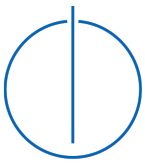
DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Robotics, Cognition, Intelligence

Automated Medical Dataset Curation: A Case Study for Zonal Detection of Clinically-Significant Prostate Cancer with Machine Learning

Patrick W. Remerscheid



DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Robotics, Cognition, Intelligence

Automated Medical Dataset Curation: A Case Study for Zonal Detection of Clinically-Significant Prostate Cancer with Machine Learning

Automatisierte Kuratierung Medizinischer Datensätze: Eine Fallstudie für die Zonale Erkennung von klinisch-signifikantem Prostatakrebs mit Maschinellem Lernen

Author:	Patrick W. Remerscheid
Supervisor:	Prof. Daniel Rückert
Advisors:	Prof. Tina Kapur, Prof. Sandy Wells, Prof. Andriy Fedorov
Submission Date:	15th of June 2023

I confirm that this master's thesis in robotics, cognition, intelligence is my own work and I have documented all sources and material used.

Boston, MA (USA), 15th of June 2023 Patrick W. Remerscheid

Acknowledgments

I would like to thank my main supervisor's at the Harvard Medical School & Brigham and Women's Hospital Tina, Sandy and Andriy for all their time, effort, encouragement and guidance over the last months.

I would like to say special thanks to Ron and Tina for accepting me as a visiting student and Daniel for accepting to supervise me from the Technical University of Munich.

Finally, I want to thank Mum, Dad, Becki, Steph and foremost my brother Nico, for always being at my side.

I want to dedicate this work to all people I know and all I don't, who have suffered from cancer, still do, or will at some point in their lives.

Abstract

The limited availability of structured and labeled medical image datasets still poses a significant hurdle for training, validation, and testing of machine learning algorithms for any task related to medical imaging, including outcome prediction and disease diagnosis [**willemink2020Prep**]. Especially the semi- and unstructured nature of medical data, as well as costly annotations - usually necessitating additional time of medical experts outside of their clinical routine - contribute significantly to the lack of medical datasets for supervised ML development.

For the detection and classification of prostate cancer, there are currently only a total of 3,096 prostate MRI scans publicly-available, acquired across multiple medical centers with varying sample sizes, image quality, and annotation standards [**sunoqrot2022**]. It is of great importance to correctly aggregate and structure all data generated during the clinical diagnosis process, to ensure adequate image quality, establish a trustworthy reference standard for all annotations, and provide as much "context" - in the form of clinical metadata - for algorithm development as possible.

In the frame of this work, we introduce a new prostate cancer image dataset - the BWH-MRIGPB dataset - comprising 86 patient studies, with 105 histopathology-confirmed point annotations of lesion candidates inside the prostate gland. Based on an in-depth analysis of the clinical workflow for MR-guided prostate biopsies at the Brigham and Women's Hospital, we introduce the task of "zonal" prostate cancer detection and provide "lesion bounding boxes" which can be extracted and constructed automatically from raw radiology reports. Furthermore, we introduce an automated data curation pipeline, leveraging a combination of simple "syntactic" and more sophisticated "semantic" modern text mining techniques, to generate a complete structured and labeled dataset from multiple, locally-available, raw input data sources without the direct need of medical experts. With the proposed framework we hope to contribute to making medical data curation more efficient, transparent, and easier maintainable, to increase the amount of publicly-available, well-curated medical data.

Contents

1 Introduction

1.1 Prostate cancer: a public health issue

Prostate cancer is the second most common cancer in American men after skin cancer [**ACSpctatistics2023**]. For the year 2023 the American Cancer Society estimates 288,300 new cases of prostate cancer and 34,700 deaths from prostate cancer in the United States alone [**ACSpctatistics2023**]. While about 1 in 8 men will be diagnosed with prostate cancer at some point in their lives, only 1 in 48 actually die from prostate cancer [**ACSpctatistics2023**].

1.2 The diagnosis of clinically-significant prostate cancer: a non-trivial problem

1.2.1 Prostate biopsies

In standard clinical practice, men with elevated PSA levels (>= 3 ng per milliliter) undergo non-targeted systematic, ultrasonography-guided, transrectal (TRUS) biopsying of the prostate, which significantly reduces mortality [**lancet2014**]. Typically 12-15 biopsy samples (cores) are taken bilaterally from the main anatomical regions of the prostate gland - the apex, base and midgland [**penzkofer2015**]. In figure 1.1 a TRUS-biopsy is schematically visualized. Studies have shown, however, that TRUS-biopsies can have significant downsides, such as being prone to diagnosing clinically-insignificant cancers, missing significant cancers, and posing a significant risk for infectious complications [**eau2013**][**jou2015**][**lancet2017**]. Alternative prostate biopsy procedures have been proposed, focusing primarily on altering the access route to the prostate gland that is used - accessing through the perineum instead of the rectum - as well as the sampling strategy - using a targeted (image-guided) sampling approach instead of a systematic one [**penzkofer2015**][**lancet2017**]. At the Brigham and Women's Hospital transperineal in-bore 3-T magnetic resonance (MR) imaging-guided prostate biopsies are performed [**penzkofer2015**], representing the primary source of data used in the curation of the prostate dataset presented in chapter 3. "MR imaging-guided"

[1] `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transrectal-biopsy`

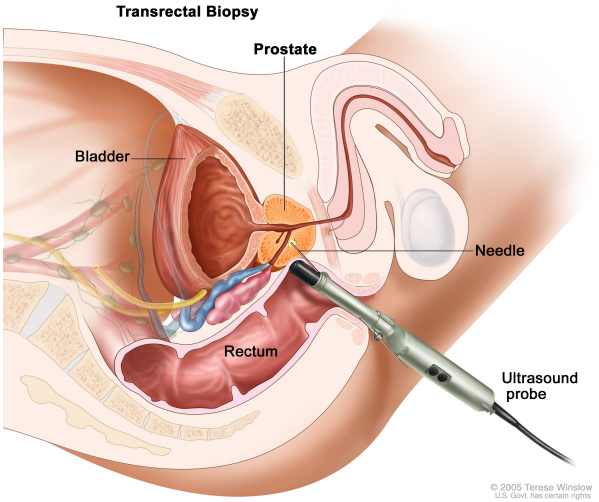


Figure 1.1: The prostate is located below the bladder and in front of the rectum. In above image an ultrasonography-guided transrectal biopsy of the prostate is visualized. Picture: National Cancer Institute[1]

refers in this context to both the use of MR imaging for the pre-biopsy identification of high risk targets - **pre-procedural**, as well as using MR imaging for the localization of potential clinically-significant prostate cancer lesions during a biopsy procedure - **intra-procedural**. Recent studies show that MRI-directed targeted biopsies can reduce the risk of overdiagnosis by half while maintaing the same recall, in comparison to standard of care systematic biopsies [**nejm2022**]. The tissue samples that are extracted at any prostate biopsy are used to evaluate cancer aggressiveness based on pathological analysis, to then decide on the significance of the prostate cancer and fitting treatment options if necessary. Traditionally, cancerous cells in the prostate tissue samples are assigned a Gleason Grade - ranging from 1 (normal prostate tissue) to 5 (mutated "high-grade" prostate tissue). A Gleason Score (GS) is derived by a Pathologist from the Gleason Grade, describing the sum of the two highest Gleason Grades in the prostate tissue sample, and respectively ranging from 2-10 [**PCFgrading2023**]. A revised grading system was proposed in 2014 by the International Society of Urological Pathology introducing so-called Grade Groups (GrG), also ranging from 1-5, that are additionally used in clinical practice for prostate cancer staging [**ajsp2014**] [**ajsp2019**](see Table1 for comparison of GS and GrG). The main question to answer from a clinical perspective

[2] https://www.mayoclinic.org/medical-professionals/urology/news/ultrasound-guided-transperineal-prostate-biopsy/mac-20473283

[3] https://www.urologynews.uk.com/features/features/post/the-role-of-transperineal-template-biopsies-in-the-

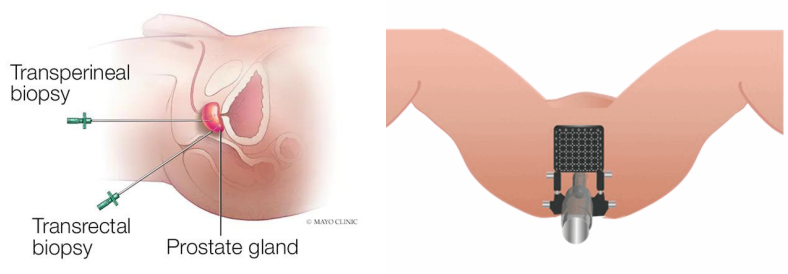


Figure 1.2: At the Brigham and Women's Hospital, transperineal MRI-guided prostate biopsies are performed. In contrast to a TRUS-biopsy the prostate is access through the perineum. On the right the "template" used by the interventional radiologist to guide the insertion of the biopsy needle is visualized. The template consists of a rigid polymer with equidistant holes through which the needle can be inserted. Left: Mayo Clinic[2], Right: Urology News[3]

is whether clinically-significant prostate cancer (csPCa) can be detected. In the frame of this thesis, prostate lesions or biopsy targets with GS >= 7 (4+3) and GrG >= 3 are defined as clinically significant.

1.2.2 Prostate MRI

Prostate Magnetic resonance (MR) imaging can significantly reduce the amount of unnecessary biopsies (around 33% in [**elwenspoek2019**]), tackling overdiagnosis and potential downsides of biopsies (see previous subsection). The Prostate Imaging Reporting & Data System (PI-RADS, most recent version at time of writing 2.1) describes the current standard for the interpretations of multiparametric MRI by Radiologists [**pirads2019**]. Here, the term "multiparametric" refers to the existence of T2-weighted images (T2W), diffusion-weighted images (DWI) and apparent diffusion coefficient maps (ADC) and dynamic contrast-enhanced (DCE) sequences in the MR imaging study. Each sequence provides different information about the imaged tissue and is combined into a multi-channel 3D-volume [**pellicer2022**]. Various parameters for image acquisition exist, including the magnetic field strength of the MR scanner used (usually either 1.5 or 3 Tesla) and the type of "coil" used for enhancing the signal (usually either a surface coil or an endorectal coil).

PI-RADS v2.1 assessment follows a 5-point scoring system, where each point refers to a likelihood of clinically significant prostate cancer for any given lesion in the prostate

gland on multiparametric MRI. The clinical significance of a lesion in the frame of the PI-RADS v2.1 assessment is based on pathological analysis with Gleason Scores >= 7, as described in the previous subsection [**pirads2019**].

- PI-RADS 1: Very low (clinically significant cancer is highly unlikely to be present)
- PI-RADS 2: Low (clinically significant cancer is unlikely to be present)
- PI-RADS 3: Intermediate (the presence of clinically significant cancer is equivocal)
- PI-RADS 4: High (clinically significant cancer is likely to be present)
- PI-RADS 5: Very high (clinically significant cancer is highly likely to be present)

While PI-RADS does not explicitly include decision support, in terms of for example proposing suitable further treatment options and patient management based on its scoring system, it is suggested that for PI-RADS scores 4 and 5 histopathology confirmation in form of prostate biopsies are conducted, while the opposite is suggested for PI-RADS scores of 1 and 2. A PI-RADS score of 3 is in general regarded as indecisive and further clinical information needs to be taken into consideration for deciding on next steps in patient treatment [**pirads2019**]. PI-RADS assessment - while being an important step in the direction of standardized analysis and interpretation of prostate MRI - still has important shortcomings, such as relatively low positive predicted values and significant performance differences across centers (estimated PPV of 49% for PI-RADS of >=4 according to [**westphalen2020**]), motivating the development of computer aided diagnosis (CAD) and detection algorithms for clinically-significant prostate cancer.

1.2.3 Standard of care at the Brigham and Women's Hospital

The clinical workflow for the analysis and diagnosis of clinically-significant prostate cancer at the Brigham and Women's Hospital, consists of following steps:

1. **Pre-procedural MRI interpretation:** first, an expert radiologist looks at the diagnostic - pre-procedural - MRI scans of a patient with increased probability of having clinically-significant prostate cancer. Reasons range from increased PSA levels, to previous inconclusive prostate biopsies (Gleason Score of 6) or even a follow-up check after previous treatment of the prostate. The radiologist defines in the findings and impression sections of a pre-procedural radiology report suspicious lesions. For characterization of the lesion candidates the anatomical zone (see 1.3), the estimated lesion size and the aggressiveness in-terms of PI-RADS score are usually noted down.

