# Computer-aided diagnosis of prostate cancer with multiparametric MRI

Doctoral Thesis

University of Minnesota

**Ethan Yize Leng**

*Submitted in partial fulfillment of the requirements*
*for the degree of Doctor of Philosophy*

Adviser: Dr. Gregory Metzger

July 2020

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| **Clinical terms** | |
|---|---|
| **AFS** | **A**nterior **F**ibromuscular **S**troma |
| **AS** | **A**ctive **S**urveillance |
| **BPH** | **B**enign **P**rostatic **H**yperplasia |
| **CAD** | **C**omputer **A**ided **D**etection/**D**iagnosis |
| **CADe** | **C**omputer **A**ided **D**etection |
| **CADx** | **C**omputer **A**ided **D**iagnosis |
| **CG** | **C**entral **G**land (prostate) |
| **EPE** | **E**xtra**P**rostatic **E**xtension |
| **GG** | (Gleason) **G**rade **G**roup |
| **GS** | **G**leason **S**core |
| **HGPIN** | **H**igh-**g**rade **P**rostatic **I**ntraepithelial **N**eoplasia |
| **PCa** | **P**rostate **Ca**ncer |
| **PI-RADS** | **P**rostate **I**maging-**R**eporting **a**nd **D**ata **S**ystem |
| **PSA** | **P**rostate **S**pecific **A**ntigen |
| **SVI** | **S**eminal **V**esicle **I**nvasion |
| **PIN** | **P**rostatic **I**ntraepithelial **N**eoplasia |
| **PZ** | **P**eripheral **Z**one (prostate) |
| **RP** | **R**adical **P**rostatectomy |
| **TRUS** | **T**rans**R**ectal **U**ltra**S**ound |
| **TZ** | **T**ransition **Z**one (prostate) |

**Imaging terms**

| | |
|---|---|
| **2D** | **T**wo-**D**imensional |
| **3D** | **T**hree-**D**imensional |
| **ADC** | **A**pparent **D**iffusion **C**oefficient |
| **BW** | **B**and**W**idth |
| **CHESS** | **CHE**mical-shift **S**elective **S**aturation |
| **CPMG** | **C**arr-**P**urcell-**M**eiboom-**G**ill |
| **DCE** | **D**ynamic **C**ontrast **E**nhanced |
| **DWI** | **D**iffusion **W**eighted **I**maging |
| **EPI** | **E**cho **P**lanar **I**maging |
| **ERC** | **E**ndo**R**ectal **C**oil |
| **ETL** | **E**cho **T**rain **L**ength |
| **FID** | **F**ree **I**nducation **D**ecay |
| **FOV** | **F**ield **o**f **V**iew |
| **FSE/TSE** | **F**ast/**T**urbo **S**pin **E**cho |
| **GBCA** | **G**adolinium-**B**ased **C**ontrast **A**gent |
| **MRI** | **M**agnetic **R**esonance **I**maging |
| **mpMRI** | **m**ulti**p**arametric **MRI** |
| **PI** | **P**arallel **I**maging |
| **PDW** | **P**roton **D**ensity-**W**eighted |
| **qMRI** | **q**uantitative **MRI** |
| **RF** | **R**adio**F**requency |
| **SAC** | **S**urface **A**rray **C**oil |
| **SAR** | **S**pecific **A**bsorption **R**ate |
| **SNR** | **S**ignal-to-**N**oise **R**atio |
| **SSFP** | **S**teady-**S**tate **F**ree **P**recession |
| **SPGR** | **SP**oiled **GR**adient-echo |
| **T1W** | **T1**-**W**eighted |

| | |
|---|---|
| **T2W** | **T2-W**eighted |
| **TE** | **E**cho **T**ime |
| **TR** | **R**epetition **T**ime |

**Mathematical and statistical terms**

| | |
|---|---|
| **AUC** | **A**rea **U**nder (the ROC) **C**urve |
| **CI** | **C**onfidence **I**nterval |
| **CNN** | **C**onvolutional **N**eural **N**etwork |
| **DSC** | **D**ice **S**imilarity **C**oefficient |
| **GA** | **G**enetic **A**lgorithm |
| **MI** | **M**utual **I**nformation |
| **NMI** | **Normalized M**utual **I**nformation |
| **NPV** | **N**egative **P**redictive **V**alue |
| **PPV** | **P**ositive **P**redictive **V**alue |
| **ROC** | **R**eceiver **O**perating **C**haracteristic |
| **RMSE** | **Root M**ean **S**quared **E**rror |
| **SVMs** | **S**upport **V**ector **M**achines |

**Other terms**

| | |
|---|---|
| **AIF** | **A**rterial **I**nput **F**unction |
| **AMACR** | $\alpha$-**M**ethyl**A**cyl **C**oA **R**acemase |
| **DOF** | **D**egree **o**f **F**reedom |
| **EES** | **E**xtracellular **E**xtravascular **S**pace |
| **H&E** | **H**ematoxylin & **E**osin |
| **HMWCK** | **H**igh-**M**olecular **W**eight **C**yto**K**eratin |
| **IHC** | **I**mmuno**H**isto**C**hemistry |
| **OD** | **O**ptical **D**ensity |
| **ROI** | **R**egion **o**f **I**nterest |

| | |
|---|---|
| **VOI** | **V**olume **o**f **I**nterest |
| **WSI** | **W**hole **S**lide **I**mage |

# Chapter 1

# Introduction

## 1.1 Prostate anatomy and histology

The prostate is an organ in men that is situated inferior to the bladder and encircles the proximal urethra and ejaculatory duct. From superior to inferior, it can be grossly divided into the base, the midgland, and the apex. It can also be divided into four distinct zones: the peripheral zone (PZ), central zone (CZ), transition zone (TZ), and the anterior fibromuscular stroma (AFS) (Fig.1.1); the combination of the CZ and TZ plus the prostatic urethra is often referred to as the central gland (CG). The prostate is bound not by a true capsule, but rather by a band of fibromuscular tissue that is often referred to as the pseudocapsule, or even just capsule for short.[1,2]

Histologically, the prostate is composed of glands that are responsible for producing prostatic fluid. The glands normally have two layers of epithelial cells: an outer (basal) layer and an inner secretory layer closer to the gland lumen (Fig.1.2).

## 1.2 Prostate cancer

### 1.2.1 Epidemiology

PCa is the 2nd leading cause of cancer death among men in the U.S. It is estimated that 1 in 7 men will be diagnosed with PCa in his lifetime, and 1 in 34 men will die of the

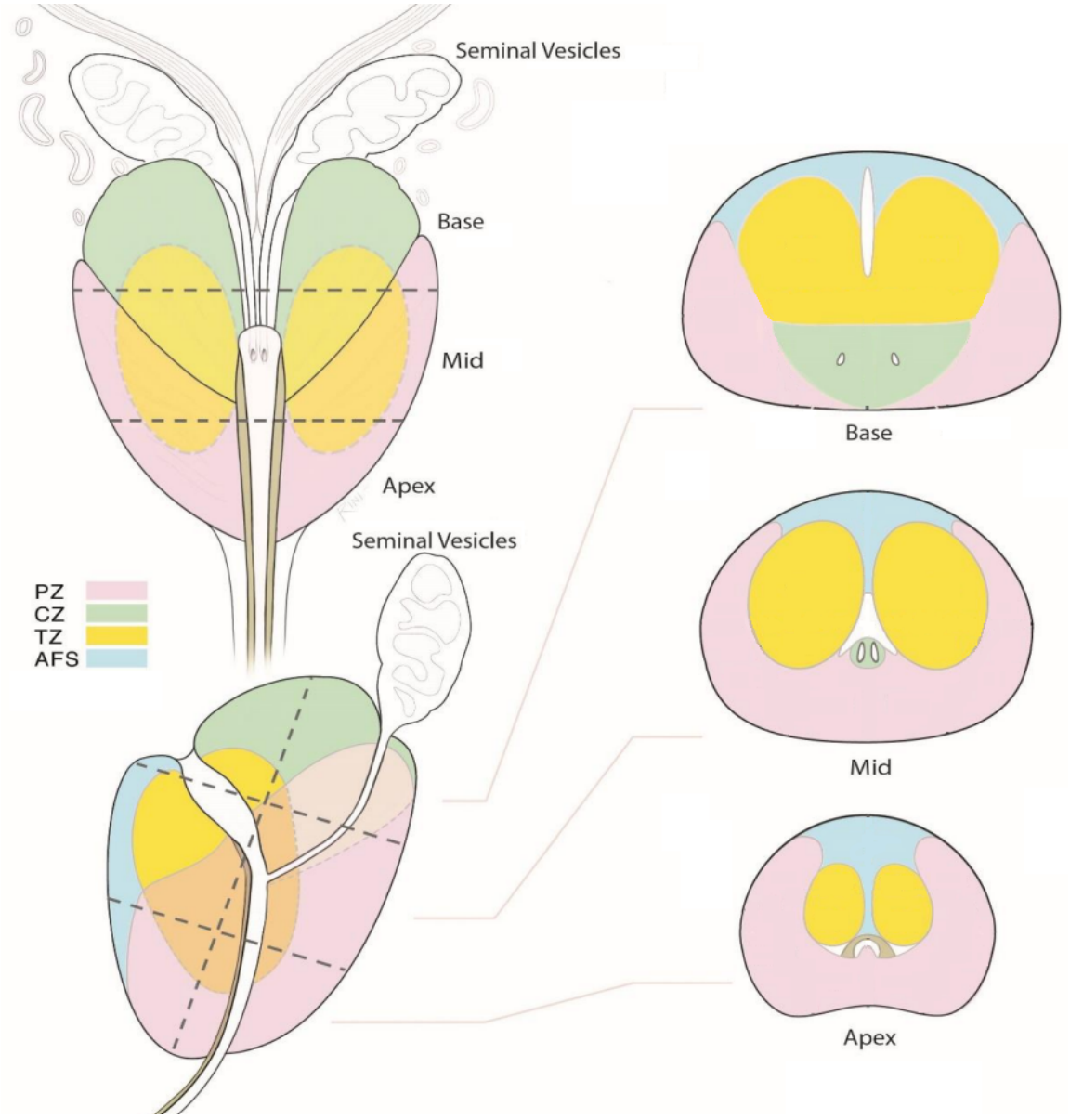



FIGURE 1.1: Coronal (left-top), sagittal (left-bottom), and axial (right) views of the prostate with the zonal anatomy labeled.[2]

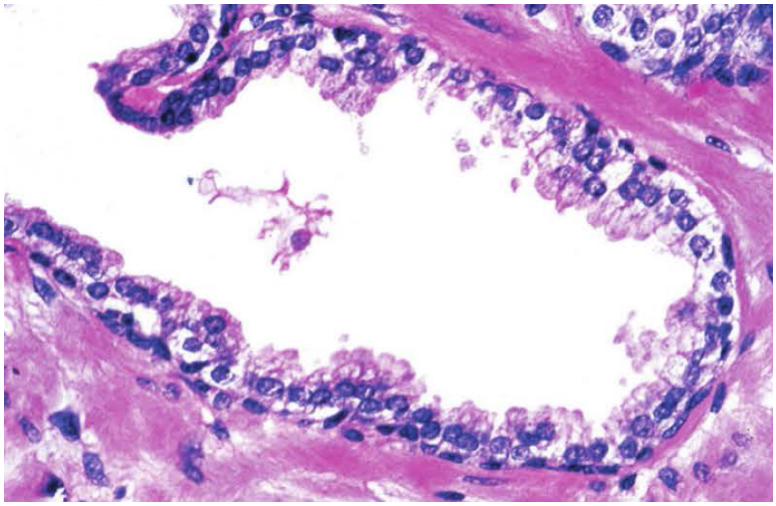


FIGURE 1.2: A single normal prostate gland (upper left) with both epithelial layers in clear view.[1]

disease.[3] PCa is generally an indolent disease, and data suggest that most men with PCa die of other causes before the disease ever becomes clinically advanced.[4] Indeed, with appropriate management, the five- and ten-year survival rates are nearly 100% for men with localized disease. At the same time, it is important to identify the rare cases of rapidly-progressing, aggressive cancers, as the five-year survival rate drops to 31% once distant metastases are found.[3]

### 1.2.2 Histopathology

The widely-differing degrees of cancer aggressiveness that is observed clinically can be explained by the pathologically complex nature of PCa. It tends to develop multifocally within the organ, and it is estimated that 60–90% of prostates contain two or more foci of cancer at time of clinical diagnosis. There is also significant heterogeneity in the synchronicity, genetic alterations, and progression of the multiple foci of disease.[5–7]

Prostate cancers are predominantly adenocarcinomas that arise from the prostate glands. Histologically, PCa is defined simply by the presence of glands without the outer basal cell layer (Fig.1.3). As the PZ contains the highest concentration of glands, it is unsurprisingly the origin of 70-80% of all diagnosed cases of prostate cancer.[1,8]

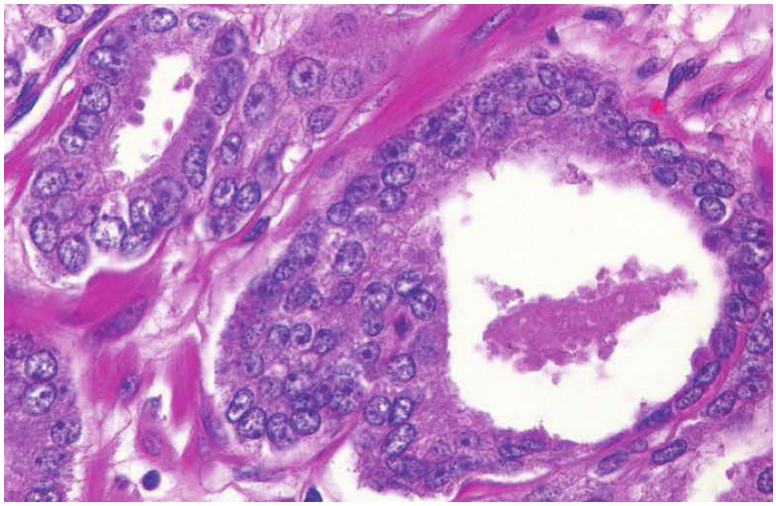


FIGURE 1.3: A pair of malignant glands within a single focus of prostate cancer.[1]

TABLE 1.1: Correspondence between Grade Groups and Gleason scores.[13]

| Grade Group | Gleason score | Gleason patterns |
|---|---|---|
| 1 | 6 | 3 + 3 |
| 2 | 7 | 3 + 4 |
| 3 | 7 | 4 + 3 |
| 4 | 8 | 4 + 4; 5 + 3; 3 + 5 |
| 5 | 9–10 | 4 + 5; 5 + 4; 5 + 5 |

The *grade* of a cancer describes the degree of abnormality of the cells. Grade in PCa is particularly important as a prognostic factor. It is determined by the Gleason system, which traditionally assigns cancers one of five grades depending on the patterns of the glands that are observed. As cancers usually contain more than one grade, grades are assigned to the most and second most common patterns. The Gleason score (GS) is then reported as the sum of the two, e.g., GS = 3 + 4 = 7.[1] In current clinical practice, Gleason grades 1 and 2 are rarely assigned, so Gleason scores range from 6–10, which is counterintuitive. There is also concern that cancers with GS = 3 + 4 and GS = 4 + 3 have disparate outcomes despite both being GS = 7.[9–11] To address these issues, the Grade Group (GG) system was developed using the Gleason system as a basis (Table 1.1). It's been shown to better stratify PCa in terms of progression and post-treatment recurrence, and has been adopted worldwide.[12,13]

The *stage* of a cancer describes the extent and spread of the cancer, and is an important prognostic factor as well in PCa. Staging includes assessments for unilateral vs. bilateral presence of cancer, extraprostatic extension (EPE) of cancer, invasion into contiguous anatomical structures — most notably the seminal vesicles (SVI), metastasis to regional lymph nodes, and metastasis to distant lymph nodes and anatomical structures.

## 1.3 Clinical management of prostate cancer

The clinical management of PCa begins with screening for the disease with the prostate-specific antigen (PSA) blood test. Patients with a positive screen as indicated by abnormally-elevated PSA levels generally undergo further diagnostic testing with transrectal ultrasound (TRUS)-guided prostate biopsy. Patients with positive findings on the PSA test and/or prostate biopsy may then undergo definitive treatment (i.e., with curative intent) of their disease.[2]

### 1.3.1 PSA-based screening

Despite the clear public health problem posed by PCa, routine screening for the disease is controversial. While it is clear that screening with the PSA blood test increases the detection rate of PCa,[3,14,15] it has at best a small benefit in reducing disease-specific mortality.[16,17] On the other hand, PSA-based screening has several downsides. First, PSA-based screening is generally prone to false-positives, as an elevated PSA may result from a number of benign conditions such as chronic prostatitis or benign prostatic hyperplasia (BPH).[2] Even when the PSA test correctly identifies cancer, the degree of PSA elevation generally does not correlate well with cancer aggressiveness,[18] and most cases of detected PCa are not and will never become clinically significant. As there is no established method for further stratifying positive screening results, many men with a positive screen will be biopsied unnecessarily, a scenario commonly referred to as

*overdiagnosis*. Due to the fact that TRUS-guided biopsy has an estimated 1-2% risk of complications that require hospitalization (e.g., persistent pain, bleeding, infection),[19] there is significant morbidity associated with the overdiagnosis of PCa.

### 1.3.2 TRUS-guided biopsy

The standard protocol for TRUS-guided biopsy involves randomly taking tissue samples from sextants of the prostate (left and right; apex, midgland, and base). Thus, the procedure inherently has low detection sensitivity,[20] and is also known to undergrade[21,22] and understage[23] disease. At the same time, standard TRUS-guided biopsies often discover low-volume, low-grade disease that are unlikely to become life-threatening.[20,24] Similar to the case with PSA-based screening, a positive diagnosis often prompts patients and their providers to take action by electing for definitive treatment, even though it may be unnecessary for the majority of men (*overtreatment*). As all forms of invasive treatment for PCa carry the risk of life-long side effects of urinary and/or sexual dysfunction that greatly diminish quality of life, there is significant morbidity associated with the overtreatment of PCa.[20]

### 1.3.3 Treatment

The conventional definitive treatment options for men with localized disease include radical prostatectomy (RP — surgical removal of the entire prostate) and external beam radiation therapy. Of the two, RP is performed more often because additional information can be obtained from pathological examination of the surgical specimen (Fig. 1.4). These prognostic factors include estimation of the percentage of the prostate involved by cancer, a more accurate assessment of cancer grade, determination of EPE and SVI, and determination of lymph node involvement, all of which can inform the need for additional therapy following surgery.

Recently, several minimally-invasive therapies for PCa, which include brachytherapy (insertion of radioactive seeds into the prostate)and cryotherapy,[26–29] have

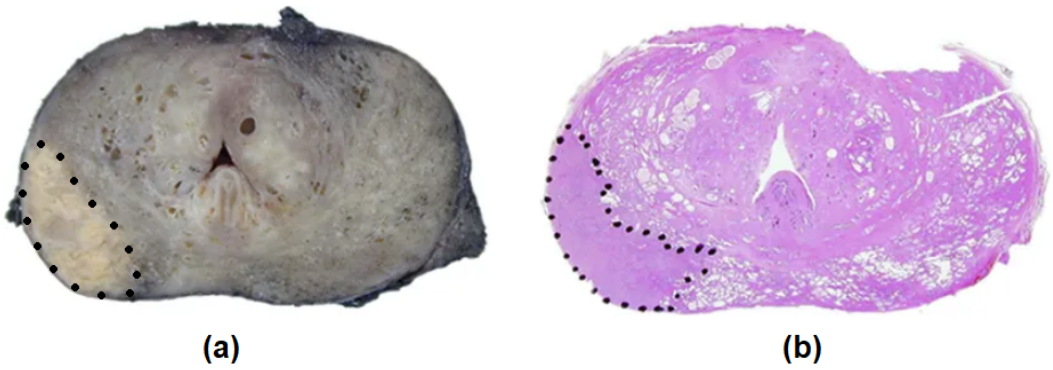



FIGURE 1.4: **(a)** Axial section of the *ex vivo* prostate after RP. **(b)** Section stained with hematoxylin and eosin (H&E). Visible cancer outlined in black.[25]

emerged as potential alternative treatments for localized disease. They offer the possibility of focal therapy (i.e., treatment of only the portions of the prostate involved by cancer), but the viability of focal therapy depends on tools that can more accurately identify and localize all of the potential foci of cancer are present.

Active surveillance (AS) has also emerged as a viable option for patients with low-grade, low-volume, localized disease. It is not simply observation of the disease; rather, patients on AS are monitored closely for signs of cancer progression with periodic PSA testing and prostate biopsy, and offered treatment if there is evidence of disease progression. AS is considered to be preferable to treatment for these types of slow-growing indolent cancers primarily because it delays treatment and the associated side-effects, sometimes indefinitely. While there does not appear to be a significant difference in the long-term mortality rate of AS versus treatment, 40–50% patients do show clinical progression and go on to receive treatment within five years of starting AS.[30,31] Therefore, the continued use and possible expansion of AS to higher-risk patients depends on tools that can identify foci of cancer as they appear and reliably track the progression of the detected foci.

## 1.4 Role of MRI in prostate cancer

In recent years, numerous studies have shown that magnetic resonance imaging (MRI) can provide valuable information for the clinical management of PCa.[32–37] In particular, multiparametric MRI (mpMRI), the combination of both anatomic and functional imaging techniques, has demonstrated significant added value in the diagnosis, localization, and staging of PCa compared to anatomic imaging alone.[2,38,39] While it is clear that mpMRI is a mainstay, there is ongoing discussion regarding how best to incorporate mpMRI into the clinical workflow, and the guidelines regarding the clinical indications for mpMRI are continuously being updated. Below is a discussion of some of the major applications, current or potential, for prostate mpMRI.

### 1.4.1 Radiologic diagnosis

For detection of PCa, meta-analysis studies have shown that mpMRI has estimated pooled sensitivity and specificity of 0.7–0.85 and 0.8–0.95, respectively.[40–43] The large variability in the reported results of individual studies can be explained by the changing mpMRI interpretation guidelines over time, degree of blinding of the clinicians involved, as well as the use of different ground truths, some of which may be more biased (e.g., biopsy more biased than RP). Sub-analyses showed improved sensitivity (0.8–0.9) but decreased specificity (0.6–0.8) for detection of clinically-significant cancer (defined broadly as higher-grade, larger-volume disease, though definitions varied among individual studies).[42,43]

### 1.4.2 Guidance for prostate biopsy

An important and growing application for positive prostate mpMRIs is guidance for biopsy. Whereas systematic biopsies randomly sample the prostate, candidate regions of interest (ROIs), i.e., regions suspicious for cancer, can first be identified on MRI and then targeted for sampling during the procedure (*MRI-targeted biopsy*). While in theory MRI-targeted

biopsy should be significantly better, its performance can be somewhat limited by both the quality of the registration between MRI and real-time ultrasound as well as the detection sensitivity of mpMRI. Nevertheless, meta-analysis studies have shown that MRI-targeted biopsies improve the sensitivity of PCa detection, reduce the identification of clinically-insignificant cancers, and reduce the number of biopsy cores required to achieve comparable performance to random biopsies.[44,45]

Another biopsy-related application for prostate mpMRI is triage, i.e., mpMRIs without radiologically-detected cancer could obviate the need for biopsy, which in turn would reduce overdiagnosis and subsequent overtreatment. Meta-analysis studies have shown the the estimated NPV of mpMRI is 0.65–0.85 for all cancers and 0.8–0.95 for clinically-significant cancer,[46–49] with large variability again due to the aforementioned factors. Although the NPV is likely not sufficiently high to warrant the use of mpMRI for triage in all cases, it could be justifiable to use negative mpMRI scans to exclude from follow-up biopsy men who are *a priori* at low risk of having clinically-significant cancer.

### 1.4.3 Active surveillance

As more and more men opt for active surveillance of their diagnosed cancer, the use of mpMRI in the setting of AS has also increased. For example, one meta-analysis study reports that an estimated 70% of men on AS have a positive mpMRI on record.[50] While it would be ideal to replace periodic biopsies with mpMRIs for monitoring the disease progression of patients on AS, studies have found that an estimated 20% patients on AS with no change on consecutive MRIs demonstrated disease progression on subsequent biopsy,[51] partially due to lack of standardized guidelines specifically for interpretation of repeat MRIs. To address this, the PRECISE guidelines were established in 2017,[52] and early results using the guidelines have shown lower rates of missing disease progression with mpMRI within one year of starting AS.[53] It is anticipated that the role of mpMRI will only continue to expand for AS.

### 1.4.4 Guidance for focal therapy

Focal therapies for PCa are typically performed under TRUS guidance, and are mostly limited to the setting of locally-recurrent cancer after definitive treatment.[28] Although there is interest in focal therapy with curative intent, as it has the potential to significantly mitigate the side effects typically associated with definitive treatment, the multifocal nature of PCa makes it challenging to be certain that all of the cancer is treated. MRI offers superior visualization of regions suspicious for cancer, and a couple of studies have used MRI-guided cryotherapy with good outcomes.[26,29] Increasing the usage and scope of focal therapy will depend on continued improvement in cancer-detection and localization accuracy with mpMRI.

## 1.5 MRI basics

### 1.5.1 MRI physics

This section provides a very brief summary of the relevant MR physics.

The phenomenon of magnetic resonance is based on interactions between nuclei and externally-applied magnetic fields, and images are formed from the measurements of these interactions. In clinical MRI of the human body, the signal is measured from hydrogen nuclei ($^1$H), i.e., protons, which are most abundant in water and fat molecules. Nuclear particles possess a quantum mechanic property called *spin angular momentum* **S**, or *spin* for short. Protons, for example, have a spin of 1/2. Particles with non-zero spin (also called *spins*, for short) possess a magnetic moment $\mu$, and the two are related as:

$$\mu = \gamma \mathbf{S} \tag{1.1}$$

where $\gamma$ is the gyromagnetic ratio.

When spins are placed in the presence of a strong, static magnetic field $\mathbf{B}_0$, the magnetic moments of the spins align with the $\mathbf{B}_0$ field (by convention, the $\mathbf{B}_0$ field is

aligned along the positive $z$-axis, and the plane containing the other two axes is referred to as the transverse plane). The spins will also precess about the $z$-axis with Larmor frequency $\omega_0$ given by the Larmor equation:

$$\omega_0 = \gamma B_0 \tag{1.2}$$

where $B_0$ is the strength of the static magnetic field. The Larmor frequency is typically expressed as a linear frequency $(f_0)$, and so the gyromagnetic ratio is often quoted as $\bar{\gamma} = \gamma/2\pi$. For protons, $\bar{\gamma} \approx 42.58\,\text{MHz/T}$.

The sum of all the magnetic moments produces a small macroscopic magnetization **M**. At equilibrium, the magnetic moments of the spins can be aligned with the $\mathbf{B}_0$ field in two ways: parallel (lower energy state) and anti-parallel (higher energy state). As the lower energy state is slightly favored, at equilibrium **M** is aligned parallel with the $\mathbf{B}_0$ field with magnitude given by:

$$M_0 \propto \frac{n\gamma^2 B_0}{T} \tag{1.3}$$

where $n$ is the number of spins and $T$ is temperature.

Since $M_0 \ll B_0$, **M** cannot be observed while it is aligned with the $\mathbf{B}_0$ field. A second magnetic field $\mathbf{B}_1$ rotating at the Larmor frequency that is also orthogonal to $\mathbf{B}_0$ may be applied to excite the spins, and it can be shown from the Bloch equations that **M** will be tipped out of alignment and begin precessing coherently about the positive $z$-axis at the Larmor frequency.[54] Since the Larmor frequency is in radiofrequency range of electromagnetic waves, the $\mathbf{B}_1$ field is also commonly referred to as a radiofrequency (RF) pulse.

The angle between **M** and $\mathbf{B}_0$ after application of an RF pulse is called the *flip angle* of the RF pulse, and is dependent on the duration and amplitude of the RF pulse. For example, a 90° RF pulse would flip **M** completely into the transverse plane; immediately afterwards, the *longitudinal magnetization* ($M_z$ — component of **M** along the

$z$-axis) would be $M_z = 0$, and the *transverse magnetization* ($M_{xy}$ — component of **M** in the transverse plane) would be $M_{xy} = M_0$.

As **M** precesses, the transverse components of **M** induce an alternating current at the Larmor frequency in the RF receive coils based on Faraday's Law of induction, and this current is the MR signal. However, after the $\mathbf{B}_1$ field is turned off, **M** will progressively return to the original equilibrium state where there is no transverse magnetization. This results in the decay of the MR signal over time, a process called free induction decay (FID).

The return to equilibrium happens through two separate processes. *T1 relaxation* is the recovery of longitudinal magnetization that occurs due the spins losing energy to the external environment, and is typically modeled as an exponential:[54]

$$M_z(t) = M_0 \left(1 - e^{-t/\mathrm{T1}}\right) \tag{1.4}$$

where $M_z(t)$ is the longitudinal magnetization over time, $M_0$ is the initial magnetization at equilibrium, and T1 is the time constant of the process. *T2 relaxation* is the irreversible loss of transverse magnetization that occurs due to the dephasing, or loss of coherence, of the spins, and is also typically modeled as an exponential:[54]

$$M_{xy}(t) = M_0 e^{-t/\mathrm{T2}} \tag{1.5}$$

where $M_{xy}(t)$ is the transverse magnetization over time, $M_0$ is the initial magnetization at equilibrium, and T2 is the time constant of the process.

In practice, the transverse magnetization decays much more quickly than would be expected for spins with known T2s. This is because there is additional, reversible dephasing from $\mathbf{B}_0$ inhomogeneities that scales with $B_0$. The term *T2$^*$ relaxation* encompasses both the irreversible T2 relaxation and the reversible dephasing; the FID signal decays exponentially with time constant T2$^*$, where T2$^* \ll$ T2.

### 1.5.2 Pulse sequences

This section is largely based on the text from *Handbook of MRI Pulse Sequences* by Bernstein et al.[55]

A pulse sequence consists of the application of RF pulses and magnetic gradients, which are performed with specific order and timing. The MR signal is sampled during one or more *echoes* that are produced by the partial rephasing of the spins over the course of the pulse sequence. All pulse sequences begin with an excitation RF pulse that tips **M** into the transverse plane; the time between consecutive excitation pulses is called the repetition time (TR), and the time between the excitation pulse and the center of the echo of interest is called the echo time (TE).

There are two fundamental pulses sequences used in MRI that form the basis for many other sequences: spin echo and gradient echo.

#### 1.5.2.1 Spin echo

A spin echo sequence begins with a 90° excitation pulse. Immediately after, the spins begin to dephase with T2$^*$ relaxation. A 180° refocusing pulse is then applied at time TE/2, which flips all the spins 180° and reverses their phase. As a result, continued dephasing of the spins causes them to come back into phase, forming the spin echo at time TE. This removes the reversible components of T2$^*$ relaxation, and thus the signal measured at time TE depends on T2, not T2$^*$. The refocusing pulse also inverts the longitudinal magnetization that has recovered through T1 relaxation between the application of the pulses.

For repeated applications of the sequence, the signal also depends on the amount of longitudinal magnetization that has recovered during the preceding TR interval through T1 relaxation. Assuming that TR $\gg$ T2 so that $M_{xy} = 0$ just before each excitation pulse, it can be shown that the general signal equation for the spin echo sequence is:

$$S = M_0 \left(1 - 2e^{-(\mathrm{TR}-\mathrm{TE}/2)/\mathrm{T1}} + e^{-\mathrm{TR}/\mathrm{T1}}\right) e^{-\mathrm{TE}/\mathrm{T2}} \tag{1.6}$$

| | Short TE ($< 40$ ms) | Long TE ($> 75$ ms) |
|---|---|---|
| Short TR ($< 750$ ms) | T1-weighted | — |
| Long TR ($> 1{,}500$ ms) | PD-weighted | T2-weighted |

TABLE 1.2: Dependence of image contrasts on TR and TE for spin echo sequences.

In the case where TR $\gg$ TE/2, the T1 relaxation between the application of the RF pulses can be ignored, reducing the signal equation to:

$$S = M_0 \left(1 - e^{-\mathrm{TR}/\mathrm{T1}}\right) e^{-\mathrm{TE}/\mathrm{T2}} \tag{1.7}$$

Spin echo sequences can produce T1-weighted (T1W), T2-weighted (T2W), or proton density-weighted (PDW) images depending on the TR and TE (Table 1.2).

#### 1.5.2.2 Gradient echo

The excitation pulse used for the gradient echo sequence usually has flip angle $\alpha < 90°$. A dephasing gradient, typically the frequency-encode gradient (see section 1.5.3 below), is applied with negative polarity shortly after the excitation pulse, which introduces variations in the Larmor frequency of the spins and in effect enhances the T2$^*$ relaxation. The polarity of the gradient is then reversed and applied for the same time, which undoes the effects of the dephasing gradient, causing spins to rephase and form the gradient echo. In contrast with the spin echo sequence, there is no refocusing pulse to compensate for $B_0$ inhomogeneities, and therefore the signal measured at time TE depends on T2$^*$.

As with spin echo, the signal for repeated applications of the sequence also depends on T1 relaxation. However, due to the use of low flip angles, the time required for sufficient recovery of longitudinal magnetization is very short. Consequently, gradient echo sequences can use much shorter TRs than spin echo sequences, resulting in faster imaging. The use of shorter TRs also means that there is generally residual transverse magnetization at the end of each TR interval. If it is allowed to persist, $M_{xy}$ eventually reaches a steady-state value, and the pulse sequence would be classified as steady-state free precession (SSFP) sequence. If it is *spoiled*, i.e., eliminated completely so that

$M_{xy} = 0$ just before each excitation pulse, the pulse sequence would be classified as a spoiled gradient-echo (SPGR) sequence. It can be shown that the steady-state signal equation for a SPGR sequence is:

$$S = k \left[ \frac{\sin\alpha \left(1 - e^{-\mathrm{TR/T1}}\right)}{1 - \cos\alpha \left(e^{-\mathrm{TR/T1}}\right)} \right] e^{-\mathrm{TE/T2^*}} \tag{1.8}$$

where $k$ is a constant and $\alpha$ is the flip angle.[55] T1 weighting increases with increased flip angles and shorter TRs. Short TEs $(< 15\,\mathrm{ms})$ produce images with more T1 weighting and less T2$^*$ weighting, while long TEs $(> 30\,\mathrm{ms})$ have the opposite effect.

### 1.5.3 Image formation

Image formation depends on the localization of signals from different points in space. In MRI, it involves the repeated application of pulse sequences with spatial encoding using magnetic gradients that alter the strength of the $\mathbf{B}_0$ field, and in turn the resonant frequency of the spins, in a spatially-dependent manner.

In two-dimensional (2D) imaging, the imaging volume is acquired as a stack of 2D slices. To acquire one 2D slice, only the spins within the slice are excited. This is achieved by applying a *slice-select* gradient (by convention along the $z$-axis) simultaneously with the excitation pulse, during which the resonant frequencies of the spins along the $z$-axis vary linearly with $z$:

$$f(z) = f_0 + \bar{\gamma} G_z z \tag{1.9}$$

where $f_0$ is the Larmor frequency and $G_z$ is the gradient amplitude. As discussed in section 1.5.1, only spins with resonant frequencies within the range of frequencies carried by the RF pulse are excited. It follows from equation 1.9 that the resulting thickness of the imaging slice $\Delta z$ is given by:

$$\Delta z = \frac{2\pi \Delta f}{\gamma G_z} \tag{1.10}$$

where $\Delta f$ is the transmit bandwidth (BW) of the RF pulse.[55]

For rectilinear data acquisition, localization within a selected slice is accomplished with the combination of *frequency-encode* and *phase-encode* gradients. The frequency-encode gradient is applied (by convention along the $x$-axis) during sampling of the MR signal. Similar to equation 1.9, the resonant frequencies along the $x$-axis are made to vary linearly with $x$ when the gradient is turned on:

$$f(z) = f_0 + \bar{\gamma} G_x x \tag{1.11}$$

where $G_x$ is the gradient amplitude. In this way, the location of the sampled signal along the $x$-axis is encoded in its frequency.

The phase-encode gradient is applied (by convention along the $y$-axis) between the application of the slice-select and frequency-encode gradients. When the gradient is turned on, the resonant frequency changes, causing the precessing spins to temporarily speed up or slow down. When the gradient is turned off, the spins all return to the Larmor frequency. The overall effect is that the spins accumulate phase $\phi$ given by:

$$\phi = \gamma G_y t \tag{1.12}$$

where $G_y$ is the gradient amplitude and $t$ is the duration of the gradient. In this way, the location of the sampled signal along the $y$-axis is encoded in its phase. In *spin-warp* imaging, phase-encoding is achieved by varying the gradient amplitude while keeping $t$ constant, which allows the TE to be kept the same for all phase-encoding steps.

The acquired signals are placed into a 2D raw-data space ($k$-space) with the range of frequencies of the signals on one axis (the $k_x$-axis, or the frequency-encode direction) and the range of their phases on the other (the $k_y$-axis, or the phase-encode direction). Each point in $k$-space corresponds to a signal acquired with a unique frequency and phase. Under the $k$-space formalism, the frequency-encoding and sampling of the MR signal corresponds to acquiring a single line of $k$-space (parallel to

the $k_x$-axis), and the phase-encoding determines the location along $k_y$ of the line that is acquired. A 2D Fourier transform is then used to transform the spatial frequencies in $k$-space to reconstruct the image.

In three-dimensional (3D) imaging, the slice-selection step is replaced with a second phase-encode step, and phase-encoding is performed so that the phase of a signal uniquely localizes the corresponding point in 3D $k$-space along both phase-encode directions. A 3D Fourier transform is then used to reconstruct the image.

## 1.6 Acquisition of mpMRI data

### 1.6.1 Magnetic field strength

The signal-to-noise ratio (SNR) of an image is often used to quantify image quality. Increasing SNR is generally desirable, as increased SNR can be used to increase image resolution and/or decrease imaging times. In MRI, the primary methods for boosting SNR include increasing $B_0$ and improving the performance of the RF receive coils.

Currently, scanners with static magnetic field strengths of either 1.5 T or 3 T are typically used in clinical practice for prostate MRI. The higher field strength at 3 T is primarily used to enable increased spatial resolution that translates to better visualization of anatomic structures and cancer. The disadvantages of increased field strength include more pronounced image artifacts and increased deposition of RF energy, though these are largely addressed with modern image acquisition techniques.

In practice, the increased SNR at 3 T significantly improves the perceived image quality as compared to at 1.5 T.[56] This SNR can be leveraged to produce images with comparable quality when using an external surface array coil (SAC) alone at 3 T when compared to the use of the combination of a SAC plus an invasive endorectal coil (ERC) at 1.5 T.[57–59] On the other hand, mpMRI at 1.5 T has demonstrated non-inferior diagnostic performance in terms of detection, localization, and localized staging of PCa as compared to mpMRI at 3 T.[56–59] However, given that the advantages of 3 T far outweigh any

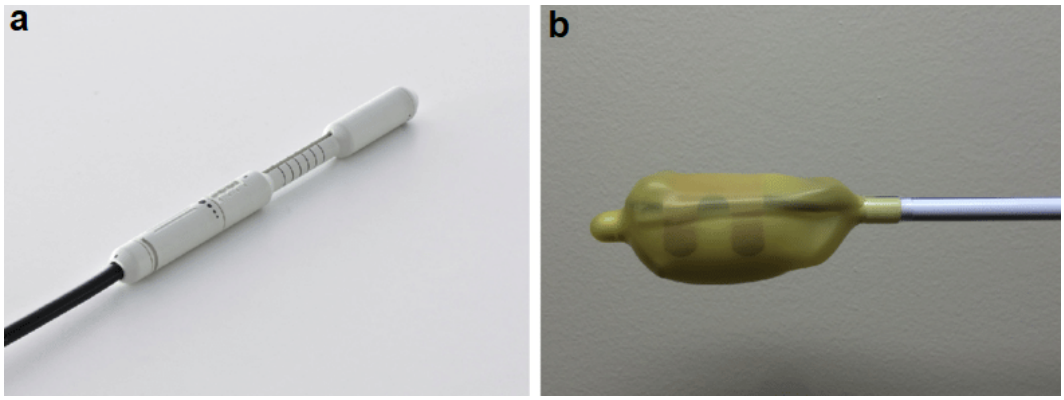



FIGURE 1.5: Examples of **(a)** a solid ERC, and **(b)** an inflatable balloon ERC.[60]

disadvantages, the use of 3 T scanners for prostate mpMRI is strongly recommended when available.[2]

### 1.6.2 Receiver coils

As mentioned above, prostate mpMRI is performed with a SAC, either alone or in combination with an ERC. The receiver element of the ERC is typically either housed in plastic (solid ERC), or in a balloon (balloon ERC) that can be inflated to hold it in place (Fig. 1.5). While the use of an ERC significantly increases SNR due to its proximity to the prostate, there are several downsides as well. First, the ERC deforms the natural shape of the organ (more pronounced for balloon ERCs) and imposes an inhomogeneous signal intensity profile on the images, which may interfere with image interpretation.[60] Use of the ERC can accentuate motion artifacts, and air in the balloon ERC can cause susceptibility mismatch at the rectal wall boundary, though this can be greatly diminished by filling the ERC with a perfluorocarbon (which better matches the magnetic susceptibility of the prostate tissue) instead of air.[61] Lastly, use of an ERC increases the time and cost of an MRI exam, and can be uncomfortable for patients.

