UNIVERSITY OF CALIFORNIA

Los Angeles

Enabling the Use of Clinically Generated Datasets to Improve Diagnostic Methods in Multiparametric MRI of the Prostate

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioengineering

by

Karthik Venkataraman Sarma

2021

ABSTRACT OF THE DISSERTATION

Enabling the Use of Clinically Generated Datasets
to Improve Diagnostic Methods in Multiparametric MRI
of the Prostate

by

Karthik Venkataraman Sarma
Doctor of Philosophy in Bioengineering
University of California, Los Angeles, 2021
Professor Corey W. Arnold, Chair

In this work, we aimed to develop methods and approaches to enable the use of unannotated or weakly annotated clinically generated datasets in clinical data science and deep learning, in the clinical context of prostate cancer and multiparametric MRI of the prostate. Specifically, we demonstrate: 1) The development of an optimized regional targeted biopsy strategy that could reduce the number of biopsies that need to retrieved in a targeted biopsy procedure, by creating a combined MRI, ultrasound, and histopathological evaluation dataset from the clinical record, 2) the creation of a state-of-the-art prostate organ segmentation model using unrefined clinically-generated annotations as well as an evaluation of the utility of those annotations to improve model training on small strongly annotated datasets, 3) the training of a high performance segmentation model on private data originating from three different healthcare institutions using the federated learning approach, without requiring any data to be transferred across institutional boundaries, and 4) the creation of patient-level predictive models for prostate cancer risk stratification from multiparametric MRI of the prostate, and

an evaluation of the relative contribution of pretrained voxel-level feature extractors using unannotated, weakly annotated, and strongly annotated data with the finding that even an unannotated data-based pretrained model is effective. The contributions of this dissertation demonstrate the potential uses of unannotated and weakly annotated clinically generated data in clinical data science and machine learning model development for healthcare, and enable the development of clinical tools for the prostate cancer clinical workflow.

The dissertation of Karthik Venkataraman Sarma is approved.

Holden H. Wu

William F. Speier

Ryan J. Ribeira

William Hsu

Denise R. Aberle

Corey W. Arnold, Committee Chair

University of California, Los Angeles

2021

Dedicated to my family and friends

who have supported me through good times and bad

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to give my appreciation to my PhD advisor and committee chair, Dr. Corey W. Arnold, for his support and guidance throughout my graduate studies. I would also like to thank Dr. William Speier for his continuous support and advice during my research. I am also very grateful for the assistance of my entire committee, Dr. Denise Aberle, Dr. William Hsu, Dr. Holden Wu, and Dr. Ryan Ribeira, for their insight and feedback on my thesis work, as well as all of the faculty of the UCLA Medical Imaging Informatics group, including Dr. Alex Bui and Dr. Ricky Taira.

I would also like to thank my fellow graduate students, whose friendship, support, and insight have been with me throughout my PhD, including Jennifer Polson, Alex Raman, Jiayun Li, Wenyuan Li, Panayiotis Petousis, Nicholas Matiasz, Nova Smedley, Edgar Rios-Piedra, Harry Zhang, Tianran Zhang, Mary Zide, and many more.

I would like to acknowledge my collaborators at UCLA and other institutions who have helped make my research possible, including Nikhil Dhinagar, Leonard Marks, Adam Kinnaird, Steve Raman, Alan Priester, Ely Felker, Anthony Sisk, Stephanie Harmon, Thomas Sanford, Sherif Mehralivand, Baris Turkbey, Holger Roth, Mona Flores, Rushikesh Kulkarni, Dieter Enzmann.

I would like to thank the UCLA-Caltech Medical Scientist Training Program and the NIH for providing research support throughout my work, which was supported by grants NIH NCI F30CA210329, NIGMS GM08042, NCI R21CA220352, NCI P50CA092131, NCI R01CA195505, NCI R01CA158627, and an NVIDIA Corporation Academic Hardware Grant.

Finally, I would like to thank my family and friends who have supported me throughout my life.

VITA

2007–2011	B.S. Computer Science (Hons), California Institute of Technology.
2011	Bhansali Prize in Computer Science, California Institute of Technology.
2011–	MD-PhD Student, UCLA-Caltech Medical Scientist Training Program.
2013–	Member, Subcommittee on Health Information Technology, California Medical Association.
2014	Member, Board of Trustees, California Medical Association
2015–2021	Graduate Student, Computation Diagnostics Lab, Department of Radiological Sciences, David Geffen School of Medicine at UCLA.
2016	F30 Fellowship Awardee, National Cancer Institute, National Institutes of Health.
2016	Student Research Seed Grant Recipient, AMA Foundation.
2016	Trainee Research Award Recipient, RSNA.
2016	M.S. Bioengineering, University of California Los Angeles.
2017–2019	Member, Board of Trustees, American Medical Association.
2020	Trainee Research Award Recipient, RSNA.

SELECTED PUBLICATIONS

Sarma KV, Raman AG, Dhinagar N, Priester AM, Harmon S, Sanford T, Mehralivand S, Turkbey B, Marks LS, Raman SS, Speier W, Arnold CW. Development and Generalizability of a Deep Learning-based Automated Prostate Segmentation Model. PLOS ONE. 2021.

Raman AG*, **Sarma KV***, Raman SS, Priester AM, Riskin-Jones H, Dhinagar N, Speier W, Mirak SA, Felker E, Lu D, Kinnaird A, Reiter RE, Marks LS, Arnold CW. "Optimizing Retrieved Biopsy Count for the Detection of Clinically Significant Prostate Cancer." J Urology. 2021.

Sarma KV, Harmon S, Sanford T, Roth H, Xu Z, Tetreault J, Xu D, Flores MG, Raman AG, Kulkarni R, Wood BJ, Choyke PL, Priester AM, Marks LS, Raman SS, Enzmann D, Turkbey B,

* denotes equal contribution.

Speier W, Arnold CW. "Federated Learning Improves Site Performance In Multi-Center Privacy-Preserving Deep Learning." JAMIA. 2021.

Dhinagar NJ, Speier W, **Sarma KV**, Raman AG, Kinnaird A, Raman SS, Marks LS, Arnold CW. "Semi-Automated PIRADS Scoring via mpMRI Analysis." Journal of Medical Imaging. 2020;7(6).

Sarma KV, Harmon SA, Sanford TH, Roth HR, Flores MG, Kulkarni R, et al. Data-Distributed Deep Learning using Federated Learning: A Case Study. In: Proceedings of the 2020 Annual Meeting of the Radiological Society of North America; Chicago, Illinois. *Awarded RSNA Trainee Research Prize.*

Ho KC, Scalzo F, **Sarma KV**, Speier W, El-Saden S, Arnold C. Predicting ischemic stroke tissue fate using a deep convolutional neural network on source magnetic resonance perfusion images. J Med Imaging. 2019.

Sarma KV, Spiegel BMR, Reid MW, Chen S, Merchant RM, Seltzer E, Arnold CW. "Estimating the Health-Related Quality of Life of Twitter Users Using Semantic Processing." In: Proceedings of the MEDINFO 2019 Meeting; Lyon, France.

Shi W, **Sarma KV**, Raman AG, Priester AM, Natarajan S, Speier W, Raman SS, Marks LS, Arnold CW. "Prediction of Clinically Significant Prostate Cancer in MR/Ultrasound guided Fusion Biopsy using Multiparametric MRI." Poster presented at: Medical Imaging Meets NeurIPS; 2018 Dec 8; Montreal, Canada.

Li W, Li J, **Sarma KV**, Ho KC, Shen S, Knudsen BS, Gertych A, Arnold CW. "Path R-CNN for Prostate Cancer Diagnosis and Gleason Grading of Histological Images." IEEE Transactions on Medical Imaging. 2018;8(1):14429.

Li J, Speier W, Ho KC, **Sarma KV**, Gertych A, Knudsen BS, et al. An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. Comput Med Imaging Graph. 2018.

Li J, **Sarma KV**, Ho KC, Gertych A, Knudsen BS, Arnold CW. "A Multi-scale U-Net for Semantic Segmentation of Histological Images from Radical Prostatectomies." AMIA Annual Symposium. Washington, DC; 2017.

Ho KC, Scalzo F, **Sarma KV**, El-Saden S, Arnold CW. "A Temporal Deep Learning Approach for MR Perfusion Parameter Estimation in Stroke." ICPR. Cancun, Mexico; 2016.

Sarma KV, Zhong X, Ho KC, Margolis DJA, Raman S, Scalzo F, Sung KH, Tan N, Arnold CW. An Investigational Patch-based Convolutional Neural Network Model for the Detection of Clinically Significant Prostate Cancer using Multiparametric MRI. RSNA Annual Meeting. Chicago, IL; 2016. *Awarded RSNA Trainee Research Prize.*

CHAPTER 1

Introduction

1.1 Motivations

The field of artificial intelligence (AI) in medicine has garnered explosive interest over the course of the last decade, driven in part by rapid advances in fields like deep learning (DL) as well as the deployment of AI-enhanced technologies across a wide variety of functional domains across society, from finance, to social media, to the automotive industry and beyond [ED19]. In this context of rapidly advancing utilization of AI across society, significant attention has been paid to the potential of AI-enhanced technologies to "transform" healthcare, with some media reporting that the deployment of AI may herald the end of entire medical specialties, such as radiology [CE16, Muk17]. A number of major results over the last five years in the fields of ophthalmology [GPC16, QCB17, BCK20], dermatology [YCL17, Har18, EKN17], pathology [BPB20, EVJ17, COS18], and radiology [CGT18, MSG20] have contributed to the growing interest in how AI may change the future of medicine.

It is beyond the scope of this dissertation to predict what the long term future may hold for the practice of medicine. However, it is notable that despite the fanfare and predictions of widespread change, the practical impact of AI on the day-to-day practice of medicine remains minimal today. This stands in stark contrast to the advent of computerization and the electronic medical record, which over the course of the last two decades have changed almost every imaginable clinical workflow and significantly altered the day-to-day lives of every healthcare provider in the United States.

Medicine is not without successfully deployed AI-based systems; over 200 AI-based medical devices have been approved by the FDA, with the pace of new device applications increasing significantly year over year [MDV21]. Over half of these devices are targeted at radiology, such as the Arterys Cardio AI system for cardiac landmark detection and segmentation on cardiac MRI [RMG20] and the iCAD ProFound AI for breast density measurement and malignancy detection on tomosynthesis [CTP19]. These devices largely focus on clinical

workflow enhancements for well-circumscribed problems, with the aim of saving clinicians time and improving the reliability and consistency of interpretation. Though these are laudable goals with an important impact on cost and quality, they are far from the life-changing, specialty-destroying specter that some have heralded.

There are many challenges that have prevented AI from having the same impact in healthcare as it has in some other fields [KKS19]. The healthcare industry is naturally cautious, and medical professionals demand a high standard of evidence before the introduction of new clinical tools and practices. Computerization is also relatively new to the industry, with fewer than 50% of medical practices having adopted any electronic health record system as recently as 2010 [Hea19]. There is no doubt that these general industry factors have contributed to a slowdown in the pace of development and acceptance of AI-based technologies.

However, several unique factors regarding medical data also present significant challenges to the development of medical AI. Despite advances in the use of evidence-based medicine over the last three decades, medical practice is still highly variable, with different institutions and providers following different approaches and protocols. Additionally, differences in enrolled patient populations at different care centers can lead to highly heterogeneous disease patterns, diagnoses, and prognoses. This complicates efforts to make use of AI-based models trained at one institution at a different institution, a barrier that has affects even well-funded, large-scale projects like IBM Watson for Oncology [SSJ21, CCK19].

Additionally, the full scope of information used to make a decision is not always available for machine learning use. For example, a radiologist may evaluate an entire MRI volume for cancer suspicion, but may only annotate a few slices, or even only annotate the regions of highest suspicion during routine clinical practice, as there is not a clinical need for a comprehensive annotation. This means that efforts to train machine learning models may require time-consuming and expensive re-annotation of clinical data to produce "gold standard" annotations, making access to large quantities of data challenging at best. For example, one of the largest publicly available medical imaging datasets is the RSNA 2019 Brain CT Hemorrhage Challenge dataset, consisting of 25,312 CT studies (874,345 slices) [FPS20], and most challenge datasets in healthcare are much smaller. In comparison, the ImageNet challenge dataset has approximately 10,000,000 images [RDS15]. Because machine learning (and especially deep learning) models require large quantities of data to train, with more data being required for "harder" problems, this presents a significant challenge.

One potential solution to this problem is to collate data from multiple institutions, as was done in the RSNA challenge [FPS20]. This would enable the creation of larger datasets and

potentially allow the training of models that generalize more effectively across institutions. However, medical data is highly controlled for regulatory and ethical purposes, as the protection of patient privacy is fundamental obligation of healthcare providers. This makes the development of multi-institutional datasets challenging, as data must be carefully evaluated and de-identified before it can be shared across institutional boundaries. In addition, data sharing can be politically complex because of the inherent value of medical data; because large private corporations are now willing to pay for access to data, the free sharing of data for research purposes can be seen as a waste of potential resources as once shared, data cannot be reclaimed.

Continued advancement in the field of medical artificial intelligence in order to harness the full potential of the technology to improve clinical workflows, reduce costs, and improve care quality will require overcoming some of these data challenges in order to enable the development of the best possible models. It is this goal that has motivated the work presented in this dissertation.

1.2 Contributions

In this work, we aimed to develop methods and approaches to enable the use of unannotated or weakly annotated clinically generated datasets in clinical data science and deep learning, in order to address the barriers described in Section 1.1. Because clinical data science cannot be separated from the context of a clinical domain, we chose the domain of prostate cancer and multiparametric MRI of the prostate, and aimed to make contributions that were both helpful to the development of prostate-specific clinical tools and shed light on the broader problem of how to unlock the potential of clinical data. The main contributions of this dissertation can be summarized in the following specific aims:

Aim 1 To combine MRI and ultrasound imaging data with histopathological diagnoses from the clinical record in order to develop an optimized strategy for biopsy sampling that can reduce the number of biopsies that need to be obtained in a single procedure while maintaining the same level of diagnostic performance.

Aim 2 To develop a state-of-the-art prostate organ segmentation model using unrefined clinically-generated annotations, to evaluate the utility of such annotations to enable better training of models using small gold standard datasets, to evaluate the impact of the quantity of data used on the utility of the dataset, and to

evaluate the relative benefit of the use of a domain-relevant weakly annotated dataset over a generic pretraining dataset.

Aim 3 To train a high performing segmentation model using private data from multiple institutions and the federated learning approach which demonstrates generalizability both across those institution and to external data, without requiring any of the private data to be transferred across institutional boundaries.

Aim 4 To create patient-level predictive models for prostate cancer risk stratification from multiparametric MRI of the prostate using pretrained voxel-level feature extractors, and to evaluate the relative contribution of using unannotated, weakly annotated, or strongly annotated data on the final performance of that model.

1.3 Outline

This dissertation is organized as follows:

Chapter 1 is the chapter you are currently reading, and provides an introduction and outline for the dissertation.

Chapter 2 provides a background on prostate cancer, risk stratification, multiparametric MRI, prostate biopsy, deep learning, and the literature in the field of machine learning for prostate cancer.

Chapter 3 provides a deep dive into MRI-ultrasound fusion targeted biopsy, and our findings from a detailed analysis of clinical data at our institution that the use of a regional targeted biopsy strategy could enable clinicians to achieve similar diagnostic performance while requiring the retrieval of fewer cores.

Chapter 4 details our efforts to build a high performance generalizable prostate segmentation model using "noisy" clinical annotations, rather than gold-standard annotations, as well as the contribution of dataset size to the performance and generalizability of the model, finding that "noisy" annotations can be used effectively to develop state-of-the-art models.

Chapter 5 reports on our efforts to train a high-performance prostate segmentation model using data from multiple institutions, without requiring that data

to be collocated in a single location, thus overcoming the regulatory and privacy challenges of multi-institutional dataset generation via Federated Learning.

Chapter 6 describes our work to develop patient-level predictive models for prostate cancer risk stratification using pretrained voxel-level feature extractors with and without radiologist-generated cancer and/or prostate annotations, finding that unannotated data can still be used to develop equally effective patient-level models.

Chapter 7 summarizes the findings and contributions of this dissertation, and potential future directions of research motivated by this work.

CHAPTER 2

Background

2.1 Prostate Cancer

Prostate cancer is the second leading cause of cancer death in American men, accounting for 26% of new cancer diagnoses and 9% of cancer deaths in men [SMJ15]. The discrepancy between these rates creates a need for risk stratification to avoid subjecting patients with clinically indolent cancers to unnecessary interventions, which can be the cause of significant morbidity and cost. Several methodologies have been developed to perform risk stratification of diagnosed prostate cancer, such as the Cancer of the Prostate Risk Assessment (CAPRA) score, which integrates information from risk factors like age and prostate-specific antigen (PSA) with biopsy results to produce an overall estimator of cancer severity and risk [SMJ15, CPE05, SSE06]. These methods have been used successfully to predict outcomes like recurrence after prostatectomy and mortality [DWM98, LBI10, HSB11, BLB15, MSM14, SKK14, SYT15, CDC15, MGF13, BIT12, IHN11]. Unfortunately, the requirement of biopsy information for these methods (which were developed for presurgical and postsurgical use) precludes their utility as noninvasive tools for screening.

The appropriate methodology for prostate cancer screening has been an area of considerable debate. The primary components currently involved in screening are the digital rectal exam (DRE) and serum biomarkers, such as PSA, PSA density, PCA3 [HTP08], PHI [CPS11], and 4Kscore [VCR10]. Over the course of the last decade, recommendations on when to deploy these screening methods (i.e., when to routinely screen with serum PSA tests) have evolved, with current recommendations from the USPSTF [US18] and AUA [CAB13] to make individualized decisions regarding serum PSA screening between the ages of 55 and 69 years, and to avoid screening at 70 years and older. Unfortunately, despite advances in screening approach as well as new serum tests, overdiagnosis still remains a major problem due to limited specificity [SYT15, VCA08, CC13, PPP15, KAS15, LUV08, LCK].

2.1.1 Prostate Biopsy

Due to the limitations of noninvasive screening for prostate cancer, positive screening results are generally followed up by transrectal biopsy of the prostate to obtain pathological confirmation, but this methodology remains limited. In the transrectal prostate biopsy procedure, a biopsy sampling needle is used to obtain biopsy cores from the prostate. In a standard "template" biopsy procedure, 6-12 cores are obtained from the prostate in a distributed manner, with recent guidelines recommending 10-12 cores ("extended core biopsy") in order to optimize detection while minimizing morbidity.

Historically, this procedure was done with manual guidance (i.e. the urologist used clinical experience to determine how to insert the sampling needle in order to obtain the needed cores); however, the predominant method is now ultrasound-guided biopsy ("TRUS biopsy"), in which the needle is rigidly attached to an ultrasound probe which is used to image the prostate in real-time and provide feedback to the urologist as to the correct positioning of the probe. The procedure is typically performed in the outpatient clinic setting under local anesthesia.

Ultrasound-guided biopsy improves targeting, but has limited utility in discriminating suspicious regions from benign changes in the prostate, leading to a substantial risk of undergrading due to poor biopsy localization [GAA15, LS02, KMW09]. As a result, 90% of patients diagnosed with prostate cancer receive treatment, even though up to 60% of those patients could be candidates for active surveillance [CC13, DES07, JMC15]. Notably, such treatment often results in long-term reductions in functional outcomes [DAB12]. In addition, the biopsy procedure itself can rarely result in adverse events such as hospital admissions due to infection [RKF13, SKM10, BGS04, NSL13].

2.1.2 Gleason System

Once biopsy cores are retrieved, they are sent for histopathological analysis in order to receive tissue diagnosis. This assessment is the most critical in determining the final diagnosis and treatment approach, and is performed by a clinical pathologist who examines the specimen microscopically in order to assess for metaplastic or neoplastic patterns.

The Gleason grading system is used to provide a standardized assessment of prostate pathology [GMA74]. This system is designed to categorize the architectural features of prostate cells based on growth pattern and degree of differentiation. Patterns are graded from 1-5, and for each sample a "primary" and "secondary" score are provided (i.e. "3+4"), with

the two scores often being added together to create a composite score. Generally, composite scores range from 6-10 as scores under 6 are usually instead reported as benign [Eps10]. Occasionally, a tertiary score may also be provided if a small area of high grade tumor is present; this tertiary score is reported alongside the primary, secondary, and composite scores.

Over the last decade, the ISUP grade group system has been adopted to further categorize Gleason scores based on risk stratification [EEA15]. This system classifies Gleason scores into five grade groups (1-5) with increasing risk of cancer mortality with increasing grade group [BBF16, HCF17]. The grade groups are as follows:

- Grade group 1: Gleason score ≤ 6
- Grade group 2: Gleason score = 3+4
- Grade group 3: Gleason score = 4+3
- Grade group 4: Gleason score = 8
- Grade group 5: Gleason score ≥ 9

The Gleason scoring system and ISUP grade group system are generally the more critical component of final diagnosis, and are part of the major clinical guidelines for prostate cancer staging and treatment [Nat15]. The Gleason system can, however, suffer from poor inter-rater reliability [BEM08]. Additionally, the Gleason grade assigned is only as good as the sample provided to the pathologist, and as such biopsy results must be interpreted with consideration of the risk of sampling error. A discussion of score "upgrading" on whole-prostate histopathology vs. prostate biopsy core histopathology is presented in Chapter 3.

2.1.3 Prostate Multiparametric MRI

The use of magnetic resonance imaging in the diagnosis and management of prostate cancer has steadily garnered interest due to the limited capability of ultrasound to assess the prostate. When imaging the prostate, generally multiple different magnetic resonance imaging (MRI) pulse sequences spanning both anatomic and functional parameters are acquired, including T2-weighted imaging (T2W), diffusion-weighted imaging (DWI), dynamic contrast-enhanced imaging (DCE), and magnetic resonance spectroscopic imaging (MRSI). This methodology, multiparametric MRI (mp-MRI), has been studied for targeted biopsy,

active surveillance, and screening [MYN13, WSG15, CLD15, EPB15, TC12, AAA14, SAM14, TLP15, AKA14, SNM13].

In the area of targeted biopsy, several major clinical trials have investigated the use of real-time ultrasound-MRI image fusion to allow for more precise acquisition of biopsy cores from areas of interest. In this approach, mpMRI of the prostate is obtained, and a trained abdominal radiologist evaluates the images and highlights one or more regions of interest ("ROIs") for suspicion of cancer. The prostate itself is then contoured on MRI, and the contour and ROIs are then transmitted to the TRUS biopsy workstation. During the biopsy procedure, the workstation performs a real-time image fusion between the previous MRI and the current ultrasound view in order to enable the urologist to visualize the ROIs delineated on MRI within ultrasound space. This allows the urologist to guide the biopsy sampling needle towards the ROIs and obtain targeted samples (Figure 2.1). Generally, a "combined" biopsy procedure is performed, in which the standard cores are obtained from the systematic sampling locations, as well as additional cores from each ROI. A more in-depth discussion about the choice of sampling targets for MRI-ultrasound targeted biopsy as well as the relative benefits of targeted biopsy vs. systematic biopsy is presented in Chapter 3.

While this approach can mitigate the risk of missing prostate cancers due to poor biopsy sampling, it still relies on accurate identification of regions of interest via expert interpretation. In order to perform the procedure, mpMRI of the prostate must be available, as well as a trained abdominal radiologist, and a biopsy workstation capable of performing MRI-ultrasound fusion (alternatively, some targeted biopsy procedures are doing entirely within the bore of an MR scanner, thus eliminating the need for ultrasound and potentially improving targeting). These requirements, as well as the relative novelty of the approach and the increased cost, have limited access to the procedure.

Another major area of interest in prostate mp-MRI is the use of imaging to risk stratify patients, avoiding biopsy in low-risk patients by identifying only clinically significant tumors for further workup [MYN13, WSG15, EPB15, TC12, TLP15, SNM13, CDM12, ME15, PMR15]. In this paradigm, patients with a positive initial screening test (DRE or serum biomarker) are referred for mp-MRI. The resulting images would be analyzed and then depending on the computed risk, the patient may be referred for systematic and/or targeted biopsy, or for active surveillance with serial imaging. For this paradigm to be safe and effective, mp-MRI must achieve a sufficient negative predictive value for clinically significant prostate cancer. Based on research over the last decade, the NPV of mpMRI is between 83 and 95% for clinically significant prostate cancer ("csPCA") [TMA13, MVS17], which has

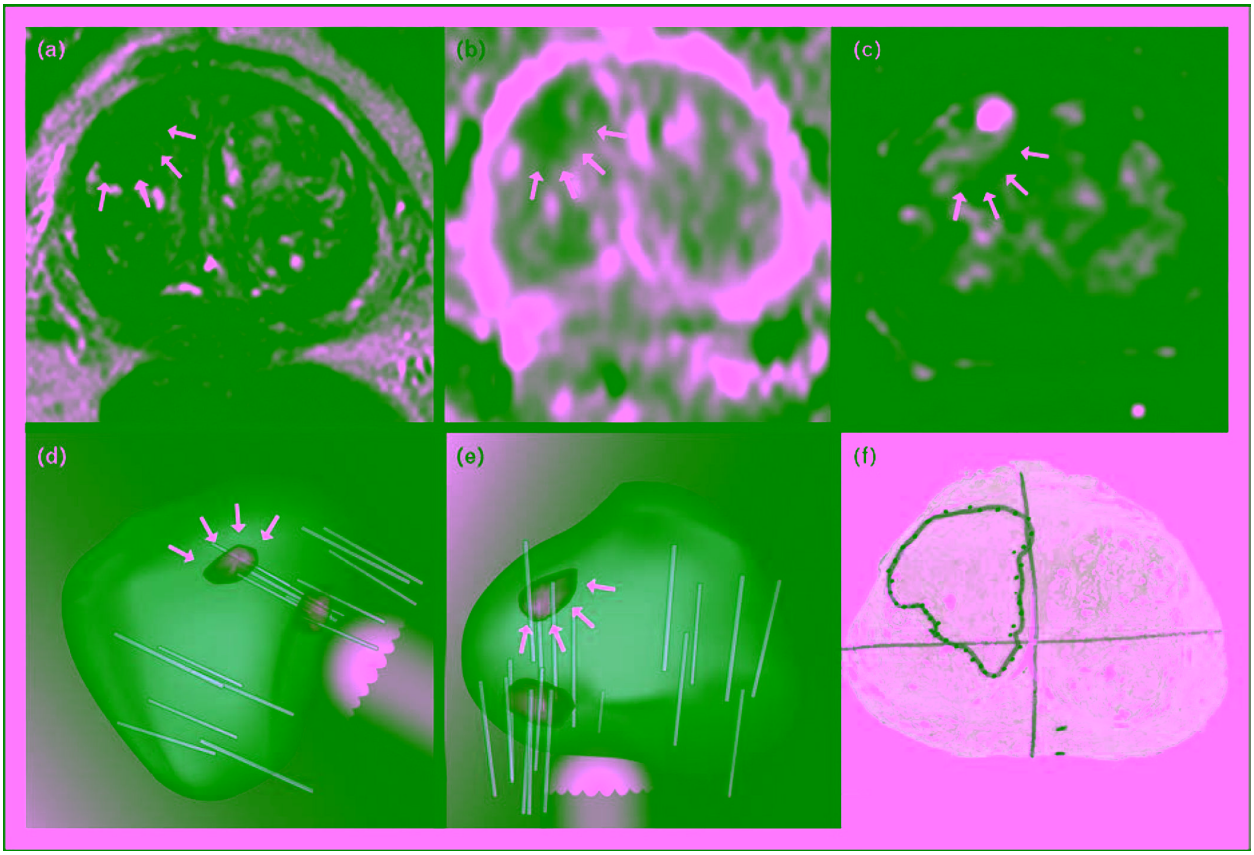


Figure 2.1: Example of Targeted Prostate Biopsy. Reproduced from Marks, et al. [MYN13]. a-c) T2, ADC, and DCE images from mpMRI scan obtained from patient in advance of biopsy, with radiologist-defined region of interest designated by arrows. d-e) Schematic diagram of biopsy device with ultrasound probe being used to collect cores from the region of interest. f) Whole-mount histopathology specimen from radical prostatectomy with lesion associated with MR lesion of interest outlined by pathologist.

enabled the development of clinical guidelines for active surveillance.

One major limitation of mp-MRI identified in the literature is the poor to moderate inter-reader reliability for identifying potentially clinically significant cancers on mp-MRI outside of a small set of major academic medical centers with experience in prostate MRI [RGC16, GSL19]. In addition, reliability amongst experts may be poor specifically for DCE-MRI. A potential driver of these limitations is the difficulty of interpreting the raw images, which requires "cognitive" fusion of several modalities and temporal data in the context of relatively few voxels for analysis.

2.1.4 PI-RADS

In order to provide a standardized assessment of prostate mpMRI, the Prostate Imaging Reporting and Data System (PI-RADS) was developed [TRH19, BWV16, DAA11, BRC12]. This system is designed to categorize focal prostate nodules seen on MRI based on the level of suspicion for cancer, with standardized criteria for assigning a score between 1 and 5 as suspicion of cancer increases. The system is periodically updated to improve the standardization and utility of the categorization, and in addition to scoring criteria also provides standard image acquisition parameters, as well as terminology for referring to prostate findings. The scores are defined as follows:

- PI-RADS 1: Clinically significant cancer is highly unlikely to be present.
- PI-RADS 2: Clinically significant cancer is unlikely to be present.
- PI-RADS 3: The presence of clinically significant cancer is equivocal.
- PI-RADS 4: Clinically significant cancer is likely to be present.
- PI-RADS 5: Clinically significant cancer is highly likely to be present.

The scores integrate information from multiple imaging parameters, including T2-weighted, diffusion-weighted (using an ADC map based on high b-value DWI), and dynamic contrast-enhanced (DCE, also known as "perfusion-weighted"). Different criteria are used for lesions in the peripheral zone and the transition zone; in the peripheral zone the diffusion-weighted images are the primary factor, and in the transition zone, the T2-weighted images are the primary factor. In both zones, DCE images are used only to differentiate between images with intermediate scores based on the primary factor, with focal and early enhancement being the factor contributing to higher score.

2.1.4.1 UCLA Score

Before the release of the PI-RADS v2 scoring system, a quantitative PI-RADS v1 based scoring system was developed and used at UCLA. This system used the PI-RADS v1 criteria to assign individual parameter scores, and then used the formula (T2W + 2*ADC + DCE - 0.25*TZ)/4, where T2W, ADC, and DCE were the single-parameter scores, and TZ was 1 if the lesion was in TZ and 0 if not. After the release of PI-RADS v2, a retrospective assessment was performed to compare the performance of the UCLA score against PI-RADS v2 and found similar performance for prostate cancer detection, grading and staging [MBH19].

2.2 Deep Learning

2.2.1 Convolutional Neural Networks

Deep convolutional neural networks (CNNs) have been shown to outperform other learning systems models (such as shallow perceptrons, support vector machines, regression, and k-means clustering) in large-scale image classification tasks [LTT15, SZ14, KSH12, DCM12, MKS15, RDS15, CBG14, JTL15, HZR15, SZ15, EKN17, CAL16, IPK18].

CNNs are extensions of traditional deep neural networks comprised of a hierarchy of functional layers classified into different possible types, including convolutional, non-linear activation, pooling, and fully-connected.

In a convolutional layer, each neuron in the layer is only connected to a subset of its input (such as a 3x3 region, or "patch," out of a 21x21 input matrix) at a time. This is the "receptive field" of that particular neuron. The learned weights for these neurons are tied together such that there is a fixed number of sets of weights (the "filters") that are applied to every patch in the input. The result of this operation is an output set consisting of a map of the outputs of each filter upon each patch, referred to as the layers "feature map." Each convolutional layer will learn filters that represent a high-level feature over the output of the previous layer. This approach allows for the filters to be shared across all of the possible locations (i.e. receptive fields) for the entire image, which reduces the number of parameters vs. a fully connected network and enables the creation of a shared, hierarchical representation of learned information. This approach is optimized for input data with meaningful location-independent "spatial"-type relationships within each sample, such as natural images (i.e. an "edge" is a spatial-type relationship which is independent of location).

A non-linear layer takes the output of a layer and applies a non-linear "activation function" in order to introduce non-linearity to the network, allowing it to capture more complex functions [HZR15, GBB11, HN10]. Commonly used nonlinearities include sigmoid, tanh, and ReLU.

A pooling layer takes patches from input values and groups them together via a mathematical operation, such as "max" or "mean." The result is a smaller output layer that also represents features in a more translation-invariant manner. For example a 2x2 max-pooling layer might take an input of size 4x4, and return a 2x2 matrix that consists of the maximum element of each 2x2 submatrix of the input.

A fully-connected layer is a traditional neural network layer in which every input neuron is connected to every output neuron, allowing for complex dependencies across the input to be learned. These layers are generally used at the end of a network that produces classifications in order to condense all of the learned features into set of outputs, and to enable input of related non-imaging information (i.e. to combine other features such as age or test results with an imaging result).

Regularization layers, such as dropout [HSH14] and batch normalization [IS15] are used to avoid overfitting and thus improve generalizability. These layers generally work through explicitly or implicitly penalizing coefficients in order to incentivize models to learn sparse representations of learned information. Dropout is an implicit regularizer that randomly (with some specified probability) selects nodes for each and removes them. The resulting randomly sparse activations encourage the network to learn a sparse representation. Batch normalization is also an implicit regularizer that normalizes its inputs for every batch in order to reduce internal covariate shift. As a side effect, with small batch sizes a large number of separate normalizations occur per epoch, essentially adding different noise to each batch. This reduces the information content per batch, which incentivizes a sparse model representation.

CNN models are often based on popular architectures which have been demonstrated for image analysis in the literature. These architectures include AlexNet [KSH12], ResNet [HZR16], DeepLab [CPK16], U-Net [RFB15], and many more.

2.2.1.1 U-Net

The U-Net [RFB15] and 3D U-Net [CAL16] are examples of "fully convolutional" CNNs, which have an output shape equal to the input shape. These networks use a encoder-

decoder pattern, in which serial encoders are first used to create an embedding of the input sample using chains of convolutional layers followed by pooling layers and activation layers (the downslope of the "U"). After the encoders have processed the input, the resulting intermediate output has low spatial resolution and high context and information density. This intermediate output is then fed through a series of decoders, which consist of up-convolutional layers followed by convolutional layers. In order to restore access to spatial information, "skip"-connections are used. At each decoder level, the intermediate output of the encoder at the same level is concatenated onto the input from the previous up-convolution before being input into the convolutional layers at that level. This combination of high context information with high localization information can enable higher performance from the network (Figure 4.1).

