that relies on serum PSA, pathologic Gleason sum, surgical margin status, ECE, seminal vesicle invasion, and LNI.

Our best postoperative ProBCR model only exceeded Waltz's nomogram performance during training phase (median AUC of 0.895 vs. 0.82). Yet, it fell far behind during validation, with AUC of 0.691. Still, it performed better than our preoperative model during validation, possibly indicating that the model could extract discriminatory information from postoperative variables.

Overall, pre- and postoperative ProBCR models do not meet the discriminatory power of the available clinical nomograms, indicating that radiomics data did not add value for BCR prediction.

We could conclude that our pre- and postoperative models' performances were below the current clinical nomograms discriminatory power revealed on independent or external validation sets.

However, we can argue that the pre- and postoperative clinical nomograms of reference do not seem highly valuable. As we can observe on Table 5.1, clinical nomograms were evaluated with cohorts with very low BCR incidence that ranged from 11.1% to 17.2%. Thus, this implies that one would expect the minimum values of AUC ranging from 88.9% to 82.8%, respectively, corresponding to the BCR-negative class representation percentage. As the AUCs achieved by the clinical nomograms were lower, it means that clinicians would achieve higher accuracy in their pre- or postoperative BCR predictions simply by naively considering that a patient would not suffer from BCR than when using these nomograms. Thus, further studies are still needed to determine whether these pre- or postoperative clinical nomograms should be used for clinical decision-making.

Yet, to our knowledge, there are no pre-biopsy clinically-used or proposed nomograms. Thus, our pre-biopsy model that incorporates baseline PSA level and MRI radiomic features for assessing 2-yr BCR occurrence is unique. This model achieved median AUC 0.739, BCR positive-specific measures with values around 0.4 – 0.45, but with BCR negative metrics ranging from 0.799 to 0.873 (see Table 4.2).

5.2 Patients Clinicohistopathological Characteristics

Different candidate clinicohistopathologic parameters were evaluated in an attempt to predict biochemical recurrence following radical prostatectomy, Gleason score being considered one of the most powerful predictive factors, used indirectly through ISUP prognostic grade grouping system, that has been validated as predictive of BCR, response to therapy, and cancer-related mortality in several large-scale studies [26, 27].

Our study also included the analysis of other known prognostic parameters in radical prostatectomy-treated patients. Variables that revealed to be significantly different between BCR groups, as shown on Table 4.1, are initial PSA level (continuous and stratified by clinically-used thresholds), type of radical prostatectomy (with/without lymphadenectomy), prognostic grade group of both overall prostate and of index tumour, histological subtype (acinar versus mixed), surgical resection margin status and its linear extension, extraprostatic linear extension, perineural and lymphovascular invasion status (positive/negative), the adapted EAU-ESUR Risk Group and, finally, the Risk stratification by overall prostate group grade. These results are in line with the current knowledge, and they most probably reflect an adequate therapy selection based on guidelines and validated nomograms.

Our observation that the type of radical prostatectomy is an important predictor of BCR is in agreement with the expected results since radical prostatectomy with extended lymphadenectomy is indicated in patients with more aggressive clinicopathologic features. Also, it is worth noting that there was not a statistically significant association between surgery type (robotic or non-robotic) and BCR occurrence, confirming what it is stated in the most recent EAU guidelines: ”no surgical approach (open-, laparoscopic- or robotic radical prostatectomy) has clearly shown superiority in terms of functional or oncological results” [17].

It is worth noting that all BCR positive cases had perineural invasion, as determined by histopathologic assessment of the postoperative specimen, and none belonged to the Low-risk categories of (adapted) EAU-ESUR Group Risk or Risk stratification by overall prostate group grade. Despite these significant differences found between clinicohistopathologic features of recurrent and non-recurrent patients of this study’s cohort, we highlight that none derived from biopsy assessment, but from whole-prostate specimen histopathology evaluation. This result emphasises the difficulty in discriminating patients solely based on clinicohistopathological information available before the treatment.

Tumour Index Contemporarily defined index tumor nodule (larger prostate cancer nodule in multifocal disease), which frequently harbors the highest Gleason score, has been shown to be predictive of BCR. It is, therefore, considered an important prognostic parameter in prostate cancer after radical prostatectomy [125, 126, 205]. In fact, recent studies suggest that index tumor nodule can be identified in about 90% of radical prostatectomy samples and may be identified in 90% patients using mpMRI and targeted biopsies [125, 205–207].

Hypothetically, this observed good correlation could give support to emerging clinically oriented proposals of using mpMRI to summarise the aggressive features based on index tumor nodule features. This could lower unnecessary surgical procedures or provide a rational for focal therapy applications within the frame of grade group assessed in index tumor. Along this line, Radtke et al. was able to identify over 90% of index tumor nodules in a series of radical prostatectomies aiming toward focal therapy and concluded that mpMRI could identify 92% of index lesions [128]. Kasivisvanatha et al. was able to identify higher-grade tumors as compared to standard biopsy, a fact that might be related to a higher detection of index tumor lesions because of MRI guiding targets [207]. In support of this is our finding of higher-grade group categories in index tumor nodule as compared with all prostate (see Table 4.1).

5.3 Image Normalisation

This study's main experimental design was preceded with a primary analysis of the effect of image intensity normalisation methodology on models' performances, not only due to the fact that (i) MR image intensity is usually relative and not directly comparable between images, but also because (ii) estimation of radiomic features requires prior intensity normalisation, that can be be performed using different methods. Also, (iii) normalisation can make computation more efficient, as it forces the pixel intensity distribution to be normal and belonging to the interval $[0, 1]$.

We found in all of the best performing model configurations that normalising the whole-prostate gland with respect to that same organ (”prostate-only” method) performed significantly better than when considering the organ and its immediate surrounding tissues contained in the limiting bounding-box (see Section 4.2). Performance was measured in terms of *Fmax*, as it was the metric of interest for the subsequent comparisons.

We had the expectation that normalisation could influence the computation of certain radiomic features, mainly the textural ones, as the normalisation output influences the posteriorly computed intensity bin size (as seen in Table D.3) used to quantise image intensities into a discrete number of grey levels. However, it was not expected that one particular normalisation strategy (prostate-only) would lead to a consistent and significantly better performance on the top configurations of the three studied computer-based algorithms. In fact, the calculation of bin-width using prostate-based normalised images led to a smaller value than relying on bounding-box normalised images (Table D.3). Thus, the former technique led to a greater textural contrast on the whole-prostate region (the VOI used for radiomic features extraction), which apparently led to greater model performance. Moreover, normalisation with ”bounding-box” made the procedure dependent on the location and size of the volume of interest, which possibly introduced unwanted noise.

The need for this comparison has risen from the fact that intensity normalisation methods for images of the prostate region and their evaluation were lacking in the literature, contrary to the many methods developed and proven successful for the analysis of brain pathologies. In medical imaging processing, it is a common practice to perform intensity normalisation based on a comparable reference tissue region, which is assumed to be stable across time points and patients, such as fat or muscle region. In fact, the utility of normalisation to the muscle reference region has been demonstrated in prostate DCE imaging studies (e.g. see Huang et al. [208]). However, as it will be further discussed, this study's cohort had a heterogeneous image dataset, thus we did not expect to find a commonly homogenous region for all images, especially with the variety of FOV sizes.

Hence, our normalisation methods' results held true for this specific problem and for the comparison of the best performing model configurations regarding maximum number of variables (*pmax*) or block priority sequences of the regularised logistic regression models.

5.4 Limitations

5.4.1 Cohort Size

The design of our single-centre, retrospective study has some limitations. Due to our stringent approach to acquire consecutive controlled data, applying clinical history, treatment, follow-up and imaging criteria (see study population flow chart in Appendix C), our sample size was reduced from 250 consecutively collected RP-treated PCa patients to a quite moderate size of 93 cases. Although comparable to similar exploratory studies related with this topic [83, 119, 120, 123, 194, 198, 209–211], only a relatively low number of BCR positive cases were eligible (n = 20). Small sample size are known to increase both type-I (incorrectly detecting a difference) and type-II (not detecting an actual difference) error rates, therefore, our results should be interpreted with the necessary caution.

The combination of small cohort size with high number of features not only contributes to the model instability, as observed and discussed before, but even to model development failure. Abundant healthcare data is not readily available for machine learning supervised tasks, being an enormous obstacle for the development of reliable models in clinical settings.

5.4.2 Model Overfitting

Another limitation of the developed models was the fact that the models exhibited overfitting. Overfitting is a modeling error that occurs when the learning function is too closely fit to a limited set of data points, being the model overly complex model to explain idiosyncrasies in the data under study. Yet, the data often has some degree of error or random noise within it. Thus, attempting to make the model conform too closely to slightly inaccurate data can infect the model with substantial errors and reduce its predictive power.

To avoid overfitting and discover the relevant features, feature selection was integrated into the model fitting process (embedding). The premise is that reduction in the dimensionality of the space of the explanatory variables leads to smaller and simpler models and thus tries to avoid overfitting, which potentially leads to a better performance.

Moreover, we implemented two other techniques that are known to lessen the chance of, or amount of, overfitting. Regularisation with LASSO was performed to penalise overly complex models, and cross-validation was used for parameter hypertuning and to test the model's ability to generalise by evaluating its performance on a set of data not used for training (assumed to approximate the typical unseen data that a model will encounter).

We restricted the maximum number of selected variables during regularisation with LASSO and Priority-LASSO, as an adequate number of events per variable (EPV) is required to generate accurate estimates. Ten events per variable (EPV) is a widely advocated minimal criterion for sample size considerations in logistic regression analysis. Thus, as 93 patients were involved in this study, we applied the ten EPV, enabling a maximum of 9 features to be selected for the final model.

However, model development was based on the training data comprised of 66 cases, therefore the threshold for the number of variables possibly should have been lower or equal to 6. Yet, the ceiling number of 9 was a mere indication during model fitting, as in fact only a fraction of the total developed final models used 8 or 9 variables (see Table 4.3). Moreover, some studies state that the ten EPV criterion is not strict and may depend on other data conditions [212, 213].

Still, the general opinion is that low EPV may lead to major problems such as bias and low model precision [214].

Also, it is possible that the 2$^{\text{nd}}$ block configuration in Priority-LASSO with $pmax = (1,7,0,1)$ was too unrestricted for the total number of allowed T2w variables, causing overfitting of the model regarding this image modality's characteristics.

We thereby conclude that the combination of these implemented strategies was not enough to prevent model overfitting. We also want to emphasise that, according to machine learning best practices, the model tuning and choice of final model configuration is based only on the training data. Therefore, the test set is only used at the end. Although our workflow was performed in a systematic way, with the referred strategies, using statistical tests to help with the necessary choices to select the final model, the used software did not allow us to access cross-validation intermediate results and, therefore, recognise model overfitting.

Contrary to R software, *scikit-learn* [215], a Python-based library, allows the user to have access to cross-computed metrics of cross-validation training and testing folds. With such information, we could have evaluated training and testing metrics' closeness or discrepancy, and thus determined if the model was tending to overfitting, without relying on the final hold-out set to evaluate so. However, Priority-LASSO algorithm was not implemented on *scikit-learn* or any other Python library, only in R.

Both LASSO and Priority-LASSO perform tuning of logistic regression parameters by cross-validation through the optimisation of a specified metric. A limitation of both methods' implementations on R is that parameter tuning for binary problems is limited to the optimisation of model accuracy or AUC. Such metrics were not ideal for characterising the performance of a model developed on imbalanced data, nor to optimise the classification of minority classes, as we intended. Instead, *scikit-learn* library allows to specify and even costumise optimal tuning search in regard to any metric, such as F-score.

5.4.3 Data Imbalance

Another potential challenge of our study was that it was based on a relatively imbalanced cohort, although being this also the case in prior similar studies [87, 123, 216, 217]. We addressed the imbalance problem by using the SMOTE method of randomly oversampling of the minority class [145] (BCR positive) during classifier training, which yielded relevant performance improvements during model development but worsened it during validation (Table 4.2).

However, the difficulty in separating the small class from the prevalent class is the key issue of the small class problem. If highly discriminative patterns exist among each class, then not very sophisticated rules are required to distinguish class objects. However, if patterns among each class are overlapping at different levels in some feature space, discriminative rules are hard to induce. In fact, the class imbalance distribution, by itself, may not be problematic, but when combined with highly overlapped classes, it can significantly decrease the number of minority (small) class examples correctly classified. Linearly separable domains are not sensitive to any amount of imbalance. As a matter of fact, as the degree of concept complexity increases, so does the system's sensitivity to imbalance.

Also, in many classification problems, a single class can be composed of various subclasses, or subconcepts. Usually, samples of a class are collected from different subconcepts. These subclasses or subconcepts do not always contain the same number of examples. This phenomenon is referred to as within-class imbalance, corresponding to the imbalanced class distribution among subclasses. The presence of within-class subconcepts worsens the imbalance distribution problem (no matter between or within class) in two aspects: (1) the presence of within-class subconcepts increases the learning concept complexity on the data set; and (2) the presence of within-class subconcepts is implicit in most cases.

We conclude that the usage of stratified cross-validation setup and SMOTE-based augmentation of the training minority data was not enough to prevent model difficulties in correctly classifying the minority class of BCR positive.

5.4.4 BCR Definition: Follow-up Time Cut-off and Threshold Level

We found heterogeneity in some clinicohistopathological variables characterising BCR positive patients, and even no statistical difference between 2-year BCR negative and positive cases in many of commonly assessed variables. This depicts the clinical homogeneity of the cohort in the majority of evaluated aspects, thus highlighting the complexity of this problem.

However, one can speculate that, despite the low number of minority cases in this study, that surely complicates the classification task, there could be a distribution overlap on the characteristics of non-BCR and BCR patients deriving from the premises used to define BCR, i.e. (1) the time-period of 2 years to assess PSA values and (2) the PSA threshold of 0.2 ng/mL.

It is important to recall that the impact of BCR on oncological outcomes of men treated for PCa with curative intent remains controversial, as there is a huge variety seen in studies investigating the natural history of PSA rising or persistence. As mentioned on Section 1.2.5.1, only a subgroup of BCR patients will proceed to local recurrence or metastatic progression, while others will have an indolent disease course. For instance, Rogers et al. assessed the clinical outcome of 160 men with a persistently detectable PSA level after RP [218] and found that 38% of patients had no evidence of metastases for > 7 years while 32% of the patients were reported to develop metastases within 3 years. A study by Vencloves et al. [219] observed that at the end of their 7-year study, 47.8% harboured BCR, yet with the majority (61%) occurring within the first year after RP. In their relatively long study with 207 BCR patients, it was concluded that the most informative time until BCR cut-off ($\leq$ 1, 1–2, 2–3, 3–4, and 4–5 years) for prediction of clinical progression and cancer-related death was one year following RP.

However, Pound et al. [54] reported a 5-yr metastatic progression (MP)-free survival rate of 64% among 304 RP BCR patients who were observed until metastatic progression. In their case, the time from RP to BCR (2 vs.$>$ 2 yr) was an independent predictor of MP in their analysis.

These study examples, reaching different conclusions, convey both the disease stage heterogeneity as well as the necessity to investigate BCR patients throughout different time-lengths. However, a great effort was made to guarantee a follow-up of minimum of 24 months, as it was the longest period possible to be implemented given the characteristics of data trackability of the institution. We believed it would be a reasonable time-period to capture BCR occurrence, as approximately two-thirds of BCRs occur within the first 2 years of surgery, and earlier BCR may be associated with increased risk of prostate cancer-specific mortality [54, 67]. Yet, the remaining one-third of BCRs might occur after the first 2 years of surgery, and if our study design had a long-term follow-up period such as 5 years, our results might have varied. Adding to this depiction of the complexity of this disease stage is the non-existence of definite consensus regarding the PSA cutoff point for defining BCR recurrence after RP. As stated in the most current EAU guidelines, there are several definitions of PSA recurrence after RP being investigated in the last years, tending towards threshold values between 0.2 and 0.4 ng/ml; yet the usage of 0.2 ng/ml in conventional assays seems to be the most acceptable threshold for PSA recurrence based on a clinical point of view [17] – in fact, the one used in this study.

These observations lead us to state that indication for further treatments should not be based on meeting a threshold PSA recurrence as defined here and in other studies, but perhaps should depend on the individualised risk of progression. Thus, additional stratification of patients with BCR is crucial to ensure timely commencement (generally before meeting the BCR threshold) or deferral of salvage treatment.

Nonetheless these are encouraging findings as they provide pilot evidence of the relevance of imaging in outcome stratification of clinically homogeneous patients. The use of whole-prostate imaging characteristics to obtain information about two year biochemical recurrence risk can potentially be used to develop individualised treatment strategies.

5.4.5 Image Data Heterogeneity

Our cohort comprised patients examined within 2014 and 2017, a period in which the MRI protocol underwent slight changes given the update on PI-RADS guidelines in 2016 [73]. Thus, this study's images were originated in three different scanners from the same vendor, yet with different field strengths, sequence parameters, image resolution and FOVs. However, given that these scanners were routinely used in clinical practice at the institution, with image acquisitions and reconstructions tailored to give out the same visual information to radiologists (precluding the need for any visual adaption), and adding to the fact that acquisition of images is time-consuming and costly, our approach in this study was to explore standard-of-care images acquired in the institution for PCa clinical diagnosis, regardless of the imminent heterogeneous data to be found. Thus, we did not apply additional imaging criteria so as to not severely diminish our dataset, instead expecting that a higher number of instances could be able to overcome some of the heterogeneities inherent in PCa clinical imaging. And, if so, this could generate more clinical impact compared to more controlled and dedicated prospective image acquisitions.

MR image appearance, quality, and the presence of artifacts can be affected by different scanners, and its influence on the extracted radiomic features is still under-investigated, but it is well-recognised that radiomic features need to be reliable, i.e. reproducible and repeatable across different imaging and image-processing protocols, as well as scanners.

As image variety affects the information being extracted by image feature algorithms, which in turn can influence the performance of computerised algorithms [110], it raises challenges for effective ways of aggregating heterogeneous data. Particularly, radiomics analysis of MRI data poses additional significant difficulties due to the inherent lack of signal standardisation of this modality and consequent scarcity of effective normalisation techniques to facilitate image comparisons.

Without the presence of a large standardised repository (as is the case of our study), setting performance benchmarks for effectiveness of image feature algorithms and classifier models built upon those features becomes difficult [130].

However, we knew that the differences that were entailed in this study's cohort could possibly be addressed in an attempt to keep features as comparable as possible across patients. Thus, following Image Biomarker Standardisation Initiative recommendations for radiomic biomarker extraction [161], we applied image resampling to the lowest resolution encountered for each image modality, as well as VOI normalisation and one costumised image quantisation method, aiming to leverage the variety of the image data characteristics in this study cohort. The image resolution downsampling is an inevitable but huge limitation of our study, as we had to downsample high-resolution MRI exams to make them comparable to old ones with much worse in- and out-of-plane resolutions, causing a great information loss.

Moreover, we aimed at introducing the most recent PI-RADS recommendations to the processing and analysis of DW imaging and ADC maps [73]. As not all patients had ADC maps calculated from the scanner software, nor all DWI series when acquired with the same b-values (given the changes in image acquisition throughout time), we chose to use DWI series with the b-values that were more common in the study cohort, so as to avoid further reduction of the dataset. The most common b-values were 0 and 1000, which was partially in accordance with technical specifications found in PI-RADS v2 [73]. There, it is stated that "in the case that only two b-values can be acquired due to time or scanner constraints, it is preferred that the lowest b-value should be set at 50–100 s/mm^2 and the highest should be 800–1000 s/mm^2". Also, it is specified that "the reason for preferably starting with a b50 instead of a b0 is to prevent shine-through of the vessels, that is, to exclude the vascular signals". For this reason, we did not use DWI b0 series for radiomic extraction, also because it mimics, with poorer resolution, T2w imaging.

This careful selection of DWI series led to the computation of ADC maps that were motion-corrected and derived from a set of images acquired with two values for the parameter *b*. We were aware that more accurate ADC maps could be computed by acquiring a set of images with more than two values for the parameter *b*, however, there was not a third commonly used *b* value in this dataset.

We chose to use DW b1000 image series as well as ADC maps for radiomics feature extractions, to compare their utility, knowing that the latter derives from several DWI images, thus making the ADC maps inherently more susceptible to cumulative errors or artifacts than individual DWI images.

5.4.6 bpMRI vs. mpMRI

For this study, we did not make use Dynamic Contrast Enhancement MR imaging series, that is part of the mpMRI prostate examination implemented in clinical practice. The most recent PI-RADS guidelines state that "when T2-weighted imaging and DWI are of diagnostic quality, DCE MRI plays a minor role in determining the PI-RADS assessment category" [73]. Specifically, DCE does not contribute to the PI-RADS overall assessment when the finding has a low (PI-RADS 1 or 2) or high (PI-RADS 4 or 5) likelihood of clinically significant cancer. However, in the specific case of when DWI results in a PI-RADS score 3 in the peripheral zone, a positive DCE may increase the likelihood that the finding corresponds to a clinically significant cancer and may upgrade the assessment category to PI-RADS 4.

Although DCE is, currently, an essential component of the mpMRI prostate examination, there is the tendency to simplify multiparametric protocols, as the latest development in this field is the so called "biparametric approach" to MRI.

DCE imaging within mpMRI protocols not only brings an incremental cost, but adds an invasive nature to the MRI examination. In fact, patients with renal dysfunction or with an allergy to the contrast agent are not allowed to undergo DCE MRI, so indeed it exists the clinical need to evaluate the utility of bpMRI in detecting prostate cancer, as done by Takeuchi et al. who confirmed the ability of DWI and T2-weighted imaging with bpMRI when assessing the PI-RADS v2 score to contribute to the prediction of BCR after radical prostatectomy [220].

5.4.7 Segmentation

Segmentation is the stage where a significant commitment is made during automated analysis by delineating structures of interest and discriminating them from background tissue. This separation, which is generally effortless and swift for the human visual system, is the single most problematic aspect of conventional radiomics workflows. Since measurements and other processing steps are based on segmented regions, in many cases the segmentation approach dictates the outcome of the entire analysis [221].

Particularly for studies of the prostate, manual segmentation is still the most common way to accurately segment the gland and its regions [222]. For this reason, in this study, the prostate gland was contoured slice-by-slice, by a non-expert (the author). It was not feasible to follow the recommendations of having segmentations curated by one or more experienced radiologists [222], given that it is a large time-consuming task. In fact, the delineation of one whole-prostate done by a specialist radiologist can take from 30 to 45 minutes, as a typical prostate covers 15 or more slices on T2w images acquired with less than 3mm slice thickness. We believe that using non-curated manual segmentation is a considerable limitation of this study, despite the efforts made to do it in a semi- or automatic way (see Appendix D.5).

In any case, whether curated or not, manual delineations always suffer from high interreader variability, thus it has been concluded that this type of segmentation is not feasible for radiomics analysis requiring very large data sets.