2. **Interventional radiologist decides on regions of the prostate to biopsy:** before the biopsy procedure, the interventional radiologist reads the pre-procedural radiology report and looks at some of the pre-procedural MRI scans himself, to then decide what region of the prostate to biopsy (see 1.3). All biopsy regions are noted down as abbreviated anatomical zones of the prostate on a physical piece of paper and handed to an assistant, operating the software tool - Slicer[4] - used to set the "biopsy targets" on the intra-procedural MRI during the procedure. In addition to this standard clinical workflow, for the development of the BWH-MRIGPB dataset, additionally so-called "pre-procedural biopsy targets" were set as point coordinates on the axial T2-weighted image series of the pre-procedural MRI study. These biopsy coordinates are not necessary for the actual clinical routine but can be used as point of references for the later histopatholgy results of the extracted tissue.

3. **Targeted prostate biopsy:** as first step of the MRI-guided prostate biopsy procedure, the so-called "intra-procedural targets" are set by an assistant on a generated intra-procedural MRI, guided by the interventional radiologst. These point coordinates are saved by the assistant / operator and stored as .fcsv files with the abbreviated anatomical zone description as label. As second step, the insertion of the biopsy needle is planned by using a digital replica of the actual "template" used by the interventional radiologist to guide the insertion process (see 1.2 on the right). The virtual replica of the actual template is overlayed over the MRI scan and the intra-procedural biopsy target and the closest coordinates on the template are used by the interventional radiologist for a first insertion attempt. Usually multiple needle insertions are necessary until the required prostate region is attained (after every insertion a new MRI scan is generated and inspected to look at the current position of the needle in the prostate). It should be noted, that after every insertion the prostate tissue is significantly deformed, making the original pre-procedural targets significantly less accurate as point of reference for the histopathology of the extracted tissue.

The above clinical workflow description for diagnosing and biopsying suspicious prostate cancer lesions was derived on-site from two in-person biopsy case observations.

[4] `https://www.slicer.org/`
[5] `https://radiologyassistant.nl/abdomen/prostate/prostate-cancer-pi-rads-v2`

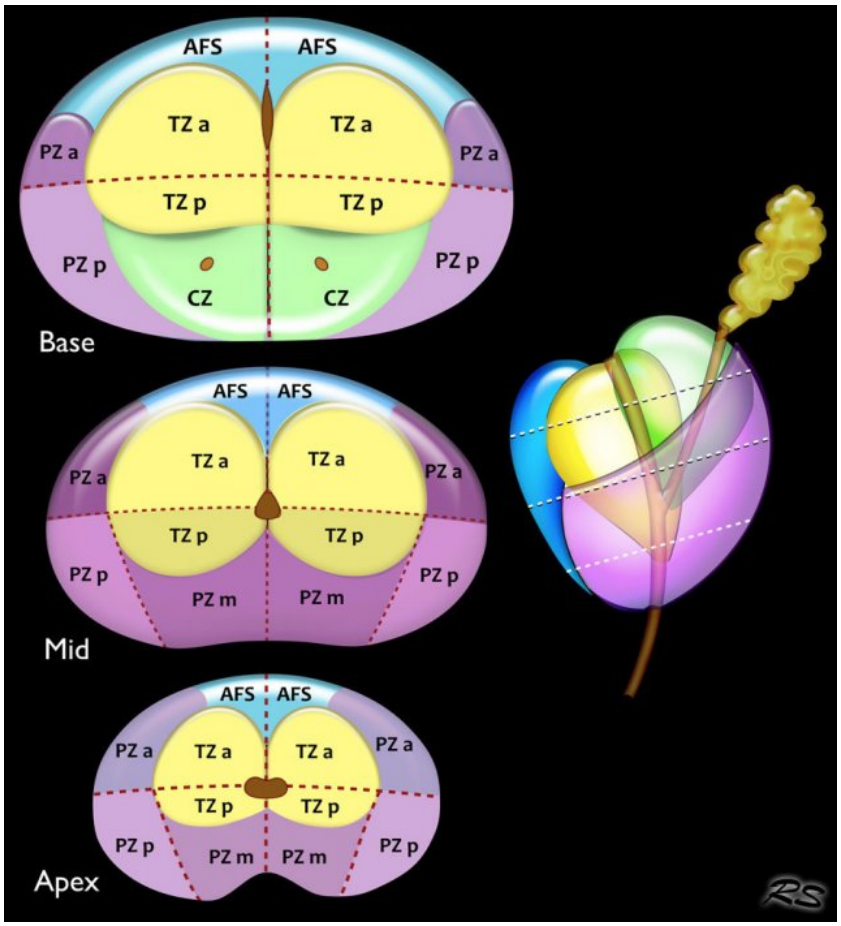


Figure 1.3: Anatomical zones of the prostate[5], used during an MRI-guided targeted biopsy for labeling the set biopsy "targets" with an abbreviated version, for example "RPZpmMid" for "Right Peripheral Zone posterior medial Mid".

1.2.4 Computer-aided diagnosis of clinically significant prostate cancer

The PI-RADS v2 assessment of prostate MRI has a moderate intra-reader reproducibility (scoring agreement within a single Radiologist) and a poor inter-reader reproducibility (scoring agreement between multiple Radiologists) [**smith2019**]. Results are dependant on the expertise of the Radiologist, as well as on the location of the lesion within the prostate gland (higher inter-reader reproducibility for the periphal zone (PZ) of the prosate than for transitional zone (TZ) [**muller2015**]).

Modern computer-aided detection and diagnosis (CAD) systems are machine learning

based approaches that use past image and / or non-image data from a patient population to model, assess and predict certain disease outcomes [**chan2020**]. In clinical practice they can for instance be used as a "second opinion" by a Radiologist in the image interpretation process [**chan2020**], and have the potential to increase the diagnostic accuracy of detecting clinically-significant prostate cancer lesions, reduce inter-reader variability and shorten reading times for Radiologists [**winkel2021**]. Especially, with the introduction of Convolutional Neural Networks (CNNs) to medical image analysis tasks [**saha2021**], deep learning-based computer-aided detection algorithms (DL-CAD) have achieved up to human-level performances. Examples include classification of skin cancer [**esteva2017**] on images of skin cancer lesions and breast cancer prediction [**mckinney2020**] using a large collection of mammograms.

1.3 The importance of well-curated medical datasets

1.3.1 Medical data curation is hard

Non-medical image datasets such as Pascal Visual Object Classes (VOC) introduced by Everingham et al. in 2005 (20 classes introduced in 2007 in [**pascal-voc-2007**]) and the following ImageNet dataset [**deng15imagenet**] introduced in 2009 by Deng et al. have had a dramatic impact on the evolution of Computer Vision Machine Learning models. As analyzed by Russakovsky et al. in 2015 [**russakovsky2015imagenet**] the error rate on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) containing nearly 15 million hand-labeled images reduced by 4.2x for image classification and by 1.7x for single-object localization error from 2010 to 2014.

The importance of curated and labeled datasets for the training of medical machine learning models is uncontested, yet the number and size of medical datasets is much more limited (see next subsection). A main reason for the reduced amount of medical datasets is the significantly more challenging data curation task. Data curation challenges in the medical domain include the sensitive nature of imaging data and necessary privacy regulations, as well as a complex and time-consuming labelling process [**simpson2019large**]. In contrast to the ImageNet dataset it is non-trivial to access medical images in a privacy-preserving manner and crowdsourcing the labeling process via tools such as Amazon Mechanical Turk[6] is significantly more difficult due to the complex nature of the medical phenomena themselves (identifying a clinically-significant prostate cancer lesion on multiple MRI sequences is not as straight forward as distinguishing a cat from a dog). While for the curation of ImageNet, annotators on

[6] Accessed on 3th of June 2023: `https://www.mturk.com/`

Amazon Mechanical Turk did not require expert knowledge and were able to annotate **50 images per minute**, the curation of medical data requires both expert knowledge and substantially more time.

1.3.2 Lack of "suitable" image datasets related to prostate cancer

There are currently a total of 3,369 prostate MRI cases available in 17 public datasets with a total size of 253 GB (incl. metadata and annotations) [**sunoqrot2022**]. A number of complications exist for using the data for training and evaluation of algorithms however, including a varying reference standard for annotations, data overlap between public datasets, relatively small sample sizes, varying levels of de-identification and available clinical variables, and potential bias in datasets (e.g. due to varying patient inclusion criteria) [**sunoqrot2022**].

1.4 Key challenges addressed in this work

Dataset related:

- **lacking clear and verifiable annotation reference standard:** annotations in medical datasets are usually difficult for non-medical experts to verify. As computer-scientists are in general not able to verify themselves whether a lesion segmentation correctly delineates a cancerous lesion, reference standards are important indicators of the validity of provded lesion annotations. For prostate cancer lesion segmentations on MRI scans a good reference standard would be including in addition to lesion segmentations the histopathology results from prostate biopsies, as these are used in clinical practice to decide upon further treatment. While the PI-CAI challenge dataset [**PICAI_BIAS**] provides general Gleason Scores for each case as validation for clinically-significant prostate cancer cases, the majority of publicly available prostate dataset do not include histopathology (see prostate158 next chapter).
- **limited considerations about "AI readiness":** while the three prostate datasets described in chapter 2 were specifically curated for the training of deep learning-based algorithms, a significant amount of publicly-available datasets are not. The lack of "AI readiness" manifests itself for example in datasets with very small sample sizes that make training of "generalizable" models challenging (e.g. TCGA-PRAD n=10 [**sunoqrot2022**]).
- **lacking considerations about clinical "context":** the majority of publicly-available prostate cancer datasets do not provide any clinical metadata, such as PSA

levels, prostate volume or patient age. While still not a lot of publicly-available deep learning-based algorithms leverage clinical metadata, we hypothesize that providing algorithms with similar information as expert radiologist have to their disposal when interpreting pre-procedural MRI scans, might increase overall model performance.

- **lacking considerations about clinical "implementation":** the eventual goal of training and testing algorithms on medical datasets for various tasks, should be the actual implementation of solutions in the clinical workflow. Hence, we argue that the curation of datasets should be coupled to a specific "machine learning task", and in addition consider possible applications in clinical workflows (for example MRI-guided prostate biopsies). No currently available prostate cancer dataset further describes or considers clinical workflows in which the trained algorithms might be used in.

- **no considerations about (manual) data exploration:** with a strong focus of the research community on training and testing deep learning-based algorithm, the importance of also visually and manually being able to easily explore and inspect specific cases or parts of the dataset gets neglected. All currently publicly-available prostate cancer datasets do not provide an easy way to visualize the dataset, and most do not provide in-depth characteristics or quality considerations.

- **no considerations about expand-ability:** with every prostate biopsy procedure at the Brigham and Women's Hospital new imaging studies and clinical reports are generated. Hence, we argue, that curated medical imaging datasets for algorithm development should be made significantly easier to "grow continuously", to adapt to the way medical imaging data is generated. None of the currently provided prostate cancer datasets demonstrate an easy expand-ability, or in fact have significantly changed in size since their initial publication.

Dataset curation related:

- **time-consuming dataset structuring and labeling:** curating medical imaging datasets is still a very time-consuming and tedious work. The high amount of different data elements that need to be structured and matched (for example image series, diagnostic scores, patient demographics, etc.), as well as the manual annotation process, take a significant amount of time.

- **medical expert needed for dataset annotation:** medical diagnoses are per definition non-trivial tasks and require in general medical experts. All currently

publicly-available prostate cancer datasets with clinically-significant prostate cancer lesion segmentations, require expert radiologists for annotation on provided MRI studies.