2.3 Machine Learning in Prostate

2.3.1 Organ Segmentation

Automated prostate segmentation is an active area of research, and substantial published work exists on the development of machine learning models for the purpose. However, these state of the art prostate segmentation algorithms [JXS19, JYF21, WLT19, TLZ20, WWW21, ZWY20] are often trained on small research-quality annotated datasets curated specifically for machine learning. Examples include the 100 patient Prostate MR Image Segmentation (PROMISE12) challenge dataset [LTV14] and the 60 patient NCI-ISBI (National Cancer Institute International Symposium on Biomedical Imaging) Automated Segmentation of Prostate Structures (ASPS13) challenge dataset [Blo15]. Other algorithms have been trained on institutionally developed local datasets that include between 100 and 650 studies [SZH20, CLR19, CZB17, SHS21]. Unfortunately, the development of research-quality prostate boundary annotations is challenging. For example, for the PROMISE12 dataset, segmentations were created by an experienced radiologist, verified by a second experienced radiologist, and then re-annotated by a third nonclinical observer – a complex and expensive process. This is necessitated by the fact that interrater reliability for prostate segmentation is poor to moderate and current commercial software-based tools perform poorly [GOM12, BMF18]. The performance of the deep learning-based models described above hovers around 0.9 Dice coefficient, which may be the upper limit due to the inherent uncertainty in the "true" segmentation. An in-depth discussion about deep learning-based prostate segmentation can be found in Chapter 4.

2.3.2 Lesion Detection and Characterization

2.3.2.1 Detection

Lesion detection models predict the presence of cancer (either any cancer, or clinically significant prostate cancer, generally considered to be grade group $/geq$ 3) from an input MRI volume. Because these models do not require a radiologist's interpretation before use, they have significant potential in increasing access to MR evaluation of the prostate and MR-targeted biopsy, as well as the potential to lower cost and improve consistency of MR-based interpretation. Though no lesion detection computer-aided diagnosis tools have yet transitioned to the clinic, a significant body of research exists in the area.

Many efforts in CAD have focused on voxel-based feature generation using support vector machines, manually engineered features and statistics, and deep learning [LWT13, WBT14, PBV07, VBG08, VBR09, MAB11, SKP11, VBK12, HVH13, LTG17, TLR17, SLT18]. These voxel-based features are then used to produce a volume-level prediction using an aggregation methodology. For example Lay et al. [LTG17] produce a voxel-wise cancer prediction over the prostate volume, and then aggregate these predictions into cells of 3mm x 3mm x 3mm and then measure the 90th percentile cancer scores in each score to determine whether to designate the cell as positive for cancer or not. Based on this methodology, their random forest model achieves a cancer detection AUC of 0.93 using a dataset of 224 cases and T2, ADC, and high b-value volumes. In follow-on work, the team found that the use of deep learning-based edge detection models improved their performance to 0.97 [SLT18, TLR17] on the same dataset.

Alternatively, Xu et al. [XBA19] use a "hit-or-miss" methodology, in which the model is evaluated by determining if any voxels over the 90th percentile cancer scores overlap with voxels designated as cancerous by a radiologist. Using this approach, their deep residual net model achieves a cancer detection AUC of 97% using a dataset of 346 cases and T2, ADC, and high b-value volumes.

2.3.2.2 Characterization

Lesion characterization models predict information about a prostate cancer when input an MRI volume as well as the localization of a lesion within the volume (i.e. a point or region of interest). Thus, they differ from the detection problem in that a radiologist must evaluate the image and highlight suspicion and cannot be used on undifferentiated images. Despite

this limitation, characterization models are of considerable interest, in part because of the limitations inherent in current methods for accurate characterization of lesions from MRI; as described above, the most widely used methodology for manual characterization of the risk level of MRI-visible lesions is the PI-RADS v2.1 system, which can suffer from inconsistency in some settings and has highest performance as a negative predictor [WMA20, MAL20, MSS15, GGG19, BDS20, PBB17].

Significant advancement in lesion characterization was driven by the ProstateX-1 and ProstateX-2 [AHD18] grand challenges, which provided approximately 180 mpMRI scans with a spatial indicator of the locus of cancer suspicion within the lesion and the biopsy-proven Gleason grade group. A wide variety of papers have investigated prediction of lesion character (generally dichotomized as grade group ≥ 3), including deep learning-based methods [SZY18, STK17, KXW15, CHL19, MSG17, LZF17, WLC18, LBK15]. For example, Yang et al. [SZY18] obtained an AUC of 0.94 using a VGGNet-based deep learning model.

2.3.3 Challenges in Interpretation

One challenge in interpreting the literature of ProstateX-based models (and generally for all challenge dataset-based models) is the inherent "meta-overfitting" that occurs with the release of serial models over time. This issue is a generalization of the machine learning data leakage problem. In order to avoid data leakage on an individual experiment level, data is often split into "training," "validation" and "test" sets. The training set is used to train models, the validation set is used to optimize over hyperparameters (i.e. model design), and the test set is used to perform a final evaluation. This ensures that the model is not inadvertently fit to the test set through hyperparameter optimization (i.e. by choosing a model design that happens to work particularly well for that specific test set). During the course of a challenge period, the test set is generally withheld in order to avoid fitting models to the test set and final evaluation is performed at the end of the challenge. However, after the completion of a challenge and the release of the test set, subsequent published models are necessarily published because of their favorable performance on the test set. This may over time lead to upward trending performance statistics that may in part be due to the gradual fitting of model design to the test set generated by the publication solely of models that perform well on that set. This "meta-overfitting" makes comparison of "best-in-class" challenge results to results from separate datasets difficult.

CHAPTER 3

Optimizing Spatial Biopsy Sampling for the Detection of Prostate Cancer

3.1 Overview

The current gold standard for prostate cancer diagnosis involves a targeted biopsy of suspicious MRI regions of interest combined with a systematic template biopsy; together, they form a combined biopsy procedure consisting of, at our institution, an average of 17 retrieved biopsy cores.

When compared to MRI-targeted biopsy, systematic biopsy has been shown to detect higher rates of clinically insignificant cancer, defined using the International Society of Urological Pathology (ISUP) prostate cancer grading system as grade group 1, and lower rates of clinically significant cancer, defined as grade group ≥ 2 (these same grade group designations are used in this study) [KRB18, EFK19]. Nevertheless, combined biopsy is widely recommended since studies have shown that in 14-16% of patients who underwent both procedures and received a csPCa diagnosis, the csPCa was detected by systematic biopsy alone [HWM19, RPR19].

The combined biopsy approach requires obtaining significantly more biopsy cores than either systematic biopsy or MRI-targeted biopsy alone, increasing the cost, length, and discomfort of the procedure as well as the risk for sepsis, hematospermia, and pelvic and perineal pain [SKM10, BGS04]. In order to reduce these risks, it is prudent to retrieve the minimal number of biopsy cores required to adequately assess the patients current cancer status.

Although a precedent has been set establishing combined biopsy as the most robust prostate biopsy protocol [AWR20], no study to date has rigorously investigated the spatial

---

The work described in this chapter was published as: Raman AG*, Sarma KV*, Raman SS, et al. Optimizing Spatial Biopsy Sampling for the Detection of Prostate Cancer. J Urol. 2021. doi:10.1097/ju.0000000000001832. * indicates equal contribution.

relationship between systematic biopsy cores and MRI targets using the measured locations of obtained cores. As a result, little evidence is available to guide the determination of the optimal total number and location of biopsy cores that should be obtained from a patient; instead, most attention has been focused on determining the appropriate number of cores sampled from each ROI in the targeted biopsy component [TFS20, KRR18, SDT20, LSG19].

Tschirdewahn et al. used a retrospective analysis to examine the use of a targeted saturation biopsy strategy in which biopsies were taken only from the MRI target and adjacent areas [TWB20], an approach suggested in some scenarios by the Prostate Imaging Reporting and Data System (PI-RADS) committee [PWR19]. This analysis found that restricting sampling to targeted locations within ROIs and systematic biopsy locations within adjacent Ginsburg sectors (which segment the prostate into zones from which each systematic biopsy should be sampled [KWC13]) was superior to targeted or systematic biopsy alone. However, the true biopsy retrieval coordinates were not available to enable a complete analysis of the relationship between distance and yield. In addition, without prostatectomy data, upgrading and downgrading rates could not be assessed, making a full sensitivity assessment impossible.

In this study, we propose a biopsy site selection strategy which we refer to as "regional targeted biopsy" (RTB). This strategy optimizes the selection of additional biopsy sites by focusing on regions of the prostate located within the two-centimeter penumbra of a radiologist-designated ROI with a high suspicion index (i.e. PI-RADS score). Prior work that places MRI underestimation of prostate cancer tumors when compared to whole mount at a median of 13.5 mm per tumor, along with clinical intuition from the urologists and radiologists involved in this study helped inform the decision of using a 2-cm margin as the basis for constructing a RTB [PNK17]. A sensitivity analysis of this threshold choice is provided. We hypothesized that this strategy would achieve equivalent detection rates for clinically significant prostate cancer while requiring the retrieval of fewer biopsy cores.

In order to evaluate the potential impact of this strategy, we retrospectively calculated the results of an RTB by discarding cores obtained from combined biopsy that are located outside of the two-centimeter ROI penumbra. This location assessment was enabled using a retrospective sensor fusion approach that provides the three-dimensional localization of each retrieved biopsy core within the prostate. We compared both csPCa detection rates across our entire cohort and grade group upgrading and downgrading rates of a subcohort who underwent radical prostatectomy across four different protocols: systematic biopsy, MRI-targeted biopsy, combined biopsy, and RTB.

3.2 Materials and Methods

3.2.1 Study Inclusion and Exclusion Criteria

We retrospectively collected data from a cohort of patients at our institution who underwent standardized 3 Tesla multiparametric MRI followed by standardized MRI-ultrasound fusion combined (both systematic and targeted) biopsy using spatially localized targets on a single system with specialized fusion software (Artemis and Profuse; Eigen Inc, Grass Valley, California) between 2011 and 2018. Patients were included regardless of how many biopsies they may have had previous to the study period or their Prostate Specific Antigen (PSA) value; however, for patients with repeat biopsies during the study period, only the first biopsy session was included for analysis. To ensure a fair comparison of cancer detection rates, we chose to include only the subset of patients who received at least 10 systematic biopsy cores. This minimum threshold of 10 systematic cores was consistent with recommendations from the European Association of Urology and others [MBB17, Pre03, HWS19]. All MRI lesions were graded by one of three experienced genitourinary radiologists (SR, DL, and EF with 22, 29, and 5 years of domain-specific experience respectively) using a published institutional score for lesions graded between 2011 and 2014 and the PI-RADS version 2 score for lesions graded between 2015 and 2018 [SNM13]. The institutional score is a 1 to 5 Likert score based on quantitative metrics that has been shown to have a similar csPCa detection rate to PI-RADS version 2 [MBH19]. Patients were excluded from analysis if their biopsy procedure was performed under a clinical trial protocol to avoid confounding from protocol differences, and were also excluded from analysis if real-time positional data was corrupt or unavailable for one or more of their biopsy cores.

3.2.2 MRI and Biopsy Protocols

As a part of routine clinical interpretation, MRI target ROI contours were drawn on axial T2-weighted scans by one of three experienced genitourinary radiologists using commercially available annotation software (DynaCad; Invivo-Philips, Gainesville, Florida). To maximize specificity, the clinical annotation protocol required ROI margins to be drawn tightly around suspicious targets. These MRI annotations were then transferred to the MRI-ultrasound fusion device to enable the biopsy procedure. During the procedure, real-time sensor fusion was used to determine the three-dimensional spatial coordinates of the tip and base of each individual biopsy core retrieved, including both targeted and template cores. Patients were anesthetized using a periprostatic nerve block of 20cc 1% xylocaine, and all cores were

retrieved by a single urologist (LSM) with 10 years of fusion biopsy experience.

Computerized targeting guidance was provided for both systematic and targeted cores. For systematic cores, a target marker was designated on the procedural console and the operator retrieved a core from the designated location. For targeted cores, a radiologist-delineated ROI was displayed on the procedural console, and the operator retrieved cores from the ROI. Our combined biopsy protocol uses the "target + standard" approach for all biopsies, wherein targeted cores were taken before systematic cores. All cores that were intended to be targeted at an ROI were designated as targeted cores and all cores that were intended to be systematic were designated as such, regardless of their position relative to the ROI. Cores were taken every 5 mm along the longest axis for irregularly-shaped ROIs, and in a cross-hair pattern for regularly shaped ROIs. The standard minimum number of cores per ROI was 2, though a single patient in our dataset received 1 core for their ROI.

All retrieved biopsy cores were interpreted by a subspecialized group of genitourinary pathologists with 5-15 years of experience in prostate cancer interpretation and assigned Gleason scores and grade groups [EEA15]. For the purposes of this study, clinically significant prostate cancer included any biopsy core assigned a ISUP grade group of 2 or higher.

3.2.3 Biopsy Distance Calculations

For each patient, the three-dimensional spatial coordinates corresponding to each biopsy core were retrieved. We then linearly interpolated 30 points between the tip and base to represent the three-dimensional trajectory of the retrieved core. The distance between a targeted or systematic biopsy core and an ROI was determined by both a distance from the edge of the ROI and the distance from the centroid of the ROI. If multiple ROIs were present, the smallest distance of the core to any ROI was used.

To determine an individual biopsy cores distance from the edge of an ROI, we first used the ray-casting algorithm to determine if the core intersected the ROI [Rot82]. A distance of 0 was assigned to biopsy cores intersecting the ROI, while the shortest three-dimensional distance between the set of points representing the biopsy core and ROI margin was assigned to biopsy cores not intersecting the ROI. In addition to the distance from the edge, we computed the shortest three-dimensional distance between each core and the ROI centroid as an alternative distance metric.

3.2.4 Statistical Methods

To compare the cancer detection rates of systematic biopsy, MRI-targeted biopsy, combined biopsy, and RTB, as well as subsequent whole-mount grade group upgrading and downgrading of each of these methods, the two-tailed, two-proportion z-test was used. All tests were evaluated at a significance level of $p < 0.05$.

3.3 Results

3.3.1 Patient Cohort

The initial study cohort included 1,705 patients. We excluded 239 patients with fewer than 10 systematic biopsy cores, 233 patients who participated in the Prospective Assessment of Image Registration for the Diagnosis of Prostate Cancer) PAIREDCAP clinical trial [EFK19], and 262 patients due to missing biopsy core positional data. The final study cohort included 971 patients who underwent 3 Tesla multiparametric MRI and MRI-ultrasound fusion biopsy between April 2011 and December 2018 (Figure 3.1) with an average age, PSA level, and prostate volume of 64.5 ± 7.4 years, 8.4 ± 7.9 ng/ml, and 49.9 ± 24.2 cm^3, respectively (Table 3.1). The average ROI volume was 0.9 ± 2.2 cm^3, and when the ROI volume was expanded by 2 cm, the average expanded ROI volume was 26.4 ± 9.0 cm^3.

Table 3.1: Clinical and Demographic Information for both Patient Cohorts. Average age, PSA, prostate volume, and number of targeted, systematic, combined biopsy, and simulated regional targeted biopsy cores are presented with their standard deviations. All other values are presented with their corresponding percentage of the cohort listed in parentheses. RT = regional targeted, ROI = region of interest, csPCA = clinically significant prostate cancer.

Attribute	**All patients** ($N = 971$)	**Prostatectomy** ($N = 102$)
Age (years)	64.5 ± 7.4	62.2 ± 6.1
Race		
Caucasian	616 (63.4%)	72 (70.6%)
Asian	54 (5.6%)	6 (5.9%)
African American	37 (3.8%)	4 (3.9%)
Hispanic	22 (2.3%)	3 (2.9%)
Native American	1 (0.1%)	0
Mixed	1 (0.1%)	0
Other	19 (2.0%)	2 (2.0%)

Attribute	**All patients** ($N = 971$)	**Prostatectomy** ($N = 102$)
Unknown	221 (22.8%)	15 (14.7%)
PSA (ng/ml)	8.4 ± 7.9	8.2 ± 6.8
Prostate Volume (cm^3)	60.8 ± 29.1	46.5 ± 18.1
ROI Volume (cm^3)	0.9 ± 2.2	0.7 ± 0.9
RT ROI Volume (cm^3)	26.4 ± 9.0	23.4 ± 7.3
Maximum ROI Score		
3	415 (42.7%)	34 (33.0%)
4	380 (39.1%)	37 (36.2%)
5	176 (18.1%)	31 (30.4%)
Previous Biopsy		
No Previous Biopsy	309 (31.8%)	39 (38.2%)
1 Previous Biopsy	413 (42.5%)	41 (40.2%)
> 1 Previous Biopsy	246 (25.3%)	22 (21.6%)
Unknown	3 (0.3%)	0
Number of		
Targets	1.3 ± 0.6	1.4 ± 0.6
Targeted Cores	5.0 ± 1.9	5.2 ± 1.9
Systematic Cores	11.9 ± 1.1	11.6 ± 0.9
Combined Biopsy Cores	17.0 ± 2.0	16.8 ± 1.9
RT Cores	13.2 ± 1.5	13.8 ± 3.5
csPCa Targeted Cores	1.0 ± 1.7	2.0 ± 1.7
csPCa Systematic Cores	0.5 ± 1.2	1.2 ± 1.4
csPCa Combined Cores	1.6 ± 2.5	3.2 ± 2.6
csPCa RT Cores	1.5 ± 2.5	3.1 ± 2.6

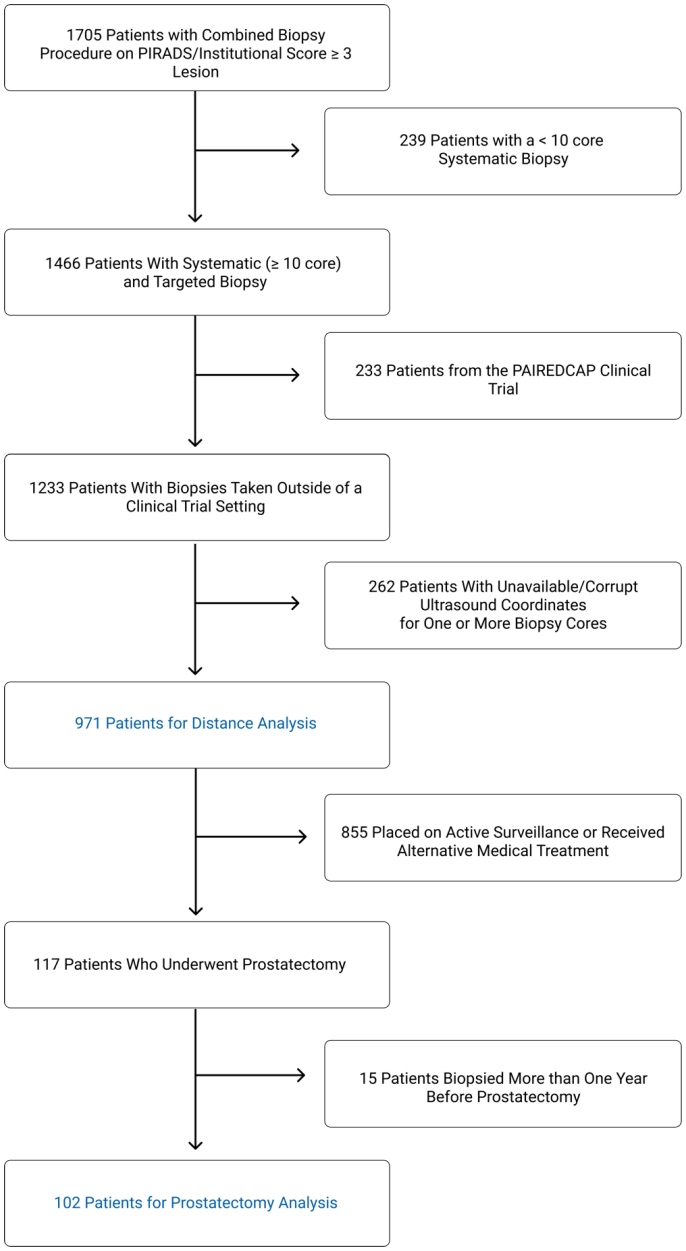


Figure 3.1: Patient Exclusion Criteria. The initial study cohort of 1,705 patients underwent combined biopsy at our institution. Patients were excluded if they received a systematic biopsy with fewer than ten cores, if they were subjects in the PAIREDCAP trial, or if they were missing coordinates for one or more biopsy cores. The primary distance analysis set therefore includes 971 patients. Among these 971 patients, 102 underwent prostatectomy less than a year after biopsy and were included in the prostatectomy subset.

Of these 971 patients, 117 patients underwent prostatectomy after biopsy, and 855 were placed on active surveillance or received other medical treatment. For our prostatectomy subset analysis, we excluded 15 patients whose biopsies occurred more than a year before prostatectomy, yielding a final prostatectomy subcohort of 102 patients. This one-year cutoff was chosen to align with our active surveillance protocol, in which repeat biopsy is not generally performed less than 12 months after the previous biopsy.

3.3.2 Biopsy Core Distance Analysis

In the primary analysis cohort of 971 patients, 16,459 cores were obtained, including 13,515 no cancer cores, 1,409 grade group 1 cores, 941 grade group 2 cores, 243 grade group 3 cores, 168 grade group 4 cores, and 183 grade group 5 cores. The cumulative proportion of cores with csPCa as a function of distance from the ROI is shown in Figure 3.2.

3.3.3 Biopsy Prostate Cancer Detection Rates

The cancer detection rates of different regional target penumbra sizes, as well as the number of cores saved for each size are shown in Table 3.2. Systematic, MRI-targeted, combined, and RTB (defined with a chosen 2 cm margin around the ROI) detected csPCa in 27.0% (262/971), 38.3% (372/971), 44.8% (435/971) and 44.0% (427/971) of patients, respectively. Although combined biopsy detected significantly more patients with csPCa compared to systematic and MRI-targeted biopsy ($p < 0.001$ and $p = 0.004$, respectively), it detected a similar number of patients with csPCa to RTB ($p = 0.71$). The RTB approach resulted in a 22.1% (3,644/16,459) decrease in the overall number of biopsy cores (an average of 3.8 cores per patient) when compared to combined biopsy (Figure 3.3, Table 3.1). MRI-targeted biopsy utilized an average of 3.97 cores per ROI while RTB, which expanded the ROI size, utilized an average of 10.58 cores per ROI.

The cancer detection rates of RTB, MRI-targeted, and systematic biopsy were additionally stratified by PI-RADS score and compared to combined biopsy (Table 3.3). RTB maintained a cancer detection rate above 95% (with the number of csPCa cases found by combined biopsy used as ground truth) for PI-RADS 3, 4, and 5 cases while MRI-targeted biopsy improved steadily from 74.5% to 85.7% to 93.3% for PI-RADS 3, 4, and 5 respectively.

Systematic biopsy, MRI-targeted biopsy, combined biopsy, and RTB detected only cancer-negative cores in 434/971 (44.7%), 446/971 (45.9%), 323/971 (33.3%), and 353/971 (36.4%) patients, respectively and detected at most grade group 1 cancer in 275/971 (28.3%), 153/971

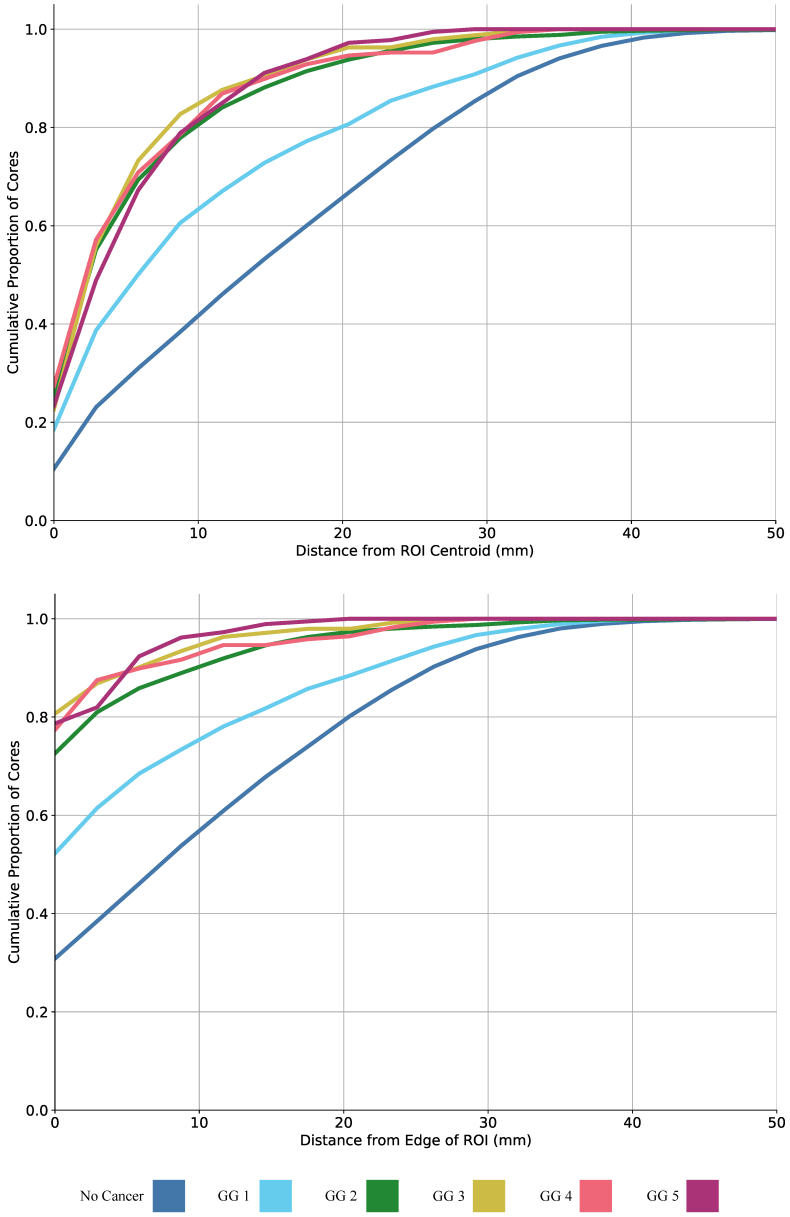


Figure 3.2: Cancer Capture with Distance from the ROI. The proportion of cores found within varying distances of the edge of the closest ROI and the centroid of the closest ROI is shown, stratified by grade group. 94.2% and 97.0% of grade group 2 or higher cores are found within 1.5 cm and 2 cm of the edge of the ROI, respectively. 86.8% and 92.7% of grade group 2 or higher cores are found within 1.5 cm and 2 cm of the ROI centroid, respectively.

Table 3.2: Cancer Detection Rates of RTB with Varying Penumbra Size. The cancer detection rates of regional targeted biopsy with an increasing ROI margin is shown, along with the average number of cores saved relative to combined biopsy. The P-value represents the results of a two-proportion z-test comparing the cancer detection rate of each regional targeted biopsy method with combined biopsy. RTB = regional targeted biopsy, csPCa = clinically significant prostate cancer.

RTB Penumbra Distance (mm)	Patients with csPCa Found	Proportion of Total csPCa Patients Detected	Average Number of Fewer Cores Relative to Combined Biopsy	p-val for RTB vs. Combined
5	396	0.91	9.658	0.074
10	413	0.949	7.45	0.314
15	421	0.968	5.501	0.522
20	427	0.982	3.753	0.715
25	432	0.993	2.211	0.891
30	434	0.998	1.064	0.964

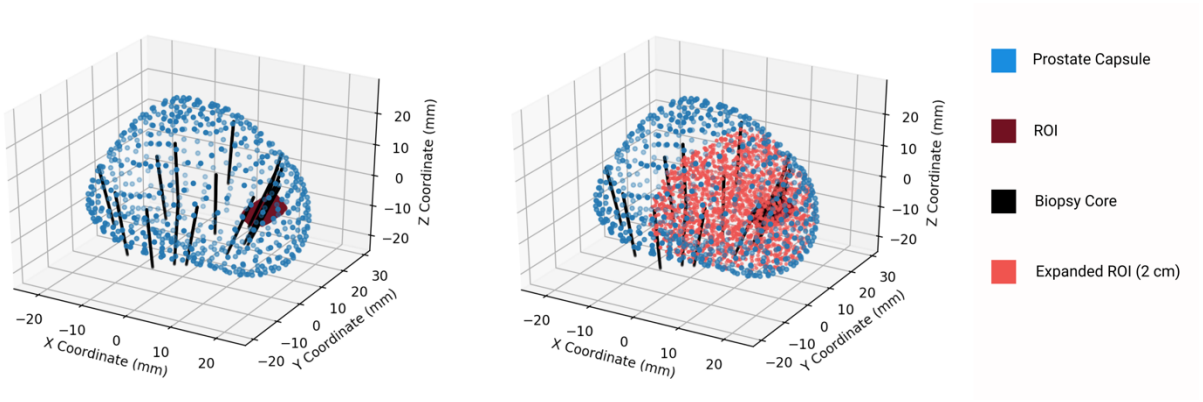


Figure 3.3: Expanded Three-Dimensional ROI for Regional Targeted Biopsy. Three-dimensional representation of a patients prostate with the original (left) radiologist-derived ROI (maroon) and regional target (right) covering 20 millimeters (mm) in all directions from the edges of the ROI.

Table 3.3: Cancer Detection Rates of RTB, MRI-targeted, Systematic, and Combined Biopsy by PI-RADS Score. The number of csPCa cases found by each of the four biopsy methods discussed are stratified by PI-RADS scores. The number in parentheses shows each biopsy methods number of csPCa cases detected as a percentage of the total number of csPCa cases detected by combined biopsy for that PI-RADS score.

PIRADS	RTB	Targeted	Systematic	Combined
3	105 (95.5%)	82 (74.5%)	69 (62.7%)	110 (100%)
4	173 (98.9%)	150 (85.7%)	111 (63.4%)	175 (100%)
5	149 (99.3%)	140 (93.3%)	82 (54.7%)	150 (100%)

(15.8%), 213/971 (21.9%), and 191/971 (19.7%) patients, respectively. Combined and RTB detected only cancer-negative cores in a similar number of patients ($p = 0.15$) but significantly fewer than MRI-targeted and systematic biopsy ($p < 0.001$). Systematic biopsy detected significantly more grade group 1 prostate cancer compared to combined, MRI-targeted, and RTB ($p = 0.001$, $p < 0.001$, $p < 0.001$, respectively).

3.3.4 Locations of Positive Biopsies Outside MRI Targets

In 63 of 971 patients (6.5%) csPCa was detected only on systematic biopsy. In 8 of these 63 cases (12.7%), a systematic core that detected csPCa was greater than 2 cm from an MRI target (i.e. outside the regional penumbra). Every csPCa systematic core found for these 8 patients was of grade group 2. Of the 63 patients for whom systematic biopsy alone found csPCa, 18 had bilateral or midline targets and 45 had unilateral targets. Within the set of 45 patients with unilateral targets, csPCa was detected only ipsilateral to the target in 25 patients (55.6%), only contralateral to the target in 16 patients (35.6%), and both ipsilateral and contralateral to the target in 4 patients (8.9%). Entirely omitting contralateral biopsy would have thus missed csPCa in 16/971 patients (1.6%). The locations and grade groups of the positive cores found outside of unilateral MRI targets are shown in 3.4.

Table 3.4: Grade Groups of Positive Cores Found Outside Unilateral MRI Targets. Positive core counts in each group are presented; all cores originate from the 45 patients who had only ipsilateral MRI targets and positive cores outside of those targets. Four of these patients had both ipsilateral and contralateral positive cores, which have been allocated to the appropriate columns.

Gleason Group	Ipsilateral Lesions ($N = 37$)	Contralateral Lesions ($N = 25$)
Group 2	28	23
Group 3	4	2
Group 4	4	0
Group 5	1	0

3.3.5 Whole Mount Histopathology Analysis

For the subcohort of 102 patients who underwent robotic prostatectomy and MRI-sectioned axial whole mount histopathology within a year of combined biopsy, 20.6% (21/102) and 12.7% (13/102) of patients were upgraded to grade groups ≥ 2 and ≥ 3, respectively, when compared to the maximum grade group assigned to any retrieved biopsy core (i.e. any combined biopsy core, Figure 3.4). When only RTB cores within 2 cm of a target were

included in the comparison, 25 (24.5%) and 14 (13.7%) were upgraded to grade group ≥ 2 and grade group ≥ 3, respectively. These upgrading rates were not significantly different when compared to combined biopsy ($p = 0.50$, $p = 0.84$, respectively). When the upgrading results of MRI-targeted biopsy alone or systematic biopsy alone were compared to combined biopsy, all upgrading rates were significantly higher, except for the grade group 3 upgrading of MRI-targeted biopsy (19 vs. 13 cases, $p = 0.25$). Downgrading on whole mount pathology occurred in relatively few cases without significant differences between biopsy protocols ($p > 0.05$ for all comparisons).

3.4 Discussion

An ideal prostate biopsy protocol would maximize the detection of csPCa using the fewest biopsy samples to optimize clinical utility while minimizing morbidity and cost. In this study, we used a retrospective analysis to evaluate a regional targeted biopsy strategy in which biopsy cores are only sampled from MRI targets (and their 2 cm margins) with a PI-RADS-related score of 3 or higher. We found that this optimized strategy performed similarly to combined biopsy in the detection of csPCa, while requiring significantly fewer biopsy cores (on average 3.8) per patient and 22% fewer cores overall.

In the entire study cohort, we found that 94.2% and 97.0% of grade group 2 or higher prostate cancers were detected even if cores retrieved more than 1.5 cm or 2 cm, respectively, from the edge of the MRI target were removed from consideration. The high csPCa detection rate of cores in the penumbral region of MRI targets confirms the importance of the MRI-derived ROI as a hub of csPCa and supports the role of an institutional Likert and PI-RADS-based ROI scoring system as a predictor of underlying csPCa [RBT17, FNM16, MBS17]. This study also confirms that MRI-targets that are drawn for specificity can underestimate the true size and extent of tumor volumes [PNK17, PEV18]. Our analysis of the relationship between RTB distance thresholds and the resulting cancer detection rates (Table 3.2) may also have implications for optimal margin size determination for focal therapy.