In fact, automatic prostate delineation on MRI exams is still an unsolved issue. First, the ambiguity of its boundaries makes it very hard to differentiate the gland from surrounding tissue with intraprostatic tissue heterogeneity further contributing to under- or oversegmentation. Second, examinations on different MR images with use of different imaging protocols lead to wide variations in signal intensity. Third, the prostate gland has a wide range of sizes, shapes and tissue types, either due to physiologic variations among patients or due to the presence of pathology. Therefore, reliable automated segmentation of the prostate gland is highly desirable in daily clinical practice and would facilitate PCa quantitative imaging studies. Alternative methods worth exploring could be U-Net convolutional neural network developed for segmentation of medical images [223], so far applied to zonal prostate segmentation in MR images [203], or capsule neural network-based segmentation [224].

5.5 Future Work

This study can be expanded and improved upon by addressing the previously mentioned limitations. This would include the employment of a larger cohort with longer follow-up period, bpMR exams with homogeneous characteristics and curation of VOIs. Also, exploring other techniques for tackling the high-dimensionality problem, testing classification algorithms beyond those available in R, and ensuring a correct model development and validation to avoid overfitting.

As mentioned, the biochemical relapse prediction assignment is not yet solved. This is reflected in the few methods used to predict this condition, which are yet not as accurate and generalisable as intended. Hence, it is likely that currently proposed approaches are too simple for the BCR prediction assignment, especially in early stages of the disease management.

Continuing work should be performed on making mpMRI assessment more objective and quantitative so as to be able to be explored with radiomics, e.g. through bpMRI protocol standardisation, image normalisation techniques, (automatic) segmentation and model development. Further research is needed to ascertain which imaging sequences, radiomic features, region of interest and algorithms would be optimal; yet, our results indicate that T2w imaging sequences are promising for BCR prediction.

The premise of radiomics is that quantitative image features can serve as a biomarker characterising the disease, allowing for prediction of response and, thus, providing decision support for patient management. Radiomic analysis is expected to excel at characterising features non- or barely-visible to the human observer [160, 225]. Yet, to reliably derive conclusions based on any biomarker, a basic requirement is that its value must remain stable between different measurements, if the conditions remain stable [175, 176]. Thus, considering the repeatability of radiomic features would therefore be a characteristic that could be used for pre-selecting features for a classification task, given the problematic of the large amount of radiomic features available in medical imaging to select from. It would be the first step towards discovering robust biomarkers to move this form of precision medicine forward.

One parallel step to improve the assessment of the real capabilities of radiomics for PCa management would be the recognition of radiologically perceived limitations of mpMRI technique. For this, image findings need to be validated with the gold-standard histopathology of prostatectomy samples.

Several studies have described the association between tumour adjacent stroma and prostate microenvironment to relapse and disease progression [200, 201]. MRI is known to have limited accuracy in the detection of small tumour foci of less than 0.5 cm^3 [226]. Thus, it is currently impossible to rule out the presence of nearby lesions, not visible on MRI and missed by biopsy sampling, in the remaining prostatic region, at the pre-biopsy or preoperative stages. Analysis of the prostate surrounding area with radiomics, separately or not, could provide important information, as recently explored by Fernandes et al. [123].

Also, obesity has been associated with aggressive PCa and higher rates of post treatment disease recurrence [227, 228]. It is still unclear if this is due to specific tumour-promoting effects of obesity or diagnostic bias. Thus, further studies could categorise patients according to their body mass index, to understand how prognostic tools may be impacted by obesity. It is noteworthy that this was not a commonly available parameter for all patients of this study.

The comprehensive collection and processing procedures (i.e. data curation) culminated in the establishment of an annotated database of bpMRI exams with the corresponding clinico-histopathological data. Apart from being used in this study for the development of LASSO logistic regression models with radiomic analysis for BCR prediction, the potential of this database has already started to be explored in this study's institution. Transfer and deep learning architectures for classification of PCa medical images have been used to try to replicate findings found in the literature, such as Gleason score prediction. Also, our created dataset has been used to explore and validate novel survival analysis techniques combining deep learning, imaging and clinical information. A wide range of hypotheses may continue to be explored, and clinical questions regarding prostate cancer clinical management, from diagnosis to prognosis, may be addressed.

6 CONCLUSIONS

Prostate cancer is the second-most common cancer amongst men, where 1 out of 8 will develop this disease. Even though it can often be treated successfully, prostate cancer relapses are still common, affecting 20–35% of patients who undergo treatment with curative intent, such as prostatectomy. Particularly, two-thirds of such cases develop biochemical recurrence within the first two years, with early-on BCR implying a more aggressive disease and poor prognosis.

Current prognostic BCR tools have the shortcoming of being bounded to biopsy or whole-prostate specimen histopathology assessment to attain higher precision, associated with difficulties in evaluating aggressiveness of PCa prior to biopsy or surgery. To date, there has not been a thorough exploration of pretreatment, readily available, standard-of-care derived data, such as PCa imaging with MRI, to study this phenomenon in early stages of clinical workup.

This dissertation aimed at implementing bpMRI radiomic analysis of the whole-prostate to explore its potentially prognostic information for classification of 2-yr post-prostatectomy BCR, in conjunction with clinicohistopathological data, in three clinical phases: pre-biopsy, pre- and postoperative. For such, we created a curated database of bpMRI data (T2w, DWI b1000 MRI sequences and ADC maps) linked with clinical and histopathological data from 93 eligible patients (out of 250), treated with RP and examined in this study's institution.

Clinicohistopathological data descriptive analysis allowed to characterise RP-treated patients, where known postoperative prognostic BCR influencing factors revealed significant differences between BCR positive and negative groups, reflecting an adequate therapy selection. Yet, it depicted the difficulty in discriminating both classes of patients solely based on clinicohistopathological information available before treatment, even though biopsy grading is accessible at such timepoint. Nevertheless, a promising result was that grading of the contemporarily defined index tumour (larger PCa nodule in multifocal disease) revealed highly significant differences between groups, suggesting tumour index potential to summarise PCa aggressiveness.

Retrospective collection of bpMRI exams revealed heterogeneity in PCa imaging protocols and acquisition characteristics, leading to the development of methodology to enable fairer image comparison. It was mandatory to perform image registration and normalisation (with two techniques), FOV regularisation and costumised grey level discretisation through binning.

Tailored radiomic features extraction from T2w, DWI b1000 MRI sequences and ADC maps, generated a high-dimensional set, with approximately 2700 variables for each of the 93 patients. The large amount of information extracted with radiomics analysis required the development of a comprehensive machine-learning framework comprised of embedded LASSO logistic regression classification, repeated and stratified k-fold cross-validation and statistical analyses, and combined SMOTE oversampling to identify the best performing classifier of 2-yr BCR cases for each of the clinical scenarios considered. A primary analysis of the effect of image intensity normalisation method on models' performance was also executed.

We implemented logistic regression with LASSO, a L_1-regularisation method, and Priority-LASSO, a novel LASSO-based procedure designed to aggregate multiple omics data into "blocks" and to establish prior preference on the relevance of such blocks for model fitting.

In total, 2.177.280 models were fitted, resulting from 2 datasets generated with different normalisation techniques, 2 training datasets (non-oversampled/oversampled), 3 types of regularisation methods applied to the logistic regression modelling (LASSO, Priority-LASSO with 2 and 4 blocks), together with 5 sparsity configurations for LASSO, 6 for P-LASSO-2-blocks and 9 for P-LASSO-4-blocks, 2 block priority sequences implemented for P-LASSO-2-blocks and 24 for Priority-LASSO-4-blocks. All of these logistic regression models were fitted optimising the amount of correctly classified BCR positive cases. This was implemented through F-measure maximisation derived from ROC curve analysis (*Fmax*).

The final model's results derived from the normalisation of the whole-prostate gland with respect to that same organ, since a primary analysis led us to conclude that this method was preferential to differentiate the two BCR groups than when alternatively considering also the signal coming from immediate surrounding tissue belonging to the segmentation-box. Yet, standard normalisation of mpMRI prostate region procedures are lacking in the literature, and further research on its impact on quantitative imaging analysis is needed.

SMOTE was applied to generate a class-balanced dataset for training purposes. This technique was overoptimistic during validation and performed worse than non-oversampled models during test. SMOTE did not provide improvements for neither the minority nor the majority classes classification, being consequently discarded. Yet, alternative implementations of this technique should be explored.

Hypertuning of models' sparsity and priority sequence configurations, and their comparisons through statistical analyses allowed for the choice of the best performing model. Across all three clinical scenarios, Priority-LASSO-4-blocks implemented with priority sequence 19: *Clinical* > *T2* > *b1000* > *ADC*, with $pmax$ = (1,7,0,1) was the best and final radiomics-and-clinicohistopathologic-based classifier for each clinical scenario.

Our top classifiers performed with reasonably high F-measure (range: 0.702 – 0.754 and 0.910 – 0.925 for BCR positive and negative classes, respectively) when classifying PCa patients in the development set. For validation, we obtained very low F-measure for the BCR positive class (0.297 – 0.400). The discrepancy between BCR-positive class train and test classification results revealed that models overfitted and that training results were overoptimistic. Yet, models attained reasonably good F-measure and Precision for BCR negative class during training and validation phases (0.779 – 0.833 and 0.821 – 0.873), making these models worth exploring.

Despite Pre-biopsy models having yielded the lowest performance during training phase, for both minority and majority classes, in validation phase they achieved the highest median values of AUC (0.739), *Fmax* (0.400), *Pmax* (0.382) and *Rmax* (0.450) for the BCR positive class, and *Pmax* (0.873) for the BCR negative cases. Thus, pre-biopsy models might have suffered less from overfitting.

The Preoperative classifier produced the most accurate predictions of both classes during training, except for the threshold-free measure of AUC. However, it was an overoptimistic performance, given that Preoperative models showed the worst generalisability among all.

Yet, majority class classification with Postoperative models was the most successful in terms of *Fmax* (0.833) and *Rmax* (0.823).

This study did not allow for the development of stable models, in terms of performance nor selected features. A small cohort size characterised by a high number of variables not only contributes to model instability but also to being prone to failure. Thus, this variability was expected for this ill-posed and complex classification problem.

The combination of several implemented strategies was not enough to prevent model overfitting. We would like to restate that this study's images were originated in three different scanners from the same vendor but with different field strengths, sequence parameters, image resolution and FOVs. Image variety affects the information being extracted, which in turn can influence the performance of computerised algorithms. Moreover, mandatory image downsampling of in- and out-of-plane led to great loss of information.

As the chosen top model configuration was the same for all clinical scenario's final model (Priority-LASSO-4-blocks with $pmax$ = (1,7,0,1)), we concluded that addition of clinico-histopathological information that successively becomes available through the normal course of PCa clinical management did not improve the classification of 2-year BCR positive cases. With the analyses allowing for model selection, we observed that, overall, models' performances not only depend on the used method but also on its configuration, sparsity restrictions and clinical scenario. Comparison of best performing implementations of LASSO and Priority-LASSO led us to conclude that treating radiomic and clinical feature groups equally (w/ LASSO) led to worse performances.

Implementing models with all possible block priority sequences of radiomics data in Priority-LASSO implementations, let us to observe that (i) different block priority sequences (i.e. usage of group information) affected the potential of radiomics and other types of data—diverging from what is stated in the original algorithm's paper; (ii) the usage of group information combined with higher prioritisation of clinical variables led, overall, to better model performances; and (iii) higher level of segregation of radiomic blocks of variables (with P-LASSO-4-blocks) provided better results than when not separating radiomic features (with P-LASSO-2-blocks).

Baseline PSA, in combination with T2w and ADC variables, was selected for our pre-biopsy and preoperative models. Against our expectations, variables derived from biopsy histopathologic assessment (such as Gleason Score), were not selected by any of the preoperative models.

Out of 14, 13 postoperative models chose surgical resection (positive) margin extension as the most valuable (postoperative) clinical variable. Yet, incomplete cancer excision cannot be attributed to the technique of the surgeon nor surgery type, as the incidence of positive surgical margins depends also on the characteristics of the cancer (e.g. aggressiveness and location).

Favouring clinical features was not enough for BCR discrimination, it is therefore not surprising that current clinical nomograms are not exceptional at this task, even in conditions of greater statistical power.

The vast majority of selected radiomic variables were texture-based and derived from T2w filtered images. These findings aligned with previous studies revealing these features' importance for PCa aggressiveness evaluation and BCR prediction. Higher resolution of T2w compared to DWI and ADC maps may explain the relevance of T2w modality in this study. The results also shine a light on the fact that textural features could serve as non-invasive markers for assessing PCa aggressiveness. The recognition that texture features cannot be identified and detected by the human eye makes the medical contribution of these associations better understood.

The findings of this study are derived from a whole-prostate approach to characterise PCa disease —in contrast with single-lesion—, supporting the idea that relevant information can be found on with a whole-organ level and that it may have potential for BCR prediction. Focal delineation uncertainty and restriction to "humanly"-visible tumours were consequently avoided.

The more advanced quantitative analysis proposed in our study appears to be an important component of optimising the potential utility of MRI in 2-yr PCa BCR risk assessment, with the desired value of non-invasiveness for prospectiveness.

To date, the existing nomograms for the prediction of BCR after RP include information derived from the biopsy findings and/or from the evaluation of the whole surgical specimen, thus implicitly requiring the patient to be biopsied and/or undergo RP. Also, the more recent nomograms incorporating image-derived information involve radiological image evaluation, which maintains the issue of inter-reader concordance and low reproducibility. Moreover, to our knowledge, there are currently no pre-biopsy models for BCR prediction after RP, reasonably making them of higher clinical value and of greater need.

In this study with knowledge leveraging being challenged by the heterogeneity of bpMRI dataset characteristics and non-optimal model development conditions (small cohort, moderate follow-up, high-dimensionality, data imbalance, model overfitting) still allowed to develop a pre-biopsy model with a negative predictive value of 87.3%. For future studies, tackling the limitations pointed out might increase this performance and possibly allow for a better guidance of patients eligible for RP in a very early-point of PCa clinical management.

In the face of growing concern about the invasiveness and overutilisation of prostate biopsy and overtreatment of clinically insignificant cancer in urologic practice, and the strong tendency of removal of subjectiveness derived from human intervention, a pre-biopsy model based on baseline PSA and standard-of-care MRI-derived radiomic features of the whole-prostate gland could provide unseen insights into the aggressiveness of the disease, functioning as a "virtual biopsy". Thus, objectiveness in prognostic assessment could be achieved, allowing for patient counseling even prior to biopsy, or even precluding the need for one. This could also translate to more effective and personalised patient treatments, wherein decisions regarding post-RP active surveillance, intensified or diminished therapy could be made more objectively and reliably.

Future research should be conducted involving a larger number of patients with pre-biopsy MRI and longer follow-up, with development of effective standardisation and normalisation of images, as well as automatic and curated delineation of the whole-prostate. Also, the wide array of proposed radiomic textural features would need to be tested for reproducibility and robustness, allowing for proper feature selection. Finally, posterior validation of a model in external cohorts would be essential to ascertain the added value of these radiomic features, and to elucidate the currently uncertain role of bpMR imaging as a means to predict of BCR.

Even though PCa clinical management has a strong need for improved characterisation of the disease using imaging, applications of quantitative analysis of mpMRI are still limited in clinical practice. Yet, both should continue to grow hand-in-hand. Radiomics is an ambitious and promising field that will prosper from the globalisation of healthcare, with growing libraries of patient data accessible to clinicians. As the molecular landscape of each cancer is different, leading to variable responses to treatment, decoding of this complexity through phenotype imaging with radiomic analysis may be used to predict sensitivity or resistance to treatment. With cancer being such a complex disease, its characterisation and treatment strategy should follow suit.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, sep 2018.

[2] K. J. Bell, C. Del Mar, G. Wright, J. Dickinson, and P. Glasziou, “Prevalence of incidental prostate cancer: A systematic review of autopsy studies,” *International Journal of Cancer*, vol. 137, pp. 1749–1757, oct 2015.

[3] E. Segal, C. B. Sirlin, C. Ooi, A. S. Adler, J. Gollub, X. Chen, B. K. Chan, G. R. Matcuk, C. T. Barry, H. Y. Chang, and M. D. Kuo, “Decoding global gene expression programs in liver cancer by noninvasive imaging,” *Nature Biotechnology*, vol. 25, pp. 675–680, jun 2007.

[4] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts, “Radiomics: Extracting more information from medical images using advanced feature analysis,” *European Journal of Cancer*, vol. 48, pp. 441–446, mar 2012.

[5] A. Waugh and A. A. W. Grant, *Ross & Wilson Anatomy and Physiology in Health and Illness*, p. 450. 13th ed.

[6] J. T. Hansen, F. H. F. H. Netter, and C. A. G. Machado, *Netter’s Clinical Anatomy*, p. 256.

[7] V. Kumar, A. Abbas, and J. Aster, *Robbins & Cotran Pathologic Basis of Disease*, pp. 959–989. Elsevier, 9th ed., 2014.

[8] J. E. McNeal, “The zonal anatomy of the prostate,” *The Prostate*, vol. 2, pp. 35–49, jan 1981.

[9] Y. J. Choi, J. K. Kim, N. Kim, K. W. Kim, E. K. Choi, and K.-S. Cho, “Functional MR Imaging of Prostate Cancer,” *RadioGraphics*, vol. 27, pp. 63–75, jan 2007.

[10] A. H. P. Epstein, Jonathan; Cubilla, *Tumors of the Prostate Gland, Seminal Vesicles, Penis, and Scrotum: 14 (AFIP Atlas of Tumor Pathology: Series 4)*, p. 3. Amer Registry of Pathology, 1st ed., 2011.

[11] N. Mottet, R. C. N. V. D. Bergh, P. C. Vice-chair, M. D. Santis, S. Gillessen, A. Govorov, J. Grummet, A. M. Henry, T. B. Lam, M. D. Mason, T. H. V. D. Kwast, O. Rouvière, T. Wiegel, G. A. T. V. D. Broeck, M. Cumberbatch, N. Fossati, T. Gross, M. Lardas, M. Liew, L. Moris, I. G. Schoots, and P. M. Willemse, “EAU-ESUR-ESTRO-SIOG Guidelines on Prostate Cancer ,” 2018.

[12] R. J. Cohen, B. A. Shannon, M. Phillips, R. E. Moorin, T. M. Wheeler, and K. L. Garrett, “Central Zone Carcinoma of the Prostate Gland: A Distinct Tumor Type With Poor Prognostic Features,” *Journal of Urology*, vol. 179, pp. 1762–1767, may 2008.

[13] G. Ploussard, J. I. Epstein, R. Montironi, P. R. Carroll, M. Wirth, M.-O. Grimm, A. S. Bjartell, F. Montorsi, S. J. Freedland, A. Erbersdobler, and T. H. van der Kwast, “The Contemporary Concept of Significant Versus Insignificant Prostate Cancer,” *European Urology*, vol. 60, pp. 291–303, aug 2011.

[14] “World Population Ageing 2019: Highlights,” tech. rep., United Nations, Department of Economic and Social Affairs, Population Division (2019).

[15] N. Mottet, J. Bellmunt, M. Bolla, E. Briers, M. G. Cumberbatch, M. De Santis, N. Fossati, T. Gross, A. M. Henry, S. Joniau, T. B. Lam, M. D. Mason, V. B. Matveev, P. C. Moldovan, R. C. van den Bergh, T. Van den Broeck, H. G. van der Poel, T. H. van der Kwast, O. Rouvière, I. G. Schoots, T. Wiegel, and P. Cornford, “EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent,” *European Urology*, vol. 71, pp. 618–629, apr 2017.

[16] R. Saunders, J. Plun-Favreau, C. Takizawa, and W. Valentine, “Clinical and Economic Burden of Prostate Cancer,” *Value in Health*, vol. 18, p. A449, nov 2015.

[17] N. Mottet, R. C. N. van den Bergh, E. Briers, L. Bourke, P. Cornford, M. De Santis, S. Gillessen, A. Govorov, J. Grummet, A. M. Henry, T. B. Lam, M. D. Mason, H. G. van der Poel, T. H. van der Kwast, O. Rouvière, and T. Wiegel, *European Association of Urology Guidelines. 2019 Edition.*, vol. presented, european association of urology guidelines. 2019 edition. EAU ESTRO. European Association of Urology Guidelines Office, 2019.

[18] J. H. Hayes and M. J. Barry, "Screening for Prostate Cancer With the Prostate-Specific Antigen Test," *JAMA*, vol. 311, p. 1143, mar 2014.

[19] O. T. Okotie, K. A. Roehl, M. Han, S. Loeb, S. N. Gashti, and W. J. Catalona, "Characteristics of Prostate Cancer Detected by Digital Rectal Examination Only," *Urology*, vol. 70, pp. 1117–1120, dec 2007.

[20] T. A. Stamey, N. Yang, A. R. Hay, J. E. McNeal, F. S. Freiha, and E. Redwine, "Prostate-Specific Antigen as a Serum Marker for Adenocarcinoma of the Prostate," *New England Journal of Medicine*, vol. 317, pp. 909–916, oct 1987.

[21] A. Semjonow, B. Brandt, F. Oberpenning, S. Roth, and L. Hertle, "Discordance of assay methods creates pitfalls for the interpretation of prostate-specific antigen values.," *The Prostate. Supplement*, vol. 7, pp. 3–16, 1996.

[22] W. J. Catalona, J. P. Richie, F. R. Ahmann, M. A. Hudson, P. T. Scardino, R. C. Flanigan, J. B. DeKernion, T. L. Ratliff, L. R. Kavoussi, B. L. Dalkin, W. B. Waters, M. T. MacFarlane, and P. C. Southwick, "Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men.," *The Journal of urology*, vol. 151, no. 5, pp. 1283–90, 1994.

[23] M. F. O'Brien, A. M. Cronin, P. A. Fearn, B. Smith, J. Stasi, B. Guillonneau, P. T. Scardino, J. A. Eastham, A. J. Vickers, and H. Lilja, "Pretreatment prostate-specific antigen (PSA) velocity and doubling time are associated with outcome but neither improves prediction of outcome beyond pretreatment PSA alone in patients treated with radical prostatectomy.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 27, pp. 3591–7, aug 2009.

[24] A. J. Vickers, C. Savage, M. F. O'Brien, and H. Lilja, "Systematic review of pretreatment prostate-specific antigen velocity and doubling time as predictors for prostate cancer.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 27, pp. 398–403, jan 2009.

[25] D. F. Gleason, "Classification of prostatic carcinomas.," *Cancer chemotherapy reports. Part 1*, 1966.

[26] J. I. Epstein, M. J. Zelefsky, D. D. Sjoberg, J. B. Nelson, L. Egevad, C. Magi-Galluzzi, A. J. Vickers, A. V. Parwani, V. E. Reuter, S. W. Fine, J. A. Eastham, P. Wiklund, M. Han, C. A. Reddy, J. P. Ciezki, T. Nyberg, and E. A. Klein, "A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score," *European Urology*, vol. 69, pp. 428–435, mar 2016.