Previous studies have demonstrated that using the combination of a SAC and an ERC for prostate mpMRI at 3 T improves visualization of the zonal anatomy and the

prostate capsule as compared to using a SAC alone, in spite of the increased motion artifacts.[62–64] Addition of the ERC was also shown to improve the sensitivity of detection of PCa, especially for foci that are located in the PZ and larger in size,[62–66] as well as the staging of PCa in terms of identification of EPE and/or SVI.[40,64,66] However, given the aforementioned issues with ERCs, current guidelines suggest that the use of an ERC is optional for prostate mpMRI.[2]

### 1.6.3 Sequences for mpMRI

Current guidelines recommend for mpMRI studies the use of anatomic T2W and T1W imaging, plus at least two functional techniques, most commonly diffusion-weighted imaging (DWI) and dynamic contrast-enhanced MRI (DCE-MRI).[2,67]

#### 1.6.3.1 T2-weighted imaging

T2W images are typically acquired with spin-echo based sequences in which refocusing pulses are used to remove the effects of inhomogeneous dephasing to achieve T2-weighted contrast. As conventional spin-echo imaging, in which one line of $k$-space is acquired per TR, is too slow for clinical use, rapid techniques that acquire multiple $k$-space lines per TR are instead used significantly reduce imaging time.

For prostate imaging, rapid 2D T2W images are acquired using a technique introduced by Hennig et al.[68] originally referred to as Rapid Acquisition with Relaxation Enhancement (RARE) and more commonly by the commercial implementation as fast spin echo (FSE) or turbo spin echo (TSE). With TSE, following a single excitation pulse, a train of refocusing pulses are applied, producing a corresponding train of echoes. The phase-encoding gradient is changed for each of the resulting echoes, and in this way, multiple lines of $k$-space are acquired for each TR interval. To limit the *specific absorption rate* (SAR — the rate of deposition of RF energy into the tissue), the flip angle of the refocusing pulses are typically less than 180°. This causes the formation of stimulated echoes that are not necessarily coincident in time or phase with the regular spin echoes,

which can cause inconsistent contrast and/or signal cancellation in the acquired images. Typically, the Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence (see section 2.5.2) is used with TSE to address these problems.[69,70]

As the signal decays with T2 over the course of the echo train, the phase-encode lines have non-uniform weighting. This effectively is a filtering or apodization of $k$-space, which produces imaging blurring that is more pronounced for spins with shorter T2s. Additionally, as echoes in the echo train have different TEs, the phase-encode lines are acquired with varied weighting. For the purposes of defining the contrast of an acquisition, the effective echo time ($TE_{eff}$) is given as the time of the echo acquired at the center of $k$-space. The imaging time with TSE, as compared to a standard spin echo sequence, is reduced by a factor equal to the number of echoes per TR, a quantity referred to as the echo train length (ETL). While the use of longer ETLs is desirable for imaging speed, the T2-related signal loss becomes more pronounced for later echoes, which in turn exacerbates the image blurring effect and decreases overall SNR. The use of longer ETLs is also commonly restricted by SAR limitations.

As T2W images are exemplary for visualizing the zonal anatomy of the prostate, they typically have the highest in-plane resolution of all the mpMRI data. The through-plane resolution, or slice thickness, is typically several fold lower in resolution. Current guidelines recommend acquiring 2D T2W TSE images in the three principal anatomic planes (axial, coronal, sagittal) with TE $\approx$ 100 ms, TR $\geq$ 5,000 ms, slice thickness $\leq$ 3 mm, and high in-plane resolution of at least 0.7 mm $\times$ 0.4 mm.[2]

One downside of the multi-slice 2D T2W TSE acquisitions in the three orthogonal planes is suboptimal visualization of the anatomy in any given view due to partial volume effects in the through-plane direction. Three-dimensional T2W TSE images acquired with isotropic resolution can address this issue, as they can be reformatted in any plane with nearly equal image quality. Commercial implementations of the 3D TSE sequence are based on the technique introduced by Mugler et al., which used a slab-selective (as opposed to a slice-selective) excitation pulse followed by a long train of non-selective

hard pulses, which allows for very small *echo spacings*, i.e., time between consecutive echoes.[71,72] The desired contrast of the sequence is achieved by optimizing the spin evolution using a variable flip angle approach. This work was further developed by Busse et al. to optimize the flip angle train, and to investigate its use for clinical imaging.[73] While the use of the 3D TSE sequence in prostate imaging has been shown to be useful for the visualization of suspicious ROIs and subsequent planning of MRI-targeted biopsies,[74] current guidelines recommend its use as an adjunct to the multi-slice T2 TSE acquisitions.[2]

#### 1.6.3.2 T1-weighted imaging

T1W images can be acquired with either spin echo or gradient echo sequences. T1W images are used primarily for evaluation of post-biopsy hemorrhage within the prostate and of the pelvic lymph nodes as opposed to the prostate itself, which has very little T1 contrast. Therefore, T1W images usually have lower resolution but a larger field of view for increased coverage of the pelvis.

#### 1.6.3.3 Diffusion-weighted imaging

DWI sequences rely on the use of the pulsed gradient spin echo technique during magnetization preparation to achieve diffusion weighting.[75] Two strong diffusion gradients are added symmetrically to either side of the refocusing pulse of a spin echo sequence. The phases of stationary spins of immobile water molecules are dephased and then rephased by the diffusion gradients, and are left unchanged. On the other hand, the spins of mobile water molecules that move between the application of the gradients are not rephased by the second, causing dephasing-related signal loss. The strength of diffusion weighting increases with the amplitude and duration of the diffusion gradients as well as the time between them, and is quantified by the $b$-value:

$$b = \gamma^2 \delta^2 G^2 \left( \Delta - \frac{\delta}{3} \right) \tag{1.13}$$

where $G$ is the gradient amplitude, $\delta$ is the duration of the gradient, $\Delta$ is the interval between the start times of the gradients, and $\gamma$ is the gyromagnetic ratio. Depending on the maximum gradient amplitude and the slew rate of the gradients, it may take a relatively long time to perform the diffusion preparation, especially for large $b$-values. As a result, incorporating diffusion weighting makes the minimum TE quite long and imposes an inherent T2 weighting on diffusion-weighted images.

In order to reduce signal loss and image artifacts resulting from acquiring motion-sensitive diffusion data, it is advantageous to acquire it rapidly. Singleshot echo planar imaging (EPI) is the fastest and most common technique to accomplish this. The entirety of $k$-space for a slice is acquired within a single TR in a rectilinear fashion using rapidly oscillating high-bandwidth frequencyencoding gradients, which generates a train of gradient echoes. Between each echo, small blipped phase-encode gradients are applied to step through the phase-encode direction of $k$-space.

While fast, the tradeoff with EPI is the presence of image artifacts resulting from increased sensitivity to off-resonance effects such as chemical shift displacement between water and fat as well as susceptibility artifacts that manifest as geometric distortions and/or T2$^*$related signal loss. While negligible in the frequency-encode direction, both of these issues are amplified in the phase-encode direction in EPI while being absent in sequences using traditional spin-warp phase encoding. Despite the relatively high readout BW, the period of spin evolution over the entire gradient echo train is quite long, allowing the evolution of off-resonance effects to occur in the phase-encode direction. To reduce these effects, EPI strongly benefits from fat suppression; common fat suppression techniques used in EPI include chemical-shift selective saturation (CHESS)[76] and slice-selection gradient reversal (SSGR).[77]

There are two strategies commonly used to increase the effective BW in the phase-encode direction while also decreasing the relatively long TEs from the use of the diffusion-encoding gradients. The first is parallel imaging (PI), which allows undersampled data to be reconstructed using the sensitivity profiles of the elements comprising a coil

array. Two popular methods for parallel imaging include SENSE (SENSitivity Encoding), which unfolds aliased images in image space,[78] and GRAPPA (GeneRalized Autocalibrating Partial Parallel Acquisition), which works by interpolating missing data in $k$-space.[79] The downside of using PI, however, is decreased SNR:

$$\text{SNR} \propto \frac{1}{g\sqrt{R}} \tag{1.14}$$

where $g$ is the spatially varying geometry factor that depends on the number and orientation of the elements of the coil array, and $R$ is the PI acceleration factor that quantifies the degree of undersampling of $k$-space (e.g., $R = 2$ would mean every other line of $k$-space is skipped). A factor of $R = 2$ is commonly used with EPI.

Another strategy is the use of 2D spatially-selective RF pulses (in both slice-select and phase-encode directions) in place of the conventional 1D slice-selective excitation pulse. As the field of view (FOV) in image space is inversely proportional to sample spacing in $k$-space, 2D selective excitation reduces the FOV and increases the effective BW in the phase-encode direction. However, unlike the case with PI, there is no SNR penalty from the geometry factor.[80]

Another challenge of EPI imaging is the presence of Nyquist ghosts. These again occur in the phase-encode direction due to phase offsets between alternating lines of kspace, which are acquired with opposite polarities of the frequency-encoding gradient; phase offsets can in turn be caused by eddy currents (induced by the rapidly-changing gradients), $\mathbf{B}_0$ inhomogeneity, and gradient imperfections. Navigators are typically used to correct for these offsets,[81] but are often confounded by the presence of fat signals that decreases their performance in applications outside the head, even when fat suppression is employed.

For DWI of the prostate, current guidelines recommend TE $\leq$ 90 ms and TR $\geq$ 3,000 ms (to minimize T2 and T1 weighting, respectively), slice thickness $\leq$ 4 mm, and in-plane resolution of at least 2.5 mm $\times$ 2.5 mm.[2] For purposes of calculating apparent

diffusion coefficient (ADC) maps (section 2.5.3), at least two, but preferably three to four diffusion-weighted images should be acquired using multiple $b$-values between 0–1,000 $\mathrm{s}/\mathrm{mm}^2$. A "high $b$-value" ($b \geq 1{,}400\,\mathrm{s}/\mathrm{mm}^2$) image is diagnostically useful,[82] but can be challenging to acquire. Due to the requirement for larger diffusion weighting, the minimum TE is increased, in turn introducing more T2 weighting and decreasing SNR. An alternative is to calculate the high $b$-value image from the acquired images with lower $b$-values (see section 2.5.3), which is also acceptable for diagnostic purposes.[2]

#### 1.6.3.4 Dynamic contrast-enhanced imaging

DCE-MRI involves the acquisition of a series of T1-weighted volumes at high temporal resolution before, during, and after the intravenous injection of a gadolinium-based contrast agent (GBCA) in order to visualize contrast enhancement of the tissue over time. Gadolinium is a paramagnetic element that shortens both T1 and T2 in a concentration-dependent manner:

$$\frac{1}{\mathrm{T1}} = \frac{1}{\mathrm{T1}_o} + r_1[\mathrm{GBCA}] \qquad \text{and} \qquad \frac{1}{\mathrm{T2}} = \frac{1}{\mathrm{T2}_o} + r_2[\mathrm{GBCA}] \tag{1.15}$$

where T1 and T2 are the observed relaxation times after contrast administration, $\mathrm{T1}_o$ and $\mathrm{T2}_o$ are the original relaxation times of the tissue, $r_1$ and $r_2$ are the relaxivity values specific to each particular GBCA, and [GBCA] is the concentration of the GBCA within the tissue. As $\mathrm{T1} \gg \mathrm{T2}$ for most tissues while the relaxivities for most GBCAs are comparable, the predominant effect of GBCAs is T1 shortening, especially at the low concentrations of GBCAs that are used in practice.[83,84]

Fast imaging methods are needed to achieve high temporal resolutions, and therefore gradient-echo based methods are generally preferred for DCE-MRI. For the prostate, 3D SPGR sequences with ultrashort TRs ($< 10\,\mathrm{ms}$) are preferred, and can achieve a reasonable balance between temporal resolution and image quality.[85,86] These sequences often rely on magnetization preparation with inversion recovery to achieve

T1-weighting. Fat suppression (e.g., with CHESS) is also commonly used to minimize the fat-water phase cancellation artifacts. Since the signal with SPGR is both T1- and T2$^*$-weighted (equation 1.8), the TE is kept as short as possible to minimize the T2$^*$ weighting, especially at higher field strengths where susceptibility effects due to the compartmentalized nature of contrast agent in tissue and blood are more pronounced.[87] For further acceleration, PI with the CAIPIRINHA (Controlled Aliasing In Parallel Imaging Results IN Higher Acceleration) technique is commonly used with PI acceleration factor $R = 4$, i.e., sampling half the lines of both phase-encode directions.[88]

For DCE-MRI of the prostate, current guidelines recommend the use of a 2D or 3D T1W gradient-echo sequence with TE $<$ 5 ms (to minimize T2$^*$) and TR $<$ 10 ms, temporal resolution $\leq$ 7 s, total scan duration $\geq$ 2 min, slice thickness $\leq$ 3 mm, and in-plane resolution of at least 2.0 mm $\times$ 2.0 mm. A set of pre-contrast images are also typically acquired over a period of 15–20 seconds before contrast injection. These can be used to produce subtraction images (obtained by subtracting the pre-contrast image acquired at the first time point from the acquired images at other time points), which are diagnostically useful.[2]

## 1.7 Interpretation of mpMRI data

### 1.7.1 T2-weighted imaging

T2W imaging provides high-resolution images of the prostate that allow for visualization of the zonal anatomy. The PZ, due to its high concentration of glands, has high water content and is therefore usually homogeneously hyperintense on T2W images. PCa is characterized by increased cellularity and decreased water content, and therefore PCa in the PZ classically appear as hypointense foci. However, this is nonspecific, as other changes like prostatitis and post-biopsy hemorrhage can have a similar appearance. T2W imaging is also useful for visualizing EPE and SVI, which are important prognostic features.[2,67]

### 1.7.2 T1-weighted imaging

T1W imaging is primarily for visualization of pelvic lymph nodes for staging, and for detection of post-biopsy hemorrhage (hyperintense on T1W images) that can mimic or obscure cancer on T2W imaging and DWI. The prostate itself otherwise has a homogeneous appearance on T1W images, and PCa is also not associated with any significant changes on T1W images.[2,67]

### 1.7.3 Diffusion-weighted imaging

DWI measures the random Brownian motion of water molecules. While water molecules are relatively free to move within normal tissue, the diffusion of water is highly restricted in cancerous tissue due to increased cellularity (an established hallmark of cancer[1]). Thus, PCa appears as focal hyperintensity on DWI with matching hypointensity on the ADC map, with decrease in ADC correlated with cancer grade.[29] The use of DWI in combination with T2WI is more sensitive and more specific for detection of PCa than use of T2WI alone.[39] In particular, DWI is useful for accurately distinguishing TZ cancers from BPH nodules. For these reasons, DWI is considered the most useful functional technique in mpMRI.[2,67]

### 1.7.4 Dynamic contrast-enhanced imaging

DCE-MRI assesses the patterns of contrast uptake and washout by prostate tissue over time. Compared to benign tissue, cancerous tissue has increased angiogenesis (another established hallmark of cancer[1]), and the newly-formed vessels are typically more permeable. Therefore, PCa demonstrates early and rapid focal contrast enhancement followed by rapid washout. DCE images can be evaluated qualitatively (often in conjunction with subtraction images), semi-quantitatively through evaluation of the contrast concentration curves, or quantitatively through pharmacokinetic modeling (see section 2.5.4).

Currently, there is debate regarding the use of DCE-MRI in prostate mpMRI. First, several studies have shown that biparametric MRI with T2WI and DWI can achieve comparable performance of cancer detection compared to mpMRI with T2WI, DWI, and DCE-MRI,[48,89] though there have not been direct comparisons of the two approaches. This likely explained by the limited sensitivity of DCE-MRI for detection of PCa.[86] On the other hand, positive findings on DCE-MRI seem to be useful for increasing suspicion of cancer when the findings on T2WI and DWI are equivocal.[2,86] There are also safety concerns regarding the use of GBCAs for DCE-MRI. GBCAs are renally cleared, and it is well-established that GBCAs can cause nephrogenic systemic fibrosis in patients with impaired kidney function, and therefore should be avoided in settings of acute kidney injury and advanced chronic kidney disease $\left(\text{eGFR} \leq 30\,\text{mL}/\text{min}/1.73\text{m}^2\right)$.[90] More recently, there is concern that GBCAs are not completely cleared, resulting in the accumulation of gadolinium ions in organs, in particular the brain.[91] Although there are no documented side effects of the deposition of gadolinium in tissues, the general consensus is that GBCAs and the imaging methods that depend on them will be used more cautiously and judiciously in the future.[92]

### 1.7.5 PI-RADS

A major challenge to using prostate mpMRI is the synthesis, interpretation, and reporting of mpMRI studies. Traditionally, mpMRI studies were assessed simply on a 1–5 Likert scale for the probability of the presence of PCa. Without any objective criteria, the assessments were decidedly subjective and heavily influenced by radiologists' experience,[93,94] and thus affected by high inter-reader variability and low reliability. To address this problem, the Prostate Imaging Reporting and Data System (PI-RADS) was developed with the purpose of standardizing mpMRI interpretation and reporting.[2,67] PI-RADS provides a set of standardized interpretation criteria for assigning a 1–5 score for each sequence (T2WI, DWI, and DCE-MRI) and an algorithm to combine the scores into a single final PI-RADS 1–5 score (PI-RADS 1 = very low suspicion, PI-RADS 5 = very

high suspicion).[2] Validation studies of PI-RADS show that the PI-RADS score correlates with increased biopsy yield, and PI-RADS 4+ with increased grade of biopsied cancers.[49]

While PI-RADS has gained significant traction since its conception, it still has similar problems with subjectivity and reader experience that cause significant interobserver variability,[95] and it is not clear that use of the PI-RADS guidelines actually leads to better cancer detection performance compared to the traditional Likert scoring.[96,97] Future development of PI-RADS may focus on addressing the interobserver variability as well as simplification of the overall PI-RADS scoring,[98] and its wider adoption will depend on the continued training of radiologists to use the guidelines.[99]

## 1.8 CAD for prostate cancer: a brief overview

A complementary approach to qualitative assessment of mpMRI data (e.g., through PI-RADS) is the use of a computer-aided detection/diagnosis (CAD) system that can automatically, quantitatively, and objectively process and analyze mpMRI data, feed the data into a predictive model, and generate information to aid clinical decision making. The use of a CAD system has the potential to reduce the time and expertise required to interpret mpMRI studies, and to improve the consistency of interpretation and reporting. There has been a myriad of works in the past 15–20 years related to the development of prostate CAD systems. While they differ in many ways, they share a number of commonalities in terms of the workflow needed to process and analyze the input data as well as the modeling approaches used to generate the desired outputs (Fig. 1.6). The numerous steps of the CAD workflow can be broken down into the following functional blocks.

### 1.8.1 Processing of mpMRI data

Given the potential heterogeneity of the mpMRI data due to differences in acquisition methods, the data need to be processed into a more standardized format that facilitates

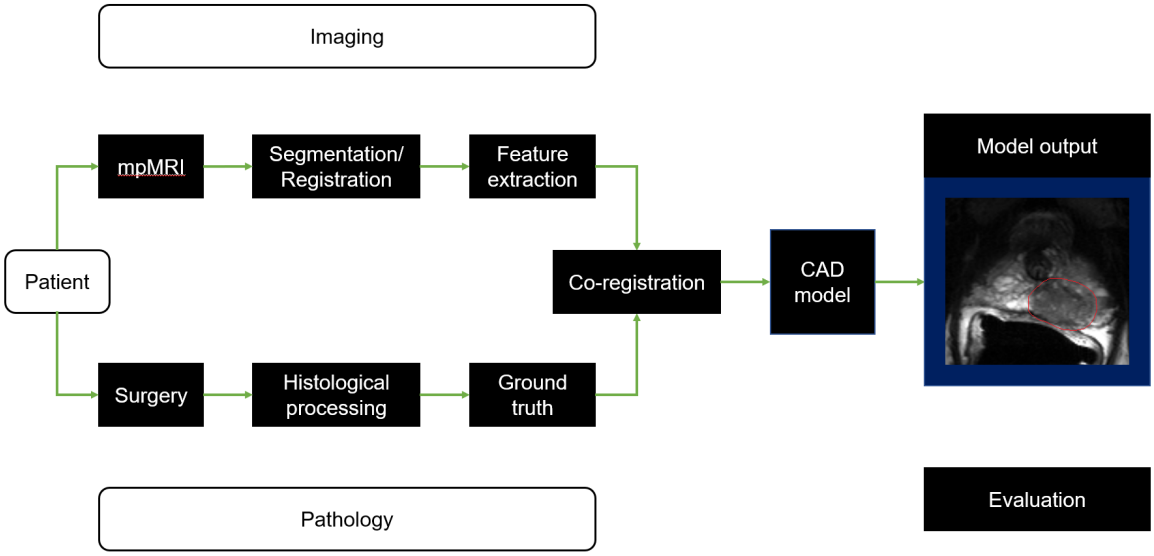



FIGURE 1.6: General workflow for prostate CAD systems.

further analysis. The major steps include the following.

#### 1.8.1.1 Intensity correction

As discussed in section 1.6.2, images acquired with an ERC have an inhomogeneous signal intensity profile. This can affect further processing steps that rely on the signal intensities. Therefore, intensity correction is an important step with mpMRI data acquired with ERCs.

#### 1.8.1.2 Image segmentation

The FOV of MR images of the prostate contain anatomical structures that are not part of the prostate itself. As further processing steps are typically only concerned with analysis of the prostate itself, the primary purpose of image segmentation is to delineate the outline of the prostate capsule. Given that the *a priori* probabilities of finding PCa differ depending on the zone (PZ $\gg$ TZ > others),[1] a secondary application of image segmentation is to delineate the zonal anatomy of the prostate to provide useful information for the CAD predictive model.

#### 1.8.1.3 Image registration

The accurate co-localization of the mpMRI data is necessary for feature extraction because predictive features are typically extracted from the same candidate region (e.g., voxels or ROIs) on different mpMRI series. However, there can be significant movement of the prostate during the 30–60 minute acquisition of the mpMRI data, which may result in the relative misalignment of the prostate volumes on the different mpMRI series. The goal of image registration is to realign the prostate volumes to ensure accuracy of the extracted features.

### 1.8.2 Processing of pathology data

#### 1.8.2.1 Ground truth

The term *ground truth* can be defined as the true pathologic state of the prostate, e.g., presence/absence of cancer, grade of the cancer. It is usually obtained from histopathologic examination of prostate tissue from prostate biopsy and/or RP. The ground truth provides the labels for training and evaluation of the CAD predictive model, and therefore accuracy of the ground truth can significantly influence the performance of the model.

#### 1.8.2.2 Mapping of the ground truth

To use the ground truth, the labels need to be mapped onto the imaging data. In the case where the ground truth is obtained from biopsy, the locations where tissue was sampled need to be identified on the imaging data. On the other hand, in the case where the ground truth is obtained from RP specimens, the individual sections need to be registered with the imaging data.

### 1.8.3 CAD predictive model

#### 1.8.3.1 Intended purpose

Broadly speaking, CAD systems can be divided into two groups depending on their intended purpose. The goal of computer-aided detection (CADe) systems is to detect and localize ROIs suspicious for cancer, while the goal of computer-aided diagnosis (CADx) systems is to characterize detected ROIs, e.g., cancer grading. Although strictly speaking CADx systems don't perform detection, CADe and CADx systems are often combined into a single framework where the output of the CADe portion is used to generate ROIs for the CADx portion.[100]

#### 1.8.3.2 Feature extraction

*Features* are explanatory variables that serve as inputs into the CAD predictive model. The term *feature extraction* encompasses the methods for obtaining and/or calculating the features from the imaging data. Examples of features that are simple to extract include signal intensity and the area/volume of an ROI, while the calculation of features like the T2 value and ADC value are more involved (see section 2.5). Choosing features that are informative, i.e., that provide good discrimination between the target classes of interest, is important to ensure good performance of the CAD model.

#### 1.8.3.3 Modeling approaches

The extracted features and labels are used to train and evaluate the CAD predictive model. These models are usually trained to perform classification, e.g., determination of cancer presence/absence, determination of cancer grade. In the past 20 years, in concordance with the growth of the machine learning and statistical modeling fields, a substantial number of techniques have been used as classifiers for prostate CAD systems. While the classifier is what generates the predictions, the general consensus is

that the quality of the features and labels contributes more to classifier performance than the choice of the classifier itself.[100]

## 1.9 Organization of the thesis

Chapter 2 focuses on a literature review of the components of a CAD system that has been outlined here. Chapters 3–6 describe projects related to four components of a CAD system: registration of mpMRI series (Chapter 3), obtaining the ground truth (Chapter 4), predictive modeling (Chapter 5), and performance evaluation (Chapter 6). Chapter 7 closes with concluding thoughts and future directions.

# Chapter 2

# Literature review

## 2.1 Intensity correction

As discussed in section 1.6.2, prostate MR images acquired with an endorectal coil (ERC) have an inhomogeneous signal intensity profile so that the signals are artificially more hyperintense closer to the coil. This effect can be strong enough to mask the signal from the posterior portion of the prostate, which includes the posterior and posterolateral peripheral zone (PZ) as well as the neurovascular bundles and rectoprostatic angles, making qualitative image interpretation difficult. It can also affect further processing steps of the computer-aided detection/diagnosis (CAD) workflow that rely on signal intensities, including segmentation, registration, and feature extraction.

The observed intensity $I_{\text{obs}}$ at a given point $p$ in images acquired with a receive coil can be modeled as the product of its true intensity $I_{\text{true}}$ and the sensitivity of the coil at that point $S(p)$:

$$I_{\text{obs}}(p) = S(p)I_{\text{true}}(p) \tag{2.1}$$

Intensity correction methods seek to estimate the sensitivity profile so that it can be divided out from equation 2.1. The earliest methods modeled $I_{\text{true}}(p)$ as a sum of a constant background $C$ and the content of interest $F(p)$ that consist of relatively

high-frequency components:[101,102]

$$I_{\text{true}}(p) = C + F(p) \tag{2.2}$$

Since the smoothly-varying sensitivity profile is composed of low-frequency components, applying a low-pass or median filter $T$ to $I_{\text{obs}}(p)$ preferentially smooths away the high-frequency components while leaving the low-frequency components relatively untouched:

$$T\left[S(p)I_{\text{true}}(p)\right] = T\left[S(p)(C + F(p))\right] \approx S(p)C \tag{2.3}$$

Dividing this out from equation 2.1 gives:

$$I_{\text{true}}(p) = 1 + \frac{F(p)}{C} \tag{2.4}$$

as the signal unaffected by the sensitivity profile. However, the filtering introduces ringing artifacts in regions where the intensities transition sharply.[101,103] For the prostate, the artifacts would appear along the capsule boundary, and thus qualitative image interpretation would not be significantly improved with this method.

Another approach is to estimate $S(p)$ using the Biot-Savart law, which calculates the magnetic field **B** at point $p$ generated by a current $I$ going through the coil:

$$\mathbf{B}(p) \propto \sum_{\text{coil}} \frac{\Delta\vec{\ell} \times \vec{R}}{|\vec{R}|^3} \tag{2.5}$$

where $\Delta\vec{\ell}$ is a unit segment of the coil and $\vec{R}$ is the vector from the segment to point $p$. Since the coil sensitivity $S(p)$ is directly proportional to the magnitude of the field $B(p)$, dividing 2.1 by the calculated $B(p)$ produces a signal unaffected by the sensitivity profile. However, this approach requires 1) the accurate localization of the coil in the images, and 2) the calculation of $\mathbf{B}(p)$ for every point in the image, which generally requires some simplifying assumptions about the geometry of the coil. Singh et al. modeled an ERC as

a single straight wire,[102] while Moyher et al. modeled a surface coil (for brain imaging) as a polyhedron of straight wire segments;[104] both used MR-visible fiducial markers for localization of the coil within the acquired images.

Liney et al. described an intensity-correction method specific for prostate imaging.[105] Using the assumption that the proton density is uniform in the prostate, they proposed to acquire a set of proton-density weighted (PDW) images $I_{\text{PD}}$ to estimate the sensitivity profile:

$$S(p) = \frac{I_{\text{PD}}(p)}{\max\limits_{p}\left(I_{\text{PD}}(p)\right)} \tag{2.6}$$

where $\max\limits_{p}\left(I_{\text{PD}}(p)\right)$ is the maximum intensity value in the PDW image. However, this method requires additional acquisition time, and also implicitly assumes that the PDW image is registered to the other mpMRI series.

## 2.2 Image segmentation

Image segmentation is the task of subdividing an image into meaningful parts. For the CAD predictive model, the meaningful partitions of an image are primarily the prostate itself, and secondarily the anatomical zones of the prostate, in particular the PZ and central gland (CG). These tasks will be referred to as capsule segmentation and zonal segmentation, respectively. Of the two, zonal segmentation is significantly more difficult due to 1) inherently low contrast between different zones of interest, and 2) the common presence yet highly-variable appearance of benign prostatic hyperplasia (BPH) and prostate cancer (PCa), which can greatly affect the boundaries and/or shapes of the zones of interest. Among the surveyed prostate CAD works, manual segmentation was actually the most common approach for both capsule and zonal segmentation.[106–114] Of course, manual segmentation not only is very time consuming and highly dependent on user experience, but also cannot be integrated into a fully-automated CAD workflow. Automatic segmentation methods can be divided into contour-based methods,

model-based methods, atlas-based methods, and supervised classification approaches; many of the surveyed segmentation works use a combination of these methods.

#### 2.2.0.1 Contour-based segmentation

Contour-based segmentation methods rely on the identification of edges of distinct regions, and have only been used for capsule segmentation. In image processing, edge enhancement and/or detection is commonly performed with the use of gradient image filters (e.g., Prewitt, Laplacian, Sobel). However, these methods are generally insufficient for capsule segmentation because they produce many false edges. Published methods have combined edge detection with pre-defined models of the capsule and user-defined control points. Samiee et al. used adaptive Canny filtering and control points to refine the boundaries of the pre-defined model on T2-weighted (T2W) images,[115] while Flores-Tapia et al. used the Haar wavelet transform in a similar fashion.[116]

#### 2.2.0.2 Model-based segmentation

Model-based segmentation methods impose some constraints on the possible shape and appearance of distinct regions. The segmentation problem is then formulated as an optimization problem with respect to these constraints. An initial guess for the segmentation is iteratively modified until it converges to some optimum of the objective function. For example, Vos et al. performed capsule segmentation through simultaneous segmentation of the prostate, bladder, and rectum.[107] The constraints on the spatial relationship of the three structures were estimated from pre-segmented images. The objective function of the optimization problem included intensity values as well as parameters that described the shape and location of the three structures on both T2W images and apparent diffusion coefficient (ADC) maps. As another example, Toth et al. performed capsule segmentation using both T2W images and dynamic contrast-enhanced (DCE) images through a two-stage approach.[117] In the first stage, texture features (see section 2.5.1.1) were calculated in neighborhoods around

user-defined control points on the prostate capsule of pre-segmented images in order to generate a texture-based model for the shape and appearance of the capsule. In the second stage, the landmark points were iteratively repositioned to new points, where the optimization criterion was maximizing the mutual information between the texture features of the original and the new landmark points.

#### 2.2.0.3 Atlas-based segmentation

An atlas is a collection of co-registered, pre-segmented images. Atlas-based segmentation methods register a prospective image to the atlas image(s), segment the registered image based on the labels of the atlas image(s), then undo the registration to obtain a segmented version of the original image. In this way, the segmentation problem is formulated as a registration problem. Dowling et al. compared two different methods of creating an atlas from multiple co-registered, pre-segmented T2W images for capsule segmentation.[118] In the first method, all of the images and their labels were combined to create a single-image atlas; prospective images were then registered to the single image. In the second method, the images and their labels were left uncombined. Prospective images were then registered to each image separately, and labels for each voxel of the registered images were determined through majority voting. It was found that the latter approach gave slightly better segmentation results as measured by the Dice similarity coefficient (DSC) of 0.83 vs 0.81 across different registration transformation models. Litjens et al. adapted this approach to creating an atlas from both T2W images and ADC maps, and used it to perform zonal segmentation (PZ and CG).[119] Other works use atlas-based methods to produce an initial segmentation, then refine the results using other approaches.[120]

#### 2.2.0.4 Classification-based segmentation

Classification-based methods perform segmentation by labeling each voxel of the image (e.g., for capsule segmentation, a voxel is either labeled as “prostate” or “non-prostate”).

Allen et al. proposed a hybrid two-stage approach for combined zonal and capsule segmentation.[121] In the first stage (classification-based segmentation), a Bayesian Gaussian mixture model was used to model the histogram of intensities in T2W images. Three Gaussians were chosen, which corresponded to the three classes of background (non-prostate), PZ, and CG. The model was fit using the expectation-maximization algorithm, and then used to predict the class for each voxel. In the second stage (model-based segmentation), the initial segmentation from the first stage was refined using the constraint that the surfaces of both the prostate and the PZ/CG boundary are smooth.

In recent years, there have been a number of works using deep learning for segmentation. Cheng et al. proposed a two-stage model using two different fully-connected convolutional neural network (CNN) architectures for capsule segmentation using T2W images.[122] The first stage applied AlexNet in a patch-based fashion to identify voxels that contained the capsule boundary to produce an initial capsule segmentation. The second approach used holistic nested networks, which is a modification of the CNN designed to improve edge detection,[123] to modify the initial segmentation. The U-net, which is a popular modification of the original fully convolutional network architecture,[124] has been used for both capsule and zonal segmentation. Whereas the traditional CNN produces a single output regardless of the dimensions of the input, the fully convolutional network architecture has additional upsampling layers so that the input and output have the same dimensions, which allows a segmented image to be produced as an output.[125] Zabihollahy et al. used U-nets for capsule and zonal segmentation (PZ and CG) on T2W images and ADC maps,[126] while Meyer et al. used U-nets for capsule and zonal segmentation of four zones (PZ, CG, anterior fibromuscular stroma, distal prostatic urethra) on T2W images.[127]

## 2.3 Registration of mpMRI data

Image registration is the task of spatially aligning two or more images to each other. Typically, one image is chosen as the *target/fixed* image, and the others are the source/moving images that are to be aligned to the target image. For a given source image, the registration problem is then posed as finding the best spatial transformation such that when it is applied to the source image, the similarity between the target and source images is maximized. Mathematically, the registration problem can be posed as an optimization problem:

$$\underset{T}{\text{argmax}} \left( Sim \left( F, T(M) \right) \right) \tag{2.7}$$

where $F$ is the target image, $M$ is the source image, $T : \mathbb{R}^n \mapsto \mathbb{R}^n$ is the $n$-dimensional spatial transformation, and $Sim$ is the similarity measure that quantifies the quality of alignment between the target and the transformed source images.

Registration algorithms can be divided into feature-based methods and intensity-based methods. The former uses image features common to both target and source images (e.g., anatomical landmarks, organ boundaries) to guide the registration, while the latter use only the intensity values of the images and/or transformed versions of the images. Intensity-based methods are preferred for automatic registration because feature-based methods require the identification and segmentation of salient features, which as discussed above is difficult to do automatically. Registration algorithms can also be classified based on the type of spatial transform $T$ that is allowed. For example, an affine transform allows for the translation, rotation, scaling, and shearing of the source image, and the application of the transformation can be efficiently calculated as matrix multiplication. A rigid transform is an affine transform that only allows translation and rotation of the source image. Non-affine transforms are also called deformable transforms; these are generally difficult to calculate and require specific similarity measures to perform the optimization.

While many works have described methods for automatic registration of

multimodal prostate images (e.g., MRI with ultrasound), relatively little has been published regarding the registration of prostate mpMR images with different contrasts. Viswanth et al. and Vos et al. carried out registration of DCE images to the T2W images using an affine transformation model and mutual information (MI) as the similarity measure.[109,128] In the latter work, Vos et al. refined the results of the global affine registration by implementing a second local deformable registration step; within each $4 \times 4 \times 4$ neighborhood of a voxel, the elastic deformation was modeled on B-splines using the global affine transformation parameters as the initial guess.[109] Chappelow et al. proposed a sophisticated deformable registration framework for the co-registration of prostate mpMR images as well as the histological ground truth.[129] Instead of registering each source image to the target as is typically done, the following steps were performed for a given set of images $I_1, I_2 \ldots I_n$:

1) Register images $I_1$ and $I_2$.

2) For $k = 3$ to $n$:

   i) Calculate a single ("multiattribute") image $I_{m,k}$ from images $I_1, I_2 \ldots I_{k-1}$.

   ii) Register image $I_k$ to $I_{m,k}$.

As a result, each image is sequentially registered to the previously-registered images, ensuring a more exact spatial alignment of the data. To accomplish this, a multivariate extension of the MI metric was developed and used. However, the algorithm is computationally inefficient, and the proposed multivariate extension of the MI metric requires the calculation of higher joint entropy terms, which exacerbates this issue.

## 2.4 Processing of pathology data

For CAD studies, the pathology data is needed to provide the ground truth for training the CAD predictive model. Obtaining and processing the pathology data is generally difficult and labor-intensive, and thus studies using primary data tend to have smaller sample

sizes. Some of the more recent prostate CAD studies, especially ones using deep-learning classification models, have been developed on publicly-available datasets (e.g., from the PROSTATEx Challenge[130]) with the ground truth included. These datasets also have larger amounts of data from multiple institutions, which are advantageous characteristics for producing predictive models that are less biased.

### 2.4.1 Obtaining the ground truth

Prostate CAD studies generally use one of two types of ground truth.

The first type of ground truth is obtained from the histologic examination of radical prostatectomy (RP) specimens, which are obtained from patients who elected for definitive treatment of their diagnosed cancer. After the prostate is removed, the tissue is sectioned, stained, and made into slides. PCa is then identified on the slides by pathologists. There is significant variation in how the cancer identification is performed. In some studies, only the *index lesion* that comprises the largest volume and/or highest grade of disease is identified,[114,131–137] while in other studies all sufficiently large foci of cancer are identified.[106,112,138–147] The identified regions of cancer in some studies are also *annotated*, i.e., outlined in detail and manually segmented from the rest of the prostate tissue.[106,112,133–135,137,138,141–146] This type of ground truth is generally considered to be best because it is the most detailed and least biased. However, it still has several drawbacks. The processing pipeline for obtaining this ground truth is more involved, which further limits the sample size of prostate CAD studies using primary data. Since RP is usually only performed on patients with at least intermediate-risk PCa, the ground truth is biased towards containing PCa with larger volumes and/or higher grades. Lastly, histologic identification of PCa is subject to significant inter-reader, experience-dependent variability.[148,149]

The second type of ground truth is obtained from the histologic examination of prostate biopsy specimens, which are obtained from patients for diagnosis of PCa.[107,119,132,150–157] The biopsy samples are made into slides, and PCa is identified on

the slides by pathologists as described above. The main disadvantage to using biopsy specimens for the ground truth is that it very likely does not sample all of the cancer that is actually present, which means the pathologic state of the vast majority of the prostate is unknown. Therefore, further processing steps of the (CAD) workflow can only reliably be applied to the locations of biopsied tissue. For similar reasons, the true volume and grade of PCa found on biopsy cannot be reliably ascertained.

### 2.4.2 Mapping the ground truth

To use the ground truth, the labels need to be mapped onto the mpMRI data. For both types of ground truth, the mapping is most commonly performed manually by a committee of pathologists, urologists, and radiologists. The mapping would ideally be performed automatically for the CAD workflow.

For automatically mapping the ground truth from biopsy, the locations where tissue was sampled need to be identified on the imaging data. A few studies have used analysis of recordings of the transrectal ultrasound (TRUS)-guided procedure to determine the trajectory of the biopsy needle deployments in order to map the ground truth onto the ultrasound images.[152,153,155,157] The ultrasound images were then registered to the T2W images, effectively transferring the ground truth mapping. Instead of trying to precisely localize the biopsied tissue, Schelb et al. mapped biopsy specimens to one of the six standardized sextants of the prostate (left and right; apex, midgland, and base).[157] While this approach is more reliable, CAD predictive models trained with this ground truth would be much less useful for cancer localization, as they would only be able to make predictions about prostate sextants.