A major advantage of this study is the use of whole-mount histopathology data to indicate the ground truth presence of csPCa. We found that the prostate cancer upgrade rates after prostatectomy for combined biopsy and for RTB did not exhibit a statistically significant difference, despite the fewer biopsy cores used for the regional strategy. In contrast, systematic biopsy and MRI-targeted biopsy alone had significantly higher upgrade rates than combined biopsy. This aligns with other studies that show that combined biopsy demonstrates

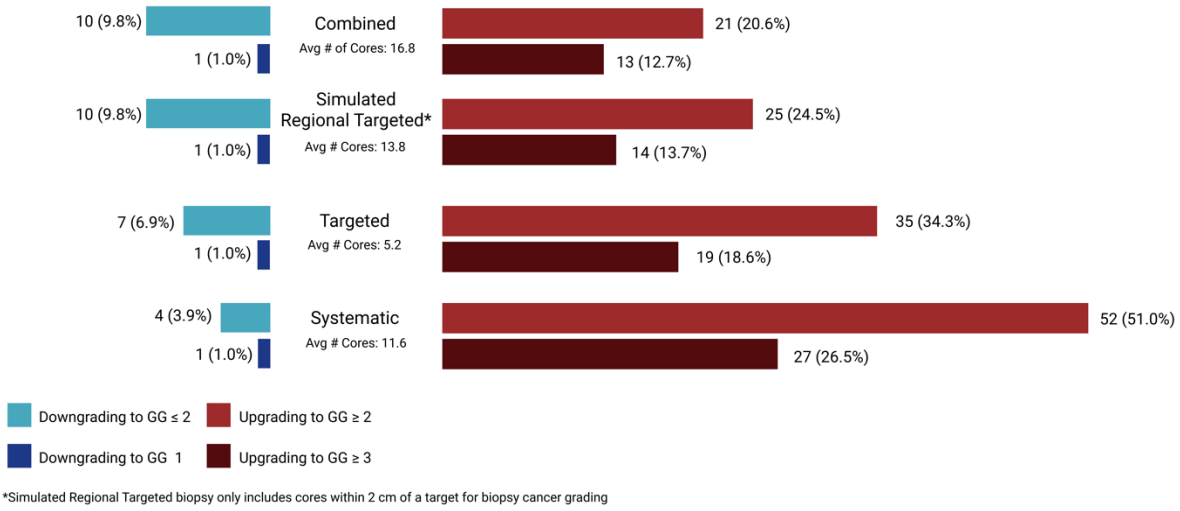


Figure 3.4: Upgrading and Downgrading of csPCa Diagnosis After Robotic Prostatectomy. The highest grade group from the biopsy and subsequent prostatectomy grade group were compared. The upgrading and downgrading of these grade groups for each of the four biopsy methods is shown, using the whole mount prostatectomy as ground truth.

the fewest upgrades on prostatectomy compared to systematic and MRI-targeted, and that MRI-targeted biopsy tends to have fewer upgrades than systematic biopsy [AWR20, DST19]. Ultimately, the results of our whole-mount analysis suggest that a regional targeted biopsy can be an effective method for maximizing csPCa yield by achieving the sensitivity benefit of combined biopsy with fewer sampled cores.

One limitation of our work is the retrospective analytic approach we used to evaluate regional targeted biopsy. In this approach, we censored certain systematic cores based on a defined distance from the ROI as a stand-in for a true regional targeted biopsy. Thus, this study cannot establish the prospective efficacy of a true regional targeted biopsy when compared to combined biopsy. Since systematic biopsy cores were obtained after targeted biopsy cores, it is also possible that the operators knowledge of the target location may have influenced the placement of systematic cores. In addition, results are from a single tertiary institution with genitourinary MRI and pathology expertise, and all biopsy procedures were performed by a single urologist (LSM) with significant MRI-ultrasound fusion biopsy experience; as such, our findings may not be representative of those obtained in other care settings. The real-time sensor fusion approach we used to determine biopsy core locations has a 2-3 mm registration uncertainty, which may have led to inaccuracies in the calculation of biopsy distances [MYN13]; additionally, the designation of ROIs was done by a single radiologist for any given patient, and may be subject to inter-reader differences in boundary delineation.

3.5 Conclusions

We found that a regional targeted biopsy strategy had statistically similar sensitivity for clinically significant prostate cancer as a combined biopsy approach while requiring fewer cores, outperforming the MRI-targeted and systematic biopsy approaches alone. The success of the strategy was driven by the propensity of the most significant biopsy cores retrieved to be in the penumbral region of MRI targets with a PI-RADS-related score of 3 or higher. These findings can be useful to clinicians when determining the optimal set of biopsy locations for an individual patient and suggest that the regional targeted biopsy approach should be further evaluated as an alternative to combined MRI-targeted and systematic biopsy.

CHAPTER 4

Harnessing Clinical Annotations to Improve Deep Learning Performance in Prostate Segmentation

4.1 Overview

Prostate segmentation is a component of the routine evaluation of prostate magnetic resonance imaging (MRI) necessary both for surveillance (through volume estimation) as well as targeted biopsy (to enable registration with real-time ultrasound). In the segmentation workflow, a clinician (generally a radiologist or urologist) will manually review the slices of a 3D T2-weighted MR image and produce a contour for each slice. In some workflows, the radiologist will use a computer-assistance tool, such as DynaCAD Prostate (Invivo-Philips, Gainesville, Florida) [Inv], to assist in segmentation, either by first producing an approximate annotation that is then edited by the radiologist, or by providing an assisted drawing tool that heuristically supports the designation of a contour. Regardless of workflow, segmentation requires a slice-by-slice analysis, which is time consuming, requires the skills of a specially trained radiologist, and is prone to intra- and inter-reader variability [BCS19]. In addition to the utility of such segmentations for these clinical applications, obtaining a precise segmentation is critical for supporting image analysis research, as incorrectly assigning image regions may impair trained classifier accuracy, particularly in the case of lesion detecting classifiers that rely on input prostate segmentations as a component of the input pathway.

Automated prostate segmentation is an active area of research, and substantial published work exists on the development of machine learning models for the purpose. However, these state of the art prostate segmentation algorithms [JXS19, JYF21, WLT19, TLZ20, WWW21, ZWY20] are often trained on small research-quality annotated datasets curated

---

The work described in this chapter is in press as: Sarma KV, Raman AG, Dhinagar NJ, et al. Harnessing clinical annotations to improve deep learning performance in prostate segmentation. PLOS ONE. 2021. doi:10.1371/journal.pone.0253829.

specifically for machine learning. Examples include the 100 patient Prostate MR Image Segmentation (PROMISE12) challenge dataset [LTV14] and the 60 patient NCI-ISBI (National Cancer Institute International Symposium on Biomedical Imaging) Automated Segmentation of Prostate Structures (ASPS13) challenge dataset [Blo15]. Other algorithms have been trained on institutionally developed local datasets that include between 100 and 650 studies [SZH20, CLR19, CZB17, SHS21]. Unfortunately, the development of research-quality prostate boundary annotations is challenging. For example, for the PROMISE12 dataset, segmentations were created by an experienced radiologist, verified by a second experienced radiologist, and then re-annotated by a third nonclinical observer – a complex and expensive process.

If, however, rough clinical annotations could be used to enable training a highly accurate segmentation model, these issues could be avoided, and substantially more data could be available. In this study, we train a prostate segmentation model using a large clinical prostate MRI dataset and rough clinical annotations created as part of the clinical workflow at our academic medical center. We then explore generalizing that model through refinement with small datasets, and the impact of original dataset size on generalizability. Finally, to confirm that it is the prostate specific features in our model that improve generalization rather than general MR features, we explore the relative utility of using our pretrained prostate model for as a basis for generalization versus a model pretrained on an MR dataset from brain cancer patients.

4.2 Materials and Methods

4.2.1 Data

Four retrospective sources of data were used for this project. For training our segmentation model with our clinical data, we used MRI data collected from patients seen at our institution during routine clinical procedures. For examining generalization, we made use of two research-quality prostate MRI challenge datasets. Finally, for determining the relative utility of using our model trained with clinical data as a pre-trained starter, we made use of a brain MRI challenge dataset for comparison. All data was used for this work under the approval of the University of California, Los Angeles (UCLA) institutional review board (IRB# 16-001087). Informed consent was waived with the approval of the IRB for this retrospective study of medical records, based on institutional guidelines, the fact that the study involved no more than minimal risk, the fact that the waiver would not adversely affect the

rights and welfare of study patients, and the impracticality of conducting the retrospective analysis in which results would not change care already delivered to study patients. Data used for this study was de-identified after collection and before analysis.

4.2.1.1 Primary Dataset

Our internal clinical population for this study consists of 1,620 MRI studies collected from 1,111 patients who underwent transrectal ultrasound-MRI fusion biopsy (TRUS biopsy) using the Artemis guided biopsy system (Eigen Systems) between 2010 and 2018 at our institution using a standardized protocol and 3T magnet (Trio, Verio, or Skyra, Siemens Healthcare). As part of the protocol, prostate MRIs were contoured in a two-part process. First, the attending radiologist for the case (the attending radiologists for the patients included in this study each had between 10-27 years of experience) created a prostate contour using the DynaCAD Prostate image analysis platform as part of the routine clinical workflow. This contour was then used by a technician to re-contour the prostate on the Profuse (Eigen Systems) platform in order to enable use with the Artemis biopsy system, as DynaCAD segmentations cannot be directly imported for use on the Artemis.

We retrospectively collected 3D axial turbo spin echo (TSE) T2-weighted images and prostate contour sets from these studies. T2 images were acquired clinically using the spatial and chemical-shift encoded excitation (SPACE, Siemens Healthcare) protocol, with field of view (FOV) 170 x 170 x 90 mm^3 and resolution 0.66 x 0.66 x 1.5 mm^3. Acquisition parameters are provided in Table 4.1. Studies were collected from our institutions picture archiving and communication system (PACS). Corresponding T2 prostate contours were collected from the Profuse image analysis platform. Imaging data were collected from every available study for each patient seen at our institution during the study period. Studies were excluded from retrieval if the T2 image or contour was missing from PACS or corrupt, or if the image exhibited a protocol deviation, such as a variance in FOV or resolution. A total of 1,620 studies were included from 1,111 patients, and 84 studies were excluded. Of the 1,620 included studies, 29 used an endorectal coil.

4.2.1.2 External Prostate Challenge Datasets

Two external challenge prostate datasets were used for this study: ProstateX-2 [AHD18] and PROMISE12 [LTV14].

The ProstateX-2 Challenge was a prostate cancer prediction challenge held in 2017. This

Table 4.1: Imaging acquisition parameters for study datasets. Full acquisition data is not available for the PROMISE12 dataset, and the counts for images acquired at different field strengths and resolutions are not available.

	UCLA (n=1620)	ProstateX-2 (n=99)	PROMISE12 (n=50)
Vendor(s)	Siemens	Siemens	Siemens, GE
Field Strength	3T	3T	1.5T, 3T
In-plane resolution (mm)	0.664	0.5	0.25-0.75
Slice thickness (mm)	1.5	3.6	2.2-4.0
TR (ms)	2200	5660	Not available
TE (ms)	201	104	Not available
Endorectal coil used (n, %)	29 (1.8%)	0 (0%)	24 (48%)

dataset consists of 99 deidentified cases collected from patients seen at Radboud University Medical Center in the Netherlands. A consistent imaging protocol was used for all cases, which was significantly different from the protocol used for the primary dataset at our institution. A variety of images and clinical variables were provided with each case. For use in our experiments, we retrieved transverse T2-weighted MR images from each case in the dataset. These images were then annotated with a research-quality prostate contour by co-author B.T., an experienced abdominal radiologist.

The PROMISE12 Grand Challenge was a prostate segmentation-specific challenge held in 2012. This dataset includes 50 deidentified cases collected from four different centers (Haukeland University Hospital in Norway, Beth Israel Deaconess Medical Center in the United States, University College London in the United Kingdom, and Radboud University Nijmegen Medical Center in the Netherlands). Each institution had unique acquisition protocols, with wide variability in the MR field strength, endorectal coil usage, and image resolution. Each case consists of a transverse T2 weighted MR image and a reference research-quality prostate contour produced by agreement of two expert radiologists (one radiologist at the institution where the image was acquired, and a second radiologist at Radboud University). Detailed acquisition parameters are not available for this dataset, but images were scanned at a variety of field strengths (1.5T or 3T), with or without endorectal coil, and with a variety of acquisition resolutions, pulse sequences, and device manufacturers [LTV14].

4.2.1.3 Brain Cancer Challenge Dataset

Transfer learning is a often-used approach for accelerating the development of deep learning models [RZK19]. In transfer learning, a pre-trained model, usually trained on an out-of-domain dataset such as ImageNet [RDS15], is used to initialize model weights before fine-tuning on the study dataset. This approach is well-known to improve convergence and model performance in natural image recognition and medical image analysis by providing pre-trained feature detection layers. Unfortunately, there is not a commonly accepted pretrained 3D model that could be used in order to facilitate comparison against our domain-specific clinically annotated data.

As such, in order to provide a non-prostate comparison, the Brain Tumor Segmentation (BraTS) 2019 [BAS17, MJB15] challenge dataset was also used for this study. The dataset includes over 300 annotated cases collected from 19 different institutions using a wide variety of protocols. These cases include T2-weighted images of the brain with tumor segmentations.

These segmentations were created manually using a multi-step protocol requiring agreement between multiple raters and final approval by an experienced neuroradiologist. Though tumor segmentation is a far more complex segmentation task than organ segmentation, this dataset provided an MRI comparison with a defined 3D segmentation task that could be leveraged as pretraining for prostate MRI segmentation. The BraTS 2019 data originates from large number of institutions and includes data collected with a variety of acquisition parameters; a specific breakdown of these parameters is not available [BRJ18].

4.2.2 Preprocessing

In order to facilitate transportability, we processed images from all three datasets using the same pipeline. Initial preprocessing was done in Python, primarily using the SimpleITK toolkit [LCI13], and included bias field correction [TAC10] and resampling to isotropic voxel size (1mm x 1mm x 1mm) for further processing; these steps were based on preprocessing done in previous work [CLR19, TAC10, SSR18, GMV15]. After initial preprocessing, we applied interquartile range (IQR)-based intra-image normalization to address the relative nature of MR image intensity values (both within and between institutions). Each image was normalized to the image-level IQR (calculated from the central 128x128 column of the volume) and then values were clipped between two IQRs below the first quartile and five IQRs above the third quartile, in order to eliminate outlying values created by imaging artifacts. The preprocessing pipeline is depicted in Figure 4.1.

4.2.3 Augmentation

For all model training in this study, real-time augmentation using the Batchgenerators package was performed [IJW20]. Three augmentation transformations were used: 1) random elastic deformation, 2) random rotation in the range $[-\pi/8, \pi/8]$ in the axial plane, and $[-\pi/4, \pi/4]$ along the axis, and 3) random mirroring across the y-axis. After augmentation, the image was cropped to the central column of the transformed image (i.e. the central 128x128 voxels in the x, y plane), which always contained the prostate for our dataset.

4.2.4 Model, Training and Evaluation

The base model used for this study was the 3D U-Net [CAL16]. For all experiments, the network was configured with four encoder levels, three decoder levels, a ReLU transfer func-

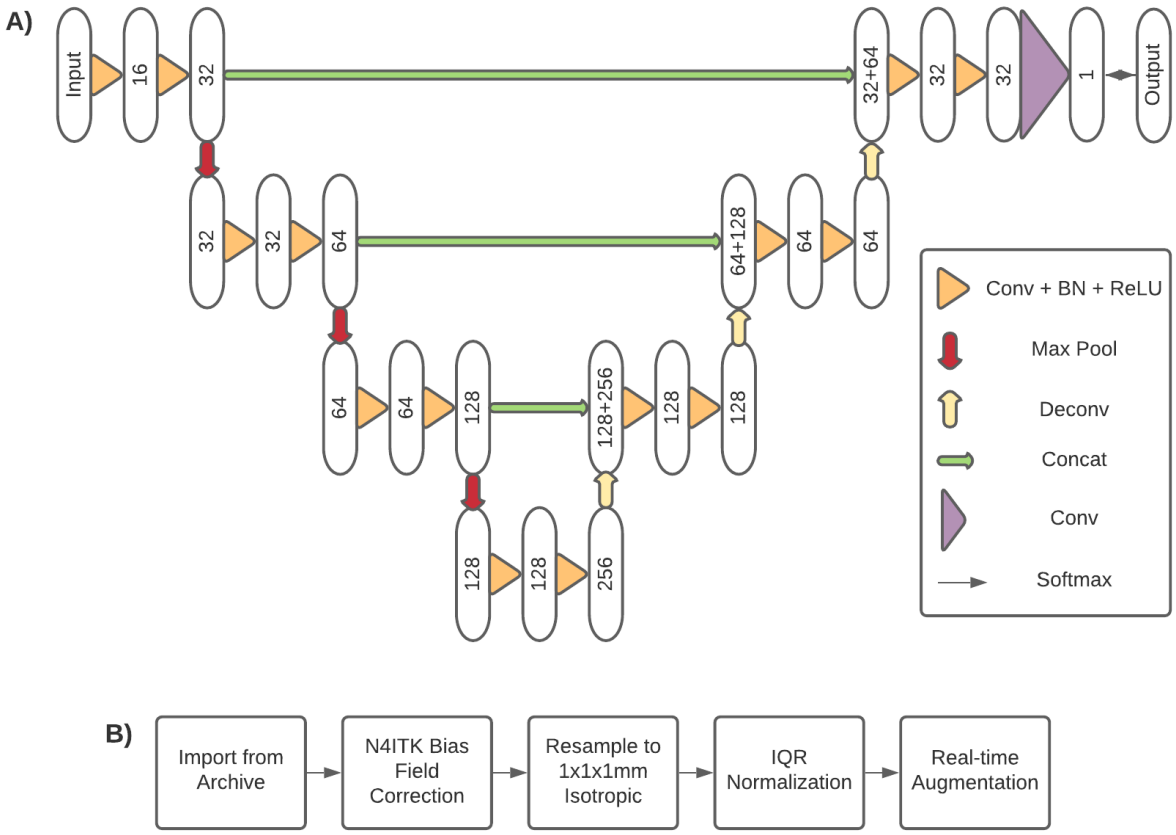


Figure 4.1: 3D U-Net Model Diagram and Preprocessing Steps. A) Network diagram of the 3D U-Net used for this study. Numbers within the ovals represent number of feature maps at that layer. Connections represent network operations, such as 3x3x3 3D convolution ("Conv"), 2x2x2 max pooling ("Max Pool"), 3x3x3 3D transposed convolution ("Deconv"), skip feature map concatenation ("Concat"), batch normalization ("BN"), rectified linear unit activation ("ReLU"), and softmax output ("Softmax"). B) Process diagram of preprocessing steps. Once images were imported from the archive (either PACS or challenge download), N4ITK bias field correction was applied. Images were then resampled to 1mm isotropic resolution and IQR normalized. During training, real-time augmentation was applied to each input image to create the training sample for that epoch.

tion and group normalization (using eight groups) following every convolutional layer, and a softmax output layer. The model architecture is depicted in Figure 4.1. All training and evaluation was done using the PyTorch framework on a DGX-1 (NVIDIA) deep learning appliance. Mixed-precision training using the NVIDIA Accelerated Mixed Precision (AMP) was used at optimization level O2, consisting of 16-bit model weights and inputs, 32-bit master weights and optimizer parameters, and dynamic loss scaling.

Network inputs consisted of the full augmented image volume (with dimension 128x128x136). Training was performed using the Adam optimizer with learning rate 10^{-5} and the soft Dice loss function. Each epoch consisted of training on a full dataset comprised of one augmented sample generated for every original input sample.

The primary evaluation metric used to compare segmented volumes was the soft Dice coefficient function as denoted in Equation 4.1, where S_{DL} is the segmentation of a deep learning model and S_m is the manual segmentation. The value of the coefficient can range between 0 (no overlap) and 1 (perfect overlap).

$$\text{DICE}(S_{DL}, S_m) = \frac{2|S_{DL} \cap S_m}{|S_{DL}| + |S_m} \tag{4.1}$$

The average Hausdorff distance (AHD) was also used as a secondary metric, as denoted in Equation 4.2, where X is the set of all points within the manual segmentation, Y is the set of all points within the segmentation of the deep learning model, and d is the Euclidean distance. The AVD is a positive real number, and smaller numbers denote better matching segmentations.

$$\text{AHD}(X, Y) = \frac{\frac{1}{|X|}\sum_X \min_Y d(x,y) + \frac{1}{|Y|}\sum_Y \min_X d(x,y)}{2} \tag{4.2}$$

The evaluation metrics were calculated for whole prostate gland segmentation on the entire uncropped volume. In addition, each slice of the segmentation mask was split into three subvolumes: the apex subvolume (consisting of the apical 25% of prostate slices), the base subvolume (consisting of the basal 25% of prostate slices), and the midgland subvolume (consisting of the remaining middle 50% of slices); the Dice evaluation metric was calculated for each subvolume. Means and standard deviations across the entire dataset were reported for performance on the whole prostate as well as each of the three subvolumes. These were calculated using the following approach: for each of the five folds, metrics were calculated for each of the images within the fold using the model trained without that folds data.

Once the metrics were calculated for every study, the mean and standard deviation of each metric across all images was computed (including whole-volume and subvolume metrics) and reported as the evaluation result.

4.3 Experiments

4.3.1 Baseline Models

To establish baseline performance, models were first trained from scratch separately on the primary dataset, the ProstateX-2 (PX2) data, and the PROMISE12 (P12) data. Training was performed using five-fold cross-validation (CV) over each entire dataset, with 324 images per fold. The evaluation metrics were then computed using the approach described above.

4.3.2 Generalizability to Challenge Datasets

To assess the utility of the baseline primary dataset model on the external challenge datasets, two sets of experiments were done for each dataset. First, the model was used to produce segmentation mask predictions for each example in the external datasets, and mean scores were reported for each dataset. Then, the model was refined for each external dataset using the baseline primary dataset model as the pretrained weight initializer. This refining was done using five-fold CV over 100 epochs, and validation soft Dice scores were calculated and reported as in the previous experiments. Results were compared for superiority against the baseline models using a one-tailed paired t-test, with $\alpha = 0.001$.

4.3.3 Impact of Dataset Ablation

To assess the impact of the size of the primary dataset on both segmentation performance and generalizability, a series of ablation experiments were conducted. First, a series of models was trained using truncated versions of the primary dataset. In these experiments, the training set for each fold was truncated to a fixed proportion of its original size, from 5% to 80%. The validation set was not truncated to ensure a fair comparison. Five-fold CV was again used over 100 epochs. The resulting models were then evaluated using the soft Dice criterion to determine model performance on the primary dataset. In order to determine the impact of ablation on generalizability, the resulting models were then refined for 100 epochs using the PX2 or P12 datasets (without truncation), and then evaluated as in the previous

experiments. These models were compared for superiority against the baseline models using one-tailed paired t-tests, with $\alpha = 0.001$.

4.3.4 Comparison to BraTS Model

In order to assess the relative importance of using the domain-specific primary baseline model as a pretrained weight initializer, a comparison model was trained using the BraTS dataset. The BraTS dataset was chosen for the comparison model because of the similar underlying data (T2-weighted imaging) and 3D nature of the desired output. The BraTS data was preprocessed using the same pipeline before training as the prostate data, and the same model architecture and training protocol was used. Five-fold CV was performed over 150 epochs. The BraTS model was then used as a pretrained weight initializer for refining PX2 and P12 segmentation models, using the same approach as in the previous experiments. These models were compared for superiority against the baseline models using one-tailed paired t-tests, with $\alpha = 0.001$; additionally, the refined ablation models were compared against the refined BraTS models for superiority using one-tailed paired t-tests, with $\alpha = 0.001$.

4.4 Results

4.4.1 Baseline Models

Training results are shown in Table 4.1; all results are reported as mean ± standard deviation in tables and text. The primary baseline model achieved a high overall performance, with a mean overall Dice coefficient of 0.909 ± 0.042 and mean AHD of 0.156 ± 0.231. This result is comparable to the top performing prostate segmentation models found in the literature. Example evaluation segmentations for the baseline model on the primary dataset are shown in Figures 4.2, 4.7, and 4.8. The PX2 and P12 models performed less well, with mean overall Dice coefficients of 0.702 ± 0.083 and 0.568 ± 0.122, and mean AHDs of 0.480 ± 0.555 and 2.155 ± 2.466, respectively. Across all three models, midgland Dice performance was the highest (0.762-0.941) and performance on the base and apex regions was more limited (0.501-0.863). The P12 model was the worst performing across every measure. Performance measures on a per-sample basis for the PX2 and P12 baseline models are shown in Figure 4.3.

Table 4.2: Evaluation results for baseline models. PX2 = ProstateX-2, P12 = PROMISE12, AHD = average Hausdorff distance; results reported as mean ± standard deviation across all images.

Dataset	Soft Dice Coefficients				AHD
	Overall	Base	Midgland	Apex	
Primary	0.909 ± 0.042	0.863 ± 0.095	0.941 ± 0.030	0.832 ± 0.094	0.156 ± 0.231
PX2	0.702 ± 0.083	0.679 ± 0.117	0.849 ± 0.051	0.702 ± 0.093	0.480 ± 0.555
P12	0.568 ± 0.122	0.501 ± 0.168	0.762 ± 0.087	0.561 ± 0.168	2.155 ± 2.466

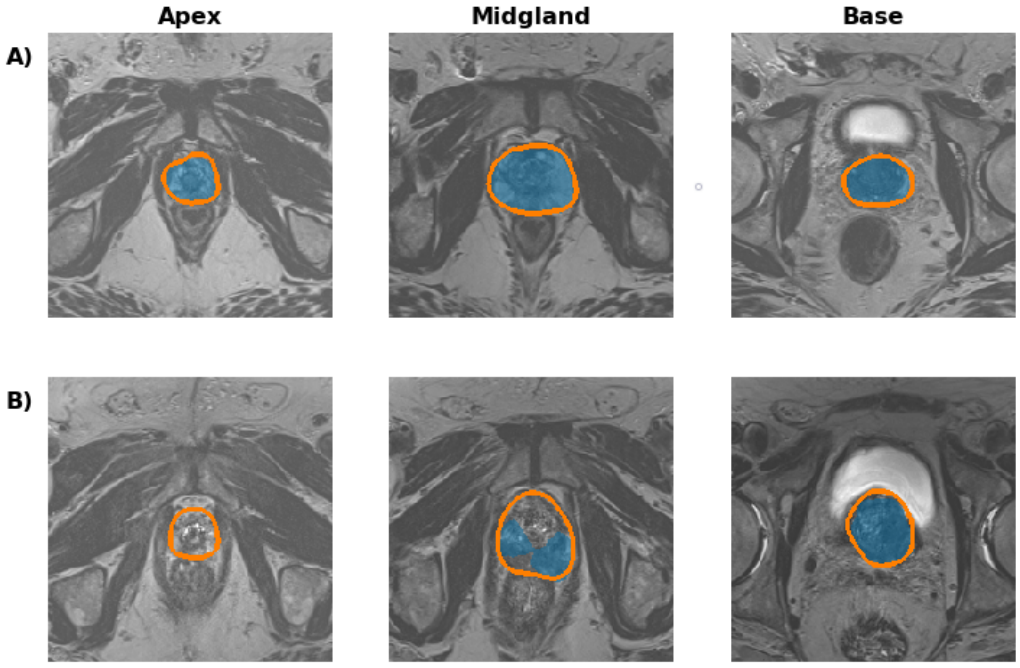


Figure 4.2: Example UCLA baseline model segmentations. The orange contour depicts ground truth segmentation and the shaded blue area depicts model segmentation. A) Example apex, midgland, and base slice from a sample in the primary dataset with a high metric on evaluation. The soft Dice coefficient for this sample was 0.928, and the average Hausdorff distance was 0.085. Images of all of the slices for this study are presented in Figure 4.7. B) Example apex, midgland, and base slice from a sample in the primary dataset with a low metric on evaluation. The soft Dice coefficient for this sample was 0.738, and the average Hausdorff distance was 0.935. Images of all of the slices for this study are presented in Figure 4.8.

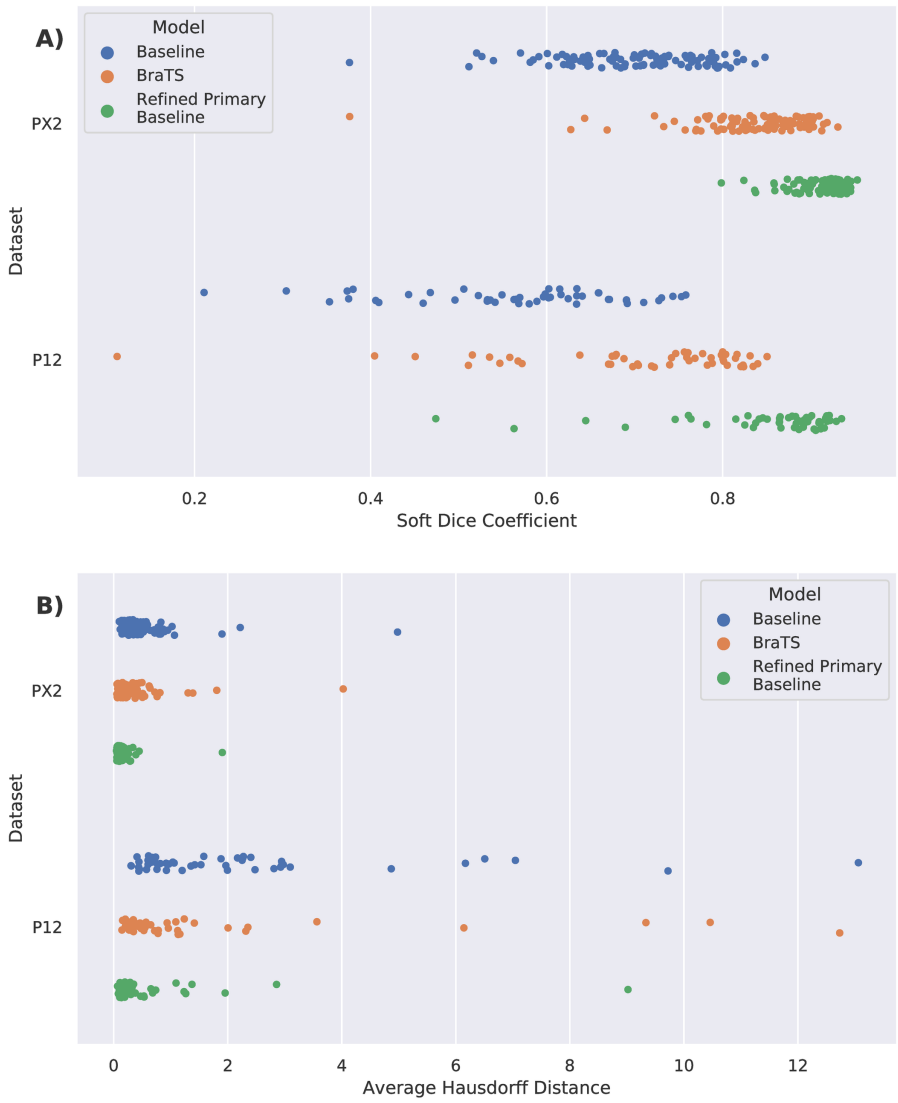


Figure 4.3: Evaluation metrics for PX2 and P12 datasets. Soft Dice coefficients (A) and average Hausdorff distances (B) for every sample in the ProstateX-2 (PX2, n=99) and PROMISE12 (P12, n=50) datasets, after model evaluation for the baseline, BraTS, and refined primary baseline models. Each solid dot represents a single training example. The models trained by refining the BraTS pretrained model or the baseline pretrained model both exhibited improved performance and reduced variance on both evaluation metrics, and with the refined primary baseline model exhibiting the highest performance and lowest variance. Detailed statistics are available in Tables 4.2, 4.3, and 4.5.

4.4.2 Generalizability to Challenge Dataset

Results are shown in Table 4.3. For the PX2 dataset, the primary baseline model exhibited a mean overall Dice coefficient of 0.465 ± 0.291 and AHD of 4.824 ± 5.920 before refining, and a coefficient of 0.912 ± 0.029 and AHD of 0.150 ± 0.192 after refining. For the P12 dataset, the primary baseline model exhibited an overall Dice coefficient of 0.708 ± 0.210 and AHD of 1.953 ± 3.747 before refining and a Dice of 0.852 ± 0.091 and AHD of 0.581 ± 1.314 after refining. Similar to the previous experiments, Dice performance in the midgland region was higher than that in the base and apex regions for all models. For both datasets, the refined primary baseline model performed significantly better ($p < 0.001$) than the baseline model trained with only the respective dataset across all measures. Though the unrefined UCLA model performed better on the P12 dataset, after refining, performance was best on the PX2 dataset. Performance measures on a per-sample basis for the PX2 and P12 refined baseline models are shown in Figure 4.3. Example segmentations before and after refining are shown in Figures 4.4 and 4.5.

Table 4.3: Evaluation results for retargeted models. * denotes significantly higher than baseline model, $p < 0.001$. PX2 = ProstateX-2, P12 = PROMISE12, AHD = average Hausdorff distance; results reported as mean ± standard deviation across all images.

Refining?	Dataset	Soft Dice Coefficients				AHD
		Overall	Base	Midgland	Apex	
No	PX2	0.465 ± 0.291	0.314 ± 0.314	0.517 ± 0.316	0.401 ± 0.312	4.824 ± 5.920
Yes	PX2	0.912*±0.029	0.851*±0.102	0.949*±0.024	0.849*±0.070	0.150*±0.192
No	P12	0.708 ± 0.210	0.475 ± 0.317	0.779 ± 0.215	0.679 ± 0.221	1.953 ± 3.747
Yes	P12	0.852*±0.091	0.744*±0.207	0.918*±0.046	0.777*±0.134	0.581*±1.314

4.4.3 Impact of Dataset Ablation

Results for these experiments are shown in Table 4.4 and Figure 4.6. We found that model performance generally increased as the proportion of data used increased, with the primary model exhibiting an overall mean Dice coefficient of 0.638 at 5% and 0.909 at 100%. Both the PX2 and P12 models exhibited significantly increased performance ($p < 0.001$) over their baseline at all ablation levels. For all three sets of models, the models trained at the 60% ablation level achieved approximately the eightieth percentile performance.