- **limited insight into details of data curation process:** for most of publicly available prostate cancer datasets no detailed information about the dataset curation process is provided. We argue, that insights about the curation process, such as number and years of experience of radiologists responsible for lesion annotations, extraction method from hospital information systems, or criteria for discarding or selecting imaging studies, are important indicators of the quality of the final resulting dataset.
- **minimal usage of existing clinical reports:** while in general there exists for every analyzed imaging study a dedicated radiology report, that is created as part of the clinical workflow, most public prostate cancer datasets do not actively integrate them into the dataset curation process. We argue, that radiology report should be used more extensively for algorithm development, as they represent "free-of-charge" - because generated as part of the clinical workflow - already existing expert knowledge.
- **not repeatable:** this challenge is related to the dataset challenge mentioned above, describing a lack of "expandable medical datasets". As the expand-ability of a dataset primarily depends on the used curation approach, it can be assumed from pure observation of the growth rate of public prostate cancer datasets, that data curation pipelines are not "efficiently repeatable" or capable of efficiently continuously adding data to the published datasets.
- **Not generalizable to other (similar) data curation tasks:** none of the currently existing prostate cancer datasets, provide in addition to the published dataset, instructions, methods or code that can be used by other institutions to simplify the curation of their raw data.

1.5 Main contributions

First, we introduce a new prostate cancer dataset - the BWH-MRIGPB dataset - with following core characteristics, addressing the key challenges described above:

- **clear reference standard for provided annotations:** all included cases in the BWH-MRIGPB dataset come with histopathology results. For a subset of cases the histopathology can even be directly compared to pre-procedurally assigned diagnostics, such as PI-RADS scores and estimated lesion diameter.

- **"AI ready" dataset:** the BWH-MRIGPB dataset was curated with specific publicly-available deep learning-based algorithms for the detection of clinically-significant prostate cancer lesions in mind. In chapter 6 an example application of an algorithm is demonstrated.
- **extensive clinical metadata:** the BWH-MRIGPB dataset provides 13 relevant clinical variables extracted form clinical reports (including prior therapy and recent prior biopsies), doubling the number of the closest publicly-avaialable prostate cancer dataset 7.2.
- **clear defined purpose and clinical application of dataset:** in chapter 3, we define a clear "algorithmic task" that the BWH-MRIGPB dataset is primarily curated for, as well as a specific clinical workflow in which algorithms that were developed with our dataset, might be applicable to.
- **web-based interactive dataset viewer:** in order to enable easy exploration and inspection of all cases in the BWH-MRIGPB dataset, an all DICOM-based version of the dataset can be visualized via OHIF[7] webviewer. No local download, or specialized software needed.
- **simple "automated" dataset expand-ability:** the BWH-MRIGPB dataset dedicated "code-based" curation pipeline described in chapter 4, enables an automatic "growing" of the final structured and curated dataset, with only changing the locally-available raw data input.

Second, an approach to automated dataset structuring and labeling - used to curate the first version of the BWH-MRIGPB dataset - is presented, with following core characteristics, addressing the key challenges described above:

- **automated dataset structuring and labeling:** we significantly reduce the time needed for structuring and labeling the BWH-MRIGPB dataset, by automating dataset curation tasks, such as image series retrieval and histopathology matching.
- **"debuggable" curation pipeline:** by providing an end-to-end code pipeline for structuring and labeling the BWH-MRIGPB dataset, we are able to visualize and inspect the dataset at all intermediary steps. All applied methods are traceable and modifiable, guaranteeing transparency.
- **report-guided annotations:** the proposed data curation pipeline in chapter 4, does not require additional time from expert radiologists. By processing radiology

[7] `https://ohif.org/`

and pathology reports and extracting clinical keywords, all provided annotations can be constructed automatically.

- **publicly-available modularized python script:** the code implementation of our data curation pipeline is planned to be published to a publicly-accessible GitHub[8] repository, after careful verifying no sensitive information remain in any of the scripts. The hope is to provide researchers from similar medical institutions with boilerplate code that can be used to simplify data curation for their specific use case as well.

[8] https://github.com/

2 Related works

The following chapter introduces the three main imaging datasets, currently available for Machine Learning model development and benchmarking related to prostate cancer diagnosis. In addition, existing methods for medical data curation are discussed.

2.1 Public datasets for detecting clinically-significant prostate cancer

2.1.1 PI-CAI challenge dataset

The "Prostate Imaging: Cancer AI" challenge - in short PI-CAI challenge - is a recent effort from Radboud University Medical Center, Ziekenhuis Groep Twente, University Medical Center Groningen and the Norwegian University of Science and Technology, for developing and validating AI-based algorithms for the detection and diagnosis of clinically-signifcant Prostate Cancer (csPCa) in biparametric MRI [**PICAI_BIAS**]. Biparametric MRI (bpMRI) omits DCE sequences in contrast to mpMRI (see chapter 1). PI-CAI is subdivided into a reader study, challenging an international group of radiologists with varying degrees of experience at csPCa diagnosis, and an AI study, enabling the training and benchmarking of DL-CAD systems for csPCa detection and diagnosis. The underlying objective is to compare human vs. machine performance, and accelerate the clinical translation of prostate AI solutions.

Imaging data and annotations

The imaging data curated from the four above mentioned institutions was acquired in a time frame from 2012-2021 [**sunoqrot2022**] with scanners from Siemens and Philips, using a field strength of 3T and 1.5T with surface coils (see chapter 1 for brief explanation of MRI acquisition protocols and sequences). Triplanar (axial, sagittal, coronal) T2W, axial high b-value DWI (>= 1000 s/mm2) and axial ADC sequences are included. The cohort is structured into four independent subsets as follows:

- **1500 cases | publicly-accessible training dataset for algorithm development** This subset can be used to train DL-CAD systems. Of the 1500 cases, 328

cases are related to the "ProstateX Challenge" [**litjens2014**][**litjens2017**] (see next subsection). All patient cases have at least axial T2W, DWI and ADC sequences. For some cases additionally, coronal and sagittal T2W images are included. No DCE sequences are present. No co-registration is performed between image sequences.
Human expert-derived annotations have been created for the Public Training and Development dataset. Out of the 1500 cases, 1075 have benign tissue or indolent PCa, while 425 cases have csPCa. Only 220 cases out of these 425 have annotations derived by human experts. The remaining 205 positive cases are not annotated. The annotations have been provided in two formats, the original annotations and resampled annotations (to T2W image series). AI-derived annotations for csPCa lesion and whole-gland segmentations of the prostate have been released, building on [**Bosma2022**].

- **7607 cases | private training dataset** Imaging data is identical to the 1500 (82 times 1.5T field strength + 1,418 3T field strength [**sunoqrot2022**]) cases of the publicly-accessible training dataset. This dataset is only accessible to the challenge organizers and is used for retraining the best five ranked algorithms after the public training phase is over. No co-registration is performed between image sequences is performed. No publicly released annotations available.

- **1000 cases | hidden testing cohort** Five image sequences per patient available: triplanar T2W, axial DWI, axial ADC. Used for identifying the final top five best performing AI algorithms and compare them to the performance of Radiologists. The reader study will include a subset of 400 cases, with occasionally added DCE sequences (only available for human / Radiologist interpretation). Include internal (unseen data from seen centers) and external test data (unseen data from unseen centers). Co-registration between sequences is performed. No publicly released annotations available.

- **100 cases | hidden validation and tuning dataset** Imaging data is identical to the 1000 cases of the hidden testing dataset. Used for displaying model performances on a public leaderboard during the open development phase. Co-registration between sequences is performed. No publicly released annotations available.

Clinical variables

In addition to the image data the following clinical variables are provided:

- PSA

- prostate volume
- PSA density (:= PSA / prostate volume)
- patient age
- MRI scanner manufacturer
- MRI scanner model name
- highest b-value of DWI scan

Lack of clinical variables for some cases are due to lack of reporting during clinical routine, and lack of automatic retrospective calculation (e.g. PSA density calculation with prostate volume and PSA level) or by expert Radiologist.

Reference standard

A strong reference standard is crucial for accurate validation of AI and human-reader performance in detecting csPCa. The PI-CAI reference standard defines the ground-truth for every case in the validation and testing cohorts, using histologically-confirmed positives (ISUP 2) and negatives (ISUP 1 or PI-RADS 2) with follow-up (3 years). Negative cases are confirmed with follow-up data and expert inspection. Training datasets are annotated using histopathology (ISUP 2) positives and negatives (ISUP 1 or PI-RADS 2) without follow-up (same reference standard as ProstateX).

Six pre-trained deep learning-based algorithms provided in the frame of the challenge. These - together with a detailed description of the evaluation metrics used for benchmarking the algorithms - will be described in subsection 2.2.

Baseline algorithms

The PI-CAI challenge introduces three baseline models together with their training dataset, each coming in two variations (supervised and semi-supervised):

- **U-Net (supervised and semi-supervised)**: Based on the original U-Net model architecture, introduced by [**unet2015**], and trained on the "Public Training and Development Dataset" with (semi-supervised) and without additional "AI-generated" labels. "AI-generated" refers to an approach of using report-guided labeling to increase the amount of labels for model training (especially relevant in the medical domain as labeling process is often expert-dependant and time-consuming), created by [**bosma2022annotationefficient**] and further described in the next subsection.

- **nnU-Net (supervised and semi-supervised)**: Based on the original nnU-Net model architecture, introduced by [**Isensee2021**], and trained on the "Public Training and Development Dataset" with (semi-supervised) and without (supervised) additional automatically generated labels.
- **nnDetection (supervised and semi-supervised)**: Based on the original nnDetection model architecture, introduced by [**Baumgartner2021**], and trained on the "Public Training and Development Dataset" with (semi-supervised) and without (supervised) additional automatically generated labels.

All provided models are required to output "detection maps" of lesion candidates, including their probability of harbouring clinically-significant prostate cancer. "Detection maps" is a losely defined term, referring to segmentations of clearly delineated 3D areas on the MRI scan, corresponding to lesion candidates. Segmentations are converted from softmax probability maps (received as model output), by setting a certain confidence / probability threshold and averaging over all values for a certain area, to obtain detection maps. In addition to lesion-level detection maps, the lesion candidate with the highest likelihood of harbouring clinically-significant prostate cancer is used as "patient-level risk score", describing the overall cancer likelihood for a patient. The created detection map can look quite different, depending on what kind of detection model is used. While the U-Net and nnU-Net models output pixel-level segmentations, the nnDetection framework outputs rectangular bounding boxes that highlight areas in the prostate with a certain likelihood of harboring clinically-significant prostate cancer.

The lesion-level detection performance is evaluated using precision-recall and free-response receiver operating characeristic curves. Object-level "hits" (True Positives) are defined with a required minimum prediction-ground truth "overlap", using Intersection over Union (IoU) as metric measuring the overlap of two 3D volumes. The default threshold for a true positive ("hit"), set by the PI-CAI project organizers is set to **IoU = 0.1**. The Dice Similarity Coefficient (DSC) - as another potential metric to measure lesion-ground truth overlap - is considered less suitable in this case, due to the small size of tumor lesions (making DSC less suitable, see [**Taha2015**]).

The pre-trained baseline version of the semi-supervised nnDetection model is selected and applied to the BWH-MRIGPB dataset in the discussion section in chapter 6.

2.1.2 PROSTATEx challenge dataset

The PROSTATEx dataset **[litjens2014]** **[litjens2017]** comprises T2-weighted, proton density-weighted, dynamic contrast enhanced, and diffusion-weighted imaging from two different types of Siemens 3T MR scanners, the MAGNETOM Trio and Skyra. The dataset contains 346 participants, 349 studies, 18,321 series, and 309,251 images, with a total image size of 15.1 GB.

The ProstateX dataset was used as a benchmark for the SPIE-AAPM-NCI Prostate MR Classification Challenge and the SPIE-AAPM-NCI Prostate MR Gleason Grade Group Challenge. The dataset has been superseded by the PI-CAI Public Training and Development dataset, which includes the ProstateX dataset and over 10,000 curated prostate MRI exams to validate modern AI algorithms and estimate radiologists' performance at clinically significant prostate cancer detection and diagnosis.

2.1.3 Prostate158 dataset

The Prostate158 is the third publicly-available prostate related imaging dataset, that we will use as direct comparison to the in chapter 3 introduced BWH-MRIGPB dataset. It comprises 158 3T acquired axial T2-weighted and diffusion-weighted (unclear b-values) image studies, with axial ADC maps, all curated at Charité-Universitätsmedizin Berlin (CUB). A train-test split of 139-19 is provided, with all image studies made available via zenodo.org platform.