[27] Jonathan I. Epstein, Lars Egevad, Mahul B. Amin, Brett Delahunt, John R. Srigley, and Peter A. Humphrey, "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," *The American Journal of Surgical Pathology*, vol. 40, pp. 244–252, feb 2016.

[28] C. J. Kane, S. E. Eggener, A. W. Shindel, and G. L. Andriole, "Variability in Outcomes for Patients with Intermediate-risk Prostate Cancer (Gleason Score 7, International Society of Urological Pathology Gleason Group 2–3) and Implications for Risk Stratification: A Systematic Review," *European Urology Focus*, vol. 3, pp. 487–497, oct 2017.

[29] J. O. Barentsz, J. Richenberg, R. Clements, P. Choyke, S. Verma, G. Villeirs, O. Rouviere, V. Logager, and J. J. Fütterer, "ESUR Prostate MR guidelines 2012," *European Radiology*, vol. 22, pp. 746–757, apr 2012.

[30] F. Bratan, E. Niaf, C. Melodelima, A. L. Chesnais, R. Souchon, F. Mège-Lechevallier, M. Colombel, and O. Rouvière, "Influence of imaging and histological factors on prostate cancer detection and localisation on multiparametric MRI: a prospective study," *European Radiology*, vol. 23, pp. 2019–2029, jul 2013.

[31] J. D. Le, N. Tan, E. Shkolyar, D. Y. Lu, L. Kwan, L. S. Marks, J. Huang, D. J. Margolis, S. S. Raman, and R. E. Reiter, "Multifocality and Prostate Cancer Detection by Multiparametric Magnetic Resonance Imaging: Correlation with Whole-mount Histopathology," *European Urology*, vol. 67, pp. 569–576, mar 2015.

[32] S. Borofsky, A. K. George, S. Gaur, M. Bernardo, M. D. Greer, F. V. Mertan, M. Taffel, V. Moreno, M. J. Merino, B. J. Wood, P. A. Pinto, P. L. Choyke, and B. Turkbey, "What Are We Missing? False-Negative Cancers at Multiparametric MR Imaging of the Prostate," *Radiology*, vol. 286, pp. 186–195, jan 2018.

[33] F.-J. H. Drost, D. F. Osses, D. Nieboer, E. W. Steyerberg, C. H. Bangma, M. J. Roobol, and I. G. Schoots, "Prostate MRI, with or without MRI-targeted biopsy, and systematic biopsy for detecting prostate cancer," *Cochrane Database of Systematic Reviews*, apr 2019.

[34] M. de Rooij, E. H. Hamoen, J. A. Witjes, J. O. Barentsz, and M. M. Rovers, “Accuracy of Magnetic Resonance Imaging for Local Staging of Prostate Cancer: A Diagnostic Meta-analysis,” *European Urology*, vol. 70, pp. 233–245, aug 2016.

[35] G. J. Jager, E. T. Ruijter, C. A. van de Kaa, J. J. de la Rosette, G. O. Oosterhof, J. R. Thornbury, and J. O. Barentsz, “Local staging of prostate cancer with endorectal MR imaging: correlation with histopathology.,” *AJR. American journal of roentgenology*, vol. 166, pp. 845–52, apr 1996.

[36] F. Cornud, T. Flam, L. Chauveinc, K. Hamida, Y. Chrétien, A. Vieillefond, O. Hélénon, and J. F. Moreau, “Extraprostatic Spread of Clinically Localized Prostate Cancer: Factors Predictive of pT3 Tumor and of Positive Endorectal MR Imaging Examination Results,” *Radiology*, vol. 224, pp. 203–210, jul 2002.

[37] L. Wang, M. Mullerad, H.-N. Chen, S. C. Eberhardt, M. W. Kattan, P. T. Scardino, and H. Hricak, “Prostate Cancer: Incremental Value of Endorectal MR Imaging Findings for Prediction of Extracapsular Extension,” *Radiology*, vol. 232, pp. 133–139, jul 2004.

[38] A. V. D’Amico, R. Whittington, B. Malkowicz, M. Schnall, D. Schultz, K. Cote, J. E. Tomaszewski, and A. Wein, “Endorectal magnetic resonance imaging as a predictor of biochemical outcome after radical prostatectomy in men with clinically localized prostate cancer,” *Journal of Urology*, 2000.

[39] M. R. Engelbrecht, G. J. Jager, and J. Severens, “Patient Selection for Magnetic Resonance Imaging of Prostate Cancer,” *European Urology*, vol. 40, pp. 300–307, sep 2001.

[40] A. M. Hövels, R. A. Heesakkers, E. M. Adang, G. J. Jager, S. Strum, Y. L. Hoogeveen, J. L. Severens, and J. O. Barentsz, “The diagnostic accuracy of CT and MRI in the staging of pelvic lymph nodes in patients with prostate cancer: a meta-analysis,” *Clinical Radiology*, 2008.

[41] B. Kiss, H. C. Thoeny, and U. E. Studer, “Current Status of Lymph Node Imaging in Bladder and Prostate Cancer,” *Urology*, vol. 96, pp. 1–7, oct 2016.

[42] H. C. Thoeny, J. M. Froehlich, M. Triantafyllou, J. Huesler, L. J. Bains, P. Vermathen, A. Fleischmann, and U. E. Studer, “Metastases in Normal-sized Pelvic Lymph Nodes: Detection with Diffusion-weighted MR Imaging,” *Radiology*, vol. 273, pp. 125–135, oct 2014.

[43] M. R. Cooperberg, D. J. Pasta, E. P. Elkin, M. S. Litwin, D. M. Latini, J. Du Chane, and P. R. Carroll, “Prostate Risk Assessment score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy.,” *The Journal of Urology*, vol. 173, pp. 1938–42, jun 2005.

[44] G. D. Coughlin, J. W. Yaxley, S. K. Chambers, S. Occhipinti, H. Samaratunga, L. Zajdlewicz, P. Teloken, N. Dungli-son, S. Williams, M. F. Lavin, and R. A. Gardiner, “Robot-assisted laparoscopic prostatectomy versus open radical retropubic prostatectomy: 24-month outcomes from a randomised controlled study.,” *The Lancet. Oncology*, vol. 19, pp. 1051–1060, aug 2018.

[45] P. A. de Carvalho, J. A. B. A. Barbosa, G. B. Guglielmetti, M. D. Cordeiro, B. Rocco, W. C. Nahas, V. Patel, and R. F. Coelho, “Retrograde Release of the Neurovascular Bundle with Preservation of Dorsal Venous Complex During Robot-assisted Radical Prostatectomy: Optimizing Functional Outcomes.,” *European urology*, vol. 0, jul 2018.

[46] V. Srougi, J. Bessa, M. Baghdadi, I. Nunes-Silva, J. B. da Costa, S. Garcia-Barreras, E. Barret, F. Rozet, M. Galiano, R. Sanchez-Salas, and X. Cathelineau, “Surgical method influences specimen margins and biochemical recurrence during radical prostatectomy for high-risk prostate cancer: a systematic review and meta-analysis,” *World Journal of Urology*, vol. 35, pp. 1481–1488, oct 2017.

[47] J. Binder, J. Jones, W. Bentas, M. Wolfram, R. Bräutigam, M. Probst, W. Kramer, and D. Jonas, “Roboterunterstützte Laparoskopie in der Urologie Radikale Prostatektomie und rekonstruktive retroperitoneale Eingriffe [Robot-assisted laparoscopy in urology. Radical prostatectomy and reconstructive retroperitoneal interventions],” *Der Urologe A*, vol. 41, pp. 144–149, mar 2002.

[48] F. C. Hamdy, J. L. Donovan, J. A. Lane, M. Mason, C. Metcalfe, P. Holding, M. Davis, T. J. Peters, E. L. Turner, R. M. Martin, J. Oxley, M. Robinson, J. Staffurth, E. Walsh, P. Bollina, J. Catto, A. Doble, A. Doherty, D. Gillatt, R. Kockelbergh, H. Kynaston, A. Paul, P. Powell, S. Prescott, D. J. Rosario, E. Rowe, and D. E. Neal, “10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer,” *New England Journal of Medicine*, vol. 375, pp. 1415–1424, oct 2016.

[49] Ö. Dillioglugil, B. D. Leibman, M. W. Kattan, C. Seale-Hawkins, T. M. Wheeler, and P. T. Scardino, “Hazard rates for progression after radical prostatectomy for clinically localized prostate cancer,” *Urology*, vol. 50, pp. 93–99, jul 1997.

[50] J. Walz, F. K.-H. Chun, E. A. Klein, A. Reuther, F. Saad, M. Graefen, H. Huland, and P. I. Karakiewicz, "Nomogram Predicting the Probability of Early Recurrence After Radical Prostatectomy for Prostate Cancer," *The Journal of Urology*, vol. 181, pp. 601–608, feb 2009.

[51] A. J. Stephenson, M. W. Kattan, J. A. Eastham, Z. A. Dotan, F. J. Bianco, H. Lilja, and P. T. Scardino, "Defining Biochemical Recurrence of Prostate Cancer After Radical Prostatectomy: A Proposal for a Standardized Definition," *Journal of Clinical Oncology*, vol. 24, pp. 3973–8, aug 2006.

[52] E. M. Horwitz, H. D. Thames, D. A. Kuban, L. B. Levy, P. A. Kupelian, A. A. Martinez, J. M. Michalski, T. M. Pisansky, H. M. Sandler, W. U. Shipley, M. J. Zelefsky, G. E. Hanks, and A. L. Zietman, "Definitions of biochemical failure that best predict clinical failure in patients with prostate cancer treated with external beam radiation alone: A multi-institutional pooled analysis," in *Journal of Urology*, 2005.

[53] T. Van den Broeck, R. C. van den Bergh, N. Arfi, T. Gross, L. Moris, E. Briers, M. Cumberbatch, M. De Santis, D. Tilki, S. Fanti, N. Fossati, S. Gillessen, J. P. Grummet, A. M. Henry, M. Lardas, M. Liew, O. Rouvière, J. Pecanka, M. D. Mason, I. G. Schoots, T. H. van Der Kwast, H. G. van Der Poel, T. Wiegel, P.-P. M. Willemse, Y. Yuan, T. B. Lam, P. Cornford, and N. Mottet, "Prognostic Value of Biochemical Recurrence Following Treatment with Curative Intent for Prostate Cancer: A Systematic Review," *European Urology*, 2018.

[54] C. R. Pound, A. W. Partin, M. A. Eisenberger, D. W. Chan, J. D. Pearson, and P. C. Walsh, "Natural History of Progression After PSA Elevation Following Radical Prostatectomy," *JAMA*, vol. 281, p. 1591, may 1999.

[55] S. A. Boorjian, R. H. Thompson, M. K. Tollefson, L. J. Rangel, E. J. Bergstralh, M. L. Blute, and R. J. Karnes, "Long-Term Risk of Clinical Progression After Biochemical Recurrence Following Radical Prostatectomy: The Impact of Time from Surgery to Recurrence," *European Urology*, vol. 59, pp. 893–899, jun 2011.

[56] T. A. Stamey, J. N. Kabalin, J. E. McNeal, I. M. Johnstone, F. Freiha, E. A. Redwine, and N. Yang, "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients.," *The Journal of urology*, vol. 141, pp. 1076–83, may 1989.

[57] L. Boccon-Gibod, B. Djavan, P. Hammerer, W. Hoeltl, M. Kattan, T. Prayer-Galetti, P. Teillac, and U. Tunn, "Management of prostate-specific antigen relapse in prostate cancer: A European consensus," *International Journal of Clinical Practice*, vol. 58, pp. 382–390, apr 2004.

[58] C. L. Amling, E. J. Bergstralh, M. L. Blute, J. M. Slezak, and H. Zincke, "Defining prostate specific antigen progression after radical prostatectomy: what is the most appropriate cut point?," *The Journal of urology*, vol. 165, pp. 1146–51, apr 2001.

[59] A. Toussi, S. B. Stewart-Merrill, S. A. Boorjian, S. P. Psutka, R. H. Thompson, I. Frank, M. K. Tollefson, M. T. Gettman, R. E. Carlson, L. J. Rangel, and R. J. Karnes, "Standardizing the Definition of Biochemical Recurrence after Radical Prostatectomy—What Prostate Specific Antigen Cut Point Best Predicts a Durable Increase and Subsequent Systemic Progression?," *The Journal of Urology*, vol. 195, pp. 1754–1759, jun 2016.

[60] "Prognostic Value of Biochemical Recurrence Following Treatment with Curative Intent for Prostate Cancer: A Systematic Review," *European Urology*, vol. 75, pp. 967–987, jun 2019.

[61] R. S. Pompe, M. Bandini, F. Preisser, M. Marchioni, E. Zaffuto, Z. Tian, G. Salomon, T. Schlomm, H. Huland, M. Graefen, D. Tilki, S. F. Shariat, and P. I. Karakiewicz, "Contemporary approach to predict early biochemical recurrence after radical prostatectomy: update of the Walz nomogram," *Prostate Cancer and Prostatic Diseases*, vol. 21, pp. 386–393, sep 2018.

[62] M. Han, A. W. Partin, C. R. Pound, J. I. Epstein, and P. C. Walsh, "Long-term biochemical disease-free and cancer-specific survival following anatomic radical retropubic prostatectomy. The 15-year Johns Hopkins experience," *Urologic Clinics of North America*, vol. 28, pp. 555–565, aug 2001.

[63] L. Budäus, J. Schiffmann, M. Graefen, H. Huland, P. Tennstedt, A. Siegmann, D. Böhmer, V. Budach, D. Bartkowiak, and T. Wiegel, "Defining biochemical recurrence after radical prostatectomy and timing of early salvage radiotherapy," *Strahlentherapie und Onkologie*, vol. 193, pp. 692–699, sep 2017.

[64] E. Brunocilla, C. Pultrone, R. Pernetti, R. Schiavina, and G. Martorana, "Preservation of the smooth muscular internal (vesical) sphincter and of the proximal urethra during retropubic radical prostatectomy: Description of the technique," *International Journal of Urology*, vol. 19, pp. 783–785, aug 2012.

[65] F. Barchetti and V. Panebianco, "Multiparametric MRI for Recurrent Prostate Cancer Post Radical Prostatectomy and Postradiation Therapy," *BioMed Research International*, vol. 2014, pp. 1–23, 2014.

[66] S. J. Freedland, E. B. Humphreys, L. A. Mangold, M. Eisenberger, and A. W. Partin, “Time to Prostate Specific Antigen Recurrence After Radical Prostatectomy and Risk of Prostate Cancer Specific Mortality,” *The Journal of Urology*, vol. 176, pp. 1404–1408, oct 2006.

[67] S. J. Freedland, E. B. Humphreys, L. A. Mangold, M. Eisenberger, F. J. Dorey, P. C. Walsh, and A. W. Partin, “Risk of Prostate Cancer–Specific Mortality Following Biochemical Recurrence After Radical Prostatectomy,” *JAMA*, vol. 294, p. 433, jul 2005.

[68] M. Kattan, “Statistical Prediction Models, Artificial Neural Networks, and the Sophism “I Am a Patient, Not a Statistic”,” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 20, pp. 885–7, feb 2002.

[69] A. V. D’Amico, R. Whittington, S. B. Malkowicz, D. Schultz, K. Blank, G. A. Broderick, J. E. Tomaszewski, A. A. Renshaw, I. Kaplan, C. J. Beard, and A. Wein, “Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer,” *JAMA*, vol. 280, p. 969, sep 1998.

[70] A. J. Stephenson, P. T. Scardino, J. A. Eastham, F. J. Bianco, Z. A. Dotan, P. A. Fearn, and M. W. Kattan, “Preoperative Nomogram Predicting the 10-Year Probability of Prostate Cancer Recurrence After Radical Prostatectomy,” *JNCI: Journal of the National Cancer Institute*, vol. 98, pp. 715–717, may 2006.

[71] M. W. Kattan, J. A. Eastham, A. M. F. Stapleton, T. M. Wheeler, and P. T. Scardino, “A Preoperative Nomogram for Disease Recurrence Following Radical Prostatectomy for Prostate Cancer,” *JNCI: Journal of the National Cancer Institute*, vol. 90, pp. 766–771, may 1998.

[72] L. Dickinson, H. U. Ahmed, C. Allen, J. O. Barentsz, B. Carey, J. J. Futterer, S. W. Heijmink, P. J. Hoskin, A. Kirkham, A. R. Padhani, R. Persad, P. Puech, S. Punwani, A. S. Sohaib, B. Tombal, A. Villers, J. van der Meulen, and M. Emberton, “Magnetic Resonance Imaging for the Detection, Localisation, and Characterisation of Prostate Cancer: Recommendations from a European Consensus Meeting,” *European Urology*, vol. 59, pp. 477–494, apr 2011.

[73] J. C. Weinreb, J. O. Barentsz, P. L. Choyke, F. Cornud, M. A. Haider, K. J. Macura, D. Margolis, M. D. Schnall, F. Shtern, C. M. Tempany, H. C. Thoeny, and S. Verma, “PI-RADS Prostate Imaging – Reporting and Data System: 2015, Version 2,” *European Urology*, vol. 69, pp. 16–40, jan 2016.

[74] B. Turkbey, A. B. Rosenkrantz, M. A. Haider, A. R. Padhani, G. Villeirs, K. J. Macura, C. M. Tempany, P. L. Choyke, F. Cornud, D. J. Margolis, H. C. Thoeny, S. Verma, J. Barentsz, and J. C. Weinreb, “Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2.,” *European urology*, vol. 76, pp. 340–351, sep 2019.

[75] C. Sato, S. Naganawa, T. Nakamura, H. Kumada, S. Miura, O. Takizawa, and T. Ishigaki, “Differentiation of noncancerous tissue and cancer lesions by apparent diffusion coefficient values in transition and peripheral zones of the prostate,” *Journal of Magnetic Resonance Imaging*, vol. 21, pp. 258–262, mar 2005.

[76] Y. Peng, Y. Jiang, C. Yang, J. B. Brown, T. Antic, I. Sethi, C. Schmid-Tannwald, M. L. Giger, S. E. Eggener, and A. Oto, “Quantitative Analysis of Multiparametric Prostate MR Images: Differentiation between Prostate Cancer and Normal Tissue and Correlation with Gleason Score—A Computer-aided Diagnosis Development Study,” *Radiology*, vol. 267, pp. 787–796, jun 2013.

[77] F. Khalvati, A. Wong, and M. A. Haider, “Automated prostate cancer detection via comprehensive multi-parametric magnetic resonance imaging texture feature models,” *BMC Medical Imaging*, vol. 15, p. 27, aug 2015.

[78] G. J. S. Litjens, R. Elliott, N. N. Shih, M. D. Feldman, T. Kobus, C. Hulsbergen-van de Kaa, J. O. Barentsz, H. J. Huisman, and A. Madabhushi, “Computer-extracted Features Can Distinguish Noncancerous Confounding Disease from Prostatic Adenocarcinoma at Multiparametric MR Imaging,” *Radiology*, vol. 278, pp. 135–145, jan 2016.

[79] A. Cameron, F. Khalvati, M. A. Haider, and A. Wong, “MAPS: A Quantitative Radiomics Approach for Prostate Cancer Detection,” *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 1145–1156, jun 2016.

[80] G. Gao, C. Wang, X. Zhang, J. Hu, X. Yang, H. Wang, J. Zhang, and X. Wang, “Quantitative analysis of diffusion-weighted magnetic resonance images: differentiation between prostate cancer and normal tissue based on a computer-aided diagnosis system,” *Science China Life Sciences*, vol. 60, no. 1, pp. 37–43, 2017.

[81] S. Gaur, N. Lay, S. A. Harmon, S. Doddakashi, S. Mehralivand, B. Argun, T. Barrett, S. Bednarova, R. Girometti, E. Karaarslan, A. R. Kural, A. Oto, A. S. Purysko, T. Antic, C. Magi-Galluzzi, Y. Saglican, S. Sioletic, A. Y. Warren, L. Bittencourt, J. J. Fütterer, R. T. Gupta, I. Kabakus, Y. M. Law, D. J. Margolis, H. Shebel, A. C. Westphalen, B. J. Wood, P. A. Pinto, J. H. Shih, P. L. Choyke, R. M. Summers, and B. Turkbey, “Can computer-aided diagnosis assist in the identification of prostate cancer on prostate MRI? a multi-center, multi-reader investigation,” *Oncotarget*, vol. 9, no. 73, pp. 33804–33817, 2018.

[82] T. Chen, M. Li, Y. Gu, Y. Zhang, S. Yang, C. Wei, J. Wu, X. Li, W. Zhao, and J. Shen, "Prostate Cancer Differentiation and Aggressiveness: Assessment With a Radiomic-Based Model vs. PI-RADS v2," sep 2018.

[83] R. Cuocolo, A. Stanzione, A. Ponsiglione, V. Romeo, F. Verde, M. Creta, R. La Rocca, N. Longo, L. Pace, and M. Imbriaco, "Clinically significant prostate cancer detection on MRI: A radiomic shape features study," *European Journal of Radiology*, vol. 116, pp. 144–149, jul 2019.

[84] V. Poulakis, U. Witzsch, R. de Vries, V. Emmerlich, M. Meves, H.-M. Altmannsberger, and E. Becht, "Preoperative neural network using combined magnetic resonance imaging variables, prostate-specific antigen, and gleason score for predicting prostate cancer biochemical recurrence after radical prostatectomy," *Urology*, vol. 64, pp. 1165–1170, dec 2004.

[85] M. H. Fuchsjäger, A. Shukla-Dave, H. Hricak, L. Wang, K. Touijer, J. F. Donohue, J. A. Eastham, and M. W. Kattan, "Magnetic resonance imaging in the prediction of biochemical recurrence of prostate cancer after radical prostatectomy," *BJU International*, vol. 104, pp. 315–320, aug 2009.

[86] K. Nishida, S. Yuen, K. Kamoi, K. Yamada, K. Akazawa, H. Ito, K. Okihara, A. Kawauchi, T. Miki, and T. Nishimura, "Incremental value of T2-weighted and diffusion-weighted MRI for prediction of biochemical recurrence after radical prostatectomy in clinically localized prostate cancer," *Acta Radiologica*, vol. 52, pp. 120–126, feb 2011.

[87] S. Y. Park, C. K. Kim, B. K. Park, H. M. Lee, and K. S. Lee, "Prediction of biochemical recurrence following radical prostatectomy in men with prostate cancer by diffusion-weighted magnetic resonance imaging: initial results," *European Radiology*, vol. 21, pp. 1111–1118, may 2011.

[88] D. L. Langer, T. H. Van Der Kwast, A. J. Evans, A. Plotkin, J. Trachtenberg, B. C. Wilson, and M. A. Haider, "Prostate tissue composition and MR measurements: Investigating the relationships between ADC, T2, Ktrans, Ve, and corresponding histologic features," *Radiology*, 2010.