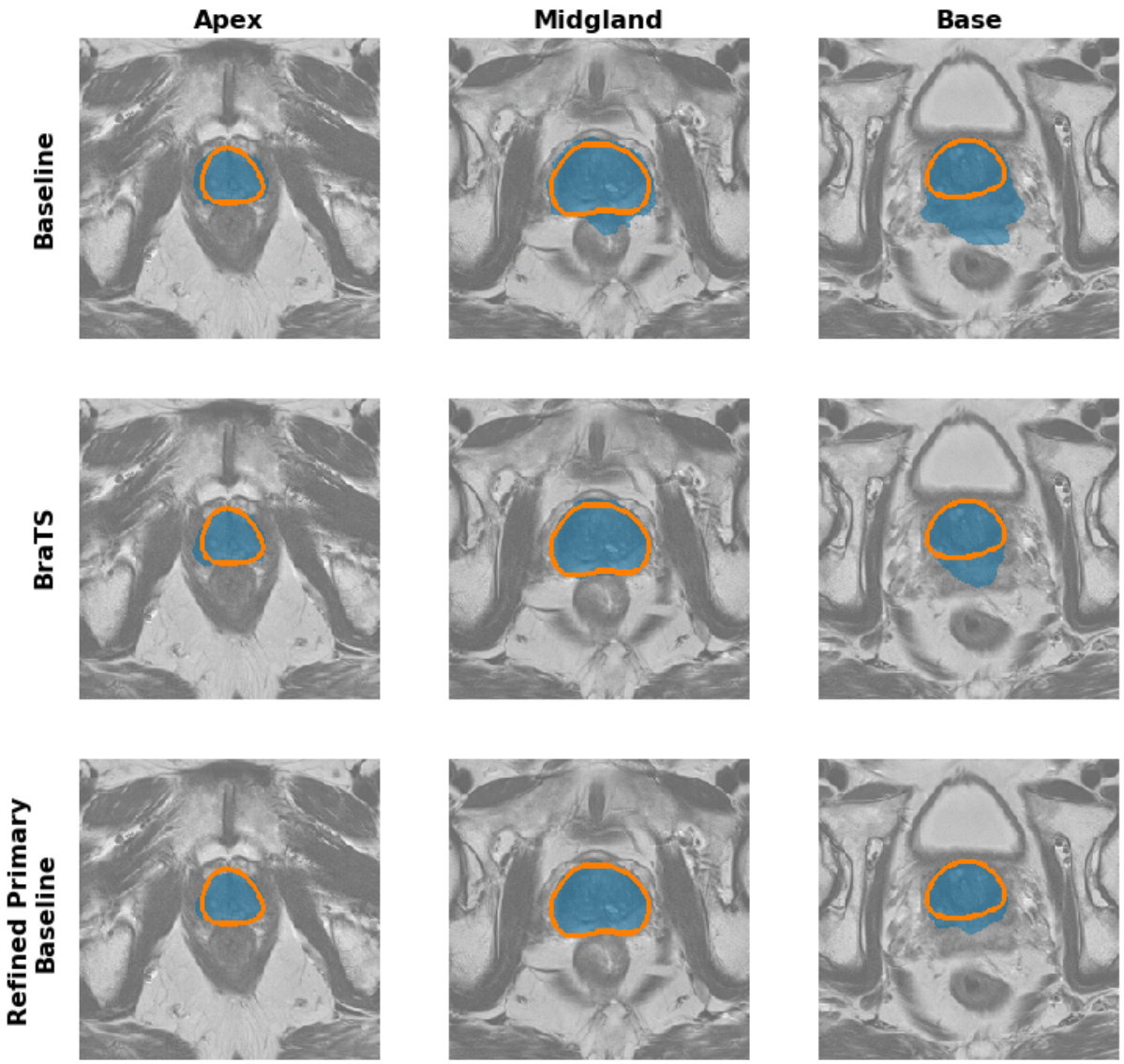


Figure 4.4: Example ProstateX-2 segmentations. Orange contour depicts ground truth segmentation. Shaded blue area depicts model segmentation. The soft Dice coefficient and average Hausdorff distance metrics were 0.645 and 1.024 for the baseline model, 0.864 and 0.167 for the BraTS model, and 0.932 and 0.079 for the refined primary baseline model.

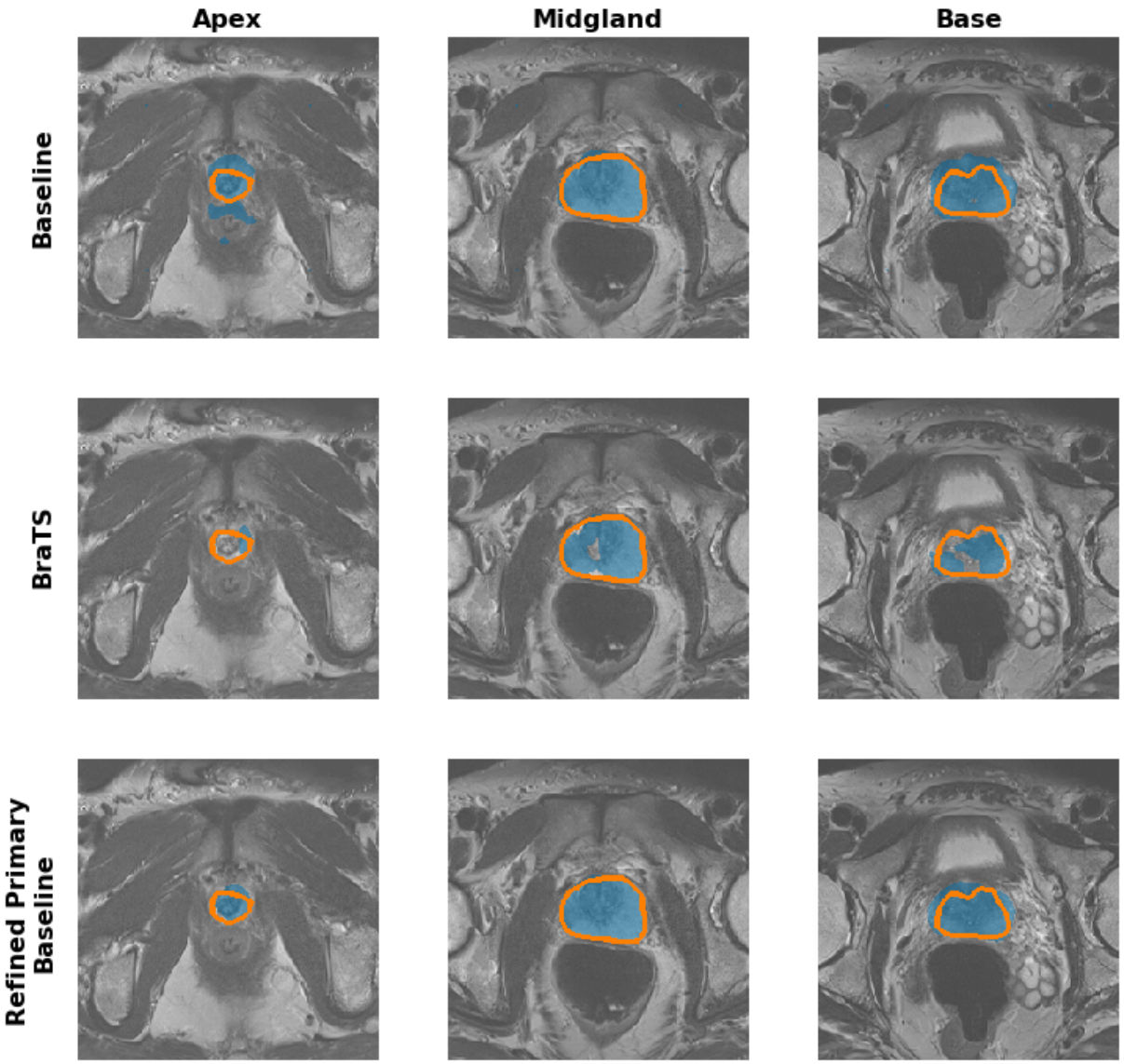


Figure 4.5: Example PROMISE12 segmentations. Orange contour depicts ground truth segmentation. Shaded blue area depicts model segmentation. The soft Dice coefficient and average Hausdorff distance metrics were 0.536 and 2.974 for the baseline model, 0.678 and 0.291 for the BraTS model, and 0.910 and 0.102 for the refined primary baseline model.

Table 4.4: Model performance using ablated primary dataset. Reported metric is overall soft Dice coefficient. * denotes significantly higher than baseline model, $p < 0.001$. PX2 = ProstateX-2, P12 = PROMISE12, FT = fine-tuned, results reported as mean across all images.

Model	5%	10%	15%	20%	40%	60%	80%	100%
Primary	0.638	0.754	0.775	0.825	0.883	0.901	0.906	0.909
PX2 FT	0.740*	0.814*	0.829*	0.861*	0.899*	0.907*	0.909*	0.912*
P12 FT	0.625*	0.721*	0.727*	0.781*	0.831*	0.848*	0.842*	0.852*

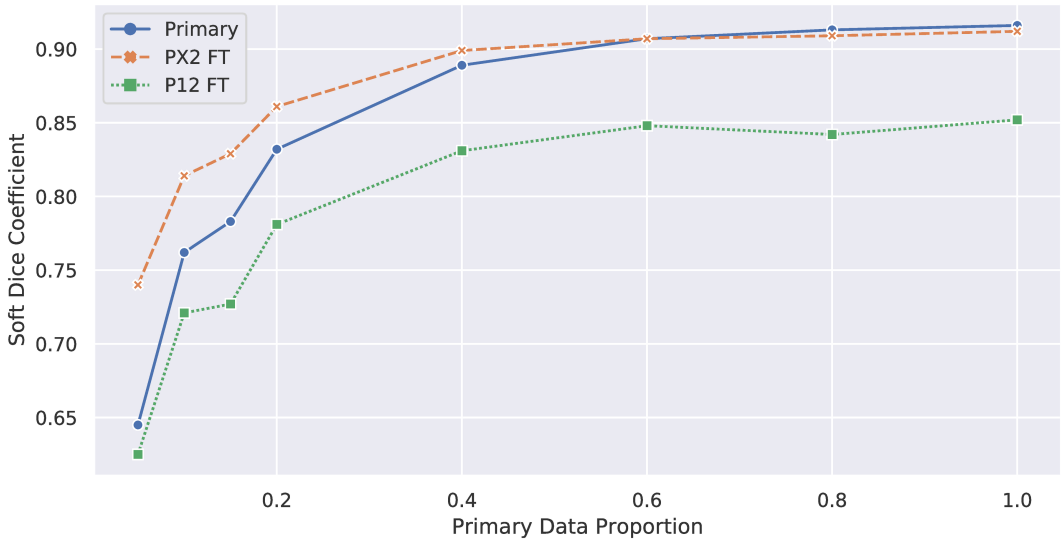


Figure 4.6: Soft Dice coefficients for models trained with ablated dataset. Soft Dice Coefficients for models trained using the ablated primary dataset ("Primary") or trained using an ablated primary model as weight initializer ("FT"). PX2 = ProstateX-2, P12 = PROMISE12, FT = fine-tuned. Significant improvements can be seen in the performance of the fine-tuned models at 5% of the primary dataset used for training the ablated primary baseline model, with the performance benefits leveling out at 60% of the dataset.

4.4.4 Comparison to BraTS Model

The results of these experiments are shown in Table 4.5. The final overall soft Dice coefficient of the resulting model on the BraTS segmentation task was 0.591. When refined on the PX2 dataset, the mean overall soft Dice coefficient was 0.834, and the AHD was 0.299 ± 0.465. When refined on the P12 dataset, the mean overall Dice coefficient was 0.704, and the AHD was 1.428 ± 2.603. In both cases, the refined BraTS model significantly outperformed the baseline model across all measures ($p < 0.001$), but was outperformed by the ablation models at 20% and higher ($p < 0.001$). Performance measures on a per-sample basis for the PX2 and P12 refined BraTS models are shown in Figure 4.3. Example segmentations are shown in Figures 4.4 and 4.5.

Table 4.5: Evaluation results for refined BraTS models. * denotes significantly higher than baseline model, $p < 0.001$. PX2 = ProstateX-2, P12 = PROMISE12, AHD = average Hausdorff distance; results reported as mean ± standard deviation across all images.

Dataset	Soft Dice Coefficients				AHD
	Overall	Base	Midgland	Apex	
PX2	0.834*±0.072	0.783*±0.126	0.903*±0.065	0.775*±0.097	0.299*±0.465
P12	0.704*±0.137	0.614*±0.208	0.820*±0.120	0.644*±0.186	1.428*±2.603

4.5 Discussion

In this study, we developed a prostate segmentation CNN model using a large clinically generated dataset, and examined the relationship between dataset size and model performance. We further explored the generalizability of the model to external datasets, and the relative contribution of using the model as a pre-trained starter for improving performance when training on limited datasets.

We found that the network trained on our institutions dataset did not perform well initially when used on outside data. However, refining the network on the external data using the initial model as a pre-trained starter yielded significantly superior performance to training using randomly initialized models. On the PX2 dataset, using our institutions model as a pre-trained starter yielded an increase in mean overall Dice coefficient of 30%, and on the P12 dataset, an increase of 49%. Using a model trained on data (BraTS 2019) completely unrelated to prostate segmentation as a pre-trained starter also yielded improvements over baseline, but was not as effective as using the primary dataset as a starter. As demonstrated in Figure 4.3, model performance improved progressively from the baseline model, to the

model trained using the non-relevant BraTS MR data, and finally to the model trained with the highly relevant UCLA prostate MR data. The final performance of the models we trained using our pre-trained prostate MR starter was comparable to other results from the literature on both datasets using more complex models [ZWY20, JSH19, IPK18], highlighting the value of creating a domain-specific starter for this task. For example, the leading model on the P12 leaderboard (submitted on 9/7/2020) has a Dice score of 0.895, which compares favorably with our final overall Dice coefficient of 0.852 [LTV14]. The leading model trained on a large private dataset (trained on 648 studies at the NIH) has a Dice score of 0.915, which compares favorably with our 0.909 [SZH20].

We also found that using truncated versions of our dataset still yielded significant improvements. Even using a model trained on only 15% of the primary dataset as a pre-trained starter yielded improvements over baseline of 18% and 28% on the PX2 and P12 datasets, and the gains from increasing dataset set saturated at approximately 60% of the primary dataset.

These findings are notable in part because our primary dataset consists of rough clinical contours that have not been carefully re-annotated to produce a machine learning-quality dataset and images that were not filtered for inclusion of only optimal quality series. We included in our primary dataset images with quality limitations, images that used endorectal coils, and images from patients who had had prostate treatments that significantly distort the visual appearance of the prostate. Despite these complications, we still found that we were able to train a state-of-the-art model and then use that model to boost the performance of models trained on gold-standard data. The performance gained through the use of our model as a pre-trained starter was greater than that obtained using an unrelated pretrained model (as is typical for transfer learning; i.e. ImageNet [RDS15]), suggesting that our model was able to learn features that were useful starters for the segmentation models trained for the external datasets.

Our work does have some limitations. Because we did not use a machine learning-quality version of our dataset, it is difficult to compare the overall performance results on our data to state-of-the-art models. In addition, the imperfections in the clinically generated ground truth segmentations we used for our primary dataset likely include both areas incorrectly annotated in the foreground and the background. As a result, some differences between model predictions and the ground truth in the primary dataset are the result of inaccurate labels, rather than model error.

Because we held the model design constant and simple in order to isolate the dependent

variables in our experiments to the datasets used for training and pretraining (and as such used data from all folds in our evaluations, rather than a single held-out fraction), we may have been prevented from realizing performance gains that other works have found through complex model designs or post-processing steps. However, our intent with these choices was to demonstrate that even a simple model with rough clinical contours can provide substantial value when contemplating model development. This finding may have significant implications for future work, in which larger datasets with lower-quality annotations may be combined with smaller datasets with high-quality annotations to maximize the value of available data without requiring the significant expenditure of re-annotation effort. We plan to further explore this hypothesis in future work using more difficult problems, such as prostate cancer segmentation, in order to determine if this approach may unlock additional potential for medical image analysis. Additionally, because this is a retrospective analysis and does not include the real-time ultrasound used for image fusion, it is not possible for us to evaluate the impact of segmentation quality from different models on registration and biopsy targeting. Future, prospective work should include such an evaluation.

4.6 Conclusion

We trained a state-of-the-art model using rough clinical annotations, producing a prostate segmentation model with a mean overall Dice coefficient of 0.909 and an AHD of 0.156. We additionally found that models trained using truncated fractions of our data were effective pre-trained starters for achieving higher performance models on external prostate segmentation challenge datasets. Our findings suggest a role for the combined use of datasets with low-quality and high-quality annotations in future medical image analysis model development in order to maximize performance while minimizing annotation effort.

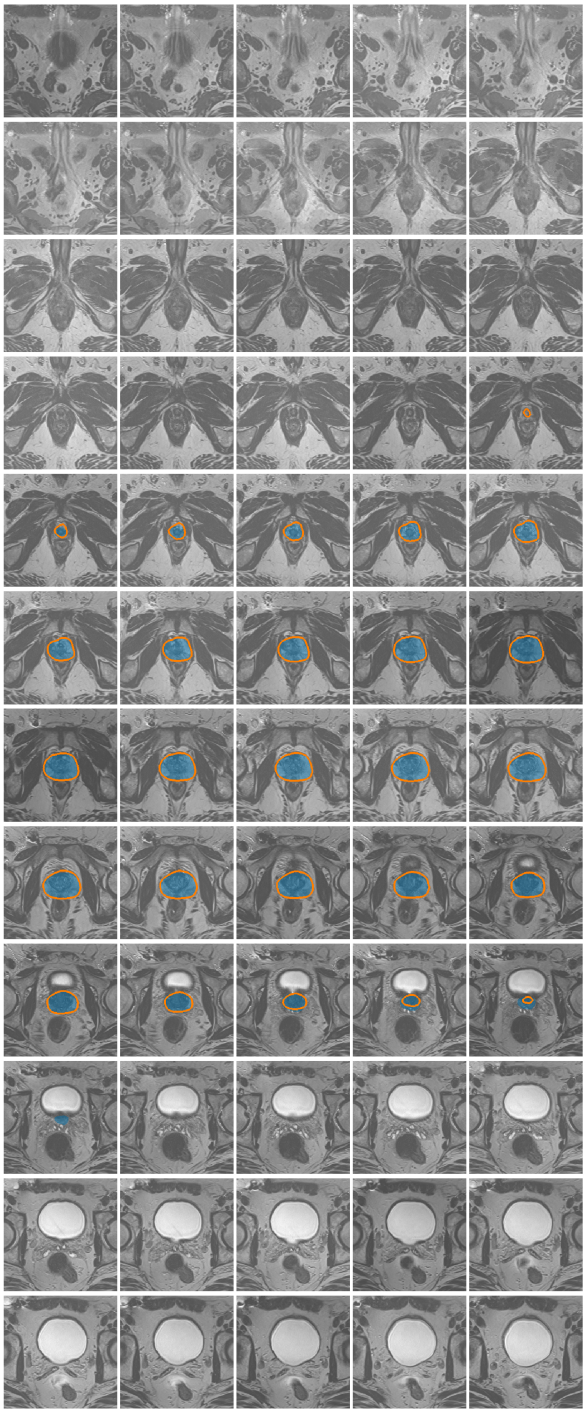


Figure 4.7: Full volume example of primary baseline dataset segmentation, high metric. Orange contour depicts ground truth segmentation. Shaded blue area depicts model segmentation. Slices depicted from apex to base. The soft Dice coefficient for this sample was 0.928, and the average Hausdorff distance was 0.085.

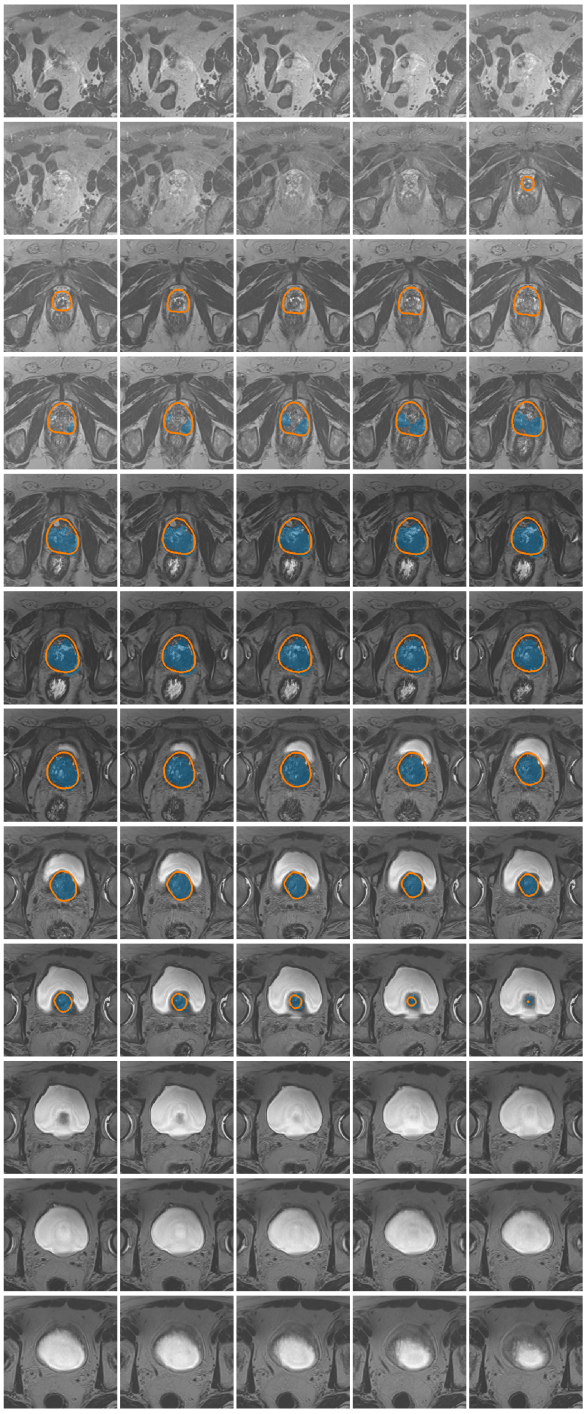


Figure 4.8: Full volume example of primary baseline dataset segmentation, low metric. Orange contour depicts ground truth segmentation. Shaded blue area depicts model segmentation. Slices depicted from apex to base. The soft Dice coefficient for this sample was 0.738, and the average Hausdorff distance was 0.935.

CHAPTER 5

Federated Learning Improves Site Performance in Multi-Center Deep Learning Without Data Sharing

5.1 Overview

The disposition of healthcare data has generated significant interest in recent years. With the rapid expansion of the use of software-enhanced medical diagnostics, devices, and other interventions, access to clinical data has become critical to innovation. Clinicians and healthcare researchers facing this new data climate are forced to balance their professions ethical directives to "protect patient privacy in all settings to the greatest extent possible" and to "contribute to the advancement of knowledge and the welfare of society and future patients" [AMA17]. When the sharing of data is contemplated, ethics committees must evaluate the relative risks of unauthorized protected health information disclosure against the benefits of performing research and innovation using healthcare data.

An important contributor to the demand for healthcare data is the rapid advent of artificial intelligence (AI)-enhanced applications. For example, the field of medical image analysis has been driven forward in recent years by the advent of deep learning (DL). DL has enabled a wave of innovation in imaging decision support, with recent major results in the fields of ophthalmology [GPC16, QCB17, BCK20], dermatology [YCL17, Har18, EKN17], pathology [BPB20, EVJ17, COS18], and radiology [CGT18, MSG20].

A major limitation of the DL approach is the need for a large volume of training data that captures the full breadth of inputs on which the model is likely to be subsequently used. In the field of natural image processing, large-scale pooled datasets with over a million images captured by a variety of different cameras are commonly used [RDS15]. This large volume is required because deep learning models are primarily interpolators, not extrapolators that is,

---

The work described in this chapter was published as: Sarma KV, Harmon S, Sanford T, et al. Federated learning improves site performance in multicenter deep learning without data sharing. J Am Med Informatics Assoc. 2021. fdoi:10.1093/jamia/ocaa341.

they perform best when presented with inputs that are similar to the data that those models were trained on. This creates the need to ensure that models intended for widespread clinical use are exposed to heterogeneous data sources that capture the full breadth of the patient populations, clinical protocols, and data acquisition devices (i.e. scanners) that they will be used on.

However, medical imaging data in most cases is siloed within provider institutions, and as a result, assembling large-scale datasets traditionally requires the transfer of data between these silos. Such transfers present ethical and legal challenges around preserving patient privacy. As a result, very few public large-scale pooled medical image datasets exist. This has led to a challenge of generalizability for deep learning models in medical imaging research, which are often trained on single-institution datasets. Such models often suffer from poor performance when transferred to other institutions with differing protocols, equipment, or patient populations [YA18, DLR18]. As a result, there is a need for methods to enable the development of general models for clinical use, without requiring the creation of pooled datasets.

An alternative methodology to centralizing multi-center datasets is known as "distributed" learning [CBL18]. In this paradigm, data is not combined into a single, pooled dataset. Instead, data at a variety of institutions is used to train the DL model by distributing the computational training operations across all sites. One such approach is federated learning (FL) [RHL20, KMR20, LMX19, RCS20, SER20]. In FL, models are trained simultaneously at each site and then periodically aggregated and re-distributed. This approach requires only the transfer of learned model weights between institutions, thus eliminating the requirement to directly share data. However, a limitation of this approach is that no single model ever "sees" a complete picture of the universe of potential inputs during the training phase, thus placing pressure on the federated aggregation function to adequately distribute knowledge from each site into the model. Previous work has demonstrated the potential utility of FL for model training, generally using publicly available data to simulate multi-institutional training. However, works that examine the practical application of FL in radiological applications are still limited [RCS20, SER20]. Our work shows that FL can be reduced to practice using real-world private clinical data across multiple institutions, and that this approach creates a model that demonstrates improved generalizability both within the participating institutions, and with outside data.

In this work, we demonstrate the application of FL at three institutions: the University of California, Los Angeles (UCLA); the State University of New York (SUNY) Upstate

Medical University; and the National Cancer Institute (NCI). For this demonstration, we used the medical image analysis task of whole prostate segmentation, an initial step for MRI diagnosis of cancer and fusion-guided interventions. We demonstrate that FL training and aggregation is able to produce a model that learns general predictive weights applicable to each institution dataset, and demonstrates improved generalizability when applied to an external validation dataset.

5.2 Materials and Methods

5.2.1 Study Design

In this study, we use data collected retrospectively from each of our institutions to train and validate DL models to perform whole prostate segmentation on MRI. At no point during this study was private data transferred or shared across institutions. Instead, training on private data was done at that datas respective institution, and model weights were iteratively aggregated by a federated server and redistributed (Figure 5.1). After training, we evaluated the generalizability of each of the models using held-out testing sets from each institution, as well as an external challenge dataset.

5.2.2 Data Governance

One of the major challenges in multi-center DL studies is data governance. Our collaboration included one industry partner (nVIDIA, Inc.), two public universities (UCLA and SUNY Upstate), and one Federal institution (NCI). For this study, UCLA, SUNY Upstate and the NCI established a two-way agreement with nVIDIA to collaborate and share model weights, but no material transfer agreement to exchange protected or private data was required. All three academic institutions had IRB approval for review and image analysis, with written informed patient consent or waiver of patient consent.

5.2.3 Datasets and Preprocessing

Each institution retrospectively collected one prostate MRI from each of a cohort of 100 patients enrolled in an IRB-approved protocol studying the use of MRI for prostate cancer diagnosis (the "private datasets"). Axial T2 weighted (T2W) images of the prostate acquired at 3T were obtained for each patient. A ground truth whole prostate segmentation was

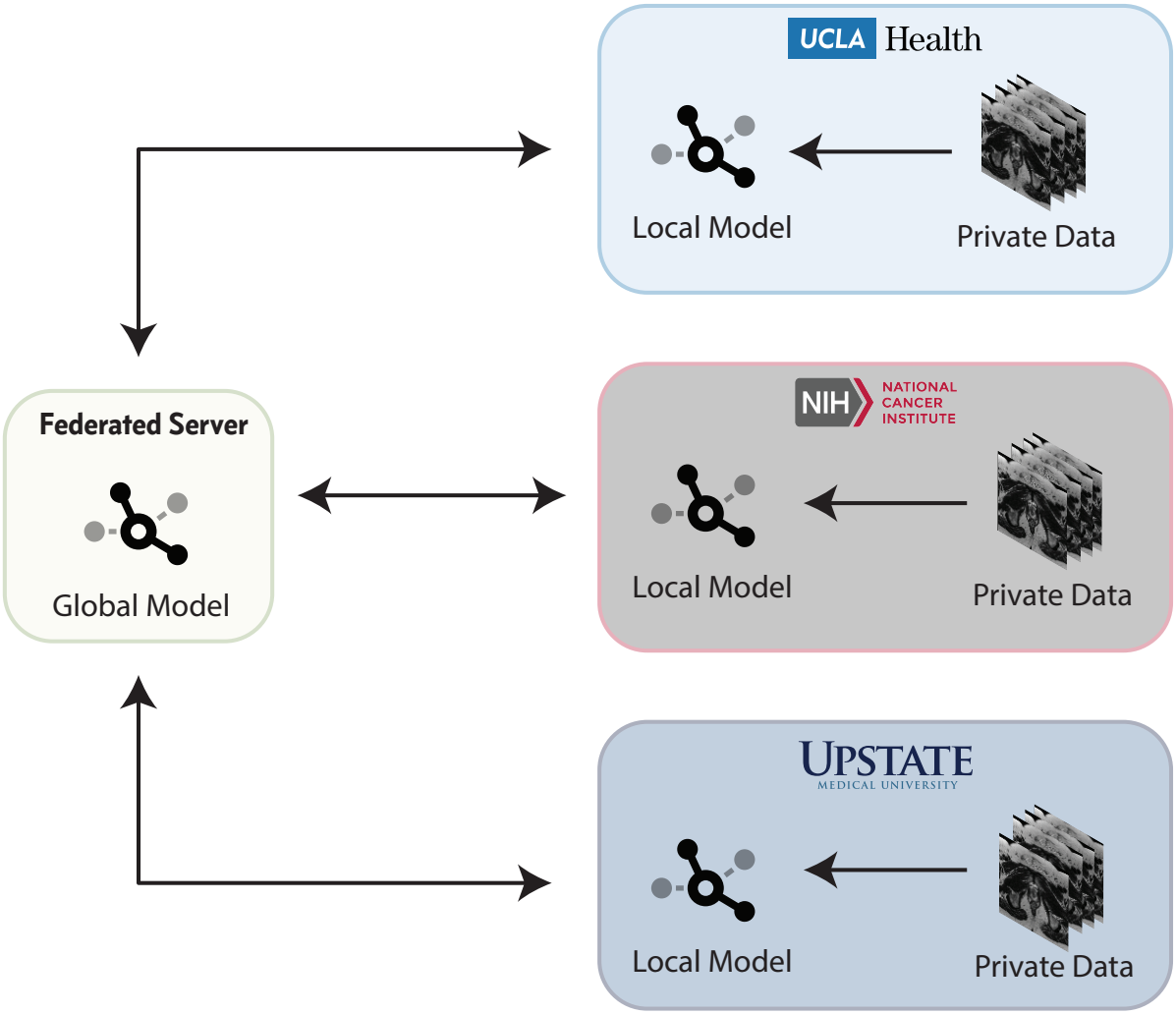


Figure 5.1: Federated Learning System Overview. Each participating institution (UCLA, NIH and SUNY Upstate) possesses a private dataset of prostate images. These images are used to train a local copy of the deep learning model using a local machine learning workstation. After each local training epoch, the local model weights are then sent to the federated server (a cloud-hosted application). Once weights are received from every institution, they are aggregated by the server, and the resulting weights are distributed to each institution. The local model is updated with the new weights, and the process restarts. At no point is any data shared between institutions; only model weights are transmitted.

produced for each patient by an expert clinician at each institution (radiologist or urologist ranging from 9-27 years of experience). Segmentations were performed under the standard manual and semi-automatic clinical methodologies in place at the individual institutions. In order to demonstrate broad generalizability, participating institutions intentionally made no effort to harmonize either the T2W acquisition protocol or segmentation methodologies. In addition, 343 axial T2W images of the prostate were obtained from the public SPIE-AAPM-NCI PROSTATEx dataset [AHD18] (the "challenge dataset"). These images were annotated with ground truth whole prostate segmentations by an expert clinician.

Each T2W image and annotation included in the study was resampled to an isotropic 1mm x 1mm x 1mm voxel size. The images were then converted to the NIFTI format [LM14] for training, and the intensity values within each image were normalized to zero mean and unit variance. Each of the private datasets was divided into a training set of 80 images and a held-out test set of 20 images.

5.2.4 Model Architecture and Data Augmentation

The 3D Anisotropic Hybrid Network [LXZ18] (3D AH-Net) was used as the DL model for this study, using an ImageNet [RDS15] pretrained ResNet50 backbone. This architecture is designed to perform well on anisotropic 3D volumes with higher slice spacing relative to in-plane voxel spacing. This network was chosen based on the image acquisition parameters (Table 5.2) for the NCI and SUNY datasets, which had high in-plane to slice spacing ratios (approx 1:10 vs 1:3 for UCLA).

Unlike the 3D U-Net, the 3D AH-Net does not use 3x3x3 convolutional kernels on input volumes in order to avoid issues caused by the anisotropy (i.e. poor alignment, loss of context, etc.). Instead, for every block, first a 2D multi-channel encoder is used on each slice, and then a 3D column encoder (1x1x3) is used on the encoded slice sets. The 2D multi-channel encoder is derived by pretraining a 2D multi-channel model (in this case ResNet50 on RGB ImageNet data), and then sets of three neighboring slices are fed into the pretrained network as channels (i.e. slice 0 = R, slice 1 = G, slice 2 = B, etc.). U-Net type skip connections are also used in the network.

The training metric was the soft Dice loss (Equation 5.1), and the Adam optimizer with validation metric-based learning rate decay was used for training. Real-time data augmentation was performed using the Deep Stacked Transformation [ZWY20] methodology. The DST approach augments data using serial "stacked" transformations from the following list:

Gaussian noise with kernel standard deviation between [0.1, 1.0], intensity scaling between [-0.1, 0.1], and contrast gamma correction with γ between [0.5, 4.5].

5.2.5 Training Strategy and Federated Model Aggregation

Each private training set of 80 images was split into five sets of 16 images each. Then, for each experiment, five sub-models were trained, each using one of the sets of 16 images as validation data, and the remainder as training data. The resulting sub-models were then combined into a single ensemble model outputting the mean of all five sub-models. The same cross-validation training sets were used for all experiments. A total of four training experiments were performed: one training run to develop a private model at each institution, and an additional training run to develop an FL model across all institutions.