Zonal prostate segmentations are provided for 139 cases, and lesion segmentations for 83 cases. All lesion segmentations were created by at least one of two radiologists on either axial T2-weighted or diffusion-weighted MRI sequences. Volumes and segmentation maps come in "nii.gz" format. Together with the curated dataset, two pre-trained U-Net base models are provided with model weights available via zenodo.org.

2.2 Automated Medical Dataset Curation

Different lines of research have been explored to facilitate medical data curation. Concerning the task to aggregate distributed data across departments and hospitals, especially efforts in Privacy-Preserving ML including Federated Learning and Differential Privacy have been explored to allow for distributed data access [**kaissis2021**; **ziller2021differentially**]. For the purpose of this thesis, however, we assume the data to be already locally available and focus on the annotation of available medical images based on unstrucured data. Specifically, we use medical text mining techniques and

semi-structured medical reports to structure and annotate medical images that can then be used to train downstream image models.

2.2.1 Report-guided training of computer vision models

Instead of using explicit labels to train image models, multiple lines of work focus on using medical reports as a supervision signal for model training. The information contained in medical reports promise to contain relevant information for computer vision models since they are used to transmit information between different medical experts in clinial workflows. As highlighted in chapter three interventional radiologists only use pre-procedural radiology reports and scans of the imaging study to locate the tumor - generally no visual annotations on the transmitted scans are provided. This suggests that semi-structured medical reports have much signal that could be used to supervise the downstream model training and hence circumvent the need for expensive manual image annotations by experts. Multiple different lines of work search to extract and use this signal to enhance the performance of downstream medical computer vision models in an un-, semi-, or self-spervised learning setting.

Bosma et al. showed that specifically for the task of prostate cancer detection semi-supervised learning techniques taking into account medical reports can lead to significant performance increases [**bosma2022annotationefficient**]. The authors assume the availability of (a) a labeled subset on which an expert model is trained, and (b) corresponding medical reports from which PI-RADS score can be automatically extracted. The number of significant PI-RADS scores is used to determine the number of top candidates to keep from the expert models preliminary prediction for an unlabeled sample. Finally, the model is trained on the original labeled subset and on the unlabeled subset with predicted annotations. Although the performance of the final model increases this approach is specifically taylored to the classification task, the extracted PI-RADS scores depict only a fraction of the information contained in the medical reports, and the regex-based extraction process can only be applied to the specific structure of radiology reports.

Other approaches, such as the work by Liao et al. in 2021 [**Liao_2021**], try to learn representations that take into account the relation between the clinical images and reports. Liao et al. show that the information in clinical texts can successfully be used to learn better image representations that can improve the performance on downstream classification tasks. Models that have been initialized with representation learning perform nearly as well as the models that have been trained on the labeled images, and models that have been both initialized with representation learning and fine-tuned on

the labeled image dataset perform the best.

These works show that the information in clinical texts can indeed be used to improve the performance of models trained on related data. However, their applicability is limited to the specific tasks they were conceived for, and given the noisy nature of medical reports it remains unclear how effective these approaches leverage the given signal encoded in the medical reports.

2.2.2 Medical text mining *applications*

Medical text mining is frequently used to exrtact important information from semi- or unstructured medical data such as the above-mentioned medical reports. As mentioned by Nair et al. and Bose et al. many classical and modern Machine Learning methods are already being applied to a multitude of text mining applications, including Information Extraction as for example Named Entity Recognition, and Classification which are also used in this work [**Nair2014ASO**][**bose2021**]. To date, applications of medical text mining focus on extracting information from scientific papers, EHR systems or unstructured medical reports to accelerate further reserach [**Zhu2013BiomedicalTM**], remove Personally Identifiable Information [**Pearson2021NamedER**], power medical decision support systems [**Sun2018DataPA**], amongst other applications.

2.2.3 Our position and background

Works that include medical reports directly into the training process of a specific computer vision model - as is done by the works mentioned in 2.2.1 - can only be partially applied to the training of other model groups or other tasks. Moreover, embedding the extraction of information into the final model training makes it challenging to extract all of the signal while discarding all of the noise from the medical reports. For instance, while Bosma et al. use only a fraction of the whole signal by considering only the number of signficant PIRADS scores, Liao et al. consider all of the signal but also all of the noise making the learning of a mapping between reports and images difficult.
In order to extract all information with as few noise as possible we consider the medical reports as a semi-structured data source which allows us to delimite noisy sections manually beforehand. We propose a trade-off between both works that uses multiple medical text mining techniques 2.2.2 to extract important information and match them with the corresponding images. Furthermore, instead of embedding these semi-manual filters into a model training pipeline we use them to create a structured image dataset with high quality annotations. This allows to convert the raw medical image and text data into a dataset whose annotations can be verified by experts and used for the

supervised learning of a multitude of models for a multitude of different tasks. To the best of our knwoledge this is the first work that applies modern medical text mining techniques to directly structure and annotate medical image datasets.

3 The BWH-MRIGPB dataset

In the following the Brigham and Women's Hospital MRI-guided prostate biopsy dataset - in short BWH-MRIGPB dataset - will be presented. The curation process itself will be described in more detail in chapter 4.

3.1 Background & summary

The complete BWH-MRIGPB dataset comprises a cohort of 86 prostate MRI exams, curated exclusively from the Brigham and Women's Hospital in Boston, USA, between July 2015 and October 2018. All image studies are of men suspected of having clinically-significant prostate cancer and having had one or more targeted transperineal MRI-guided prostate biopsies done, confirming or rejecting (or in some cases turning out inconclusive) the diagnosis of clinically-significant prostate cancer (see chapter one for standard biopsy procedure at BWH). Before biopsying, pre-procedural diagnostic MRI scans were acquired and interpreted by expert radiologists.

Imaging studies for all patients comprise of the following three MRI sequences, that are also used in current clinical practice by expert radiologists to assess (see PI-RADS section in chapter 1) the presence and severity of clinically-significant prostate cancer lesions:

- Axial T2-weighted imaging (T2W)
- Axial high b-value (1400 s/mm$\hat{2}$) diffusion-weighted imaging (DWI)
- Axial apparent diffusion coefficient maps (ADC)

Following the approach taken in the PI-CAI challenge (see chapter 2), only biparametric MRI scans are included in the current version of the dataset, excluding contrast-enhanced (DCE) sequences. Multiple scanner manufacturers and models were used to acquire the MRI sequences, including the "DISCOVERY MR750w" scanner from GE Healthcare and the "Prisma" scanner from SIEMENS (see next section for more details about used acquisition parameters). No image registration between the three MRI sequences was performed (again following the reasoning and approach applied in

the PI-CAI challenge). All imaging studies were acquired prior to the actual biopsy procedure, and refer to the diagnostic, pre-procedural, MRI scans also used for analysis by the expert radiologists. All intra-procedural imaging studies - i.e. acquired during the targeted MRI-guided biopsy procedure - were excluded, as they are in general not used for diagnosis of clinically-significant prostate cancer (but for orientation and guidance during the biopsy procedure) and hence are usually of significantly worse quality.

For all imaging studies one or multiple biopsy targets - in total 105 for 84 patients and 86 image studies - are available as point coordinates, including the histopathology result (Gleason score and Gleason Grade Group) of the extracted tissue probe during the targeted MRI-guided biopsy. The biopsy targets were placed with respect to the pre-procedural diagnostic axial T2W scan before the actual procedure took place. In addition to the biopsy targets with corresponding histopathology, 99 out of the 105 target coordinates were used to construct lesion bounding boxes with assigned PI-RADS score, using the lesion diameter and PI-RADS scoring mentioned in the radiology reports, accompanying the pre-procedural diagnostic image study before the biopsy took place. Because of the time-consuming and expensive nature of creating them, no direct expert-annotated lesion segmentations / delineations are available at this point. For details about the label creation process and quality considerations, see chapters 4 and 6.

3.2 Patient cohort & clinical metadata

3.2.1 Dataset characteristics

As visualized in table 7.2, around two thirds of the included prostate MRI exams correspond to cases of benign or indolent prostate cancer, while around one third of exams correspond to cases with clinically-significant prostate cancer. The Gleason Grade Group (described as "ISUP" in table 7.2) is used as main metric for a study to be classified as a clinically-significant versus insignificant case and is based on the histopathology result of the biopsy targets extracted during the biopsy procedure. The biopsy target with the highest ISUP-value (max is 5) determines the classification of the study and all MRI exams including at least one biopsy target with ISUP >= 2, are considered "significant". There are in total 93 positive (assigned PI-RADS score of >=3) lesions pre-procedurally identified by expert radiologists, of which 47 were assigned an ISUP score of >= 1, while the rest were classified as "benign" according to the histopathology findings. Figure 3.1, depicts respectively the distributions of pre-procedurally assigned PI-RADS scores and post-procedurally assigned Gleason scores.

While around one third of the patient population did not have any prior treatment or

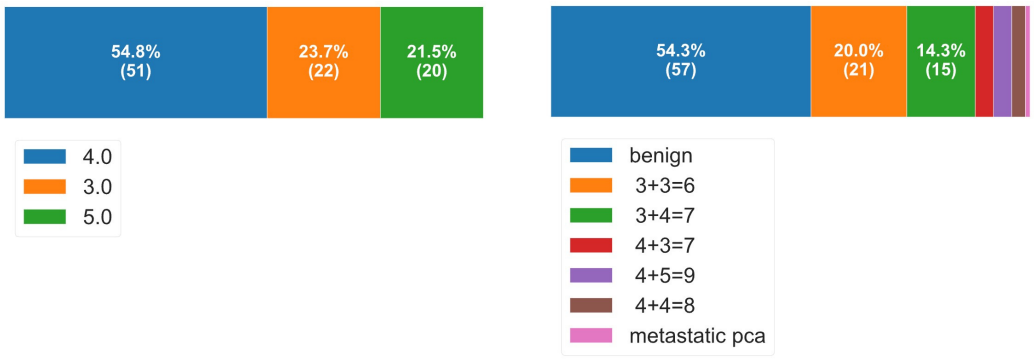


Figure 3.1: Left: Distribution of PI-RADS scores 3,4 and 5, Right: lesion Gleason scores (3+3, 3+4, 4+3, 4+4, 4+5=9) in final dataset. For one case "metastatic prostate cancer" was diagnosed.

biopsy procedure before, a significant amount of patients underwent prior biopsies or even various cancer or non-cancer related treatments of the prostate, as depicted on figure 3.2. The median age of the underlying patient population is at 66 years, the median PSA at 7.5 ng/ml and the median prostate volume prostate volume at around 52 mL (see table 7.2 for IQRs). In figure 3.3, the race distribution of the underlying patient population of the dataset is visualized, showing a majority of white patients with only a small number of black and asian patients. A complete overview of the BWH-MRIGPB dataset core characteristics are listed in:

- **Image acquisition parameters (figures 7.19, 7.20, 7.21, 7.22, 7.23, 7.24, 7.25, 7.26)**: describing the hardware that was used to acquire every curated MRI sequence, including scanner type and manufacturer, magnetic field strength and whether an endorectal coil was used or not. All information was manually extracted from the DICOM metadata of the automatically extracted imaging studies.
- **Clinical metadata (figures 7.8, 7.9, 7.10, 7.11, 7.12, 7.13, 7.14, 7.15, 7.16, 7.17, 7.18)**: describing relevant clinical metadata and patient information that was manually extracted from the DICOM metadata of all MRI sequences, and the raw pre-procedural radiology reports, accompanying the automatically extracted imaging studies (see chapter 4).