The task of automatically mapping the ground truth from RP specimens to the mpMRI data is essentially a registration problem. Here, the registration is particularly difficult because the appearance of the prostate when it is imaged *in vivo* is drastically different from the appearance of the sectioned RP specimen, especially if an ERC is used. The prostate undergoes significant shrinkage and deformation when it is removed

from the body due to loss of support from the surrounding anatomic structures (e.g., bladder, rectum) and the connective tissue. The shape is further affected by the subsequent histological processing steps, especially sectioning and formalin fixation.[158] To constrain some of the possible deformations, several studies developed custom sectioning boxes in which *ex vivo* prostates are placed and sectioned in a consistent manner to approximately match the axial imaging slices. For carrying out the registration, Kalavagunta et al. proposed a globally deformable, locally-affine transformation model for 2D co-registration of the axial T2W image and the corresponding pathology slice. The capsule as well as large, clearly-identifiable structures (e.g., BPH nodules) were manually identified and used to guide and constrain the registration.[159] Mazaheri et al. proposed a two-step approach for 2D intensity-based registration. In the first stage, the axial T2W image and the corresponding pathology slice were registered by aligning their centers of mass using a rigid transform. In the second stage, a deformable registration implemented with B-splines was carried out.[160] Losnegard et al. also proposed a two-step approach for 3D intensity-based registration. In the first stage, the capsules of the corresponding axial T2W images and pathology slices were aligned slice-by-slice using rigid transformation. After interpolating the pathology data, the 3D volumes were co-registered, first with affine registration followed by deformable registration implemented with B-splines.[161]

## 2.5 Feature extraction

Predictive features extracted from the mpMRI data can be broadly divided into those that directly measure or reflect physical properties of the tissue and those that do not. The former are derived from the quantitative analysis of the intrinsically qualitative mpMRI data and are commonly referred to as quantiative MRI (qMRI) features or parameters, while the latter are derived from the application of statistical methods and/or image analysis techniques to the images and are commonly referred to as image-based features. While image-based features are relatively simple to calculate and their use is

more prevalent among the CAD literature, a common criticism is that many of the features have little meaning in of themselves, which makes the models trained using them difficult to interpret. On the other hand, qMRI features have real-world meaning, but generally require more work to obtain, and it is unclear whether they offer unique information for predictive modeling that cannot be obtained or approximated from image-based features.

Predictive features can also be divided into region-wise features and voxel-wise features. As the names suggest, region-wise features are calculated for defined ROIs (one feature for each ROI), while voxel-wise features are extracted for each voxel location (one feature for each voxel). The qMRI features are voxel-wise features, while the image-based features can be both.

### 2.5.1 Image-based features

#### 2.5.1.1 Region-wise features

Region-wise features can be categorized into shape-based features and statistical features. Shape-based features measure the geometric properties of the regions (e.g., volume, surface area, sphericity).[119,153,162] Statistical features are the most commonly-used region-wise features in prostate CAD works, and include histogram-based properties[107,109,112,119,147,163] (e.g., median, $n$th percentile value, interquartile range) as well as statistical moments that are estimated from the histograms (e.g., mean, standard deviation, skewness, kurtosis).[140,143,147,153] Texture-based features, which describe the relationships among the intensity values within a region, are popular as well.[131,136,143,151,162] Examples include gray-level co-occurrence matrices that describe properties of a matrix created from the histogram of signal intensities within the region, and gray-level run-length matrices that describe the texture of the intensities in pre-defined orientations; Chitalia et al. provides a comprehensive overview of the different categories of textural features and their definitions.[164] Many of these textural features may also be calculated for transformed versions of the images; commonly-used

transforms include edge-enhancement filters (e.g., Prewitt, Sobel, Gabor, Laplacian of Gaussian)[140,143,144,151] and spectral transforms (e.g., discrete cosine transform,[112] wavelet transform[140]).

*Radiomics* is an emerging field in image processing and analysis.[165] The goal of radiomics is to extract a very large number of features using combinations of relatively simple image-based operations with the aim of either using an appropriately weighted combination of the features or identifying a small number of discriminatory features for the intended clinical application. Therefore, most, if not all of these aforementioned features can be considered *radiomic features*.[139] Several free open-source packages for extraction of radiomic features are available, including LIFEx[166] and PyRadiomics,[167] which greatly facilitates the ease and reproducibility of the calculation of radiomic features.

#### 2.5.1.2 Voxel-wise features

The most common image-based voxel-wise features are the intensity values of either the original images or a transformed version of the images.[108,113,119,133,141] Otherwise, most voxel-wise features are really just region-wise features that are calculated by defining for each voxel an ROI (of pre-defined dimensions) that is centered on the voxel itself. One notable exception is the use of the spatial position of a voxel within the prostate as a feature, and is likely used as a surrogate feature for the zonal position of the voxel when the zonal segmentation is unavailable.[119,132,138]

### 2.5.2 Quantitative MRI: T2 mapping

Popular methods for generating T2 maps in a timely manner rely on the use of spin-echo based sequences to acquire a series of two or more T2W images with different echo times (TEs), then fitting the signal intensities across the T2W images to the known signal equation in order to estimate the T2 for each voxel.

#### 2.5.2.1 Single-echo spin echo

The simplest and most reliable method of T2 mapping uses images acquired with the single-echo spin echo sequence introduced in section 1.5.2. By using a very long TR, the signal equation (equation 1.6) reduces to a simple monoexponential decay with respect to TE:

$$S \approx ke^{-\mathrm{TE}/\mathrm{T2}},\ \mathrm{TR} \gg \mathrm{T1} \tag{2.8}$$

for constant $k$. However, using such a long repetition time (TR) results in excessively long scan times, making it unsuitable for practical use. One solution proposed by Sussman et al.[168] is to use a different TR for each TE so that the difference TR $-$ TE is constant. Then assuming that TE $\ll$ T1, it can be shown that the signal equation reduces again to a simple monoexponential decay with respect to TE:

$$S \approx M_0 \left(1 - e^{-\frac{\mathrm{TR}-\mathrm{TE}}{\mathrm{T1}}}\right) e^{-\mathrm{TE}/\mathrm{T2}} = ke^{-\mathrm{TE}/\mathrm{T2}} \tag{2.9}$$

since $k = M_0 \left(1 - e^{-\frac{\mathrm{TR}-\mathrm{TE}}{\mathrm{T1}}}\right)$ would be a constant. However, this approach limits the range of values that can be used for the TE, which may affect the accuracy of the calculated T2 values.

#### 2.5.2.2 Multi-echo spin echo

Much faster T2 mapping can be achieved by using the multi-echo spin echo sequence, in which a train of 180° refocusing pulses are used instead of a single refocusing pulse. The sequence of refocusing pulses produces a corresponding train of spin echoes with increasing TEs, some or all of which can be sampled. As the refocusing pulses are typically applied in very rapid succession, the T1 relaxation between consecutive radiofrequency (RF) pulses can be ignored, again reducing the signal equation to a simple monoexponential decay with respect to TE (equation 1.7). The acquisition of

signals with different TEs within the same TR as well as the use of shorter TRs provide a significant speed-up compared to single-echo spin echo.

However, there are several systematic errors that can affect the T2 values calculated with multi-echo spin echo. In practice, the flip angles of the refocusing pulses are not perfectly 180°. As a result, the net magnetization vector does not lie perfectly within the transverse plane afterwards, and so a portion of the transverse magnetization is lost with each refocusing pulse. If the refocusing pulses are applied along the same direction as the excitation pulse, this effect accumulates over the course of the sequence. The observed signal then decays faster than $e^{-\mathrm{TE}/\mathrm{T2}}$, which tends to cause an underestimation of T2.[169] Another consequence of imperfect refocusing pulses is the production of stimulated echoes, which in general can be T1- and/or T2-weighted. If the time between consecutive refocusing pulses is not constant, the stimulated echoes can additionally overlap to varying degrees with the normal spin echoes and significantly contaminate the measured signal. Non-even spacing of the refocusing pulses also produces incomplete reversal of T2$^*$-related dephasing. This causes the observed rate of signal decay to be faster than $e^{-\mathrm{TE}/\mathrm{T2}}$ and also change between echoes, which makes it difficult to fit an equation to the acquired signals.[170]

These issues are addressed by the Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence,[69,70] which is the most widely used sequence for T2 quantification. The first feature of the CPMG sequence is that the refocusing pulses are evenly spaced with spacing equal to twice the spacing between the excitation pulse and the first refocusing pulse $(\tau)$. Consequently, the $n$th refocusing pulse is applied at time $(2n-1)\tau$ and the $n$th echo occurs at time $n(2\tau)$, which means the TEs are also evenly spaced. Also, any stimulated echoes that are produced will coincide both temporally and in phase with normal spin echoes; although the stimulated echoes still contribute T1-weighting to the sampled signals, the effect is predictable enough to be accounted for. A simple approach for estimating T2 is to discard the first echo, which is a normal echo without contribution from stimulated echoes and therefore has low signal relative to the other echoes, though

this results in inaccurate quantification of short T2s.[171,172] More sophisticated methods attempt to directly model the formation of stimulated echoes and then compensate for their effect, producing more reliable T2 quantification.[173,174]

Second, *phase cycling* of the RF pulses (a technique where the phase of each RF pulse can be changed) is used. A simple phase cycling method originally described by Meiboom & Gill is the application of refocusing pulses with a 90° phase shift relative to the excitation pulse;[70] for example, if the excitation pulse is applied along the positive $x$-axis, the refocusing pulses would be applied along the positive $y$-axis. The benefit of doing this is that the imperfect refocusing of one pulse is exactly compensated by the following one, so that the spins are correctly refocused after every two refocusing pulses. Therefore, usually only the signals from the even-numbered echoes of the CPMG sequence are used for T2 quantification, while those from the odd-numbered echoes (including the first echo) are discarded.

A final consideration with the use of CPMG for T2 quantification is the presence of a non-zero noise floor, which can contaminate the measured signal at later echoes due to low SNR, especially for tissues with short T2s. One approach to account for this is to add a constant term to the signal equation during fitting.[55,172]

#### 2.5.2.3 FSE/TSE

Acquisition of a series of T2W turbo spin echo (TSE) images with different $TE_{eff}$s is another popular method for T2 mapping.[106,175–179] The major advantage of this method is speed, especially since at least one of the images would be acquired anyway in a clinical setting. The most common approach is to simply fit the signal intensities to a monoexponential decay (of the same form as equation 2.8). While the variations in image contrast described in section 1.6.3.1 are not accounted for with this method, they can be made fairly similar across the acquired images by using the same acquisition parameters (e.g., TR, flip angle of the refocusing pulse, ETL) aside from the changing $TE_{eff}$, which diminishes their effect on the calculated T2 values. Image blurring is also not accounted

for, which means that the calculated T2 maps effectively have lower spatial resolution than the nominal acquisition resolution would imply. Despite these issues, Liney et al. demonstrated in phantom studies that the calculated T2s have less than 10% on average compared to the gold standard single-echo spin echo method.[169]

### 2.5.3 Quantitative MRI: ADC mapping

The ADC map is calculated from a series of two or more diffusion-weighted images acquired with different $b$-values, then fitting the signal intensities across the different diffusion-weighted images to the known signal equation. The effect of diffusion in MRI is typically modeled as a monoexponential attenuation of the measured signal:[55]

$$S = S_0 e^{-bD} \tag{2.10}$$

where $S_0$ is the intensity without diffusion. $D$ is the diffusion coefficient for the direction along which the diffusion-encoding gradients are applied, and as tissues are generally anisotropic, $D$ will vary depending on the direction. While this can be used intentionally (e.g., in diffusion tensor imaging), it is often desirable to remove this dependence on the diffusion orientation. To do this, images are acquired with diffusion gradients applied along the three orthogonal directions with the same $b$-value for each. Applying equation 2.10, the signal for diffusion along the $x$-axis would be:

$$S_x = S_0 e^{-bD_{xx}} \tag{2.11}$$

with analogous expressions for diffusion along $y$- and $z$-axes. Anisotropic effects are removed by combining the diffusion coefficients:

$$S_{xyz} = (S_x S_y S_z)^{1/3} = S_0 e^{-b(D_{\text{trace}})/3} = S_0 e^{-b \cdot \text{ADC}} \tag{2.12}$$

where $D_{\text{trace}} = D_{xx} + D_{yy} + D_{zz}$ is the trace of the diffusion tensor. Mathematically, the trace of a matrix is invariant under a change of basis, and therefore the signal $S_{xyz}$ is the same regardless of how the coordinate system is defined. The resulting images are sometimes also referred to as trace-weighted images. The ADC is then defined as $\text{ADC} = (D_{\text{trace}})/3$.[55]

The diffusion-weighted images for ADC mapping are typically acquired with the methods described in section 1.6.3.3. The lowest $b$-value that is used is typically either zero or close to zero, which generates image that is sometimes referred to as the "b0 image". The b0 image is the direct measurement of $S_0$ in equation 2.12, and using it improves the ADC fitting by removing one degree of freedom. Besides this, there is no consensus regarding the best choice of $b$-values. However, it is described in the literature that for the prostate, the choice of $b$-values has a significant effect on the calculated ADC values.[180–182] As reported in Thormer et al., prostate DWI studies vary substantially in the number and values of the $b$-values used as well as the calculated ADC values for both benign and malignant regions.[182] One clear trend is that the calculated ADC increases with the number of small $b$-values ($b < 300\,\text{s}/\text{mm}^2$),[181,182] which can be explained by the fact that at lower $b$-values, signal loss due to dephasing from perfusion effects becomes relatively stronger.[183] More recent studies have used the normalized ADC as a predictive feature instead.[184–186] The normalized ADC is simply a ratio of the calculated ADC in a region suspicious for cancer to that in a benign region of the prostate, and is not as affected by the systematic variations in ADC calculations.

Besides the changing $b$-value, the same acquisition parameters should be used. In particular, keeping the TE constant is difficult because it requires that stronger diffusion weighting be achieved with increased gradient amplitude, which may not be possible due to hardware and/or safety constraints. As a result, the highest $b$-values that are used for ADC mapping at 3 T are around $1{,}200\,\text{s}/\text{mm}^2$.

As discussed in section 1.6.3.3, the use of high $b$-value images ($b \geq 1{,}400\,\text{s}/\text{mm}^2$) in conjunction with ADC maps can increase the sensitivity of cancer detection.[82] Due to

the aforementioned gradient amplitude limitations, the acquisition of high $b$-value images usually occurs separately from the acquisition of images for ADC mapping, if they are acquired at all. An alternative to acquiring the high $b$-value image is to synthetically generate a diffusion-weighted image with an arbitrarily high $b$-value, which is often referred to as a calculated high $b$-value image. This can be done simply by applying equation 2.12 given the acquired b0 image, the calculated ADC map, and a specified $b$-value.[187]

### 2.5.4 Quantitative MRI: pharmacokinetic modeling

While DCE images can be qualitatively assessed, there is an abundance of information that can be extracted related to the uptake and washout of the gadolinium-based contrast agent (GBCA) over time. Quantitative analysis first begins by relating the signal intensity to the tissue concentration of the GBCA $(C_t(t))$. Assuming a spoiled gradient-echo (SPGR) sequence is used, the ratio of the signal with contrast to the pre-contrast signal $(S/S_0)$ can be calculated with equation 1.8 and used to estimate the shortened T1, provided that the pre-contrast T1 value of the tissue is known. Equation 1.15 can then be used to calculate $C_t(t)$.[84] While obtaining a pre-contrast T1 map is generally preferred, using a pre-defined T1 value for the prostate is acceptable as well since 1) the prostate has a very homogeneous T1 contrast, and 2) the variable presence of post-biopsy hemorrhage can result in inaccurate T1 quantification.

The two-compartment extended Tofts-Kermode (ETK) pharmacokinetic model is commonly used to model the time-varying changes of the GBCA in the tissue after injection.[188] The two compartments are the plasma (liquid portion of blood) and the interstitial space (sometimes referred to more literally as the extracellular extravascular space, or EES) of tissues; it is assumed that GBCAs do not enter cells themselves. The tissue concentration is modeled as a weighted sum of the concentrations:

$$C_t(t) = v_p C_p(t) + v_e C_e(t) \tag{2.13}$$

where $C_p(t)$ and $C_e(t)$ are the concentrations of GBCAs in the plasma and EES, and $v_p$ and $v_e$ are the fractional volumes (i.e., volume per unit volume of tissue) of the plasma and EES, respectively.

After injection, the GBCA resides exclusively in the plasma compartment, but distributes rapidly into the EES. The ETK model assumes the process is driven by passive diffusion, i.e., the rate of contrast flux is proportional to the difference between $C_p(t)$ and $C_e(t)$:[188]

$$v_e \frac{dC_e}{dt} = K^{trans} \left( C_p(t) - C_e(t) \right) \tag{2.14}$$

where $K^{trans}$ is the *transfer constant* of the process. The equation can be rewritten in terms of $C_t(t)$ using equation 2.13 in order to relate it to the DCE signal:

$$\begin{aligned} \frac{dC_t}{dt} - v_p \frac{dC_p}{dt} &= K^{trans} C_p(t) - K^{trans} \frac{C_t(t) - v_p C_p}{v_e} \\ &= K^{trans} C_p(t) - k_{ep} \left( C_t(t) - v_p C_p \right) \end{aligned} \tag{2.15}$$

where $k^{ep} = K^{trans}/v_e$ is defined as the *rate constant* of the efflux of contrast out of the EES. It was shown that the solution to this differential equation is:[188]

$$C_t(t) = v_p C_p(t) + K^{trans} \int_0^t C_p(\tau) e^{-k_{ep}(t-\tau)} \, d\tau \tag{2.16}$$

For given $C_p(t)$, this equation can be used to estimate $K^{trans}$ and $k_{ep}$, which are the most notable quantitative features that are calculated from quantitative analysis of DCE data. As discussed in section 1.7.4, the vasculature in cancer tends to be more permeable. Thus, it is unsurprising that $K^{trans}$ and $k_{ep}$ are typically increased in prostate cancer, with some studies demonstrating grade-dependent increases.[106,189–191]

In practice, $C_p(t)$, also known as the arterial input function (AIF) in the literature, is difficult to measure or estimate.[84] At the same time, the quantification of the pharmacokinetic parameters is known to depend heavily on the AIF that is used.[192] There are currently no standards or guidelines regarding the best method of obtaining the AIF.

The most accurate determination of the AIF involves directly measuring the GBCA concentration in arterial blood samples taken periodically during the course of the DCE acquisition, but the invasiveness of the approach and the preparation required make it impractical for clinical use.[193,194]

An alternative non-invasive approach is to estimate the AIF by measuring the signal from an artery within the FOV of the image, which can be converted to a GBCA concentration using the approach outlined above;[195,196] for DCE-MRI of the prostate, the femoral or iliac arteries may be used for this purpose. However, the measurement is susceptible to partial voluming as well as in-flow effects.[197,198] Reference tissue approaches use the signal measured in well-characterized reference tissues within the FOV (e.g., nearby muscles) to iteratively solve the inverse problem of estimating the AIF and other pharmacokinetic parameters necessary to produce the recorded signals, but also rely on strong assumptions (e.g., pre-definition of a subset of pharmacokinetic parameters) that may not generally be true.[199] In contrast to the previously-described methods, population-based AIFs are not determined from the subject, but rather by taking the average of AIFs that were determined through direct measurement of blood samples from a small number of subjects.[200] Population-based AIFs are widely used because they are convenient and do not require any additional work to obtain, but they have the potential to generalize poorly due to between-subject variabilities.

## 2.6 Approaches to predictive modeling

The numerous approaches to the predictive modeling portion of the CAD workflow can be categorized based on three criteria: the goal or intended purpose of the model, the regions considered by the model, and the type of the classifier.

### 2.6.1 Intended purpose

The vast majority of published CAD models were trained to detect and localize PCa, and as discussed in section 1.8.3.1 performed the task of computer-aided detection (CADe). Models for computer-aided diagnosis (CADx) have been very rare until recently, and thus far all of them have focused on the task of distinguishing between high-grade (GS $\geq$ 7) versus low-grade (GS $\leq$ 6) disease.[134,156,201,202] The model proposed by Fehr et al. performed this classification on candidate regions of cancer mapped to the ground truth obtained from RP specimens, and therefore is a rare instance of CADx-only model.[134] THe other three surveyed works have used deep learning approaches to perform both CADe and CADx.[156,201,202]

### 2.6.2 Considered regions

Published CAD models generally focus on either classification of discrete, contiguous regions of interest (region-wise classification) or classification of individual voxels of the prostate (voxel-wise classification). Region-wise classifiers require the *a priori* identification of the candidate regions themselves. For works where biopsy findings were used as the ground truth, locations where biopsies were taken served as the candidate regions.[150,153,202–205] For works where findings from RP specimens were used as the ground truth, benign regions on the ground truth that corresponded to suspicious regions on mpMRI were commonly manually identified and added to the training set for purposes of model training, which is clearly a biased approach.[109,112,163] Cameron et al. addressed this by applying a simple threshold to the ADC map to generate candidate regions.[162] While voxel-wise classifiers are less biased because effectively every voxel is considered a candidate, the outputs of the models are noisier, which makes them difficult to interpret.[106–108,113,114,132,133,135,136,138–141,143,144,156,201]

### 2.6.3 Classifier type

For a given set of feature vectors $x$ and corresponding ground truth labels $y_t$, the goal of a classifier is to produce predictions $y$ such that the degree of disagreement between $y_t$ and $y$ is minimized. Classifiers can be divided into generative models and discriminative models based on how they calculate $y$ for the given set of feature vectors, i.e., how the conditional probability $P(Y|X = x)$ is determined. Generative models estimate the joint distribution $P(X, Y)$ and use Bayes theorem to calculate $P(Y|X = x)$, while discriminate models estimate $P(Y|X = x)$ directly. In the prostate CAD literature, discriminative models have been significantly more popular.

The following is a brief overview of the different types of classifiers that have been used for prostate CAD. For more information about each, please refer to *Pattern recognition and machine learning* by Bishop CM.[206]

#### 2.6.3.1 Generative models

- **Discriminant analysis**: Linear and quadratic discriminant analysis are binary classifiers in which the separation between the two classes is found by maximizing the ratio of the interclass (between-class) variance to the intraclass (within-class) variance.[107,132,133,139,147] A non-linear separation between the two classes is achieved in quadratic discriminant analysis by computing class-specific covariance matrices (instead of assuming identical covariance matrices as in linear discriminant analysis).

- **Bayesian classifier**: Giannini et al. used a Naive Bayes classifier for voxel-wise cancer detection,[207] while Cameron et al. and Niaf et al. used a Naive Bayes classifier for region-wise cancer detection.[112,162] The Naive Bayes classifier assumes independence of the features; this makes the covariance matrix diagonal, which greatly simplifies the model. In contrast, Jin et al. developed a full Bayesian

hierarchical model that modeled the features and the probability of cancer as a function of spatial location.[138]

#### 2.6.3.2 Discriminative models

- **Support vector machines (SVMs)**: SVMs were the most popular discriminative classifier among the surveyed works.[108,109,112,113,134–136,141,143,152,153,163] SVMs is a binary classifier that aims to find the hyperplane that maximally separates the two classes. While SVMs is formulated as a linear classifier, it can produce non-linear decision boundaries through the use of kernel mapping functions; both polynomial and radial basis function (RBF) kernels have been used.
- **Random Forests**: Random Forests were the most popular ensemble classifier among the surveyed works.[119,144,204] In general, ensemble methods combine many weak classifiers (e.g., decision trees) to produce a single strong classifier. A Random Forest classifier is composed of many decision trees, each of which is trained with a random subset of the features.
- **Deep learning**: The majority of deep learning models for prostate CAD have used CNNs with ten or more encoding layers.[146,156,157,202,205,208–212] For voxel-wise classification, upsampling layers were added to produce cancer probability maps.[145,209] Wang et al. exploited the 3D imaging data by using 3D convolutions in the CNN,[145] but was the only work to do so.

Other lesser used discriminative models include logistic regression and $k$-Nearest Neighbor clustering. Logistic regression is a binary classifier that seeks to fit a logistic sigmoid function to the data[106,133] to separate it. The $k$-Nearest Neighbor algorithm is an unsupervised classifier that separates the data into a number of clusters so that examples within the same cluster are more like each other than they are to examples outside the cluster.

# Chapter 3

# Framework for intensity-based affine registration of mpMRI data

## 3.1 Introduction

As discussed in section 2.3, registration of the mpMRI data is important so that extracted features can be accurately calculated and co-localized. However, it is challenging due to the absence of rigid well-defined structures in the prostate and differences in contrast between imaging series. The registration problem is equivalent to solving the optimization problem (equation 2.7) of finding the best transformation function $T$ that maps source image $M$ to target image $F$ so that the similarity measure $Sim(F, T(M))$, which quantifies the degree of alignment between $F$ and $M$, is maximized. Therefore, there are several components of the registration method that need to be chosen or defined.

### 3.1.1 Choice of transformation model

In the most general case, the function $T$ maps each point of $M$ to a corresponding point in $F$, and the degrees of freedom (DOFs) is maximum (3 times the number of points in $M$). The transformation model can be made more simple by placing constraints on the form of $T$, which reduces the DOFs. Intuitively, the transformation model should be as simple as

possible while still being able to fully characterize the possible transformations that are actually observed.

Organs of the body are typically modeled as smooth surfaces, and undergo physiologic motions that can be described as smooth, non-linear, and reversible (i.e., invertible) transformations. Theoretically, then, physiologic motion is best described by a *diffeomorphic* transform, which provides a smooth, one-to-one mapping of each point in $M$. However, diffeomorphisms have the maximum number of DOFs, and are generally very time-consuming to compute.

Simple transformation models can reasonably approximate physiologic motion, especially when the expected non-linear components of the motion are relatively small. Therefore, *affine* transformation models are commonly used in medical image registration. An affine transformation **T** is a linear mapping of every point in $M$ to a point in $F$. For a given point $M_k = (M_x, M_y, M_z)$, the 3-dimensional (3D) affine transform can be calculated as matrix multiplication:

$$\mathbf{T}(M_k) = \begin{bmatrix} a_{00} & a_{01} & a_{02} & t_x \\ a_{10} & a_{11} & a_{12} & t_y \\ a_{20} & a_{21} & a_{22} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} M_x \\ M_y \\ M_z \\ 1 \end{bmatrix} \tag{3.1}$$

where $t_x$, $t_y$, and $t_z$ are the translations along the respective axes. An affine transformation is the composition of translation, rotation, scaling, and shearing transforms, and by convention is calculated as:

$$\mathbf{T} = \mathbf{T}_{\text{shear}} \cdot \mathbf{T}_{\text{scale}} \cdot (\mathbf{T}_{\text{rot}} + \mathbf{T}_{\text{trans}}) \tag{3.2}$$

As each component has three DOFs, an affine transform has 12 DOFs in total.

### 3.1.2 Choice of similarity measure

The similarity measure may be calculated over a number of matching features (e.g., points, regions, surfaces) present in both images or over the intensity values for both images; these correspond to feature-based and intensity-based registration, respectively. As discussed in section 2.3, automatic image registration is generally more easily implemented with intensity-based methods because they do not require the prior identification of image features. However, the two approaches can also be used together. A common example for both medical[213] and non-medical[214] images is the registration of objects of interest within the images. The objects of interest, which constitute the features, often compose a small fraction of the field of view (FOV) of the images, and can be delineated from the background in advance on both the target and source images. Intensity-based registration can then be carried out over the space common to both.

For intensity-based registration, the choice of similarity measure depends on the expected relationship between the image intensities of the target and source images. For example, the negative sum of squared differences (SSD) is often used when the aligned source image is expected to be identical to the target:

$$-\,\mathrm{SSD}(F, T(M)) = -\frac{1}{n}\sum_{k=1}^{n}(F_k - T(M_k))^2 \tag{3.3}$$

where $n$ is the number of voxels in the overlap between the two images. If the expected relationship between image intensities is linear, cross correlation or normalized cross correlation are commonly used instead. However, these measures are insufficient for the registration of medical images with different contrasts. For images of different contrasts, mutual information (MI) and its variants are commonly used. The mutual information between $F$ and $T(M)$ is defined as:

$$\mathrm{MI}(F, T(M)) = E(F) + E(T(M)) - E(F, T(M)) \tag{3.4}$$

Here, $E$ is the entropy function defined by:

$$E(F) = -\sum_{a \in F} p(a) \log(p(a)) \quad \text{and} \quad E(T(M)) = -\sum_{b \in T(M)} p(b) \log(p(b)) \tag{3.5}$$

where $p(a)$ is the probability that a voxel in $F$ has intensity $a$, and $p(b)$ is the probability that a voxel in $T(M)$ has intensity $b$. $E(F, T(M))$ is the joint entropy, and is defined similarly:

$$E(F, T(M)) = -\sum_{a \in F} \sum_{b \in T(M)} p(a, b) \log(p(a, b)) \tag{3.6}$$

where $p(a, b)$ is the joint probability that a voxel in the overlap of $F$ and $M$ has intensity $a$ and $b$, respectively. The marginal probabilities and joint probability in the definitions above are estimated empirically from the observed image intensities. Since the MI increases with increasing overlap of the images, the normalized mutual information (NMI) is an alternative, and is obtained through normalization with the joint entropy:

$$\text{NMI}(F, T(M)) = \frac{\text{MI}(F, T(M))}{E(F, T(M))} + 1 = \frac{E(F) + E(T(M))}{E(F, T(M))} \tag{3.7}$$

### 3.1.3 Choice of optimizer

The registration problem (equation 2.7) can be solved using the standard array of optimization methods, which can be divided into gradient-based and gradient-free methods. The choice of optimizer will depend heavily on the similarity measure $Sim$. If $Sim$ is differentiable, then gradient-based methods are preferred for their speed and reliability. However, the registration problem is generally non-convex, and therefore gradient-based methods are likely to converge upon local optima, especially when the target and source images have different contrasts. For non-convex problems, gradient-free methods, while not guaranteed to find the global optimum, can perform better than gradient-based methods. The major downside of gradient-free methods is that they are computationally expensive as they all involve some degree of systematic

searching of the parameter space, and thus are impractical when the number of DOFs becomes large.

### 3.1.4 Proposed registration method

Here, we propose a semi-automatic framework for the 3D affine registration of mpMR images with NMI as the similarity measure. An affine transformation model was chosen on the basis that the greatest contribution to prostate motion during the course of a scan is from patient movements, which are mostly translational and to a lesser degree rotational. Non-linear compression of the prostate also occurs due to filling of the bladder superiorly and filling of the rectum with gas posteriorly, and these are approximated by scaling and shearing components of the affine transform. NMI was a natural similarity measure due to the different contrasts of the mpMR images. The registration framework also implements a registration-based intensity correction of mpMR images acquired with an endorectal coil (ERC).

## 3.2 Methods

The mpMRI data used in this work were acquired as previously described,[106] and consist of the anatomic T2-weighted (T2W) image, multiple turbo-spin echo (TSE) images acquired at different echo times (TEs), the “b0” trace-weighted image, and the dynamic-contrast enhanced (DCE) images. Please refer to Appendix A for more details.

The anatomic T2W image was chosen as the target/fixed image, and the others chosen as the source/moving images. One of the applications of this registration work is to achieve alignment of the maps of the quantiative MRI (qMRI) features. Registration of the multi-TE TSE images is needed as a preprocessing step for T2 mapping (see section 2.5.2). As the diffusion-weighted images are acquired over a shorter period of time than the TSE images, only the b0 image is registered, and we assume that the same transformation applies to the images at other $b$-values as well as the apparent diffusion

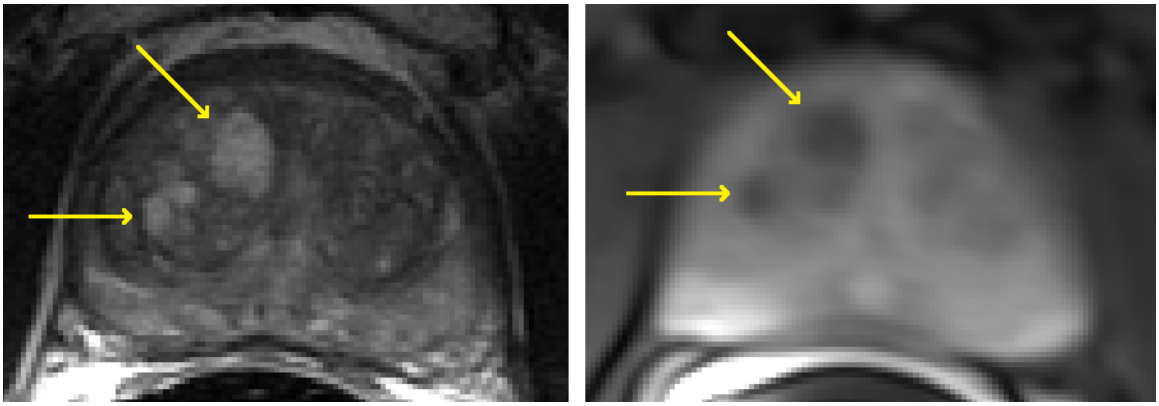


FIGURE 3.1: Comparison of corresponding slices of an anatomic T2W image (left) and a late-time point DCE image (right). Yellow arrows point to corresponding BPH nodules seen on both.

coefficient (ADC) map. The b0 image was chosen as the representative source image for the diffusion-weighted images because it is effectively a T2W image and therefore has the most similar appearance to the target image. Similarly, as the DCE images are acquired over a shorter period of time, only the image acquired at the second-to-last time point was registered, and we assume that the same transformation applies to the images at other time points as well as the pharmacokinetic maps. A late time point was chosen because the accumulation of the contrast agent in the tissue highlights certain portions of the anatomy that are also seen on the target images. Common examples of this are benign prostatic hyperplasia (BPH) nodules, which do not take up contrast well and are visibly hypointense on the late-time point DCE image (Fig. 3.1).

### 3.2.1 Definition of the initial VOI

First, a rectangular volume of interest ($\text{VOI}_0$) was defined on the target image $F$. Since we are only concerned with registration of the prostate, the purpose of defining the VOI is to specify the sub-volume containing the prostate on which the registration parameters will be optimized. Using the convention that the left-right axis is the $x$-axis, the anterior-posterior axis is the $y$-axis, and the base-apex axis is the $z$-axis, the extent of the prostate in the $x$ and $y$ dimensions were first determined on the target image through

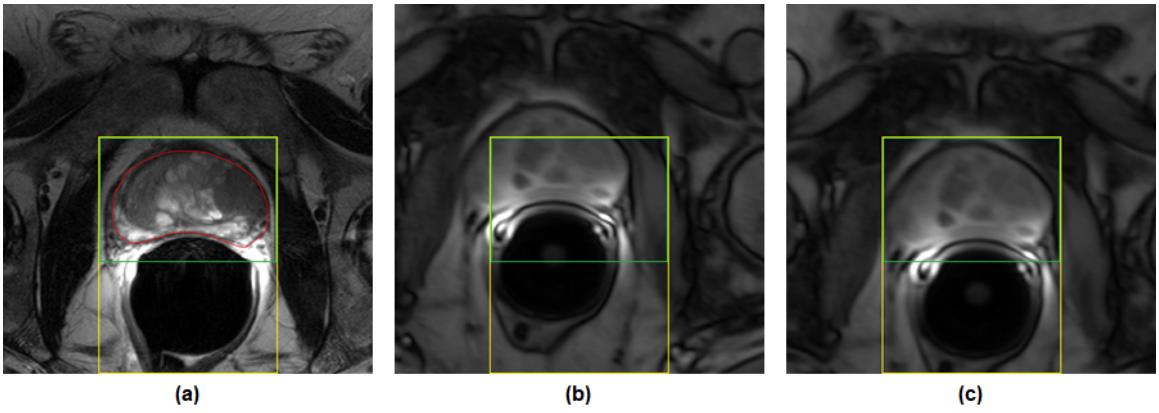



FIGURE 3.2: **(a)** Representative slice of the target image $F$ with capsule annotation in red, $\text{VOI}_0$ in green, and $\text{VOI}_1$ in yellow. **(b)** Corresponding slice of the late-time point DCE image $M$ without registration. The large degree of misalignments illustrates that the VOIs cannot simply be copied from $F$ to $M$. **(c)** Corresponding slice of $M_1 = \mathbf{T}_0(M)$ with the propagated VOIs.

manual annotation of the capsule by a radiologist. The extent of the $\text{VOI}_0$ in those directions were then chosen to be equal to the extent of the prostate in those dimensions plus a margin of error of 2.5 mm in the positive and negative directions (yellow VOI in Fig. 3.2a). The extent of $\text{VOI}_0$ in the $z$ dimension was taken to be equal to the extent of the acquired imaging volume.

As the capsule was only annotated on the target image, $\text{VOI}_0$ needs to be defined on the source images as well. To do this, the translation component of the affine transformation was first estimated over all overlapping voxels of $F$ and $M$. This translation-only pre-transformation matrix $\mathbf{T}_0$ was applied to $M$, and then $\text{VOI}_0$ was propagated to the transformed image $M_1 = \mathbf{T}_0(M)$ (yellow VOI in Fig. 3.2c).

### 3.2.2 Intensity correction

Next, the signal inhomogeneity caused by the sensitivity profile of the ERC was corrected for. This step was necessary because intensity-based registration could otherwise be biased toward matching the coil sensitivity profiles instead of the the prostate volumes. To perform the correction, the ERC was modeled as two parallel wires, and the Biot-Savart law was used to estimate the coil sensitivity profile $S$ according to

equation 2.5 (Fig. 3.3b). $S$ was registered to each image $I$ using an affine transformation model with the goal of minimizing the variance of the image intensities within $\text{VOI}_0$ of the corrected image. Therefore, the *minimum variance* similarity measure to be maximized can be stated as:

$$Sim(I, T(S)) = -\sum_{k \in \text{VOI}_0} \left[ \log\left(\frac{I_k}{T(S_k)}\right) - \bar{x} \right]^2 \tag{3.8}$$

where

$$\bar{x} = \frac{1}{n} \sum_{k \in \text{VOI}_0} \log\left(\frac{I_k}{T(S_k)}\right) \tag{3.9}$$

is the average value of $\log\left(\frac{I_k}{T(S_k)}\right)$ for within $\text{VOI}_0$. After registration, each image $I$ was then divided by the aligned sensitivity profile to obtain the intensity-corrected version $I^c$ (Fig. 3.3d).

### 3.2.3 Registration of the mpMR images

After the intensity-corrected images were obtained, each source image $M_1^c$ was registered to $F^c$. To provide more information for the registration, $\text{VOI}_0$ was extended posteriorly in the $-y$ direction to the edge of each image to obtain a larger VOI ($\text{VOI}_1$); this is the sub-volume on which the registration of the mpMR images was carried out (Fig. 3.2). The extended $\text{VOI}_1$ generally includes most of the ERC, which can be considered another feature to help guide the registration. NMI was used as the similarity measure in all cases.