[89] S. Jung, O. F. Donati, H. A. Vargas, D. Goldman, H. Hricak, and O. Akin, "Transition zone prostate cancer: Incremental value of diffusion-weighted endorectal MR imaging in tumor detection and assessment of aggressiveness," *Radiology*, 2013.

[90] O. F. Donati, Y. Mazaheri, A. Afaq, H. A. Vargas, J. Zheng, C. S. Moskowitz, H. Hricak, and O. Akin, "Prostate cancer aggressiveness: Assessment with whole-lesion histogram analysis of the apparent diffusion coefficient," *Radiology*, 2014.

[91] O. F. Donati, A. Afaq, H. A. Vargas, Y. Mazaheri, J. Zheng, C. S. Moskowitz, H. Hricak, and O. Akin, "Prostate MRI: Evaluating tumor volume and apparent diffusion coefficient as surrogate biomarkers for predicting tumor Gleason score," *Clinical Cancer Research*, 2014.

[92] S. Hattori, T. Kosaka, R. Mizuno, K. Kanao, A. Miyajima, Y. Yasumizu, S. Yazawa, H. Nagata, E. Kikuchi, S. Mikami, M. Jinzaki, K. Nakagawa, A. Tanimoto, and M. Oya, "Prognostic value of preoperative multiparametric magnetic resonance imaging (MRI) for predicting biochemical recurrence after radical prostatectomy," *BJU International*, vol. 113, pp. 741–747, may 2014.

[93] J. J. Park, C. K. Kim, S. Y. Park, B. K. Park, H. M. Lee, and S. W. Cho, "Prostate Cancer: Role of Pretreatment Multiparametric 3-T MRI in Predicting Biochemical Recurrence After Radical Prostatectomy," *American Journal of Roentgenology*, vol. 202, pp. W459–W465, may 2014.

[94] R. Ho, M. M. Siddiqui, A. K. George, T. Frye, A. Kilchevsky, M. Fascelli, N. A. Shakir, R. Chelluri, S. F. Abboud, A. Walton-Diaz, S. Sankineni, M. J. Merino, B. Turkbey, P. L. Choyke, B. J. Wood, and P. A. Pinto, "Preoperative Multiparametric Magnetic Resonance Imaging Predicts Biochemical Recurrence in Prostate Cancer after Radical Prostatectomy," *PloS one*, vol. 11, no. 6, p. e0157313, 2016.

[95] Y.-D. Zhang, C.-J. Wu, M.-L. Bao, H. Li, X.-N. Wang, X.-S. Liu, and H.-B. Shi, "MR-based prognostic nomogram for prostate cancer after radical prostatectomy," *Journal of Magnetic Resonance Imaging*, vol. 45, pp. 586–596, feb 2017.

[96] Z.-N. Zhang, C. Luo, B. Xu, H.-F. Song, B.-L. Ma, and Q. Zhang, "Preoperative PROSTATE scoring system: a potential predictive tool for the risk of biochemical recurrence after radical prostatectomy.," *Cancer Management and Research*, vol. 10, pp. 4671–4677, 2018.

[97] S. Ghose, R. Shiradkar, M. Rusu, J. Mitra, R. Thawani, M. Feldman, A. C. Gupta, A. S. Purysko, L. Ponsky, and A. Madabhushi, "Prostate shapes on pre-treatment MRI between prostate cancer patients who do and do not undergo biochemical recurrence are different: Preliminary Findings.," *Scientific Reports*, vol. 7, p. 15829, nov 2017.

[98] Y.-D. Zhang, J. Wang, C.-J. Wu, M.-L. Bao, H. Li, X.-N. Wang, J. Tao, and H.-B. Shi, “An imaging-based approach predicts clinical outcomes in prostate cancer through a novel support vector machine classification,” *Oncotarget*, vol. 7, pp. 78140–78151, nov 2016.

[99] L. A. R. Reisæter, J. J. Fütterer, A. Losnegård, Y. Nygård, J. Monssen, K. Gravdal, O. J. Halvorsen, L. A. Akslen, M. Biermann, S. Haukaas, J. Rørvik, and C. Beisland, “Optimising preoperative risk stratification tools for prostate cancer using mpMRI,” *European Radiology*, vol. 28, pp. 1016–1026, mar 2018.

[100] M. Y. Yoon, J. Park, J. Y. Cho, C. W. Jeong, J. H. Ku, H. H. Kim, and C. Kwak, “Predicting biochemical recurrence in patients with high-risk prostate cancer using the apparent diffusion coefficient of magnetic resonance imaging.,” *Investigative and Clinical Urology*, vol. 58, no. 1, pp. 12–19, 2017.

[101] L. Qi, C.-J. Wu, J. Zhang, M.-L. Bao, X. Yan, and Y.-D. Zhang, “Construction of a Preoperative Radiologic-Risk Signature for Predicting the Pathologic Status of Prostate Cancer at Radical Prostatectomy,” *American Journal of Roentgenology*, vol. 211, pp. 805–811, oct 2018.

[102] M. D. Greer, A. M. Brown, J. H. Shih, R. M. Summers, J. Marko, Y. M. Law, S. Sankineni, A. K. George, M. J. Merino, P. A. Pinto, P. L. Choyke, and B. Turkbey, “Accuracy and agreement of PIRADSv2 for prostate cancer mpMRI: A multireader study,” *Journal of Magnetic Resonance Imaging*, vol. 45, pp. 579–585, feb 2017.

[103] J. Thompson, P. van Leeuwen, D. Moses, R. Shnier, P. Brenner, W. Delprado, M. Pulbrook, M. Böhm, A. Haynes, A. Hayen, and P. Stricker, “The Diagnostic Performance of Multiparametric Magnetic Resonance Imaging to Detect Significant Prostate Cancer,” *Journal of Urology*, vol. 195, pp. 1428–1435, may 2016.

[104] R. Renard-Penna, P. Mozer, F. Cornud, N. Barry-Delongchamps, E. Bruguière, D. Portalez, and B. Malavaud, “Prostate Imaging Reporting and Data System and Likert Scoring System: Multiparametric MR Imaging Validation Study to Screen Patients for Initial Biopsy,” *Radiology*, vol. 275, pp. 458–468, may 2015.

[105] F. Chen, S. Cen, and S. Palmer, “Application of Prostate Imaging Reporting and Data System Version 2 (PI-RADS v2),” *Academic Radiology*, vol. 24, pp. 1101–1106, sep 2017.

[106] L. K. Bittencourt, J. O. Barentsz, L. C. D. De Miranda, and E. L. Gasparetto, “Prostate MRI: Diffusion-weighted imaging at 1.5T correlates better with prostatectomy Gleason grades than TRUS-guided biopsies in peripheral zone tumours,” *European Radiology*, 2012.

[107] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nature Communications*, vol. 5, p. 4006, sep 2014.

[108] T. A. Data, R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: Images Are More than Pictures, They Are Data,” vol. 278, no. 2, 2016.

[109] H. J. Aerts, “The potential of radiomic-based phenotyping in precision medicine: a review,” *JAMA Oncology*, vol. 2, no. 12, pp. 1636–1642, 2016.

[110] L. Liu, Z. Tian, Z. Zhang, and B. Fei, “Computer-aided Detection of Prostate Cancer with MRI: Technology and Applications.,” *Academic Radiology*, vol. 23, no. 8, pp. 1024–46, 2016.

[111] F. Khalvati, J. Zhang, A. G. Chung, M. J. Shafiee, A. Wong, and M. A. Haider, “MPCaD: a multi-scale radiomics-driven framework for automated prostate cancer localization and detection,” *BMC Medical Imaging*, vol. 18, p. 16, dec 2018.

[112] A. Wibmer, H. Hricak, T. Gondo, K. Matsumoto, H. Veeraraghavan, D. Fehr, J. Zheng, D. Goldman, C. Moskowitz, S. W. Fine, V. E. Reuter, J. Eastham, E. Sala, and H. A. Vargas, “Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores,” *European Radiology*, vol. 25, pp. 2840–2850, oct 2015.

[113] D. Kwon, I. M. Reis, A. L. Breto, Y. Tschudi, N. Gautney, O. Zavala-Romero, C. Lopez, J. C. Ford, S. Punnen, A. Pollack, and R. Stoyanova, “Classification of suspicious lesions on prostate multiparametric MRI using machine learning.,” *Journal of Medical Imaging (Bellingham, Wash.)*, vol. 5, p. 034502, jul 2018.

[114] D. Bonekamp, S. Kohl, M. Wiesenfarth, P. Schelb, J. P. Radtke, M. Götz, P. Kickingereder, K. Yaqubi, B. Hitthaler, N. Gählert, T. A. Kuder, F. Deister, M. Freitag, M. Hohenfellner, B. A. Hadaschik, H.-P. Schlemmer, and K. H. Maier-Hein, “Radiomic Machine Learning for Characterization of Prostate Lesions with MRI: Comparison to ADC Values,” *Radiology*, vol. 289, pp. 128–137, oct 2018.

[115] J. Wang, C.-J. Wu, M.-L. Bao, J. Zhang, X.-N. Wang, and Y.-D. Zhang, "Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer," *European Radiology*, vol. 27, pp. 4082–4090, oct 2017.

[116] X. Min, M. Li, D. Dong, Z. Feng, P. Zhang, Z. Ke, H. You, F. Han, J. Tian, and L. Wang, "Multi-parametric MRI-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: Cross-validation of a machine learning method," *European Journal of Radiology*, vol. 115, pp. 16–21, 2019.

[117] D. Fehr, H. Veeraraghavan, A. Wibmer, T. Gondo, K. Matsumoto, H. A. Vargas, E. Sala, H. Hricak, and J. O. Deasy, "Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. E6265–73, nov 2015.

[118] A. Chaddad, T. Niazi, S. Probst, F. Bladou, M. Anidjar, and B. Bahoric, "Predicting Gleason Score of Prostate Cancer Patients Using Radiomic Analysis," *Frontiers in Oncology*, vol. 8, p. 630, dec 2018.

[119] R. Shiradkar, S. Ghose, I. Jambor, P. Taimen, O. Ettala, A. S. Purysko, and A. Madabhushi, "Radiomic features from pretreatment biparametric MRI predict prostate cancer biochemical recurrence: Preliminary findings," *Journal of Magnetic Resonance Imaging*, vol. 48, pp. 1626–1636, dec 2018.

[120] K. Gnep, A. Fargeas, R. E. Gutiérrez-Carvajal, F. Commandeur, R. Mathieu, J. D. Ospina, Y. Rolland, T. Rohou, S. Vincendeau, M. Hatt, O. Acosta, and R. de Crevoisier, "Haralick textural features on T2-weighted MRI are associated with biochemical recurrence following radiotherapy for peripheral zone prostate cancer," *Journal of Magnetic Resonance Imaging*, vol. 45, pp. 103–117, jan 2017.

[121] V. Bourbonne, M. Vallières, F. Lucia, L. Doucet, D. Visvikis, V. Tissot, O. Pradier, M. Hatt, and U. Schick, "MRI-Derived Radiomics to Guide Post-operative Management for High-Risk Prostate Cancer," *Frontiers in Oncology*, vol. 9, p. 807, aug 2019.

[122] V. Bourbonne, G. Fournier, M. Vallières, F. Lucia, L. Doucet, V. Tissot, G. Cuvelier, S. Hue, H. Le Penn Du, L. Perdriel, N. Bertrand, F. Staroz, D. Visvikis, O. Pradier, M. Hatt, and U. Schick, "External Validation of an MRI-Derived Radiomics Model to Predict Biochemical Recurrence after Surgery for High-Risk Prostate Cancer," *Cancers*, vol. 12, p. 814, mar 2020.

[123] C. Dinis Fernandes, C. V. Dinh, I. Walraven, S. W. Heijmink, M. Smolic, J. J. van Griethuysen, R. Simões, A. Losnegård, H. G. van der Poel, F. J. Pos, and U. A. van der Heide, "Biochemical recurrence prediction after radiotherapy for prostate cancer with T2w magnetic resonance imaging radiomic features," *Physics and Imaging in Radiation Oncology*, vol. 7, pp. 9–15, jul 2018.

[124] R. V. Moch H, Humphrey PA, Ulbright TM, *WHO Classification of Tumours of the Urinary System and Male Genital Organs, WHO/IARC Classification of Tumours, 4th Edition, Volume 8*. 4th ed. ed., 2016.

[125] M. Karavitakis, M. Winkler, P. Abel, N. Livni, I. Beckley, and H. U. Ahmed, "Histological characteristics of the index lesion in whole-mount radical prostatectomy specimens: implications for focal therapy," *Prostate Cancer and Prostatic Diseases*, vol. 14, pp. 46–52, mar 2011.

[126] N. Vau, V. Henriques, L. Cheng, A. Blanca, J. Fonseca, R. Montironi, A. Cimadamore, and A. Lopez-Beltran, "Predicting biochemical recurrence after radical prostatectomy: the role of prognostic grade group and index tumor nodule," *Human Pathology*, vol. 93, pp. 6–15, nov 2019.

[127] D. Sugano, A. Sidana, A. L. Jain, B. Calio, S. Gaur, M. Maruf, M. Merino, P. Choyke, B. Turkbey, B. J. Wood, and P. A. Pinto, "Index tumor volume on MRI as a predictor of clinical and pathologic outcomes following radical prostatectomy," *International Urology and Nephrology*, pp. 1–7, may 2019.

[128] J. P. Radtke, C. Schwab, M. B. Wolf, M. T. Freitag, C. D. Alt, C. Kesch, I. V. Popeneciu, C. Huettenbrink, C. Gasch, T. Klein, D. Bonekamp, S. Duensing, W. Roth, S. Schueler, C. Stock, H.-P. Schlemmer, M. Roethke, M. Hohenfellner, and B. A. Hadaschik, "Multiparametric Magnetic Resonance Imaging (MRI) and MRI–Transrectal Ultrasound Fusion Biopsy for Index Tumor Detection: Correlation with Radical Prostatectomy Specimen," *European Urology*, vol. 70, pp. 846–853, nov 2016.

[129] F. Russo, D. Regge, E. Armando, V. Giannini, A. Vignati, S. Mazzetti, M. Manfredi, E. Bollito, L. Correale, and F. Porpiglia, "Detection of prostate cancer index lesions with multiparametric magnetic resonance imaging (mp-MRI) using whole-mount histological sections as the reference standard," *BJU International*, vol. 118, pp. 84–94, jul 2016.

[130] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. W. L. Aerts, A. Dekker, D. Fenstermacher, D. B. Goldgof, L. O. Hall, P. Lambin, Y. Balagurunathan, R. A. Gatenby, and R. J. Gillies, "Radiomics: the process and the challenges," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[131] C. Parmar, J. D. Barry, A. Hosny, J. Quackenbush, and H. J. W. L. Aerts, "Data Analysis Strategies in Medical Imaging.," *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 24, pp. 3492–3499, aug 2018.

[132] R. E. Bellman, *Adaptive Control Processes*, pp. 94, 197. Princeton University Press, 1961.

[133] "Ascent of machine learning in medicine," *Nature Materials*, vol. 18, p. 407, may 2019.

[134] D. Cnns, "Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique," vol. 35, no. 5, pp. 1153–1159, 2016.

[135] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine Learning for Medical Imaging," *Radiographics: a review publication of the Radiological Society of North America, Inc*, vol. 37, no. 2, pp. 505–515, 2017.

[136] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, feb 2017.

[137] C. Zhang, X. Sun, K. Dang, K. Li, X. Guo, J. Chang, Z. Yu, F. Huang, Y. Wu, Z. Liang, Z. Liu, X. Zhang, X. Gao, S. Huang, J. Qin, W. Feng, T. Zhou, Y. Zhang, W. Fang, M. Zhao, X. Yang, Q. Zhou, Y. Wu, and W. Zhong, "Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network," *The Oncologist*, vol. 24, pp. 1159–1165, sep 2019.

[138] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. De Fauw, and S. Shetty, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, pp. 89–94, jan 2020.

[139] T. Hastie, R. Tibshirani, and J. Friedman, pp. 9, 29, 43–69, 101–126, 219–222, 241–244, 649–651, 661–663.

[140] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, pp. 267–288, jan 1996.

[141] S. Klau, V. Jurinovic, R. Hornung, T. Herold, and A.-L. Boulesteix, "Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data," *BMC Bioinformatics*, vol. 19, p. 322, dec 2018.

[142] N. A. Obuchowski, "ROC Analysis," *American Journal of Roentgenology*, vol. 184, pp. 364–372, feb 2005.

[143] S. H. Park and K. Han, "Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction," *Radiology*, vol. 286, pp. 800–809, mar 2018.

[144] Y. Zhao, Z. S.-Y. Wong, and K. L. Tsui, "A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection," *Journal of Healthcare Engineering*, vol. 2018, pp. 1–11, 2018.

[145] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002.

[146] M. Mendonça Silva, "GitHub Repository 'ProBCR'." `https://github.com/mlmms/ProBCR`, 2020.

[147] S. J. Freedland, M. E. Sutter, F. Dorey, and W. J. Aronson, "Defining the ideal cutpoint for determining PSA recurrence after radical prostatectomy. Prostate-specific antigen.," *Urology*, vol. 61, pp. 365–9, feb 2003.

[148] K. Y. E. Aryanto, M. Oudkerk, and P. M. A. van Ooijen, "Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy," *European Radiology*, vol. 25, pp. 3685–3695, dec 2015.

[149] "Horos Project - Free DICOM Medical Image Viewer." `www.horosproject.org`.

[150] M. B. Amin, S. B. Edge, and American Joint Committee on Cancer, *AJCC Cancer Staging Manual.* Springer International Publishing, 8 ed., 2017.

[151] D. Palumbo, B. Yee, P. O'Dea, S. Leedy, S. Viswanath, and A. Madabhushi, "Interplay between bias field correction, intensity standardization, and noise filtering for T2-weighted MRI," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 5080–5083, 2011.

[152] J. D. Gispert, S. Reig, J. Pascau, J. J. Vaquero, P. García-Barreno, and M. Desco, "Method for bias field correction of brain T1-weighted magnetic resonance images minimizing segmentation error," *Human Brain Mapping*, vol. 22, pp. 133–144, jun 2004.

[153] U. Bağci, J. K. Udupa, and L. Bai, "The role of intensity standardization in medical image registration," *Pattern Recognition Letters*, vol. 31, no. 4, pp. 315–323, 2010.

[154] B. Dawant, A. Zijdenbos, and R. Margolin, "Correction of intensity variations in MR images for computer-aided tissue classification," *IEEE Transactions on Medical Imaging*, vol. 12, no. 4, pp. 770–781, 1993.

[155] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: Improved N3 Bias Correction," *IEEE transactions on medical imaging*, vol. 29, pp. 1310–20, jun 2010.

[156] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, and R. Kikinis, "3D Slicer as an image computing platform for the Quantitative Imaging Network," *Magnetic Resonance Imaging*, vol. 30, pp. 1323–1341, nov 2012.

[157] F. P. Oliveira and J. M. R. Tavares, "Medical image registration: a review," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 17, pp. 73–93, jan 2014.

[158] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration," *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 574–582, jun 2016.

[159] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The Design of SimpleITK," *Frontiers in Neuroinformatics*, vol. 7, p. 45, dec 2013.

[160] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, pp. e104–e107, nov 2017.

[161] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative," dec 2016.

[162] L. Torgo, *Data Mining with R, learning with case studies.* Chapman and Hall/CRC, 2010.

[163] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 01 2006.

[164] R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2019.

[165] T. Pohlert, *The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR)*, 2014. R package.

[166] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, dec 1965.

[167] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, vol. 33, pp. 1–22, feb 2010.

[168] R. B. A. Klau, Simon; Hornung, *prioritylasso: Analyzing Multiple Omics Data with an Offset Approach*, 2019. R package version 0.2.2.

[169] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "AUC: a misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography*, vol. 17, pp. 145–151, mar 2008.

[170] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLOS ONE*, vol. 10, p. e0118432, mar 2015.

[171] J. Lever, M. Krzywinski, and N. Altman, "Classification evaluation," *Nature Methods*, vol. 13, pp. 603–604, aug 2016.

[172] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "proc: an open-source package for r and s+ to analyze and compare roc curves," *BMC Bioinformatics*, vol. 12, p. 77, 2011.

[173] C. Thiele and G. Hirschfeld, "cutpointr: Improved Estimation and Validation of Optimal Cutpoints in R," feb 2020.

[174] J. T. Kwak, S. Xu, B. J. Wood, B. Turkbey, P. L. Choyke, P. A. Pinto, S. Wang, and R. M. Summers, "Automated prostate cancer detection using T2-weighted and high-b-value diffusion-weighted magnetic resonance imaging," *Medical Physics*, vol. 42, pp. 2368–2378, may 2015.

[175] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and Reproducibility of Radiomic Features: A Systematic Review," *International Journal of Radiation Oncology Biology Physics*, vol. 102, no. 4, pp. 1143–1158, 2018.

[176] M. Schwier, J. van Griethuysen, M. G. Vangel, S. Pieper, S. Peled, C. Tempany, H. J. W. L. Aerts, R. Kikinis, F. M. Fennessy, and A. Fedorov, "Repeatability of Multiparametric Prostate MRI Radiomics Features," *Scientific Reports*, vol. 9, p. 9441, dec 2019.

[177] S. Van De Geer, "l1-regularization in high-dimensional statistical models," in *Proceedings of the International Congress of Mathematicians 2010, ICM 2010*, pp. 2351–2369, 2010.

[178] H. D. Bondell and B. J. Reich, "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR," *Biometrics*, vol. 64, pp. 115–123, mar 2008.

[179] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," no. April, 1999.

[180] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, pp. 559–572, nov 1901.

[181] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv*, feb 2018.

[182] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[183] G. Singh, F. Memoli, and G. Carlsson, "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition," in *Eurographics Symposium on Point-Based Graphics* (M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, eds.), The Eurographics Association, 2007.

[184] S. F. Crone and S. Finlay, "Instance sampling in credit scoring: An empirical study of sample size and balancing," *International Journal of Forecasting*, vol. 28, pp. 224–238, jan 2012.

[185] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, "Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [Research Frontier]," *IEEE Computational Intelligence Magazine*, vol. 13, pp. 59–76, nov 2018.

[186] P. Kupelian, J. Katcher, H. Levin, C. Zippe, and E. Klein, "Correlation of clinical and pathologic factors with rising prostate-specific antigen profiles after radical prostatectomy alone for clinically localized prostate cancer," *Urology*, vol. 48, pp. 249–260, aug 1996.

[187] S. García-Barreras, I. Nunes, V. Srougi, F. Secin, M. Baghdadi, R. Sánchez-Salas, E. Barret, F. Rozet, M. Galiano, and X. Cathelineau, "Predictors of early, intermediate and late biochemical recurrence after minimally invasive radical prostatectomy in a single-center cohort with a mean follow-up of 8 years," *Actas Urológicas Españolas (English Edition)*, vol. 42, pp. 516–523, oct 2018.