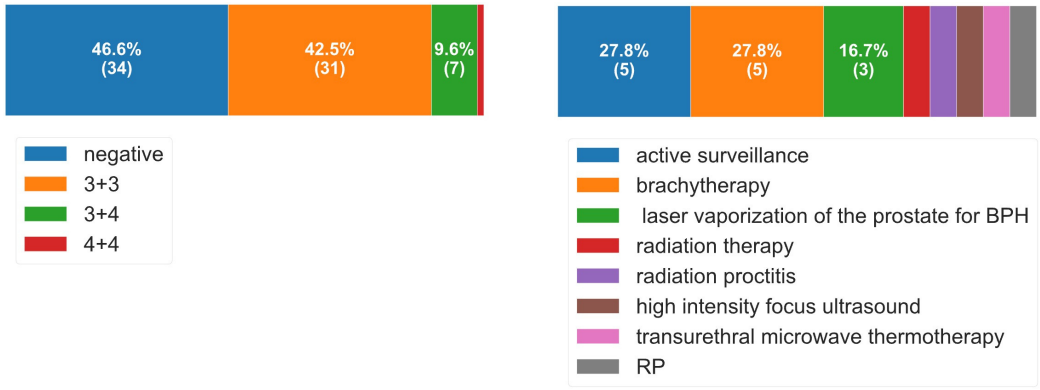


Figure 3.2: Left: a significant amount of the patient population underwent prior prostate biopsies (all Gleason scores documented in pre-procedural radiology reports), Right: cancer/non-cancer related treatment of the prostate. Note: absolute numbers given, relate to the number of lesions identified and biopsied (n=105), not to the number of studies (n=86) or patients (n=84) in the dataset.

3.2.2 Dataset comparison

In figure 7.2 in the attachments the BWH-MRIGPB dataset is compared to all major currently publicly available prostate cancer datasets, related to machine learning model training for the detection or classification of clinically-significant prostate cancer (see related work chapter). PROSTATEx has officially been superseded by the PI-CAI challenge as a benchmark for prostate AI development, also including cases from the PROSTATEx dataset. However, as the lesion locations (biopsy target coordinates), that are part of the original PROSTATEx dataset, are no longer available in the PI-CAI challenge and the original dataset is similar to the BWH-MRIGPB dataset, it is explicitly listed in 7.2.

As depicted, the BWH-MRIGPB dataset is with 86 curated MRI exams only about half the size of the prostate158 dataset, and counts <10% of the "Public Training and Development Dataset", published as part of the PI-CAI challenge with 1500 cases. Also, no direct expert-derived lesion delineations are available for the BWH-MRIGPB dataset (except constructed lesion bounding boxes, see next section for details). Lesion segmentations for prostate158 exist for 83 cases and were performed by two radiologists, in addition to zonal segmentation of the prostate. The PI-CAI public

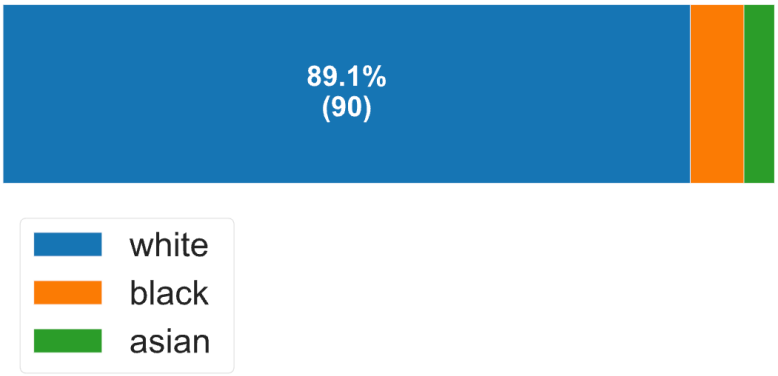


Figure 3.3: Around 90% of the patient population is identified as white, while black and asian patients make only about 10% of the sample size of the dataset. Note: the absolute numbers given relate to the number of lesions identified and biopsied (n=105), not to the number of studies (n=86) or patients (n=84) in the dataset.

training dataset counts in total 1,295 annotations, of which only 220 (17%) have actual clinically-significant lesion annotations, while the remaining 1075 cases carry empty lesion annotations. Lesion delineations were created using ITK-SNAP by "10 trained investigators or 1 radiologist, supervised by 3 expert radiologists" at RUMC or UMCG. Interestingly, medians for patient age, PSA levels and prostate volume are very similar between the PI-CAI public training dataset and the BWH-MRIGPB dataset, they also have a very similar relative fraction of clinically-significant vs benign or indolent cases (30%/70%). A slightly higher percentage of ISUP-scores of four and five are observable in the BWH-MRIGPB dataset (12% vs 15%). All datasets include axial T2-weighted, diffusion-weighted and ADC image modalities. The majority of MRI exams across all datasets were acquired with magnetic field strength of 3 Tesla, on either Siemens, GE or Philips MRI scanners, using surface coils. The BWH-MRIGPB dataset is the only dataset with the majority of scans acquired using an endorectal coil (64%). The PI-CAI, PROSTATEx and the BWH-MRIGPB dataset all provide clinical variables in addition to imaging data. All provided clinical variables / clinical metadata provided by the BWH-MRIGPB dataset are listed in 7.2 and can be reviewed in detail for every case in 7.8 in the attachments.

3.2.3 A few example illustrations

Figure 3.4 and 3.5 depict example cases of the final BWH-MRIGPB dataset, visualized via the Open Health Imaging Foundation (OHIF) webviewer. All imaging studies, together with point annotations (relating to pre-procedural biopsy target coordinates) and lesion bounding boxes (see next section for details) are available in DICOM format and stored in a DICOM data store on Google Cloud Platform. The DICOM data store can directly be visualized in the OHIF web-based viewer for an easy and fast visualization of the dataset without the need of specialized software to inspect the data.

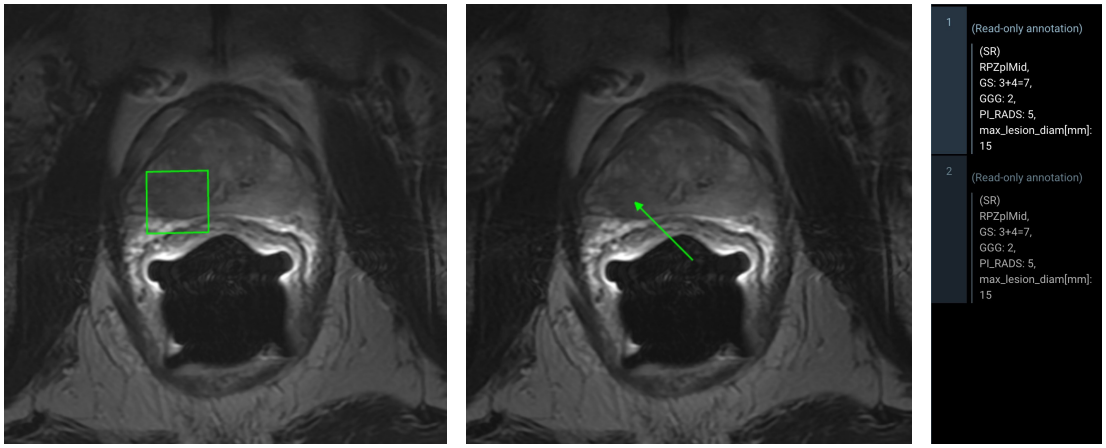


Figure 3.4: Axial T2-weighted scan, acquired with surface coil, with PI-RADS 5, ISUP 3, lesion located in the right peripheral zone posterior lateral mid (RPZplMid) region of the prostate. Left: image with pre-procedural set biopsy target coordinate, middle: image with constructed lesion bounding box, approximating lesion size, right: annotation description from DICOM Structed report, visualized in OHIF webviewer (see full example of interface in attachments).

For more details about the curation process and the used data types, see chapter 4.

3.3 Purpose - "Zonal" detection of clinically-significant prostate cancer

As described in chapter 1, detecting lesions on biparametric MRI with a high likelihood of harbouring clinically-significant prostate cancer is a very challenging task, even for experienced Radiologists. As can be seen in this very dataset, not all from Radiologists

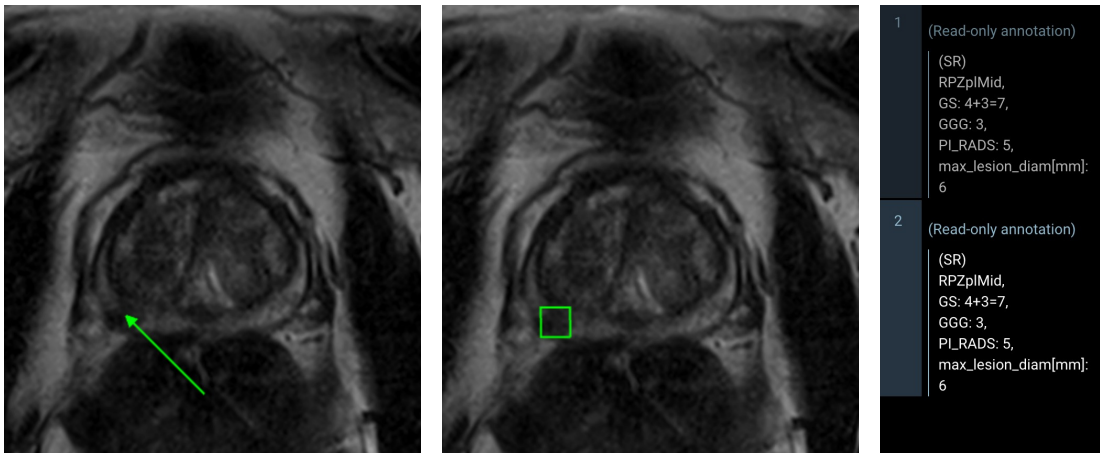


Figure 3.5: Axial T2-weighted scan, acquired with endorectal coil, with PI-RADS 5, ISUP 2, lesion located in the right peripheral zone posterior lateral mid (RPZplMid) region of the prostate. Left: image with constructed lesion bounding box, approximating lesion size, middle: image with pre-procedural set biopsy target coordinate, right: annotation description from DICOM Structed report, visualized in OHIF webviewer (see full example of interface in attachments).

estimated significant lesions (PI-RADS >=4) actually lead to an ISUP score >= 2 (see 7.8 for direct comparison of PI-RADS scores assigned by Radiologists and histopathology results, provided by Pathologists).

3.3.1 Requirements for BWH-MRIGPB dataset annotations

In 1 we mention two desired properties for our final dataset.

First, we want it to be "AI-ready". Hence, it should be curated with a specific "algorithmic task" in mind and provide the needed input and labels for convenient and simple model ingestion. In order to be able to apply the PI-CAI challenge baseline models, introduced in 2, we originally define our algorithmic task identical to the one presented within the challenge, namely "the detection of clinically-significant prostate cancer lesions on biparametric MRI". Biparametric MRI refers in this context to the required input of T2-weighted images and diffusion-weighted images, as well as apparent diffusion coefficient maps. "Detection" refers to simultaneous localization and

classification of prostate cancer lesions, as only "significant lesions" should be predicted.

Second, we want our dataset to be used for developing machine learning algorithms that are "implementable" in actual clinical workflows. We will for the frame of this thesis focus on the extensively analyzed and observed MRI-guided prostate biopsy procedure, conducted as standard of care at the Brigham and Women's Hospital and described in 1. Making even a well-trained and thoroughly benchmarked algorithm clinically implementable is a very challenging task and we will not consider the many hurdles involved in the actual downstream model development task. However, considering the nature of our raw data and the clinical workflow it was generated from, we may be able to derive requirements for our annotations that on one side can be phrased as a concrete enough algorithmic task, and at the same time be of concrete use during an MRI-guided biopsy procedure.

3.3.2 Constructing lesion bounding boxes for zonal detection of clinically-significant prostate cancer

To come up with annotations for our dataset that fulfill the requirements defined above, we will start by looking at the raw data provided.

First, we have pre-procedural targeted biopsy coordinates available for a significant amount of biopsy cases, stored in the manually maintained biopsy case archive (more information in 4). These biopsy targets can be used as point annotations on pre-procedural MRI and when matched with the correct histopathology result, can be used to indicate a ground truth (as Gleason score or Gleason Grade Group) for a single pixel coordinate on the axial T2-weighted MRI series.

Point annotations alone are, however, not enough for a **complete** evaluation of lesion detection maps, as for example produced by the baseline algorithms introduced as part of the PI-CAI challenge (see 2). While it is possible to evaluate the approximate location of predicted detection maps with point annotations (see 5.11), a complete evaluation requires in addition to a point annotation a measure of the "size" of the predicted lesion. Different approaches have been proposed in the literature to do "automatic lesion segmentation" [**pellicer2022**] and [**Liu2019**], mostly based on using point annotations as "seed", defining a certain threshold, and growing the segmentation based on intensity values and intensity gradients in the image. As intensity values in MRI are, however, not standardized this method might lead to significant error in estimating a correct lesion size and is not based on any "medical insights" about the lesion.