As mentioned in the introduction, the marginal and joint probabilities in the definition of MI are estimated empirically from the image intensities of the images. In this work, a histogram-based approach was used, where the count of each bin is the number of occurrences (for marginal probability) or co-occurrences (for joint probability) of the corresponding range(s) of intensity values. The choice for the number of bins is known to be an important parameter that has a significant effect on the calculated values of MI. In this work, the number of bins $(N_I)$ for a VOI defined on image $I$ was determined using the

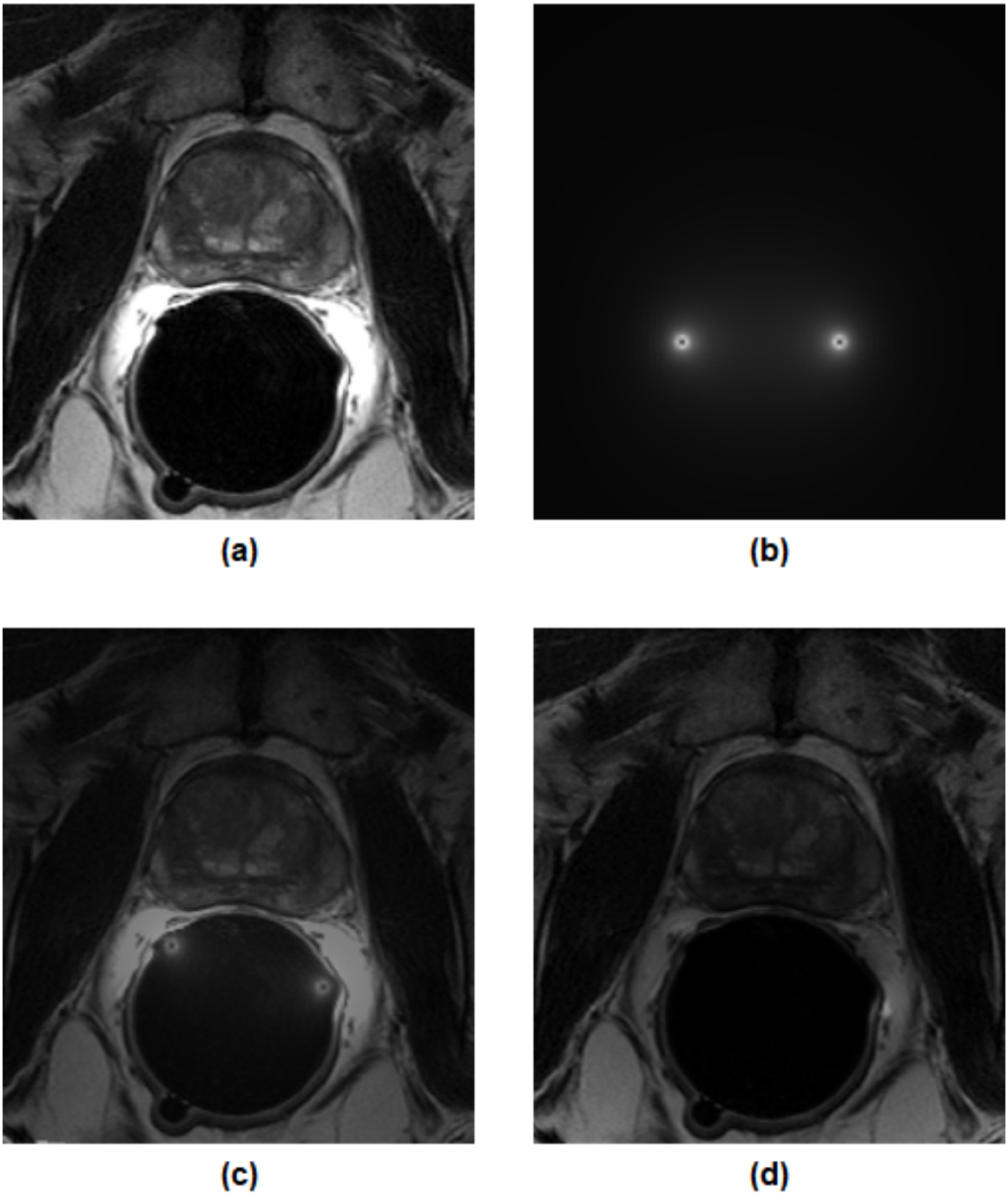



FIGURE 3.3: **a**: Representative slice of the original target image $F$. **b**: Corresponding slice of the intensity profile of the ERC calculated from the Biot-Savart law. **c**: Registered intensity profile overlaid on top of $F$. **d**: Intensity-corrected target image $F^c$ obtained by dividing the registered intensity profile.

Freedman-Diaconis rule:[215]

$$N_I = \left\lceil \frac{\max(I) - \min(I)}{2 \cdot \mathrm{IQR}(I) \cdot n^{-1/3}} \right\rceil \tag{3.10}$$

where $n$ is the number of voxels in the VOI and IQR is the interquartile range of intensity values within the VOI defined on image $I$.

#### 3.2.3.1 Genetic algorithm for optimization

To find the best affine transform in the described registration tasks, a genetic algorithm (GA) was used as the optimizer. A GA is a gradient-free optimization method that iteratively evolves a population of candidate solutions toward better solutions using concepts borrowed from evolutionary theory such as natural selection, genetic crossover, and random mutations. For more details about GAs, please refer to Hajela et al.[215] and Whitley et al.[216]

Procedurally, a GA is composed of the following steps:

1. **Initialization**: The GA starts with the initialization of a population of random guesses. As an affine transform has 12 DOFs, each guess is represented by a vector of 12 parameters. A population size of $N_p = 1,000$ members was used.

2. **Fitness calculation**: At each iteration, the value of the similarity measure corresponding to the transformation encoded by each member is calculated as the reproductive fitness of the member (higher similarity = more fit). The number of "offspring" (i.e., members of the next iteration) a member is allowed to contribute is directly proportional to its fitness. Specifically, the number of offspring produced by the $i$th member ($N_{c,i}$) was determined by:

$$N_{c,i} = \left\lfloor N_p \left( \frac{x_i}{\sum_{k=1}^{N_p} x_k} \right) \right\rfloor \tag{3.11}$$

where $x_i$ is the value of the similarity measure achieved by the $i$th member.

3. **Offspring generation**: There are numerous ways to generate offspring involving one or more members of the current iteration. Here, a fairly simple scheme was chosen where each offspring is generated from only one member. For the $j$th parameter of the $i$th member $(\theta_{j,i})$, the value of the $j$th parameter of its offspring is sampled from the normal distribution with $\mu = \theta_{j,i}$ and $\sigma^2 = 1 - x_i$ for NMI (equation 3.7) or $\sigma^2 = -x_i$ for minimum variance (equation 3.8). The random sampling procedure is analogous to introducing random mutations to the population. The effect of the choice for $\sigma^2$ is that as $x_i$ improves, the spread of the parameter values of the offspring generated from member $i$ decreases, which approximates convergence.

Due to the random nature of GAs, the optimization step was repeated five times for each registration task (except for the calculation of $\mathbf{T}_0$ where it was performed only once) with different initial populations each time, and the calculated parameters were averaged. A multi-resolution registration approach was also implemented (except for the calculation of $\mathbf{T}_0$ which was only performed at the original resolution). Target and source images were downsampled to 25%, 50% and 100% of their original resolution, then sequentially registered at increasing resolution levels, starting with the lowest. The final population of the GA calculated at one resolution level was used as the initial population of the GA at the next level.

## 3.3 Experiments

To characterize the consistency of the proposed methods, ten source images of each type (TSE image with TE = 36 ms, b0 image, late-time point DCE image) were randomly selected from the available data. Each was registered to the corresponding target using the described registration procedure, and the registered image $\mathbf{T}(M)$ as well as the composite affine transformation matrix $\mathbf{T}$ were obtained.

For each source image, five random affine pre-transformations $\mathbf{X}_1$ to $\mathbf{X}_5$ were created. The value for each of the 12 parameters was randomly and uniformly sampled from the range of values observed in the calculated affine transformation matrices for the same type of source image (across all 34 cases). For each $\mathbf{X}_i$, the image $\mathbf{X}_i(M)$ was calculated first, then registered to the corresponding target. Ideally, $\mathbf{T}(M) = \mathbf{T}\mathbf{X}_i^{-1}\mathbf{X}_i(M)$ should again be found as the registered image, and thus the transformation matrix $\mathbf{T}\mathbf{X}_i^{-1}$ would ideally be recovered. However, this is generally not the case. Let $\mathbf{T}_i$ be the transformation that is actually found. The difference between $\mathbf{T}_i\left(M^c\right)$ and $\mathbf{T}\left(M^c\right)$ was quantified by the normalized mean square error, which for two images $Y$ and $Y'$, where $Y'$ is an approximation for $Y$, is defined as:

$$\mathrm{NMSE}(Y, Y') = \frac{\mathrm{MSE}(Y, Y')}{\mathrm{MSE}(Y, 0)} = \frac{\sum_k \left(Y_k - Y'_k\right)^2}{\sum_k Y_k^2} \tag{3.12}$$

for voxels $k$. $\mathbf{T}_i\left(M^c\right)$ was considered to be an approximation to $\mathbf{T}\left(M^c\right)$, and the NMSE was calculated for voxels $k \in \mathrm{VOI}_0$ instead of over the entire overlap between the two. Note that the intensity-corrected images $M^c$ were used for the calculation of the NMSE.

The difference between $\mathbf{T}_i$ and $\mathbf{T}\mathbf{X}_i^{-1}$ was quantified by $\left\|\mathbf{T}_i - \mathbf{T}\mathbf{X}_i^{-1}\right\|_F$. $\|\mathbf{A}\|_F$ is the Frobenius norm of matrix $\mathbf{A}$:

$$\|\mathbf{A}\| = \sqrt{\sum_{i,j} \left(a_{ij}\right)^2} \tag{3.13}$$

where $a_{ij}$ are the entries of $\mathbf{A}$, and is equivalent to the L2 norm of the vectorized version of $\mathbf{A}$. For each image type, the results were averaged over both the five different source images and the ten random affine transformations for each source image.

## 3.4 Results

From the experiments, it was observed that the proposed registration method gives qualitatively good results when the estimated parameters of the affine transformation are

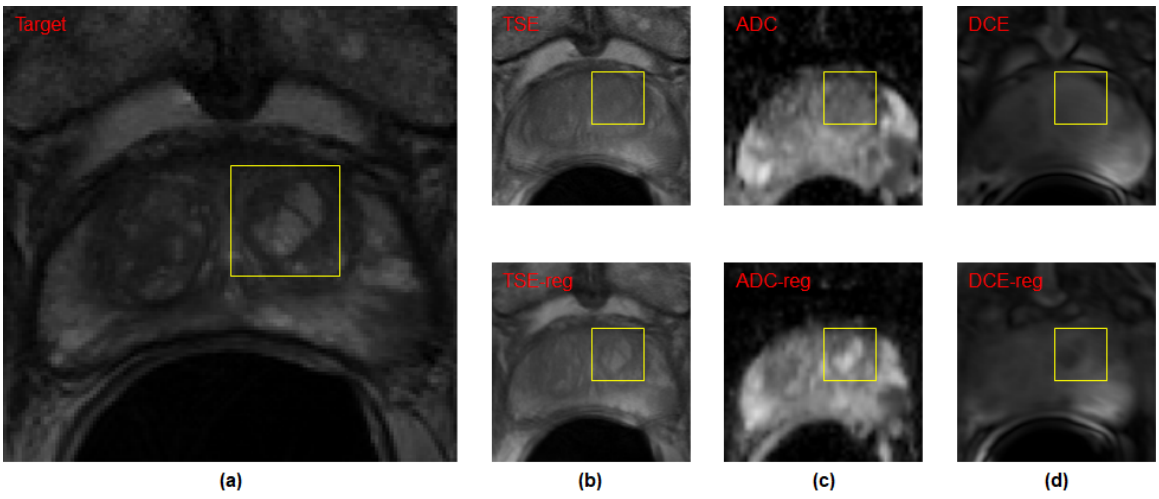



FIGURE 3.4: Example of a successful application of the proposed registration method. The FOV of all of the images here are spatially co-localized, and all images are intensity corrected. A slice of the target image in (a) with the yellow ROI containing a prominent BPH nodule. The feature is not seen on any of the original source images ((b)–(d), top row) due to sizeable through-plane translations of the prostate, but is clearly identified on all of the registered images ((b)–(d), bottom row).

relatively small (e.g., translation $< 10$ mm, rotations $< 0.25$ rad$(\approx 15°)$, shear factor $< 0.1$, and scaling factor $< 5\%$). An example of a successfully-registered case is shown in Figure 3.4. Initially, the BPH nodule identified on the target image was not found on the same slice of any of the source images, but is approximately co-localized after registration on all of the source images with calculated through-plane translations of 2.5 mm, 4.2 mm and 8.7 mm for the TSE, b0, and late-time point DCE images, respectively.
In cases with bigger estimated ranges of motions, the GA sometimes struggled to converge to a good solution, which required multiple re-initializations of the starting population and/or many more iterations of the GA to resolve.

Experimental results are shown in Table 3.1. For each type of source image, the mean and standard deviation of the values of the error measures were calculated over all 50 trials (10 images $\times$ 5 pre-transformations $\mathbf{X}_i$). As expected, the error measures were the smallest for the TSE images due to the fact that they have similar contrasts and contents as compared to target image, and were also acquired immediately following the acquisition of the target image. The NMSE was noticeably higher for the DCE images

TABLE 3.1: Mean values of the error measures for the three types of source images. Numbers in parentheses indicate the standard deviation.

| Error measure | TSE images (TE = 36) | b0 images | Late-time point DCE images |
|---|---|---|---|
| NMSE $(\mathbf{T}_i(M^c), \mathbf{T}(M^c))$ | 0.096 (0.131) | 0.134 (0.187) | 0.331 (0.290) |
| $\left\|\mathbf{T}_i - \mathbf{T}\mathbf{X}_i^{-1}\right\|_F$ | 1.26 (1.04) | 2.50 (1.71) | 2.69 (3.75) |

than for the other two; the contrast of the DCE images is most different from the target image, and the time between its acquisition and the acquisition of the target image is also the longest of the source images. In all cases, the standard deviation of the error measure is close to or exceeds the mean. This likely reflects the variability in both the difficulty of the registration problem across different subjects and the results generated by the GA optimizer.

## 3.5 Discussion

We presented here a framework for the 3D intensity-based affine registration of mpMR images, with normalized mutual information as the similarity measure and genetic algorithms as the optimizer. Notable features of the method include 1) the identification of VOIs on which to carry out the registration, 2) estimation of the translation component with a pre-transformation matrix $\mathbf{T}_0$ to facilitate propagation of the VOIs from the target to the source images and to help the later registration step, and 3) registration-based intensity correction to compensate for the signal reception profile of the ERC. While the method is fairly reliable for correction of small movements and deformations, it tends to struggle with larger, non-linear deformations, requiring many more iterations and/or multiple restarts of the GA. The transformations found by the method are also less reproducible due in large part to the random nature of GAs.

There are three potential areas of improvement for the proposed registration framework. First, the choice of a single globally affine transformation model could be refined. While an affine (or even rigid) model for patient motion is sufficient, the deformations of the prostate from surrounding structures as well as geometric distortions

seen on the MR images (e.g., seen on the diffusion-weighted data — Fig. 3.4c) are not adequately described by linear transformations. A potential solution is to calculate a global affine transform, then define a deformable transformation model over the VOIs that contain the prostate; the same or similar concepts have been described in previous works.[87,109] Other structures within the images (e.g., vessels, muscles) that are currently considered background information may then also be used to guide the global registration.

Another aspect of the method that can be improved is the calculation of mutual information (and in turn, NMI), which currently is performed with a histogram-based method. As discussed above, histogram methods require the selection of the number/size of bins for constructing the histograms, and the calculated values are heavily influenced by that choice. Although the number of bins was picked using the Freedman-Diaconis rule in this work, it is difficult to characterize how different choices for the bin number/size would ultimately affect the registration results. Additionally, histogram-based estimates of MI are not differentiable. Using one of the numerous differentiable estimators of entropy and mutual information[217] may provide more consistent results, while also permitting the use of gradient-based optimizers.

Genetic algorithms, while flexible, are generally not the best choice for optimization problems due to long computation times, poor reproducibility, and slow convergence. Gradient-based optimizers would certainly be faster and more reliable than the GA that is currently used. While the registration problem would still be non-convex, it may be the case that local optima are, on average, good enough for this particular registration problem. There are also approaches to combine both gradient-based and gradient-free optimizers for non-convex optimization.[218]

In conclusion, the methods described here warrant further characterization and development. In particular, it may be worthwhile to investigate whether the registration methods may be applied to registration of mpMR data acquired with only surface array coils and/or at higher field strengths.

# Chapter 4

# Quantitative digital pathology for automatic identification of PCa

## 4.1 Introduction

While traditionally the diagnosis of prostate cancer (PCa) relies on the examination of prostate biopsy specimens, there is a wealth of clinically-significant information that can be gathered from the assessment of radical prostatectomy (RP) specimens, which include the refinement of diagnoses made on biopsy specimens as well as the assessment of surgical margins and extraprostatic extension (EPE).[219,220] As discussed in section 2.4.1, prostate cancer (PCa) identified on RP specimens also serves as the ideal ground truth for the development of computer-aided detection/diagnosis (CAD) systems. However, accomplishing this requires the detailed examination of hematoxylin and eosin (H&E) stained sections of RP specimens by trained pathologists, which involves the manual annotation of PCa, i.e., the detection and delineation of cancerous tissue from benign tissue. This process is not only tedious and time-consuming, but also associated with significant inter-reader, experience-dependent variability.[148,149]

The recent advent of digital pathology and whole-slide imaging systems provides an opportunity to improve the pathology annotation process.[221] Stained sections digitized by whole-slide imaging systems at high resolution can be processed and analyzed by a

variety of image analysis algorithms to extract and assess features such as stain intensity and nuclei density,[222–224] which relate to the likelihood of disease being present. These features can then be used to build a computational model to estimate the spatial distribution of disease on each whole-slide image (WSI), in effect automating the annotation process. In particular, deep learning techniques have been applied in recent years to digitized histopathologic images for the detection of a variety of cancers, including lung, prostate, breast, kidney, bladder, skin, and gastric cancer.[225–229] One limitation common to these works is that they rely solely on the analysis of H&E-stained slides for cancer detection. While H&E staining offers information about tissue morphology and architecture, it does not capture the gene expression profiles of the cells, which provides functional information that can inform disease likelihood. Therefore, image analysis of immunohistochemical (IHC) slides is a promising approach to extend and improve upon existing work.

As described in section 1.2.2, prostate adenocarcinoma is histologically defined simply by the presence of glands without the outer basal cell layer. However, the accurate annotation of PCa is challenging. PCa tends to be locally infiltrative, and distinguishing malignant glands from surrounding benign glands can be tedious. The presence of the basal cell layer is often difficult to ascertain on H&E alone, which leads to underdiagnosis.[230] Additionally, there are several pathologic entities that are mimics of PCa. The most prominent of these is prostatic intraepithelial neoplasia (PIN). While PIN itself is considered benign, high-grade PIN (HGPIN) is suggestive of the presence of invasive carcinoma.[231] To further complicate matters, HGPIN is difficult to distinguish from intraductal carcinoma of the prostate (IDC-P), which is a malignant entity that usually represents the invasion of PCa into benign glands.[232,233] For these reasons, IHC is often used in aiding pathologic diagnosis of PCa. In particular, the triple-antibody cocktail specific for high-molecular weight cytokeratin (HMWCK), p63, and $\alpha$-methylacyl CoA racemase (AMACR) is routinely used.[234] HMWCK and p63 are basal cell markers that act as negative cancer markers, i.e., the lack of immunoreactivity is indicative of the absence

of the basal cell layer.[230,235,236] On the other hand, AMACR is a positive cancer marker that is usually highly overexpressed in PCa as well as HGPIN and IDC-P.[230,237,238] The combination of these three IHC markers has been shown to be superior for demonstrating PCa than any of them individually.[234,239]

Given that IHC staining for HMWCK + p63 + AMACR has a well-established role in aiding the histological diagnosis of PCa, we developed in this work methods for automated annotation of PCa on digitized WSIs of RP specimens stained with H&E and the triple-antibody cocktail. Features were extracted from colorimetric image analysis of both H&E and IHC slides, and a regression model was trained to predict the extent and distribution of cancerous epithelium within each slide. The model was then applied to a large number of test cases, and the outputs were evaluated against slide-level manual annotation of PCa by pathologists.

## 4.2 Methods

### 4.2.1 Ethics statement

All experiments were approved under IRB protocol 0601M79888 with the University of Minnesota Institutional Review Board and carried out in accordance with approved guidelines. The IRB waived the need for informed consent for this retrospective analysis of de-identified samples.

### 4.2.2 Patient cohort

A total of 184 prostate specimens were obtained from a cohort of 63 patients who underwent radical prostatectomy for definitive treatment of biopsy-proven prostate adenocarcinoma at the University of Minnesota between November 2009 and January 2012. A summary of the patient characteristics is detailed in Table 4.1.

TABLE 4.1: Summary of the clinical and pathologic characteristics of the patient cohort.

| Parameter | Data | |
|---|---|---|
| | **Training set (n = 10)** | **Test set (n = 53)** |
| Mean age (yrs) | 61 (range: 55–72) | 63 (range: 47–76) |
| Mean serum PSA at surgery (ng/mL) | 11.3 (range: 2.5–19.4) | 7.85 (range: 0.40–37.60) |
| **Pathologic stage** | | |
| T2a | 0 | 9 |
| T2b | 0 | 4 |
| T2c | 4 | 26 |
| T3a | 5 | 10 |
| T3b | 1 | 4 |
| **Gleason score** | | |
| 3+3 | 1 | 13 |
| 3+4 | 4 | 21 |
| 4+3 | 4 | 8 |
| 4+4 | 1 | 5 |
| 4+5 | 0 | 4 |
| 5+4 | 0 | 2 |

### 4.2.3 Histopathology processing and staining

The prostate specimens were fixed and paraffin-embedded using a previously described protocol and sliced into 4 µm-thick axial sections.[106,222] From each tissue block, two sections were selected from tissue levels no more than 100 µm apart, and were de-paraffinized and rehydrated using standard methods. H&E and IHC staining was performed on the two sections, respectively. H&E staining was performed in three batches using routine clinical protocols. IHC staining was performed using the Ventana BenchMark ULTRA automated immunostainer platform (Ventana Medical Systems, Tucson, AZ). Antigen retrieval and blocking were performed as previously described.[240] Slides were incubated for 32 minutes with the triple-antibody cocktail containing primary antibodies to the basal cocktail of HMWCK + p63 (monoclonal mouse; clones 34$\beta$E12 and 4A4 respectively; prediluted; Ventana, Tucson, AZ) and AMACR (monoclonal rabbit; clone 13H4; prediluted; Dako, Glostrup, Denmark). Detection was performed with the Ventana ultraView Universal DAB Detection Kit and ultraView Universal Alkaline Phosphatase Red Detection Kit (Ventana, Tucson, AZ) according to manufacturer's

instructions. This was followed by rinsing, counterstaining with hematoxylin, dehydrating, and coverslipping. In summary, HMWCK + p63 expression in benign basal epithelium was demonstrated as brown by 3,3-diaminobenzidine (DAB), AMACR expression in malignant epithelium was demonstrated as red by Fast Red chromogen, and stroma was demonstrated as blue by hematoxylin counterstain.

### 4.2.4 Slide digitization and slide-level annotations.

Both H&E and IHC slides were digitized at 20x magnification (0.5 µm/pixel) using a whole slide scanner (Aperio ScanScope CS, Leica Biosystems, Buffalo Grove, IL). Digitized H&E WSIs were annotated at the slide-level for PCa by pathology trainees (B.M.B., A.D.J., N.P.R.) under the supervision of a board-certified pathologist (S.C.S.) using Aperio's ImageScope software (Leica Biosystems, Buffalo Grove, IL) and a pen-tablet screen (Wacom Cintiq 22HD, Saitama, Japan). The slide-level annotations were carried out by demarcating the borders of distinct regions of cancer and assigning a Gleason score (GS) to each region (Fig. 1c). Using the same tools, negative annotations, defined as regions containing artifacts of the histological processing (e.g., tissue folds, debris, irregular staining), were demarcated on the IHC WSIs by technologists (A.E.R., J.C.H.). Regions of negative annotations were ultimately excluded from analysis, and typically comprised no more than 5% of a given slide. Digitized WSIs and annotations were stored and managed as previously described.[241]

SigMap software was used to further process the digitized WSIs.[222] First, it was used to register the IHC WSI to the H&E WSI using a rigid transformation (Fig. 4.1a & 4.1b). Next, binary masks of the slide-level cancer annotations and the negative annotations were created to transfer the annotations between H&E WSIs and IHC WSIs (Fig. 4.1c). A virtual grid composed of analysis squares (dimensions 1,000 $\times$ 1,000 pixels, area of 0.25 $mm^2$) was then generated by SigMap and added to both WSIs (Fig. 4.1d). Analysis squares whose areas overlapped at least 75% with the cancer annotation mask were labeled as cancer and assigned the GS of the corresponding annotation (Fig.

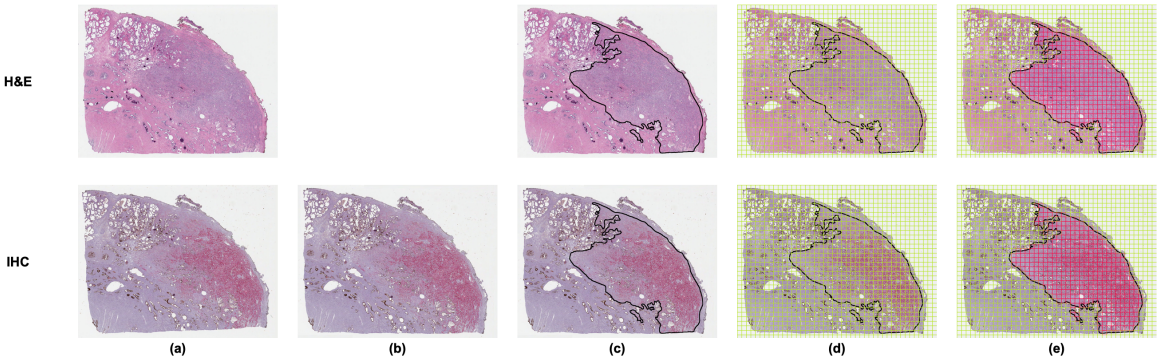



FIGURE 4.1: Use of SigMap software for initial processing of WSIs. This ensures the accurate spatial co-localization of H&E and IHC WSIs, and in turn the co-localization of image features extracted from both. **(a)** Digitized WSIs. **(b)** IHC WSI after rigid registration to the H&E WSI. **(c)** Regions of manually-annotated cancer outlined in black on the H&E WSI (GS 3+4 in this example). These regions were copied to the registered IHC WSI by SigMap. **(d)** Grid of analysis squares generated by SigMap overlaid on H&E and IHC WSIs. **(e)** Analysis squares with $\geq$ 75% overlap with the slide-level annotation identified by SigMap (in red). These analysis squares were subsequently labeled cancer and assigned the GS of the annotation.

4.1e). Analysis squares whose areas overlapped at least 75% with the negative annotation mask were excluded from further analysis.

### 4.2.5 Colorimetric image analysis algorithms

The following three quantitative image analysis algorithms (Aperio Brightfield Image Analysis Toolbox, Leica Biosystems, Buffalo Grove, IL) were configured by a technologist (J.C.H.), then applied to H&E and IHC WSIs in order to extract features for prediction of cancer.

The Positive Pixel Count (PPC) algorithm was applied to H&E WSIs. Briefly, the PPC algorithm counts the number of stained pixels within each analysis square that falls within and out of a specified range of hue-saturation-brightness (HSB) color values (positive and negative pixels, respectively). HSB values were sampled from three types of regions that predominantly contained a single histological feature of interest (nuclei, cytoplasm, or stroma). Fifteen of each type of region were manually identified on control H&E WSIs and sampled. Ranges of HSB values were calculated for each type of region and were manually adjusted to eliminate overlap between ranges. A separate PPC

algorithm was configured for each type of region and its corresponding range of HSB values. The three configured PPC algorithms were then applied prospectively to analysis squares of H&E WSIs. The resulting numbers of positive pixels were converted to percentages of the analysis square occupied by nuclei, cytoplasm, and stroma (% nuclei, % cytoplasm, and % stroma, respectively), which were in turn used as predictive features. The unstained percentage of each analysis square was also calculated as % unstained = 100% – (% nuclei + % cytoplasm + % stroma), and analysis squares with % unstained > 99% were excluded from further analysis on the basis that they are taken from regions outside of the tissue boundaries. To account for variations in H&E staining intensity across the three batches, a different set of PPC algorithms was configured and applied to each batch.

Color Deconvolution (CD) and Co-expression (CE) algorithms were applied to IHC WSIs to measure the colorimetric features of the IHC stain. Briefly, the CD algorithm isolates individual staining components of IHC WSIs for quantification, while the CE algorithm quantifies how often the staining components occur separately and together. These algorithms were first configured on control slides. Three control slides were cut, processed, and singly-stained with either DAB chromogen (brown), Fast Red chromogen (red), or hematoxylin counterstain (blue), using the same protocols as the triple-stained IHC slides described above. The average red-green-blue (RGB) optical density (OD) values of the three components were sampled from the corresponding WSIs of the control slides and were measured as Fast Red (R: 0.283, G: 0.949, B: 0.757), DAB (R: 0.461, G: 0.826, B: 1.0), and hematoxylin (R: 0.21, G: 0.276, B: 0.176), and intensity thresholds were manually configured for each component to define positively-stained pixels. The configured CD and CE algorithms were then applied prospectively to analysis squares of IHC WSIs, from which the percentage of each analysis square that was positively staining (%Pos) was calculated. As previously described, the OD quantifies the stain intensity, as it is linearly related to the amount of staining.[223,240–242]

Using the configured RGB OD and intensity threshold values, IHC WSIs were

TABLE 4.2: Summary of the extracted features. Features are calculated on an analysis-square level.

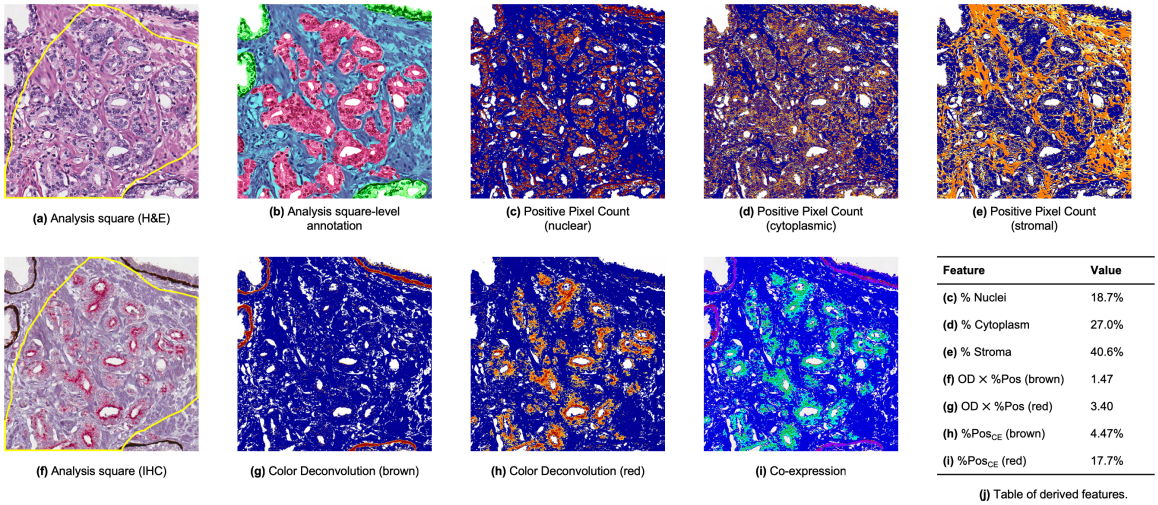
| Feature | Source | Algorithm |
|---|---|---|
| % Nuclei | H&E | Positive Pixel Count (nuclear) |
| % Cytoplasm | H&E | Positive Pixel Count (cytoplasmic) |
| % Stroma | H&E | Positive Pixel Count (stromal) |
| OD $\times$ %Pos (brown) | IHC | Color Deconvolution (brown) |
| OD $\times$ %Pos (red) | IHC | Color Deconvolution (red) |
| $\%Pos_{CE}$ (brown) | IHC | Co-expression |
| $\%Pos_{CE}$ (red) | IHC | Co-expression |

then separated into brown, red, and blue color channels corresponding to each staining component. The brown and red staining were separately quantified by the CD algorithm as previously described.[241] Specifically, the average OD and %Pos were measured by the CD algorithm for both brown and red components, and the products OD $\times$ %Pos were calculated and used as predictive features. The co-localization of brown and red staining was quantified by the CE algorithm, which was then used to calculate the percentage of the analysis square that was positively staining for only red or only brown, but not both ($\%Pos_{CE}$). $\%Pos_{CE}$ for red and brown components were used as predictive features.

In summary, seven features were extracted from each analysis square (Table 4.2). The features derived from H&E WSIs were the percentages of nuclei, cytoplasm, and stroma, while the features derived from IHC WSIs were the percentages and stain intensities (quantified by the OD) of brown and red staining, which corresponded to the characteristics of the basal cell staining (HMWCK + p63) and the AMACR staining, respectively.

### 4.2.6 Training data and analysis square-level annotations

Ten of the 63 patients in our cohort were randomly selected, and one pair of WSIs was created from each for purposes of training the regression model. Forty analysis squares were randomly selected from each of the ten pairs of WSIs (400 analysis squares in total) and were manually annotated in much greater detail than usual (S.C.S.).




| Feature | Value |
|---|---|
| **(c)** % Nuclei | 18.7% |
| **(d)** % Cytoplasm | 27.0% |
| **(e)** % Stroma | 40.6% |
| **(f)** OD × %Pos (brown) | 1.47 |
| **(g)** OD × %Pos (red) | 3.40 |
| **(h)** $\%Pos_{CE}$ (brown) | 4.47% |
| **(i)** $\%Pos_{CE}$ (red) | 17.7% |

**(j)** Table of derived features.

FIGURE 4.2: Examples of pseudo-color outputs of Aperio image analysis algorithms from which the predictive features were calculated. **(a)** Analysis square from an H&E WSI in the training set (75% overlap with the slide-level annotated cancer outlined in yellow). **(b)** Analysis square-level annotation of (a), with benign epithelium in green, malignant epithelium in red, gland lumens in white, and stroma in blue. The percentage of malignant epithelium is used as the ground truth for training. **(c – e)** Outputs of the PPC algorithms. Positive pixels are in red/orange/yellow, and negative pixels in blue. The percentages of positive pixels is taken to be the percentages of the analysis square occupied by nuclei, cytoplasm, or stroma. **(f)** Analysis square corresponding to (a) taken from the corresponding IHC WSI. **(g & h)** Outputs of the CD algorithms. Positive pixels are in yellow/orange/red, and negative pixels in blue. **(i)** Output of the CE algorithm. Positive pixels for red and brown components are in green-cyan and red-purple, respectively. **(j)** Table of features derived from the outputs of the image analysis algorithms that are used as inputs for the regression model.

The analysis square-level annotations were carried out by meticulously delineating the benign and cancerous epithelium, gland lumens, stroma, and regions of clear glass within the 1,000 $\times$ 1,000 pixel-area of each analysis square. The fractional areas of each of the aforementioned components were then summed for each annotated analysis square (Fig. 4.2b). The percentage of cancerous epithelium within each analysis square-level annotation was taken to be the ground truth. Details on the slides can be found in Table 4.3.

### 4.2.7 Regression model training and evaluation

Elastic net regression models were trained on these data using 10-fold cross validation, with each fold containing the 40 analysis squares from a single pair of WSIs. The elastic net is a generalized linear regression model with both L1 and L2 regularization, and its corresponding objective function to be minimized is

$$\min_{w} \frac{1}{2m} \|Xw - y\|_2^2 + \alpha\rho \|w\|_1 + \frac{\alpha(1-\rho)}{2} \|w\|_2^2 \tag{4.1}$$

where $m$ is the number of training examples, $n$ is the number of features, $X$ is the $m$-by-$n$ matrix of training examples, $w$ is the $n$-by-$1$ vector of feature weights, $y$ is the $m$-by-$1$ vector of labels, and $\alpha$ and $\rho$ are parameters that determine the strengths of the regularization terms. Elastic net regression was implemented with the Scikit-learn package in Python.[243]

Models were trained on four different sets of features: 1) features from H&E WSIs alone (the H&E model), 2) features from IHC WSIs alone (the IHC model), 3) features from both H&E and IHC WSIs, but without the two $\%Pos_{CE}$ features (the $full_{\text{-}CE}$ model), and 4) all features from both H&E and IHC WSIs (the full model). Given the similarity of %Pos from the CD algorithm and $\%Pos_{CE}$ from the CE algorithm, both the $full_{\text{-}CE}$ model and the full model were included in order to test if the inclusion of the two $\%Pos_{CE}$ features would provide any benefit to cancer identification accuracy.

TABLE 4.3: Breakdown of the distribution of the analysis squares of the training and test data by cancer presence and Gleason score. An analysis square was labeled cancer if it overlapped at least 75% with the slide-level annotation. Excluded analysis squares (i.e., those that overlapped at least 75% with the negative annotation, or were found to be > 99% unstained on H&E staining) are not tabulated here. Numbers in parentheses indicate the number of pairs of WSIs containing cancer with the corresponding Gleason score. Note that some WSIs contained no annotated cancer (1 in the training set, 39 in the test set).

| **Type** | **Training (10 total)** | **Test (174 total)** |
|---|---|---|
| **Cancer** | **84** | **23,757** |
| **3+3** | 13 (1) | 2,849 (31) |
| **3+4** | 37 (4) | 6,146 (47) |
| **4+3** | 34 (4) | 4,452 (22) |
| **4+4** | 0 (0) | 2,790 (15) |
| **4+5** | 0 (0) | 6,146 (16) |
| **5+4** | 0 (0) | 1,374 (4) |
| **Benign** | **316** | **189,629** |
| **Totals** | **400** | **213,386** |

For each model, the coefficients of the two regularization terms ($\alpha$ and $\rho$) were treated as hyperparameters and selected by cross-validation to minimize the mean value of the objective function averaged across the ten folds. Trained models were then applied to the analysis squares of the other 174 pairs of slides to produce predicted maps of cancerous epithelium. Model outputs were compared to the slide-level annotations on a per-analysis square level using receiver operating characteristic (ROC) curve analysis. Sensitivities and specificities were calculated using the optimum cut-off points for the ROC curves that corresponded to the maxima of the Youden indices. The 95% confidence intervals (CIs) were calculated using a bootstrap procedure that resampled both WSIs and analysis squares from the training set, and only WSIs from the test set. Two-sided $p$-values were found by inverting the 95% bootstrap CIs.

TABLE 4.4: Comparison of the cross-validation performance for the four regression models.

| Model | Root mean square error | Median absolute error | Maximum absolute error |
|---|---|---|---|
| **H&E model** | 15.4 | 8.37 | 52.3 |
| **IHC model** | 9.36 | 3.55 | 49.9 |
| **Full$_{\text{-CE}}$ model** | 11.9 | 5.59 | 49.4 |
| **Full model** | 8.37 | 3.09 | 38.8 |

## 4.3 Results

### 4.3.1 Outputs of colorimetric image analysis algorithms

Figures 4.2a and 4.2f illustrate the detail at the level of an analysis square for H&E and IHC WSIs, respectively. Figures 4.2c–e show sample outputs of the PPC algorithms. Figures 4.2g and 4.2h show sample outputs of the CD and CE algorithms, respectively. As the degree of co-localization between the brown and red components of the deconvolved IHC slides is generally small (typically $<$ 5% of the analysis square), the overlap is difficult to visualize in Figure 4.2i.

### 4.3.2 Quantitative evaluation of model performance

Cross-validation performance for the four models was evaluated by plotting cumulative scatterplots of predicted % cancer epithelium vs. the actual % cancer epithelium across the 10 cross-validation folds (Figure 4.3). The cross-validation root mean square error, median absolute error, and maximum absolute error for each model are shown in Table 4.4.

Performance on the test set for the four models was evaluated by plotting the ROC curves (Figure 4.4). The area under the ROC curves (AUCs), as well as the sensitivities and specificities calculated at the respective maxima of the Youden indices, are shown in Table 4.5.