[188] X.-H. Hu, H. Cammann, H.-A. Meyer, K. Jung, H.-B. Lu, N. Leva, A. Magheli, C. Stephan, and J. Busch, "Risk prediction models for biochemical recurrence after radical prostatectomy using prostate-specific antigen and Gleason score.," *Asian journal of andrology*, vol. 16, no. 6, pp. 897–901, 2014.

[189] P. H. Tan, L. Cheng, J. R. Srigley, D. Griffiths, P. A. Humphrey, T. H. Van Der Kwast, R. Montironi, T. M. Wheeler, B. Delahunt, L. Egevad, and J. I. Epstein, "International society of urological pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. Working group 5: Surgical margins," jan 2011.

[190] J. Silberstein and J. Eastham, "Significance and management of positive surgical margins at the time of radical prostatectomy," in *Indian Journal of Urology*, vol. 30, pp. 423–428, Medknow Publications, oct 2014.

[191] I. Evren, A. Haciislamoğlu, M. Ekşi, A. H. Yavuzsan, F. Baytekin, Y. Çolakoğlu, D. Canoğlu, and V. Tugcu, "The impact of single positive surgical margin features on biochemical recurrence after robotic radical prostatectomy," *International Braz J Urol*, vol. 45, pp. 45–53, jan 2019.

[192] K. Yoneda, T. Utsumi, T. Somoto, K. Wakai, R. Oka, T. Endo, M. Yano, N. Kamiya, N. Hiruta, and H. Suzuki, "External validation of two web-based postoperative nomograms predicting the probability of early biochemical recurrence after radical prostatectomy: a retrospective cohort study," *Japanese Journal of Clinical Oncology*, vol. 48, pp. 195–199, feb 2018.

[193] P. Tiwari, S. Viswanath, J. Kurhanewicz, A. Sridhar, and A. Madabhushi, "Multimodal wavelet embedding representation for data combination (MaWERiC): integrating magnetic resonance imaging and spectroscopy for prostate cancer detection.," *NMR in biomedicine*, vol. 25, pp. 607–19, apr 2012.

[194] S. J. Hectors, M. Cherny, K. K. Yadav, A. T. Beksaç, H. Thulasidass, S. Lewis, E. Davicioni, P. Wang, A. K. Tewari, and B. Taouli, "Radiomics Features Measured with Multiparametric Magnetic Resonance Imaging Predict Prostate Cancer Aggressiveness," *Journal of Urology*, vol. 202, pp. 498–505, sep 2019.

[195] V. Giannini, S. Rosati, D. Regge, and G. Balestra, “Texture Features and Artificial Neural Networks: A Way to Improve the Specificity of a CAD System for Multiparametric MR Prostate Cancer,” pp. 296–301, Springer, Cham, 2016.

[196] H. Wang, S. Viswanath, and A. Madabhushi, “Discriminative Scale Learning (DiScrn): Applications to Prostate Cancer Detection from MRI and Needle Biopsies,” *Scientific Reports*, vol. 7, 2017.

[197] G. Nketiah, M. Elschot, E. Kim, J. R. Teruel, T. W. Scheenen, T. F. Bathen, and K. M. Selnæs, “T2-weighted MRI-derived textural features reflect prostate cancer aggressiveness: preliminary results,” *European Radiology*, vol. 27, pp. 3050–3059, jul 2017.

[198] A. Vignati, S. Mazzetti, V. Giannini, F. Russo, E. Bollito, F. Porpiglia, M. Stasi, and D. Regge, “Texture features on T2-weighted magnetic resonance imaging: new potential biomarkers for prostate cancer aggressiveness,” *Physics in Medicine and Biology*, vol. 60, pp. 2685–2701, apr 2015.

[199] P. Kuess, P. Andrzejewski, D. Nilsson, P. Georg, J. Knoth, M. Susani, J. Trygg, T. H. Helbich, S. H. Polanec, D. Georg, and T. Nyholm, “Association between pathology and texture features of multi parametric MRI of the prostate,” *Physics in Medicine & Biology*, vol. 62, pp. 7833–7854, sep 2017.

[200] P. Wikström, J. Marusic, P. Stattin, and A. Bergh, “Low stroma androgen receptor level in normal and tumor prostate tissue is related to poor outcome in prostate cancer patients,” *Prostate*, vol. 69, pp. 799–809, jun 2009.

[201] D. A. Leach, E. F. Need, R. Toivanen, A. P. Trotta, H. M. Palenthorpe, D. J. Tamblyn, T. Kopsaftis, G. M. England, E. Smith, P. A. Drew, C. B. Pinnock, P. Lee, J. Holst, G. P. Risbridger, S. Chopra, D. B. DeFranco, R. A. Taylor, and G. Buchanan, “Stromal androgen receptor regulates the composition of the microenvironment to influence prostate cancer outcome,” *Oncotarget*, vol. 6, no. 18, pp. 16135–16150, 2015.

[202] S. B. Ginsburg, A. Algohary, S. Pahwa, V. Gulani, L. Ponsky, H. J. Aronen, P. J. Boström, M. Böhm, A.-M. Haynes, P. Brenner, W. Delprado, J. Thompson, M. Pulbrock, P. Taimen, R. Villani, P. Stricker, A. R. Rastinehad, I. Jambor, and A. Madabhushi, “Radiomic features for prostate cancer detection on MRI differ between the transition and peripheral zones: Preliminary findings from a multi-institutional study,” *Journal of Magnetic Resonance Imaging : JMRI*, vol. 46, no. 1, pp. 184–193, 2017.

[203] C. Jensen, K. S. Sørensen, C. K. Jørgensen, C. W. Nielsen, P. C. Høy, N. C. Langkilde, and L. R. Østergaard, “Prostate zonal segmentation in 1.5T and 3T T2W MRI using a convolutional neural network.,” *Journal of medical imaging (Bellingham, Wash.)*, vol. 6, p. 014501, jan 2019.

[204] K. Takamatsu, K. Matsumoto, K. Shojo, N. Tanaka, T. Takeda, S. Morita, T. Kosaka, R. Mizuno, T. Shinojima, E. Kikuchi, H. Asanuma, and M. Oya, “The prognostic value of zonal origin and extraprostatic extension of prostate cancer for biochemical recurrence after radical prostatectomy,” *Urologic Oncology: Seminars and Original Investigations*, apr 2019.

[205] C. C. Huang, F. M. Deng, M. X. Kong, Q. Ren, J. Melamed, and M. Zhou, “Re-evaluating the concept of ”dominant/index tumor nodule” in multifocal prostate cancer,” *Virchows Archiv*, vol. 464, pp. 589–594, mar 2014.

[206] H. U. Ahmed, A. El-Shater Bosaily, L. C. Brown, R. Gabe, R. Kaplan, M. K. Parmar, Y. Collaco-Moraes, K. Ward, R. G. Hindley, A. Freeman, A. P. Kirkham, R. Oldroyd, C. Parker, and M. Emberton, “Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study,” *The Lancet*, vol. 389, pp. 815–822, feb 2017.

[207] V. Kasivisvanathan, A. S. Rannikko, M. Borghi, V. Panebianco, L. A. Mynderse, M. H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R. G. Hindley, M. J. Roobol, S. Eggener, M. Ghei, A. Villers, F. Bladou, G. M. Villeirs, J. Virdi, S. Boxler, G. Robert, P. B. Singh, W. Venderink, B. A. Hadaschik, A. Ruffion, J. C. Hu, D. Margolis, S. Crouzet, L. Klotz, S. S. Taneja, P. Pinto, I. Gill, C. Allen, F. Giganti, A. Freeman, S. Morris, S. Punwani, N. R. Williams, C. Brew-Graves, J. Deeks, Y. Takwoingi, M. Emberton, and C. M. Moore, “MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis,” *New England Journal of Medicine*, mar 2018.

[208] W. Huang, Y. Chen, A. Fedorov, X. Li, G. H. Jajamovich, D. I. Malyarenko, M. P. Aryal, P. S. LaViolette, M. J. Oborski, F. O’Sullivan, R. G. Abramson, K. Jafari-Khouzani, A. Afzal, A. Tudorica, B. Moloney, S. N. Gupta, C. Besa, J. Kalpathy-Cramer, J. M. Mountz, C. M. Laymon, M. Muzi, P. E. Kinahan, K. Schmainda, Y. Cao, T. L. Chenevert, B. Taouli, T. E. Yankeelov, F. Fennessy, and X. Li, “The Impact of Arterial Input Function Determination Variations on Prostate Dynamic Contrast-Enhanced Magnetic Resonance Imaging Pharmacokinetic Modeling: A Multicenter Data Analysis Challenge,” *Tomography*, vol. 2, pp. 56–66, mar 2016.

[209] J. Toivonen, I. Montoya Perez, P. Movahedi, H. Merisaari, M. Pesola, P. Taimen, P. J. Boström, J. Pohjankukka, A. Kiviniemi, T. Pahikkala, H. J. Aronen, and I. Jambor, “Radiomics and machine learning of multisequence multiparametric prostate MRI: Towards improved non-invasive prostate cancer characterization,” *PLOS ONE*, vol. 14, p. e0217702, jul 2019.

[210] D. Germanese, E. Bertelli, S. Agostini, L. Mercatelli, S. Colantonio, V. Miele, M. A. Pascali, C. Caudai, N. Zoppetti, R. Carpi, and A. Barucci, "Radiomics to Predict Prostate Cancer Aggressiveness: A Preliminary Study," in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 972–976, IEEE, oct 2019.

[211] B. Varghese, F. Chen, D. Hwang, S. L. Palmer, A. L. De Castro Abreu, O. Ukimura, M. Aron, M. Aron, I. Gill, V. Duddalwar, and G. Pandey, "Objective risk stratification of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images," *Scientific Reports*, vol. 9, p. 1570, dec 2019.

[212] M. Van Smeden, J. A. De Groot, K. G. Moons, G. S. Collins, D. G. Altman, M. J. Eijkemans, and J. B. Reitsma, "No rationale for 1 variable per 10 events criterion for binary logistic regression analysis," *BMC Medical Research Methodology*, vol. 16, pp. 1–12, nov 2016.

[213] M. van Smeden, K. G. Moons, J. A. de Groot, G. S. Collins, D. G. Altman, M. J. Eijkemans, and J. B. Reitsma, "Sample size for binary logistic prediction models: Beyond events per variable criteria," *Statistical Methods in Medical Research*, vol. 28, pp. 2455–2474, aug 2019.

[214] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstem, "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 49, pp. 1373–1379, dec 1996.

[215] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[216] S. Y. Park, Y. T. Oh, D. C. Jung, N. H. Cho, Y. D. Choi, K. H. Rha, and S. J. Hong, "Prediction of biochemical recurrence after radical prostatectomy with PI-RADS version 2 in prostate cancers: initial results," *European Radiology*, vol. 26, pp. 2502–2509, aug 2016.

[217] C. Wei, Y. Zhang, H. Malik, X. Zhang, S. Alqahtani, D. Upreti, M. Szewczyk-Bieda, S. Lang, and G. Nabi, "Prediction of Postprostatectomy Biochemical Recurrence Using Quantitative Ultrasound Shear Wave Elastography Imaging," *Frontiers in Oncology*, vol. 9, p. 572, jul 2019.

[218] C. G. Rogers, M. A. Khan, M. Craig Miller, R. W. Veltri, and A. W. Partin, "Natural history of disease progression in patients who fail to achieve an undetectable prostate-specific antigen level after undergoing radical prostatectomy," *Cancer*, vol. 101, pp. 2549–2556, dec 2004.

[219] Z. Venclovas, M. Jievaltas, and D. Milonas, "Significance of Time Until PSA Recurrence After Radical Prostatectomy Without Neo- or Adjuvant Treatment to Clinical Progression and Cancer-Related Death in High-Risk Prostate Cancer Patients," *Frontiers in Oncology*, vol. 9, p. 1286, nov 2019.

[220] N. Takeuchi, S. Sakamoto, A. Nishiyama, T. Horikoshi, Y. Yamada, J. Iizuka, M. Maimaiti, Y. Imamura, K. Kawamura, T. Imamoto, A. Komiya, Y. Ikehara, K. Akakura, and T. Ichikawa, "Biparametric Prostate Imaging Reporting and Data System version2 and International Society of Urological Pathology Grade Predict Biochemical Recurrence after Radical Prostatectomy.," *Clinical Genitourinary Cancer*, vol. 16, pp. e817–e829, aug 2018.

[221] *Handbook of Medical Image Processing and Analysis*, p. 67. Elsevier, 2009.

[222] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau, "Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review," *Computers in Biology and Medicine*, vol. 60, pp. 8–31, may 2015.

[223] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," pp. 234–241, Springer, Cham, 2015.

[224] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic Routing Between Capsules," tech. rep.

[225] J. H. Thrall, X. Li, Q. Li, C. Cruz, S. Do, K. Dreyer, and J. Brink, "Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success," *Journal of the American College of Radiology*, vol. 15, pp. 504–508, mar 2018.

[226] P. Steenbergen, K. Haustermans, E. Lerut, R. Oyen, L. De Wever, L. Van den Bergh, L. G. Kerkmeijer, F. A. Pameijer, W. B. Veldhuis, J. R. van der Voort van Zyp, F. J. Pos, S. W. Heijmink, R. Kalisvaart, H. J. Teertstra, C. V. Dinh, G. Ghobadi, and U. A. van der Heide, "Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation," *Radiotherapy and Oncology*, vol. 115, pp. 186–190, may 2015.

[227] K. Chow, S. Mangiola, J. Vazirani, J. S. Peters, A. J. Costello, C. M. Hovens, and N. M. Corcoran, "Obesity suppresses tumor attributable PSA, affecting risk categorization," *Endocrine-Related Cancer*, vol. 25, pp. 561–568, may 2018.

[228] C. Zeigler-Johnson, A. Hudson, K. Glanz, E. Spangler, and K. H. Morales, "Performance of prostate cancer recurrence nomograms by obesity status: a retrospective analysis of a radical prostatectomy cohort.," *BMC cancer*, vol. 18, p. 1061, nov 2018.

Appendices

A Classification and Staging Systems

The objective of a tumour classification system is to combine patients with a similar clinical outcome. Here we present 2017 Tumour, Node, Metastasis (TNM) classification for staging of PCa (Table A.1) [?] and the EAU risk group classification (Table 1.3) [43]. The latter classification is based on the grouping of patients with a similar risk of biochemical recurrence after radical prostatectomy (RP) (or external beam radiotherapy).

Clinical T stage only refers to DRE findings; imaging findings are not considered in the TNM classification. Pathological staging (pTNM) is based on histopathological tissue assessment and largely parallels the clinical TNM, except for clinical stage T1c and the T2 substages. All histopathologically confirmed organ-confined PCas after RP are pathological stage T2, as pT2 substages are no longer recognised [?].

Table A.1: Clinical Tumour Node Metastasis (TNM) classification of PCa [**?**].

T - Primary Tumour (stage based on DRE only)		
TX	Primary tumour cannot be assessed	
T0	No evidence of primary tumour	
T1	Clinically inapparent tumour that is not palpable	
	T1a	Tumour incidental histological finding in 5% or less of tissue resected
	T1b	Tumour incidental histological finding in more than 5% of tissue resected
	T1c	Tumour identified by needle biopsy
T2	Tumour that is palpable and confined within the prostate	
	T2a	Tumour involves one half of one lobe or less
	T2b	Tumour involves more than half of one lobe, but not both lobes
	T2c	Tumour involves both lobes
T3	Tumour extends through the prostatic capsule	
	T3a	Extracapsular extension (unilateral or bilateral)
	T3b	Tumour invades seminal vesicle(s)
T4	Tumour is fixed or invades adjacent structures other than seminal vesicles	
N - Regional (pelvic) Lymph Nodes		
NX	Regional lymph nodes cannot be assessed	
N0	No regional lymph node metastasis	
N1	Regional lymph node metastasis	
M - Distant Metastasis		
M0	No distant metastasis	
M1	Distant metastasis	
	M1a	Non-regional lymph node(s)
	M1b	Bone(s)
	M1c	Other site(s)

B Ethics Committee Submission Documents

B.1 Study Summary

Title: Prostate Cancer Biochemical Recurrence Prediction using Machine Learning Analysis of Multiparametric Magnetic Resonance and Histopathology

Acronym: ProBCR

Portuguese Synopsis:
Neste estudo pretende-se utilizar métodos de aprendizagem automática para analisar imagens médicas de ressonância magnética multiparamétrica e de histopatologia digital, de pacientes com cancro da próstata tratados com prostatectomia radical. O objetivo é verificar se estes métodos são capazes de recolher informação das imagens que consiga prever a recorrência bioquímica de cancro da próstata. A capacidade de identificar indivíduos com essas características teria um impacto clínico importante.

English Synopsis:
The purpose of this study is to use machine learning methods to analyse medical images of multiparametric magnetic resonance and digital pathology, of prostate cancer patients treated with radical prostatectomy. It is intended to verify if these methods can extract relevant image information to predict the occurrence of biochemical recurrence of prostate cancer after surgical treatment. The ability to identify individuals with such characteristics would have an important clinical impact.

Type of study:
Pilot retrospective study for patients already treated (with request for Informed Consent waiver).

Involved Institutes/Collaborations:
No other institutes will be involved or collaborations established for this study.

CR/CCC participation units/groups:
CR - Computational Clinical Imaging Group
CCC - Urology Unit
CCC - Radiology Clinical Service
CCC - Pathology Clinical Service

Number of patients:
120-200

Study duration:
1 year

Funding:
No funding required for the execution of this project.

B.2 Research Protocol

Project Title

Prostate Cancer Biochemical Recurrence Prediction using Machine Learning Analysis of Multiparametric Magnetic Resonance and Histopathology

Investigators and Affiliation

Principal Investigator: Nickolas Papanikolaou, PhD[a]

Co-Investigators: Ana Gaivão, MD[b], António Beltran, MD, PhD[b]; Jorge Fonseca, MD[b]

Project Executors: Carolina Seabra, MSc Student[a]; Mónica Silva, MSc Student[a]

[a] Champalimaud Research, Lisbon, Portugal
[b] Champalimaud Foundation, Lisbon, Portugal

1 Background

1.1 Prostate Cancer and Biochemical Recurrence

Malignant neoplasms of the prostate, hereafter referred to as prostate cancer (PCa), usually originate in the glandular tissue. While these cancers are often indolent, there is a subset of men who are diagnosed with highly malignant prostate cancers associated with poor prognosis[1].

The disease poses a substantial public health burden worldwide: PCa is the most frequently diagnosed cancer among men in over-half of the countries in the world, and it is the leading cause of oncological death among men in 46 countries, with nearly 1.3 million new cases and 359000 associated deaths estimated for this year of 2018[2]. Selecting the optimal treatment for each patient is an important aspect in order to improve PCa clinical management.

Currently, PCa early clinical detection depends on the prostate specific antigen (PSA) serum level, as well as on the examination of multiparametric magnetic resonance imaging (mpMRI) that provides information on both morphological and physiological properties of tissues[3]. Nevertheless, its definitive diagnosis is based on histopathologic biopsy verification[4]. The pathologist attributes a prognostic predictor, the Gleason Score. This scoring system, developed in the 1960s, consists of two sub-grades: primary and secondary grades. The former is assigned to the dominant pattern of the tumour or the most common cell morphology, while the latter is assigned to the subordinate pattern. Each of the grades is defined on the scale from 1 to 5, according to the cellular and architectural appearance of recognisable glands, with lower grades corresponding to more normal prostate tissue[5].

In clinical practice, different treatment options are currently available for PCa patients, including active surveillance, adjuvant therapy, radiation therapy and radical prostatectomy (RP). The latter is, yet, the standard first-line curative procedure for the management of localised PCa[4,6–9], given the effectiveness of this therapeutic option for such patients[7,10,11]. It consists of surgically removing the prostate gland, the seminal vesicles, and surrounding tissue sufficient to ensure a negative surgical margins[12].

Besides being measured at the diagnostic stage, PSA serum level, an organ- but not cancer-specific biomarker, is also indispensably assessed in follow-up after a curative treatment, such as RP[13,14]. Its previous elevated value is expected to reach undetectable levels within 4 weeks after RP[15].

However, despite technical improvements in the surgical procedures for PCa treatments, there is a significant risk of cancer recurrence after therapy[7]. Of all patients undergoing RP, between 25% and 35%[7,16–20] present a rising detectable serum PSA level greater than 0,2 ng/ml[21] - a state known as biochemical recurrence (BCR).

Approximately two-thirds of BCRs occur within the first 2 years of surgery[22] and earlier BCR may be associated with increased risk of prostate cancer-specific mortality[23]. BCR is widely used as an end point to assess RP efficacy, and it most likely represents the first sign of progression after surgery[24]. In some cases, its occurrence precedes cancer recurrence, representing, simultaneously, a marker of metastatic progression and PCa-specific mortality[6].

Thus, for men with PCa and candidates to receive curative treatment, such as RP, the risk of development of cancer recurrence after treatment is a main concern. Hence, an early BCR identification and treatment is of paramount importance to improve long-time survival.

1.2 Prediction of Biochemical Recurrence

One of the most acute needs in PCa management, nowadays, is higher precision in prediction of clinical outcomes for more effective decision-making[25].

Nomograms and probability graphs for BCR prediction following surgery have been constructed, such as the D'Amico et al. risk stratification scheme[26], the Stephenson et al. nomogram[27], and the Cancer of the Prostate Risk Assessment (CAPRA) score[28]. All of these models rely on commonly available clinical and histopathological variables, such as preoperative PSA, clinical stage, and biopsy Gleason score[6,29]. Although these nomograms have been internationally validated, only a few of them have predicted the probability of BCR with more than 70% accuracy[30–32].

Recent BCR prediction tools are incorporating mpMRI-derived variables to improve the outcome of the prediction models. This imaging modality has been investigated as promising way not only for detecting and staging the cancer, but also for risk stratification, since it is considered the most sensitive and most specific imaging technique to detect and characterise the clinical aggressiveness of the PCa tumours[7], which is related with BCR occurrence. On the other hand, including histopathologic digital images from the prostate resection might improve the accuracy of the tools which rely on biopsy information, since biopsy is not representative of the whole prostate lesion[33,34].

Given the current used methods, alternative approaches such as logistic regression models, support vector machines, classification and regression trees analysis, and artificial neural networks, are believed to further improve sensitivity, specificity and accuracy of the referred tools[35]. The combination of such computational approaches with data harvested from different imaging modalities might enhance BCR prediction for prostate cancer.

2 Project Aim

This study aims to develop classification models that can predict prostate cancer BCR.

One of the approaches will use imaging features that can be harvested from preoperative mpMRI data, while the other will be based on histopathologic prostate resection digital images. Both approaches will use preoperative clinical variables acquired in the current diagnostic protocol, and implement machine learning methods to predict BCR occurrence in patients who underwent radical prostatectomy for prostate cancer treatment.