In addition to biopsy target coordinates, we have access to semi-structure radiology and pathology reports. Hence, circumventing the need for expensive expert-derive lesion delineations, we propose a method for finding approximations of clinically-significant prostate cancer, based on the automatic extraction of key diagnostics in pre-procedural radiology reports. As described in 1 these key diagnostics usually consist of an estimated lesion size, a PI-RADS score and the anatomical prostate region the lesion candidate is located in.

An intuitive procedure for creating so-called "lesion bounding boxes" out of these three diagnostic scores in combination with the mentioned pre-procedural biopsy target coordinates, would look as follows:

1. Use pre-procedural biopsy target coordinate as center, assuming the tissue extraction took place inside the estimated lesion.
2. Take half the maximal indicated lesion size in the pre-procedural radiology report, and use it to construct a 3D cube around the target coordinates.
3. Assign PI-RADS score and histopathology scores as constant value to entire cube-shaped lesion bounding box

Having created another "proposal" for lesion annotations - as lesion bounding boxes - for the BWH-MRIGPB dataset, we re-visit the defined requirements for our dataset:

- **AI readiness:** Now, in addition to having a specific location in the prostate, the constructed lesion bounding box also approximates the "size" of a lesion, therefore fulfilling the requirements for evaluating and potentially training the PI-CAI baseline models (see 5.11 for qualitative example). We therefore classify the annotation as "AI ready".
- **Clinically "implementable" and useful:** in addition to the "algorithmic perspective", we require the lesion bounding boxes as annotations of the BWH-MRIGPB dataset to be able to train and correctly benchmark algorithms, that are of clinical use. From a clinical perspective the question remains of whether "lesion bounding boxes" - instead of exact lesion delineations - are sufficiently accurate to be of clinical use in the MRI-guided biopsy procedure clinical workflow. Relating back to the extensive analysis of the workflow in 1, we **hypothesize** that in this specific case "zonal" indications of clinically-significant prostate cancer lesion candidates are sufficient to be of clinical value. Specifically, as in MRI-guided biopsies the interventional radiologist in the standard of care only receives the anatomical zone **description** as location of a lesion candidate, an exact lesion delineation / shape

might not even be necessary. Lesion bounding boxes could hence, in the workflow of MRI-guided prostate biopsies, be implemented as "zonal" ground truth of clinically-significant prostate cancer. Algorithms trained and benchmarked on the resulting dataset, would be able to predict prostate zones with high probability of harbouring clinically-significant prostate cancer without providing exact "shapes" / "segmentations" of lesions. As depicted in 5.11 the nnDetection framework (a baseline model from the PI-CAI challenge) in any case does not predict lesion augmentations, but bounding boxes, fitting our "ground truth" as lesion bounding boxes. Without further clinical validation, we assume, in the frame of this thesis, requirement two for our annotations also fulfilled (see further considerations about lesion bounding boxes in 5).

Next to the derived lesion bounding box annotations, we keep the biopsy target point annotations with matched histopathology as additional type of annotation for the BWH-MRIGPB dataset, due to their limited capacity of evaluating prostate cancer detection models 5.11 and their similarity to the point annotations, provided in the PROSTATEx dataset 2 (for interoperability reasons).

In 3.6, two examples of "IMPRESSION" sections of raw pre-procedural radiology reports are depicted. In 4, more details will be given on the automatic extraction process of the key pre-procedural diagnostics, that are used for the construction of the lesion bounding box annotations.

```
IMPRESSION
IMPRESSION:
1.    1.4 cm peripheral zone lesion in the left, apex,
posterolateral region and 1.2 cm lesion arising from the the
peripheral zone and extending to the transitional zone in the
right, mid, anterior region with MR characteristics consistent
with PI-RADS 4 lesions. There is no definite extraprostatic
involvement. No lymphadenopathy.
2.    Multiple BPH nodules in the transitional zone and one in
the central zone as detailed above.
```

```
IMPRESSION
IMPRESSION:

1.    1.9 cm right anterior peripheral and transitional zone
focal lesion mid to apex region, crossing the midline and with
likely extraprostatic extension along the anterior margin:
PI-RADS 5.
2.    Status post brachytherapy.
```

Figure 3.6: Two examples of raw impression sections from pre-procedural radiology reports, used to construct lesion bounding boxes in combination with biopsy target coordinates. Potentially clinically-significant lesion candidates are described as bullet points, including anatomical prostate zone description, observed lesion diameter and assigned PI-RADS score. All three entities are automatically extracted in the automated medical data curation approach, outline in chapter 4.

3.4 Usage notes and dataset availability

After expert validation, the BWH-MRIGPB dataset will be submitted to "The Cancer Imaging Archive" (TCIA). Alternatively, following the approach outline in the frame of the PI-CAI challenge, the imaging data will be made publicy-available via the Zenodo platform with lesions hosted on Github. Should a TCIA submission be successful, the dataset is planned to be ingested into the National Cancer Institute "Imaging Data Commons[1]" (IDC).

[1] `https://datacommons.cancer.gov/repository/imaging-data-commons`

4 An automated approach for medical data curation

In the following the automated curation process for the BWH-MRIGPBX dataset will be described in more detail. The motivation and reasoning behind each step and method used, as well as general considerations of why and how automated medical data curation can be used, will be the focus of the discussion in chapter 6.

4.1 Curation objective

In the following chapter we will assume **locally** available - already extracted and aggregated - data. They will be referred to as "raw data sources" and we will not go into detail on how the aggregation process for getting the data looked like. Instead the focus is on the automated data structuring and labeling process that was developed for creating the BWH-MRIGPB dataset described in chapter 3.

Figure 4.1 provides an overview of the complete medical dataset curation process. We distinguish between three main parts: extracting data, aggregating data and finally structuring and labelling data. The three main data sources for the curation of the BWH-MRIGPB dataset are:

- **Clinical reports:** all reports (pathology and radiology) need to be extracted from a hospital-internal server system, including the Electronic Health Records (EHR) of all patients. To reduce the total number of clinical reports as starting point for our curation task, a subset of reports can be queried from the internal IT infrastructure by providing Medical Record Numbers (MRN), Accession Numbers and the relevant institution (Brigham and Women's Hospital in this case). The original number of relevant MRNs and Accession Numbers was based on a **manually maintained** "enrollment"-sheet from MRI-guided biopsy cases at the Brigham and Women's Hospital over a significant amount of years.
- **Image studies:** image studies generated during clinical routines are stored in remote hospital servers and need to be first requested (similarly to clinical texts) and then "extracted" from a remote server. Extracting the image data is an

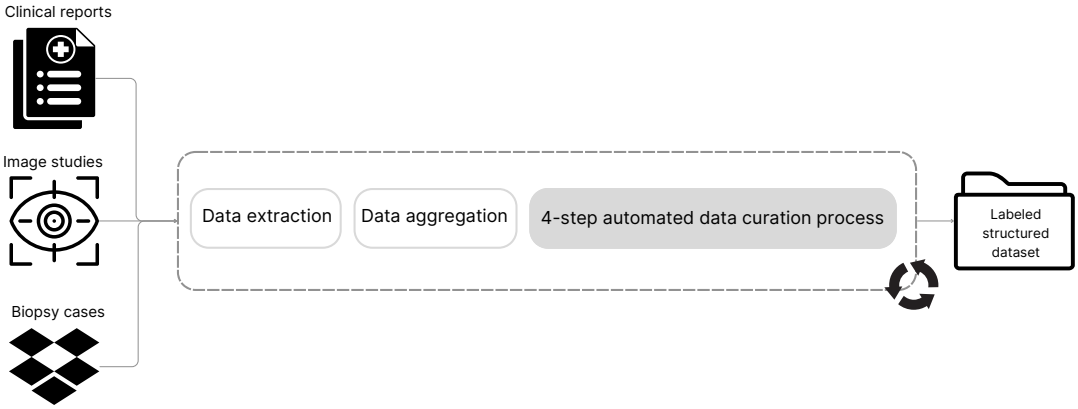


Figure 4.1: Complete "black-box" curation overview.

extremely tedious process, involving accessing the remote server via an unreliable API and transferring the de-identified imaging data to first a personal account on the internal hospital-wide computing network and finally either downloading the data to a local device or transferring it to another remote cloud location. In our case the imaging data - about 230GB - was transferred and downloaded onto a local SSD using "s5cmd" (local filesystem execution tool) and also transferred to a Google Cloud Platform bucket.

- **Biopsy cases:** a biopsy case archive was **manually maintained** inside a restricted-access Dropbox folder and needed to be replicated locally for further efficient processing. Similarly to the image extraction and aggregation process, downlaoding the biopsy cases consists of a very time-consuming in-efficient process and required a full week for downloading the biopsy case archive - about 416GB - to a local SSD.

4.2 Process overview: 4-step automated data curation process

4.2.1 Requirements for medical data curation

We define the following three main requirements / performance metrics for medical data curation:

- **Transparency:** all data transformations and operations that are applied to get from the raw data input, to the final dataset should be clearly traceable.

- **Efficiency:** data curation in general is a very time-consuming and tedious task. Any efficiency increases are therefore welcome.
- **Maintainability:** with every MRI-guided biopsy procedure new MRI imaging studies, clinical reports and biopsy targets are created. By ensuring an easy maintainability of the data curation pipeline, we hope to simplify the task of "continuously" growing the resulting BWH-MRIGPB dataset.

Further, we argue that by using a primarily "code-based" pipeline for data curation we are able to:

- Significantly reduce curation time
- Increase transparency of the entire curation process
- Allow for an "re-usability" to continuously grow the resulting dataset

Results and findings will further be discussed in chapter 6.

4.2.2 Conceptional overview

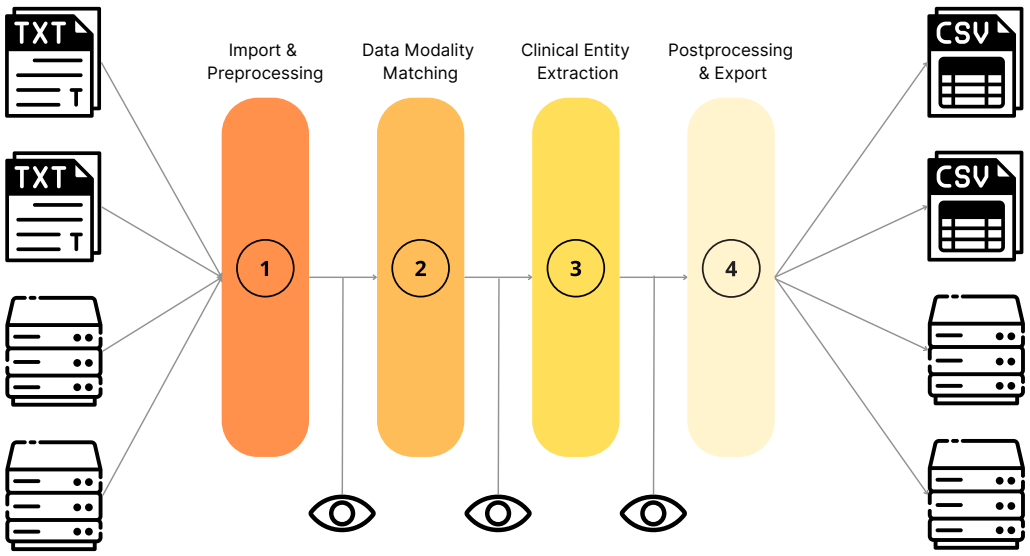


Figure 4.2: 4-step automated data curation process

The curation process can be subdivided into four main sections (see figure 4.2): import and pre-processing of the raw input data, matching of all available data modalities

using specific unique identifiers, extracting the desired relevant clinical entities and finally post-processing and exporting the final dataset in the desired data format.

For this cohort four data modalities were part of the raw data input:

- Radiology reports
- Pathology reports
- Prostate biopsy case archive
- Imaging studies

It is assumed for this chapter, that all clinical reports (pathology and radiology) are available in two separate ".txt" files, and both the requested imaging studies and the biopsy case archive are available on local SSDs. The imaging studies are additionally accompanied by two ".csv" files, one specifying a complete table of de-identified medical record numbers, study numbers and dates, as well as series descriptions of the MRI sequences, and the second file specifying the mapping between anonymized identifiers and original ones.