The AUC for the full model was significantly higher than that of the H&E model ($p$ = 0.026), while the specificity for the full model was significantly higher than those of the H&E and full-CE models ($p$ < 0.001 for both). The AUC and specificity for the full model

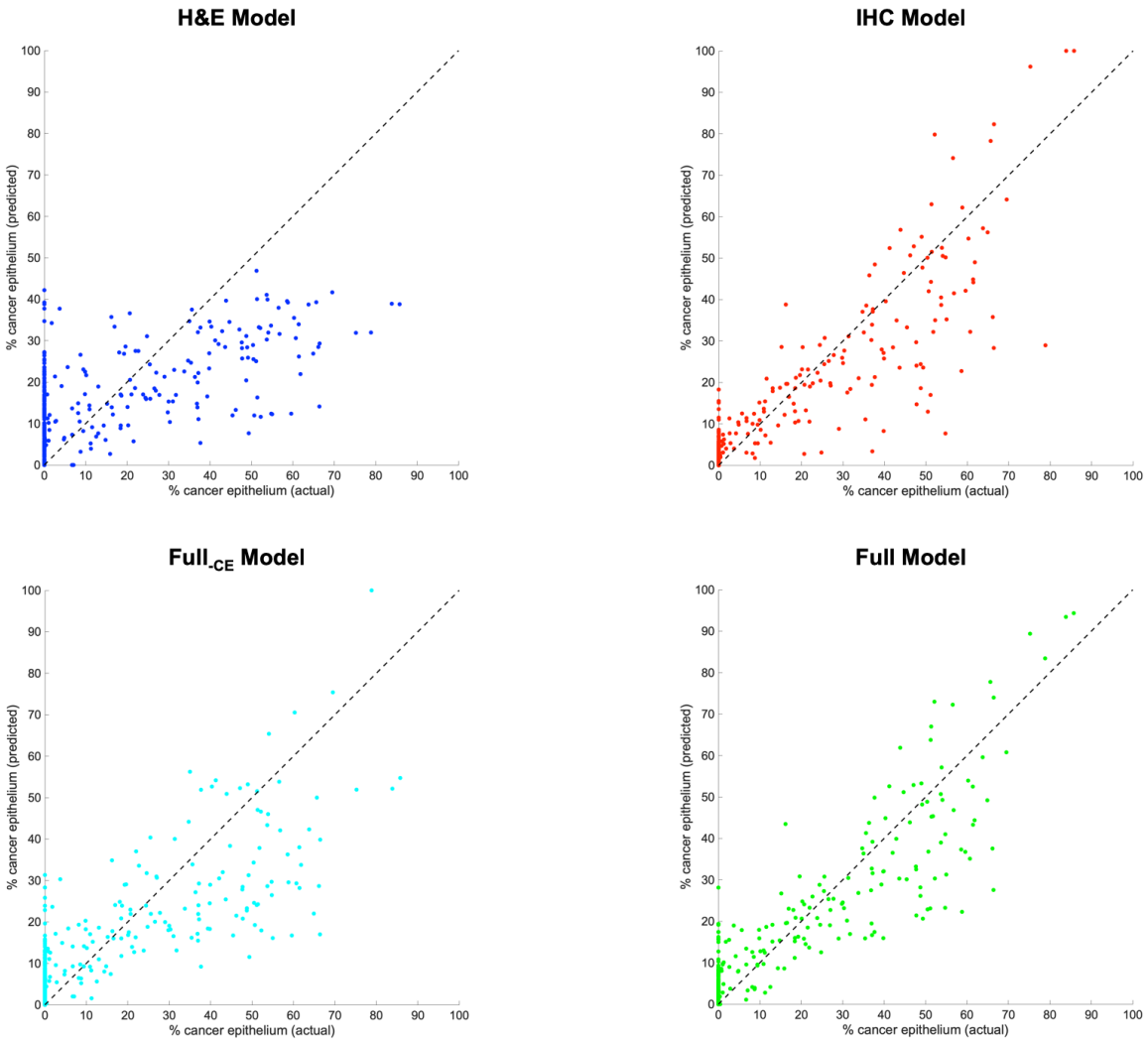



FIGURE 4.3: Scatterplots of the predicted vs. actual % cancer epithelium for the four regression models trained with different feature sets. Data points in each plot were accumulated across the ten cross-validation folds.

TABLE 4.5: Comparison of classification performance for the four regression models. Numbers in brackets are the 95% bootstrap confidence intervals generated from 1,000 bootstrap samples.

| Model | AUC | Sensitivity | Specificity |
|---|---|---|---|
| **H&E model** | 0.755 [0.582, 0.867] | 0.661 [0.562, 0.898] | 0.760 [0.665, 0.803] |
| **IHC model** | 0.937 [0.692, 0.961] | 0.918 [0.661, 0.931] | 0.920† [0.780, 0.938] |
| **Full-CE model** | 0.911 [0.682, 0.943] | 0.907 [0.683, 0.924] | 0.809 [0.765, 0.864] |
| **Full model** | 0.951† [0.832, 0.964] | 0.871 [0.753, 0.934] | 0.907†,* [0.894, 0.959] |

† Significant at $p < 0.05$ compared to the H&E model.

*Significant at $p < 0.05$ compared to the full-CE model.

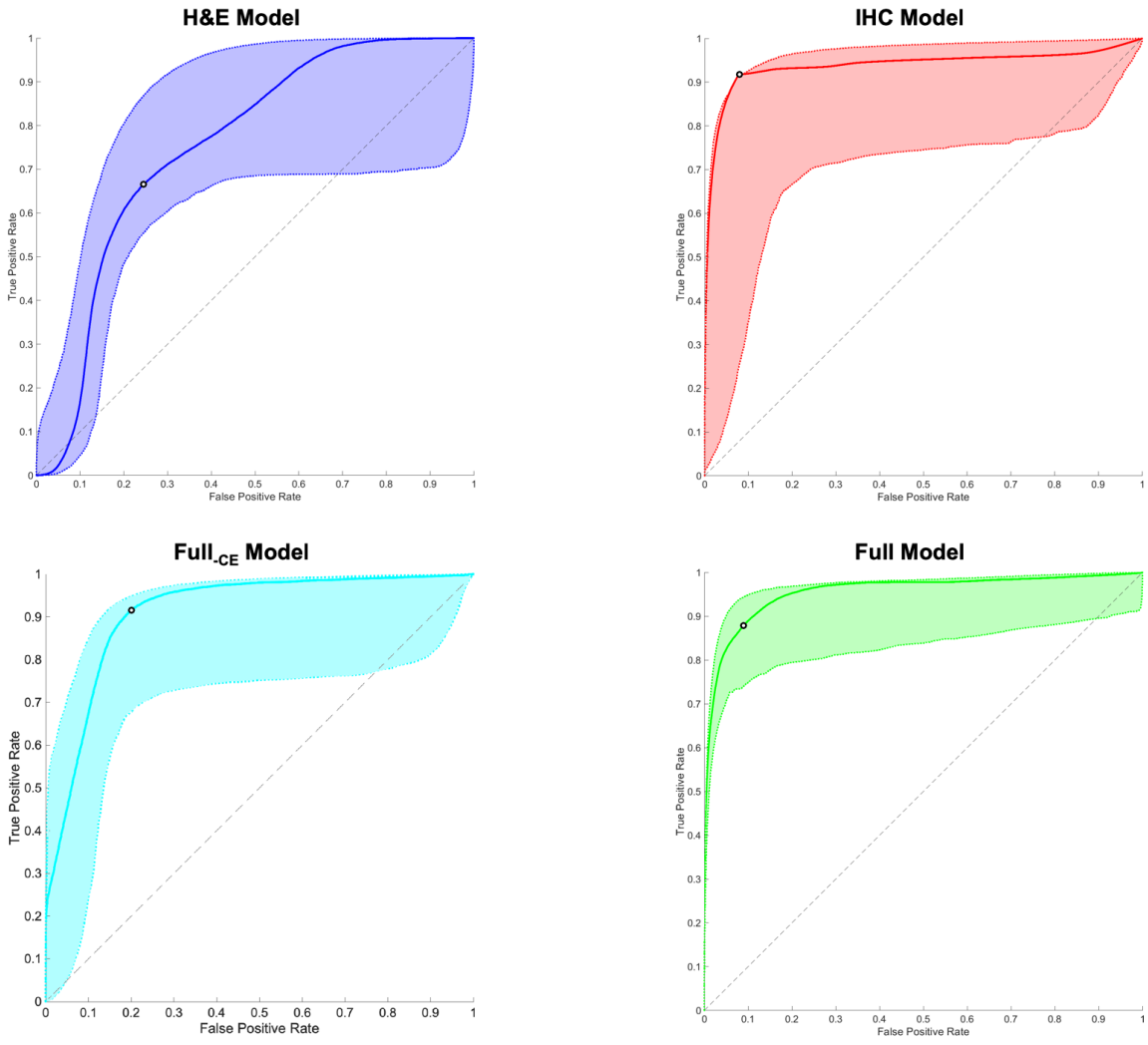



FIGURE 4.4: Receiver operating characteristic (ROC) curves for the regression models trained with different feature sets. Shaded regions correspond to the 95% bootstrap confidence intervals generated from 1,000 bootstrap samples. Black circles indicate the maxima of the Youden indices, which were chosen as the cutoff points.

TABLE 4.6: Sensitivity of the full model broken down by Gleason score and Gleason grade groups. Numbers in brackets are the 95% bootstrap confidence intervals generated from 1,000 bootstrap samples.

| Type | Number of Analysis Squares | Number Correctly Labeled | Sensitivity |
|---|---|---|---|
| **3+3** | 2,849 | 2,411 | 0.846 [0.784, 0.957] |
| **3+4** | 6,146 | 5,539 | 0.901 [0.721, 0.954] |
| **GG $\leq$ 2** | 8,995 | 7,950 | 0.884 [0.792, 0.971] |
| **4+3** | 4,452 | 4,098 | 0.921 [0.788, 0.956] |
| **4+4** | 2,790 | 2,246 | 0.805 [0.732, 0.976] |
| **4+5** | 6,146 | 5,135 | 0.836 [0.601, 0.891] |
| **5+4** | 1,374 | 1,274 | 0.927 [0.715, 1] |
| **GG $\geq$ 3** | 14,762 | 12,753 | 0.864 [0.727, 0.934] |
| **Totals** | 23,757 | 20,703 | 0.871 [0.742, 0.929] |

were not significantly different than those of the IHC model ($p$ = 0.542 and $p$ = 0.108, respectively). The sensitivity of the full model was also not significantly different than those of the H&E, IHC, and full$_{\text{-CE}}$ models ($p$ = 0.134, $p$ = 0.748, and $p$ = 0.939, respectively). The CIs of these summary statistics for the full model were notably narrower than those of the other three models, suggesting that the performance of the full model will likely be closer to what is reported here when it is applied prospectively.

For the full model, the sensitivity of PCa detection was broken down (Table 4.6) by both Gleason score and Gleason grade group (GG). Using the convention that GG $\leq$ 2 (GS 3+3 or 3+4) is low to intermediate-grade and GG $\geq$ 3 (GS 4+3, 4+4, 4+5, or 5+4) is high-grade, the sensitivity of detecting low to intermediate-grade cancers was 0.884, while it was 0.864 for high-grade cancers; this difference was found to be not significant ($p$ = 0.107).

### 4.3.3 Comparison of model-generated annotations to manual slide-level annotations.

Figure 4.5 shows representative H&E and IHC slides with slide-level, manually-annotated cancer by pathologists compared with maps generated by the full model. Note the high

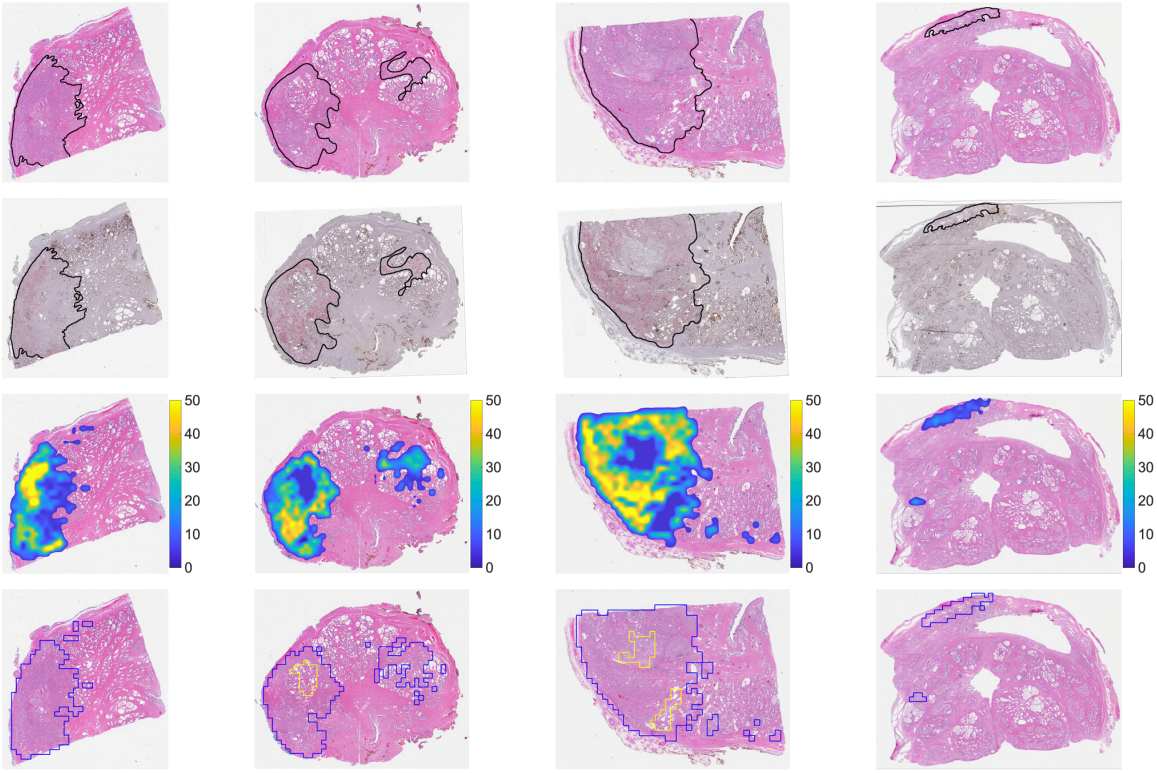



FIGURE 4.5: Representative comparisons of slide-level annotations to model-generated prediction maps. **Row 1**: H&E WSIs with slide-level annotations outlined in black. **Row 2**: IHC WSIs corresponding to the H&E WSIs in Row 1. Sigmap software was used to perform registration to the H&E WSIs, and to copy the annotated cancer. **Row 3**: Model-generated maps of the predicted distribution of malignant epithelium overlaid on the H&E WSIs. Colorbars correspond to the percentage of malignant epithelium. **Row 4**: Thresholded versions of prediction maps shown in Row 3, with the Youden index (2.61%) chosen as the threshold. Predicted slide-level annotations are outlined in blue, with internal benign regions outlined in yellow.

degree of correlation between the annotated cancer, distribution of AMACR staining (red on IHC slides), and predicted distribution of malignant epithelium.

## 4.4 Discussion

In this work, we show that a predictive model that uses features derived from colorimetric analysis of both digitized H&E and IHC slides is able to detect and delineate PCa on WSIs with accuracy comparable to pathologists' slide-level annotations. The performance of the full model was found to be superior to those of the other three models, individually

(Table 4.5), though this difference was only significant in comparison to the H&E model. Furthermore, despite the relatively small amount of training data that included predominantly low and intermediate-grade cancers, the model performed well across a large number of test set slides. Its sensitivity was also largely consistent across cancers with different Gleason scores, and was only slightly worse for high-grade cancers (0.864 vs. 0.871 for all cancers, Table 4.6).

In contrast to most published works, the regression model described here uses a compact set of seven features that were calculated from the outputs of standard image analysis algorithms applied to H&E WSIs (% nuclei, % cytoplasm, % stroma) and IHC WSIs (OD $\times$ %Pos and $\%Pos_{CE}$, for brown and red). These features were ultimately chosen for their simplicity and interpretability. Although %Pos from the CD algorithm and $\%Pos_{CE}$ from the CE algorithm appeared to be redundant, the results demonstrate that excluding the two $\%Pos_{CE}$ features resulted in worse specificity of cancer detection ($full_{\text{-}CE}$ model vs. full model, Table 4.5). This is most likely due the fact that some slides contained a larger fraction of non-cellular components (e.g., intraglandular cellular debris, corpora amylacea) that stained both brown and red with IHC staining. This would cause %Pos (red) as calculated by the CD algorithm to be falsely elevated, but would not affect $\%Pos_{CE}$ (red) as calculated by the CE algorithm as the CE algorithm excludes regions that stained both brown and red. Therefore, inclusion of the two $\%Pos_{CE}$ features increased the specificity of cancer detection for slides containing a significant fraction of such regions, and in turn increased the overall specificity of cancer detection.

Another notable aspect of the work is that the full model was trained to predict the percentage of cancerous epithelium, which was made possible by the unique ground truth obtained from the meticulous annotation of individual analysis squares. For purposes of model training, this ground truth is superior to the slide-level annotations, as those are known to have finite accuracy and precision.[148,149]

The trained models were evaluated against the slide-level annotations on a per-analysis square level using ROC curve analysis. However, despite the high AUC

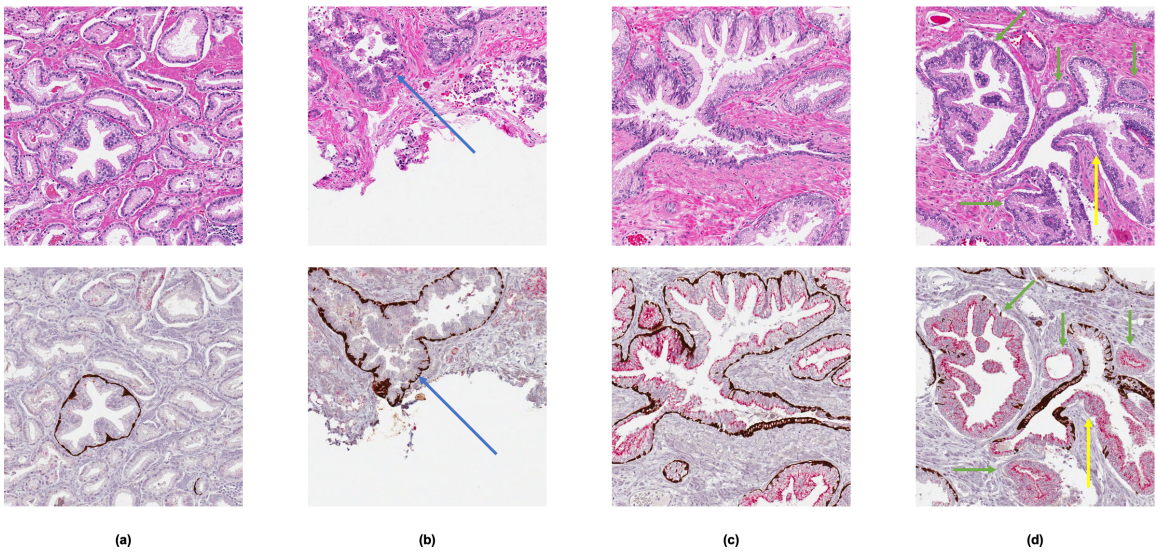



FIGURE 4.6: Illustrative examples of analysis squares with discrepancies between the slide-level annotation and the model output. The top row shows H&E slides, and the bottom row shows the corresponding IHC slides. **(a)** Analysis square with predominantly malignant glands that have poor AMACR staining. 100% overlap with the slide-level annotation, but incorrectly labeled as non-PCa by the model. **(b)** Analysis square with PIN that has poor AMACR staining (blue arrow) and a large cystic region (bottom half). Correctly labeled non-PCa by the model, but had 100% overlap with slide-level annotation. **(c)** Analysis square with PIN that has strong AMACR staining. Did not overlap with the slide-level annotation, but was incorrectly labeled as PCa by the model. **(d)** Analysis square with small malignant glands (green arrows), including an example of HGPIN/IDC-P (yellow arrows). Correctly labeled PCa by the model, but did not overlap with the slide-level annotation.

achieved by the full model, it is difficult to assess the true performance of the model due again to the limitations of the slide-level annotations. Accurate assessment would require analysis square-level annotations of all the slides in the test set, which would be prohibitive. Qualitatively, visual comparison of the model-generated maps of cancerous epithelium with the slide-level annotations shows generally good concordance (Fig. 4.5). Sources of disagreements between the two can be divided into four categories, which are illustrated in Figure 4.6:

1. Cancer missed by the model (Fig. 4.6a). This was most often due to cancer with poor AMACR staining; while AMACR is a sensitive positive marker of PCa, it is well-documented that some variants of PCa do not exhibit increased expression of AMACR.[238,239,244,245] Alternatively, inconsistencies in the staining procedure may

have caused variabilities in AMACR staining, and these variabilities would be amplified in regions of cancer.

2. Cancer incorrectly annotated by the pathologist (Fig. 4.6b). This was most often due to large regions of glass (e.g., cystic areas, luminal areas of malignant glands) that are looped in with the slide-level annotations. More rarely, benign glands were incorrectly annotated as cancer; usually, these were examples of PIN with low AMACR expression.

3. Cancer incorrectly labeled by the model (Fig. 4.6c). This was most often due to PIN with high AMACR expression.

4. Cancer missed by the pathologist (Fig. 4.6d). This was most often due to small, isolated regions of cancer that were not annotated. More rarely, HGPIN and/or glands with IDC-P were missed by the pathologist but identified as cancer by the model due to high AMACR expression.

In summary, accurately distinguishing the different possible presentations of PCa from PIN is a challenge for both pathologists and the full predictive model. Although in theory glands with PIN are characterized by the presence of an intact basal cell layer, the basal cell layer may be quite fragmented, which would make it difficult to assess by either visual inspection or by quantitative assessment of brown staining intensity.

There are three major limitations of the features. First, since they are calculated from colorimetric analysis of stained WSIs, their consistency is highly-dependent on the reproducibility of the staining procedure and digitization process. As noted in previous works, the use of different histology protocols and/or different slide scanners can cause large variations in the morphological features of WSIs, which in turn degrade the predictive performance of trained models.[246] In our work, the three separate batches of H&E staining presented a major source of potential variability in the calculated H&E features, as the stain intensities were visibly different between H&E WSIs of different batches. To compensate, a different set of PPC algorithms was configured for each batch,

though this was not ideal. In order to minimize batch effects in the future, H&E staining will also be performed on an automated platform using a standardized protocol, like what was done for the IHC staining. Additionally, algorithms like the PPC algorithm that rely on the analysis of intensity values (e.g., RGB or HSB values) are naturally quite prone to variations in stain intensity. Therefore, it would be worth extending the use of color deconvolution algorithms for the analysis of H&E WSIs, as proposed in previous works.[247] To expand this work to larger datasets, further methods for normalization of WSIs and/or derived features may also be investigated.[246,247]

Another limitation of the features is that they are derived from WSIs of different tissue levels of the tissue block. Due to technological constraints, H&E and IHC staining were not performed on the same tissue section, and thus two sections were taken from each specimen. Although the sections were spatially adjacent (separated by $\leq 100\,\mu m$), there were sometimes noticeable differences between the two when digitized and viewed at the magnification level of individual analysis squares (Fig. 4.2a and 4.2f; Fig. 4.6b). However, these differences were relatively minor and unlikely to significantly affect the calculated features, and can further be minimized in the future by always selecting serial levels of the tissue block for staining. Methods for visualizing multiple stains on the same tissue section may also be considered.[248,249]

Lastly, the features only characterize the composition within each analysis square, and not the arrangement (i.e., cellular architecture) of the components. Therefore, the predictive model has difficulties distinguishing between analysis squares containing PIN and those containing a mixture of benign and malignant glands. This limitation may be addressed in future work by identifying additional IHC markers that are differentially expressed in cancer and PIN; for example, IDC-P is characterized by decreased expression of PTEN, which can be used to distinguish HGPIN from IDC-P.[250] An alternative approach could be to develop custom algorithms for object detection and segmentation (e.g., for identification of whole prostate glands). A more straightforward approach could be to supplement the training data with examples of PIN or PIN-like

entities. Augmenting the training data may also allow the use of deep learning approaches such as convolutional neural networks that can learn features that account for differences in the glandular architecture within analysis squares.

In summary, the methods introduced in this work can be modularly integrated into digital pathology frameworks for detection of prostate cancer on whole-slide images of histopathology slides. The unique aspect of this work is that it incorporates information from slides with conventional H&E staining as well as those with IHC staining, and as demonstrated in this work, the combination of both allows for more accurate identification of prostate cancer. Given the number of previously identified and characterized genetic markers in other types of cancers, the methods presented here may be extended naturally to other types of cancer as well.

# Chapter 5

# Detection and grading of PCa using qMRI and radiomic features

## 5.1 Introduction

As discussed in section 2.6, while numerous works have described approaches for computer-aided detection (CADe) of prostate cancer (PCa) with mpMRI data, far fewer have addressed the problem of predicting the aggressiveness of detected cancer, which falls under computer-aided diagnosis (CADx). However, as discussed in section 1.3, determining whether detected cancer is aggressive (i.e., likely to become life-threatening if left untreated) is of much higher clinical significance than simply detecting it.

We developed in this work a novel two-stage classifier for simultaneous cancer detection and assessment of cancer aggressiveness. The first-stage voxel-wise classifier uses a combination of radiomic and quantitative MRI (qMRI) features to perform CADe. After processing the voxel-wise predictions to generate discrete candidate regions, the second-stage region-wise ordinal classifier performs CADx by categorizing the derived regions as non-cancer, low-grade cancer, or high-grade cancer. Through this approach, the candidate regions are identified automatically from voxel-wise predictions, avoiding the bias of manually-identifying candidate regions (section 2.6). We demonstrate that using the combination of both radiomic and qMRI features in the first stage improves

classification performance, compared to using either alone. We also demonstrate that the overall performance of the classifier is comparable to that of a radiologist in terms of accuracy of detecting high-grade PCa.

## 5.2 Methods

The mpMRI data used for this work were acquired as previously described,[106] In total, 46 axial slices of interest from 34 patients were included for the development of the methods. Please refer to Appendix A for more details.

### 5.2.1 Derivation of qMRI and radiomic features

The calculation of the 5 qMRI features (T2 value, ADC, $K^{trans}$, $k^{ep}$, and AUGC90) is described in Appendix A.

Prior to extraction of radiomic features, intensity correction was performed on the data using the methods described in Chapter 3. Both voxel-wise and region-wise radiomic features were calculated using the PyRadiomics package in Python.[167] Features were extracted from each axial slice on the original versions of the T2W image, the ADC map, and the calculated high $b$-value diffusion-weighted image, as well as on the edge-enhanced versions of each obtained through the application of a Laplacian of Gaussian (LoG) filter with standard deviation $\sigma = 1$. Features extracted include first-order statistics ($n$ = 18); texture features calculated from the gray-level co-occurrence matrix (GLCM, $n$ = 24), the gray-level run length matrix (GLRLM, $n$ = 16), the gray-level size zone matrix (GLSZM, $n$ = 16), the gray-level dependence matrix (GLDM, $n$ = 14), and the neighboring gray tone difference matrix (NGTDM, $n$ = 5); and shape-based features ($n$ = 10) from regions only. A detailed description of these features can be found in the documentation for PyRadiomics.[167]

In summary, a total of 563 voxel-wise features (5 qMRI + 558 radiomic) and 588 region-wise features were extracted.

### 5.2.2 Feature selection

Feature selection was performed to discard features with poor predictive power. The feature selection pipelines for the two classification stages were identical. Unpaired t-tests between the feature values of cancer-labeled voxels and those of non-cancer voxels were first performed for each feature. Pearson correlation coefficients ($\rho$) were then calculated for all pairs of features, and for each pair with $|\rho|$ > 0.75, the feature with the larger $p$-value on the t-test was discarded; this process was repeated iteratively until $|\rho| \leq 0.75$ for all pairs of remaining features. Lastly, the minimum redundancy maximum relevance (mRMR) algorithm[251] was applied to select the final set of features, where the number of features was determined through cross validation (see below).

### 5.2.3 Model training: voxel-wise classifier

For the voxel-wise classification stage, Random Forest classifiers were trained using leave-one-patient-out cross validation with the Scikit-learn package in Python.[243] Classifiers were trained on qMRI features alone (without feature selection), radiomic features alone, and the combination of both. Receiver operating characteristic (ROC) curves were constructed using the probabilistic outputs of the classifiers, and the number of features selected by mRMR as well as the model hyperparameters (number of trees, tree depth, and number of features to consider for splitting) were chosen to maximize the detection sensitivity at a fixed specificity of 0.90. The relatively high specificity was chosen so as to minimize the appearance of small isolated candidate regions for the second stage. This trained model was then used to generate maps of predicted cancer for the 46 axial slices.

### 5.2.4 Automated derivation of candidate regions from voxels

To generate candidate regions from cancer-labeled voxels, 2-dimensional (2D) morphological opening (i.e., image erosion followed by image dilation) with a 2 $\times$ 2 square

structuring element was applied to each prediction map. Distinct contiguous voxels on the resulting image were identified as candidate regions. A size threshold of 75 voxels was applied to remove small regions, as they are more likely to be benign tissue or low-grade, clinically-insignificant disease.[252] Labels were then assigned to each remaining candidate region by comparing the overlap of the voxels of the region to the labeled voxels of the ground truth. A label of low-grade PCa or high-grade PCa was assigned if the majority of the overlapping voxels had Gleason score (GS) = 3+3 or GS ≥ 3+4, respectively, while a label of non-cancer was assigned if there were no overlapping voxels (i.e., false positive).

### 5.2.5 Model training: region-wise classifier

To augment the number of examples for training the region-wise classifier, 4,600 synthetic prediction maps (100 for each of the 46 ground truth maps) were randomly generated with random voxel-wise sensitivity and specificity using similar methods to those described in Appendix C. Candidate regions present in these synthetic prediction maps (6,853 in total) were identified and labeled, and radiomic features were extracted and selected from each candidate region. Two Random Forest-based classifiers were trained using these features, both using leave-one-patient-out cross validation. First, ensembles of Random Forest classifiers were used to construct a three-class ordinal model (with ordering non-PCa — low-grade PCa — high-grade PCa) using previously described methods.[253] Next, to focus on the identification of high-grade PCa, a binary classifier was trained to classify each candidate region as high-grade PCa or not.

### 5.2.6 Model evaluation

The two-stage model was applied to the candidate regions derived from the predictions of the first-stage voxel-wise classifier and evaluated against the ground truth labels of each region. The cross-validation performance of the multi-class region-wise classifier was evaluated using the quadratic weighted kappa score. The cross-validation performance of the binary region-wise classifier was evaluated using ROC curve analysis, and the

TABLE 5.1: Comparison of classification performance for the voxel-wise classifiers trained on three different set of features. Cut-offs of the ROC curves were set at a specificity of 0.90.

| Feature set | AUC | Sensitivity |
|---|---|---|
| **qMRI features** | 0.788 | 0.490 |
| **Radiomic features** | 0.749 | 0.387 |
| **qMRI + radiomic features** | 0.818 | 0.526 |

sensitivity and specificity was calculated using the cut-off point that corresponded to the maximum of the Youden index.

The two-stage model was also evaluated against the retrospective, manual annotations of radiologically-significant disease on the 34 cases by a single attending radiologist (B.S.) with 7 years of experience in prostate MRI, as described in 6.2.4. Briefly, all cases were interpreted by the radiologist in accordance with PI-RADS v2 guidelines.[2] A total of 0-3 clinically-significant lesions, as defined by a PI-RADS score of 3+, were annotated on each of the 46 axial slices. The total numbers of true positive, false negative, and false positive lesions were tabulated for the two-stage model and the radiologist.

## 5.3 Results

The cross-validation performance of the voxel-wise classifier on the three different sets of features is summarized in Table 5.1. The corresponding ROC curves are shown in 5.1. It was found that the AUC and voxel-wise sensitivity of the classifier at a fixed specificity of 0.90 was higher with the combination of both qMRI and radiomic features as compared to with either alone.

A quadratic weighted kappa score of 0.374 and an overall accuracy (across all three classes) of 0.458 were achieved by the ordinal region-wise classifier. The corresponding confusion matrix that summarizes the results is shown in Figure 5.2. The predictions were notably less precise for low-grade (GS = 3+3) disease compared to non-cancer and high-grade (GS $\geq$ 3+4) disease.

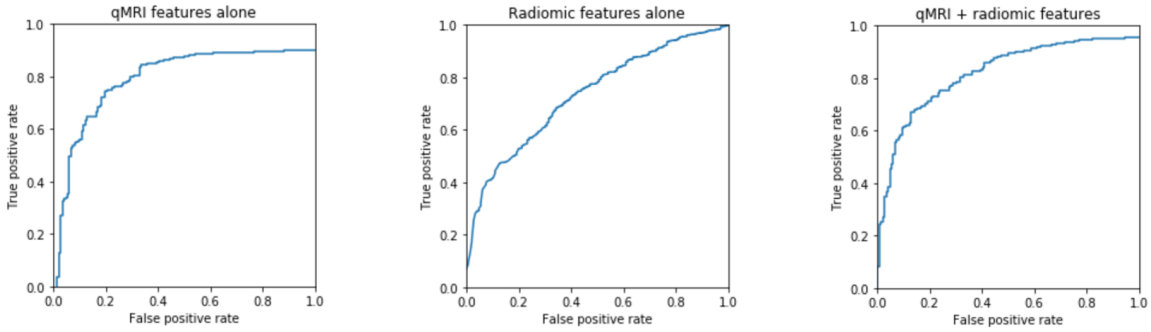



FIGURE 5.1: ROC curves for the voxel-wise classifiers trained with different feature sets.

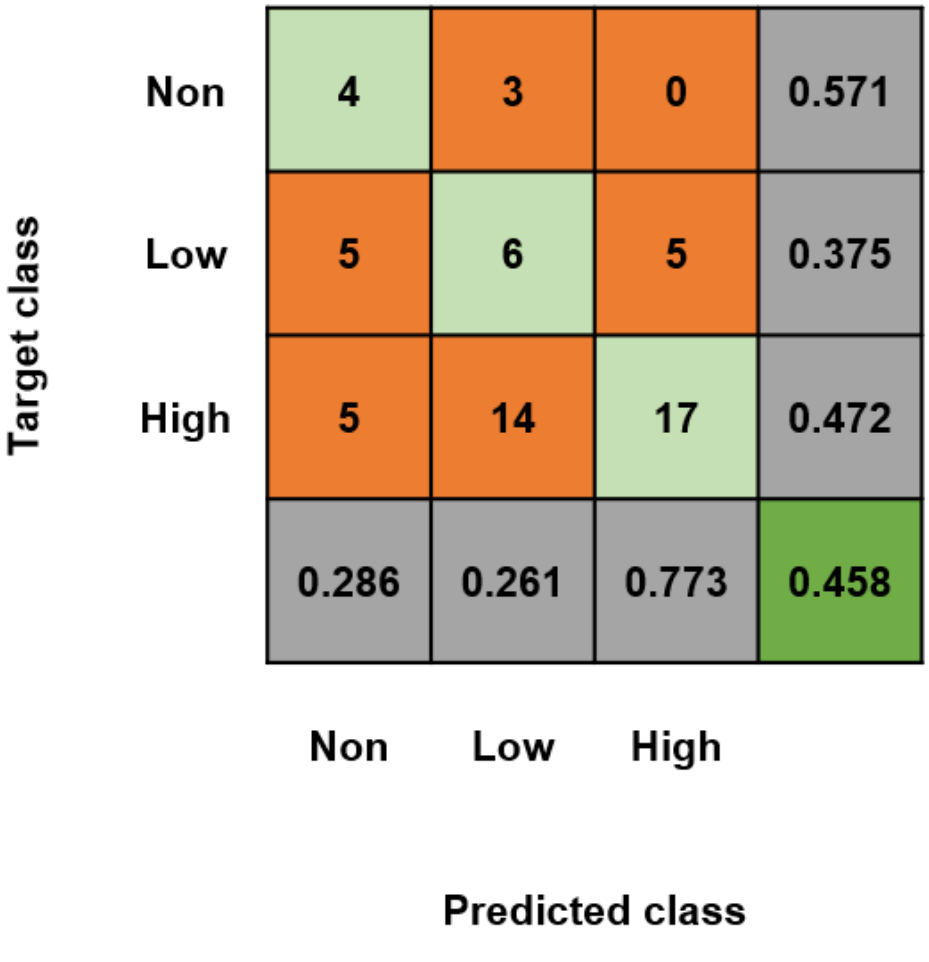



FIGURE 5.2: Confusion matrix for the ordinal classifier.

TABLE 5.2: Comparison of lesion classification accuracy for the two-stage model and radiologist annotations. A true positive (TP) lesion was defined as a high-grade (GS $\geq$ 3+4) lesion for the purposes of this analysis.

| | **Two-stage model** | **Radiologist annotation** |
|---|---|---|
| **TP lesions** | 31 | 34 |
| **FP lesions** | 11 | 7 |
| **FN lesions** | 10 | 7 |

The binary region-wise classifier achieved an AUC of 0.805 as well as sensitivity and specificity of 0.747 and 0.838, respectively, corresponding to the maximum of the Youden index. A comparison of the two-stage model to radiologist annotations is summarized in Table 5.2.

## 5.4 Discussion

In this work, we developed a novel two-stage classifier for detection and assessment of prostate cancer aggressiveness using a combination of qMRI and radiomic features derived from prostate MRI data. The performances of both classifier stages were on par with those of previously-published studies, and the overall performance of classifier was comparable to that of a radiologist in terms of accuracy of detecting high-grade lesions.

The first-stage voxel-wise classifier achieved a voxel-wise AUC of 0.818, which is comparable to the performance reported by previous studies.[106,108,132,133,135,138,139] While there is a myriad of different set of features that have been used for voxel-wise classification, to our knowledge our work was the first that considered the combination of both qMRI and radiomic features. We found that both the AUC and voxel-wise sensitivity of the classifier at a fixed specificity of 0.90 was higher with the combination of both qMRI and radiomic features as compared to with either alone (0.788 and 0.749, respectively), suggesting that there is value to using both kinds of features for prostate CAD models.

The ordinal region-wise classifier achieved a quadratic weighted kappa score of 0.374 and an overall accuracy of 0.458 across all three classes, which for purposes of evaluating prediction quality can be considered fair performance[254] and is only slightly

inferior to the reported performance of several other models[145,201,210] that were trained with two to three times more data. The performance of the ordinal classifier was notably worse for low-grade (GS = 3+3) disease. This is likely due to the fact that 1) the differences between low-grade disease and benign prostate tissue are known to be more subtle than the differences between low- and high-grade disease,[2] and 2) a relatively small percentage of low-grade lesions (27%) was present in the modeling data. Both factors would be expected to cause a worse performance for classifying low-grade disease.

The binary region-wise classifier achieved an AUC of 0.805 as well as sensitivity and specificity of 0.747 and 0.838, respectively, corresponding to the maximum of the Youden index. Compared to the results of previous studies, these appear to be less impressive.[119,150,153,163,204] However, this can be explained by the fact that the candidate regions in previous studies were manually identified by expert clinicians, while the candidate regions in this work were automatically identified. Therefore, the two-stage model here is likely to be less biased and more generalizable when applied prospectively. After thresholding, the high-grade lesion detection performance of the model was compared to that of a radiologist, though the model detected fewer true-positive (25 vs. 30) and more false-positive lesions (11 vs. 7) overall.

There are several limitations of this work. First, the classifiers were developed using a relatively small amount of data from only 34 patients at a single institution. As a result, we were unable to quantify the performance of the classifiers in a prospective setting, and so the results reported here are likely to be optimistic.

Another limitation is the manner in which cancer aggressiveness was quantified from the Gleason score of identified lesions. Gleason score is quite limited as a surrogate measure of aggressiveness, as it is assessed at a single point in time that is fairly early in the disease course of PCa. Better surrogate measure of aggressiveness include time to biochemical recurrence or time to metastasis, but follow-up data were not uniformly available for the enrolled patients. Assignment of the Gleason score was also performed

by pathologists, which is known to be quite subjective and therefore subject to significant interobserver variability.[148,149] Using a cut-off of GS = 3+3 to distinguish low-grade from high-grade cancer is also not well-established, with justifications for and against the practice.[255]

Lastly, there is concern regarding the reproducibility of the features used to train the classifiers. There are numerous options and methods for performing the calculation and/or extraction of the qMRI and radiomic features, many of which could affect the numerical values of the features. The variability in the features could then translate into variability in the performance of the classifiers. This highlights the general lack of consensus regarding the best practices for performing feature extraction from prostate MRI data.[256]

In summary, the two-stage model introduced in this work can be integrated into a CAD system for the simultaneous detection and grading of PCa on mpMRI images. The unique aspects of this work is that 1) it combines both qMRI and radiomic features for voxel-wise cancer detection, and 2) performs region-wise cancer detection and grading using automatically-identified candidate regions. As a result, the proposed model is less biased and therefore more generalizable than previous works.

# Chapter 6

# Metric for evaluating lesion-wise performance of CAD models

## 6.1 Introduction

The development of computer-aided detection/diagnosis (CAD) systems is one of the most active areas of research in medical imaging. While most of the literature in the field has focused on the development of novel techniques and methods for constructing predictive models, relatively little attention has been paid to the evaluation of these models.[257] For example, for the task of binary classification (i.e., assessment of the presence or absence of disease), performance is almost always evaluated using receiver operating characteristic (ROC) curve analysis and/or confusion matrix statistics like sensitivity, specificity, and F1-score.[106,107,114,119,131–138,145,146,150–157] However, while metrics like AUC measure classification performance and may be useful for model development, they do not adequately reflect localization performance, i.e., the ability to accurately identify *lesions* (distinct foci of disease). This is because they only describe the frequency of the occurrence of correctly and incorrectly classified voxels, but not where they occur. In reality, the spatial distribution of predictions can greatly affect the perception of a predictive map. Consider the example shown in Figure 6.1, which compares two different binary predictive maps to the ground truth map. While the two


FIGURE 6.1: Comparison of two hypothetical predictive maps, PRED1 and PRED2, that have the same voxel-wise sensitivity and specificity (sensitivity = 0.70, specificity = 0.867) with respect to the ground truth. In the maps, white = cancer, black = non-cancer.

predictive maps have the same voxel-wise sensitivity and specificity, the spatial distribution of false positive voxels creates a perceived false positive lesion in PRED1 and not in PRED2. Consequently, the two predictive maps would be perceived very differently in a clinical scenario. This example illustrates the fact that voxel-wise evaluation of model performance is insufficient to provide a clinically-meaningful, quantitative comparison of different model outputs to each other and to the ground truth.