All models were trained for 300 epochs. For the FL training, a cloud-based federated weight aggregation server ("federated server") was deployed by UCLA on a secure Amazon Web Services instance using the Clara application framework (nVIDIA, Inc.). Bilateral websocket connections (over SSL) were established during training between each institutions training server and the cloud-based aggregation server. After each training epoch, model weights and validation metrics from each institution for that epoch were sent to the server, where an aggregation function [MMR17] was used to combine them into a single set of model weights which were then sent back to each institution. These weights were then used as the basis for the next training epoch, and the process was repeated until all epochs had elapsed. The aggregation function used a weighted average of input models to produce the combined model. Each institutions input was weighted based on the validation metric (mean Dice coefficient) from the most recent training epoch reported by the corresponding institution on the validation set for that fold. The FL training framework was implemented using the nVIDIA Clara Train SDK [NVI], and training at each site was performed using single nVIDIA GPUs.

5.2.6 Statistical Analysis

Each of the ensemble models was evaluated at each institution using its held-out test set, producing an evaluation for each model at each institution. In addition, each of the models was also evaluated on the challenge dataset. The evaluation metric used to compare segmented volumes was the Dice coefficient function as denoted in Equation 5.1, where S_{DL} is the segmentation of a deep learning model and S_m is the manual segmentation. The value

of the coefficient can range between 0 (no overlap) and 1 (perfect overlap).

$$\text{DICE}(S_{DL}, S_m) = \frac{2|S_{DL} \cap S_m}{|S_{DL}| + |S_m} \tag{5.1}$$

The mean Dice coefficient was then compared for each model on each of the individual private test sets, as well as the overall mean Dice coefficient for each model (across all of the test set data). The mean Dice coefficient was also separately computed for each model on the challenge dataset. Finally, 2-sided paired t-tests were used to compare the mean Dice coefficients from each private model to the FL model, for both the "combined" private test set and the held-out challenge dataset.

5.3 Results

Patient and imaging characteristics of the three private datasets are shown in Tables 5.1 and 5.2. Tables 5.3 and 5.4 show all experimental results. The private models performed well on their own private test sets (Dice coefficient range: 0.883-0.925) but had diminished performance on the other private test sets (Dice coefficient range: 0.575-0.887). This led to overall mean Dice coefficients between 0.745 and 0.833 for the private models.

In comparison, the FL model performed well on all three test sets. The FL model exhibited private test set mean Dice coefficients between 0.880 and 0.920, yielding an overall result of 0.895. The statistical analysis using 2-sided paired t-tests demonstrated that the FL model was significantly superior to any of the private models ($p < 0.001$ for all comparisons).

The private models exhibited varied performance on the challenge dataset (Dice coefficient range: 0.812-0.872). The generic FL model outperformed each of the private models, with an overall mean Dice coefficient of 0.889. The statistical analysis again demonstrated that the FL model was significantly superior to any of the private models ($p < 0.001$).

Table 5.1: Patient Demographics.

	Private Test Set Institution		
	NCI	SUNY	UCLA
Age (years)	66 (47-83)	66 (49-81)	65 (50-83)
Prostate Size (cc)	65.5 (21.7-231)	72.9 (26.8-210)	52.1 (15.8-147)

Table 5.2: Image Acquisition Parameters.

	Private Test Set Institution			
	NCI		SUNY	UCLA
	with endorectal coil (n=50)	without endorectal coil (n=50)		
Vendor(s)	Philips Medical Systems		Siemens	Siemens
Field Strength	3T		3T	3T
In-plane resolution	0.273mm	0.352mm	0.625mm	0.664mm
Slice thickness	3mm	3mm	3mm	1.5mm
TR (ms)	4775	3686	5500	2230
TE (ms)	120	120	136	204

Table 5.3: Model Evaluation Results: Private Test Sets.
** Significantly lower than FL model ($p < 0.001$).

		Private Test Set Institution			
		NCI (n=20)	SUNY (n=20)	UCLA (n=20)	Overall (n=60)
Private Models	NCI	0.925 ± 0.016	0.854 ± 0.050	0.720 ± 0.165	0.833 ± 0.131**
	SUNY	0.887 ± 0.027	0.906 ± 0.018	0.768 ± 0.064	0.854 ± 0.074**
	UCLA	0.777 ± 0.102	0.575 ± 0.177	0.883 ± 0.069	0.745 ± 0.178**
FL Model		0.920 ± 0.029	0.880 ± 0.034	0.885 ± 0.032	0.895 ± 0.036

Table 5.4: Model Evaluation Results: Public Test Set.
** Significantly lower than FL model ($p < 0.001$).

		ProstateX (n=343)
Private Models	NCI	0.872 ± 0.062**
	SUNY	0.838 ± 0.043**
	UCLA	0.812 ± 0.136**
FL Model		0.889 ± 0.036

5.4 Discussion

We sought to demonstrate that data-distributed learning can be successfully operationalized across multiple institutions with real patient data using federated learning, and that the resulting model would gain the benefit of having learned from each of the private datasets without ever needing to transfer or pool data at a single location.

Since no transfer of protected health information (or even de-identified health information) was required, we were able to address the privacy and data governance limitations inherent to multi-center studies through the use of simplified two-way collaboration agreements, rather than requiring the negotiation of a complex four-way collaboration and material transfer agreement that would have been required if data was shared across institutions. This allowed for expedited ethics and compliance reviews because of the minimal risk posed by the FL paradigm, and enabled us to be assured that our patients privacy was maintained.

The FL model that we trained performed well across all of the private datasets, yielding an overall performance level that was significantly better than that of any of the private models alone. This suggests that the FL model was able to benefit from the advantage of learning important institution-specific knowledge through the FL aggregation paradigm, without requiring any individual training site to "see" the full breadth of inputs.

Additionally, our results showed that the FL model performed significantly better than any of the individual private models on the held-out challenge dataset, suggesting that the model also attained the expected advantages inherent in training with more data through the FL aggregation method, even though the full dataset was not seen at any single training site.

Our work does have limitations. In this work, we did not attempt to address the potential for an inside actor (i.e. one of the participating institutions) to attempt to recover the underlying patient data through a model inversion attack on the trained weights shared during federated learning. Future enhancements to the federated approach could include the addition of calibrated distortion to shared model weights in order to suppress the potential for inversion. However, we believe the method we demonstrate in this paper significantly better protects the privacy of patients than the current standard of direct sharing of data. In addition, though model inversion is a technical risk that cannot be ruled out, we empirically believe that the practical risk of inversion outside of crafted malintent on the part of study designers to be low due to the weight averaging scheme in place. Finally, we note that the sharing of trained model weights is an accepted practice within healthcare [SR20, LM20], and

in the worst case our method is no less secure as only model weights are ever transmitted.

Secondly, the task we used (prostate segmentation on T2-weighted MRI) is relatively simple and all private models achieved high performance on their own institutional datasets. Thus, we were unable to demonstrate the expected benefit that an FL-trained model would significantly outperform a single-site-trained model on that single sites data. In addition, because we used similarly sized private datasets at each institution, we did not explore the potential in varying the federated model aggregation methodology, which could be extended to differentially weight model weights from institutions based on data quantity, quality, or other metrics. Thirdly, adding additional institutions to the federation may present new challenges in heterogeneity of imaging data quality, governance, intellectual property, and model generalizability. In order to ensure that the FL model performs well at each institution in a large federation, it may be necessary in future work to explore adding an additional private fine-tuning step at each institution, though care must be taken to avoid losing generalizability through overfitting.

5.5 Conclusion

The power of federated learning was successfully demonstrated across three academic institutions using real clinical prostate imaging data. The federated model demonstrated improved performance across both held-out test sets from each institution and an external test set, validating the FL paradigm. This methodology could be applied to a wide variety of DL applications in medical image analysis, and merits further study to enable accelerated development of DL models across institutions, enabling greater generalizability in clinical use.

CHAPTER 6

Developing Patient-Level Predictive Models Using Pretrained Voxel-Level Feature Extractors for Prostate mpMRI

6.1 Overview

Prostate cancer is the second most frequent cancer diagnosis and the fifth leading cause of death for men worldwide [SMJ15]. However, there is a significant discrepancy between the incidence (26% of new diagnoses in men) and mortality (9% of cancer deaths in men) due to the heterogeneous nature of prostate cancer [SMJ15]. This discrepancy creates the need for risk stratification to avoid subjecting patients with indolent cancers to unnecessary interventions, which can be the cause of significant morbidity in cost.

This need is currently addressed through systematized grading systems that assign scores to patients based on the clinical impression of an appropriate diagnostic. For patients who undergo prostate biopsy, the Gleason grading system [GMA74] (and more recently, the ISUP grade group system [EEA15]) is used to classify the patient's prostate cancer based on histopathological examination of the samples retrieved. This pathologic diagnosis is the most accurate, definitive diagnosis and allows for the best prognostication. However, it does require an invasive biopsy procedure (which, as all invasive procedures, is subject to potential morbidity). Additionally, there is a risk of sampling error. Because a limited number of biopsy cores can be obtained in any given procedure, the cores retrieved may not pass through the most severe areas of cancer within the prostate, resulting in an inaccurate patient-level risk classification [RSR21]. Though advanced targeted biopsy techniques (such as MR-ultrasound targeted biopsy [MYN13]) improve the probability that a representative biopsy will be sampled, a significant misclassification risk still remains.

These two major limitations of pathological diagnosis via biopsy (the invasive nature of the diagnostic and the risk of sampling error) motivate the desire for enhanced non-invasive approaches for risk assessment. Reliable imaging-based diagnostics for prostate cancer could

enable more accurate stratification and the avoidance or deferral of invasive diagnostics for some patients, reducing morbidity and cost.

The clinically accepted approach for mpMRI-based risk stratification is the PI-RADS v2 scoring system. This system enables radiologists to report their impressions of imaging using a five-point grading system which correlates with Gleason grade [BWV16]. However, this approach suffers from a major limitation: a prostate-trained abdominal radiologist is generally required to review images and generate a PI-RADS score, which eliminates the possibility at many centers due to a lack of availability. Additionally, though use of this score enables some patients to defer biopsy, the population of patients who proceed to biopsy based on high PI-RADS score still includes patients with low risk cancer.

The development of a machine learning-based predictive model for patient-level risk scores (such as PI-RADS or ISUP grade group) is thus of considerable interest. A high performance patient-level prediction (PLP) model for PI-RADS could enable the use of prostate mpMRI even in the absence of a specially trained abdominal radiologist, thus expanding access to noninvasive diagnosis. Additionally, a PLP model for ISUP grade group could have the potential to produce a higher quality risk stratification than the PI-RADS system, enabling the use of "watchful waiting" rather than invasive intervention for a broader group of patients.

Traditionally, such a model would be trained by training a network to locate the areas of suspicion that would be located by a radiologist, an approach that is taken by prior work in the field [TLR17, LTG17, SLT18]. However, this approach requires the use of training data that has been annotated by a highly qualified radiologist. This poses a number of challenges. In the traditional clinical workflow, radiologists do not attempt to annotate an entire visualized defect within an MRI volume. Instead, the areas of maximal suspicion are annotated in order to maximize the probability that a biopsy obtained from that area will be representative. Though clinically useful, such annotations are of limited utility in training a machine learning model because they falsely represent abnormal-appearing areas on an MRI as "normal" since they not annotated as suspicious. This false-negative ground truth problem can limit the capability of a network to learn appropriate features. To address this problem, prior studies have had trained radiologists re-annotate manually, an expensive and time-consuming process.

If instead an entire network could be trained using simply the unannotated mpMRI images as inputs and the risk score (PI-RADS or ISUP grade group) as targets, the need for radiologist re-annotation of training data would be eliminated, unlocking substantial additional data for use. However, attempting to directly train a PLP model in this manner

is challenging because of the "vanishing gradients" problem [LXL19]. This problem, which generally affects deep learning models due to the need to propagate error information through the entire network, is partially alleviated when voxel-wise supervision is provided (i.e. a region of interest mask of the volume), and exacerbated when limited supervision is provided (i.e. a single patient-level risk score). Several approaches have been proposed in the literature to address this issue, such as the use of short and/or long residual connections [HZR16].

Here, we propose an approach to successfully train deep learning-based PLP models using domain-specific voxel-level pretraining. Specifically, we demonstrate both PI-RADS and Gleason PLP models developed by first pretraining a fully convolutional voxel-level network on one of three different voxel-level targets in order to generate a deep prostate feature encoder, and then using that encoder to train a fully connected attention network to produce the final PLP model. We also demonstrate how this approach can facilitate the addition of non-imaging clinical variables (such as PSA) into the model to improve the final result.

6.2 Materials and Methods

6.2.1 Data

Our internal clinical population for this study consists of 1,785 MRI studies collected from 1,534 patients who underwent transrectal ultrasound-MRI fusion biopsy (TRUS biopsy) using the Artemis guided biopsy system (Eigen Systems) between 2010 and 2018 at our institution using a standardized protocol and 3T magnet (Trio, Verio, or Skyra, Siemens Healthcare). As part of the protocol, prostate MRIs were contoured in a two-part process. First, the attending radiologist for the case (the attending radiologists for the patients included in this study each had between 10-27 years of experience) created a prostate contour using the DynaCAD Prostate image analysis platform as part of the routine clinical workflow. The radiologist then contoured any regions of interest (ROIs) for targeted biopsy sampling. These regions of interest were selected based on either the PI-RADS v2 criteria [BWV16] or, before the development of the PI-RADS v2 criteria, the UCLA score [MBH19], a comparable imaging-based risk stratification score. These contours were then used by a technician to re-contour the prostate on the Profuse (Eigen Systems) platform in order to enable use with the Artemis biopsy system, as DynaCAD segmentations cannot be directly imported for use on the Artemis.

We retrospectively collected 3D T2-weighted (T2W) images, apparent diffusion coefficient (ADC) maps, and prostate and ROI contour sets from these studies. Imaging volumes were collected from our institutions picture archiving and communication system (PACS) and corresponding contours were collected from the Profuse image analysis platform. T2 images were acquired clinically using the spatial and chemical-shift encoded excitation (SPACE, Siemens Healthcare) protocol.

Full-text radiology reports associated with the studies were collected from our institution's electronic health record (EHR) system, and histopathological results from each obtained biopsy core were acquired from the laboratory informatics system (LIS). Data were collected from every available study for each patient seen at our institution during the study period. Studies were excluded from retrieval if the T2 image, ADC map or contour was missing from PACS or corrupt, if the full-text radiology report was not retrievable, or if the image exhibited a protocol deviation, such as a variance in FOV or resolution.

After retrieval of the studies, clinical data was extracted manually from the full-text radiology reports and pathology results. This included the study quality assessment, prior prostate cancer treatment status, presence of endorectal coil, overall PI-RADS or UCLA score and sequence-level scores, patient age and biopsy Gleason scores. Studies were excluded if the prior prostate treatment had occurred, such as transurethral resection of the prostate (TURP), laser interstitial thermal therapy (LITT), or radiation therapy (RT), if an endorectal coil was used, or if study quality limitations were noted (such as motion artifacts or prosthesis susceptibility artifacts). A total of 1,103 studies were included from 999 patients, and 682 studies were excluded. All data was used for this work under the approval of the University of California, Los Angeles (UCLA) institutional review board (IRB# 16-001087).

6.2.2 Preprocessing

6.2.2.1 MRI Volumes

T2W and ADC volumes were preprocessed using a common pipeline. Initial preprocessing was done in Python, primarily using the SimpleITK toolkit [LCI13]. First, N4ITK bias field correction [TAC10] was performed. Then, T2W volumes were resampled to isotropic voxel spacing using a B-spline approach, and ADC volumes were resampled to the same voxel spacing as the resulting isotropic T2W volume; these steps were based on preprocessing done in previous work [CLR19, TAC10, SSR18, GMV15].

After initial preprocessing, we applied a previously used interquartile range (IQR)-based intra-image normalization [SRD21] to address the relative nature of MR image intensity values (both within and between institutions). Each image was normalized to the image-level IQR (calculated from the central 128x128 column of the volume) and then values were clipped between two IQRs below the first quartile and five IQRs above the third quartile, in order to eliminate outlying values created by imaging artifacts. Volumes were then cropped to the central 128x128 in the x, y plane, which includes the prostate.

6.2.2.2 Clinical Variables

For every study, the maximum PI-RADS v2 score for any region of interest was used to label the study. Similarly, for every biopsy procedure, assigned Gleason grades were converted to ISUP grade groups, and the maximum grade group from the procedure was used the label the study. PI-RADS v2 scores were dichotomized as ≥ 4 and ISUP grade groups were dichotomized as ≥ 2.

Based on the UCSF-CAPRA risk scoring system [CPE05], age was dichotomized as either <50 or ≥50 years old, and PSA was categorized into five ordinal categories (Table 6.1).

Table 6.1: UCSF-CAPRA PSA Categories.

PSA level	Points
PSA ≤ 6	0
$6 <$ PSA ≤ 10	1
$10 <$ PSA ≤ 20	2
$20 <$ PSA ≤ 30	3
PSA > 30	4

6.2.3 Data Augmentation and Model Architecture

We adopted a multi-stage model approach motivated by [ZPN21] for this effort in order to overcome the challenges posed by limited patient-level supervision. In the first stage of training, an end-to-end fully convolutional neural network (FCN), the "base model," based on the 3D U-Net [CAL16] was trained using one of three voxel-level ground truth options: autodidactic models genesis, organ segmentation, or cancer region of interest. Once all three FCNs were trained, the second stage of model development began. For the second stage PLP models, the initial 3D U-Net FCN was truncated to the bottom of the encoder hierarchy, and the output of this base encoder was then fed into the second stage fully connected model.

This second stage was then trained using one of two patient-level dichotomized ground truth options: maximum PI-RADS v2 score or maximum ISUP grade group, with varying levels of "freezing" applied to the encoder layers of the base model. All base and PLP models were trained by stacking the T2W and ADC volumes as input channels.

6.2.3.1 Autodidactic Models Genesis

We used the generic autodidactic models ("Models Genesis") approach [ZSS19] as one of our pretrained model creation methodologies. The models genesis approach is a variation on the traditional autoencoder approach. Instead of training the model to recreate input as output, a Models Genesis autodidactic model is trained to restore an original image from a perturbed version of the image. The random perturbation applied for our work included flipping on the x and y axes, local pixel shuffling (in which random small windows were selected and pixel shuffled randomly within each window), nonlinear intensity transformation (in which intensities are resampled using a nonlinear monotonic intensity transformation), in-painting (in which randomly selected small rectangular patches are masked to fixed random intensity), and out-painting (in which random rectangular edge regions are cropped out of the image). This overall approach is intended to train a model to develop an understanding of appearance, texture, geometry, and context of the input distribution in order to create an informative pretrained model without requiring semantic labeling of the input space.

6.2.3.2 Model Architecture

The base model used for this study was the 3D U-Net [CAL16]. For all experiments, the network was configured with four encoder levels, three decoder levels, a ReLU transfer function and batch normalization following every convolutional layer, and a softmax output layer. The number of features per encoder level was 64, 128, 256, and 512, with the encoder levels reducing the number of features by a factor of two at each level until the final set of 64 feature maps, which was then fed into a final convolutional layer and a sigmoid output layer. For the autodidactic model, two output volumes were produced (one for each input), and for the segmentation and ROI models, one output volume was produced.

The PLP model used a variable architecture with multiple optimized hyperparameters and a shared basic structure. First, the 512-feature output of the final encoder of the base model was average pooled across all dimensions, resulting in a 512-element encoded vector. Optionally, one or two soft attention modules [SOS18] were also used, modeled after [LZG20].

These attention modules were fed the 128 feature and/or 256 feature encoder level outputs as inputs and the 512 feature encoder level outputs as gating channels, and the resulting softmax output was then also pooled by sum into 128 or 256 elements respectively.

The 512 (and optionally 256 and 128) element vectors were then concatenated, and optionally, the scalar variables for age and/or PSA were also appended to this intermediate vector. The resulting intermediate output was then optionally fed through a batch normalization and/or dropout layout before being fed through a fully connected layer to produce the final output.

The complete model architecture is depicted in Figure 6.1.

6.2.3.3 Hyperparameters

A number of configurable hyperparameters were used in the patient-level model architecture in order to enable experimentally investigating the optimal PLP model design. Configurable hyperparameters included the inclusion of age and PSA as model inputs, the use of one or both attention blocks, the use of a final batch normalization and/or dropout layer, and the depth to which the base model was frozen. A table of all the configurable model and training hyperparameters is displayed in Table 6.2.

6.2.3.4 Data Augmentation

For all model training in this study, real-time augmentation using the Batchgenerators package was performed [IJW20]. Three augmentation transformations were used: 1) random elastic deformation using a thin plate spline, 2) random rotation in the range $[-\pi/8, \pi/8]$ in the axial plane, and $[-\pi/4, \pi/4]$ along the axis, and 3) random mirroring across the y-axis. Data augmentation was performed once for every training example for each epoch (i.e., for each epoch, new augmented data was generated in real time and used for training).

6.2.4 Training and Hyperparameter Optimzation

6.2.4.1 Hyperparameter Optimization

Extensive hyperparameter optimization was found in order to characterize the relative impact of various configurations on final performance, and to obtain the best final performance results. Optimzation was performed using the Ray Tune framework and Pytorch, using

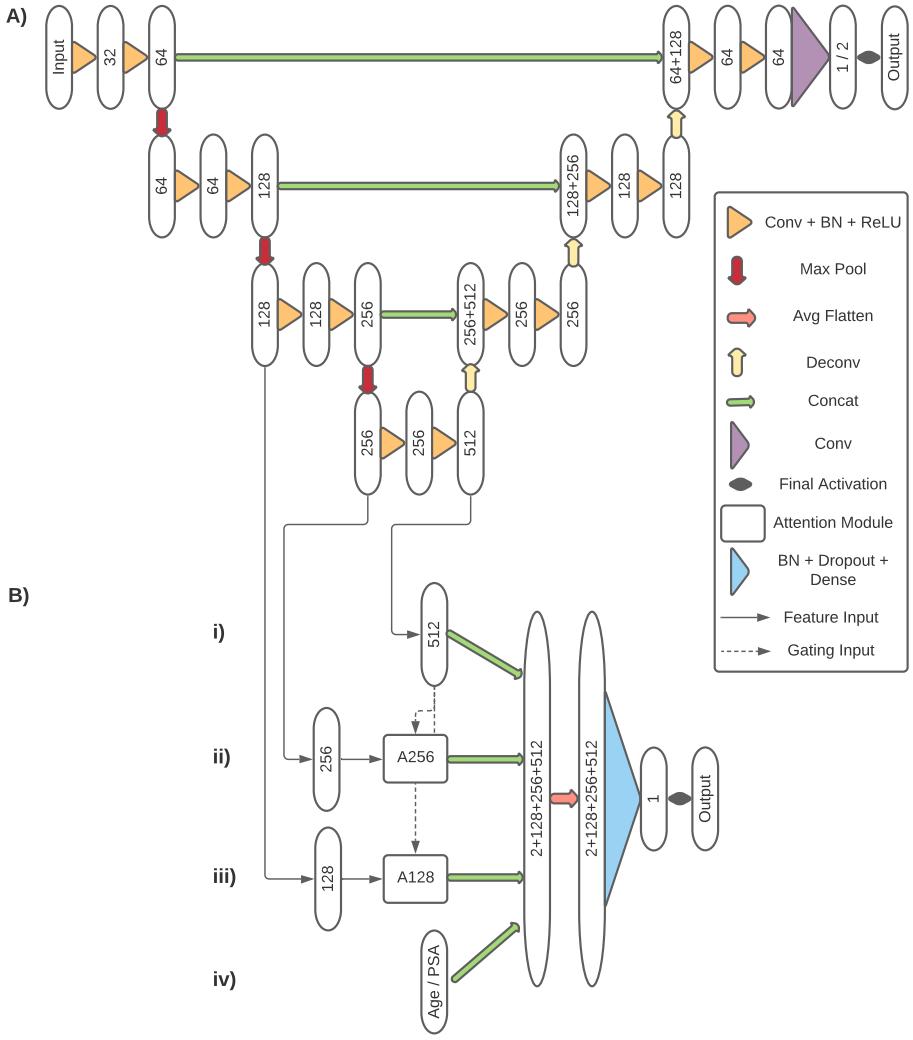


Figure 6.1: PLP and Base Model Architecture. A) The base model 3D U-Net architecture used in this work. The model was configured for either single or double channel output, depending target. Numbers within the ovals represent the number of feature maps at that layer. Connections represent network operations, such as 3D convolution ("Conv"), max pooling ("Max Pool"), 3D transposed convolution ("Deconv"), feature map concatenation ("Concat"), batch normalization ("BN"), rectified linear unit activation ("ReLU"), 3D average flattening ("Avg Flatten"), attention modules, and dropout. B) The PLP model architecture used in this work. The PLP model shares encoding layers with the base model, with feature outputs from the 512, 256 and 128 feature encoder layers becoming feature inputs for the PLP model. Depicted model is with all hyperparameters "on." Model column i) was always present and represented the most encoded 512 feature layer. Columns ii) and iii) represent the 256 and 128 feature attention modules; encoder outputs were used as feature input, and the 512 feature layer was used as gating input. Column iv) represents the clinical variable inputs of Age and PSA. All used columns had their outputs concatenated before passing through optional BN and dropout layers and a final densely connected layer before final activation and output.

a random hyperparameter space sampling approach. For space sampling, the HyperOpt random search algorithm was used, which iteratively adjusts the sampling distribution based on performance results of experimental trials to increase the weight of areas of the parameter space more associated with better performance [BB12]. Individual hyperparameter trials were scheduled and run using the async successive halving algorithm (ASHA) scheduler [LJR18], which asynchronously runs trials while pruning poor performers early in order to minimize wasted time while training. Hyperparameter optimization was configured to run a total of 100 trials.

The hyperparameter space was initialized using the random distributions depicted in Table 6.2.

6.2.4.2 Model Training and Evaluation

All training and evaluation was done using the PyTorch framework on NVIDIA Quadro 8000 GPUs with 48 GB of onboard memory. Mixed-precision training using the NVIDIA Accelerated Mixed Precision (AMP) was used at optimization level O2, consisting of 16-bit model weights and inputs, 32-bit master weights and optimizer parameters, and dynamic loss scaling.

Base model training was performed using the Adam optimizer with learning rate 10^{-5} and either the soft Dice loss function (for the segmentation model), the focal loss function (for the ROI model [LGG18]), or the mean squared error (MSE) loss (for the autodidactic Models Genesis model). Each epoch consisted of training on a full dataset comprised of one augmented sample generated for every original input sample, and training was performed for 100 epochs without early stopping after 50 epochs of no improvement of validation loss. For base model training, data was split in a 80% train, 20% validate configuration.

PLP model training was performed using the AdaBoundW optimizer [LXL19] with initial learning rate (LR), final LR and weight decay configured as hyperparameters. The binary cross-entropy (BCE) loss was used for optimization, and the area under the receiver operating characteristic curve (AUC) was used as the evaluation metric.

For PLP model training, data was split in a 60% train, 20% validate, and 20% test configuration. Optimal hyperparameter configurations for each choice of base model were selected using the validation AUC, and final results are reported using the test AUC.

Table 6.2: Hyperparameter Options and Distributions. Hyperparameter search was intialized to a uniform random distribution of the parameters as described by the table. Distributions used were *Choice*, which randomly selected one of the arguments with equal probability, *Uniform[a, b, c]*, which randomly sampled the space between a and b in increments of c, and *LogUniform[a, b, c]* which randomly sampled the log-transformed space between a and b in increments of c.

Hyperparameter	**Distribution**
Base Model	Choice[Autodidactic, Segmentation, ROI]
Use PSA	Choice[True, False]
Use Age	Choice[True, False]
Use Attention Module (128)	Choice[True, False]
Use Attention Module (256)	Choice[True, False]
Use Final Batch Normalization	Choice[True, False]
Use Final Dropout	Uniform[0, 0.5, 0.1]
Base Model Encoder Freeze Depth	Choice[1, 2, 3, 4]
Initial Learning Rate	LogUniform[10^{-5}, 10^{-1}, 5 * 10^{-6}]
Final Learning Rate	LogUniform[10^{-5}, 10^{-1}, 5 * 10^{-6}]
Weight Decay	LogUniform[10^{-6}, 10^{-3}, 5 * 10^{-7}]

6.3 Results

6.3.1 Dataset Characteristics

1,103 studies were included as part of the analysis dataset, including 225 in the validation set, 215 in the held-out test set, and the remainder in the training set. The mean age of included patients was 64.6 ± 7.3. The mean PSA of included patients was 8.8 ± 9.4. Maximum PI-RADS v2 scores were distributed as follows: 1: 1 study (<1%), 2: 21 (2%), 3: 409 (37%), 4: 373 (34%), 5: 299 (27%). ISUP grade groups were distributed as follows: 1: 509 studies (46%), 2: 303 (27%), 3: 128 (12%), 4: 70 (6%), 5: 93 (8%). Characteristics of the training, validation and test sets are summarized in Table 6.3.

Table 6.3: Dataset Characteristics. Summary of characteristics of the dataset, including breakdown of training, validation and test sets. PSA Group refers to the categorization described in Table 6.1. ISUP GG = ISUP Grade Group.

Characteristic	Training	Validation	Test	Overall
n	663	225	215	1103
Age	64.5 ± 7.3	64.5 ± 7.2	65.2 ± 7.7	64.6 ± 7.3
Age ≥ 50	650 (98%)	221 (98%)	212 (99%)	1083 (98%)
PSA	9.0 ± 10.5	8.5 ± 7.5	8.4 ± 7.2	8.8 ± 9.4
PSA Group				
0	284 (43%)	106 (47%)	101 (47%)	491 (44%)
1	233 (35%)	63 (28%)	63 (29%)	359 (33%)
2	110 (17%)	41 (18%)	41 (19%)	192 (17%)
3	11 (2%)	10 (4%)	5 (2%)	26 (2%)
4	25 (3%)	5 (2%)	5 (2%)	35 (3%)
PI-RADS v2				
1	0 (0%)	1 (<1%)	0 (0%)	1 (<1%)
2	16 (2%)	4 (2%)	1 (<1%)	21 (2%)
3	242 (37%)	86 (38%)	81 (38%)	409 (37%)
4	216 (33%)	76 (34%)	81 (38%)	373 (34%)
5	189 (29%)	58 (26%)	52 (24%)	299 (27%)
ISUP GG				
1	306 (46%)	111 (49%)	92 (43%)	509 (46%)
2	177 (27%)	60 (27%)	66 (31%)	303 (27%)
3	82 (12%)	28 (12%)	18 (8%)	128 (12%)
4	46 (7%)	9 (4%)	15 (7%)	70 (6%)
5	52 (8%))	17 (8%)	24 (11%)	93 (8%)

6.3.2 Base Model Pretraining

Pretrained base models for the autodidactic, segmentation and ROI targets were successfully trained using T2 and ADC volumes. The autodidactic model achieved a minimum validation MSE loss of 0.0001 at epoch 81. The segmentation model achieved a validation soft Dice coefficient of 0.895 at epoch 100, in line with our previous work [SRD21]. The ROI model achieved a validation soft Dice coefficient of 0.627, with early stopping at epoch 20.

6.3.3 Hyperparameter Optimization

Hyperoptimization was successfully performed to select optimal hyperparameters configuration for the PI-RADS v2 and ISUP Grade Group ("Gleason") PLP models, with 100 trials for each of the two targets. An overview of all 200 trials, broken down by target and base model, is available in Figure 6.2. The highest performance configurations (by maximum validation AUC) for PI-RADS v2 exhibited test set AUCs of 0.606 (autodidactic), 0.618 (segmentation), and 0.621 (ROI). The highest performance configurations for ISUP Grade Group exhibited test set AUCs of 0.692 (autodidactic), 0.696 (segmentation), and 0.702 (ROI). A summary and comparison of configurations for the best performance models is available in Table 6.4. All three base models produced PLP models with approximately equivalent maximum performance for both PI-RADS v2 and ISUP Grade Group. However, the ROI model performed slightly better by trial mean for PI-RADS v2, and the prostate segmentation model performed slightly better by trial mean for ISUP Grade Group (see Figure 6.2).

Due to the adaptive probabilistic sampling strategy, it is not possible to rigorously evaluate the impact of individual hyperparameters on performance. However, based on evaluation of trial means, hyperparameters that appear strongly associated with better performance included the use of PSA (for both PI-RADS v2 and ISUP Grade Group, Figure 6.3), and *not* using a final batch normalization layer (for both targets, Figure 6.4). The use of the 128 feature attention module appeared to have slightly positive impact for PI-RADS v2 and negative impact for ISUP Grade Group (Figure 6.5), and the use of the 256 feature attention module did not appear to have significant impact (Figure 6.6). The use of age appeared to have little impact for PI-RADS v2 and a negative impact for ISUP Grade Group (Figure 6.7).

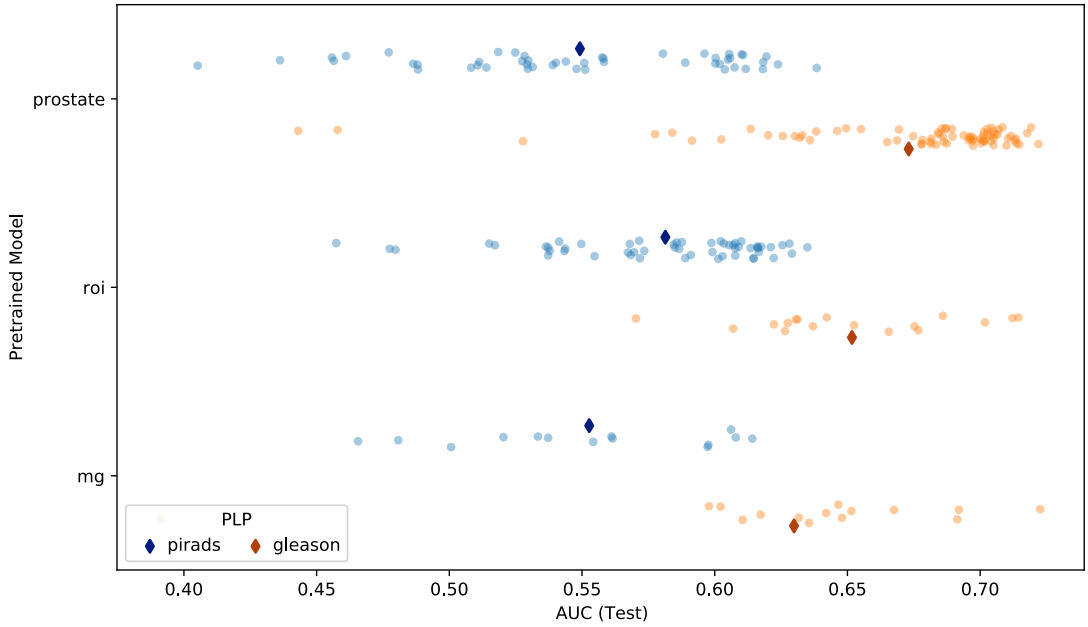


Figure 6.2: Trial Test AUCs by Base Model. Individual trial test AUC values, by target (color) and base model (row). Each point represents a single trial test AUC. Solid diamonds represent mean values across the row. Gleason (ISUP Grade Group) models generally performed better than PI-RADS models. For Gleason models, the prostate segmentation base model had the most positive impact on test AUC, though the highest performing models across all three base models are similar. For the PI-RADS models, the ROI base model had the most positive impact, but again the highest performing models were similar across all three base models, with slightly poorer performance from the MG model.