For the former, the main objective would be to provide an initial evaluation, before a curative attempt, to patients diagnosed with PCa. Meaning, at the diagnostic time-point, and without the need of performing a biopsy, indirectly evaluate the biological aggressiveness of the patient's tumour - in the sense of relapsing or not after a RP surgery. The latter aims to take advantage of a novel way of analysing pathologic information, which is digital pathology. It presents several functionalities, such as the extraction of quantitative data, resulting from direct measurements like dimensions, number of pixels, or intensities, which cannot be accomplished with optical microscopy.

We hypothesise that there are different mpMR and histopathology digital image characteristics capable of differentiating between BCR-free and BCR patients, after undergoing radical prostatectomy.

3 Materials and Methods

3.1 Study Type

This pilot study will analyse retrospective data: preoperative clinical parameters, standard-of-care mpMR images routinely acquired at the diagnostic phase of prostate cancer management - without any changes to the clinical and imaging protocols -, and digital pathology images acquired from radical resection of the patients with PCa - without any changes to the clinical procedure.

3.2 Study Population

Regarding the histopathological imaging based approach, all patients who underwent radical prostatectomy at Champalimaud Clinical Center between 2016 and 2018 (200 patients) will be considered to be included in this study. As for the mpMR imaging based approach, all patients examined with preoperative prostate mpMRI and treated with radical prostatectomy at Champalimaud Clinical Centre, between 2016 and 2018 (approximately 120-150 patients) will be considered in the inclusion of this study. The imaging protocol used in these patients was the oncological standard of care in the management of prostate cancer.

Patients who received pre- or postoperative hormone or radiation therapy will be excluded from this study.

3.3 Patient Data Collection and Protection

The collection of the retrospective data will be performed by institutional radiologists, urologists and pathologists, reviewing, retrospectively, the clinical information of patients, and selecting the patients to be included on the analysis.

Regarding the mpMRI approach, the researchers and collaborators responsible for the project execution will be responsible for the retrieval of the patients' radiographic images from the institution Picture Archiving and Communication System (PACS) and de- identification of images and databases containing patient identifiable information.

For the histopathologic approach, the researcher responsible for the project execution will be responsible for the retrieval and scanning of the histopathologic digital images and de-identification of images and databases containing patient identifiable information.

For both procedures, a password protected encrypted file with the de-identification correspondence keys will be saved in a password protected workstation located within a room accessible only with the door key.

3.4 Data Processing, Analysis and Model Development

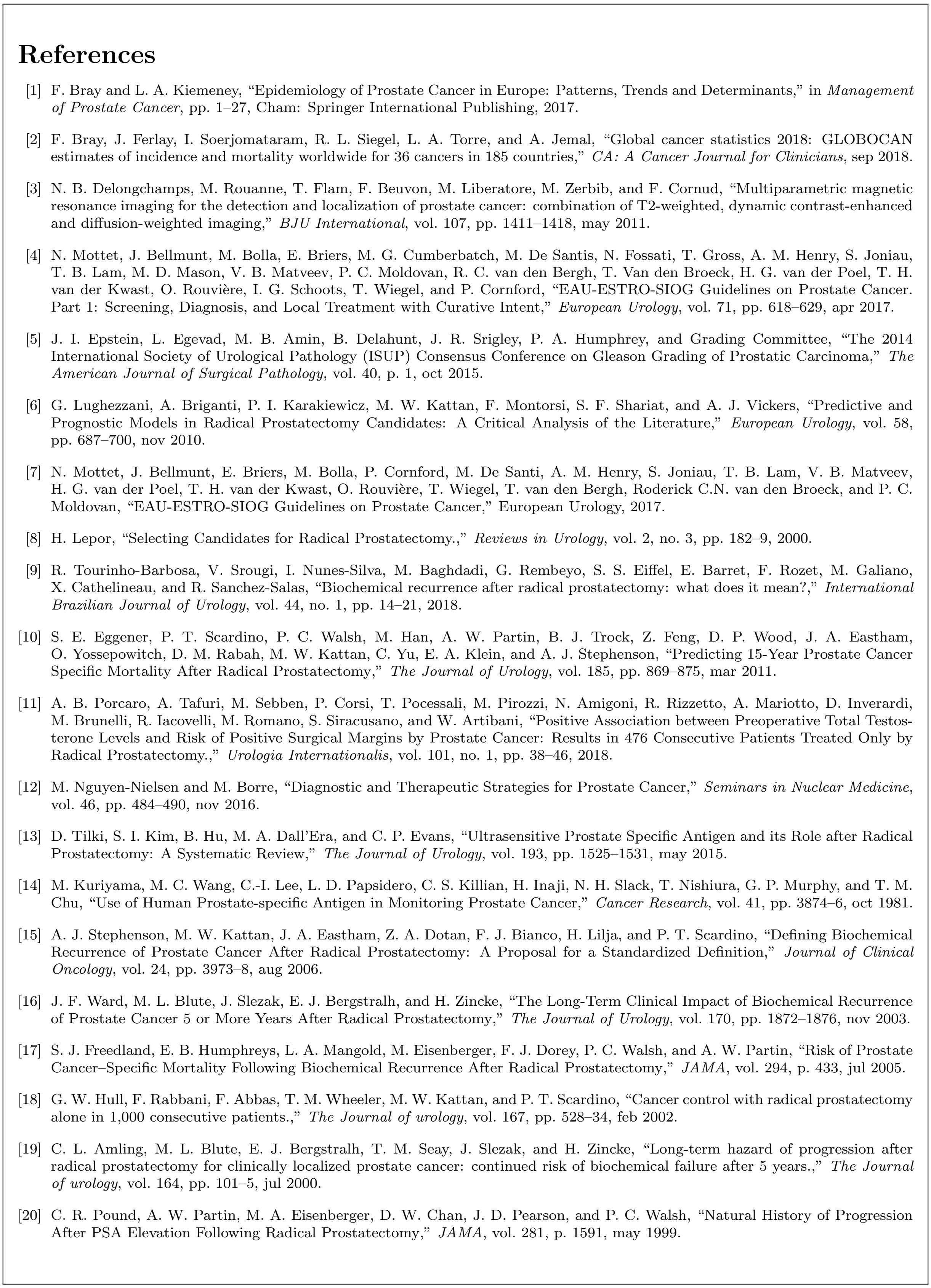
The collected data will be de-identified to ensure patient anonymity.

Imaging preprocessing methods will be explored for the enhancement of the image data. Machine learning algorithms will be applied to the set of data for the processing phase, since it is a powerful technique for recognising patterns on medical images[36]. This will allow the description of visible and invisible textures to the human eye in a numerical and objective way - known as radiomics features[37] -, followed by the model development to classify PCa patients based on their BCR occurrence. Lastly, model evaluation will be based on the outcome prediction performance.

4 Request for waiver of Informed Consent

The development of models to predict BCR will require the analysis of prostatectomy specimens digital images, mpMR images, as well as preoperative clinical information. The results of this study will be stronger by working with a larger number of data, and discarding the patients already imaged or followed in the past would delay the results. We will only use anonymised data from patients already treated, whose retrospective data analysis could benefit the treatment of future patients.

This pilot study might make possible the prediction of BCR, and this could improve the patients' treatment and/or prognosis control.



References

[1] F. Bray and L. A. Kiemeney, “Epidemiology of Prostate Cancer in Europe: Patterns, Trends and Determinants,” in *Management of Prostate Cancer*, pp. 1–27, Cham: Springer International Publishing, 2017.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, sep 2018.

[3] N. B. Delongchamps, M. Rouanne, T. Flam, F. Beuvon, M. Liberatore, M. Zerbib, and F. Cornud, “Multiparametric magnetic resonance imaging for the detection and localization of prostate cancer: combination of T2-weighted, dynamic contrast-enhanced and diffusion-weighted imaging,” *BJU International*, vol. 107, pp. 1411–1418, may 2011.

[4] N. Mottet, J. Bellmunt, M. Bolla, E. Briers, M. G. Cumberbatch, M. De Santis, N. Fossati, T. Gross, A. M. Henry, S. Joniau, T. B. Lam, M. D. Mason, V. B. Matveev, P. C. Moldovan, R. C. van den Bergh, T. Van den Broeck, H. G. van der Poel, T. H. van der Kwast, O. Rouvière, I. G. Schoots, T. Wiegel, and P. Cornford, “EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent,” *European Urology*, vol. 71, pp. 618–629, apr 2017.

[5] J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, P. A. Humphrey, and Grading Committee, “The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma,” *The American Journal of Surgical Pathology*, vol. 40, p. 1, oct 2015.

[6] G. Lughezzani, A. Briganti, P. I. Karakiewicz, M. W. Kattan, F. Montorsi, S. F. Shariat, and A. J. Vickers, “Predictive and Prognostic Models in Radical Prostatectomy Candidates: A Critical Analysis of the Literature,” *European Urology*, vol. 58, pp. 687–700, nov 2010.

[7] N. Mottet, J. Bellmunt, E. Briers, M. Bolla, P. Cornford, M. De Santi, A. M. Henry, S. Joniau, T. B. Lam, V. B. Matveev, H. G. van der Poel, T. H. van der Kwast, O. Rouvière, T. Wiegel, T. van den Bergh, Roderick C.N. van den Broeck, and P. C. Moldovan, “EAU-ESTRO-SIOG Guidelines on Prostate Cancer,” European Urology, 2017.

[8] H. Lepor, “Selecting Candidates for Radical Prostatectomy.,” *Reviews in Urology*, vol. 2, no. 3, pp. 182–9, 2000.

[9] R. Tourinho-Barbosa, V. Srougi, I. Nunes-Silva, M. Baghdadi, G. Rembeyo, S. S. Eiffel, E. Barret, F. Rozet, M. Galiano, X. Cathelineau, and R. Sanchez-Salas, “Biochemical recurrence after radical prostatectomy: what does it mean?,” *International Brazilian Journal of Urology*, vol. 44, no. 1, pp. 14–21, 2018.

[10] S. E. Eggener, P. T. Scardino, P. C. Walsh, M. Han, A. W. Partin, B. J. Trock, Z. Feng, D. P. Wood, J. A. Eastham, O. Yossepowitch, D. M. Rabah, M. W. Kattan, C. Yu, E. A. Klein, and A. J. Stephenson, “Predicting 15-Year Prostate Cancer Specific Mortality After Radical Prostatectomy,” *The Journal of Urology*, vol. 185, pp. 869–875, mar 2011.

[11] A. B. Porcaro, A. Tafuri, M. Sebben, P. Corsi, T. Pocessali, M. Pirozzi, N. Amigoni, R. Rizzetto, A. Mariotto, D. Inverardi, M. Brunelli, R. Iacovelli, M. Romano, S. Siracusano, and W. Artibani, “Positive Association between Preoperative Total Testosterone Levels and Risk of Positive Surgical Margins by Prostate Cancer: Results in 476 Consecutive Patients Treated Only by Radical Prostatectomy.,” *Urologia Internationalis*, vol. 101, no. 1, pp. 38–46, 2018.

[12] M. Nguyen-Nielsen and M. Borre, “Diagnostic and Therapeutic Strategies for Prostate Cancer,” *Seminars in Nuclear Medicine*, vol. 46, pp. 484–490, nov 2016.

[13] D. Tilki, S. I. Kim, B. Hu, M. A. Dall’Era, and C. P. Evans, “Ultrasensitive Prostate Specific Antigen and its Role after Radical Prostatectomy: A Systematic Review,” *The Journal of Urology*, vol. 193, pp. 1525–1531, may 2015.

[14] M. Kuriyama, M. C. Wang, C.-I. Lee, L. D. Papsidero, C. S. Killian, H. Inaji, N. H. Slack, T. Nishiura, G. P. Murphy, and T. M. Chu, “Use of Human Prostate-specific Antigen in Monitoring Prostate Cancer,” *Cancer Research*, vol. 41, pp. 3874–6, oct 1981.

[15] A. J. Stephenson, M. W. Kattan, J. A. Eastham, Z. A. Dotan, F. J. Bianco, H. Lilja, and P. T. Scardino, “Defining Biochemical Recurrence of Prostate Cancer After Radical Prostatectomy: A Proposal for a Standardized Definition,” *Journal of Clinical Oncology*, vol. 24, pp. 3973–8, aug 2006.

[16] J. F. Ward, M. L. Blute, J. Slezak, E. J. Bergstralh, and H. Zincke, “The Long-Term Clinical Impact of Biochemical Recurrence of Prostate Cancer 5 or More Years After Radical Prostatectomy,” *The Journal of Urology*, vol. 170, pp. 1872–1876, nov 2003.

[17] S. J. Freedland, E. B. Humphreys, L. A. Mangold, M. Eisenberger, F. J. Dorey, P. C. Walsh, and A. W. Partin, “Risk of Prostate Cancer–Specific Mortality Following Biochemical Recurrence After Radical Prostatectomy,” *JAMA*, vol. 294, p. 433, jul 2005.

[18] G. W. Hull, F. Rabbani, F. Abbas, T. M. Wheeler, M. W. Kattan, and P. T. Scardino, “Cancer control with radical prostatectomy alone in 1,000 consecutive patients.,” *The Journal of urology*, vol. 167, pp. 528–34, feb 2002.

[19] C. L. Amling, M. L. Blute, E. J. Bergstralh, T. M. Seay, J. Slezak, and H. Zincke, “Long-term hazard of progression after radical prostatectomy for clinically localized prostate cancer: continued risk of biochemical failure after 5 years.,” *The Journal of urology*, vol. 164, pp. 101–5, jul 2000.

[20] C. R. Pound, A. W. Partin, M. A. Eisenberger, D. W. Chan, J. D. Pearson, and P. C. Walsh, “Natural History of Progression After PSA Elevation Following Radical Prostatectomy,” *JAMA*, vol. 281, p. 1591, may 1999.

[21] P. Cornford, J. Bellmunt, M. Bolla, E. Briers, M. De Santis, T. Gross, A. M. Henry, S. Joniau, T. B. Lam, M. D. Mason, H. G. van der Poel, T. H. van der Kwast, O. Rouvière, T. Wiegel, and N. Mottet, "EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer," *European Urology*, vol. 71, pp. 630–642, apr 2017.

[22] J. Walz, F. K.-H. Chun, E. A. Klein, A. Reuther, F. Saad, M. Graefen, H. Huland, and P. I. Karakiewicz, "Nomogram Predicting the Probability of Early Recurrence After Radical Prostatectomy for Prostate Cancer," *The Journal of Urology*, vol. 181, pp. 601–608, feb 2009.

[23] S. J. Freedland, E. B. Humphreys, L. A. Mangold, M. Eisenberger, and A. W. Partin, "Time to Prostate Specific Antigen Recurrence After Radical Prostatectomy and Risk of Prostate Cancer Specific Mortality," *The Journal of Urology*, vol. 176, pp. 1404–1408, oct 2006.

[24] L. Budäus, J. Schiffmann, M. Graefen, H. Huland, P. Tennstedt, A. Siegmann, D. Böhmer, V. Budach, D. Bartkowiak, and T. Wiegel, "Defining biochemical recurrence after radical prostatectomy and timing of early salvage radiotherapy," *Strahlentherapie und Onkologie*, vol. 193, pp. 692–699, sep 2017.

[25] S. F. Shariat, M. W. Kattan, A. J. Vickers, P. I. Karakiewicz, and P. T. Scardino, "Critical review of prostate cancer predictive tools," *Future Oncology*, vol. 5, pp. 1555–1584, dec 2009.

[26] A. V. D'Amico, R. Whittington, S. B. Malkowicz, D. Schultz, K. Blank, G. A. Broderick, J. E. Tomaszewski, A. A. Renshaw, I. Kaplan, C. J. Beard, and A. Wein, "Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer," *JAMA*, vol. 280, p. 969, sep 1998.

[27] A. J. Stephenson, P. T. Scardino, J. A. Eastham, F. J. Bianco, Z. A. Dotan, P. A. Fearn, and M. W. Kattan, "Preoperative Nomogram Predicting the 10-Year Probability of Prostate Cancer Recurrence After Radical Prostatectomy," *JNCI: Journal of the National Cancer Institute*, vol. 98, pp. 715–717, may 2006.

[28] M. R. Cooperberg, D. J. Pasta, E. P. Elkin, M. S. Litwin, D. M. Latini, J. Du Chane, and P. R. Carroll, "The University of California, San Francisco Cancer of the Prostate Risk Assessment score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy.," *The Journal of Urology*, vol. 173, pp. 1938–42, jun 2005.

[29] K. Nishida, S. Yuen, K. Kamoi, K. Yamada, K. Akazawa, H. Ito, K. Okihara, A. Kawauchi, T. Miki, and T. Nishimura, "Incremental value of T2-weighted and diffusion-weighted MRI for prediction of biochemical recurrence after radical prostatectomy in clinically localized prostate cancer," *Acta Radiologica*, vol. 52, pp. 120–126, feb 2011.

[30] V. Poulakis, U. Witzsch, R. de Vries, V. Emmerlich, M. Meves, H.-M. Altmannsberger, and E. Becht, "Preoperative neural network using combined magnetic resonance imaging variables, prostate-specific antigen, and gleason score for predicting prostate cancer biochemical recurrence after radical prostatectomy," *Urology*, vol. 64, pp. 1165–1170, dec 2004.

[31] G. Lughezzani, L. Budäus, H. Isbarn, M. Sun, P. Perrotte, A. Haese, F. K. Chun, T. Schlomm, T. Steuber, H. Heinzer, H. Huland, F. Montorsi, M. Graefen, and P. I. Karakiewicz, "Head-to-Head Comparison of the Three Most Commonly Used Preoperative Models for Prediction of Biochemical Recurrence After Radical Prostatectomy," *European Urology*, vol. 57, pp. 562–568, apr 2010.

[32] J. Morote, J. del Amo, A. Borque, E. Ars, C. Hernández, F. Herranz, A. Arruza, R. Llarena, J. Planas, M. J. Viso, J. Palou, C. X. Raventós, D. Tejedor, M. Artieda, L. Simón, A. Martínez, and L. A. Rioja, "Improved Prediction of Biochemical Recurrence After Radical Prostatectomy by Genetic Polymorphisms," *The Journal of Urology*, vol. 184, pp. 506–511, aug 2010.

[33] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, and N. Rajpoot, "Histopathological Image Analysis: A Review NIH Public Access," *IEEE Rev Biomed Eng*, vol. 2, pp. 147–171, 2009.

[34] J. L. Fine, D. M. Grzybicki, R. Silowash Bs, J. Ho, J. R. Gilbertson, L. A. Ma, R. W. Ma, A. V. Parwani, S. I. Bastacky, J. I. Epstein, and D. M. Jukic, "Evaluation of whole slide image immunohistochemistry interpretation in challenging prostate needle biopsies B," *Human Pathology*, vol. 39, pp. 564–572, 2008.

[35] M. W. Kattan, A. M. Stapleton, T. M. Wheeler, and P. T. Scardino, "Evaluation of a nomogram used to predict the pathologic stage of clinically localized prostate carcinoma.," *Cancer*, vol. 79, pp. 528–37, feb 1997.

[36] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine Learning for Medical Imaging," *Radiographics: a review publication of the Radiological Society of North America, Inc*, vol. 37, no. 2, pp. 505–515, 2017.

[37] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, pp. e104–e107, nov 2017.

B.3 Informed Consent Form

This informed consent form is directed to patients who underwent radical prostatectomy.

The title of the research project is *“Prostate Cancer Biochemical Recurrence Prediction using Machine Learning Analysis of Multiparametric Magnetic Resonance and Histopathology”*.

This is a national project developed by Champalimaud Foundation.

Principal Investigator

Nickolas Papanikolaou

Introduction and Purpose

The Urology Unit and Computational Clinical Imaging group together with the Pathology and Radiology Clinical Service from Champalimaud Foundation, have a research project on biochemical recurrence following radical prostatectomy. The main purpose is to use multiparametric magnetic ressonace (mpMR) imaging and histopathology to predict prostate cancer biochemical recurrence after radical prostatectomy surgical treatment.

After hearing the procedure information, we present a brief description about the study. Please read carefully the following information. After reading, if you have any doubts, do not hesitate to ask them to your physician so they can be clarified.

Type of intervention

This study did not change the clinical and therapeutic procedure previously defined. We solely ask you the permission for the researcher to use your mp-MR and histopathologic images to apply new analysis methods.

Patient selection

All patients with prostate cancer who were examined through mp-MR imaging and underwent radical prostatectomy, at the Champalimaud Foundation may be included in the study. In case you received any type of pre- or postoperative treatment, your enrolment in this study should be disregarded.

Voluntary participation

The participation in this study is voluntary. The patient is free to choose whether or not to participate. The patient’s choice did not change the quality of the treatment he received. In case you want to participate, you are free to dismiss at any moment, needing only to inform your urologist.

Procedure

The participation in this study did not involve additional procedures. The follow-up of the patient was done according with the institution’s standard of care of oncological patients.

Risks

No risks were defined for the participation on this study.

Benefits

The results of this study can potentially improve the treatment of prostate cancer patients who underwent radical prostatectomy.

Compensations
Your participation in this study will not have any additional costs for you. No compensation, either monetary or other, will be given to you for your participation in this study.

Confidentiality
All the data collected during this study will be confidential and protected. In case you participate, a unique identifier will be assigned to you and only your physician will know that this unique identifier corresponds to you.

Sharing of the results
The results of this study, ensuring the correct de-identification, can be shared on scientific meetings and published in scientific journals.

Right to refuse or dismiss
Your participation in this study is your choice. Whether you chose or not to participate in this study the quality and the treatments that you receive will not change. In case you decide to dismiss, you can do it at any moment by contacting your urologist.