Calculating localization performance from voxel-wise predictions can be described as a two-step process. The first step is the creation of lesions from groups of voxels on both the ground truth and prediction maps, and the second is the definition of rules or metrics that describe how well predicted lesions localize with ground truth lesions. In the CAD literature, the latter step is sometimes referred to as *mark-labeling* (mark being synonymous with predicted lesion). A large number of mark-labeling methods have been described for non-prostate applications,[257] and include the following:

- Degree of overlap between ground truth and predicted lesions, which may be thresholded to produce a binary result.
- Containment of the centroid of one lesion within the other (and vice versa).
- Visual inspection by a trained radiologist.

Kallergi et al. showed that different mark-labeling methods in the context of comparing models for microcalcification detection on digital mammography can yield drastically different conclusions about their relative performance.[258] Therefore, it would be best to choose or to design an evaluation method that could be applied consistently across different models. Here, we propose novel lesion-wise evaluation metrics that emphasize both overlap and co-localization of predicted lesions with ground truth lesions. We then demonstrate the stable behavior of the metrics and characterize how they correlate with voxel-wise performance. Lastly, we demonstrate the use of the proposed metrics in 1) comparing predictive models of PCa on mpMRI to a radiologist's annotations of disease, and 2) quantifying the degree to which viewing model-predicted lesions improves the quality of the radiologist's annotations.

## 6.2 Methods

### 6.2.1 Data description

The mpMRI data used for this work were acquired as previously described,[106] In total, 46 axial slices of interest from 34 patients were included for the development of the methods. Please refer to Appendix A for more details.

### 6.2.2 Description of the proposed metrics

Refer to Table 6.1 for the notations used in describing the proposed evaluation measures. Additionally, for a given $\ell_{tr}$, an $\ell_p$ is defined to be *associated* with $\ell_{tr}$ if there exist voxels $v \in \ell_p$ and $w \in \ell_{tr}$ such that the Euclidean distance $\|v - w\| \leq z$, for some maximum allowable distance $z$. An associated $\ell_p$ is defined to be *overlapping* with $\ell_{tr}$ if any voxel is labeled cancer in both.

Discrete lesions need to be identified in both $m_{tr}$ and $m_p$ before a lesion-wise measure can be applied. The series of steps we used for automatic lesion identification is diagrammed in Figure 6.2. Compared to radiologists' annotations, $m_p$s produced by

TABLE 6.1: Notations used in the proposed methods.

| Notation | Definition |
|---|---|
| $m_{tr}$ | Binary ground truth map |
| $m_p$ | Binary prediction map |
| $\ell_{tr}$ | Lesion identified in $m_{tr}$ |
| $\ell_p$ | Lesion identified in $m_p$ |
| $d = d(\ell_{tr}, \ell_p)$ | Euclidean distance between the centroids of $\ell_{tr}$ and $\ell_p$ |
| $\text{TP}_\ell$, $\text{FN}_\ell$, $\text{FP}_\ell$ | Sets of true positive (TP), false negative (FN), and false positive (FP) voxels with respect to an $\ell_{tr}$ and its associated $\ell_p$s |
| $\lvert \cdot \rvert$ | Set size (i.e., number of voxels) |
| $s_\ell$ | Lesion-wise score |
| $s_\sigma$ | Lesion-summary score |

voxel-wise classification may be quite noisy, i.e., they may have small, isolated, and non-contiguous regions of cancer-labeled voxels among many negative voxels (or vice versa). Depending on how the ground truth was derived and processed, $m_{tr}$ may also have non-continuous regions of annotated cancer (Fig. 6.2a). To address these irregularities, binary dilation was first applied to both maps. Connected voxels were labeled as the same lesion (Fig. 6.2c), and then the masks of the original maps were applied (Fig. 6.2d). For $m_p$s, median filtering was also applied to reduce the spatial noise (Fig. 6.2b). A size threshold of 75 voxels was then applied (Fig. 6.2e) with the rationale that smaller lesions would likely represent benign, clinically-insignificant disease.[252]

Next, associations between $\ell_{tr}$s and $\ell_p$s were determined (Fig. 6.2e) using a maximum allowable distance of $z = 3$ mm. For each $\ell_{tr}$, all associated $\ell_p$s were found with the condition that each $\ell_p$ is associated with at most one $\ell_{tr}$. In the case where a lesion would otherwise be associated with multiple $\ell_{tr}$s (i.e., $n$ $\ell_{tr}$s, where $n > 1$), it is instead partitioned into $n$ $\ell_p$s where the $i$th $\ell_p$ $\left(\ell_p^i\right)$ is associated with the $i$th $\ell_{tr}$ $\left(\ell_{tr}^i\right)$. Specifically, a given voxel $v \in \ell_p^i$ if the following is true:

$$\underset{1 \leq k \leq n}{\arg\min} \left( \min_{w \in \ell_{tr}^k} \|v - w\| \right) = i \tag{6.1}$$

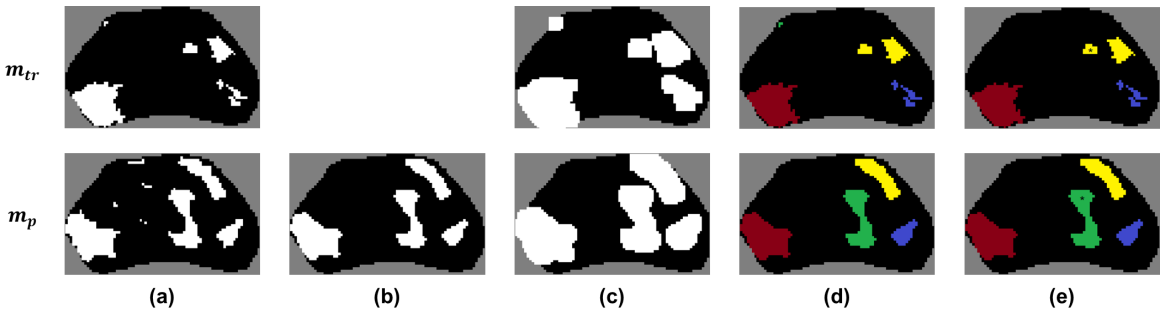



FIGURE 6.2: Demonstration of the workflow for identification of discrete lesions in $m_{tr}$ and $m_p$. **(a)** Original maps: black = non-cancer, white = cancer. **(b)** Median filtering (3×3 kernel) of $m_p$. **(c)** Dilation (5×5 square kernel) and identification of connected cancer voxels. **(d)** Application of the masks of original maps to the dilated regions. Distinct lesions identified by this procedure are labeled in separate colors. **(e)** Size-thresholding of sufficiently small lesions ($\leq$ 75 voxels) in $m_{tr}$ and $m_p$ produces the final $m_{tr}$ and $m_p$. In this case, the blue, cyan, and yellow lesions in $m_{tr}$ and $m_p$ are overlapping (and therefore also associated). While the brown lesion in $m_p$ does not overlap with any in $m_{tr}$, it is associated with the yellow lesion in $m_{tr}$ due to their proximity (minimum voxel-to-voxel distance between the two is $\leq$5 voxels).

where $\|v - w\|$ is the Euclidean distance between two voxels. An example of this situation is illustrated in Figure 6.3.

After this step of lesion identification, the proposed *lesion-wise score* $s_\ell$ was calculated for each $\ell_{tr}$. $s_\ell$ was designed to satisfy the following:

- $0 \leq s_\ell \leq 1$, with $s_\ell = 0$ when the $\ell_{tr}$ has no overlapping $\ell_p$s and $s_\ell = 1$ when $\ell_p = \ell_{tr}$.
- $s_\ell$ increases as degree of overlap and co-localization between the $\ell_{tr}$ and its associated $\ell_p$s improve, where co-localization is quantified by $d$.

$s_\ell$ is based on the Jaccard similarity coefficient $(J_c)$, which is a similarity measure between two samples. In the setting of binary classification, $J_c$ measures the degree of overlap between $\ell_{tr}$ and its associated $\ell_p$s, and is defined as:

$$J_c = \frac{|\mathrm{TP}_\ell|}{|\mathrm{TP}_\ell| + |\mathrm{FN}_\ell| + |\mathrm{FP}_\ell|} \tag{6.2}$$

$J_c$ is closely related to the Dice similarity coefficient (DSC). For comparison, the DSC is defined as:

$$\mathrm{DSC} = \frac{2\,|\mathrm{TP}_\ell|}{2\,|\mathrm{TP}_\ell| + |\mathrm{FN}_\ell| + |\mathrm{FP}_\ell|} \tag{6.3}$$

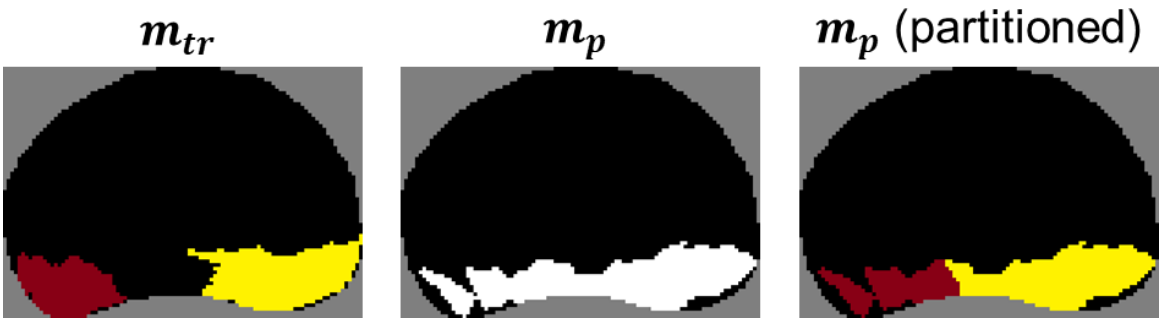



FIGURE 6.3: Demonstration of the handling of an edge case where a lesion in $m_p$ would overlap with multiple $\ell_{tr}$s (two $\ell_{tr}$s in this case). The lesion is partitioned into two $\ell_p$s such that the voxels of the $i$th $\ell_p$ are closer to those of the $i$th $\ell_{tr}$ than any other $\ell_{tr}$. The end result is that the $i$th $\ell_p$ is associated with the $i$th $\ell_{tr}$. For example, in the partitioned $m_p$, the green and blue $\ell_p$s are associated with the same colored $\ell_{tr}$s.

The proposed $s_\ell$ is defined as:

$$s_\ell = \max\left(1 + \log_{10}\left(s_o\right), 0\right) \tag{6.4}$$

where $s_o$ is the untransformed score given by:

$$s_o = f\left(\omega, \ell_{tr}, \ell_p\right) g(d) = \left[\frac{\sum_{v \in \mathrm{TP}_\ell} \omega(r(v))}{\sum_{v \in \mathrm{TP}_\ell} \omega(r(v)) + \sum_{v \in \mathrm{FN}_\ell} \omega(r(v)) + |\mathrm{FP}_\ell|}\right]\left[\frac{a_1}{(a_1 - 1) + a_2^d}\right] \tag{6.5}$$

In the definition for $s_o$, two modifications to $J_c$ were introduced to account for co-localization. The first is a weighting function $\omega$ that weights the voxels of $\ell_{tr}$ such that voxels closer to the centroid of the lesion contribute more heavily to $s_o$ than those at the periphery, which rewards co-localization of TPs. Thus, $\omega$ should be radially symmetric with respect to the centroid of $\ell_{tr}$, i.e., defined as a function of $r(v)$, the Euclidean distance of the voxel $v \in \ell_{tr}$ to the centroid of $\ell_{tr}$. Furthermore, $\omega(r(v))$ should be a non-negative, monotonically-decreasing function defined on the domain $[0, m]$, where $m = \max\left(r(v)\right)$. Here, we chose:

$$\omega(r(v)) = (m - r(v))^{a_\omega} \tag{6.6}$$

where $a_\omega > 0$ is a constant that controls how heavily co-localization influences $s_o$. The second modification is a function $g(d)$ that penalizes poor co-localization. Thus, $g(d)$ should be a non-negative, monotonically-decreasing function of $d$ with $g(0) = 1$ (i.e., no penalty for perfect co-localization of the centroids). Here, we chose:

$$g(d) = \frac{a_1}{(a_1 - 1) + a_2^d} \tag{6.7}$$

where $a_1 > 0$ and $a_2 > 1$ are constants. Lastly, a log-transform is applied to $s_o$ so that $s_\ell$ more fully spans the range of values from 0 to 1.

While $s_\ell$ is a continuous score, it may be thresholded to produce a mark-labeling rule, i.e., a given $\ell_p$ would be considered a TP lesion if the score is above the threshold, and a FN lesion otherwise. Alternatively, if a single summary statistic is desired, we propose that a *lesion-summary score* $s_\sigma$ be defined as a weighted average of a number of $s_\ell$s:

$$s_\sigma = \frac{\sum_i s_\ell^i \left|\ell_{tr}^i\right|}{\sum_i \left|\ell_{tr}^i\right| + \sum_j \left|\ell_{p,\mathrm{FP}}^j\right|} \tag{6.8}$$

where $|\ell_{tr}| = |\mathrm{TP}_\ell| + |\mathrm{FN}_\ell|$ and $\ell_{p,\mathrm{FP}}$ denotes an unassociated $\ell_p$ (i.e., FP lesion). The lesion-summary score $s_\sigma$ may be calculated for any number of ground truth lesions across any number of cases.

In the definition of $s_\ell$ in equation 6.4, there are three constants that need to be chosen: $a_1$, $a_2$, and $a_\omega$. While the choice of constants may be customized to the specific clinical application, here a quantitative optimization procedure was carried out to select the constants (see Appendix B). Briefly, $s_\ell$s were calculated over a subset of 25 $m_{tr}$s with only one $\ell_{tr}$ and their corresponding model-generated $m_p$s (see link) as the constants were varied. The goal was to find the set of constants that maximized the *discriminatory power* and *stability* of $s_\ell$, where discriminatory power was quantified by the standard deviation in $s_\ell$s (large standard deviation = high discriminatory power) and stability was quantified by the change in $s_\ell$ with respect to the constants (small changes = high stability). As a result, constants of $a_1 = 14$, $a_2 = 1.25$ and $a_\omega = 2.7$ were chosen and used

in further experiments.

### 6.2.3 Characterization of the proposed measures using synthetic prediction maps

To characterize the application of the proposed measures in a variety of scenarios, synthetic $m_p$s were generated using a custom algorithm (see Appendix C). Briefly, to generate a synthetic map for a given $m_{tr}$, $m_p$ is first initialized with randomly seeded TP and FP voxels. A series of morphological operations (including opening, closing, and filling) as well as median filtering are then iteratively applied until specified voxel-wise sensitivity and specificity are achieved, with maximum error of $\pm 1\%$. Randomness in the strength of these operations as well as in the order that they are applied produces a variety of $m_p$s for a given $m_{tr}$.

To generate an $m_p$ with a specified $s_\sigma$, $m_p$s were synthesized with randomly-generated voxel-wise sensitivity and specificity pairs until one with the desired $s_\sigma$ ($\pm 0.02$) was produced. Similarly, to generate $m_p$s with a specified number of $\ell_p$s (including unassociated $\ell_p$s), $m_p$s were synthesized until one with the desired characteristics was produced.

To characterize $s_\ell$, synthetic $m_p$s with only one $\ell_p$ were generated for each of the 25 $m_{tr}$s with only one $\ell_{tr}$. First, to demonstrate that $s_\ell$ correlates with prediction quality, $m_p$s achieving a range of $s_\ell$s (from 0.2 to 0.8) were generated for each $m_{tr}$. Next, to quantify how $s_\ell$ varies with prediction quality, a series of $m_p$s with a single circular $\ell_p$ was generated for each of the aforementioned 25 $m_{tr}$s. Initially, the centroid of the $\ell_p$ was co-localized with the centroid of its $\ell_{tr}$, and the radius of $\ell_p$ was chosen to maximize $s_\ell$. $s_\ell$ and DSC were then calculated 1) as the distance between the centroids of $\ell_p$ and $\ell_{tr}$ was increased, and 2) as the radius of $\ell_p$ was varied.

To characterize the proposed measures vs. voxel-wise sensitivity and specificity, 50 synthetic $m_p$s achieving a specified voxel-wise sensitivity and specificity pair were generated for each of the 46 $m_{tr}$s. For each voxel-wise sensitivity and specificity pair, the

average $s_\sigma$ as well as the average numbers of TP, FP, and FN lesions over all $m_{tr}$s were calculated. $s_\ell = 0.5$ was used as the threshold for lesion detection, i.e., for a given $\ell_{tr}$, an $\ell_p$ was defined to be a TP lesion if $s_\ell \geq 0.5$ and a FN lesion otherwise. Unassociated $\ell_p$s were defined to be FP lesions. These definitions for TP/FN/FP lesions were used in further experiments.

### 6.2.4 Demonstration of the clinical utility of the proposed measures

The proposed measures were used to 1) compare $m_p$s from a predictive model to a radiologist's annotations, and 2) assess the degree to which the model aids radiologic diagnosis of PCa. The model used for this work was described previously.[259] Briefly, the model is a voxel-wise classifier that is an extension of the L1-regularized logistic regression model used in Metzger et al.[106] with added predictive features that are related to the spatial location of voxels. A case-based leave-one-out cross-validation scheme was used to train and assess performance. ROC curve analysis and the DSC were used to assess the voxel-wise performance, while the proposed measures were used to assess the lesion-wise performance. For calculation of $s_\sigma$ and DSC, a sensitivity and specificity pair was chosen from the ROC curve of the model (58.8% sensitivity, 88.4% specificity) corresponding to the maximum of the Youden index.

For the 34 cases, the mpMRIs were interpreted in accordance with PI-RADS v2 guidelines[2] by a board-certified, fellowship-trained body imaging sub-specialized radiologist (B.S.) with 5 years of experience in prostate MRI. Initially, only the imaging data, consisting of the T2-weighted (T2W) images, the diffusion-weighted images (including a calculated high b-value image), the ADC map, the DCE time series, and the T1-weighted images, were available for review. Specifically, the radiologist was blinded to the clinical data, the $m_{tr}$s, and the model-generated $m_p$s. 0-3 ROIs were drawn by the radiologist using the DynaCAD program (Invivo) on each of the 46 slices to subjectively outline the maximum extent of disease. Only ROIs receiving a PI-RADS score of $\geq 3$ (representing at least an intermediate risk of malignancy) were annotated.

Next, the radiologist was selectively un-blinded to the $m_p$s from the model; using this new information in conjunction with the mpMRI data, the radiologist subjectively modified the original ROIs. These modifications included changing the contour of existing ROIs, drawing new ROIs, and/or removing existing ROIs. Specifically, a new ROI was drawn when the $m_p$ either aided the detection of a previously unrecognized lesion or changed the subjective probability of cancer from low (the equivalent of PI-RADS = 2) to intermediate (the equivalent of PI-RADS = 3). Similarly, an existing ROI was removed when the $m_p$ changed the subjective probability of cancer from intermediate to low. The radiologist also had the option of leaving the original annotation intact if it was felt that the $m_p$ was unhelpful (e.g., if identified $\ell_p$s were concordant with existing ROIs) or inaccurate (e.g., if identified $\ell_p$s were deemed unlikely to represent malignancy).

The annotations were then exported into MATLAB (MathWorks) and processed in the same way that $m_p$s were processed. The overall quality of the original and the adjusted annotations were compared with that of the model in terms of $s_\sigma$ and DSC as well as lesion detection accuracy. The same analysis was also carried out on the select subset of slices in which the annotations were modified.

## 6.3 Results

Three representative $m_p$s achieving a range of $s_\ell$s for three different $m_{tr}$s are shown in Figure 6.4. From these results, it appears that increases in $s_\ell$ reflect the progressive improvement in the quality of the $m_p$s.

Plots characterizing how $s_\ell$ and DSC change with degrees of co-localization and overlap between $\ell_p$ and $\ell_{tr}$ are shown in Figure 6.5. It appears that $s_\ell$ decreases in a mostly linear fashion with both decreasing co-localization (Fig. 6.5a) and overlap (Fig. 6.5b) over a wide range of values, approximately between $s_\ell = 0.1$ and $s_\ell = 0.8$. The results also show that the range of values for $s_\ell$ is noticeably larger than that for DSC over

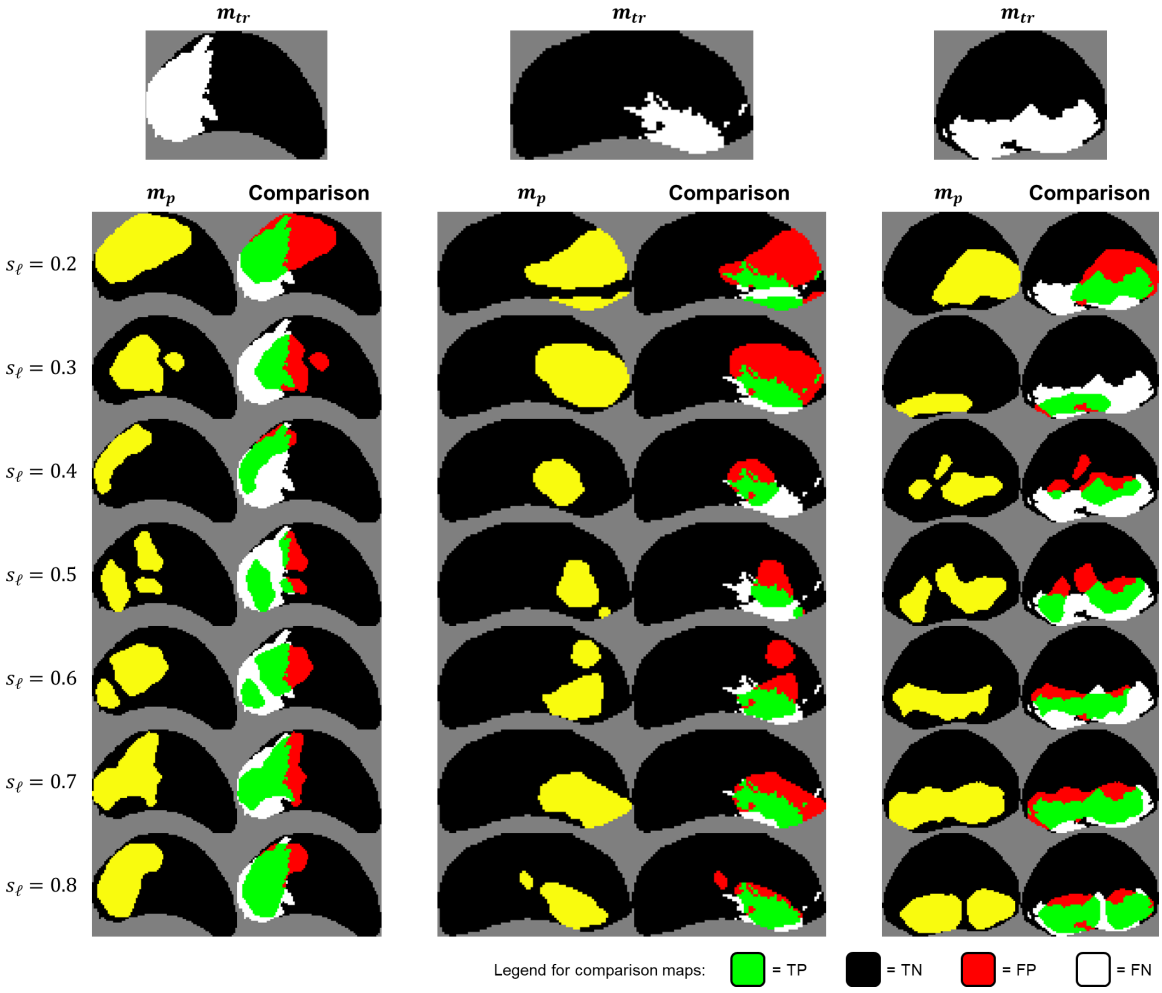



FIGURE 6.4: Representative $m_p$s for the same $m_{tr}$s that achieve scores of $s_\ell = 0.2$ to $s_\ell = 0.8$. In the original maps, $\ell_{tr}$s are in white, $\ell_p$s are in yellow, and black = non-cancer. In the comparison maps, green = TP, black = TN, red = FP, white = FN.

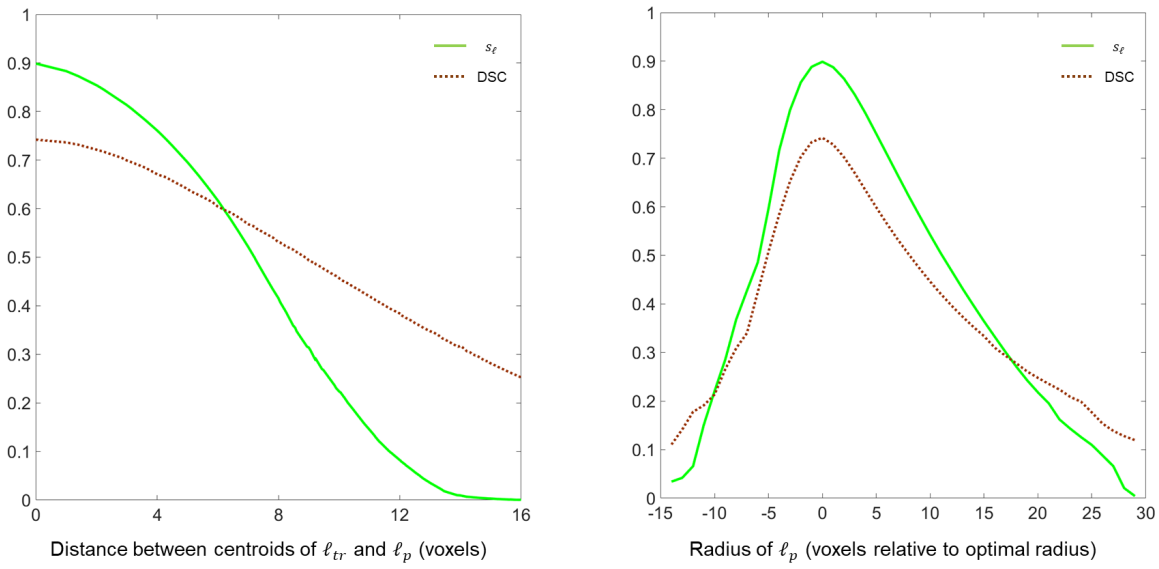



FIGURE 6.5: **(a)** Plot of $s_\ell$ and DSC vs. average distance between the centroids of a circular $\ell_p$ and $\ell_{tr}$ (calculated across 25 slices with a single $\ell_{tr}$). Radius of $\ell_p$ in all cases was chosen to maximize the corresponding measure when the centroids are co-localized. **(b)** Plot of $s_\ell$ and DSC vs. radius of $\ell_p$ averaged across the 25 slices. In all cases, the centroids of $\ell_p$ and $\ell_{tr}$ are co-localized.

the same range of prediction qualities, which suggests that $s_\ell$ is superior to DSC for comparison of prediction maps.

Comparisons of $s_\sigma$ vs. voxel-wise sensitivity and specificity are shown in Table 6.2. Increases in $s_\sigma$ are approximately linear with respect to sensitivity and quadratic with specificity, meaning that improvements in specificity are more effective than improvements in sensitivity at increasing $s_\sigma$ (Table 6.2a). Table 6.2b catalogues the lesion detection statistics for select voxel-wise sensitivity and specificity pairs taken from Table 6.2a. Again, it is apparent that improvements in specificity are more effective for increasing lesion detection accuracy.

The top half of Table 6.3 shows the cumulative AUC, DSC, and $s_\sigma$ for the model and radiologist annotations across all 46 slices, while the bottom half summarizes the lesion detection performance.

The Pearson correlation coefficient between AUC and $s_\sigma$ for the model was $\rho = 0.68$ (95% CI [0.49, 0.81]), while the correlation between DSC and $s_\sigma$ was $\rho = 0.88$

TABLE 6.2: Characterization of the lesion-summary score. 50 synthetic $m_p$s were generated for each slice. Results shown are averaged across the 50 $m_p$s and across all 46 slices. **(a)** $s_\sigma$ vs. voxel-wise sensitivity and specificity. **(b)** Lesion detection statistics for representative voxel-wise sensitivity and specificity pairs shown in (a). Accuracy of lesion detection was defined to be TP/(TP + FN + FP).

| | | sensitivity | | | | |
|---|---|---|---|---|---|---|
| **$s_\sigma$** | | **0.50** | **0.60** | **0.70** | **0.80** | **0.90** |
| | **0.50** | 0.11 | 0.14 | 0.17 | 0.20 | 0.21 |
| | **0.60** | 0.14 | 0.18 | 0.23 | 0.25 | 0.29 |
| **specificity** | **0.70** | 0.18 | 0.25 | 0.29 | 0.35 | 0.41 |
| | **0.80** | 0.30 | 0.36 | 0.45 | 0.51 | 0.56 |
| | **0.90** | 0.50 | 0.58 | 0.66 | 0.72 | 0.78 |

**(a)**

| specificity | sensitivity | $s_\sigma$ | Average TP lesions | Average FN lesions | Average FP lesions | Detection accuracy |
|---|---|---|---|---|---|---|
| **0.50** | **0.50** | 0.11 | 3.1 | 48.9 | 9.3 | 0.05 |
| **0.50** | **0.70** | 0.17 | 7.9 | 44.1 | 11.6 | 0.12 |
| **0.50** | **0.90** | 0.21 | 10.5 | 41.5 | 13.9 | 0.16 |
| **0.70** | **0.50** | 0.18 | 8.9 | 43.1 | 12.8 | 0.14 |
| **0.70** | **0.70** | 0.29 | 14.1 | 37.9 | 10.6 | 0.23 |
| **0.70** | **0.90** | 0.41 | 18.1 | 33.9 | 11.2 | 0.29 |
| **0.90** | **0.50** | 0.50 | 29.0 | 23.0 | 10.7 | 0.46 |
| **0.90** | **0.70** | 0.66 | 42.7 | 9.3 | 7.9 | 0.71 |
| **0.90** | **0.90** | 0.78 | 47.5 | 4.5 | 3.4 | 0.86 |

**(b)**

TABLE 6.3: Comparison of measures of PCa detection for the model and radiologist annotations before (ANN) and after (ANN + Model) seeing the $m_p$s. For calculation of DSC and $s_\sigma$, a sensitivity and specificity pair was chosen from the ROC curve of the model corresponding to the maximum of the Youden index.

| | Model | ANN | ANN + Model |
|---|---|---|---|
| **AUC** | 0.81 | — | — |
| **DSC** | 0.56 | 0.56 | 0.59$^\dagger$ |
| **$s_\sigma$** | 0.58 | 0.55 | 0.58$^\dagger$ |
| **TP lesions** | 30 | 33 | 37 |
| **FN lesions** | 22 | 19 | 15 |
| **FP lesions** | 12 | 7 | 6 |

$^\dagger$ significant at $p < 0.05$ compared to ANN.

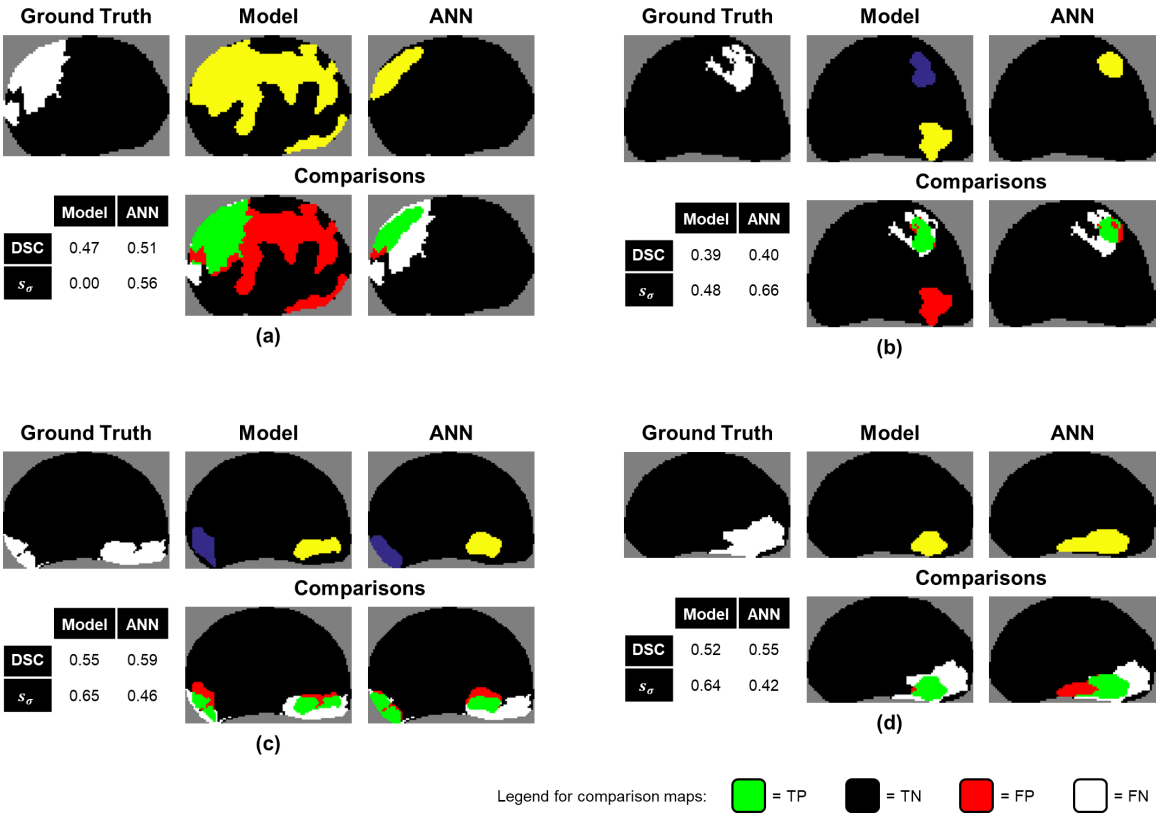



FIGURE 6.6: Comparisons of the model-produced predictive maps vs. the original radiological annotations (ANN) on four select slices in which similar DSCs but noticeably different $s_\sigma$s were achieved. In the original maps, $\ell_{tr}$s are in white, distinct $\ell_p$s are in separate colors, and black = non-cancer. In the comparison maps, green = TP, black = TN, red = FP, white = FN.

(95% CI [0.83, 0.92]). Despite generally good agreement between DSC and $s_\sigma$, there were several slices in which the model prediction and radiologist annotation achieved similar DSCs but very different $s_\sigma$s. The comparisons are illustrated in Figure 6.6 for four such representative slices. Paired two-tailed t-tests were performed to compare $s_\sigma$ and DSC for 1) model vs. the original annotations, and 2) original vs. adjusted annotations. While there was no significant difference in $s_\sigma$ ($p$ = 0.93) and DSC ($p$ = 0.36) between the model and the original annotations, the small improvement in radiologist performance after viewing the $m_p$s was statistically significant as quantified by both $s_\sigma$ ($p$ = 0.01) and DSC ($p$ = 0.04).

These improvements are magnified when considering only the 15 slices in which

TABLE 6.4: Comparison of the cumulative DSC and $s_\sigma$ of PCa detection for the model and radiologist annotations (ANN) on the 15 slices in which the annotations were modified after seeing the model predictions (ANN + Model).

| | **Model** | **ANN** | **ANN + Model** |
|---|---|---|---|
| **DSC** | 0.52 | 0.54 | 0.62$^\dagger$ |
| **s**$_\sigma$ | 0.54 | 0.54 | 0.70$^\dagger$ |
| **TP lesions** | 10 | 9 | 13 |
| **FN lesions** | 6 | 7 | 3 |
| **FP lesions** | 2 | 1 | 0 |

$^\dagger$ significant at $p$ < 0.05 compared to ANN.

the annotations were actually modified (Table 6.4). Again, while there was no significant difference in $s_\sigma$ ($p = 0.79$) and DSC ($p = 0.58$) between the model and the original annotations, the improvement in radiologist performance was statistically significant as quantified by both $s_\sigma$ ($p$ = 0.0086) and DSC ($p$ = 0.02).

Of the 15 slices, nine involved adjusting the extent of ROI(s), while six involved the addition or deletion of ROI(s). Figure 6.7 illustrates the types of modifications that the radiologist made to the original annotations as well as the usage of the proposed $s_\sigma$ to quantify the effect of the modifications for six representative slices.

## 6.4 Discussion

A majority of the surveyed literature on predictive models of PCa on mpMRI reported only the voxel-wise classification performance. Of the seven studies that did report localization performance:

- Four studies reported the mean DSC[38,108,133,153] as a measure of the overlap between $\ell_{tr}$s and $\ell_p$s.
- Litjens et al. used the rule that an $\ell_p$ is considered to be a TP lesion if its centroid is within 10 mm of the centroid of an $\ell_{tr}$, and a FP lesion otherwise.[119]

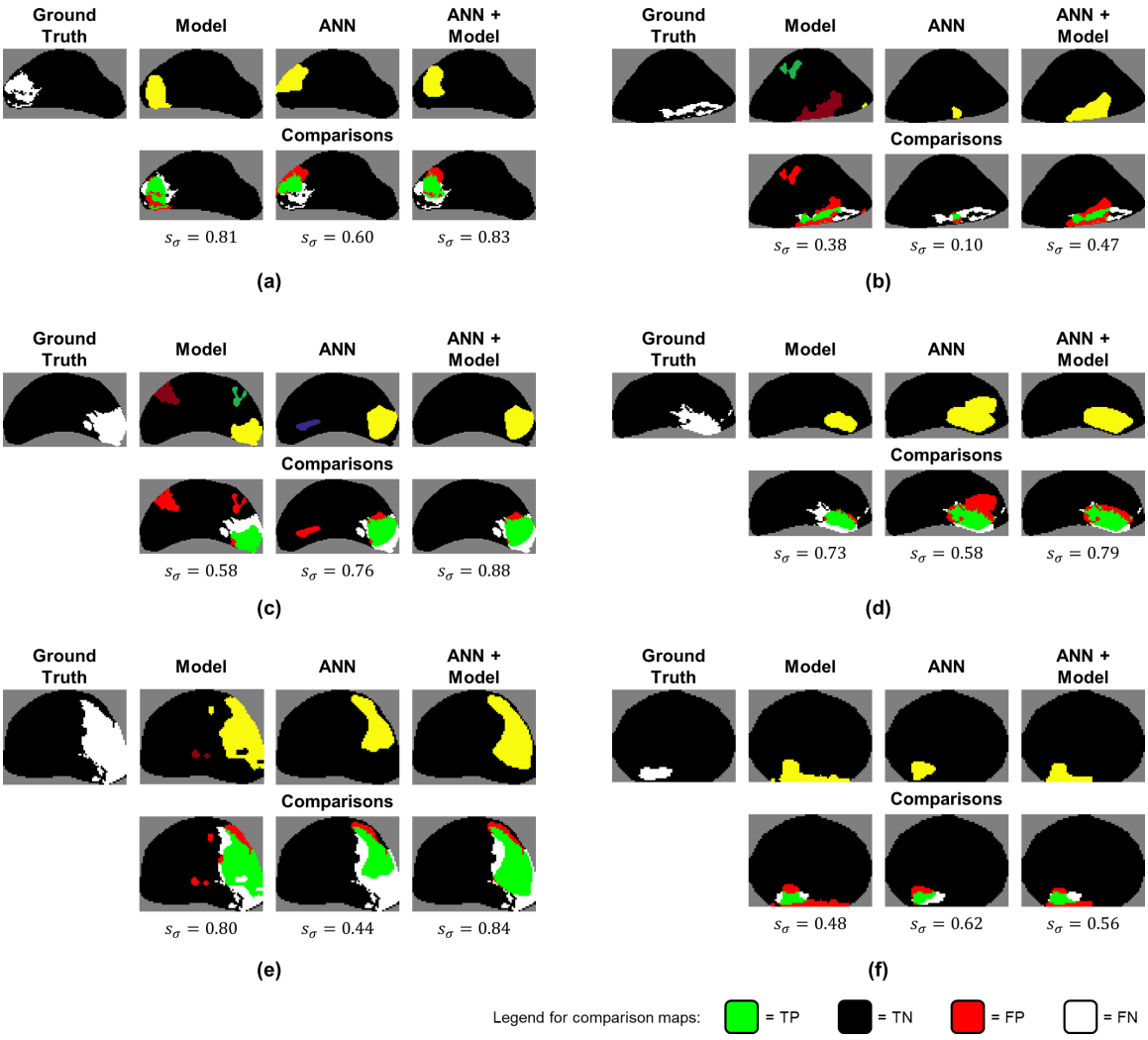



FIGURE 6.7: Illustration of the usage of the proposed measures to quantitatively evaluate the clinical utility of a predictive model in aiding radiological diagnosis of PCa. In the original maps, $\ell_{tr}$s are in white, distinct $\ell_p$s are in separate colors, and black = non-cancer. In the comparison maps, green = TP, black = TN, red = FP, white = FN. The $s_\sigma$ for the modified annotations (ANN + Model) in these cases clearly reflect the adjustment of disease extent and/or removal of FP lesions, with respect to either the model output or the original annotations (ANN).