Table 6.4: Optimal Hyperparameter Configurations for PI-RADS v2 and ISUP Grade Group PLP Models. Best hyperparameter configurations for PI-RADS v2 and ISUP Grade Group PLP models (by maximum validation AUC). Att128 = 128 feature attention module, Att256 = 256 feature attention module, BN = batch normalization.

	PI-RADS v2			ISUP Grade Group		
Base Model	AD	Seg	ROI	AD	Seg	ROI
Use Age	Yes	Yes	Yes	Yes	Yes	Yes
Use PSA	Yes	Yes	Yes	Yes	No	Yes
Use Att128	Yes	Yes	Yes	No	No	No
Use Att256	No	Yes	Yes	Yes	Yes	Yes
Use BN	No	No	No	No	No	No
AUC	0.606	0.618	0.621	0.692	0.696	0.702

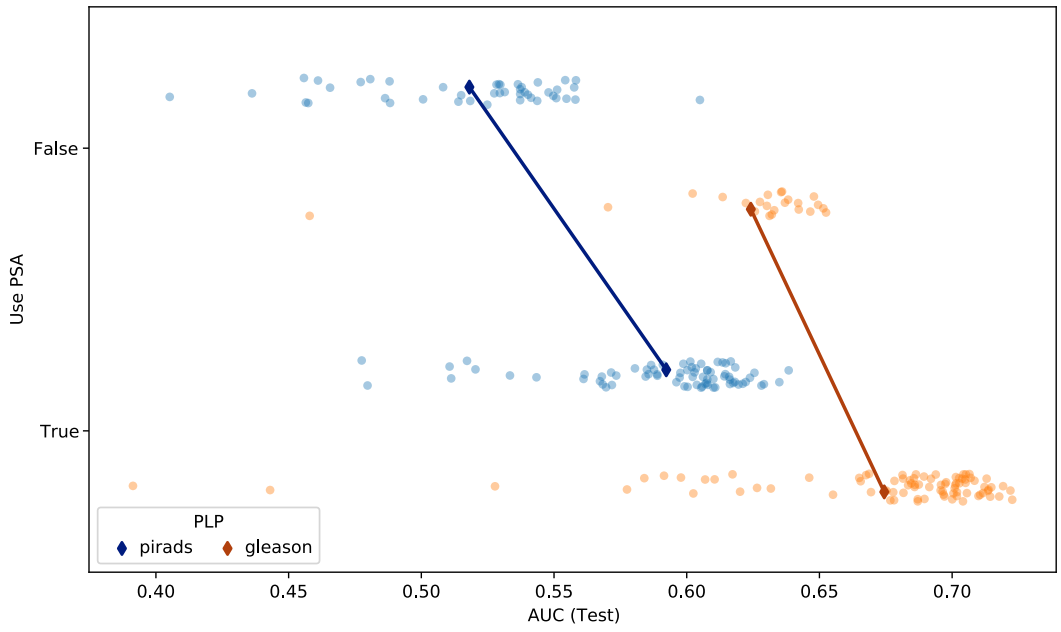


Figure 6.3: Trial Test AUCs by Use PSA. Individual trial test AUC values, by target (color) and Use PSA hyperparameter (row). Each point represents a single trial test AUC. Solid diamonds represent mean values across the row, and solid lines connect diamonds for the same target across rows. For both targets, models that used PSA generally performed better than models that did not, with both the mean and best case models with PSA outperforming those without.

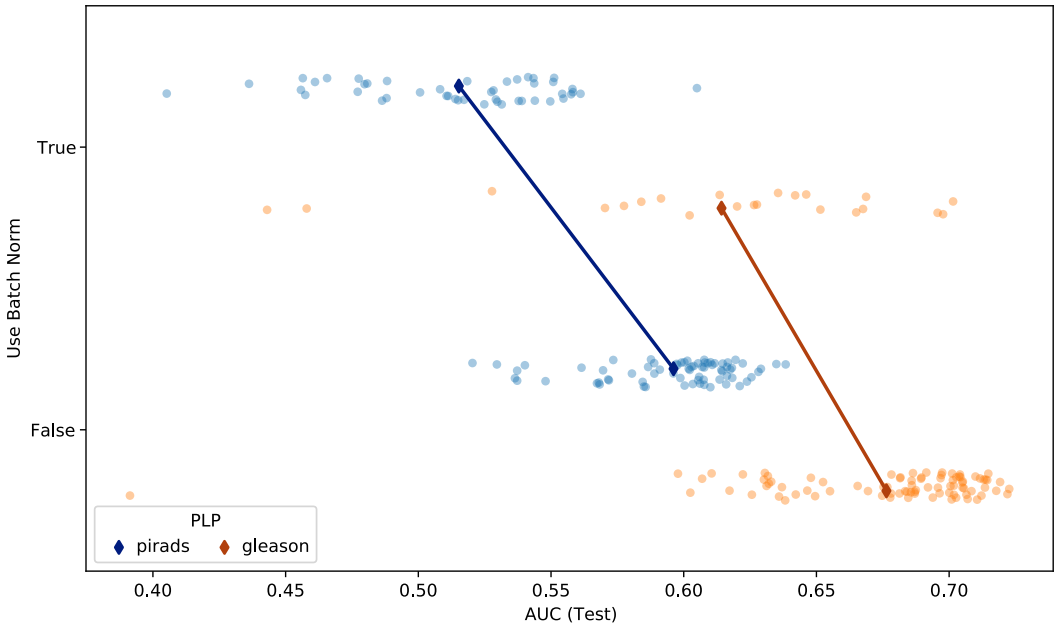


Figure 6.4: Trial Test AUCs by Use Batch Normalization. Individual trial test AUC values, by target (color) and Use Batch Normalization hyperparameter (row). Each point represents a single trial test AUC. Solid diamonds represent mean values across the row, and solid lines connect diamonds for the same target across rows. The use of batch normalization (BN) generally led to worse performance for both targets, with the mean and best no BN models outperforming those with BN.

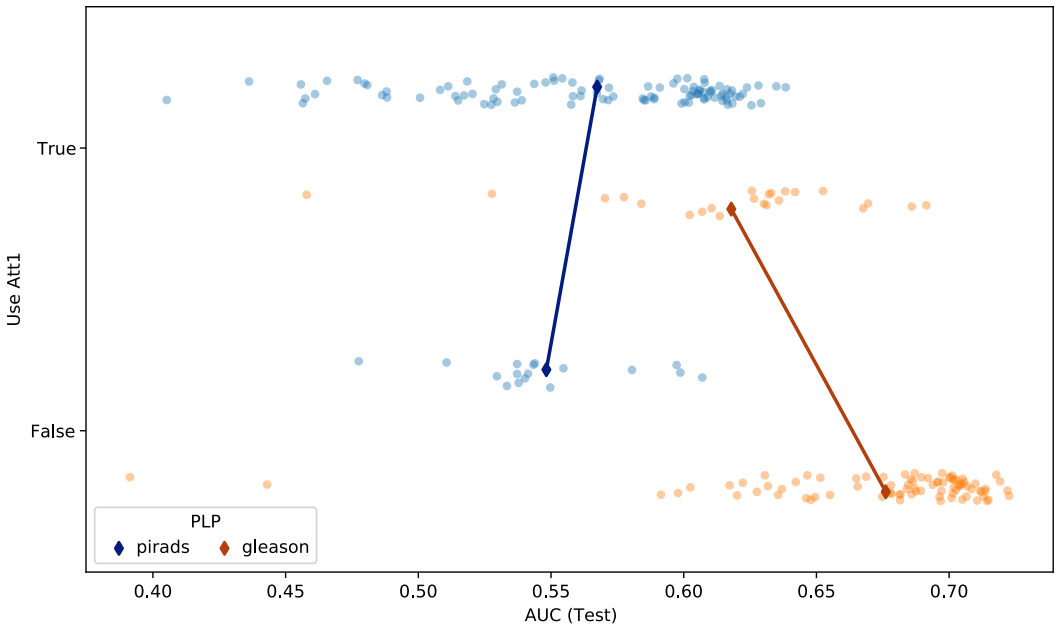


Figure 6.5: Trial Test AUCs by Use 128 Feature Attention. Individual trial test AUC values, by target (color) and Use 128 Feature Attention hyperparameter (row). Each point represents a single trial test AUC. Solid diamonds represent mean values across the row, and solid lines connect diamonds for the same target across rows. This hyperparameter provided mixed results, though interpretability is limited by oversampling of the no BN component of the hyperoptimization space for Gleason, and the use BN component for PI-RADS. The use of BN appears to provide limited benefit for PI-RADS but reduces performance for Gleason.

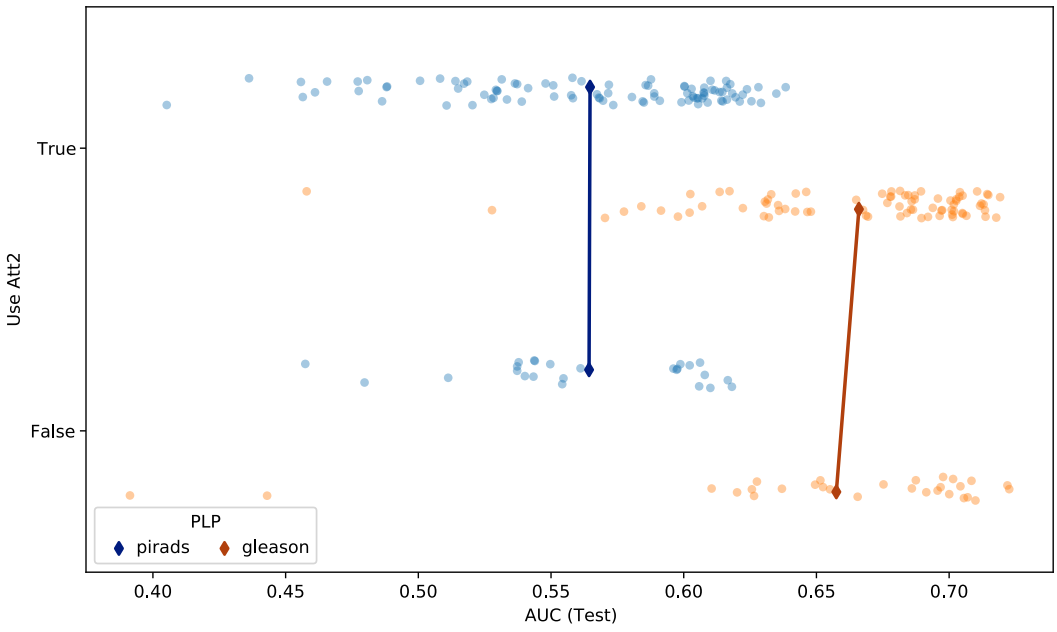


Figure 6.6: Trial Test AUCs by Use 256 Feature Attention. Individual trial test AUC values, by target (color) and Use 256 Feature Attention hyperparameter (row). Each point represents a single trial test AUC. Solid diamonds represent mean values across the row, and solid lines connect diamonds for the same target across rows. For both targets, the use of 256 feature attention module seems to have little impact on mean or maximum performance.

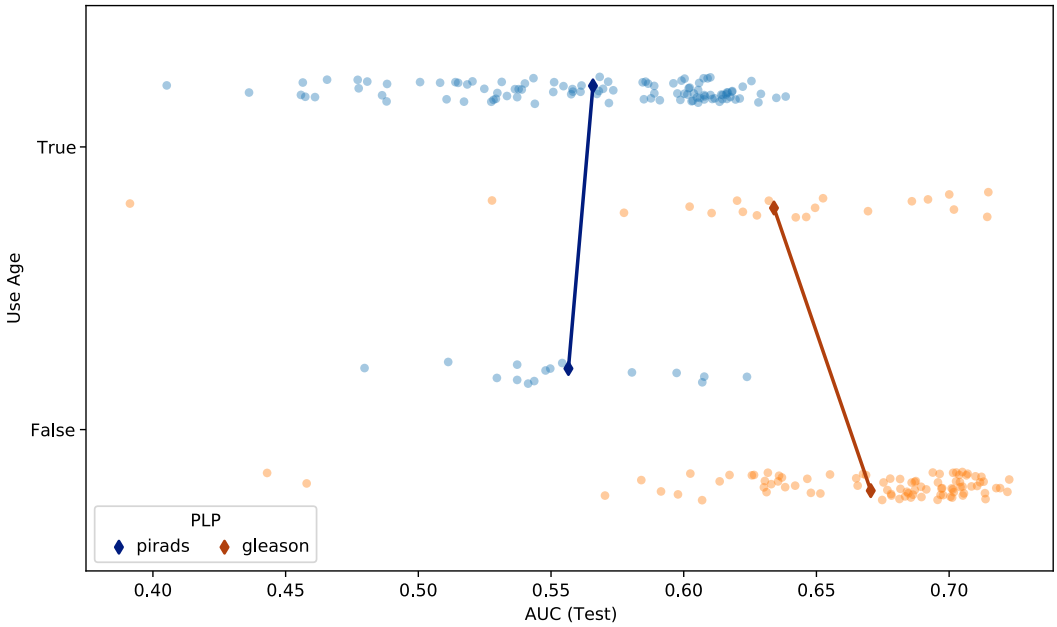


Figure 6.7: Trial Test AUCs by Use Age. Individual trial test AUC values, by target (color) and Use Age hyperparameter (row). Each point represents a single trial test AUC. Solid diamonds represent mean values across the row, and solid lines connect diamonds for the same target across rows. This hyperparameter provided mixed results, though interpretability is limited by oversampling of the no age component of the hyperoptimization space for Gleason, and the use age component for PI-RADS. The use of age appears to provide limited benefit for PI-RADS but reduces performance for Gleason. Notably, the age variable is extremely skewed, with 98% of study patients in the high age category, which may explain the relative lack of contribution for this variable.

6.3.4 Best Performing Models

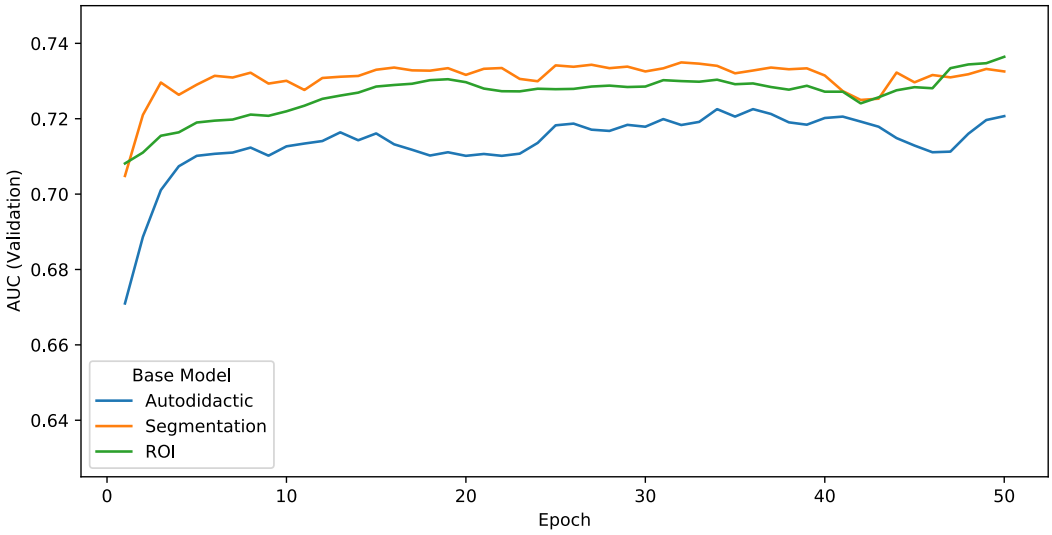
The best performing models for each target (ISUP Grade Group and PI-RADS v2), for each choice of base model, were chosen based on the maximum validation AUC, and the test set AUC is reported. A summary of results is available in Table 6.4. The overall best PI-RADS v2 model had a test AUC of 0.62, and the overall best ISUP Grade Group model had a test AUC of 0.70. Training curves for all six models are available in Figure 6.8. All models experienced approximate training saturation within the first 10 epochs with minimum further improvement.

ROC curves for the PI-RADS v2 and ISUP Grade Group models are available in Figure 6.9. Smoothed precision-recall curves are available in Figure 6.10 for both targets. The best model for PI-RADS v2 achieved a precision of 71% at 90% recall, and the best model for ISUP Grade Group achieved a precision of 65% at 90% recall. The average precision (AP) metrics for the PI-RADS v2 models were 0.716 (autodidactic), 0.729 (segmentation), and 0.721 (ROI). The AP metrics for the ISUP Grade Group models were 0.754 (autodidactic), 0.720 (segmentation), and 0.723 (ROI). Overall, the three models for each target exhibited comparable performance.

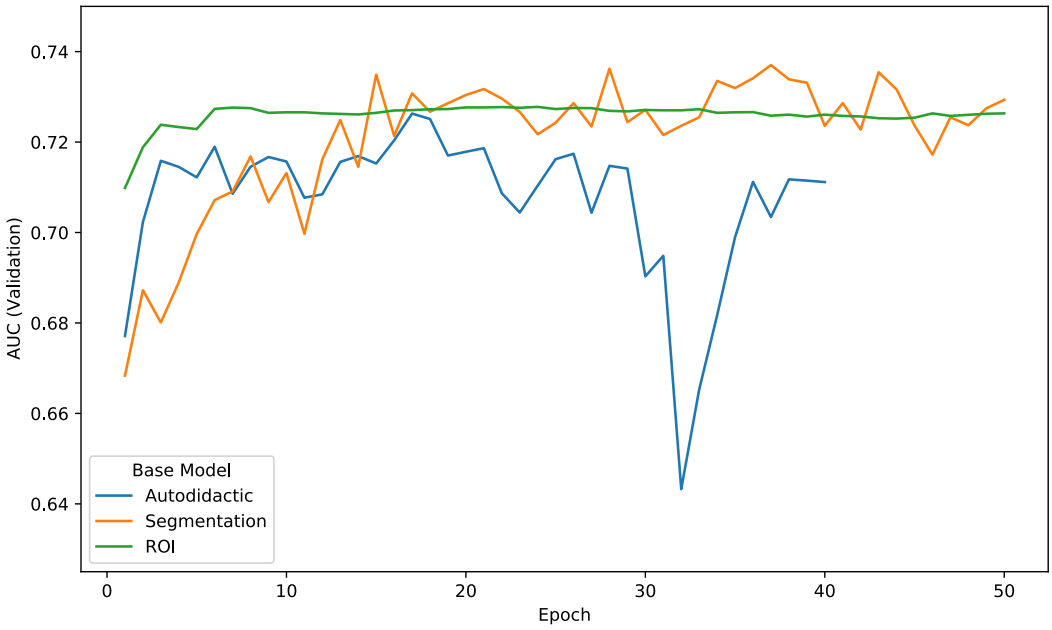
6.4 Discussion

In this study, we developed novel imaging-based patient-level predictive models for dichotomized PI-RADS v2 and ISUP Grade Group using three different types of pretrained base models, including models based on radiologist-generated ROI annotations, prostate organ segmentations, and no annotations at all. We further explored the contribution of two clinical variables, age and PSA, to the performance of these models.

We found that our models exhibited better performance for ISUP Grade Group prediction (AUC 0.7) than for PI-RADS v2 prediction (AUC 0.6). This surprising result likely is due to the fact that PI-RADS v2 scores are assigned by radiologists who make use of both high b-value images as well as DCE images, neither of which were available for this study. Because the PI-RADS v2 score is a composite metric of sequence-level evaluations, lack of access to some of these sequences may have harmed the predictive power of our models. We also found that despite extensive hyperparameter tuning, the three base models produced approximately equivalent best-case models by test set performance. This surprising result suggests that deep learning can be used to develop successful patient-level models without

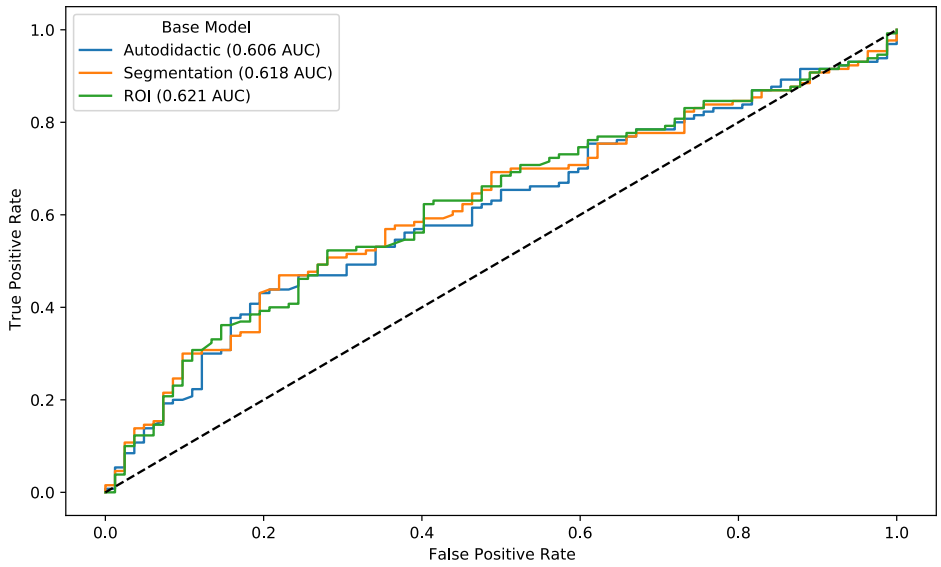


(a) PI-RADS v2

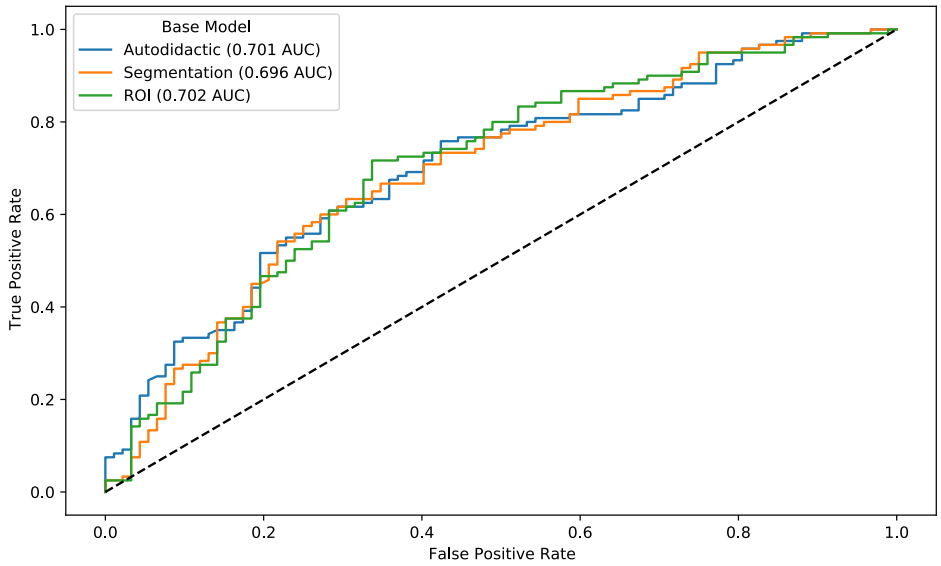


(b) ISUP Grade Group

Figure 6.8: Training Curves for Best PI-RADS v2 and ISUP Grade Group Models. Curves of validation set AUC vs epoch for the best models for each target, by base model.

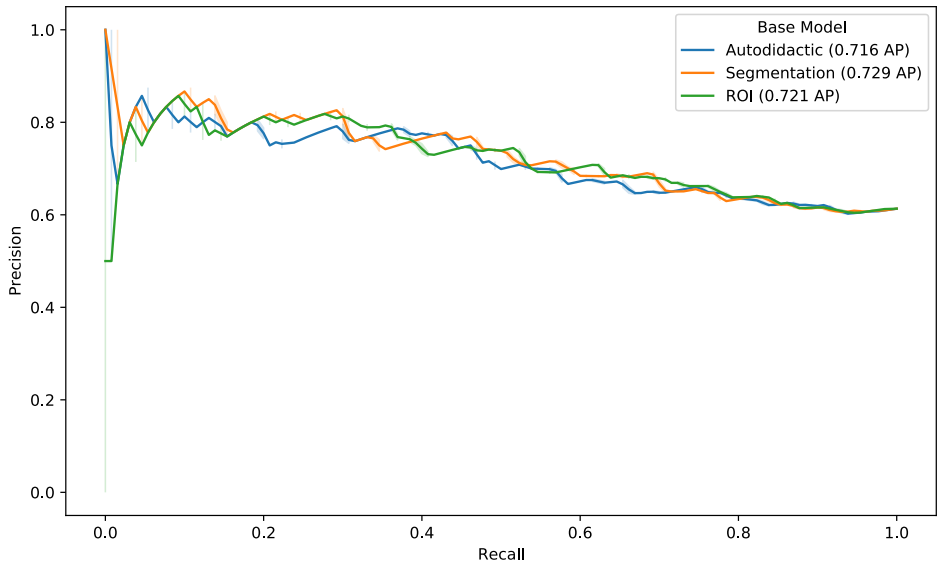


(a) PI-RADS v2

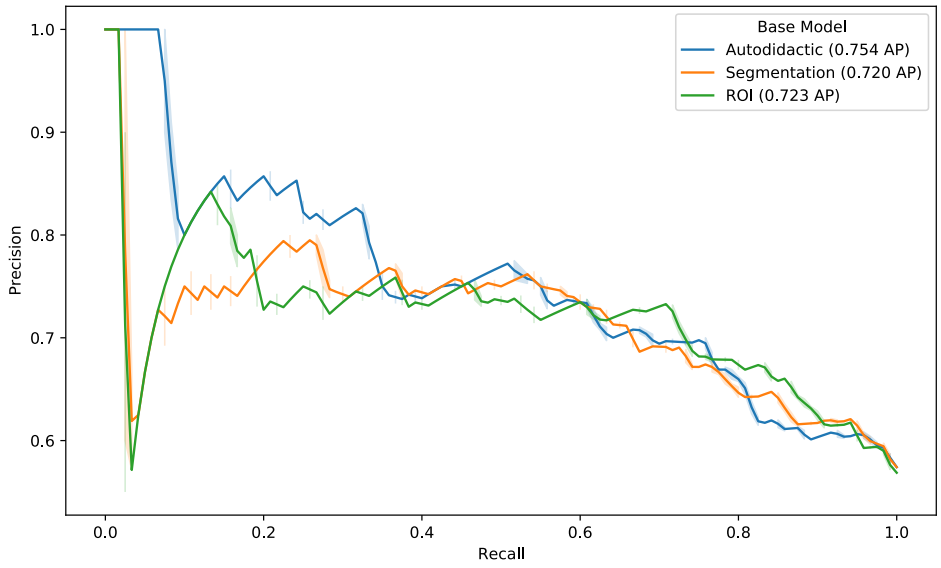


(b) ISUP Grade Group

Figure 6.9: ROC Curves for Best PI-RADS v2 and ISUP Grade Group Models. Test set performance reciever operating characteristic (ROC) curves for the best models for each target, by base model. Dotted black line is the identity function. Area under the curve (AUC) for each trial is listed in the figure legend.



(a) PI-RADS v2



(b) ISUP Grade Group

Figure 6.10: PR Curves for Best PI-RADS v2 and ISUP Grade Group Models. Test set performance precision-recall (PR) curves for the best models for each target, by base model. Opaque curves have been smoothed, lightened curves are original. Average precision (AP) for each trial is listed in the figure legend.

requiring access to a radiologist's annotations, and could have significant implications for future development. Access to radiologist annotations (i.e. cancer ROIs) is one of the most challenging barriers to deep learning research in medical imaging, and significantly more data without annotations (but with patient-level risk scores) could be available for future work.

We also found that adding PSA information to our models was helpful for performance, but adding age was not. This is likely because our dataset was heavily skewed towards an over 50 years old population, which is the cutoff we used for age dichotomization in order to harmonize with the accepted CAPRA scoring model [CPE05]. Future work could explore the benefit of a more granular encoding of age.

This work has some limitations. We discarded 38% of the initial dataset (by study count) due to our exclusion factors, the predominant of which was the inability to automatically retrieve either the T2W or ADC images from our institution's PACS or the inability to retrieve the full text radiology report from our institution's EMR. It is not possible to exclude the possibility that these challenges do not occur at random, and future studies should attempt to recover additional data through a more in-depth search in order to maximize the included population. We also did not make use of high b-value images or DCE images, both of which are traditional components of the evaluation of prostate mpMRI. Though significant debate exists about the necessity of using DCE [KBK17], it remains a component of the PI-RADS scoring system [TRH19], and established consensus agrees that high b-value images are of significant value. Future studies should consider including these sequences as potential drivers of enhanced performance. Finally, the ROI base model trained for this experiment did not reach state-of-the-art performance (Dice 0.627), likely in part because of the need to use identical base model architectures for all three pretraining approaches. This may have limited the contribution of the ROI-based encoded features, and future studies should explore different base model designs, such as the 3D AH-NET [LXZ18], as well as alternative training and optimization approaches, to enhance performance.

6.5 Conclusion

We developed novel imaging-based patient-level predictive models using T2W and ADC images for dichotomized PI-RADS v2 and ISUP Grade Group using three different types of pretrained base models, achieving a best AUC of 0.702 for ISUP Grade Group and 0.621 for PI-RADS v2. We found that all three of our pretraining approaches were effective in

facilitating PLP model training, suggesting a role for future, larger-scale studies with unannotated datasets in prostate cancer model development. Additionally, we found that the combination of clinical features like PSA with encoded imaging-based features can be an effective approach in improving patient-level model performance.

CHAPTER 7

Conclusion

The aim of this dissertation was to develop methods and approaches to enable the use of clinically generated datasets in the clinical domain of multiparametric MRI of the prostate, and with a focus on enabling the use of unannotated or weakly annotated data for machine learning. The technical chapters of this dissertation describe four major scientific efforts towards this goal, and the contributions to science are summarized as follows:

Contribution 1 We developed a regional targeted biopsy strategy using a combination of MRI and ultrasound data and histopathological evaluations from the clinical record, and found that it had statistically similar sensitivity for clinically significant prostate cancer as a combined biopsy approach while requiring fewer cores, outperforming the MRI-targeted and systematic biopsy approaches alone. This finding can be useful to urologists when determining the optimal set of biopsy locations for an individual patient and suggest that the regional targeted biopsy approach should be further evaluated as an alternative to combined MRI-targeted and systematic biopsy, and demonstrate the value of a clinical data fusion approach for clinical procedure optimization.

Contribution 2 We trained a state-of-the-art prostate segmentation model using rough clinical annotations without re-annotation. We additionally found that models trained using truncated fractions of our data were effective pre-trained starters for achieving higher performance models on external prostate segmentation challenge datasets. Our findings suggest a role for the combined use of datasets with low-quality and high-quality annotations in future medical image analysis model development in order to maximize performance while minimizing annotation effort.

Contribution 3 We made a state-of-the-art federated learning model using private

clinical prostate imaging data from three academic institutions without transferring any imaging data across institutional boundaries. The federated model demonstrated improved performance across both held-out test sets from each institution and an external test set, validating the FL paradigm. This methodology could be applied to a wide variety of DL applications in medical image analysis in order to unlock access to much larger clinical datasets while maintaining regulatory and ethical controls on patient data.

Contribution 4 We developed novel imaging-based patient-level predictive models using T2W and ADC images for dichotomized PI-RADS v2 and ISUP Grade Group using three different types of pretrained base models: un-annotated (autodidactic), weakly annotated (organ segmentation), and highly annotated (cancer ROI). We found that all three of our pretraining approaches were effective in facilitating PLP model training, suggesting a role for future, larger-scale studies with un-annotated datasets in prostate cancer model development. Additionally, we found that the combination of clinical features like PSA with encoded imaging-based features can be an effective approach in improving patient-level model performance.

The contributions of this dissertation also suggest numerous opportunities for follow-on future work to continue to develop the methods and approaches, as well as to make use of them to develop clinically useful tools. The regional targeted biopsy strategy we propose in Chapter 3 may require prospective clinical validation in order to provide the necessary level of evidence to transform clinical practice. Additionally, the combination of this regional targeted biopsy approach with machine-learning based predictive models, such as those developed in Chapter 6 could allow for even better site selection, potentially further reducing the total number of biopsies that may need to be retrieved for some patients. The federated learning approach we demonstrated for prostate segmentation in Chapter 5 can be easily adapted to enable training of models for more difficult problems, such as the detection and delineation of cancer regions of interest on mpMRI. These models could be used to accelerate the prostate targeted biopsy process, and could also provide even better pretrained features for patient-level risk prediction models using the approach described in Chapter 6. Additionally, the use of latest emerging model architectures, such as models designed for inherently anisotopic data like MR imaging, could enable both better performance for fully

convolutional image segmentation and improved feature generation for downstream models. Finally, the imaging-based patient-level predictive models could benefit from the integration of additional data types, such as the inclusion of high b-value imaging, DCE imaging, and other clinically measured biomarkers (such as PSA density) in order to better replicate the data environment used by a urologist to make a decision. The addition of these additional knowledge sources could yield improved performance of the PLP models, and thus enable better stratification of patients into treatment and watchful waiting categories. Though all of the methods and approaches developed over the course of this work are specifically targeted within the clinical domain of prostate cancer, they could also be transferred to other disease domains with similar problems, such as the detection and staging of renal and hepatic cancers.

It is beyond the scope of this dissertation to predict what the long term future may hold for the practice of medicine. However, it seems likely that the next few decades will see the introduction of increasing numbers of data science and AI-based tools into medical practice. Though the pace of change in healthcare can at times feel slow, the art and science of medicine is nevertheless constantly changing and improving. It is our hope that the contributions of this work, in combination with the work in progress at research groups like ours around the world, will enable this process of improvement to continue by unlocking the potential of clinical data for the development of better predictive models. In doing so, we hope to help build a healthier future for new generations, just as the physicians and scientists of the past built one for us.