Each curation step adds specific data transformations necessary for creating the final dataset. In fig. 7.5 a simplified data flow is visualized, describing the number of data points per modality and per curation step. As observable, all modalities start at a rather high number of data points, that are significantly reduced during pre-processing, matching, extraction and post-processing, to finally result in the curated dataset.

Although the curation process is referred to as being "automated", meaning the actual data curation is conducted by lines of code, instead of manual work, human validation is still crucial. It is important to note here, that automating data curation does not replace visual inspection and exploration of the data. On the contrary, as depicted in fig. 4.2 with the "eye" symbol, a crucial advantage of partitioning the data curation into separate automated modules, is that every step can be retraced and debugged. In contrast to a manually curated dataset, the credibility and correctness is openly verifiable, as long as the code is openly accessible (see section on "Usage notes and code availability").

It is assumed in this regard (and also estimated, see discussion in chapter 6) that manual inspection and validation of the automatically generated dataset takes significantly less time than conducting the entire curation manually. Next to the time efficiency perspective, other reasons (e.g. maintainability) justifying the use of automated data curation over manual data curation are discussed in chapter 5.

The following sections will deep dive into each of the four main curation steps, methods used for data transformation and their respective performance. For a complete overview of the most important curation steps used, the methods applied at each step and additional information about their location in the code, see attachments 7.6, 7.7. Three "task categories" are distinguished, with "manual inspection & correction" referring to the validation steps described above, and "syntactic" and "semantic" referring to the two main categories the used methods can be classified into. Both categories are related to ways of processing natural language with computer algorithms, "syntactic" referring to the set of used methods that focus purely on the syntax of the analyzed text (e.g. regular expressions), and "semantic" referring to methods that try to infer a specific "meaning" of the provided text.

4.3 Step 1: raw data import & pre-processing

Modality No.1: radiology reports

Both pathology and radiology reports come in a "semi-structured" format - consisting of unstructured natural text sections divided by standardized section headers or delimiters. As highlighted in figure 8.1 by the red boxes, the main sections of interest in the radiology reports are the general patient information (usually marked as "Additional Information:" or as "FINDINGS") noted down during the visit (e.g. PSA score, patient age, volume of prostate gland, etc.), representing important clinical metadata, and the "IMPRESSION" section, specifying all "important" observations done by the radiologist and noted down as brief bullet points in unstructured natural text. In our case we define every bullet point in the impression section as a "Single Diagnosis", ideally consisting of an assigned PI-RADS score (see chapter 1), an estimated lesion diameter in millimeter or centimeter, and a description of the exact anatomical zone of the prostate the lesion was observed in. During the "lesion-level" matching our goal will be to match each single diagnosis in an impression section with the diagnostic scores from the other modalities, to get a complete picture of the evaluation of a specific lesion, for a specific patient (more in following section). Radiology reports are available for a time frame from June 2003 to July 2022 with a total number of 10788 reports (see fig. 7.5).

Modality No.2: pathology reports

The main region of interest in pathology reports is the pathologic diagnosis section, comprising usually of a list of all biopsy locations and their diagnostic scores with varying length (depending on number of biopsies taken during procedure). Usually this section is clearly delimited by two section headers "PATHOLOGIC DIAGNOSIS"

and "CLINICAL DATA". Each section usually starts with an alphabetically ordered letter and the anatomical region of the prostate the biopsy was taken from, followed by a brief diagnosis section, describing whether the extracted tissue was benign or showed at least partly signs of cancerous cells. As seen on fig. 7.1, the pathologic diagnosis section is more standardized and structured than the radiologic impression section, in the cancerous case usually comprising of the diagnosis itself (in the majority of cases "PROSTATIC ADENOCARCENOMA"), the Gleason Score, Gleason Grade Group, the number of cores taken and how much cancerous tissue involvement was observed. There still however is a significant variance in documenting the results, complicating the diagnosis extraction and lesion-level matching (see following sections). Pathology reports are available for a time frame from November 2001 to February 2022 with a total number of 3286 reports (see fig. 7.5).

Modality No.3: biopsy target coordinates

The biopsy case archive consists of about 875 biopsy cases (exact number difficult to estimate due to inconsistent numbering), that all relate to prostate biopsies that were conducted at the Brigham and Women's Hospital in an approximate time frame from June 2010 to October 2020. In contrast to the other data modalities (clinical reports and imaging studies), this archive is not part of the general electronic health records or the hospital IT infrastructure, but all case folders were manually added and uploaded (see details about data transfer in previous chapter). The case folders vary substantially in their internal sub-folder structure, naming conventions and file content making rule-based extraction approaches difficult. The main time period of interest can be primarily limited to the years from 2016-2017 with a few outliers to 2015 and 2020, where a specific 3D slicer module was used (mpReview) for creating preoperative biopsy targets. Preoperative biopsy targets correspond to image studies that were conducted before the actual biopsy procedure took place. Hence, they relate to preoperative MRI scans in contrast to intraprocedural biopsy targets that corresond to MRI scans that were performed during the biopsy procedure. Intraprocedural biopsy targets need to be discarded for the dataset creation (see justification in description of next data modality). All biopsy target coordinates are stored in ".fcsv" files in varying locations of the case directories but are themselves documented in a structured way (see more details in matching section).

Modality No.4: MRI scans

All requested imaging studies (see chapter 3 for details on requesting, extracting and transferring the data) come in a de-identified folder structure, grouped by anonymized

medical record number, anonymized study number and date, and the original series descriptions of various MRI sequences. The MRI volumes themselves are stored per slice in DICOM format. The number of MRI sequences per study varies substantially (ranging from around five sequences to up to 200), as well as the series descriptions themselves. Series descriptions have no clear naming convention and often are non-trivial to assign to a specific used MRI sequence type without inspecting the volumes visually. The following pie chart gives a good impression of the high variability in series descriptions:

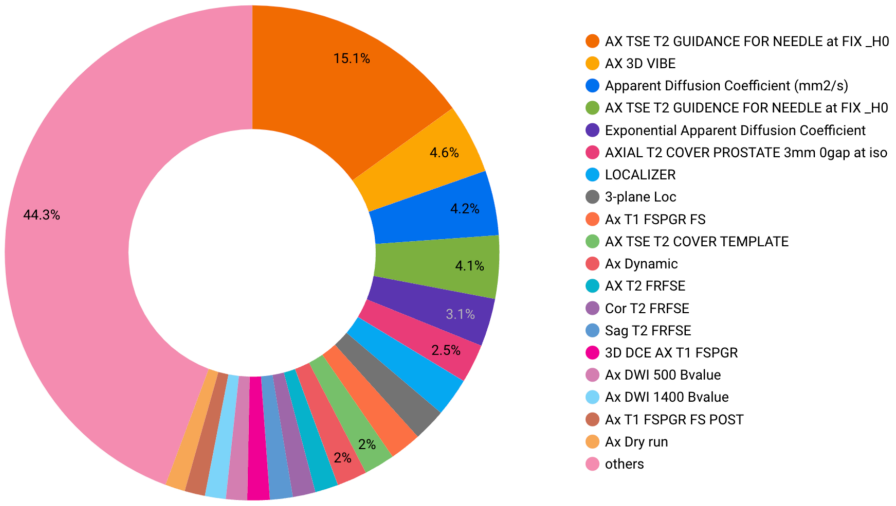


Figure 4.3: Variability of series descriptions for imaging studies

The high variability of series descriptions significantly complicates the selection of the relevant MRI sequences (T2, HBV, ADC), that are relevant for the final dataset.

Key characteristics

Both text files containing the clinical reports are first split into individual reports, loaded into pandas dataframes, all prostate related reports are selected and crucial matching keywords are extracted from standardized headers, including medical record number (MRN), accession number and accession date (see methods marksheet in attachments). The individual single diagnosis sections containing the lesion-level anatomical prostate region and diagnostic scores are extracted by further filtering and splitting the raw text input and storing each "diagnosis candidate" in a new separate dataframe row.

A similar procedure is implemented for the imaging studies and biopsy targets. After

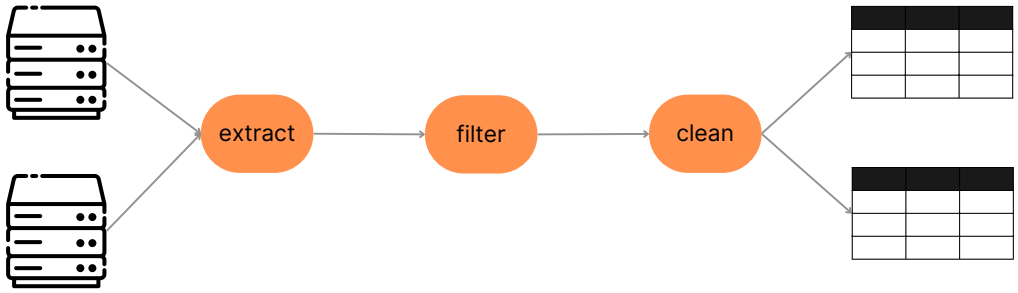


Figure 4.4: Clinical report preprocessing steps

the initial request and transfer of all imaging studies (chapter 1) to a local SSD, as well as to Google Cloud Platform (GCP), a table of all distinct values for the columns "SeriesDescription", "StudyDate", "PatientID" and "AccessionNumber", grouped by the DICOM attribute "SeriesInstanceUID" is queried using Bigquery and exported as csv file. The csv file is loaded as a datframe and all rows are mapped form their anonymized values to the original values (using an additional conversion table). The translated csv file is also laoded as pandas dataframe and optionally filtering and cleaning operations can be implemtend. A similar dataframe is created for all biopsy target candidates, by recursively crawling through the biopsy case archive folder structure, selecting all ".fcsv" files, filtering for all preoperative targets by looking at the filenaming and extracting important keywords for the following matching step, including medical record number, procedure date and case number.

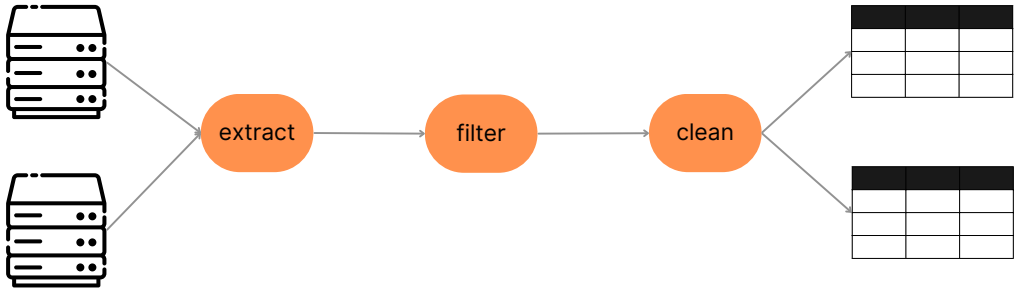


Figure 4.5: Images and targets preprocessing steps

As output of the import and pre-processing step of the curation pipeline, four primary dataframes are created with candidate biopsy targets, candidate diagnosis sections and candidate image seriesdescriptions. It should be noted, that a "permissive" strategy

is applied for this first step of the pipeline, referring to the goal of employing very basic syntax based natural language processing techniques (e.g. regular expressions) for saving computational cost on one side and allowing as many candidates from the different data modalities as possible for the following matching. The hopes are to increase the number of matches and hence the size of the final dataset. As a consequene of this strategy, the error rate (the number of wrongly extracted candidates) is relatively high, which is tolerable, as the majority of nonsense candidates will organically be discarded during the upcoming steps of the curation pipeline.

4.4 Step 2: matching data modalities

4.4.1 Step 2a: patient-level matching

The matching of different data elements across multiple modalities represents the key challenge in the curation process. Because of the complexity of the task we propose a hierarchical matching procedure in two steps where we first match all information for a specific patient (patient-level matching), and then for a specific lesion located in the prostate gland (lesion-level matching) of the respective specific patient.