Contacts
In case you have any questions in the future, either during or after this study, please contact the urologist in charge:

Dr. **Contact**

____
Name of the patient

______________________________ ____________
Signature of the patient Date (dd/mm/yyyy)

____
Name of the urologist

______________________________ ____________
Signature of the urologist Date (dd/mm/yyyy)

C Study Flowchart

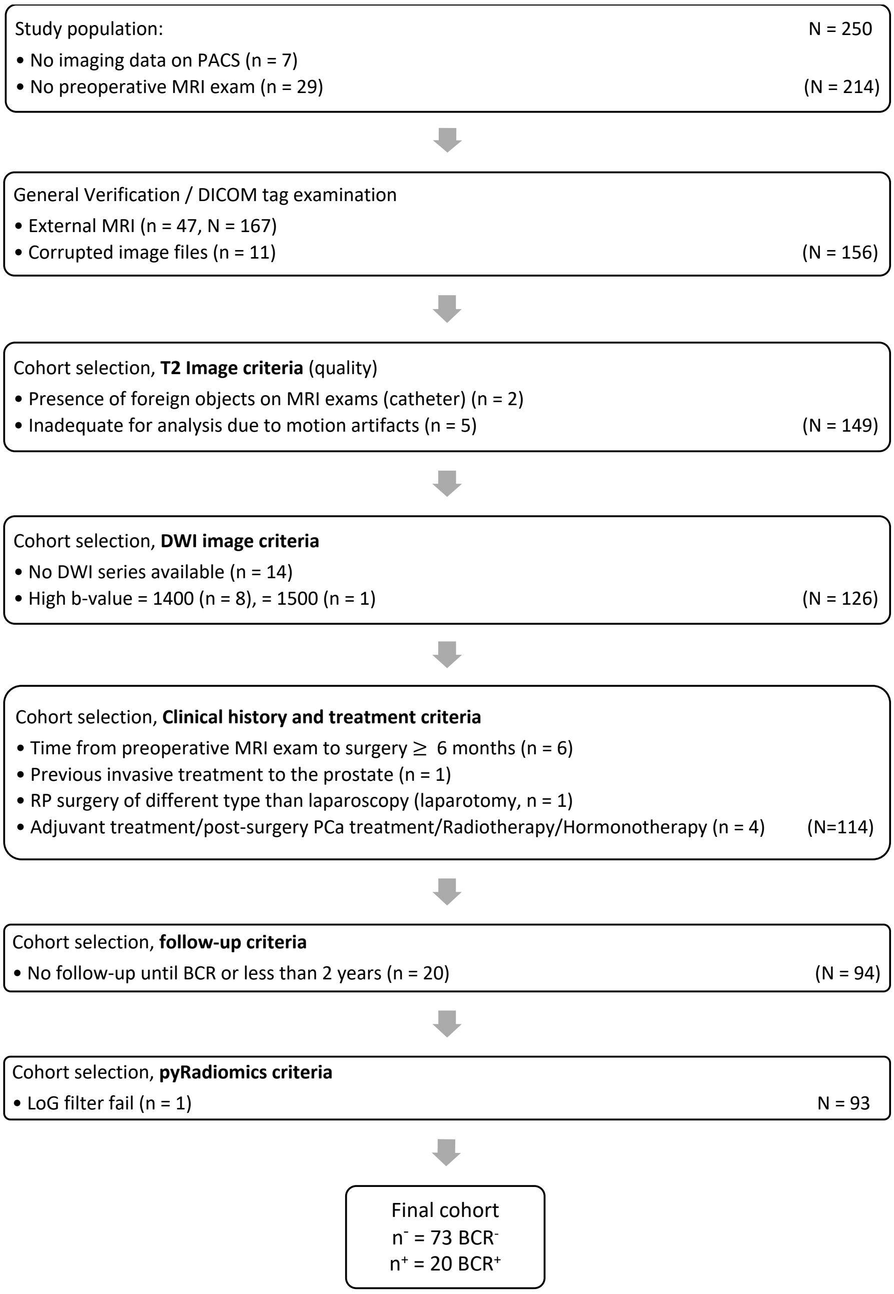

Study population: N = 250
• No imaging data on PACS (n = 7)
• No preoperative MRI exam (n = 29) (N = 214)
General Verification / DICOM tag examination
• External MRI (n = 47, N = 167)
• Corrupted image files (n = 11) (N = 156)
Cohort selection, **T2 Image criteria** (quality)
• Presence of foreign objects on MRI exams (catheter) (n = 2)
• Inadequate for analysis due to motion artifacts (n = 5) (N = 149)
Cohort selection, **DWI image criteria**
• No DWI series available (n = 14)
• High b-value = 1400 (n = 8), = 1500 (n = 1) (N = 126)
Cohort selection, **Clinical history and treatment criteria**
• Time from preoperative MRI exam to surgery ≥ 6 months (n = 6)
• Previous invasive treatment to the prostate (n = 1)
• RP surgery of different type than laparoscopy (laparotomy, n = 1)
• Adjuvant treatment/post-surgery PCa treatment/Radiotherapy/Hormonotherapy (n = 4) (N=114)
Cohort selection, **follow-up criteria**
• No follow-up until BCR or less than 2 years (n = 20) (N = 94)
Cohort selection, **pyRadiomics criteria**
• LoG filter fail (n = 1) N = 93
Final cohort
n^- = 73 BCR^-
n^+ = 20 BCR^+

Figure C.1: A flowchart of our retrospective study design.

D Detailed Methodology

D.1 MR Image Data De-identification

To follow the principle of providing the minimum amount of confidential information (i.e. patient identifiers) necessary to accommodate downstream analysis of imaging data, and to start to create the image data set of study, we performed de-identification to raw DICOM images.

We chose pseudonymisation as the method of de-identification of patient-related information in image DICOM headers as opposed to complete anonymisation. Pseudonymisation is most frequently used in clinical analysis, processing and research since it makes possible to track the real identity of the patient, should it be necessary to inform him or her about additional findings encountered, as mandatory by good clinical practice.

The selection of element tags to be modified or replaced was based considering direct and indirect information fields, as recommended [148]. Direct data directly point to patient identity, whereas indirect data refers to elements containing date, name or time related to data acquisition.

The complete de-identification pipeline involved reading DICOM files with *pydicom* [**?**], accessing the targeted fields with *deid* Python module [**?**] and saving them into a file where the final de-identification correspondence keys would be stored. The targeted fields were handled as configured in a costumed "recipe file". Unique identifiers (UIDs) were generated for each unique patient, and dummy dates, names and times values replaced the appropriate elements.

With this de-identification procedure, we end up with the final image dataset with de-identified metadata and a file with the de-identification correspondence keys. The pipeline and recipe file used can be found in [146].

The de-identification method was also applied on the subsequent created databases, where each UID allowed the linkage between imaging, clinical and histopathological data.

D.2 Characterisation of Axial T2w- and DW-MRI Exams

Characterisation of axial T2w and DW MRI exams and their acquisition parameters, performed with automatic DICOM metadata retrieval is available in Table D.1.

pydicom [**?**] Python package was used to read and to access DICOM file metadata, to apply inclusion and exclusion criteria.

Table D.1: T2w and DW MRI sequences and acquisition parameters of the final image dataset.

Image Modality	#	Model name*	Magnetic Field (T)	b-value (s/mm^2)	Repetition time (ms)	Echo time (ms)	Rows, Columns	Pixel spacing (x/y)	Slice thickness (mm)	Spacing between slices (mm)
T2w	32	Achieva	1.5	NA	[2500, 3245]	[85, 120]	{512, 704, 1024}	[0.176, 0.496]	{3, 4}	4.3
	52	Ingenia	1.5	NA	[2320, 3876]	[85, 140]	{240, 288, 512, 704, 720}	[0.352, 0.667]	{3, 4}	{3, 4.3}
	9	Ingenia	3.0	NA	[3453, 6335]	{125, 135}	{512, 640}	{0.313, 0.544}	3	{3, 3.3}
DW	32	Achieva	1.5	{0, 1000}	[1394, 2000]	80	{192, 224}	{1.429, 1.953}	{4, 6}	{4.5, 6.137, 7, 7.8}
	52	Ingenia	1.5	{0, 1000}	[2000, 5157]	[73, 90]	{144, 224, 256, 288, 320}	[0.695, 1.429]	{3, 4, 5, 6}	{3.3, 4, 5, 6, 7}
	9	Ingenia	3.0	{0, 1000}	[3593, 4128]	[73, 87]	{224, 256}	{0.781, 1.429}	4	{4, 5}

*Philips Healthcare

D.3 Image Preparation

Before initiating any image preprocessing, and having retrieved the necessary metadata from DICOM images headers, we chose to use *dicom2nifti* Python package [**?**] to convert DICOM images to NIfTI files, which eased the following image processing steps. The main difference between the image types is that instead of raw image data being saved as several DICOM files (one per each 2D image slice of the 3D volume), in NIfTI the image is saved as a 3D image. Besides, NIfTI, as a simpler image format, retains only a limited, relevant set of the images' metadata.

DWI image series required a precedent file handling. In a DW sequence, the body volumes being studied are acquired with different b-values. This leads to a final sequence that is four-dimensional, i.e. 3D volume + b-value parameter. Accessing the b-value DICOM attribute, we automatically separated each DWI folder by low- and high-b-value, creating two image volumes, b0 and b1000. After this, each image dicom stack was converted to NIfTI 3D image. The conversion of T2w dicom files was simpler to automatically convert to nifti, given its structure. The code for each case is available in this project's GitHub Repository [146].

D.4 Image Preprocessing

Given this study's goal of quantification of radiologic image characteristics with radiomics analysis, and knowing that correction of artifacts facilitates application of computerised analysis techniques, such as segmentation [152], registration [153] and tissue classification [154], it was mandatory to recognise these common artifacts, and minimise or eliminate them. A description of each category of artifacts and the process of their handling is given.

Bias-Field Correction A bias field is a low frequency smooth undesirable signal that corrupts MR imaging. This artifact, also known as bias, inhomogeneity, illumination nonuniformity, or gain field [**?**] can lead to blurred images (reduction of the high frequency contents of the image such as edges as contours) as well as to inhomogeneities in intensity values [**?**]. It can also lead to signal fall-off, where the same tissue has different grey level distribution across the image (e.g. fat or muscle). These phenomena is known to be caused mainly due to the effects of radio frequency field nonuniformity, patient anatomy, and scanner differences [**?**].

We corrected the bias field distortion of T2w images through the application of N4ITK algorithm [155], a bias field model based correction approach. This algorithm is available to the public through the Insight Toolkit of the National Institutes of Health (ITK). Even though the bias field inhomogeneity on our volume of interest (prostate) was surmountable to a human observer, the whole field-of-view of T2w images were affected by this artifact. Therefore, it was necessary to correct the whole image.

As the application of this algorithm was a compute-intensive task (one image series taking approximately 40 to 60 minutes to be corrected), we used the computational cluster available in Champalimaud Research Centre[1] to remotely perform the job. The job submission required a Python script and a configuration file customising the Python execution environment, where *SimpleITK* [159] package installation was necessary to implement N4ITK bias field filter. Both files can be found in [146].

T2w images were visually and qualitatively evaluated pre- and post-correction to determine if the correction was enough to improve fine anatomic and structural detail within the image and reduce the signal fall-off.

Image Intensity Non-Standardness Intensity non-standardness (drift in image intensities across different acquisitions) arises from the same causes as the bias-field artifact [151]. Usually this is a starting step in image analysis but, given the heterogeneity of field of view (FOV) that images were acquired (see Table D.1), we had to approach this artifact posteriorly (after FOV regularisation (Section D.8.1).

Gaussian Noise Gaussian noise is typically addressed via noise filtering, where the goal is to smooth image intensities within a tissue region while preserving (or accentuating) tissue boundaries [151]. This artifact was minimised during the MR image reconstruction process, therefore it was not a necessary to implement this processing *post hoc* step.

D.5 Segmentation

The most important and challenging problem in image processing nowadays is image segmentation, i.e. separation of structures of interest from the background and/or from each other. After identification of the pixels or voxels that belong to the object, 2D or 3D structure sets can be extracted, forming the region- or volume- of-interest (ROI, VOI, respectively) from which the image analysis will be limited to.

In this study, we aimed at defining prostate boundaries in the different MR images sequences: axial T2w, high-b-value DWI (b = 1000 s/mm^2) and further computed ADC maps.

[1] http://htcondor.champalimaud.pt/

As the whole-prostate segmentation output impacts the subsequent image analysis, we tested and evaluated different types of segmentation methods on our T2w image dataset. T2w imaging provides the best resolution and contrast to show the anatomy of the prostate and has a very high sensitivity for prostate cancer [29, 72, 110].

Specifically, we chose the T2w axial plane for determining the contours of prostate (either straight axial to the patient or in an oblique axial plane matching the long axis of the prostate), as this plane is compulsory to be acquired in a T2w MRI prostate exam and is used for interpretation and recent PI-RADS v2.1 scoring [74].

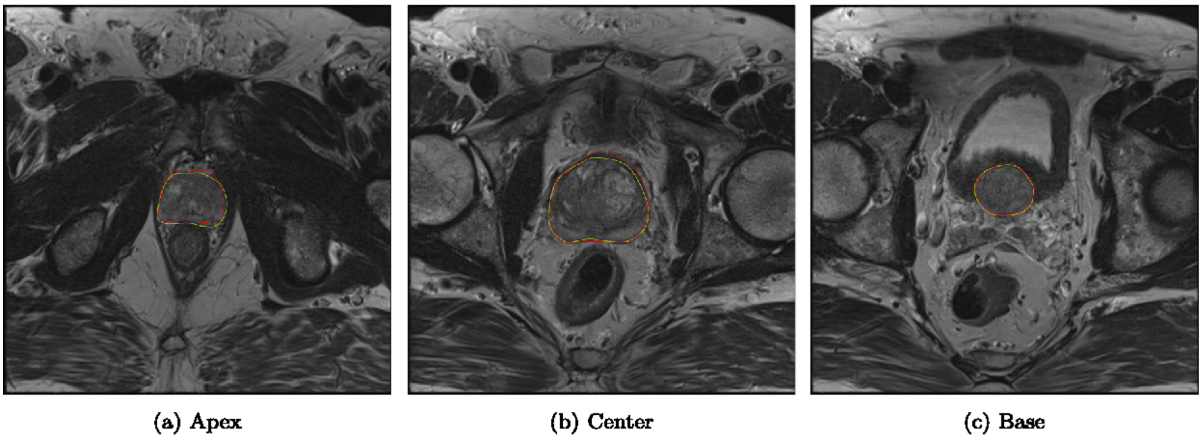
We tested and qualitatively evaluated three types of whole-prostate gland segmentation methods: automatic, semi-automatic, and, finally, manual.

Automatic segmentation methods Automatic methods remove the user interaction in order to fully automate the segmentation process, using information from the image features or previous learned information to segment the prostate. We tested two tools designed for creating contours of the prostate on T2w MR images: DeepInfer [**?**], a tool that deploys a deep learning model on *3D Slicer* [156] medical imaging software; and PelvisML, a machine learning-based module on Microsoft Radiomics App v1.0 [**?**], a software approved by the USA Food & Administration in 2007.

Semi-automatic segmentation methods Semi-automatic methodologies require the user to initiate the segmentation, contouring a subset of image slices or provide “seed points” that can be used by an algorithm to complete the segmentation. Both threshold and region growing filtering available at *3D Slicer* [156] were tested for segmentation of prostate in our T2w set of images. Also, the Active Contour (also known as “Snake”) Segmentation Mode available at ITK Snap [**?**] software application was implemented to segment the prostate gland in our 3D medical images.

The evaluation of the performance of each aforementioned method, whether semi- or automatic, was dependent on the accuracy of segmentation acquired. As these methods performed poorly, we opted to do manual delineation of the prostate.

Manual segmentation method We used axial T2w to tackle this task, as T2w sequence makes possible for a non-expert reader to differentiate the prostate from the surrounding tissues. Training was provided by an experienced and specialised urologic radiologist on how to localise the prostate within the pelvic area, involving the study of the anatomy of the pelvic region and of the prostate on MR imaging. Besides, a comprehensive study of radiological guidelines on how to delineate the prostate on T2w-MRI [**?**], initial supervision and posterior revision of the first set of delineated volumes were ensured.


(a) Apex

(b) Center

(c) Base

Figure D.1: Examples of manually segmented prostate tissue, by the organ's regions: (a) Apex, (b) Center and (c) Base.

Slicer (v. 4.10.1) [156], an open-source software platform for medical imaging processing and three-dimensional visualisation, was used for manual segmentation. We defined the axial T2 sequence as the "master" volume, i.e. the input background volume, from where the final product of the segmentation, a "label map" volume, would be created. The whole prostate was delineated using a small radius paint brush, slice-by-slice, from the base to the apex (Figure D.1). The final "label map" volume was a 3D scalar volume where each voxel contained a number indicating the type of tissue (1: prostate and 0: other).

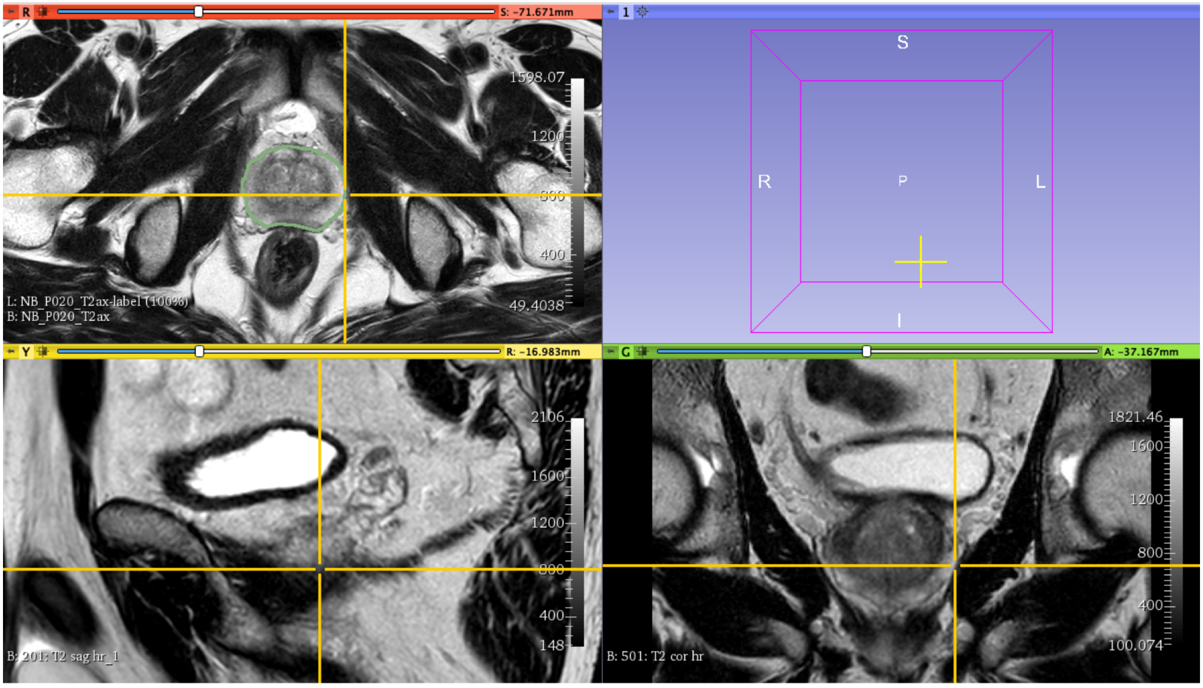


Figure D.2: Axial T2w Prostate segmentation procedure. The pixels from three T2-weighted multiplanar (axial, coronal and sagittal) sequences were cross-correlated with each other by the position of the toggle crosshair (in yellow).

As the prostate segmentation defined the final object for future analysis, great care was taken when drawing each ROI. Specifically, instead of viewing only the axial T2w sequence for navigating this task, we concomitantly visualised the three T2-weighted multiplanar (axial, coronal and sagittal) high-resolution sequences, using a toggle crosshair for precise cross-correlation between them (see Figure D.2). This technique is also used in clinical practice when radiologists aim to acquire more spatial information from the structure. With such three-dimensional cross-visualisation, we aimed at obtaining a higher confidence of whether one or more pixels, in the axial plane, comprised prostate tissue or not.

D.6 Registration

A critical assumption in the implementation of the presented image processing framework was that the prostate volume segmented on T2w imaging could be applied or transferred to remaining image sequences we wished to extract information from: b-1000 DW sequences and ADC maps.

The assessment of differing information conveyed in the MRI sequences (as the high-b-value DWI and further computed ADC maps) required definition of VOIs, or “label map volumes” in each respective image, as performed previously in the axial T2 images with manual segmentation.

However, in DW imaging, both the prostate tissue contrast and image resolution are lower than in T2w, making the manual delineation of the prostate in DWI very challenging for a non-expert, even if training had been provided. Besides, manual delineation would be highly time-consuming and demanding great manual labour, as it was the case of the preceding T2w segmentation procedure. Thus, we aimed at using intermodal image registration to automatically create a “DWI prostate label map” that would target the prostate in both high-b-value DWI and derived ADC maps.

Image registration is an iterative process of aligning an unregistered image (moving image) into a template image (fixed image) via a geometric transformation. It aims at finding the optimal transform that maximizes the similarity between the structures of interest in the input images [157], measured by a similarity metric.

It comprises two steps: (i) image matching, as the process of establishing the correspondences among the structures in input images without explicitly aligning them, i.e. only defining the mathematical transform that links both images, and (ii) image warping, as the application of such (optimised) geometric transformation on an input image, transforming an image and pixels mapped to non-integer points via an interpolator. Image matching is usually preceded by an image pre-alignment step that roughly aligns images by their geometric or gravity centre. This process is illustrated in Figure D.3.

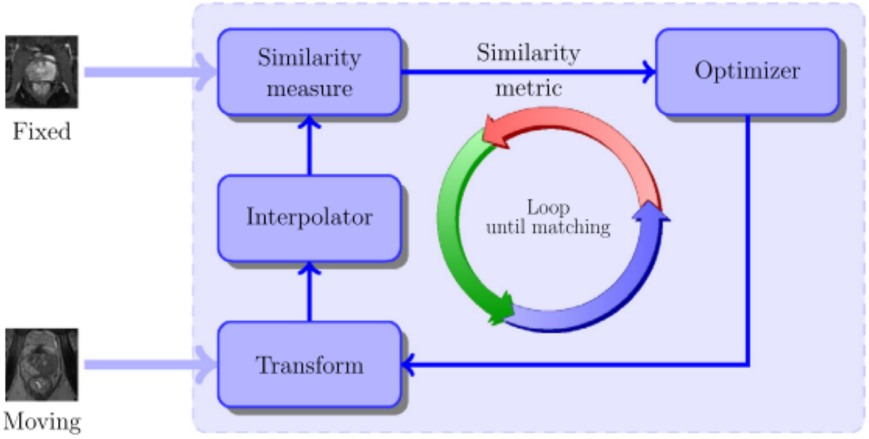


Figure D.3: Typical framework involved in image registration problem solving (from [222]).

Registration algorithms were implemented in Python, using the object-oriented interface from SimpleElastix [158]. SimpleElastix is an extension of *SimpleITK* [159] that uses Elastix [**?**], a modular collection of robust image registration algorithms that is widely used in the literature. The in-house written Python script can be found in [146].

Registration I: Transferring Segmented Volumes from T2w to b0-DWI The first registration scheme aimed at transferring the label map volumes manually delineated in axial T2w into DWI-b0 intrasubject series.

We defined T2w as the moving image and b0 as the fixed one, as it is a good practice to apply image transformations on a higher resolution (T2) onto a lower resolution image (b0), thus not creating artificially interpolated data. Also, a priori, this registration would work well as DW image acquisition with a parameter b = 0 s mm^{-2} is equivalent to a T2w MRI acquisition, however, without lower resolution.

We applied a binary thresholded mask on the fixed image aiming to limit the space of registration to the body tissue and ease algorithm convergence, as recommended in Elastix [**?**] software manual. Thus, we eliminated the background pixels using a threshold of intensity value 50, as Philips MRI systems typically produce background signal intensities below this value.