- Vos et al. used the rule that an $\ell_p$ is considered to be a TP lesion if it is wholly contained within an $\ell_{tr}$, and a FP lesion otherwise. Notably, the size of the $\ell_p$ was apparently not considered.[107,109]

These evaluation rules may not always accurately reflect the quality of the predictions, especially for targeted applications. The DSC measures overlap but does not explicitly reflect the goodness of co-localization, while the opposite is true for the rule proposed by Litjens et al. To address these deficiencies, novel lesion-wise evaluation measures were developed to more accurately evaluate the quality of predictive models of PCa on mpMRI.

The lesion-wise score $s_\ell$ is derived from the Jaccard similarity coefficient with modifications that emphasize overlap and co-localization of ground truth and predicted lesions. The lesion-summary score $s_\sigma$ is derived from $s_\ell$ and evaluates overall model performance. Our experiments demonstrate that $s_\ell$ accurately reflects the quality of the $\ell_p$ over a wide range of values, and that the behavior of $s_\ell$ is stable and predictable with respect to small changes in the quality of the $\ell_p$. $s_\ell$ varies in an approximately linear fashion with both overlap and co-localization between $\ell_p$s and $\ell_{tr}$s, and does so over a wide range of values (0.1 to 0.8) where the score of most $\ell_p$s will likely reside in practice. Similarly, our experiments demonstrate that on average, $s_\sigma$ correlates predictably with voxel-wise measures of sensitivity, specificity, and DSC. The proposed measures address the problem that it is difficult, if not impossible, to correlate the voxel-wise performance of a predictive model with clinically-relevant outcomes. Lesion-wise sensitivity and specificity can be determined via thresholding $s_\ell$, while $s_\sigma$ provides a single summary statistic of lesion-detection performance. Returning to the example illustrated in the introduction (Fig. 6.8), it is apparent that the proposed methods provide a much more complete and accurate comparison of the two predictive maps than voxel-wise sensitivity and specificity do.

The proposed measures were used to compare $m_p$s generated by a predictive model to radiological annotations of PCa and to quantify the effect that a predictive model

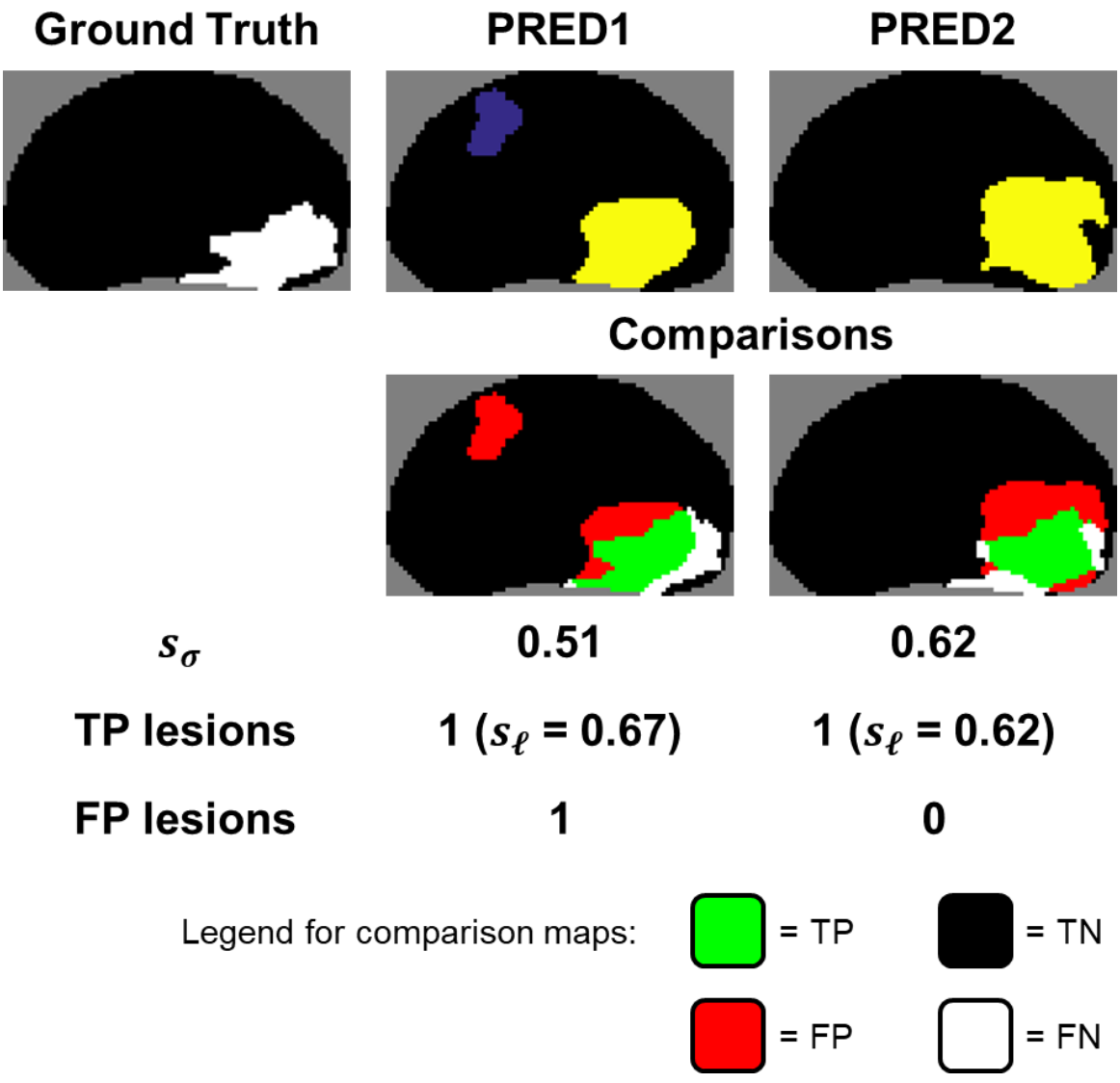



FIGURE 6.8: Analysis of the $m_p$s (with the same voxel-wise sensitivity and specificity) shown in Fig. 6.1 using the proposed methods. In the original maps, $\ell_{tr}$s are in white, distinct $\ell_p$s are in separate colors, and black = non-cancer. In the comparison maps, green = TP, black = TN, red = FP, white = FN.

has on annoation quality. In the cases where annotations were modified after seeing the $m_p$s, it appears that changes in $s_\sigma$ accurately reflect the changes in the extent of drawn ROIs and/or improvements made via addition of TP lesions or removal of FP lesions. Interestingly, the results show that an $m_p$ can improve annotation quality (as quantified by $s_\sigma$) even when the $m_p$ is worse than the original annotation (Figure 6.7c). This supports the idea that validation of predictive performance is best done in the context of the clinical workflow, and our experiments demonstrate that the proposed measures are well-suited for this kind of evaluation.

In several of the results shown, the DSC was used to as a reference for comparison to $s_\sigma$. Not only were the two measures well-correlated, but they also achieved similar values. This is unsurprising as the proposed measures are based off of the Jaccard similarity index, which in turn is very similar to the DSC. The main differences between the two are that:

- $s_\sigma$ is a weighted average of $s_\ell$s that assesses lesion-wise performance, while the DSC is calculated on an entire slice and therefore cannot be used to assess lesion-detection performance. This limitation of the DSC is demonstrated in the comparisons shown in Figures 6.6a and 6.6b.

- $s_\ell$ weights predicted voxels differently depending on their distance to the centroid of the $\ell_{tr}$s, while the DSC has no such weighting. This difference is demonstrated in the comparisons shown in Figures 6.6c and 6.6d.

Additionally, looking at the results of Table 6.4, we see that for the slices in which radiologist's annotations were modified, $s_\sigma$ = DSC for the original annotations, but DSC $> s_\sigma$ for the modified annotations. This means that the TP voxels and/or the FP voxels in the modified annotations were closer to the centroid of the $\ell_{tr}$s, which in turn suggests that the model aided the radiological assessment of the localization and/or extent of disease.

It appears that using the proposed measures to select predictive models for purposes of targeted intervention is superior to using commonly-reported voxel-wise

measures. However, there are several provisional aspects to the methods we have described, and these, detailed below, constitute potential limitations of the application of the proposed measures.

### 6.4.1 Identification of lesions

Before $s_\ell$ can be calculated, discrete lesions need to be identified on the $m_{tr}$s and $m_p$s. Due to the fact that $m_{tr}$s and $m_p$s may be quite irregular, a series of morphological operations was applied to produce maps in which lesions could more readily be identified. It is evident that this can affect the calculation of proposed measures; in particular, the strength of the binary dilation operation affects the numbers of $\ell_p$s and $\ell_{tr}$s that are initially identified, which in turn affects the associations between them and the calculated $s_\ell$s. This was partially the motivation for proposing the lesion-summary score, as $s_\sigma$ is a weighted sum of $s_\ell$s that also takes into account FP lesions and therefore is not greatly affected by how lesions are determined. It is possible that these steps would be made more superfluous by principled modeling approaches that explicitly enforce spatial smoothness. However, some degree of spatial noise is likely to be present with any voxel-wise modeling approach. Therefore, an automated approach for lesion identification similar to the one we described would likely still be necessary.

### 6.4.2 Definition of the proposed measures

The proposed lesion-wise measure is defined as a log-transform of the product of two functions ($s_\ell = \max\left(1 + \log_{10}\left(s_o\right), 0\right)$, where $s_o = f\left(\omega, \ell_{tr}, \ell_p\right) g(d)$). The function $f$ quantifies the overlap and co-localization of TP voxels between $\ell_{tr}$ and its $\ell_p$s, and includes a weighting function $\omega$ that affects how heavily co-localization is weighted in the calculation of $s_\ell$. $g$ also quantifies co-localization by considering a weighted value of the distance $d$ between the centroids of $\ell_{tr}$ and $\ell_p$, and is primarily included to account for the effect of FP voxels. The specific choices for $\omega$ and g required the selection of three constants ($a_1$, $a_2$ and $a_\omega$). As described in the methods and in Appendix A, a quantitative

optimization procedure was carried out using the $m_{tr}$s and model-generated $m_p$s to choose the constants that maximized both the discriminatory power of $s_\ell$ (as quantified by the standard deviation of $s_\ell$s) and the stability of $s_\ell$ (as quantified by the change in $s_\ell$ with respect to its defining constants).

While this optimization procedure is still somewhat arbitrary, we believe that the degrees of freedom in the definition of $s_\ell$ also offer the opportunity to customize the measure for the specific targeted clinical application. For example, for performing targeted prostate biopsy, accurate co-localization of $\ell_{tr}$s and $\ell_p$s would likely be more important than accurate assessment of the extent of disease (at least up to a point). In this case, $a_\omega$ could be increased to more strongly emphasize the accurate classification of TP voxels near the centroid of $\ell_{tr}$s, while $a_1$ could be increased and/or $a_2$ could be decreased to de-emphasize the effect of $g(d)$. On the other hand, for focal therapies such as cryotherapy, adequate coverage of the diseased regions is more important than precise localization of all disease. In this case, adjustments of the constants in the opposite direction of what was suggested above might be appropriate. Either way, the adjusted definition for $s_\ell$ could then be used to select the best predictive models for the specific application.

Furthermore, if it is felt that selecting the appropriate constants is too arbitrary or cumbersome, the form of $s_\ell$ may be altered to decrease the number of constants in its definition. For example, a weighting function similar to $\omega$ could be applied to the FP voxels such that FPs further away from the centroid of $\ell_{tr}$ are penalized more heavily, while the function $g$ could be removed altogether. In this case, if the weighting function for FP voxels were to have the same form as the current $\omega$ and share the same $a_\omega$, there would only be one constant to adjust in the definition of $s_\ell$.

### 6.4.3 Characterization of radiological annotation

There are a few potential issues with the way that the radiological annotations were obtained and processed. First, only one observer was involved in the study, which

introduces heavy bias into the original annotations as well as the types of modifications made based on the $m_p$s. Also, cases are typically read and annotated on the entire 3D volume of the prostate, and adjustment of ROIs on individual slices is not part of the typical clinical workflow. Had $m_p$s for entire 3D volumes been available, the model outputs likely would have been perceived differently in the way they were used to adjust the annotations. This is unfortunately a limitation of the modeling data.

However, the goal of these experiments was to demonstrate that the proposed measures could successfully be used in the clinical assessment of the performance of predictive models, and not necessarily to attempt to evaluate the described predictive models. The application of the proposed measures was not affected by any of the aforementioned issues, and therefore we believe they are not as severe as they otherwise would be. In particular, although the proposed methods were not developed using 3D volumes, they could readily be applied to 3D volumes; the image processing methods to identify lesions from voxels naturally extend to three dimensions, as do the concepts of associated and overlapping lesions used in the definition of $s_\ell$.

While the usage of the described measures was demonstrated specifically in the setting of detection of prostate cancer on mpMRI, the methods described here could conceivably be adapted to similar CAD applications for different diseases and different modalities. In the immediate future, rigorous comparisons of model vs. radiologist performance using the proposed measures may help further refine the definitions of the measures and yield additional insights on their clinical utility.

## Chapter 7

# Future directions

Given the numerous steps in the CAD pipeline (Fig. 1.6), standardization of data acquisition and data processing methods will also be critical. This will entail acquiring the data using the same experimental setup and protocols, analyzing and processing the data using the same methods, and minimizing user input as much as possible. The image registration and digital pathology works described in Chapters 3 and 4, respectively, specifically aimed to address a couple of tasks of standardization of data processing.

While the predictive modeling work described in Chapter 5 aimed to evaluate the utility of two different yet complementary imaging features, it is still unclear whether image-based radiomic features, qMRI features, or a combination of both should be pursued in future prostate CAD works. In any case, the standardization of the feature extraction pipeline and the rigorous quantification of the predictive power of imaging features will be necessary moving forward. One notable initiative is the Quantitative Imaging Biomarkers Alliance (QIBA),[260] which seeks to develop qMRI standards, protocols, and imaging phantoms that facilitate the measurement of the stability and reproducibility of qMRI features across different subjects, scanner conditions (e.g., manufacturers, field strengths), and data processing pipelines. The efforts of the initiative may be extended to radiomic features as well, as the repeatability and reproducibility of radiomic features is largely uncertain.[261]

Future development of CAD systems for prostate cancer detection and diagnosis

will depend on the continued availability of large quantities of high-quality modeling data. Deep learning will also likely play an increasingly prominent role in future developments of all aspects of the CAD pipeline. However, as deep learning algorithms have many more degrees of freedom than conventional modeling approaches, the success of deep learning approaches depends even more heavily on the availability of modeling data. This will in turn likely require modeling data to be pooled from multiple institutions, and ideally made publicly-available to researchers. One notable initiative in this area is The Cancer Imaging Archive (TCIA),[262] which has repositories of radiological, histopathological, and genomic datasets for multiple types of cancers.

# Bibliography


1. Kumar V, Abbas A, and Aster J. Robbins & Cotran Pathologic Basis of Disease. Elsevier Health Sciences, 2014. URL: https://books.google.com/books?id=5NbsAwAAQBAJ.
2. Weinreb JC, Barentsz JO, Choyke PL, et al. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. Eur Urol 2015.
3. Siegel RL, Miller KD, and Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69:7–34.
4. Bell KJ, Del Mar C, Wright G, Dickinson J, and Glasziou P. Prevalence of incidental prostate cancer: A systematic review of autopsy studies. Int J Cancer 2015;137:1749–57.
5. Cheng L, Poulos CK, Pan CX, et al. Preoperative prediction of small volume cancer (less than 0.5 ml) in radical prostatectomy specimens. The Journal of urology 2005;174:898–902.
6. Arora R, Koch MO, Eble JN, Ulbright TM, Li L, and Cheng L. Heterogeneity of Gleason grade in multifocal adenocarcinoma of the prostate. Cancer: Interdisciplinary International Journal of the American Cancer Society 2004;100:2362–6.
7. Greene D, Wheeler T, Egawa S, Weaver R, and Scardino P. Relationship between clinical stage and histological zone of origin in early prostate cancer: morphometric analysis. British journal of urology 1991;68:499–509.

8. McNeal JE, Redwine EA, Freiha FS, and Stamey TA. Zonal distribution of prostatic adenocarcinoma. Correlation with histologic pattern and direction of spread. Am J Surg Pathol 1988;12:897–906.
9. Burdick MJ, Reddy CA, Ulchaker J, et al. Comparison of biochemical relapse-free survival between primary Gleason score 3 and primary Gleason score 4 for biopsy Gleason score 7 prostate cancer. International Journal of Radiation Oncology* Biology* Physics 2009;73:1439–45.
10. Chan TY, Partin AW, Walsh PC, and Epstein JI. Prognostic significance of Gleason score 3+ 4 versus Gleason score 4+ 3 tumor at radical prostatectomy. Urology 2000;56:823–7.
11. Kang DE, Fitzsimons NJ, Presti Jr JC, et al. Risk stratification of men with Gleason score 7 to 10 tumors by primary and secondary Gleason score: results from the SEARCH database. Urology 2007;70:277–82.
12. Epstein JI, Zelefsky MJ, Sjoberg DD, et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. Eur Urol 2016;69:428–35.
13. Pierorazio PM, Walsh PC, Partin AW, and Epstein JI. Prognostic G leason grade grouping: data based on the modified G leason scoring system. BJU international 2013;111:753–60.
14. Ilic D, Neuberger MM, Djulbegovic M, and Dahm P. Screening for prostate cancer. Cochrane Database Syst Rev 2013:Cd004720.
15. Martin RM, Donovan JL, Turner EL, et al. Effect of a Low-Intensity PSA-Based Screening Intervention on Prostate Cancer Mortality: The CAP Randomized Clinical Trial. Jama 2018;319:883–95.
16. Andriole GL, Crawford ED, Grubb R. L. r, et al. Mortality results from a randomized prostate-cancer screening trial. N Engl J Med 2009;360:1310–9.

17. Pinsky PF, Prorok PC, Yu K, et al. Extended mortality results for prostate cancer screening in the PLCO trial with median follow-up of 15 years. Cancer 2017;123:592–9.

18. Lee DK, Park JH, Kim JH, et al. Progression of prostate cancer despite an extremely low serum level of prostate-specific antigen. Korean J Urol 2010;51:358–61.

19. Pinsky PF, Parnes HL, and Andriole G. Mortality and complications after prostate biopsy in the Prostate, Lung, Colorectal and Ovarian Cancer Screening (PLCO) trial. BJU Int 2014;113:254–9.

20. Chou R, Croswell JM, Dana T, et al. Screening for prostate cancer: a review of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 2011;155:762–71.

21. King CR, McNeal JE, Gill H, and Presti J. C. J. Extended prostate biopsy scheme improves reliability of Gleason grading: implications for radiotherapy patients. Int J Radiat Oncol Biol Phys 2004;59:386–91.

22. Moussa AS, Kattan MW, Berglund R, Yu C, Fareed K, and Jones JS. A nomogram for predicting upgrading in patients with low- and intermediate-grade prostate cancer in the era of extended prostate sampling. BJU Int 2010;105:352–8.

23. Sundi D, Ross AE, Humphreys EB, et al. African American men with very low-risk prostate cancer exhibit adverse oncologic outcomes after radical prostatectomy: should active surveillance still be an option for them? J Clin Oncol 2013;31:2991–7.

24. Welch HG and Albertsen PC. Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986-2005. J Natl Cancer Inst 2009;101:1325–9.

25. Samaratunga H, Montironi R, True L, et al. International Society of Urological Pathology (ISUP) Consensus Conference on Handling and Staging of Radical

Prostatectomy Specimens. Working group 1: specimen handling. Modern Pathology 2011;24:6–15.

26. Bomers JG, Yakar D, Overduin CG, et al. MR imaging-guided focal cryoablation in patients with recurrent prostate cancer. Radiology 2013;268:451–60.
27. Nicolae AM, Venugopal N, and Ravi A. Trends in targeted prostate brachytherapy: from multiparametric MRI to nanomolecular radiosensitizers. Cancer Nanotechnol 2016;7:6.
28. Eisenberg ML and Shinohara K. Partial salvage cryoablation of the prostate for recurrent prostate cancer after radiotherapy failure. Urology 2008;72:1315–8.
29. Gangi A, Tsoumakidou G, Abdelli O, et al. Percutaneous MR-guided cryoablation of prostate cancer: initial experience. European radiology 2012;22:1829–35.
30. Donovan JL, Hamdy FC, Lane JA, et al. Patient-Reported Outcomes after Monitoring, Surgery, or Radiotherapy for Prostate Cancer. New England Journal of Medicine 2016;375:1425–37.
31. Hamdy FC, Donovan JL, Lane JA, et al. 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. New England Journal of Medicine 2016;375:1415–24.
32. Coakley FV, Teh HS, Qayyum A, et al. Endorectal MR imaging and MR spectroscopic imaging for locally recurrent prostate cancer after external beam radiation therapy: preliminary experience. Radiology 2004;233:441–8.
33. D'Amico AV, Whittington R, Malkowicz SB, et al. Combination of the preoperative PSA level, biopsy gleason score, percentage of positive biopsies, and MRI T-stage to predict early PSA failure in men with clinically localized prostate cancer. Urology 2000;55:572–7.
34. Futterer JJ, Heijmink SW, Scheenen TW, et al. Prostate cancer: local staging at 3-T endorectal MR imaging–early experience. Radiology 2006;238:184–91.

35. Hricak H. MR imaging and MR spectroscopic imaging in the pre-treatment evaluation of prostate cancer. Br J Radiol 2005;78 Spec no 2:S103–11.

36. Jackson AS, Parker CC, Norman AR, et al. Tumour staging using magnetic resonance imaging in clinically localised prostate cancer: relationship to biochemical outcome after neo-adjuvant androgen deprivation and radical radiotherapy. Clin Oncol (R Coll Radiol) 2005;17:167–71.

37. Kurhanewicz J, Vigneron D, Carroll P, and Coakley F. Multiparametric magnetic resonance imaging in prostate cancer: present and future. Curr Opin Urol 2008;18:71–7.

38. Jung SI, Donati OF, Vargas HA, Goldman D, Hricak H, and Akin O. Transition zone prostate cancer: incremental value of diffusion-weighted endorectal MR imaging in tumor detection and assessment of aggressiveness. Radiology 2013;269:493–503.

39. Wu LM, Xu JR, Ye YQ, Lu Q, and Hu JN. The clinical value of diffusion-weighted imaging in combination with T2-weighted imaging in diagnosing prostate carcinoma: a systematic review and meta-analysis. AJR Am J Roentgenol 2012;199:103–10.

40. Rooij M de, Hamoen EH, Futterer JJ, Barentsz JO, and Rovers MM. Accuracy of multiparametric MRI for prostate cancer detection: a meta-analysis. AJR Am J Roentgenol 2014;202:343–51.

41. Futterer JJ, Briganti A, De Visschere P, et al. Can Clinically Significant Prostate Cancer Be Detected with Multiparametric Magnetic Resonance Imaging? A Systematic Review of the Literature. Eur Urol 2015;68:1045–53.

42. Hamoen EHJ, Rooij M de, Witjes JA, Barentsz JO, and Rovers MM. Use of the Prostate Imaging Reporting and Data System (PI-RADS) for Prostate Cancer Detection with Multiparametric Magnetic Resonance Imaging: A Diagnostic Meta-analysis. Eur Urol 2015;67:1112–21.

43. Woo S, Suh CH, Kim SY, Cho JY, and Kim SH. Diagnostic Performance of Prostate Imaging Reporting and Data System Version 2 for Detection of Prostate Cancer: A Systematic Review and Diagnostic Meta-analysis. Eur Urol 2017;72:177–88.

44. Schoots IG, Roobol MJ, Nieboer D, Bangma CH, Steyerberg EW, and Hunink MG. Magnetic resonance imaging-targeted biopsy may enhance the diagnostic accuracy of significant prostate cancer detection compared to standard transrectal ultrasound-guided biopsy: a systematic review and meta-analysis. Eur Urol 2015;68:438–50.

45. Valerio M, Donaldson I, Emberton M, et al. Detection of Clinically Significant Prostate Cancer Using Magnetic Resonance Imaging-Ultrasound Fusion Targeted Biopsy: A Systematic Review. Eur Urol 2015;68:8–19.

46. Ahmed HU, El-Shater Bosaily A, Brown LC, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet 2017;389:815–22.

47. Moldovan PC, Broeck T Van den, Sylvester R, et al. What Is the Negative Predictive Value of Multiparametric Magnetic Resonance Imaging in Excluding Prostate Cancer at Biopsy? A Systematic Review and Meta-analysis from the European Association of Urology Prostate Cancer Guidelines Panel. Eur Urol 2017;72:250–66.

48. Jambor I, Bostrom PJ, Taimen P, et al. Novel biparametric MRI and targeted biopsy improves risk stratification in men with a clinical suspicion of prostate cancer (IMPROD Trial). J Magn Reson Imaging 2017;46:1089–95.

49. Richenberg J, Logager V, Panebianco V, Rouviere O, Villeirs G, and Schoots IG. The primacy of multiparametric MRI in men with suspected prostate cancer. Eur Radiol 2019;29:6940–52.

50. Schoots IG, Petrides N, Giganti F, et al. Magnetic Resonance Imaging in Active Surveillance of Prostate Cancer: A Systematic Review. European Urology 2015;67:627–36.

51. Velasquez MC, Prakash NS, Venkatramani V, Nahar B, and Punnen S. Imaging for the selection and monitoring of men on active surveillance for prostate cancer. Translational Andrology and Urology 2017;7:228–35.

52. Moore CM, Giganti F, Albertsen P, et al. Reporting Magnetic Resonance Imaging in Men on Active Surveillance for Prostate Cancer: The PRECISE Recommendations-A Report of a European School of Oncology Task Force. Eur Urol 2017;71:648–55.

53. Osses DF, Drost FH, Verbeek JFM, et al. Prostate cancer upgrading with serial prostate MRI scans and repeat biopsy in men on active surveillance: are confirmatory biopsies still necessary? BJU Int 2020.

54. Bloch F. Nuclear induction. Physical review 1946;70:460.

55. Bernstein MA, King KF, and Zhou XJ. Handbook of MRI pulse sequences. Elsevier, 2004.

56. Ullrich T, Quentin M, Oelers C, et al. Magnetic resonance imaging of the prostate at 1.5 versus 3.0T: A prospective comparison study of image quality. Eur J Radiol 2017;90:192–7.

57. Beyersdorff D, Taymoorian K, Knosel T, et al. MRI of prostate cancer at 1.5 and 3.0 T: comparison of image quality in tumor detection and staging. AJR Am J Roentgenol 2005;185:1214–20.

58. El-Shater Bosaily A, Arya M, Punwani S, et al. Re: Multiparametric magnetic resonance imaging guided diagnostic biopsy detects significant prostate cancer and could reduce unnecessary biopsies and over detection: a prospective study: J. E. Thompson, D. Moses, R. Shnier, P. Brenner, W. Delprado, L. Ponsky, M.

Pulbrook, M. Bohm, A.-M. Haynes, A. Hayen and P. D. Stricker J Urol 2014;192:67-74. J Urol 2015;193:735–6, discussion 736.

59. Shah ZK, Elias SN, Abaza R, et al. Performance comparison of 1.5-T endorectal coil MRI with 3.0-T nonendorectal coil MRI in patients with prostate cancer. Acad Radiol 2015;22:467–74.

60. Wang J, Tanderup K, Cunha A, et al. Magnetic resonance imaging basics for the prostate brachytherapist. Brachytherapy 2017;16:715–27.

61. Choi H and Ma J. Use of perfluorocarbon compound in the endorectal coil to improve MR spectroscopy of the prostate. AJR Am J Roentgenol 2008;190:1055–9.

62. Gawlitza J, Reiss-Zimmermann M, Thormer G, et al. Impact of the use of an endorectal coil for 3 T prostate MRI on image quality and cancer detection rate. Sci Rep 2017;7:40640.

63. Baur AD, Daqqaq T, Wagner M, et al. T2- and diffusion-weighted magnetic resonance imaging at 3T for the detection of prostate cancer with and without endorectal coil: An intraindividual comparison of image quality and diagnostic performance. Eur J Radiol 2016;85:1075–84.

64. Heijmink SW, Futterer JJ, Hambrock T, et al. Prostate cancer: body-array versus endorectal coil MR imaging at 3 T–comparison of image quality, localization, and staging performance. Radiology 2007;244:184–95.

65. Costa DN, Yuan Q, Xi Y, et al. Comparison of prostate cancer detection at 3-T MRI with and without an endorectal coil: A prospective, paired-patient study. Urol Oncol 2016;34:255.e7–255.e13.

66. Turkbey B, Merino MJ, Gallardo EC, et al. Comparison of endorectal coil and nonendorectal coil T2W and diffusion-weighted MRI at 3 Tesla for localizing prostate cancer: correlation with whole-mount histopathology. J Magn Reson Imaging 2014;39:1443–8.

67. Barentsz JO, Richenberg J, Clements R, et al. ESUR prostate MR guidelines 2012. Eur Radiol 2012;22:746–57.

68. Hennig J, Nauerth A, and Friedburg H. RARE imaging: A fast imaging method for clinical MR. Magnetic Resonance in Medicine 1986;3:823–33.

69. Carr HY and Purcell EM. Effects of diffusion on free precession in nuclear magnetic resonance experiments. Physical review 1954;94:630.

70. Meiboom S and Gill D. Modified spin-echo method for measuring nuclear relaxation times. Review of scientific instruments 1958;29:688–91.

71. Mugler J and Brookeman J. Ultra-long echo trains for Rapid 3D T2-weighted turbo-spin-echo imaging. Proc. 11th Int. Soc. Magn. Reson. Med 2003;970.

72. Mugler J, Meyer H, and Kiefer B. Practical implementation of optimized tissue-specific prescribed signal evolutions for improved turbo-spin-echo imaging. In: *Proceedings of the International Society for Magnetic Resonance in Medicine 11th Meeting*. Vol. 203.

73. Busse RF, Hariharan H, Vu A, and Brittain JH. Fast spin echo sequences with very long echo trains: Design of variable refocusing flip angle schedules and generation of clinical T2 contrast. Magnetic Resonance in Medicine 2006;55:1030–7.

74. Rosenkrantz AB, Neil J, Kong X, et al. Prostate Cancer: Comparison of 3D T2-Weighted With Conventional 2D T2-Weighted Imaging for Image Quality and Tumor Detection. American Journal of Roentgenology 2010;194:446–52.

75. Stejskal EO and Tanner JE. Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. Journal of Chemical Physics 1965;42:288.

76. Haase A, Frahm J, Hanicke W, and Matthaei D. 1H NMR chemical shift selective (CHESS) imaging. Phys Med Biol 1985;30:341–4.

77. Nagy Z and Weiskopf N. Efficient fat suppression by slice-selection gradient reversal in twice-refocused diffusion encoding. Magn Reson Med 2008;60:1256–60.

78. Pruessmann KP, Weiger M, Scheidegger MB, and Boesiger P. SENSE: sensitivity encoding for fast MRI. Magn Reson Med 1999;42:952–62.

79. Griswold MA, Jakob PM, Heidemann RM, et al. Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn Reson Med 2002;47:1202–10.

80. Rieseberg S, Frahm J, and Finsterbusch J. Two-dimensional spatially-selective RF excitation pulses in echo-planar imaging. Magn Reson Med 2002;47:1186–93.

81. Bruder H, Fischer H, Reinfelder HE, and Schmitt F. Image reconstruction for echo planar imaging with nonequidistant k-space sampling. Magnetic resonance in medicine 1992;23:311–23.

82. Rosenkrantz AB, Chandarana H, Hindman N, et al. Computed diffusion-weighted imaging of the prostate at 3 T: impact on image quality and tumour detection. European radiology 2013;23:3170–7.

83. Kanal E, Maravilla K, and Rowley HA. Gadolinium contrast agents for CNS imaging: current concepts and clinical evidence. AJNR Am J Neuroradiol 2014;35:2215–26.

84. Khalifa F, Soliman A, El-Baz A, et al. Models and methods for analyzing DCE-MRI: A review. Medical Physics 2014;41:124301.

85. Franiel T, Hamm B, and Hricak H. Dynamic contrast-enhanced magnetic resonance imaging and pharmacokinetic models in prostate cancer. Eur Radiol 2011;21:616–26.

86. Verma S, Turkbey B, Muradyan N, et al. Overview of dynamic contrast-enhanced MRI in prostate cancer diagnosis and management. AJR Am J Roentgenol 2012;198:1277–88.

87. Kalavagunta C, Michaeli S, and Metzger GJ. In vitro Gd-DTPA relaxometry studies in oxygenated venous human blood and aqueous solution at 3 and 7 T. Contrast Media Mol Imaging 2014;9:169–76.

88. Yu MH, Lee JM, Yoon JH, Kiefer B, Han JK, and Choi BI. Clinical application of controlled aliasing in parallel imaging results in a higher acceleration (CAIPIRINHA)-volumetric interpolated breathhold (VIBE) sequence for gadoxetic acid-enhanced liver MR imaging. J Magn Reson Imaging 2013;38:1020–6.

89. Boesen L, Norgaard N, Logager V, et al. Assessment of the Diagnostic Accuracy of Biparametric Magnetic Resonance Imaging for Prostate Cancer in Biopsy-Naive Men: The Biparametric MRI for Detection of Prostate Cancer (BIDOC) Study. JAMA Netw Open 2018;1:e180219.

90. Schieda N, Blaichman JI, Costa AF, et al. Gadolinium-Based Contrast Agents in Kidney Disease: A Comprehensive Review and Clinical Practice Guideline Issued by the Canadian Association of Radiologists. Canadian journal of kidney health and disease 2018;5:136–50.

91. McDonald RJ, McDonald JS, Kallmes DF, et al. Gadolinium Deposition in Human Brain Tissues after Contrast-enhanced MR Imaging in Adult Patients without Intracranial Abnormalities. Radiology 2017;285:546–54.

92. Kanal E and Tweedle MF. Residual or retained gadolinium: practical implications for radiologists and our patients. Radiology 2015;275:630–4.

93. Villers A, Lemaitre L, Haffner J, and Puech P. Current status of MRI for the diagnosis, staging and prognosis of prostate cancer: implications for focal therapy and active surveillance. Curr Opin Urol 2009;19:274–82.

94. Mueller-Lisse U, Mueller-Lisse U, Scheidler J, Klein G, and Reiser M. Reproducibility of image interpretation in MRI of the prostate: application of the sextant framework by two different radiologists. Eur Radiol 2005;15:1826–33.

95. Rosenkrantz AB, Ginocchio LA, Cornfeld D, et al. Interobserver Reproducibility of the PI-RADS Version 2 Lexicon: A Multicenter Study of Six Experienced Prostate Radiologists. Radiology 2016;280:793–804.

96. Rosenkrantz AB, Kim S, Lim RP, et al. Prostate cancer localization using multiparametric MR imaging: comparison of Prostate Imaging Reporting and Data System (PI-RADS) and Likert scales. Radiology 2013;269:482–92.

97. Brizmohun Appayya M, Sidhu HS, Dikaios N, et al. Characterizing indeterminate (Likert-score 3/5) peripheral zone prostate lesions with PSA density, PI-RADS scoring and qualitative descriptors on multiparametric MRI. Br J Radiol 2018;91:20170645.

98. Barrett T, Rajesh A, Rosenkrantz AB, Choyke PL, and Turkbey B. PI-RADS version 2.1: one small step for prostate MRI. Clin Radiol 2019;74:841–52.

99. Gupta RT, Spilseth B, and Froemming AT. How and why a generation of radiologists must be trained to accurately interpret prostate mpMRI. Abdominal Radiology 2016;41:803–4.

100. Suzuki K. A review of computer-aided diagnosis in thoracic and colonic imaging. Quantitative imaging in medicine and surgery 2012;2:163.

101. Narayana PA and Borthakur A. Effect of radio frequency inhomogeneity correction on the reproducibility of intra-cranial volumes using MR image data. Magnetic Resonance in Medicine 1995;33:396–400.

102. Singh M and NessAiver M. Accurate intensity correction for endorectal surface coil MR imaging of the prostate. In: *IEEE Conference on Nuclear Science Symposium and Medical Imaging*. IEEE:1307–9.

103. Haselgrove J and Prammer M. An algorithm for compensation of surface-coil images for sensitivity of the surface coil. Magnetic Resonance Imaging 1986;4:469–72.

104. Moyher SE, Vigneron DB, and Nelson SJ. Surface coil MR imaging of the human brain with an analytic reception profile correction. Journal of Magnetic Resonance Imaging 1995;5:139–44.

105. Liney GP, Turnbull LW, and Knowles AJ. A simple method for the correction of endorectal surface coil inhomogeneity in prostate imaging. Journal of Magnetic Resonance Imaging 1998;8:994–7.

106. Metzger GJ, Kalavagunta C, Spilseth B, et al. Detection of Prostate Cancer: Quantitative Multiparametric MR Imaging Models Developed Using Registered Correlative Histopathology. Radiology 2016;279:805–16.

107. Vos PC, Barentsz JO, Karssemeijer N, and Huisman HJ. Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis. Phys Med Biol 2012;57:1527–42.

108. Artan Y, Haider MA, Langer DL, et al. Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields. IEEE Trans Image Process 2010;19:2444–55.

109. Vos PC, Hambrock T, Barenstz JO, and Huisman HJ. Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MRI. Phys Med Biol 2010;55:1719–34.

110. Vos EK, Kobus T, Litjens GJ, et al. Multiparametric Magnetic Resonance Imaging for Discriminating Low-Grade From High-Grade Prostate Cancer. Invest Radiol 2015;50:490–7.

111. Niaf E, Flamary R, Rouviere O, Lartizien C, and Canu S. Kernel-based learning from both qualitative and quantitative labels: application to prostate cancer diagnosis based on multiparametric MR imaging. IEEE Trans Image Process 2014;23:979–91.

112. Niaf E, Rouviere O, Mege-Lechevallier F, Bratan F, and Lartizien C. Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. Phys Med Biol 2012;57:3833–51.

113. Ozer S, Langer DL, Liu X, et al. Supervised and unsupervised methods for prostate cancer segmentation with multispectral MRI. Medical Physics 2010;37:1873–83.

114. Matulewicz L, Jansen JF, Bokacheva L, et al. Anatomic segmentation improves prostate cancer detection with artificial neural networks analysis of 1H magnetic resonance spectroscopic imaging. Journal of Magnetic Resonance Imaging 2014;40:1414–21.

115. Samiee M, Thomas G, and Fazel-Rezai R. Semi-Automatic Prostate Segmentation of MR Images Based on Flow Orientation. In: *2006 IEEE International Symposium on Signal Processing and Information Technology*:203–7. DOI: 10.1109/ISSPIT.2006.270797.

116. Flores-Tapia D, Thomas G, Venugopal N, McCurdy B, and Pistorius S. Semi automatic MRI prostate segmentation based on wavelet multiscale products. In: *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*:3020–3. DOI: 10.1109/IEMBS.2008.4649839.

117. Toth R, Chappelow J, Rosen M, Pungavkar S, Kalyanpur A, and Madabhushi A. Multi-attribute non-initializing texture reconstruction based active shape model (MANTRA). Med Image Comput Comput Assist Interv 2008;11:653–61.

118. Dowling JA, Fripp J, Chandra S, et al. Fast Automatic Multi-atlas Segmentation of the Prostate from 3D MR Images. In: Prostate Cancer Imaging. Image Analysis and Image-Guided Interventions. Springer Berlin Heidelberg:10–21.

119. Litjens G, Debats O, Barentsz J, Karssemeijer N, and Huisman H. Computer-aided detection of prostate cancer in MRI. IEEE Trans Med Imaging 2014;33:1083–92.

120. Martin S, Daanen V, and Troccaz J. Atlas-based prostate segmentation using an hybrid registration. International Journal of Computer Assisted Radiology and Surgery 2008;3:485–92.

121. Allen PD, Graham J, Williamson DC, and Hutchinson CE. Differential segmentation of the prostate in MR images using combined 3D shape modelling and voxel classification. In: *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.*410–3. DOI: 10.1109/ISBI.2006.1624940.