REFERENCES

[AAA14] Mohamed Abd-Alazeez, Hashim U Ahmed, Manit Arya, Susan C Charman, Eleni Anastasiadis, Alex Freeman, Mark Emberton, and Alex Kirkham. "The accuracy of multiparametric MRI in men with negative biopsy and elevated PSA level–can it rule out clinically significant prostate cancer?" *Urologic Oncology*, **32**(1):45.e17–22, 2014.

[AHD18] Samuel G. Armato, Henkjan Huisman, Karen Drukker, Lubomir Hadjiiski, Justin S. Kirby, Nicholas Petrick, George Redmond, Maryellen L. Giger, Kenny Cha, Artem Mamonov, Jayashree Kalpathy-Cramer, and Keyvan Farahani. "PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images." *Journal of Medical Imaging*, **5**(04):1, 2018.

[AKA14] M Abd-Alazeez, A Kirkham, H U Ahmed, M Arya, E Anastasiadis, S C Charman, A Freeman, and M Emberton. "Performance of multiparametric MRI in men at risk of prostate cancer before the first biopsy: a paired validating cohort study using template prostate mapping biopsies as the reference standard." *Prostate Cancer and Prostatic Diseases*, **17**(1):40–6, 2014.

[AMA17] AMA Council on Ethics and Judicial Affairs. *Code of Medical Ethics of the American Medical Association*. American Medical Association, Chicago, 2017.

[AWR20] Michael Ahdoot, Andrew R. Wilbur, Sarah E. Reese, Amir H. Lebastchi, Sherif Mehralivand, Patrick T. Gomella, Jonathan Bloom, Sandeep Gurram, Minhaj Siddiqui, Paul Pinsky, Howard Parnes, W. Marston Linehan, Maria Merino, Peter L. Choyke, Joanna H. Shih, Baris Turkbey, Bradford J. Wood, and Peter A. Pinto. "MRI-Targeted, Systematic, and Combined Biopsy for Prostate Cancer Diagnosis." *New England Journal of Medicine*, **382**(10):917–928, 2020.

[BAS17] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features." *Scientific Data*, **4**(1):170117, 2017.

[BB12] James Bergstra and Yoshua Bengio. "Random Search for Hyper-Parameter Optimization." Technical Report 10, 2012.

[BBF16] Daniel M. Berney, Luis Beltran, Gabrielle Fisher, Bernard V. North, David Greenberg, Henrik Møller, Geraldine Soosay, Peter Scardino, and Jack Cuzick. "Validation of a contemporary prostate cancer grading system using prostate cancer death as outcome." *British Journal of Cancer*, **114**(10):1078–1083, 2016.

[BCK20] Niranjan Balachandar, Ken Chang, Jayashree Kalpathy-Cramer, and Daniel L Rubin. "Accounting for data variability in multi-institutional distributed deep learning for medical imaging." *Journal of the American Medical Informatics Association*, **27**(5):700–708, 2020.

[BCS19] Anton S. Becker, Krishna Chaitanya, Khoschy Schawkat, Urs J. Muehlematter, Andreas M. Hötker, Ender Konukoglu, and Olivio F. Donati. "Variability of manual segmentation of the prostate in axial T2-weighted MRI: A multi-reader study." *European Journal of Radiology*, **121**:108716, 2019.

[BDS20] Giorgio Brembilla, Paolo Dell'Oglio, Armando Stabile, Anna Damascelli, Lisa Brunetti, Silvia Ravelli, Giulia Cristel, Elena Schiani, Elena Venturini, Daniele Grippaldi, Vincenzo Mendola, Paola Maria Vittoria Rancoita, Antonio Esposito, Alberto Briganti, Francesco Montorsi, Alessandro Del Maschio, and Francesco De Cobelli. "Interreader variability in prostate MRI reporting using Prostate Imaging Reporting and Data System version 2.1." *European Radiology*, **30**(6):3383–3392, 2020.

[BEM08] M. Burchardt, R. Engers, M. Müller, T. Burchardt, R. Willers, J. I. Epstein, R. Ackermann, H. E. Gabbert, A. De La Taille, and M. A. Rubin. "Interobserver reproducibility of Gleason grading: Evaluation using prostate cancer tissue microarrays." *Journal of Cancer Research and Clinical Oncology*, **134**(10):1071–1078, 2008.

[BGS04] Andreas P. Berger, Christian Gozzi, Hannes Steiner, Ferdinand Frauscher, John Varkarakis, Hermann Rogatsch, Georg Bartsch, and Wolfgang Horninger. "Complication rate of transrectal ultrasound guided prostate biopsy: A comparison among 3 protocols with 6, 10 and 15 cores." *Journal of Urology*, **171**(4):1478–1481, 2004.

[BIT12] Lars Budäus, Hendrik Isbarn, Pierre Tennstedt, Georg Salomon, Thorsten Schlomm, Thomas Steuber, Alexander Haese, Felix Chun, Margit Fisch, Uwe Michl, Hans Heinzer, Hartwig Huland, and Markus Graefen. "Risk assessment of metastatic recurrence in patients with prostate cancer by using the Cancer of the Prostate Risk Assessment score: results from 2937 European patients." *BJU International*, **110**(11):1714–20, 2012.

[BLB15] Katharina Boehm, Alessandro Larcher, Burkhard Beyer, Zhe Tian, Derya Tilki, Thomas Steuber, Pierre I Karakiewicz, Hans Heinzer, Markus Graefen, and Lars Budäus. "Identifying the Most Informative Prediction Tool for Cancer-specific Mortality After Radical Prostatectomy: Comparative Analysis of Three Commonly Used Preoperative Prediction Models." *European Urology*, 2015.

[Blo15] Nicolas B. Bloch. "NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures.", 2015.

[BMF18] Adam Bezinque, Andrew Moriarity, Crystal Farrell, Henry Peabody, Sabrina L. Noyes, and Brian R. Lane. "Determination of Prostate Volume: A Comparison of Contemporary Methods." *Academic Radiology*, **25**(12):1582–1587, 2018.

[BPB20] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. "Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study." *The Lancet Oncology*, **21**(2):233–241, 2020.

[BRC12] Jelle O. Barentsz, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager, and Jurgen J. Fütterer. "ESUR prostate MR guidelines 2012." *European Radiology*, **22**(4):746–757, 2012.

[BRJ18] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, Marcel Prastawa, Esther Alberts, Jana Lipkova, John Freymann, Justin Kirby, Michel Bilello, Hassan Fathallah-Shaykh, Roland Wiest, Jan Kirschke, Benedikt Wiestler, Rivka Colen, Aikaterini Kotrotsou, Pamela Lamontagne, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Marc-Andre Weber, Abhishek Mahajan, Ujjwal Baid, Elizabeth Gerstner, Dongjin Kwon, Gagan Acharya, Manu Agarwal, Mahbubul Alam, Alberto Albiol, Antonio Albiol, Francisco J. Albiol, Varghese Alex, Nigel Allinson, Pedro H. A. Amorim, Abhijit Amrutkar, Ganesh Anand, Simon Andermatt, Tal Arbel, Pablo Arbelaez, Aaron Avery, Muneeza Azmat, Pranjal B., W Bai, Subhashis Banerjee, Bill Barth, Thomas Batchelder, Kayhan Batmanghelich, Enzo Battistella, Andrew Beers, Mikhail Belyaev, Martin Bendszus, Eze Benson, Jose Bernal, Halandur Nagaraja Bharath, George Biros, Sotirios Bisdas, James Brown, Mariano Cabezas, Shilei Cao, Jorge M. Cardoso, Eric N Carver, Adrià Casamitjana, Laura Silvana Castillo, Marcel Catà, Philippe Cattin, Albert Cerigues, Vinicius S. Chagas, Siddhartha Chandra, Yi-Ju Chang, Shiyu Chang, Ken Chang, Joseph Chazalon, Shengcong Chen, Wei Chen, Jefferson W Chen, Zhaolin Chen, Kun Cheng, Ahana Roy Choudhury, Roger Chylla, Albert Clérigues, Steven Colleman, Ramiro German Rodriguez Colmeiro, Marc Combalia, Anthony Costa, Xiaomeng Cui, Zhenzhen Dai, Lutao Dai, Laura Alexandra Daza, Eric Deutsch, Changxing Ding, Chao Dong, Shidu Dong, Wojciech Dudzik, Zach Eaton-Rosen, Gary Egan, Guilherme Escudero, Théo Estienne, Richard Everson, Jonathan Fabrizio, Yong Fan, Longwei Fang, Xue Feng, Enzo Ferrante, Lucas Fidon, Martin Fischer, Andrew P. French, Naomi Fridman, Huan Fu, David Fuentes, Yaozong Gao, Evan Gates, David Gering, Amir Gholami, Willi Gierke, Ben Glocker, Mingming Gong, Sandra González-Villá, T. Gros. "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression

Assessment, and Overall Survival Prediction in the BRATS Challenge." *arXiv*, **124**, 2018.

[BWV16] Jelle O. Barentsz, Jeffrey C. Weinreb, Sadhna Verma, Harriet C. Thoeny, Clare M. Tempany, Faina Shtern, Anwar R. Padhani, Daniel Margolis, Katarzyna J. Macura, Masoom A. Haider, Francois Cornud, and Peter L. Choyke. "Synopsis of the PI-RADS v2 Guidelines for Multiparametric Prostate Magnetic Resonance Imaging and Recommendations for Use.", 2016.

[CAB13] H Ballentine Carter, Peter C Albertsen, Michael J Barry, Ruth Etzioni, Stephen J Freedland, Kirsten Lynn Greene, Lars Holmberg, Philip Kantoff, Badrinath R Konety, Mohammad Hassan Murad, David F Penson, and Anthony L Zietman. "Early detection of prostate cancer: AUA Guideline." *The Journal of Urology*, **190**(2):419–26, 2013.

[CAL16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation." *arXiv*, 2016.

[CBG14] Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks." *Proceedings of the SPIE*, **9041**(216):904103–904115, 2014.

[CBL18] Ken Chang, Niranjan Balachandar, Carson Lam, Darvin Yi, James Brown, Andrew Beers, Bruce Rosen, Daniel L Rubin, and Jayashree Kalpathy-Cramer. "Distributed deep learning networks among institutions for medical imaging." *Journal of the American Medical Informatics Association*, **25**(8):945–954, 2018.

[CC13] K Clint Cary and Mathew R Cooperberg. "Biomarkers in prostate cancer surveillance and screening: past, present, and future." *Therapeutic Advances in Urology*, **5**(6):318–29, 2013.

[CCK19] Youn I. Choi, Jun Won Chung, Kyoung Oh Kim, Kwang An Kwon, Yoon Jae Kim, Dong Kyun Park, Sung Min Ahn, So Hyun Park, Sun Jin Sym, Dong Bok Shin, Young Saing Kim, Ki Hoon Sung, Jeong Heum Baek, and Uhn Lee. "Concordance Rate between Clinicians and Watson for Oncology among Patients with Advanced Gastric Cancer: Early, Real-World Experience in Korea." *Canadian Journal of Gastroenterology and Hepatology*, **2019**, 2019.

[CDC15] Matthew R Cooperberg, Elai Davicioni, Anamaria Crisan, Robert B Jenkins, Mercedeh Ghadessi, and R Jeffrey Karnes. "Combined value of validated clinical and genomic risk stratification tools for predicting prostate cancer mortality in a high-risk prostatectomy cohort." *European Urology*, **67**(2):326–33, 2015.

[CDM12] F Cornud, N B Delongchamps, P Mozer, F Beuvon, A Schull, N Muradyan, and M Peyromaure. "Value of multiparametric MRI in the work-up of prostate cancer." *Current Urology Reports*, **13**(1):82–92, 2012.

[CE16] Katie Chockley and Ezekiel Emanuel. "The End of Radiology? Three Threats to the Future Practice of Radiology." *Journal of the American College of Radiology*, **13**(12):1415–1420, 2016.

[CGT18] Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G. Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. "Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study." *The Lancet*, **392**(10162):2388–2396, 2018.

[CHL19] Quan Chen, Shiliang Hu, Peiran Long, Fang Lu, Yujie Shi, and Yunpeng Li. "A Transfer Learning Approach for Malignant Prostate Lesion Detection on Multiparametric MRI." *Technology in Cancer Research & Treatment*, **18**, 2019.

[CLD15] François Cornud, Paul Legmann, and Nicolas Barry Delongchamps. "Can multiparametric MRI rule in or rule out significant prostate cancer?" *Current Opinion in Urology*, **25**(6):490–7, 2015.

[CLR19] Ruida Cheng, Nathan Lay, Holger R. Roth, Baris Turkbey, Dakai Jin, William Gandler, Evan S. McCreedy, Tom Pohida, Peter Pinto, Peter Choyke, Matthew J. McAuliffe, and Ronald M. Summers. "Fully automated prostate whole gland and central gland segmentation on MRI using holistically nested networks with short connections." *Journal of Medical Imaging*, **6**(02):1, 2019.

[COS18] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. "Classification and mutation prediction from nonsmall cell lung cancer histopathology images using deep learning." *Nature Medicine*, **24**(10):1559–1567, 2018.

[CPE05] Matthew R Cooperberg, David J Pasta, Eric P Elkin, Mark S Litwin, David M Latini, Janeen Du Chane, and Peter R Carroll. "The University of California, San Francisco Cancer of the Prostate Risk Assessment score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy." *The Journal of Urology*, **173**(6):1938–42, 2005.

[CPK16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[CPS11] William J. Catalona, Alan W. Partin, Martin G. Sanda, John T. Wei, George G. Klee, Chris H. Bangma, Kevin M. Slawin, Leonard S. Marks, Stacy Loeb, Dennis L. Broyles, Sanghyuk S. Shin, Amabelle B. Cruz, Daniel W. Chan, Lori J. Sokoll, William L. Roberts, Ron H.N. N van Schaik, and Isaac A. Mizrahi. "A Multicenter Study of [-2]Pro-Prostate Specific Antigen Combined With Prostate Specific Antigen and Free Prostate Specific Antigen for Prostate Cancer Detection in the 2.0 to 10.0 ng/ml Prostate Specific Antigen Range." *The Journal of Urology*, **185**(5):1650–5, 2011.

[CTP19] Emily F. Conant, Alicia Y. Toledano, Senthil Periaswamy, Sergei V. Fotin, Jonathan Go, Justin E. Boatsman, and Jeffrey W. Hoffmeister. "Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis." *Radiology: Artificial Intelligence*, **1**(4):e180096, 2019.

[CZB17] Tyler Clark, Junjie Zhang, Sameer Baig, Alexander Wong, Masoom A. Haider, and Farzad Khalvati. "Fully automated segmentation of prostate whole gland and transition zone in diffusion-weighted MRI using convolutional neural networks." *Journal of Medical Imaging*, **4**(04):1, 2017.

[DAA11] Louise Dickinson, Hashim U. Ahmed, Clare Allen, Jelle O. Barentsz, Brendan Carey, Jurgen J. Futterer, Stijn W. Heijmink, Peter J. Hoskin, Alex Kirkham, Anwar R. Padhani, Raj Persad, Philippe Puech, Shonit Punwani, Aslam S. Sohaib, Bertrand Tombal, Arnauld Villers, Jan Van Der Meulen, and Mark Emberton. "Magnetic resonance imaging for the detection, localisation, and characterisation of prostate cancer: Recommendations from a European consensus meeting." *European Urology*, **59**(4):477–494, 2011.

[DAB12] Marc A Dall'Era, Peter C Albertsen, Christopher Bangma, Peter R Carroll, H Ballentine Carter, Matthew R Cooperberg, Stephen J Freedland, Laurence H Klotz, Christopher Parker, and Mark S Soloway. "Active surveillance for prostate cancer: a systematic review of the literature." *European Urology*, **62**(6):976–83, 2012.

[DCM12] Jeffrey Dean, Greg S Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V Le, Mark Z Mao, Marc Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, and Andrew Y Ng. "Large Scale Distributed Deep Networks." *Neural Information Processing Systems*, pp. 1–11, 2012.

[DES07] Rauf Taner Divrik, Akin Eroglu, Ali Sahin, Ferruh Zorlu, and Haluk Ozen. "Increasing the number of biopsies increases the concordance of Gleason scores of needle biopsies and prostatectomy specimens." *Urologic Oncology*, **25**(5):376–82, 2007.

[DLR18] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cían O. Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T. Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A. Keane, and Olaf Ronneberger. "Clinically applicable deep learning for diagnosis and referral in retinal disease." *Nature Medicine*, **24**(9):1342–1350, 2018.

[DST19] Abdullah Demirta, Gökhan Sönmez, evket Tolga Tombul, Türev Demirta, and Hülya Akgün. "Comparison of the Upgrading Rates of International Society of Urological Pathology Grades and Tumor Laterality in Patients Undergoing Standard 12-Core Prostate Biopsy versus Fusion Prostate Biopsy for Prostate Cancer." *Urologia Internationalis*, **103**(3):256–261, 2019.

[DWM98] Anthony V. D'Amico, Richard Whittington, S. Bruce Malkowicz, Delray Schultz, Kenneth Blank, Gregory A. Broderick, John E. Tomaszewski, Andrew A. Renshaw, Irving Kaplan, Clair J. Beard, Alan Wein, Gleason DF, Cox DR, Cox JD, Kupelian P, D'Amico AV, Oesterling JE, Beahrs OH, Zagars GK, Pisansky TM, Lee WR, Pisansky TM, Zietman AL, Hanks GE, Partin AW, Lerner SE, Zietman AL, D'Amico AV, D'Amico AV, Ragde H, Blasko JC, Wallner K, Neter J, Kaplan EL, Chodak GW, and Adolfsson J. "Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer." *JAMA*, **280**(11):969, 1998.

[ED19] Organization for Economic Co-operation and Development. *Artificial Intelligence in Society*. 2019.

[EEA15] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. "The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma." *The American Journal of Surgical Pathology*, **40**(2):1, 2015.

[EFK19] Fuad F. Elkhoury, Ely R. Felker, Lorna Kwan, Anthony E. Sisk, Merdie Delfin, Shyam Natarajan, and Leonard S. Marks. "Comparison of Targeted vs Systematic Prostate Biopsy in Men Who Are Biopsy Naive." *JAMA Surgery*, **154**(9):811, 2019.

[EKN17] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature*, **542**(7639):115–118, 2017.

[EPB15] A. El-Shater Bosaily, C. Parker, L.C. Brown, R. Gabe, R.G. Hindley, R. Kaplan, M. Emberton, and H.U. Ahmed. "PROMIS Prostate MR imaging study: A paired validating cohort study evaluating the role of multi-parametric MRI in men with clinical suspicion of prostate cancer." *Contemporary Clinical Trials*, **42**:26–40, 2015.

[Eps10] Jonathan I. Epstein. "An Update of the Gleason Grading System." **183**(2):433–440, 2010.

[EVJ17] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, Meyke Hermsen, Quirine F. Manson, Maschenka Balkenhol, Oscar Geessink, Nikolaos Stathonikos, Marcory CRF van Dijk, Peter Bult, Francisco Beca, Andrew H. Beck, Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Aoxiao Zhong, Qi Dou, Quanzheng Li, Hao Chen, Huang-Jing Lin, Pheng-Ann Heng, Christian Haß, Elia Bruni, Quincy Wong, Ugur Halici, Mustafa Ümit Öner, Rengul Cetin-Atalay, Matt Berseth, Vitali Khvatkov, Alexei Vylegzhanin, Oren Kraus, Muhammad Shaban, Nasir Rajpoot, Ruqayya Awan, Korsuk Sirinukunwattana, Talha Qaiser, Yee-Wah Tsang, David Tellez, Jonas Annuscheit, Peter Hufnagl, Mira Valkonen, Kimmo Kartasalo, Leena Latonen, Pekka Ruusuvuori, Kaisa Liimatainen, Shadi Albarqouni, Bharti Mungal, Ami George, Stefanie Demirci, Nassir Navab, Seiryo Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Hady Ahmady Phoulady, Vassili Kovalev, Alexander Kalinovsky, Vitali Liauchuk, Gloria Bueno, M. Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Daniel Racoceanu, and Rui Venâncio. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer." *JAMA*, **318**(22):2199, 2017.

[FNM16] Christopher P. Filson, Shyam Natarajan, Daniel J.A. Margolis, Jiaoti Huang, Patricia Lieu, Frederick J. Dorey, Robert E. Reiter, and Leonard S. Marks. "Prostate cancer detection with magnetic resonance-ultrasound fusion biopsy: The role of systematic and targeted biopsies." *Cancer*, **122**(6):884–892, 2016.

[FPS20] Adam E. Flanders, Luciano M. Prevedello, George Shih, Safwan S. Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T. Mongan, Anouk Stein, Felipe C. Kitamura, Matthew P. Lungren, Gagandeep Choudhary, Lesley Cala, Luiz Coelho, Monique Mogensen, Fanny Morón, Elka Miller, Ichiro Ikuta, Vahe Zohrabian, Olivia McDonnell, Christie Lincoln, Lubdha Shah, David Joyner, Amit Agarwal, Ryan K. Lee, and Jaya Nath. "Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge." *Radiology: Artificial Intelligence*, **2**(3):e190211, 2020.

[GAA15] Henrik Grönberg, Jan Adolfsson, Markus Aly, Tobias Nordström, Peter Wiklund, Yvonne Brandberg, James Thompson, Fredrik Wiklund, Johan Lindberg, Mark

Clements, Lars Egevad, and Martin Eklund. "Prostate cancer screening in men aged 5069 years (STHLM3): a prospective population-based diagnostic study." *The Lancet Oncology*, 2015.

[GBB11] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Deep Sparse Rectifier Neural Networks." *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011*, **15**:315–323, 2011.

[GGG19] Rossano Girometti, Gianluca Giannarini, Franco Greco, Miriam Isola, Lorenzo Cereser, Giuseppe Como, Stefano Sioletic, Stefano Pizzolitto, Alessandro Crestani, Vincenzo Ficarra, and Chiara Zuiani. "Interreader agreement of PI-RADS v. 2 in assessing prostate cancer with multiparametric MRI: A study using whole-mount histology as the standard of reference." *Journal of Magnetic Resonance Imaging*, **49**(2):546–555, 2019.

[GMA74] D. F. Gleason, G. T. Mellinger, and L. J. Ardving. "Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging." *Journal of Urology*, **111**(1):58–64, 1974.

[GMV15] Valentina Giannini, Simone Mazzetti, Anna Vignati, Filippo Russo, Enrico Bollito, Francesco Porpiglia, Michele Stasi, and Daniele Regge. "A fully automatic computer aided diagnosis system for peripheral zone prostate cancer detection using multi-parametric magnetic resonance imaging." *Computerized Medical Imaging and Graphics*, **46**:219–226, 2015.

[GOM12] Soumya Ghose, Arnau Oliver, Robert Martí, Xavier Lladó, Joan C. Vilanova, Jordi Freixenet, Jhimli Mitra, Désiré Sidibé, and Fabrice Meriaudeau. "A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images." *Computer Methods and Programs in Biomedicine*, **108**(1):262–287, 2012.

[GPC16] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs." *JAMA*, **316**(22):2402, 2016.

[GSL19] Matthew D. Greer, Joanna H. Shih, Nathan Lay, Tristan Barrett, Leonardo Bittencourt, Samuel Borofsky, Ismail Kabakus, Yan Mee Law, Jamie Marko, Haytham Shebel, Maria J. Merino, Bradford J. Wood, Peter A. Pinto, Ronald M. Summers, Peter L. Choyke, and Baris Turkbey. "Interreader variability of prostate imaging reporting and data system version 2 in detecting and assessing prostate cancer lesions at prostate MRI." *American Journal of Roentgenology*, **212**(6):1197–1205, 2019.

[Har18] Balazs Harangi. "Skin lesion classification with ensembles of deep convolutional neural networks." *Journal of Biomedical Informatics*, **86**:25–32, 2018.

[HCF17] Won Sik Ham, Heather J. Chalfin, Zhaoyong Feng, Bruce J. Trock, Jonathan I. Epstein, Carling Cheung, Elizabeth Humphreys, Alan W. Partin, and Misop Han. "New Prostate Cancer Grading System Predicts Long-term Survival Following Surgery for Gleason Score 810 Prostate Cancer." *European Urology*, **71**(6):907–912, 2017.

[Hea19] Office of the National Coordinator for Health Information Technology. "Office-based Physician Electronic Health Record Adoption, Health IT Quick-Stat 50.", 2019.

[HN10] Geoffrey E. Hinton and Vinod Nair. "Rectified Linear Units Improve Restricted Boltzmann Machines." In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814. Omnipress, 2010.

[HSB11] Schuyler Halverson, Matthew Schipper, Kevin Blas, Vivien Lee, Aaron Sabolch, Karin Olson, Howard M Sandler, Felix Y Feng, and Daniel A Hamstra. "The Cancer of the Prostate Risk Assessment (CAPRA) in patients treated with external beam radiation therapy: evaluation and optimization in patients at higher risk of relapse." *Radiotherapy and Oncology*, **101**(3):513–20, 2011.

[HSH14] Geoffrey E Hinton, Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout : A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research*, **15**:1929–1958, 2014.

[HTP08] Alexander Haese, Alexandre de la Taille, Hendrik van Poppel, Michael Marberger, Arnulf Stenzl, Peter F A Mulders, Hartwig Huland, Clément-Claude Abbou, Mesut Remzi, Martina Tinzl, Susan Feyerabend, Alexander B Stillebroer, Martijn P M Q van Gils, and Jack A Schalken. "Clinical utility of the PCA3 urine assay in European men scheduled for repeat biopsy." *European Urology*, **54**(5):1081–8, 2008.

[HVH13] Thomas Hambrock, Pieter C Vos, Christina A Hulsbergen-van de Kaa, Jelle O Barentsz, and Henkjan J Huisman. "Prostate cancer: computer-aided diagnosis with multiparametric 3-T MR imaging–effect on observer performance." *Radiology*, **266**(2):521–30, 2013.

[HWM19] Nawar Hanna, Matthew F. Wszolek, Amirkasra Mojtahed, Edouard Nicaise, Bo Wu, Francisco J. Gelpi-Hammerschmidt, Keyan Salari, Douglas M. Dahl, Michael L. Blute, Mukesh Harisinghani, and Adam S. Feldman. "Multiparametric Magnetic Resonance Imaging-Ultrasound Fusion Biopsy Improves but Does

Not Replace Standard Template Biopsy for the Detection of Prostate Cancer." *Journal of Urology*, **202**(5):944–951, 2019.

[HWS19] Zhengming Hu, Jinrui Wang, Desheng Sun, Ligang Cui, and Weiqiang Ran. "How Many Cores Does Systematic Prostate Biopsy Need?: A LargeSample Retrospective Analysis." *Journal of Ultrasound in Medicine*, **38**(6):1491–1499, 2019.

[HZR15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In *IEEE International Conference on Computer Vision*, pp. 1026–1034. IEEE, 2015.

[HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016.

[IHN11] Fumio Ishizaki, Md Aminul Hoque, Tsutomu Nishiyama, Takashi Kawasaki, Takashi Kasahara, Noboru Hara, Itsuhiro Takizawa, Toshihiro Saito, Yasuo Kitamura, Kohei Akazawa, and Kota Takahashi. "External validation of the UCSF-CAPRA (University of California, San Francisco, Cancer of the Prostate Risk Assessment) in Japanese patients receiving radical prostatectomy." *Japanese Journal of Clinical Oncology*, **41**(11):1259–64, 2011.

[IJW20] Fabian Isensee, Paul Jäger, Jakob Wasserthal, David Zimmerer, Jens Petersen, Simon Kohl, Justus Schock, Andre Klein, Tobias Roß, Sebastian Wirkert, Peter Neher, Stefan Dinkelacker, Gregor Köhler, and Klaus Maier-Hein. "batchgenerators - a python framework for data augmentation." 2020.

[Inv] Invivo-Philips. "DynaCAD Prostate Advanced visualization for prostate MRI analysis — Philips Healthcare.".

[IPK18] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. "nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation." *arXiv*, 2018.

[IS15] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." *arXiv*, **37**, 2015.

[JMC15] Mohamed Jalloh, Frank Myers, Janet E. Cowan, Peter R. Carroll, and Matthew R. Cooperberg. "Racial Variation in Prostate Cancer Upgrading and Upstaging Among Men with Low-risk Clinical Characteristics." *European Urology*, **67**(3):451–457, 2015.

[JSH19] Haozhe Jia, Yang Song, Heng Huang, Weidong Cai, and Yong Xia. "HD-Net: Hybrid Discriminative Network for Prostate Segmentation in MR Images." In *MICCAI*. Springer, 2019.

[JTL15] Arjun Jain, Jonathan Tompson, Yann LeCun, and Christoph Bregler. "MoDeep: A Deep Learning Framework Using Motion Features for Human Pose Estimation." In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *ACCV 2014*, volume 9004 of *Lecture Notes in Computer Science*, pp. 302–315. Springer International Publishing, 2015.

[JXS19] Haozhe Jia, Yong Xia, Yang Song, Donghao Zhang, Heng Huang, Yanning Zhang, and Weidong Cai. "3D APA-Net: 3D Adversarial Pyramid Anisotropic Convolutional Network for Prostate Segmentation in MR Images." *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019.

[JYF21] Yao Jin, Guang Yang, Ying Fang, Ruipeng Li, Xiaomei Xu, Yongkai Liu, and Xiaobo Lai. "3D PBV-Net: An automated prostate MRI data segmentation method." *Computers in Biology and Medicine*, **128**:104160, 2021.

[KAS15] Christos K Kontos, Panagiotis G Adamopoulos, and Andreas Scorilas. "Prognostic and predictive biomarkers in prostate cancer." *Expert Review of Molecular Diagnostics*, pp. 1–10, 2015.

[KBK17] Christiane K. Kuhl, Robin Bruhn, Nils Krämer, Sven Nebelung, Axel Heidenreich, and Simone Schrading. "Abbreviated Biparametric Prostate MR Imaging in Men with Elevated Prostate-specific Antigen." *Radiology*, p. 170129, 2017.

[KKS19] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. "Key challenges for delivering clinical impact with artificial intelligence.", 2019.

[KMR20] Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. "Secure, privacy-preserving and federated machine learning in medical imaging." *Nature Machine Intelligence*, **2**(6):305–311, 2020.

[KMW09] Rune Kvåle, Bjørn Møller, Rolf Wahlqvist, Sophie D Fosså, Aasmund Berner, Christer Busch, Anne E Kyrdalen, Aud Svindland, Trond Viset, and Ole J Halvorsen. "Concordance between Gleason scores of needle biopsies and radical prostatectomy specimens: a population-based study." *BJU International*, **103**(12):1647–54, 2009.

[KRB18] Veeru Kasivisvanathan, Antti S. Rannikko, Marcelo Borghi, Valeria Panebianco, Lance A. Mynderse, Markku H. Vaarala, Alberto Briganti, Lars Budäus, Giles Hellawell, Richard G. Hindley, Monique J. Roobol, Scott Eggener, Maneesh Ghei, Arnauld Villers, Franck Bladou, Geert M. Villeirs, Jaspal Virdi, Silvan Boxler,

Grégoire Robert, Paras B. Singh, Wulphert Venderink, Boris A. Hadaschik, Alain Ruffion, Jim C. Hu, Daniel Margolis, Sébastien Crouzet, Laurence Klotz, Samir S. Taneja, Peter Pinto, Inderbir Gill, Clare Allen, Francesco Giganti, Alex Freeman, Stephen Morris, Shonit Punwani, Norman R. Williams, Chris Brew-Graves, Jonathan Deeks, Yemisi Takwoingi, Mark Emberton, and Caroline M. Moore. "MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis." *New England Journal of Medicine*, **378**(19):1767–1777, 2018.

[KRR18] Alexander P. Kenigsberg, Audrey Renson, Andrew B. Rosenkrantz, Richard Huang, James S. Wysock, Samir S. Taneja, and Marc A. Bjurlin. "Optimizing the Number of Cores Targeted During Prostate Magnetic Resonance Imaging Fusion Target Biopsy." *European Urology Oncology*, **1**(5):418–425, 2018.

[KSH12] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Geoffret E Hinton, IIya Sulskever, and Geoffrey E. Geoffret E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[KWC13] Timur H. Kuru, Karan Wadhwa, Richard Tsung Meng Chang, Lina Maria Carmona Echeverria, Matthias Roethke, Alexander Polson, Giles Rottenberg, Brendan Koo, Edward M. Lawrence, Jonas Seidenader, Vincent Gnanapragasam, Richard Axell, Wilfried Roth, Anne Warren, Andrew Doble, Gordon Muir, Rick Popert, Heinz Peter Schlemmer, Boris A. Hadaschik, and Christof Kastner. "Definitions of terms, Processes and a minimum dataset for transperineal prostate biopsies: A standardization approach of the Ginsburg Study Group for enhanced prostate diagnostics." *BJU International*, **112**(5):568–577, 2013.

[KXW15] Jin Tae Kwak, Sheng Xu, Bradford J Wood, Baris Turkbey, Peter L Choyke, Peter A Pinto, Shijun Wang, and Ronald M Summers. "Automated prostate cancer detection using T2-weighted and high-b-value diffusion-weighted magnetic resonance imaging." *Medical Physics*, **42**(5):2368–78, 2015.

[LBI10] Giovanni Lughezzani, Lars Budäus, Hendrik Isbarn, Maxine Sun, Paul Perrotte, Alexander Haese, Felix K Chun, Thorsten Schlomm, Thomas Steuber, Hans Heinzer, Hartwig Huland, Francesco Montorsi, Markus Graefen, and Pierre I Karakiewicz. "Head-to-head comparison of the three most commonly used preoperative models for prediction of biochemical recurrence after radical prostatectomy." *European Urology*, **57**(4):562–8, 2010.

[LBK15] Geert J S Litjens, Jelle O Barentsz, Nico Karssemeijer, and Henkjan J Huisman. "Clinical evaluation of a computer-aided diagnosis system for determining cancer aggressiveness in prostate MRI." *European Radiology*, **25**(11):3187–99, 2015.

[LCI13] Bradley C. Lowekamp, David T. Chen, Luis Ibáñez, and Daniel Blezek. "The Design of SimpleITK." *Frontiers in Neuroinformatics*, **7**:45, 2013.

[LCK] Kenneth Lin, Jennifer M Croswell, Helen Koenig, Clarence Lam, and Ashley Maltz. "Prostate-Specific Antigen-Based Screening for Prostate Cancer.".

[LGG18] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. "Focal Loss for Dense Object Detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**(2):318–327, 2018.

[LJR18] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. "A System for Massively Parallel Hyperparameter Tuning." *Conference on Machine Learning and Systems 2020*, 2018.