An intra-subject registration was performed to learn a transformation X between the T2w and b0-DWI, initialising the transform with pre-alignment of images with calculation of the centre of gravity, as it gave better results than the geometric centre. We applied Powell's optimisation method [**?**] for rigid image registration through advanced mattes mutual information metric, as it is the most appropriate for intramodal imaging [157]. The commonly used gradient descent optimisation method was also tested, but the visual alignment results were worst compared to Powell's. As for the resample interpolator, we used the default: B spline, the most popular interpolator for free-form transformations used in medical image registration [157].

With the calculated transform between each axial T2w to b0 DW images, we subsequently warped the binary axial T2 "label map volumes", to obtain a new label that would be valid on the b0 DWI space. We used, instead, the nearest neighbour interpolator, ensuring the created volume label map pixel values were binary: either 0 or 1, with no other values in between.

We finalised this step with image resampling of the new DWI label, using as image reference the b0, for all patients in the study cohort. This ensured the transformed DWI label and the corresponding DW image had the exact same image characteristics, crucial for future image analysis.

Registration II - DWI Motion Correction To correct motion between low- and high-b-value DWI sequences, we performed image registration on the b1000 (moving image) onto b0 (fixed image), by means of a translational transform. We used the centre of gravity for image pre-alignment and the default B-spline interpolator, with maximisation of mutual information similarity metric and Powell's method optimizer.

This registration scheme created a new set of b1000 images (b1000') spatially aligned with b0 images.

Evaluation and Validation The quality of the registration frameworks was assessed by visual inspection of the prostate boundaries, by an experienced observer, overlaying the original input and output images and label maps using *Slicer* [156].

Regarding the first registration scheme, we carefully evaluated and validated each patients' registration outputs, given the intermodality and heterogeneity of the involved sequences. Punctual corrections were necessary and performed manually. As for the second, given its simpler mathematical nature (intramodal and translational), we randomly selected a few patients to evaluate and validate the registration process. Corrections were not necessary.

Combined, both registration procedures allowed for (i) creation of one unique DWI label, valid for the new motion-corrected b1000' images, and (ii) DWI low- and high- b-values series that were intra-aligned. Both outcomes from the registration steps were simultaneously pre-requisites for the following computation of ADC maps, that (optimally) derive from intra-aligned b0 and b1000 DW images.

Having obtained satisfactory prostate alignments with both registration frames, we proceeded to computation of ADC maps, using DWI b0 and the new set of aligned b1000 images (b1000').

D.7 ADC Map Calculation

ADC maps were calculated using the relationship expressed in Equation D.1, on a voxel-by-voxel basis. $S_{0(i,j,k)}$ and $S(b1)_{(i,j,k)}$ respectively represent the signal intensities from the originally acquired low-b-value DWI and from the motion corrected high-b-value DWI at voxel (i,j,k). b_1 is the considered high-b-value, 1000 s mm^{-2}. ln represents the natural logarithm with base e.

$$ADC_{(i,j,k)} = -\frac{\ln\left(\frac{S(b1)_{(i,j,k)}}{S_{0(i,j,k)}}\right)}{b_1} \times 10^6 \tag{D.1}$$

Equation D.1 derives from Nuclear Magnetic Resonance principles and the particular conditions of DW MRI imaging technique, describing how ADC signal can be calculated from two DW images with different b values, one being zero.

We scaled the intensities by a factor of 10^6, so the ADC map appearance was closer to that produced for clinical radiology practice. This scaling does not change the data content of the image. The ADC map computation was performed using an in-house Python script [146].

Equation D.1 is mathematically undefined for the following conditions: $S(b1)_{(i,j,k)} \leq 0 \vee$ $\mathrm{S}_0(i,j,k) \leq 0 \vee (S(b1)_{(i,j,k)} \leq 0 \wedge S_0(i,j,k) \leq 0)$. Given such, we verified that negative or zero intensity values only occurred at the periphery of the image. Thus, we imposed null

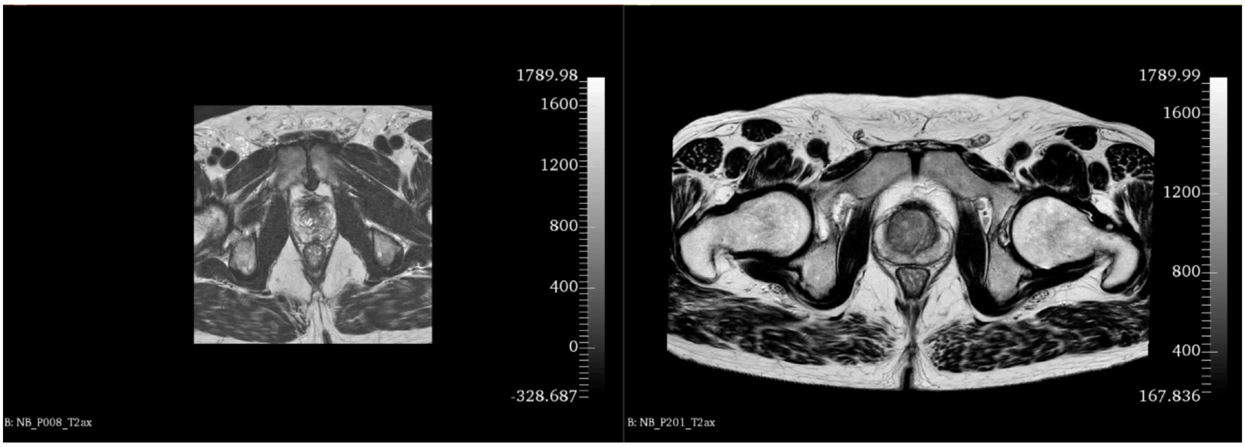


Figure D.4: Examples of two T2w MRI image slices with very different field of views: on the left, a narrow field of view, focusing on the pelvic area; on the right, a very wide field of view, comprising the whole axial body plane.

Table D.2: Range of Field of View (FOV) of the different MRI sequences of the final study cohort.

Image Modality	FOV (minimum – maximum) (mm)
T2w	160 – 350
DWI / ADC	160 – 420

$ADC_{(i,j,k)}$ value in such mathematical undefined conditions, exclusively to allow the full ADC map computation to be accomplished. This did not have quantitative influence in the future analysis.

D.8 Post-Processing

Given the high inter-patient heterogeneity in MRI exams, namely in acquisition parameters, image dimensional characteristics, FOV and used scanners (Table D.1), we implemented image preparation procedures to diminish the variability of image properties encountered and thus enable VOI comparison. This included image registration, cropping, intensity normalisation, discretisation of grey values and resampling.

D.8.1 Image Cropping/FOV Regularisation

Our image dataset was very heterogeneous regarding the axial FOVs size. It contained images acquired with small FOVs, capturing only the pelvic area where prostate resides, and also very wide FOVs, encompassing the whole-body axial plane with a lot of background, and not human tissue (see Figure D.4 and Table D.2).

For each patient, we calculated a bounding-box based on the segmented prostate volume, through the computation of a set of indices that defined it: $L_x, U_x, L_y, U_y, L_z, U_z$, where '$L$' and '$U$' are lower and upper bound, respectively, and 'x', 'y' and 'z' the three image dimensions. The bounding-box's x and y dimensions were determined by the largest prostate area found in the entire segmented volume.

As prostate size and volume vary greatly between PCa patients, the possibility of calculating an "average prostate bounding-box" to fit all patients was excluded. Each individual bounding-

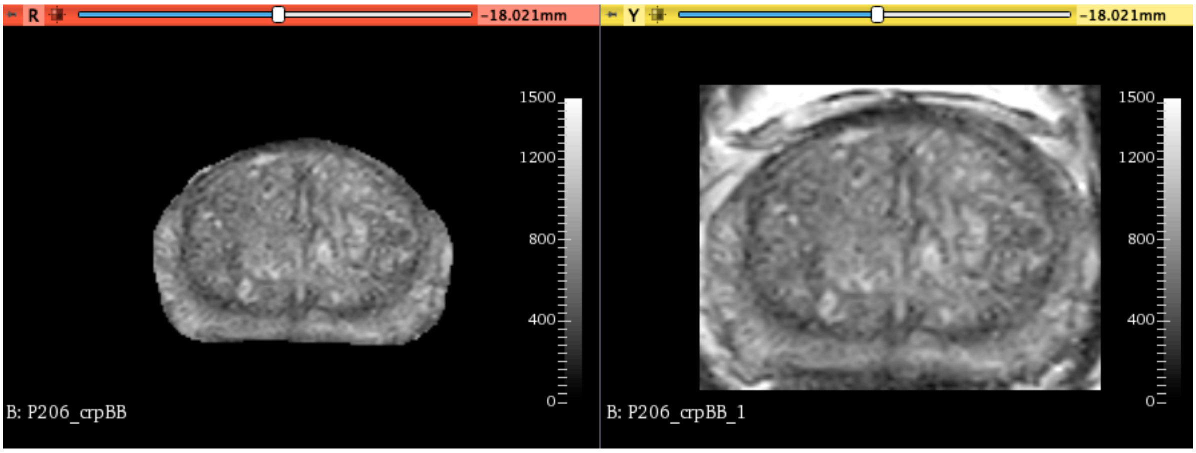


Figure D.5: Prostate volumes cropped with bounding-box (a) without and (b) with surrounding tissues, corresponding to the output of cropping methods (i) and (ii), respectively, further used for normalisation and feature extraction.

box, with a narrow rectangular shape close-fitting the segmented prostate volume, was used for subsequent image cropping of ax-T2w, DW and ADC maps images.

The cropping was performed in two different ways, such that the resulting cropped volume would contain (i) only the segmented prostate, with null background (Fig. D.5 (a)); and (ii) prostate and surrounding tissues (Fig. D.5 (b)).

Cropping method (i) required a preceding step to remove the background, i.e. to null all the external surrounding of the prostate segmentation. We achieved this through pixel-by-pixel multiplication of the original image by its binary label map volume (segmented volume). The cropping method (ii) did not require any previous step.

Thus, with the two cropping methods, we produced two different datasets. Using method (i), we produced "Dataset Prostate-only", a dataset with only the segmented prostate and without surrounding tissues. By means of the cropping method (ii), the produced images contained both the prostate and peripheral tissues inside the bounding-box, creating thus the "Dataset Bounding-box".

Image multiplication, bounding box calculation and cropping was performed in an in-house Python script, requiring both *SimpleITK* interface [159] and *pyRadiomics* package [160].

D.8.2 Image Normalisation

Variability in MR signal intensities between patients occurs during the MRI examinations even using the same scanner, protocol or sequence parameters [**?**]. Hence, normalisation of the MRI data aims at removing the variability between patients, i.e. eliminating the individual effect, leaving the data suitable for VOI comparison.

Normalisation usually takes into account all grey values composing the image, therefore it was mandatory to regularise images dimensions through FOV regularisation prior to this normalisation step.

For both datasets, prostate-only and bounding-box with T2ax and DWI-b1000 FOV regularised volumes, images intensity values were normalised according to:

$$f(x) = \frac{x - \mu_x}{\sigma_x}, \tag{D.2}$$

where x and $f(x)$ are the original and normalised intensities, μ_x and σ_x are the ROI mean and standard deviation of the intensity values. This transformation enforces the image probability density function to have a zero mean and a unit standard deviation.

ADC maps were not normalised, as their signal intensities have physical value, contrary to the two MRI sequences considered.

Thus, we obtained two different image datasets: a dataset normalised after FOV cropping with method I, “Dataset Prostate-only (normalised)”; and a dataset normalised after FOV cropping with method II,“Dataset Bounding-box (normalised)”.

It is worth noting that, given MRI signal intensity non-standardness and the heterogeneity of MRI acquisition parameters present in this dataset, the normalisation was performed to enhance the range of signal intensities and, hypothetically, also enhancing texture. Our aim was not to directly utilise pixel intensity values after such procedure - except on ADC parametric maps, given the characteristics of their signal intensities.

D.8.3 Binning

One of the cores of radiomic analysis is the computation of texture features, for which is essential to apply discretisation of the image intensities into a limited number of grey levels, a process called binning. According to the latest Image Biomarker Standardisation Initiative guidelines [161], binning allows for differing ranges in intensity in ROIs, while still keeping the texture features informative and comparable between them. Given the misalignment between intensity distributions across different MRI acquisitions, binning was crucial for further radiomic feature extraction.

Two approaches to discretisation are commonly used. One involves the discretisation to a fixed number of bins, and the other discretisation with a fixed bin width. Following the recommendation for image binning by *pyRadiomics* [160] (the Python package used for the subsequent extraction of Radiomic features), we aimed at fixing a bin width so that the resulting number of bins ranged between 30 and 130 bins.

Given that there are no specific guidelines from literature as to what constitutes an optimal bin width for a given medical imaging modality [160], we selected a fixed bin width h for each image series or maps based on the computation of the intensity distribution of each series, according to:

$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil, \tag{D.3}$$

where the number of bins, k, can be calculated from bin width h, with x as the image intensity value. k is obtained by rounding to the next nearest integer (ceiling function). Using *Minimum-MaximumImageFilter* (*SimpleITK* class), we collected the minimum and maximum intensity values for each image to obtain the intensity range (max x − min x). We then divided it by $k = 80$ (central value of the suggested number of bins: from 30 to 130 [160, 161]), subsequently getting the h value for such image. Having retrieved all the h values for each image of each modality type, we calculated the average h that, for all images from a certain modality, would ensure the number of bins $k \in [30, 130]$, the suggested interval.

The second method of binning, where a fixed bin count for an image is fixed, was discarded. If we had fixed a bin count for each image modality type, the resulting bin width would vary immensely for each image, as it was dependent on the image intensity range. However, despite

Table D.3: Calculated bin widths, according to the image modality and normalisation method, and final pixel spacing.

Image Modality	Bin Width normalisation method prostate-only	Bin Width normalisation method bounding-box	Final pixel spacing (mm)
T2w	0.0686	0.0762	0.66 × 0.66 × 4.5
DWI (b1000)	0.0482	0.0798	1.95 × 1.95 × 7
ADC	39.2684	56.9288	1.95 × 1.95 × 7

an individual image having higher or lower intensity range (defined as max x − min x), we instead calculated the average bin width h for this dataset that would produce a number of bins between 30 and 130 for all images for that image modality, as recommended [160]. Also, the fixed bin size method has the advantage of maintaining a direct relationship with the original intensity scale.

D.8.4 Image Resampling

For a consistent calculation of quantitative features, it is necessary to homogenise image resolutions, ensuring all had the same voxel properties. Thus, all T2w images were resampled to a grid of 0.66 x 0.66 x 4.5 mm voxels and 1.95 x 1.95 x 7 mm voxels for both DWI-b1000 and ADC maps, using BSplines (Table D.3). We chose to resample all images to the lowest resolution found for each modality type in our study cohort, thus avoiding image upsampling, which requires creation of artificial data. Pixel spacing of all images was automatically retrieved from image metadata, for each modality type (code in [146]).

D.9 Radiomic Feature Extraction

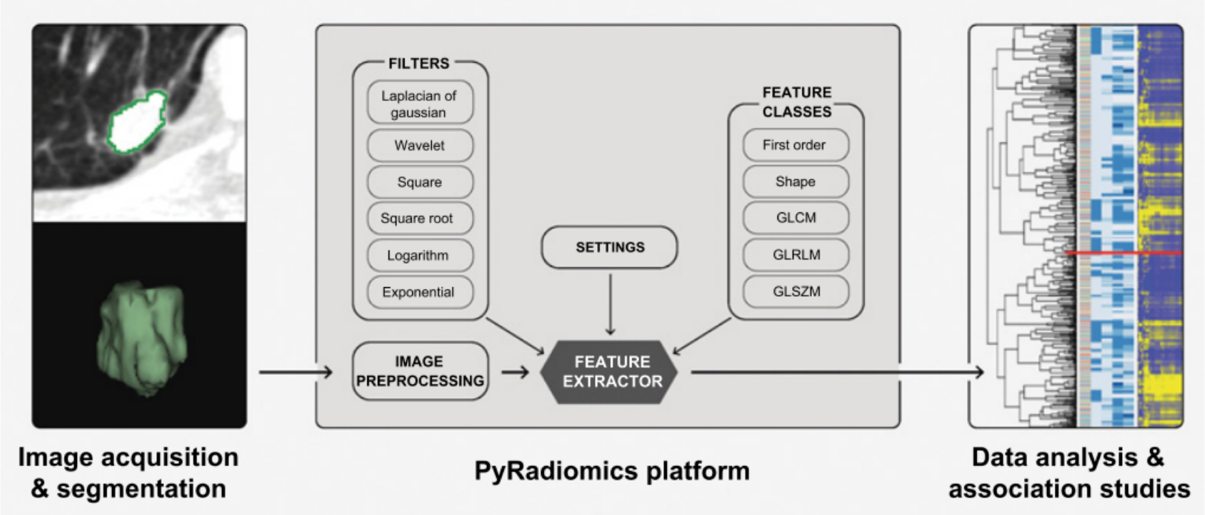


Figure D.6: Illustration of the process of PyRadiomics. First, medical images are segmented. Second, features are extracted using the PyRadiomics platform, and third, features are analysed for associations with clinical or biologic factors.

2D Radiomic analysis was conducted on the segmented volume labels of axial dimension of T2w, high-b-value DWI and ADC maps. We used *PyRadiomics* 3.0 [160], a platform available as

Table D.4: Characteristics of the datasets used for model building: types of features and total number of variables.

Clinical Scenario	Feature type				Total number of variables
	Clinicohistopathological	Radiomics			
		T2	DWI b1000	ADC	
Pre-biopsy	3	839	839	1037	2718
Preoperative	11	839	839	1037	2726
Postoperative	33	839	839	1037	2748

an open-source Python package (illustrated in Figure D.6). PyRadiomics extracts region-wise engineered features from 2D images (from a 3D volume) with their respective binary masks. This segment-based extraction computes single values per feature for a ROI.

We extracted features from (I) original, and derived images, namely (II) Laplacian of Gaussian (LoG), with kernel sizes of sigma $\sigma = \{1, 2, 3, 4, 5\}$ mm, (III) Wavelet decompositions yielding 4 derived images from all possible combinations of either a High and/or Low pass filters, (IV) Gradient, returning its magnitude (directional change in intensity), and (V) Local Binary Pattern in a by-slice operation, representing local texture. This led to a total of 11 different representations of the original image, following recommendations from the studies of Schwier [176] and Wang et al. [**?**].

The extracted features for each image can be categorised into three groups: (I) organ intensity (only extracted for ADC maps), (II) shape-based and (III) texture features. The first group quantified prostate intensity characteristics using first-order statistics, calculated from the histogram of all prostate voxel intensity values. Group 2 consists of features based on the shape of the organ (for example, sphericity or compactness). Group 3 consists of textural features that are able to quantify tissue heterogeneity differences in the texture that is observable within the prostate volume. These features are calculated in all three- dimensional directions within the prostate volume, thereby taking the spatial location of each voxel compared with the surrounding voxels into account.

Particularly for the third group, the textural features derived from the symmetrical GLCM (grey-level co-occurrence matrix), GLDM (grey-level dependence matrix) and NGTDM (neighbouring grey tone difference matrix), computed with distances of $d = \{1, 2, 3\}$ voxels between the centre voxel and the neighbour. Rotational invariant textural features derived from GLRLM (grey-level run length matrix) and GLSZM (grey-level size zone matrix) were computed with a fixed distance of 1 to define neighbours.

A total of 839 for T2w, 839 for high-b-value DWI and 1037 for ADC maps region-level features were extracted for each patient.

The code and structured parameter files used can be found in [146]. Concepts of radiomics features and their mathematical definition can be found in detail in [**?**].

D.10 Modelling

D.10.1 Data Preparation

For image types T2w, b1000, and ADC, a corresponding data set of 93 data points with radiomic and clinicohistopathologic variables were used to build models for predicting BCR

Table D.5: Clinicohistopathological variables included in each generated dataset: pre-biopsy, preoperative and postoperative, according to their clinical availability at each corresponding clinical set-up.

Dataset	Variable type	Variables
Pre-biopsy	Clinical	Baseline PSA level; PSA > 10; 10–20; > 20
Preoperative	Clinical	Baseline PSA level; PSA > 10; 10–20; > 20
	Histopathological (Biopsy)	Highest GS (sum; 1; 2)
	Other	Adapted EAU BCR risk group (low, intermediate, high); Group grade risk stratification (low, medium, high)
Postoperative	Clinical	Baseline PSA level; PSA > 10; 10–20; > 20
	Histopathological (Biopsy)	Highest GS (sum; 1; 2)
	Other	Adapted EAU BCR risk group (low, intermediate, high); Group grade risk stratification (low, medium, high)
	Treatment	Surgery type (non-robotic; robotic); Treatment type (RP; RP with lymphadenectomy)
	Histopathological (Tumour index)	Grade group (GS sum; 1; 2); Linear extension; Volume

	Histopathological (Specimen)	Perineural invasion; Specimen Linfovascular invasion;
		EPE (status; extension); Resection margin (status; extension);
		Seminal vesicle invasion; pT status; pN status;
		Histological subtype (acinar; mixed);
		Grade group (GS sum; 1; 2)

	Follow-up	2-yr BCR status

occurrence within two years, after radical prostatectomy treatment.

The dataset was transformed into three datasets named (1) Pre-biopsy, (2) Preoperative and (3) Postoperative, where the inclusion of clinical and/or histopathological variables varied according to their clinical availability at each corresponding set-up (Tables D.4 and D.5). Radiomic features were present in all three clinical settings, as PCa imaging examination is performed before biopsy.

The data points of each generated dataset were divided into two groups by their ground-truth or class, BCR negative and BCR positive (0 and 1).

Observations were randomly shuffled and then randomly split into train and test sets, in the ratio of 70:30. The training/validation dataset was used for parameter tuning, feature selection and error estimation using cross-validation, whereas the test dataset was only used at the end, to access models' generalisation capabilities.

Clinicohistopathological categorical variables were one-hot encoded. Continuous variables from the training dataset were standardised, being centered by subtraction of the mean and scaled by division by the standard deviation. The standardisation was not mandatory to run the method, although it is recommended for the purpose of interpretability and especially important when dealing with variables measured at different scales. The calculated standardisation transform for the train set was then applied to the test set, avoiding any data leakage of the test set into the training set standardisation procedure, ensuring that the test set remained independent in the model building process.

For data filtering, we search for predictors with zero- or near-zero-variance to be deleted from the training dataset. None were detected nor removed.

We aimed at evaluating the predictive performance of developed models for the three clinical scenarios specified, intending to analyse the presence of prognostic potential at each clinical phase. Therefore, three different datasets were generated, and the following steps described were performed to each dataset and normalisation method ("prostate-only" and "bounding-box"). On the final stage of this workflow, we used a resampling technique to deal with class unbalance, generating a randomly-oversampled dataset.