122. Cheng R, Roth HR, Lay N, et al. Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks. J Med Imaging (Bellingham) 2017;4:041302.

123. Xie S and Tu Z. Holistically-nested edge detection. In: *Proceedings of the IEEE international conference on computer vision*:1395–403.

124. Ronneberger O, Fischer P, and Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing:234–41.

125. Long J, Shelhamer E, and Darrell T. Fully convolutional networks for semantic segmentation. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*:3431–40. DOI: 10.1109/CVPR.2015.7298965.

126. Zabihollahy F, Schieda N, Krishna Jeyaraj S, and Ukwatta E. Automated segmentation of prostate zonal anatomy on T2-weighted (T2W) and apparent diffusion coefficient (ADC) map MR images using U-Nets. Medical Physics 2019;46:3078–90.

127. Meyer A, Rakr M, Schindele D, et al. Towards Patient-Individual PI-Rads v2 Sector Map: Cnn for Automatic Segmentation of Prostatic Zones From T2-Weighted MRI. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*:696–700. DOI: 10.1109/ISBI.2019.8759572.

128. Viswanath S, Bloch BN, Genega E, et al. A Comprehensive Segmentation, Registration, and Cancer Detection Scheme on 3 Tesla In Vivo Prostate DCE-MRI. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008. Springer Berlin Heidelberg:662–9.

129. Chappelow J, Bloch BN, Rofsky N, et al. Elastic registration of multimodal prostate MRI and histology via multiattribute combined mutual information. Med Phys 2011;38:2005–18.

130. Armato S. G. r, Huisman H, Drukker K, et al. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. J Med Imaging 2018;5:044501.

131. Viswanath S, Tiwari P, Rosen M, and Madabhushi A. A meta-classifier for detecting prostate cancer by quantitative integration of in vivo magnetic resonance spectroscopy and magnetic resonance imaging. Medical Imaging 2008: Computer-Aided Diagnosis, 6915, International Society for Optics and Photonics, 2008:69153.

132. Chan I, Wells III W, Mulkern R, et al. Detection of prostate cancer by integration of line-scan diffusion, t2-mapping and t2-weighted magnetic resonance imaging; a multichannel statistical classifier. Med Phys 2003;30:2390–8.

133. Langer D, Kwast T van der, Evans A, Trachtenberg J, Wilson B, and Haider M. Prostate cancer detection with multi-parametric MRI: logistic regression analysis of quantitative t2, diffusion-weighted imaging, and dynamic contrast-enhanced MRI. J. Magn. Reson. Imaging 2009;30:327.

134. Fehr D, Veeraraghavan H, Wibmer A, et al. Automatic classification of prostate cancer gleason scores from multiparametric magnetic resonance images. Proc. Natl. Acad. Sci 2015;112:6265–73.

135. Liu X and Yetik I. Automated prostate cancer localization without the need for peripheral zone extraction using multiparametric MRI. Med Phys 2011;38:2986–94.

136. Lopes R, Ayache A, Makni N, et al. Prostate cancer characterization on MR images using fractal features. Med Phys 2011;38:83–95.

137. Sung Y, Kwon HJ, Park BW, et al. Prostate cancer detection on dynamic contrast-enhanced MRI: computer-aided diagnosis versus single perfusion parameter maps. 2011.

138. Jin J, Zhang L, Leng E, Metzger G, and Koopmeiners J. Detection of prostate cancer with multiparametric MRI utilizing the anatomic structure of the prostate. Stat Med 2018;37:3214–29.

139. Viswanath S, Chirra P, Yim M, et al. Comparing radiomic classifiers and classifier ensembles for detection of peripheral zone prostate tumors on t2-weighted MRI: a multi-site study. In: vol. 19. 2019:22.

140. Viswanath S, Bloch N, Chappelow J, et al. Central gland and peripheral zone prostate tumors have significantly different quantitative imaging signatures on 3 tesla endorectal, in vivo t2-weighted MR imagery. Vol. 36. J. Magn. Reson. Imaging, 2012:213–24.

141. Giannini V, Mazzetti S, Vignati A, et al. A fully automatic computer aided diagnosis system for peripheral zone prostate cancer detection using multi-parametric magnetic resonance imaging. Comput. Med. Imaging Graph 2015;46:219–26.

142. Kelm B, Menze B, Zechmann C, Baudendistel K, and Hamprecht F. Automated estimation of tumor probability in prostate magnetic resonance spectroscopic imaging: Pattern recognition vs quantification. Magn. Reson Med 2007;57:150–9.

143. Tiwari P, Viswanath S, Kurhanewicz J, Sridhar A, and Madabhushi A. Multimodal wavelet embedding representation for data combination (maWERic): integrating magnetic resonance imaging and spectroscopy for prostate cancer detection. NMR Biomed 2012;25:607–19.

144. Tiwari P, Kurhanewicz J, and Madabhushi A. Multi-kernel graph embedding for detection, gleason grading of prostate cancer via MRI/MRS, Med. Image Anal. 2013;17:219–35.

145. Wang Z, Liu C, Cheng D, Wang L, Yang X, and Cheng K. Automated detection of clinically significant prostate cancer in mp-MRI images based on an end– to-end deep neural network. IEEE Trans. Med. Imaging 2018;37:1127–39.

146. Sumathipala Y, Lay N, Turkbey B, Smith C, Choyke P, and Summers R. Prostate cancer detection from multi-institution multiparametric MRIs using deep convolutional neural networks. 2018.

147. Peng Y, Jiang Y, Yang C, et al. Quantitative analysis of multiparametric prostate MR images: Differentiation between prostate cancer and normal tissue and correlation with gleason scores a computer-aided diagnosis development study. Radiology 2013;267:787–96.

148. Allam CK, Bostwick DG, Hayes JA, et al. Interobserver variability in the diagnosis of high-grade prostatic intraepithelial neoplasia and adenocarcinoma. Mod Pathol 1996;9:742–51.

149. Montironi R, Mazzuccheli R, Scarpelli M, Lopez-Beltran A, Fellegara G, and Algaba F. Gleason grading of prostate cancer in needle biopsies or radical prostatectomy specimens: contemporary approach, current clinical significance and sources of pathology discrepancies. BJU Int 2005;95:1146–52.

150. Puech P, Betrouni N, Makni N, Dewalle AS, Villers A, and Lemaitre L. Computer-assisted diagnosis of prostate cancer using DCE-MRI data: design, implementation and preliminary results. Int J Comput Assist Radiol Surg 2009;4:1–10.

151. Tiwari P, Rosen M, and Madabhushi A. A hierarchical spectral clustering and nonlinear dimensionality reduction scheme for detection of prostate cancer from magnetic resonance spectroscopy (MRS). Med Phys 2009;36:3927–39.

152. Kwak J, Xu S, Wood B, et al. Automated prostate cancer detection using t2-weighted and high-b-value diffusion-weighted magnetic resonance imaging. Med Phys 2015;42:2368–78.

153. Liu P, Wang S, Turkbey B, et al. A prostate cancer computer-aided diagnosis system using multimodal magnetic resonance imaging and targeted biopsy labels. Medical Imaging 2013: Computer-Aided Diagnosis, 8670, International Society for Optics and Photonics, 2013:86701.

154. Moradi M, Salcudean S, Chang S, et al. Multiparametric MRI maps for detection and grading of dominant prostate tumors. J. Magn. Reson. Imaging 2012;35:1403–13.

155. Shah V, Turkbey B, Mani H, et al. Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. Med Phys 2012;39:4093–103.

156. Yang X, Liu C, Wang Z, et al. Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI. Med. Image Anal 2017;42:212–27.

157. Schelb P, Kohl S, Radtke J, et al. Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. 2019.

158. Wildeboer RR, Sloun RJG van, Postema AW, et al. Accurate validation of ultrasound imaging of prostate cancer: a review of challenges in registration of imaging and histopathology. Journal of Ultrasound 2018;21:197–207.

159. Kalavagunta C, Zhou X, Schmechel SC, and Metzger GJ. Registration of in vivo prostate MRI and pseudo-whole mount histology using Local Affine Transformations guided by Internal Structures (LATIS). J Magn Reson Imaging 2015;41:1104–14.

160. Kwak JT, Sankineni S, Xu S, et al. Prostate Cancer: A Correlative Study of Multiparametric MR Imaging and Digital Histopathology. Radiology 2017;285:147–56.

161. Losnegård A, Reisæter L, Halvorsen OJ, et al. Intensity-based volumetric registration of magnetic resonance images and whole-mount sections of the prostate. Computerized Medical Imaging and Graphics 2018;63:24–30.

162. Cameron A, Khalvati F, Haider MA, and Wong A. MAPS: A Quantitative Radiomics Approach for Prostate Cancer Detection. IEEE Trans Biomed Eng 2016;63:1145–56.

163. Vos P, Hambrock T, Kaa C Hulsbergen-van de, Futterer J, Barentsz J, and Huisman H. Computerized analysis of prostate lesions in the peripheral zone using dynamic contrast enhanced MRI. Med Phys 2008;35:888–99.

164. Chitalia RD and Kontos D. Role of texture analysis in breast MRI as a cancer biomarker: A review. Journal of Magnetic Resonance Imaging 2019;49:927–38.

165. Rizzo S, Botta F, Raimondi S, et al. Radiomics: the facts and the challenges of image analysis. European radiology experimental 2018;2:1–8.

166. Nioche C, Orlhac F, Boughdad S, et al. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. Cancer Research 2018;78:4786.

167. Griethuysen JJM van, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research 2017;77:e104.

168. Sussman MS, Vidarsson L, Pauly JM, and Cheng HL. A technique for rapid single-echo spin-echo T2 mapping. Magn Reson Med 2010;64:536–45.

169. Liney GP, Knowles AJ, Manton DJ, Turnbull LW, Blackband SJ, and Horsman A. Comparison of conventional single echo and multi-echo sequences with a fast spin-echo sequence for quantitative T2 mapping: application to the prostate. J Magn Reson Imaging 1996;6:603–7.

170. Does MD and Gore JC. Complications of nonlinear echo time spacing for measurement of T (2). NMR Biomed 2000;13:1–7.

171. Poon CS and Henkelman RM. Practical T2 quantitation for clinical applications. J Magn Reson Imaging 1992;2:541–53.

172. Milford D, Rosbach N, Bendszus M, and Heiland S. Mono-Exponential Fitting in T2-Relaxometry: Relevance of Offset and First Echo. PLoS One 2015;10:e0145255.

173. Ben-Eliezer N, Sodickson DK, and Block KT. Rapid and accurate T2 mapping from multi-spin-echo data using Bloch-simulation-based reconstruction. Magn Reson Med 2015;73:809–17.

174. Lebel RM and Wilman AH. Transverse relaxometry with stimulated echo compensation. Magn Reson Med 2010;64:1005–14.

175. Mendlik T, Faber SC, Weber J, et al. T2 Quantitation of Human Articular Cartilage in a Clinical Setting at 1.5 T: Implementation and Testing of Four Multiecho Pulse Sequence Designs for Validity. Investigative Radiology 2004;39:288–99.

176. Duncan JS, Bartlett P, and Barker GJ. Technique for measuring hippocampal T2 relaxation time. AJNR Am J Neuroradiol 1996;17:1805–10.

177. Liess C, Lüsse S, Karger N, Heller M, and Glüer CC. Detection of changes in cartilage water content using MRI T2-mapping in vivo. Osteoarthritis and Cartilage 2002;10:907–13.

178. Matzat SJ, McWalter EJ, Kogan F, Chen W, and Gold GE. T2 Relaxation time quantitation differs between pulse sequences in articular cartilage. Journal of Magnetic Resonance Imaging 2015;42:105–13.

179. Chan I, Wells III W, Mulkern RV, et al. Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier. Medical Physics 2003;30:2390–8.

180. Metens T, Miranda D, Absil J, and Matos C. What is the optimal b value in diffusion-weighted MR imaging to depict prostate cancer at 3T? European Radiology 2012;22:703–9.

181. Riches SF, Hawtin K, Charles-Edwards EM, and Souza NM de. Diffusion-weighted imaging of the prostate and rectal wall: comparison of biexponential and monoexponential modelled diffusion and associated perfusion coefficients. NMR Biomed 2009;22:318–25.

182. Thörmer G, Otto J, Reiss-Zimmermann M, et al. Diagnostic value of ADC in patients with prostate cancer: influence of the choice of b values. European Radiology 2012;22:1820–8.

183. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, and Laval-Jeantet M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. Radiology 1988;168:497–505.

184. Itatani R, Namimoto T, Yoshimura A, et al. Clinical utility of the normalized apparent diffusion coefficient for preoperative evaluation of the aggressiveness of prostate cancer. Jpn J Radiol 2014;32:685–91.

185. Vargas HA, Akin O, Franiel T, et al. Diffusion-weighted endorectal MR imaging at 3 T for prostate cancer: tumor detection and assessment of aggressiveness. Radiology 2011;259:775–84.

186. Wu X, Reinikainen P, Vanhanen A, et al. Correlation between apparent diffusion coefficient value on diffusion-weighted MR imaging and Gleason score in prostate cancer. Diagnostic and Interventional Imaging 2017;98:63–71.

187. Blackledge MD, Leach MO, Collins DJ, and Koh DM. Computed diffusion-weighted MR imaging may improve tumor detection. Radiology 2011;261:573–81.

188. Tofts PS. Modeling tracer kinetics in dynamic Gd-DTPA MR imaging. J Magn Reson Imaging 1997;7:91–101.

189. Vos EK, Litjens GJ, Kobus T, et al. Assessment of prostate cancer aggressiveness using dynamic contrast-enhanced magnetic resonance imaging at 3 T. Eur Urol 2013;64:448–55.

190. Wei C, Jin B, Szewczyk-Bieda M, et al. Quantitative parameters in dynamic contrast-enhanced magnetic resonance imaging for the detection and characterization of prostate cancer. Oncotarget 2018;9:15997–6007.

191. Ma XZ, Lv K, Sheng JL, et al. Application evaluation of DCE-MRI combined with quantitative analysis of DWI for the diagnosis of prostate cancer. Oncol Lett 2019;17:3077–84.

192. Huang W, Chen Y, Fedorov A, et al. The Impact of Arterial Input Function Determination Variations on Prostate Dynamic Contrast-Enhanced Magnetic Resonance Imaging Pharmacokinetic Modeling: A Multicenter Data Analysis Challenge. Tomography 2016;2:56–66.

193. Larsson HB, Stubgaard M, Frederiksen JL, Jensen M, Henriksen O, and Paulson OB. Quantitation of blood-brain barrier defect by magnetic resonance imaging and gadolinium-DTPA in patients with multiple sclerosis and brain tumors. Magn Reson Med 1990;16:117–31.

194. Larsson HB and Tofts PS. Measurement of blood-brain barrier permeability using dynamic Gd-DTPA scanning–a comparison of methods. Magn Reson Med 1992;24:174–6.

195. Cron GO, Foottit C, Yankeelov TE, Avruch LI, Schweitzer ME, and Cameron I. Arterial input functions determined from MR signal magnitude and phase for quantitative dynamic contrast-enhanced MRI in the human pelvis. Magnetic resonance in medicine 2011;66:498–504.

196. Fritz-Hansen T, Rostrup E, Larsson HB, Søndergaard L, Ring P, and Henriksen O. Measurement of the arterial concentration of Gd-DTPA using MRI: a step toward quantitative perfusion imaging. Magnetic Resonance in Medicine 1996;36:225–31.

197. Schaaf I Van der, Vonken EJ, Waaijer A, Velthuis B, Quist M, and Van Osch T. Influence of partial volume on venous output and arterial input function. American journal of neuroradiology 2006;27:46–50.

198. Ivancevic MK, Zimine I, Montet X, et al. Inflow effect correction in fast gradient-echo perfusion imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 2003;50:885–91.

199. Yang C, Karczmar GS, Medved M, and Stadler WM. Multiple reference tissue method for contrast agent arterial input function estimation. Magnetic Resonance in Medicine 2007;58:1266–75.

200. Parker GJ, Roberts C, Macdonald A, et al. Experimentally-derived functional form for a population-averaged high-temporal-resolution arterial input function for dynamic contrast-enhanced MRI. Magn Reson Med 2006;56:993–1000.

201. Abraham B and Nair MS. Automated grading of prostate cancer using convolutional neural network and ordinal class classifier. Informatics in Medicine Unlocked 2019;17:100256.

202. Yuan Y, Qin W, Buyyounouski M, et al. Prostate cancer classification with multiparametric MRI transfer learning model. Med Phys 2019;46:756–65.

203. Kwon D, Reis I, Breto A, et al. Classification of suspicious lesions on prostate multiparametric MRI using machine learning. 2018.

204. Lay N, Tsehay Y, Sumathipala Y, et al. A Decomposable Model for the Detection of Prostate Cancer in Multi-parametric MRI. In: *BT - Medical Image Computing and Computer Assisted Intervention â MICCAI 2018*. Springer International Publishing, Cham, 2018:930–9.

205. Song Y, Zhang YD, Yan X, et al. Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI. J. Magn. Reson. Imaging 2018;48:1570–7.

206. Bishop CM. Pattern recognition and machine learning. springer, 2006.

207. Giannini V, Vignati A, Mazzetti S, et al. A prostate CAD system based on multiparametric analysis of DCE T1-w, and DW automatically registered images. Vol. 8670. SPIE Medical Imaging. SPIE, 2013. URL: https://doi.org/10.1117/12.2006336.

208. Cao R, Bajgiran A, Mirak S, et al. Joint prostate cancer detection and gleason score prediction in mp-MRI via focalnet. IEEE Trans. Med. Imaging 2019:1.

209. Kiraly A, Nader C, Tuysuzoglu A, et al. Deep Convolutional Encoder-Decoders for Prostate Cancer Detection and Classification BT - Medical Image Computing and Computer Assisted Intervention â MICCAI 2017. Vol. 4. 89–4. Springer International Publishing, 2017:97.

210. Le M, Chen J, Wang L, et al. Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. Phys. Med. Biol 2017;62:6497.

211. Wang X, Yang W, Weinreb J, et al. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. Sci. Rep 2017;7:15415.

212. Xu H, Baxter J, Akin O, and Cantor-Rivera D. Prostate cancer detection using residual networks. Int J. Comput. Assist. Radiol. Surg 2019:1–4.

213. Zhuang X, Rhode KS, Razavi RS, Hawkes DJ, and Ourselin S. A Registration-Based Propagation Framework for Automatic Whole Heart Segmentation of Cardiac MRI. IEEE Transactions on Medical Imaging 2010;29:1612–25.

214. Chen Hm, Varshney PK, and Slamani MA. On registration of regions of interest (ROI) in video sequences. In: *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.* IEEE:313–8.

215. Freedman D and Diaconis P. On the histogram as a density estimator:L2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 1981;57:453–76.

216. Hajela P. Genetic search - An approach to the nonconvex optimization problem. AIAA Journal 1990;28:1205–10.

217. Paninski L. Estimation of entropy and mutual information. Neural computation 2003;15:1191–253.

218. Alimo SR, Beyhaghi P, and Bewley TR. Optimization combining derivative-free global exploration with derivative-based local refinement. In: *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE:2531–8.

219. Swindle P, Eastham JA, Ohori M, et al. Do margins matter? The prognostic significance of positive surgical margins in radical prostatectomy specimens. J Urol 2008;179:S47–51.

220. McNeal JE, Villers AA, Redwine EA, Freiha FS, and Stamey TA. Capsular penetration in prostate cancer. Significance for natural history and treatment. Am J Surg Pathol 1990;14:240–7.

221. Gurcan MN, Boucheron L, Can A, Madabhushi A, Rajpoot N, and Yener B. Histopathological Image Analysis: A Review. IEEE Rev Biomed Eng 2009;2:147–71.

222. Metzger GJ, Dankbar SC, Henriksen J, Rizzardi AE, Rosener NK, and Schmechel SC. Development of multigene expression signature maps at the protein level from digitized immunohistochemistry slides. PLoS One 2012;7:e33520.

223. Krajewska M, Smith LH, Rong J, et al. Image Analysis Algorithms for Immunohistochemical Assessment of Cell Death Events and Fibrosis in Tissue Sections. J Histochem Cytochem 2009;57:649–63.

224. Kather JN, Weis CA, Bianconi F, et al. Multi-class texture analysis in colorectal cancer histology. Sci Rep 2016;6:27988.

225. Arevalo J, Cruz-Roa A, Arias V, Romero E, and Gonzalez FA. An unsupervised feature learning framework for basal cell carcinoma image analysis. Artif Intell Med 2015;64:131–45.

226. Cruz-Roa A, Gilmore H, Basavanhally A, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. Sci Rep 2017;7.

227. Sharma H, Zerbe N, Klempert I, Hellwich O, and Hufnagl P. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. Comput Med Imaging Graph 2017;61:2–13.

228. Litjens G, Sanchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.

229. Khosravi P, Kazemi E, Imielinski M, Elemento O, and Hajirasouliha I. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. EBioMedicine 2018;27:317–28.

230. Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. Mod Pathol 2004;17:292–306.

231. Montironi R, Mazzucchelli R, Lopez-Beltran A, Scarpelli M, and Cheng L. Prostatic intraepithelial neoplasia: its morphological and molecular diagnosis and clinical significance. BJU Int 2011;108:1394–401.

232. Shah RB and Zhou M. Atypical cribriform lesions of the prostate: clinical significance, differential diagnosis and current concept of intraductal carcinoma of the prostate. Adv Anat Pathol 2012;19:270–8.

233. Guo CC and Epstein JI. Intraductal carcinoma of the prostate on needle biopsy: Histologic features and clinical significance. Mod Pathol 2006;19:1528–35.

234. Herawi M and Epstein JI. Immunohistochemical antibody cocktail staining (p63/HMWCK/AMACR) of ductal adenocarcinoma and Gleason pattern 4 cribriform

and noncribriform acinar adenocarcinomas of the prostate. Am J Surg Pathol 2007;31:889–94.

235. Signoretti S, Waltregny D, Dilks J, et al. p63 is a prostate basal cell marker and is required for prostate development. Am J Pathol 2000;157:1769–75.

236. Wojno KJ and Epstein JI. The utility of basal cell-specific anti-cytokeratin antibody (34 beta E12) in the diagnosis of prostate cancer. A review of 228 cases. Am J Surg Pathol 1995;19:251–60.

237. Rubin MA, Zhou M, Dhanasekaran SM, et al. alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. Jama 2002;287:1662–70.

238. Luo J, Zha S, Gage WR, et al. Alpha-methylacyl-CoA racemase: a new molecular marker for prostate cancer. Cancer Res 2002;62:2220–6.

239. Ng VW, Koh M, Tan SY, and Tan PH. Is triple immunostaining with 34betaE12, p63, and racemase in prostate cancer advantageous? A tissue microarray study. Am J Clin Pathol 2007;127:248–53.

240. Rizzardi AE, Rosener NK, Koopmeiners JS, et al. Evaluation of protein biomarkers of prostate cancer aggressiveness. BMC Cancer 2014;14:244.

241. Rizzardi AE, Johnson AT, Vogel RI, et al. Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring. Diagn Pathol 2012;7:42.

242. Rizzardi AE, Vogel RI, Koopmeiners JS, et al. Elevated HA and HMMR are associated with biochemical failure in patients with intermediate grade prostate tumors. Cancer 2014;120:1800–9.

243. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825–30.

244. Dabir PD, Ottosen P, Hoyer S, and Hamilton-Dutoit S. Comparative analysis of three- and two-antibody cocktails to AMACR and basal cell markers for the immunohistochemical diagnosis of prostate carcinoma. Diagn Pathol 2012;7:81.

245. Kuefer R, Varambally S, Zhou M, et al. alpha-Methylacyl-CoA racemase: expression levels of this novel cancer biomarker depend on tumor differentiation. Am J Pathol 2002;161:841–8.

246. Kothari S, Phan JH, Stokes TH, Osunkoya AO, Young AN, and Wang MD. Removing batch effects from histopathological images for enhanced cancer diagnosis. IEEE J Biomed Health Inform 2014;18:765–72.

247. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*:1107–10. DOI: 10.1109/ISBI.2009.5193250.

248. Glass G, Papin JA, and Mandell JW. SIMPLE: a sequential immunoperoxidase labeling and erasing method. J Histochem Cytochem 2009;57:899–905.

249. Loos CM van der. Multiple immunoenzyme staining: methods and visualizations for the observation with spectral imaging. J Histochem Cytochem 2008;56:313–28.

250. Lotan TL, Gumuskaya B, Rahimi H, et al. Cytoplasmic PTEN protein loss distinguishes intraductal carcinoma of the prostate from high-grade prostatic intraepithelial neoplasia. Mod Pathol 2013;26:587–603.

251. Hanchuan P, Fuhui L, and Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 2005;27:1226–38.

252. Epstein JI, Pizov G, and Walsh PC. Correlation of pathologic findings with progression after radical retropubic prostatectomy. Cancer 1993;71:3582–93.

253. Frank E and Hall M. A Simple Approach to Ordinal Classification. In: Machine Learning: ECML 2001. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001:145–56.

254. Landis JR and Koch GG. The measurement of observer agreement for categorical data. biometrics 1977:159–74.

255. Gordetsky J and Epstein J. Grading of prostatic adenocarcinoma: current state and prognostic implications. Diagn Pathol 2016;11:25.

256. Schwier M, Griethuysen J van, Vangel MG, et al. Repeatability of Multiparametric Prostate MRI Radiomics Features. Scientific Reports 2019;9:9441.

257. Petrick N, Sahiner B, Armato S. G. r, et al. Evaluation of computer-aided detection and diagnosis systems. Med Phys 2013;40:087001.

258. Kallergi M, Carney GM, and Gaviria J. Evaluating the performance of detection algorithms in digital mammography. Med Phys 1999;26:267–75.

259. Leng E, Spilseth B, and Metzger GJ. Clinical usage and impact of predictive models of prostate cancer on multiparametric MRI: a single-observer exploratory evaluation. Proc Intl Soc Mag Reson Med 2017;26:4385.

260. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology standards for quantitative imaging biomarkers. Radiology 2015;277:813–25.

261. Traverso A, Wee L, Dekker A, and Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys 2018;102:1143–58.

262. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26:1045–57.

263. Gibbs P, Tozer DJ, Liney GP, and Turnbull LW. Comparison of quantitative T2 mapping and diffusion-weighted imaging in the normal and pathologic prostate. Magn Reson Med 2001;46:1054–8.

# Appendix A

# Data description

The methods used to obtain the data used to develop the methods described in Chapters 3, 5, and 6 are detailed in a previous work.[106] A condensed summary of the methods is given here.

The data were obtained from a cohort of 34 patients with biopsy-proven PCa who were seen at the University of Minnesota between November 2009 and January 2012 and subsequently underwent radical prostatectomy (RP) for definitive treatment of their disease. The characteristics of the patients are summarized in Table A.1.

All patients first received 3 T mpMRI scans at least 6 weeks after their last biopsy and before the surgery. The mpMRI data were acquired on a Siemens clinical 3T scanner using a combination of a surface array coil (SAC) and an inflatable endorectal coil (ERC) in accordance with PI-RADS v2 guidelines.[2] See Table A.2 for further details regarding the image acquisition parameters.

After RP, prostate specimens were then sectioned and made into slides, which in turn were digitized and manually annotated for PCa by three pathologists as previously described. Distinct foci of annotated cancer were assigned a Gleason score (GS). The annotated slides were then assembled into pseudo-whole mount sections and co-registered to the MRI data using a previously-described deformable registration method,[159] producing ground truth maps of annotated cancer on the MRI data. In summary, the modeling data consisted of mpMRI data and correlated histopathology

TABLE A.1: Summary of the clinical and pathologic characteristics of the patient cohort.

| **Parameter** | **Data** |
|---|---|
| Mean age (yrs) | 64 (range: 51–77) |
| Mean serum PSA at surgery (ng/mL) | 8.3 (range: 1.3–21.8) |
| **Pathologic stage** | |
| T2a | 3 |
| T2b | 4 |
| T2c | 16 |
| T3a | 7 |
| T3b | 4 |
| **Gleason score** | |
| 3+3 | 4 |
| 3+4 | 9 |
| 4+3 | 8 |
| 4+4 | 6 |
| 4+5 | 4 |
| 5+4 | 1 |
| 5+5 | 2 |

TABLE A.2: Acquisition parameters for the mpMRI data. DCE-MRI data were acquired over 5 minutes, resulting in 50 dynamic volumes. TSE = turbo spin echo. SE-EPI = spin echo echo planar imaging. SPGR = spoiled gradient echo.

| **Parameter** | **T2w anatomic** | **T2 mapping** | **DWI** | **DCE-MRI** |
|---|---|---|---|---|
| **Sequence** | TSE | TSE | single-shot SE-EPI | 3D SPGR |
| **Repetition time (ms)** | $\geq$ 6,000 | $\geq$ 6,000 | 3,200 | 4.09 |
| **Echo time (ms)** | 107 | 36, 71, 142 | 88 | 1.44 |
| **Echo train length** | 23 | 23 | 95 | 1 |
| **Acquisition matrix** | 256 $\times$ 230 | 256 $\times 230$ | 128 $\times$ 128 | 192 $\times$ 163 |
| **Field of view (mm)** | 140 | 140 | 180 | 220 |
| **Slice thickness (mm)** | 3 | 3 | 3 | 4 |
| **Nominal voxel size (mm)** | 0.61 $\times$ 0.55 | 0.61 $\times$ 0.55 | 1.41 $\times$ 1.41 | 1.35 $\times$ 1.45 |
| **Readout bandwidth (Hz/pixel)** | 190 | 100 | 1395 | 401 |
| **Parallel imaging $R$ factor** | 1 | 2 | 2 | 2 |
| **Temporal resolution (s)** | — | — | — | 6 |
| **$b$-values (s/mm$^2$)** | — | — | 50, 400, 800 | — |

ground truth of 46 identified axial slices of interest from the 34 patients.

Quantitative T2 maps were calculated from turbo spin echo (TSE) data sets acquired at multiple echo times using methods previously described and validated by Liney et al. and Gibbs et al.[169,263] ADC maps were calculated from DWI data acquired with 3 diffusion-encoding $b$-values (Table 2) using the methods described in section 2.5.3. Pharmacokinetic maps were generated from DCE-MRI data using an extended Tofts-Kermode model with a population-averaged arterial input function;[200] maps of the transfer constant $K^{trans}$, the efflux rate constant $k^{ep}$, and the area under the gadolinium concentration time curve at 90 s (AUGC90) were calculated. In total, five voxel-wise qMRI features were calculated from the mpMRI data.

# Appendix B

# Selection of constants in the definition of the lesion-wise score

The optimization procedure to select the constants ($a_1$, $a_2$, and $a_\omega$) for the definition of $s_\ell$ is based on the following design criteria:

- A difference in the quality of prediction maps should be reflected by a difference in their $s_\ell$s, and it is desirable that this difference is large (on average). Equivalently, if $s_\ell$s are calculated for a large number of pairs of $m_t r$s and $m_p$s, the variance in the observed $s_\ell$s should be large. We refer to this sensitivity of $s_\ell$ to differences in $m_p$s as its discriminatory power, which we seek to maximize.

- Given the uncertainty regarding the "best" values for the constants, it is desirable that $s_\ell$ does not depend too heavily on the exact values for the constants. In other words, if one of the constants changes, the resulting change in $s_\ell$ should be small. We refer to this insensitivity of $s_\ell$ to differences in $a_1$, $a_2$, and $a_\omega$ as its stability, which we seek to maximize.

The optimization procedure was conducted as follows. First, a range of values were chosen for $a_\omega([0.1, 5])$ and the ratio $a_r = a_1/a_2([1, 20])$, with $a_2 = 1.25$ fixed. We consider the ratio $a_r = a_1/a_2$ as it simplifies the optimization procedure, and because the behavior of $g(d)$ primarily depends on the ratio as opposed to the constants individually.

For each $(a_\omega, a_r)$ pair, $s_\ell$s were calculated for each of the 25 $m_{tr}$s with only one $\ell_{tr}$; $m_p$s in these calculations came from the predictive model, and were modified so that only overlapping $\ell_p$s were retained (essentially removing FP $\ell_p$s).

For each $(a_\omega, a_r)$ pair, the discriminatory power of $s_\ell$ was quantified by the standard deviation (SD) of the $s_\ell$s over the 25 cases (bigger SD = higher discriminatory power). The mean of the $s_\ell$s over the 25 cases $(\hat{s_\ell})$ was also calculated, and for each $(a_\omega, a_r)$ pair, the stability of $s_\ell$ was quantified by the magnitude of the gradient of $(\hat{s_\ell})$ (smaller gradient = higher stability). Then to maximize both discriminatory power and stability, the $(a_\omega, a_r)$ pair that maximized their ratio was taken, i.e, the solution to the following optimization problem:

$$\max_{a\omega, a_r} f(a_\omega, a_r) \tag{B.1}$$

where $f$ is the objective function given by:

$$f(a_\omega, a_r) = \frac{\text{SD of } s_\ell(a_\omega, a_r)}{\|\nabla \hat{s_\ell}(a_\omega, a_r)\|} \tag{B.2}$$

Solving this numerically gave $a_\omega = 2.7$ and $a_r = 11.2$ ($a_1 = 14$). A heatmap of $f$ is shown below.

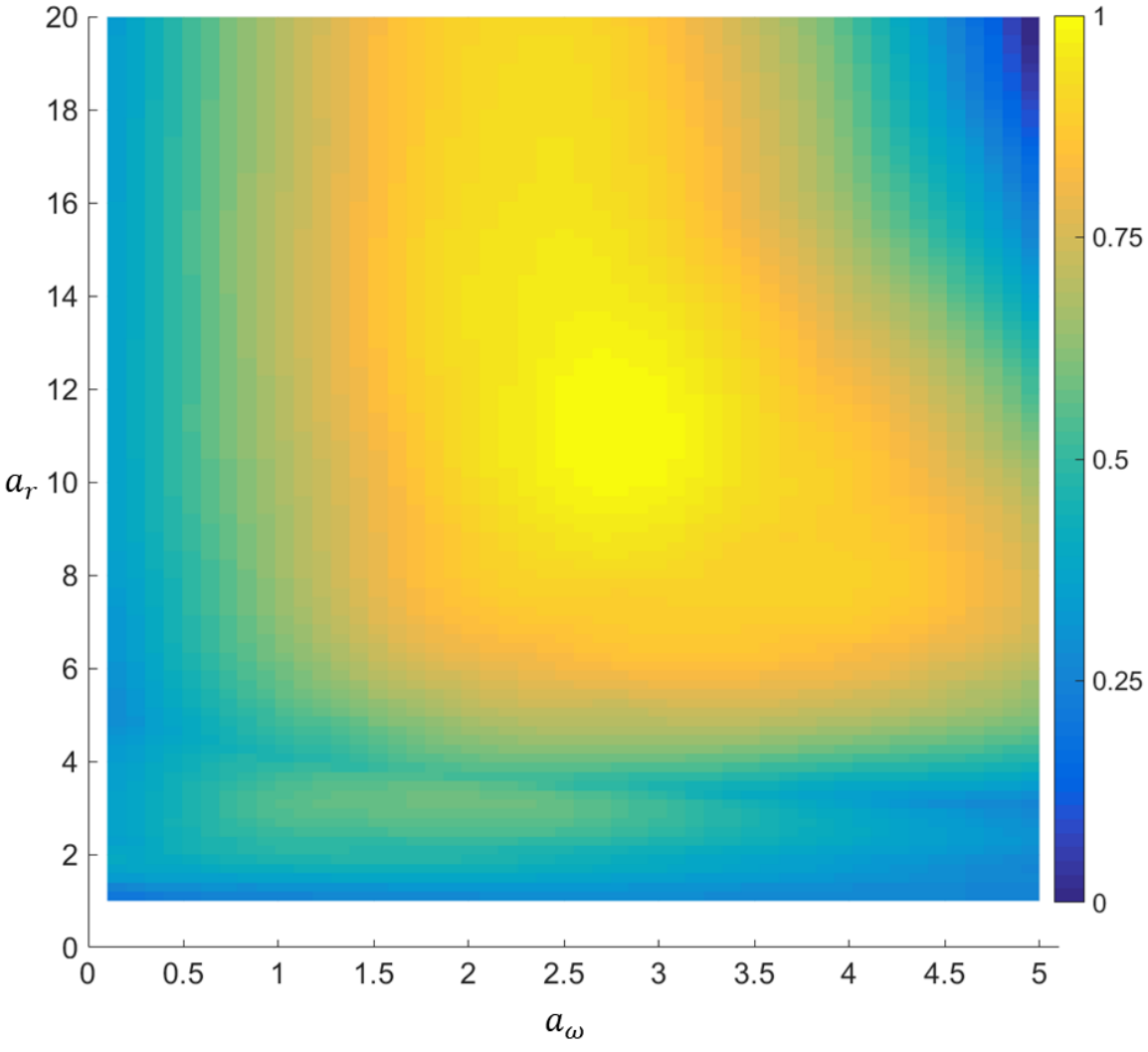



FIGURE B.1: Heatmap of $f(a_\omega, a_r)$ normalized to $[0, 1]$ over the range of values tested in the optimization procedure ($a_\omega \in [0.1, 5], a_r \in [1, 20]$).

## Appendix C

# Algorithm for generating synthetic predictive maps

The following describes the automated algorithm used to generate a synthetic predictive map with a desired voxel-wise sensitivity and specificity for a given ground truth map.
Definitions:

- $m_{tr}$: binary ground truth map.
- $m_p$: binary prediction map.
- $m_{p,\mathrm{TP}}$: an $m_p$ such that all positively-labeled voxels are TPs.
- $m_{p,\mathrm{FP}}$: an $m_p$ such that all positively-labeled voxels are FPs.
- $\mathrm{TP}_t$: target number of TP voxels in $m_p$ to achieve the desired voxel-wise sensitivity.
- $\mathrm{FP}_t$: target number of FP voxels in $m_p$ to achieve the desired voxel-wise specificity.
- $\mathrm{TP}_a$: actual number of TP voxels in $m_p$.
- $\mathrm{FP}_a$: actual number of FP voxels in $m_p$.
- $\mathrm{TP}_{ap}$: actual number of TP voxels in $m_{p,\mathrm{TP}}$.
- $\mathrm{FP}_{ap}$: actual number of FP voxels in $m_{p,\mathrm{FP}}$.

Algorithm:

1. $\text{TP}_t$ and $\text{FP}_t$ were calculated from the given $m_{tr}$ and the desired voxel-wise sensitivity and specificity.

2. $m_{p,\text{TP}}$ was initialized and randomly seeded with $\text{TP}_t$ positively-labeled TP voxels. Similarly, $m_{p,\text{FP}}$ was initialized and randomly seeded with $\text{FP}_t$ positively-labeled FP voxels.

3. Binary dilation was performed on $m_{p,\text{TP}}$ and $m_{p,\text{FP}}$ with random 3×3 structuring elements, after which $\text{TP}_{ap}$ and $\text{FP}_{ap}$ were determined.

4. While $\frac{\text{TP}_{ap}-\text{TP}_t}{\text{TP}_t} > 0.1$ or $\frac{\text{FP}_{ap}-\text{FP}_t}{\text{FP}_t} > 0.1$:

    - TP voxels were randomly added to or removed from $m_{p,\text{TP}}$ until $\text{TP}_{ap} = \text{TP}_t$. Similarly, FP voxels were randomly added to or removed from $m_{p,\text{FP}}$ until $\text{FP}_{ap} = \text{FP}_t$.
    - Morphological opening or closing (randomly chosen) was applied to $m_{p,\text{TP}}$ and $m_{p,\text{FP}}$ with random 3×3 structuring elements.

5. $m_p$ was defined as the union of $m_{p,\text{TP}}$ and $m_{p,\text{FP}}$.

6. Median filter (2×2 kernel) was applied to $m_p$, after which $\text{TP}_a$ and $\text{FP}_a$ were determined.

7. A random integer $n_r$ was chosen with $5 \leq n_r \leq 10$.

8. While $\frac{\text{TP}_a-\text{TP}_t}{\text{TP}_t} > 0.01$ or $\frac{\text{FP}_a-\text{FP}_t}{\text{FP}_t} > 0.01$:

    - TP and FP voxels were randomly added to or removed from $m_p$ until $\text{TP}_a = \text{TP}_t$ and $\text{FP}_a = \text{FP}_t$.
    - The number of 8-connected regions in $m_p$ (denoted by $n_\ell$) was determined.
        - If $n_\ell > n_r$:

    - Morphological closing was applied to $m_p$ with random 3×3 structuring elements.
    - Holes (isolated negatively-labeled voxels) within a region were filled, with the effect of reducing $n_\ell$.
  - Otherwise, morphological opening or closing (randomly chosen) was applied to $m_p$ with random 3×3 structuring elements.