[LM14] Michele Larobina and Loredana Murino. "Medical Image File Formats." *Journal of Digital Imaging*, **27**(2):200–206, 2014.

[LM20] Dianbo Liu and Tim Miller. "Federated pretraining and fine tuning of BERT using clinical notes from multiple silos." *arXiv*, 2020.

[LMX19] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M. Jorge Cardoso, and Andrew Feng. "Privacy-Preserving Federated Brain Tumour Segmentation." In *Machine Learning in Medical Imaging (MLMI) Workshop 2019*, volume 11861 LNCS, pp. 133–141. Springer, 2019.

[LS02] J.-B. Lattouf and F. Saad. "Gleason score on biopsy: is it reliable for predicting the final grade on pathology?" *BJU International*, **90**(7):694–698, 2002.

[LSG19] Amanda Jane Lu, Jamil S. Syed, Kamyar Ghabili, Walter Robert Hsiang, Kevin A. Nguyen, Michael S. Leapman, and Preston C. Sprenkle. "Role of Core Number and Location in Targeted Magnetic Resonance Imaging-Ultrasound Fusion Prostate Biopsy." *European Urology*, **76**(1):14–17, 2019.

[LTG17] Nathan Lay, Yohannes Tsehay, Matthew D. Greer, Baris Turkbey, Jin Tae Kwak, Peter L. Choyke, Peter Pinto, Bradford J. Wood, and Ronald M. Summers. "Detection of prostate cancer in multiparametric MRI using random forest with instance weighting." *Journal of Medical Imaging*, **4**(2):024506, 2017.

[LTT15] Feng Li, Loc Tran, Kim-han Thung, Shuiwang Ji, Dinggang Shen, and Jiang Li. "A Robust Deep Model for Improved Classification of AD / MCI Patients." *IEEE Journal of Biomedical Health Informatics*, **19**(5):1610–1616, 2015.

[LTV14] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip Eddie Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli

Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge." *Medical Image Analysis*, **18**(2):359–373, 2014.

[LUV08] Hans Lilja, David Ulmert, and Andrew J Vickers. "Prostate-specific antigen and prostate cancer: prediction, detection and monitoring." *Nature Reviews: Cancer*, **8**(4):268–78, 2008.

[LWT13] Peter Liu, Shijun Wang, Baris Turkbey, Kinzya Grant, Peter Pinto, Peter Choyke, Bradford J. Wood, and Ronald M. Summers. "A prostate cancer computer-aided diagnosis system using multimodal magnetic resonance imaging and targeted biopsy labels." volume 8670, p. 86701G. International Society for Optics and Photonics, 2013.

[LXL19] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. "Adaptive Gradient Methods with Dynamic Bound of Learning Rate." *7th International Conference on Learning Representations*, 2019.

[LXZ18] Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu. "3D Anisotropic Hybrid Network: Transferring Convolutional Features from 2D Images to 3D Anisotropic Volumes." In Alejandro F Frangi, Julia A Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *MICCAI*, pp. 851–858, Cham, 2018. Springer International Publishing.

[LZF17] Saifeng Liu, Huaixiu Zheng, Yesu Feng, and Wei Li. "Prostate cancer diagnosis using deep learning with 3D multiparametric MRI." In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, p. 1013428. SPIE, 2017.

[LZG20] Pradeep Lam, Alyssa H. Zhu, Iyad Ba Gari, Neda Jahanshad, and Paul M. Thompson. "3D Grid-Attention Networks for Interpretable Age and Alzheimer's Disease Prediction from Structural MRI." *arXiv*, 2020.

[MAB11] Anant Madabhushi, Shannon Agner, Ajay Basavanhally, Scott Doyle, and George Lee. "Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data." *Computerized Medical Imaging and Graphics*, **35**(7-8):506–14, 2011.

[MAL20] Hong Y. Ma, Firas S. Ahmed, Lyndon Luk, Luis A. Pina Martina, Sven Wenske, and Hiram Shaish. "The negative predictive value of a PI-RADS version 2 score of 1 on prostate MRI and the factors associated with a false-negative MRI study." *American Journal of Roentgenology*, **215**(3):667–672, 2020.

[MBB17] Nicolas Mottet, Joaquim Bellmunt, Michel Bolla, Erik Briers, Marcus G. Cumberbatch, Maria De Santis, Nicola Fossati, Tobias Gross, Ann M. Henry, Steven

Joniau, Thomas B. Lam, Malcolm D. Mason, Vsevolod B. Matveev, Paul C. Moldovan, Roderick C.N. van den Bergh, Thomas Van den Broeck, Henk G. van der Poel, Theo H. van der Kwast, Olivier Rouvière, Ivo G. Schoots, Thomas Wiegel, and Philip Cornford. "EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent." *European Urology*, **71**(4):618–629, 2017.

[MBH19] Sohrab Afshari Mirak, Amirhossein Mohammadian Bajgiran, Melina Hosseiny, Sepideh Shakeri, Afshin Azadikhah, Anthony E. Sisk, Robert E. Reiter, and Steven S. Raman. "Comparison of Performance of PI-RADSv2 and a Quantitiative PI-RADSv1 Based Protocol in 3T Multiparametric MRI for Detection, Grading and Staging of Prostate Cancer Using Whole Mount Histopathology as Reference Standard in 569 patients." In *Proceedings of the American Urological Association Annual Meeting*, 2019.

[MBS17] Sherif Mehralivand, Sandra Bednarova, Joanna H. Shih, Francesca V. Mertan, Sonia Gaur, Maria J. Merino, Bradford J. Wood, Peter A. Pinto, Peter L. Choyke, and Baris Turkbey. "Prospective Evaluation of PI-RADS Version 2 Using the International Society of Urological Pathology Prostate Cancer Grade Group System." *Journal of Urology*, **198**(3):583–590, 2017.

[MDV21] Urs J. Muehlematter, Paola Daniore, and Kerstin N. Vokinger. "Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (201520): a comparative analysis.", 2021.

[ME15] Caroline M Moore and Mark Emberton. "Will the attributes of multiparametric MRI permit the creation of a new approach to therapy?" *Current Opinion in Urology*, **25**(6):518–21, 2015.

[MGF13] Pieter Meurs, Rose Galvin, Deirdre M Fanning, and Tom Fahey. "Prognostic value of the CAPRA clinical prediction rule: a systematic review and meta-analysis." *BJU International*, **111**(3):427–36, 2013.

[MJB15] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Herve Delingette, Cagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M.S. S Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton,

Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)." *IEEE Transactions on Medical Imaging*, **34**(10):1993–2024, 2015.

[MKS15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. "Human-level control through deep reinforcement learning." *Nature*, **518**(7540):529–533, 2015.

[MMR17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. "Communication-Efficient Learning of Deep Networks from Decentralized Data." In *Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

[MSG17] Alireza Mehrtash, Alireza Sedghi, Mohsen Ghafoorian, Mehdi Taghipour, Clare M. Tempany, William M. Wells, Tina Kapur, Parvin Mousavi, Purang Abolmaesumi, and Andriy Fedorov. "Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks." In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, p. 101342A. SPIE, 2017.

[MSG20] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. "International evaluation of an AI system for breast cancer screening." *Nature*, **577**(7788):89–94, 2020.

[MSM14] Judd W Moul, Mark J Sarno, Jonathan E McDermed, Melissa T Triebell, and Mark A Reynolds. "NADiA ProsVue prostate-specific antigen slope, CAPRA-S, and prostate cancer–specific survival after radical prostatectomy." *Urology*, **84**(6):1427–32, 2014.

[MSS15] Berrend G. Muller, Joanna H. Shih, Sandeep Sankineni, Jamie Marko, Soroush Rais-Bahrami, Arvin Koruthu George, Jean J.M.C.H. De La Rosette, Maria J. Merino, Bradford J. Wood, Peter Pinto, Peter L. Choyke, and Baris Turkbey. "Prostate cancer: Interobserver agreement and accuracy with the revised

prostate imaging reporting and data system at multiparametric mr imaging1." *Radiology*, **277**(3):741–750, 2015.

[Muk17] Siddhartha Mukherjee. "A.I. Versus M.D." *The New Yorker*, 2017.

[MVS17] Paul C. Moldovan, Thomas Van den Broeck, Richard Sylvester, Lorenzo Marconi, Joaquim Bellmunt, Roderick C.N. van den Bergh, Michel Bolla, Erik Briers, Marcus G. Cumberbatch, Nicola Fossati, Tobias Gross, Ann M. Henry, Steven Joniau, Theo H. van der Kwast, Vsevolod B. Matveev, Henk G. van der Poel, Maria De Santis, Ivo G. Schoots, Thomas Wiegel, Cathy Yuhong Yuan, Philip Cornford, Nicolas Mottet, Thomas B. Lam, and Olivier Rouvière. "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." *European Urology*, **72**(2):250–266, 2017.

[MYN13] Leonard Marks, Shelena Young, and Shyam Natarajan. "MRIultrasound fusion for guidance of targeted prostate biopsy." *Current Opinion in Urology*, **23**(1):43–50, 2013.

[Nat15] National Comprehensive Cancer Network. "Prostate Cancer Early Detection (Version 2.2015).", 2015.

[NSL13] Robert K Nam, Refik Saskin, Yuna Lee, Ying Liu, Calvin Law, Laurence H Klotz, D Andrew Loblaw, John Trachtenberg, Aleksandra Stanimirovic, Andrew E Simor, Arun Seth, David R Urbach, and Steven A Narod. "Increasing hospital admission rates for urological complications after transrectal ultrasound guided prostate biopsy." *The Journal of Urology*, **189**(1 Suppl):S12–7; discussion S17–8, 2013.

[NVI] NVIDIA Corporation. "Clara Train Application Framework Documentation Clara Train Application Framework v3.0 documentation.".

[PBB17] Andrei S. Purysko, Leonardo K. Bittencourt, Jennifer A. Bullen, Thomaz R. Mostardeiro, Brian R. Herts, and Eric A. Klein. "Accuracy and interobserver agreement for prostate imaging reporting and data system, version 2, for the characterization of lesions identified on multiparametric MRI of the prostate." *American Journal of Roentgenology*, **209**(2):339–345, 2017.

[PBV07] Philippe Puech, Nacim Betrouni, Romain Viard, Arnauld Villers, Xavier Leroy, and Laurent Lemaitre. "Prostate cancer computer-assisted diagnosis software using dynamic contrast-enhanced MRI." *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, **2007**:5567–70, 2007.

[PEV18] Alan Priester, Fuad Elkhoury, Jacob Vandel, Merdie Delfin, Ely Felker, Shyam Natarajan, and Leonard Marks. "Targeted Prostate Biopsy: Cancer Extends Beyond the ROI!" In *Proceedings of the American Urological Association Annual Meeting*, volume 199, 2018.

[PMR15] Francesco Paparo, Michela Massollo, Ludovica Rollandi, Arnoldo Piccardo, Filippo Grillo Ruggieri, and Gian Andrea Rollandi. "The clinical role of multimodality imaging in the detection of prostate cancer recurrence after radical prostatectomy and radiation therapy: past, present, and future." *Ecancermedicalscience*, **9**:570, 2015.

[PNK17] Alan Priester, Shyam Natarajan, Pooria Khoshnoodi, Daniel J. Margolis, Steven S. Raman, Robert E. Reiter, Jiaoti Huang, Warren Grundfest, and Leonard S. Marks. "Magnetic Resonance Imaging Underestimation of Prostate Cancer Geometry: Use of Patient Specific Molds to Correlate Images with Whole Mount Pathology." *Journal of Urology*, **197**(2):320–326, 2017.

[PPP15] Sanoj Punnen, Nicola Pavan, and Dipen J Parekh. "Finding the Wolf in Sheep's Clothing: The 4Kscore Is a Novel Blood Test That Can Accurately Identify the Risk of Aggressive Prostate Cancer." *Reviews in Urology*, **17**(1):3–13, 2015.

[Pre03] Joseph C. Presti. "Prostate biopsy: how many cores are enough?" *Urologic Oncology: Seminars and Original Investigations*, **21**(2):135–140, 2003.

[PWR19] Anwar R. Padhani, Jeffrey Weinreb, Andrew B. Rosenkrantz, Geert Villeirs, Baris Turkbey, and Jelle Barentsz. "Prostate Imaging-Reporting and Data System Steering Committee: PI-RADS v2 Status Update and Future Directions." *European Urology*, **75**(3):385–396, 2019.

[QCB17] Gwenolé Quellec, Katia Charrière, Yassine Boudi, Béatrice Cochener, and Mathieu Lamard. "Deep image mining for diabetic retinopathy screening." *Medical Image Analysis*, **39**:178–193, 2017.

[RBT17] Andrew B. Rosenkrantz, James S. Babb, Samir S. Taneja, and Justin M. Ream. "Proposed Adjustments to PI-RADS Version 2 Decision Rules: Impact on Prostate Cancer Detection." *Radiology*, **283**(1):119–129, 2017.

[RCS20] Holger R. Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C. Bizzo, Yuhong Wen, Varun Buch, Meesam Shah, Felipe Kitamura, Matheus Mendonça, Vitor Lavor, Ahmed Harouni, Colin Compas, Jesse Tetreault, Prerna Dogra, Yan Cheng, Selnur Erdal, Richard White, Behrooz Hashemian, Thomas Schultz, Miao Zhang, Adam McCarthy, B. Min Yun, Elshaimaa Sharaf, Katharina V. Hoebel, Jay B. Patel, Bryan Chen, Sean Ko, Evan Leibovitz, Etta D. Pisano, Laura

Coombs, Daguang Xu, Keith J. Dreyer, Ittai Dayan, Ram C. Naidu, Mona Flores, Daniel Rubin, and Jayashree Kalpathy-Cramer. "Federated Learning for Breast Density Classification: A Real-World Implementation." In *Domain Adaptation and Representation Transfer (DART), and Distributed and Collaborative Learning (DLC) Workshops 2020*, volume 12444 LNCS, pp. 181–191. Springer Science and Business Media Deutschland GmbH, 2020.

[RDS15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision*, **115**(3):211–252, 2015.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *MICCAI*, 2015.

[RGC16] Andrew B. Rosenkrantz, Luke A. Ginocchio, Daniel Cornfeld, Adam T. Froemming, Rajan T. Gupta, Baris Turkbey, Antonio C. Westphalen, James S. Babb, and Daniel J. Margolis. "Interobserver reproducibility of the PI-RADS version 2 lexicon: A multicenter study of six experienced prostate radiologists." *Radiology*, **280**(3):793–804, 2016.

[RHL20] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett Landman, Klaus Maier-Hein, Sebastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. "The Future of Digital Health with Federated Learning." *arXiv*, 2020.

[RKF13] Matthew J Resnick, Tatsuki Koyama, Kang-Hsien Fan, Peter C Albertsen, Michael Goodman, Ann S Hamilton, Richard M Hoffman, Arnold L Potosky, Janet L Stanford, Antoinette M Stroup, R Lawrence Van Horn, and David F Penson. "Long-term functional outcomes after treatment for localized prostate cancer." *The New England Journal of Medicine*, **368**(5):436–45, 2013.

[RMG20] Tara A. Retson, Evan M. Masutani, Daniel Golden, and Albert Hsiao. "Clinical Performance and Role of Expert Supervision of Deep Learning for Cardiac Ventricular Volumetry: A Validation Study." *Radiology: Artificial Intelligence*, **2**(4):e190064, 2020.

[Rot82] Scott D. Roth. "Ray casting for modeling solids." *Computer Graphics and Image Processing*, **18**(2):109–144, 1982.

[RPR19] Olivier Rouvière, Philippe Puech, Raphaële Renard-Penna, Michel Claudon, Catherine Roy, Florence Mège-Lechevallier, Myriam Decaussin-Petrucci, Marine Dubreuil-Chambardel, Laurent Magaud, Laurent Remontet, Alain Ruffion, Marc

Colombel, Sébastien Crouzet, Anne-Marie Schott, Laurent Lemaitre, Muriel Rabilloud, Nicolas Grenier, Nicolas Barry Delongchamps, Romain Boutier, Flavie Bratan, Serge Brunelle, Philippe Camparo, Pierre Colin, Jean-Michel Corréas, François Cornélis, François Cornud, Fanny Cros, Jean-Luc Descotes, Pascal Eschwege, Gaelle Fiard, Jean-Philippe Fendler, Hocine Habchi, Philippe Hallouin, Ahmed Khairoune, Hervé Lang, Yann Lebras, Frédéric Lefèvre, Bernard Malavaud, Paul Cezar Moldovan, Nicolas Mottet, Pierre Mozer, Pierre Nevoux, Gaele Pagnoux, Gilles Pasticier, Daniel Portalez, Eric Potiron, Athivada Soto Thammavong, Marc-Olivier Timsit, Arnault Viller, and Jochen Walz. "Use of prostate systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-naive patients (MRI-FIRST): a prospective, multicentre, paired diagnostic study." *The Lancet Oncology*, **20**(1):100–109, 2019.

[RSR21] Alex G. Raman, Karthik V. Sarma, Steven S. Raman, Alan M. Priester, Sohrab Afshari Mirak, Hannah H. Riskin-Jones, Nikhil Dhinagar, William Speier, Ely Felker, Anthony E. Sisk, David Lu, Adam Kinnaird, Robert E. Reiter, Leonard S. Marks, and Corey W. Arnold. "Optimizing Spatial Biopsy Sampling for the Detection of Prostate Cancer." *Journal of Urology*, 2021.

[RZK19] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. "Transfusion: Understanding Transfer Learning for Medical Imaging." In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[SAM14] Lucy A.M. Simmons, Hashim Uddin Ahmed, Caroline M. Moore, Shonit Punwani, Alex Freeman, Yipeng Hu, Dean Barratt, Susan C. Charman, Jan Van der Meulen, and Mark Emberton. "The PICTURE study Prostate Imaging (multi-parametric MRI and Prostate HistoScanning) Compared to Transperineal Ultrasound guided biopsy for significant prostate cancer Risk Evaluation." *Contemporary Clinical Trials*, **37**(1):69–83, 2014.

[SDT20] Gokhan Sonmez, Turev Demirtas, Sevket T. Tombul, Figen Ozturk, and Abdullah Demirtas. "What is the ideal number of biopsy cores per lesion in targeted prostate biopsy?" *Prostate International*, **8**(3):112–115, 2020.

[SER20] Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, and Spyridon Bakas. "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data." *Scientific Reports*, **10**(1):12598, 2020.

[SHS21] Karthik V Sarma, Stephanie Harmon, Thomas Sanford, Holger R Roth, Ziyue Xu, Jesse Tetreault, Daguang Xu, Mona G Flores, Alex G Raman, Rushikesh

Kulkarni, Bradford J Wood, Peter L Choyke, Alan M Priester, Leonard S Marks, Steven S Raman, Dieter Enzmann, Baris Turkbey, William Speier, and Corey W Arnold. "Federated learning improves site performance in multicenter deep learning without data sharing." *Journal of the American Medical Informatics Association*, 2021.

[SKK14] Won Ik Seo, Pil Moon Kang, Dong Il Kang, Jang Ho Yoon, Wansuk Kim, and Jae Il Chung. "Cancer of the Prostate Risk Assessment (CAPRA) Preoperative Score Versus Postoperative Score (CAPRA-S): ability to predict cancer progression and decision-making regarding adjuvant therapy after radical prostatectomy." *Journal of Korean Medical Science*, **29**(9):1212–6, 2014.

[SKM10] Adnan Simsir, Erkan Kismali, Rashad Mammadov, Gurhan Gunaydin, and Cag Cal. "Is It Possible to Predict Sepsis, the Most Serious Complication in Prostate Biopsy?" *Urologia Internationalis*, **84**(4):395–399, 2010.

[SKP11] Yu Sub Sung, Heon-Ju Kwon, Bum-Woo Park, Gyunggoo Cho, Chang Kyung Lee, Kyoung-Sik Cho, and Jeong Kon Kim. "Prostate cancer detection on dynamic contrast-enhanced MRI: computer-aided diagnosis versus single perfusion parameter maps." *AJR. American journal of roentgenology*, **197**(5):1122–9, 2011.

[SLT18] Yohan Sumathipala, Nathan Lay, and Baris Turkbey. "Prostate cancer detection from multi-institution multiparametric MRIs using deep convolutional neural networks." *Journal of Medical Imaging*, **5**(04):1, 2018.

[SMJ15] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. "Cancer statistics, 2015." *CA: A Cancer Journal for Clinicians*, **65**(1):5–29, 2015.

[SNM13] Geoffrey A Sonn, Shyam Natarajan, Daniel J.A. Margolis, Malu MacAiran, Patricia Lieu, Jiaoti Huang, Frederick J Dorey, and Leonard S Marks. "Targeted Biopsy in the Detection of Prostate Cancer Using an Office Based Magnetic Resonance Ultrasound Fusion Device." *Journal of Urology*, **189**(1):86–92, 2013.

[SOS18] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. "Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images." *Medical Image Analysis*, **53**:197–207, 2018.

[SR20] Sarvesh Soni and Kirk Roberts. "Evaluation of Dataset Selection for Pre-Training and Fine-Tuning Transformer Language Models for Clinical Question Answering." In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 5532–5538, Marseille, France, 2020. European Language Resources Association.

[SRD21] Karthik V. Sarma, Alex G. Raman, Nikhil J. Dhinagar, Alan M. Priester, Stephanie Harmon, Thomas Sanford, Sherif Mehralivand, Baris Turkbey, Leonard S. Marks, Steven S. Raman, William Speier, and Corey W. Arnold. "Harnessing clinical annotations to improve deep learning performance in prostate segmentation." *PLOS ONE*, **16**(6):1–15, 2021.

[SSE06] Andrew J Stephenson, Peter T Scardino, James A Eastham, Fernando J Bianco, Zohar A Dotan, Paul A Fearn, and Michael W Kattan. "Preoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy." *Journal of the National Cancer Institute*, **98**(10):715–7, 2006.

[SSJ21] Suthida Suwanvecho, Harit Suwanrusme, Tanawat Jirakulaporn, Surasit Issarachai, Nimit Taechakraichana, Palita Lungchukiet, Wimolrat Decha, Wisanu Boonpakdee, Nittaya Thanakarn, Pattanawadee Wongrattananon, Anita M. Preininger, Metasebya Solomon, Suwei Wang, Rezzan Hekmat, Irene Dankwa-Mullan, Edward Shortliffe, Vimla L. Patel, Yull Arriaga, Gretchen Purcell Jackson, and Narongsak Kiatikajornthada. "Comparison of an oncology clinical decision-support system's recommendations with actual treatment decisions." *Journal of the American Medical Informatics Association*, **28**(4):832–838, 2021.

[SSR18] Wen Shi, Karthik V Sarma, Alex G Raman, Alan M Priester, Shyam Natarajan, William Speier, Steven S Raman, Leonard S Marks, and Corey W Arnold. "Prediction of Clinically Significant Prostate Cancer in MR/Ultrasound Guided Fusion Biopsy using Multiparametric MRI." In *Medical Imaging Meets NeurIPS Workshop*, 2018.

[STK17] Jarrel C. Y. Seah, Jennifer S. N. Tang, and Andy Kitchen. "Detection of prostate cancer on multiparametric MRI." In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, p. 1013429. SPIE, 2017.

[SYT15] Masaki Shiota, Akira Yokomizo, Ario Takeuchi, Kenjiro Imada, Keijiro Kiyoshima, Junichi Inokuchi, Katsunori Tatsugami, and Seiji Naito. "The oncological outcome and validation of Japan Cancer of the Prostate Risk Assessment score among men treated with primary androgen-deprivation therapy." *Journal of Cancer Research and Clinical Oncology*, **141**(3):495–503, 2015.

[SZ14] Karen Simonyan and Andrew Zisserman. "Two-Stream Convolutional Networks for Action Recognition in Videos." In *Advances in Neural Information Processing Systems*, pp. 568–576, 2014.

[SZ15] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In *International Conference on Learning Representations*, 2015.

[SZH20] Thomas H. Sanford, Ling Zhang, Stephanie A. Harmon, Jonathan Sackett, Dong Yang, Holger Roth, Ziyue Xu, Deepak Kesani, Sherif Mehralivand, Ronaldo H. Baroni, Tristan Barrett, Rossano Girometti, Aytekin Oto, Andrei S. Purysko, Sheng Xu, Peter A. Pinto, Daguang Xu, Bradford J. Wood, Peter L. Choyke, and Baris Turkbey. "Data Augmentation and Transfer Learning to Improve Generalizability of an Automated Prostate Segmentation Model." *American Journal of Roentgenology*, **215**(6):1403–1410, 2020.

[SZY18] Yang Song, Yu Dong Zhang, Xu Yan, Hui Liu, Minxiong Zhou, Bingwen Hu, and Guang Yang. "Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI." *Journal of Magnetic Resonance Imaging*, **48**(6):1570–1577, 2018.

[TAC10] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. "N4ITK: Improved N3 Bias Correction." *IEEE Transactions on Medical Imaging*, **29**(6):1310–1320, 2010.

[TC12] Baris Turkbey and Peter L Choyke. "Multiparametric MRI and prostate cancer diagnosis and risk stratification." *Current Opinion in Urology*, **22**(4):310–5, 2012.

[TFS20] Chad R. Tracy, Kevin J. Flynn, Daniel D. Sjoberg, Paul T. Gellhaus, Catherine M. Metz, and Behfar Ehdaie. "Optimizing MRI-targeted prostate biopsy: the diagnostic benefit of additional targeted biopsy cores." *Urologic Oncology: Seminars and Original Investigations*, 2020.

[TLP15] Panu P. Tonttila, Juha Lantto, Eija Pääkkö, Ulla Piippo, Saila Kauppila, Evelina Lammentausta, Pasi Ohtonen, and Markku H. Vaarala. "Prebiopsy Multiparametric Magnetic Resonance Imaging for Prostate Cancer Diagnosis in Biopsy-naive Men with Suspected Prostate Cancer Based on Elevated Prostate-specific Antigen Values: Results from a Randomized Prospective Blinded Controlled Trial." *European Urology*, 2015.

[TLR17] Yohannes K. Tsehay, Nathan S. Lay, Holger R. Roth, Xiaosong Wang, Jin Tae Kwak, Baris I. Turkbey, Peter A. Pinto, Brad J. Wood, and Ronald M. Summers. "Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images." In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, p. 1013405. SPIE, 2017.

[TLZ20] Zhiqiang Tian, Xiaojian Li, Yaoyue Zheng, Zhang Chen, Zhong Shi, Lizhi Liu, and Baowei Fei. "Graph-convolutional-network-based interactive prostate segmentation in MR images." *Medical Physics*, **47**(9):4164–4176, 2020.

[TMA13] Baris Turkbey, Haresh Mani, Omer Aras, Jennifer Ho, Anthony Hoang, Ardeshir R. Rastinehad, Harsh Agarwal, Vijay Shah, Marcelino Bernardo, Yuxi

Pang, Dagane Daar, Yolanda L. McKinney, W. Marston Linehan, Aradhana Kaushal, Maria J. Merino, Bradford J. Wood, Peter A. Pinto, and Peter L. Choyke. "Prostate cancer: Can multiparametric mr imaging help identify patients who are candidates for active surveillance?" *Radiology*, **268**(1):144–152, 2013.

[TRH19] Baris Turkbey, Andrew B. Rosenkrantz, Masoom A. Haider, Anwar R. Padhani, Geert Villeirs, Katarzyna J. Macura, Clare M. Tempany, Peter L. Choyke, Francois Cornud, Daniel J. Margolis, Harriet C. Thoeny, Sadhna Verma, Jelle Barentsz, and Jeffrey C. Weinreb. "Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2.", 2019.

[TWB20] Stephan Tschirdewahn, Manuel Wiesenfarth, David Bonekamp, Lukas Püllen, Henning Reis, Andrej Panic, Claudia Kesch, Christopher Darr, Jochen Heß, Francesco Giganti, Caroline M. Moore, Nika Guberina, Michael Forsting, Axel Wetter, Boris Hadaschik, and Jan Philipp Radtke. "Detection of Significant Prostate Cancer Using Target Saturation in Transperineal Magnetic Resonance Imaging/Transrectal Ultrasonographyfusion Biopsy." *European Urology Focus*, 2020.

[US18] U.S. Preventative Services Task Force. "Screening for Prostate Cancer: Recommendation Statement." *American Family Physician*, **98**(8), 2018.

[VBG08] Satish Viswanath, B Nicolas Bloch, Elisabeth Genega, Neil Rofsky, Robert Lenkinski, Jonathan Chappelow, Robert Toth, and Anant Madabhushi. "A comprehensive segmentation, registration, and cancer detection scheme on 3 Tesla in vivo prostate DCE-MRI." *MICCAI*, **11**(Pt 1):662–9, 2008.

[VBK12] P C Vos, J O Barentsz, N Karssemeijer, and H J Huisman. "Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis." *Physics in Medicine and Biology*, **57**(6):1527–42, 2012.

[VBR09] Satish Viswanath, B Nicolas Bloch, Mark Rosen, Jonathan Chappelow, Robert Toth, Neil Rofsky, Robert Lenkinski, Elisabeth Genega, Arjun Kalyanpur, and Anant Madabhushi. "Integrating Structural and Functional Imaging for Computer Assisted Detection of Prostate Cancer on Multi-Protocol In Vivo 3 Tesla MRI." *Proceedings of the SPIE*, **7260**:72603I, 2009.

[VCA08] Andrew J Vickers, Angel M Cronin, Gunnar Aus, Carl-Gustav Pihl, Charlotte Becker, Kim Pettersson, Peter T Scardino, Jonas Hugosson, and Hans Lilja. "A panel of kallikrein markers can reduce unnecessary biopsy for prostate cancer: data from the European Randomized Study of Prostate Cancer Screening in Göteborg, Sweden." *BMC Medicine*, **6**(1):19, 2008.

[VCR10] Andrew Vickers, Angel Cronin, Monique Roobol, Caroline Savage, Mari Peltola, Kim Pettersson, Peter T Scardino, Fritz Schröder, and Hans Lilja. "Reducing unnecessary biopsy during prostate cancer screening using a four-kallikrein panel: an independent replication." *Journal of Clinical Oncology*, **28**(15):2493–8, 2010.

[WBT14] Shijun Wang, Karen Burtt, Baris Turkbey, Peter Choyke, and Ronald M Summers. "Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research." *BioMed Research International*, **2014**:789561, 2014.

[WLC18] Zhiwei Wang, Chaoyue Liu, Danpeng Cheng, Liang Wang, Xin Yang, and Kwang Ting Cheng. "Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network." *IEEE Transactions on Medical Imaging*, **37**(5):1127–1139, 2018.

[WLT19] Bo Wang, Yang Lei, Sibo Tian, Tonghe Wang, Yingzi Liu, Pretesh Patel, Ashesh B. Jani, Hui Mao, Walter J. Curran, Tian Liu, and Xiaofeng Yang. "Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation." *Medical Physics*, **46**(4):1707–1718, 2019.

[WMA20] Antonio C. Westphalen, Charles E. McCulloch, Jordan M. Anaokar, Sandeep Arora, Nimrod S. Barashi, Jelle O. Barentsz, Tharakeswara K. Bathala, Leonardo K. Bittencourt, Michael T. Booker, Vaughn G. Braxton, Peter R. Carroll, David D. Casalino, Silvia D. Chang, Fergus V. Coakley, Ravjot Dhatt, Steven C. Eberhardt, Bryan R. Foster, Adam T. Froemming, Jurgen J. Fütterer, Dhakshina M. Ganeshan, Mark R. Gertner, Lori Mankowski Gettle, Sangeet Ghai, Rajan T. Gupta, Michael E. Hahn, Roozbeh Houshyar, Candice Kim, Chan Kyo Kim, Chandana Lall, Daniel J.A. Margolis, Stephen E. McRae, Aytekin Oto, Rosaleen B. Parsons, Nayana U. Patel, Peter A. Pinto, Thomas J. Polascik, Benjamin Spilseth, Juliana B. Starcevich, Varaha S. Tammisetti, Samir S. Taneja, Baris Turkbey, Sadhna Verma, John F. Ward, Christopher A. Warlick, Andrew R. Weinberger, Jinxing Yu, Ronald J. Zagoria, and Andrew B. Rosenkrantz. "Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: Experience of the society of abdominal radiology prostate cancer disease-focused panel." *Radiology*, **296**(1):76–84, 2020.

[WSG15] Annerleim Walton Diaz, Nabeel Ahmad Shakir, Arvin K George, Soroush Rais-Bahrami, Baris Turkbey, Jason T Rothwax, Lambros Stamatakis, Cheng William Hong, Mohummad Minhaj Siddiqui, Chinonyerem Okoro, Dima Raskolnikov, Daniel Su, Joanna Shih, Hui Han, Howard L Parnes, Maria J Merino, Richard M Simon, Bradford J Wood, Peter L Choyke, and Peter A Pinto. "Use of serial multiparametric magnetic resonance imaging in the management of patients with prostate cancer on active surveillance." *Urologic Oncology*, **33**(5):202.e1–7, 2015.

[WWW21] Wei Wang, Gangmin Wang, Xiaofen Wu, Xie Ding, Xuexiang Cao, Lei Wang, Jingyi Zhang, and Peijun Wang. "Automatic segmentation of prostate magnetic resonance imaging using generative adversarial networks." *Clinical Imaging*, **70**:1–9, 2021.

[XBA19] Helen Xu, John S.H. Baxter, Oguz Akin, and Diego Cantor-Rivera. "Prostate cancer detection using residual networks." *International Journal of Computer Assisted Radiology and Surgery*, **14**(10):1647–1650, 2019.

[YA18] Koichiro Yasaka and Osamu Abe. "Deep learning and artificial intelligence in radiology: Current applications and future directions." *PLOS Medicine*, **15**(11):e1002707, 2018.

[YCL17] Yading Yuan, Ming Chao, and Yeh-Chi Lo. "Automatic Skin Lesion Segmentation Using Deep Fully Convolutional Networks With Jaccard Distance." *IEEE Transactions on Medical Imaging*, **36**(9):1876–1886, 2017.

[ZPN21] Haoyue Zhang, Jennifer S. Polson, Kambiz Nael, Noriko Salamon, Bryan Yoo, Suzie El-Saden, Fabien Scalzo, William Speier, and Corey W. Arnold. "Intra-domain task-adaptive transfer learning to determine acute ischemic stroke onset time." *Computerized Medical Imaging and Graphics*, **90**:101926, 2021.

[ZSS19] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. "Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis." *MICCAI*, 2019.

[ZWY20] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J. Wood, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. "Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation." *IEEE Transactions on Medical Imaging*, **39**(7):2531–2540, 2020.