First - in the patient-level matching - we attempt to create a subset of all patients represented in the imported and processed raw data input, for whom at least the following information is available in **all** of the four data modalities:

- Medical Record Number (MRN)
- Study date (corresponding to either the procedure date of the biopsy or the accession date of the report creation)

As every selected patient can have multiple studies, and within these studies multiple lesion locations in the prostate that were either noted down in the impression section of a radiology report and-or appear in a diagnosis section of a pathology report with or without a corresponding biopsy target coordinate (depending on whether a systematic or targeted biopsy was conducted, see chapter 1), an additional "lower-level" matching step is necessary for creating the final labels for every imaging study.

In this next step, the lesion-level matching, all selected patients and respective studies are further decomposed, by taking the already extracted (see pre-processing) single diagnosis candidates from the clinical reports, as well as the biopsy target candidates into consideration.

Similar to the pre-processing step of the curation pipeline, the reasoning behind this two-step matching process is, to first employ easy-to-implement methods (based on the little structured data available across all modalities) to reduce the number of data points as much as possible with little effort, discarding for example patients that do not have all the above information available, to simplify and increase the efficiency of the much more complex and computationally expensive lesion-level matching.

The patient-level matching process itself consists of three main steps, as depicted on figure 4.6.

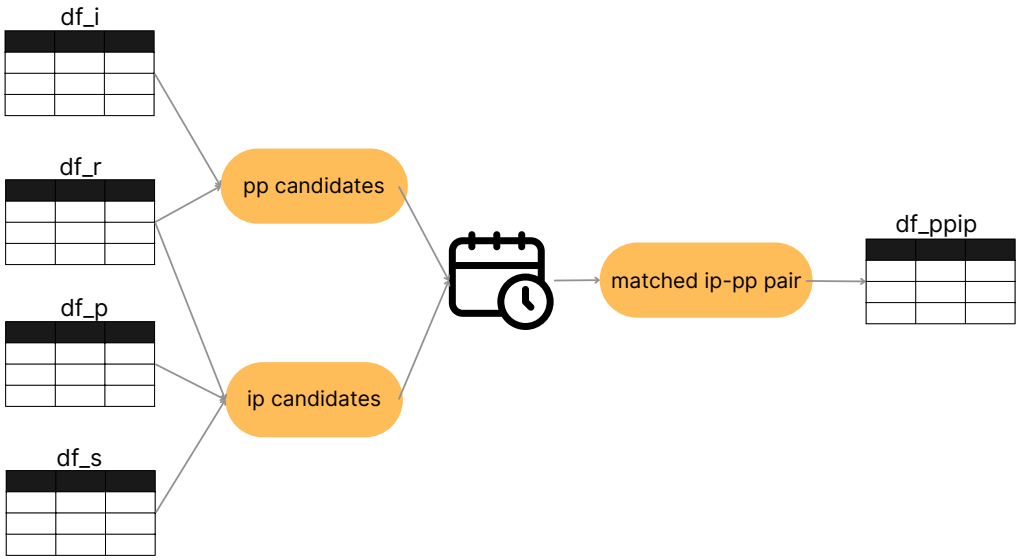


Figure 4.6: 3-step patient-level matching

First all preprocedural candidates ("pp candidates) are created by merging (inner join) the processed radiology dataframe, with the processed image dataframe on the extracted MRN and study date. The MRN and study date have proven to be the most reliable combination of keys, mainly as they can be automatically converted into non-string datatypes ("int64" for MRN, "datetime64[ns]" for study date) and are therefore less susceptible to human input errors. A significant number of accession numbers for example (not generably convertable to int datatype, because of a starting letter, e.g. "E123456" and hence stored as string datatype) include a varying number of spaces (assumed to be created during manual input of the number in the report header), causing unexpected merge issues between dataframes although the keys seem to be identical. Similar phenomena are observable with accession numbers from imaging

studies after de-identification, making the accession number in general unreliable for joining dataframes. The resulting preprocedural candidates refer to all studies for which an image study could be queried from GCP, and a prostate related diagnostic radiology report was available, containing diagnostic scores (lesion size and PI-RADS) given for a specific subset of lesion for the respective study. All reports that do not have a matching are automatically discarded after merging the dataframes.

The intraprocedural candidates ("ip candidates") are created by merging the processed radiology dataframe, with the processed pathology dataframe and target candidates dataframe. Identical to the preprocedural candidates an inner join is performed to merge the dataframes, based on MRN and study date. The study date in the pathology report corresponds to the procedure date explicitly mentioned in the report and not necessarily to the accession date of the report. The reason for a mismatch between the date of the biopsy procedure and the creation date of the report is simply because the analysis of the extracted tissue by a Pathologist is not necessarily done on the date of the procedure itself but can be delayed by several days, depending on various factors in the clinical workflow. Hence, the study date is extracted as the procedure date from within the report (slightly complicating the extraction process, due to human errors in the documentation process, e.g. missing procedure dates). All reports that do not have a matching are automatically discarded after merging the dataframes. For the intraprocedural candidates, the mentioned three dataframes are relevant, as for the construction of a valid label ("point annotations", as defined in section 3), at least one set of biopsy target coordinates, the corresponding histopathology and a set of triplaner **preproceduraldiagnostic** MRI are needed.

As step two of the patient-level matching, the pp and ip candidates are grouped into pairs of corresponding ip-pp studies. The goal is to loop through all ip studies, and by comparing the study dates find the most recent preprocuderal study to every intraprocedural study (depicted by the calender symbol in figure 4.6). This "temporal matching" relies on the completeness of the radiology reports, as common dataframe between both the pp and ip candidates (see 4.6).

It is important to note here, that the temporal matching approach is strongly dependant on the correct assignment of the "type" of the prostate radiology reports into ip and pp. For example, in case of a "repeat biopsy" - being the case if a prostate MRI scan showed one or more suspicious lesions, a prostate biopsy was conducted, but the tissue analysis was inconclusive (e.g. Gleason Score 3+3=6), and the patient was set on active surveillance, another prostate biopsy procedure might directly follow the first one, without another preprocedural diagnostic scan. In this case in order to not

match two ips, by just looking at the study date, the report header is used to estimate the reports type. For the edge case of a repeat biopsy the majority of ip reports can be detected by looking for the regex "(?i)biopsy" and hence can be clearly assigned to the ip candidates, instead of the pp candidates. In the case of a repeat biopsy the most recent pp report candidate will be matched to the last (hopefully conclusive) prosate biopsy procedure.

A second edge case that can affect a correct assignment of ip-pp matches, are duplicate reports. For a small subset of radiology reports, the exact same report occurs multiple time with different header names and sometimes even varying study dates (reason unknown, but potentially linked to billing). In the majority of cases these edge cases can again be discarded by using the header information for classifying the reports into ip and pp, then sorting all ip-pp candidates by study date, and finally dropping all duplicate rows in the final created dataframe df_pp (only keeping first occurrence).

The resulting (df_ppip), consisting of all matched ip-pp studies for a certain MRN, counts 164 cases in the baseline version of the BWH-MRIGPBX dataset.

It is important to note at this point, that even after matching there still is no guarantee that the matched "pp candidates" and "ip candidates" will actually enable the construction of valid point or bounding box annotations. The final number of cases for which we will be able to construct at least one valid label type is now depending on the actual content of the extracted single diagnosis sections from pathology and radiology reports, as well as the content of the ".fcsv" files in which the biopsy target coordinates are stored. The content of the matched data modalities will be processed and analyzed in the lesion-level matching of the following subsection.

4.4.2 Step 2b: lesion-level matching

The lesion-level matching represents the pinnacle of the entire data curation process and includes following main tasks:

- Extract T2W, HBV and ADC MRI sequences from the preprocedural image study of every ip-pp match (discard ip-pp match if not all three image modalities available)
- Match the lesion-level single-diagnosis sections of intraprocedural pathology and preprocedural radiolgoy reports with the preprocedural biopsy target coordinates

The extraction of the correct MRI sequences for every preprocedural image study, could concept-wise also be assigned to the patient-level matching however, as a very similar

"semantic matching" approach is used for the extraction, the approach is presented in this subsection.

A "semantic" approach to clinical keyword extraction & matching

While the import, pre-processing and patient-level matching was feasible (with some errors) with a purely "syntactic" natural language processing approach, the lesion-level matching requires a more sophisticated processing, due to two main reasons.

First, because of a lack of usable standardized (unique) identifiers. While being able to rely to an acceptable extent on standardized clinical report headers, section headers and unique identifiers (or a combination of non-unique identifiers, e.g. MRN and study date) used across all input modalities, the same is no longer the case for the lesion-level matching. For finding the correct pathological diagnosis, radiological assessment and 3D coordinates of a biopsy target, no specific clinical keyword or identifier is available. Hence, we have to rely on the description of the anatomical prostate zone the lesion is located in, that is mentioned across all input modalities for the majority of cases. Similar "content-based" approaches need to be taken for extracting the needed MRI sequences from all image studies (by looking at the series description), and extracting some of the preprocedural and intraprocedural diagnostics described in the following subsections. Important to note here is, that none of these "content-based identifiers" (anatomical zone, series description etc.) comes in a standardized structure or syntax, making the approaches used in the patient-level matching very time-consuming to implement and also error prone. More details on the respective challenges faced with every matching or extraction task are described in the following subsections.

Secondly, the tolerated error in the lesion-level matching compared to the patient-level matching should be significantly lower, to ensure the quality of the final dataset and minimize necessary manual corrections.

Natural text embeddings, i.e. finding a numerical representation of tokens, words or sequences of words in form of vectors, enables us to focus more on the "semantic meaning" of natural text, rather then focusing purely on structure and syntax. In this use case, the hope would be to find meaningful vector representations of anatomical prostate regions, image series descriptions and medical diagnostic scores, so that we can extract these keywords from their unstructured original text, based on (limited) semantic understanding, instead of looking at significantly varying text structure. In the frame of this work two main approaches to medical semantic keyword extraction are explored and applied to our use case of automated medical data curation.

In the first approach to semantic keyword extraction (figure 4.7), first a number n of candidate text snippets (e.g. the single diagnosis sections from the clinical reports containing the anatomical prostate region) and a number m of natural text targets is projected into the same vector space. As second step, the cosine similarity as standard metric to measure the similarity of two vectors is computed for all n candidates for all m targets. Finally, all m targets are stored as keys in a dictionary, with the "candidate" text snippet with the highest similarity score as values. In the following subsections and figures this approach is referred to as "es_keyword_extractor()" (corresponding to the naming of the implemented function in the python code).

The second approach to semantic keyword extraction explored in the frame of

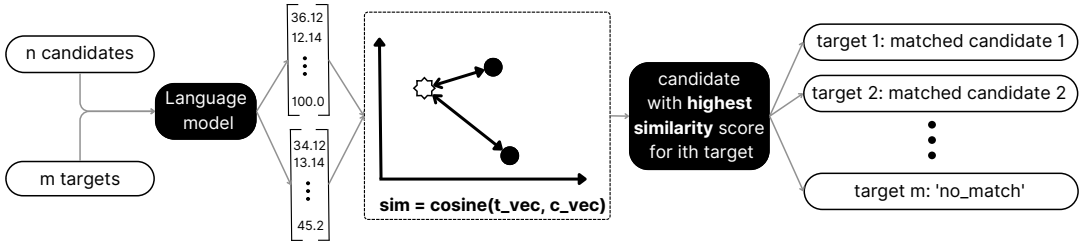


Figure 4.7: Using cosine similarity of embedded natural text vectors to match a list of candidate strings with one or multiple target strings. A dictionary with targets as keys and a list of the best candidate with the respective score is returned.

this thesis (figure 4.8) also relies on vectorizing a number n of candidate text snippets (referred to as "hypotheses" in figure 4.8) and comparing them to m vectorized natural text targets (referred to as "premises" in figure 4.8). In addition to this vectorization step, we compute the probability of a natural text candidate being "entailed" in a vectorized target, in accordance with the Natural Language Inference (NLI) terminology. Computing the entailement probability substitutes the computation of a similarity score and can be used in an identical way to match each target with the "best" candidate and store them in a dictionary. In the following subsections and figures this approach is referred to as "nli_keyword_extractor()" (corresponding to the naming of the implemented function in the python code). A main advantage of both approaches is that they are "generalizable" and applicable to multiple medical keyword extraction matching tasks. As visualized in the following subsections, we are able to use both keyword extractors for five different matching and extraction tasks (anatomical region matching in pathol-

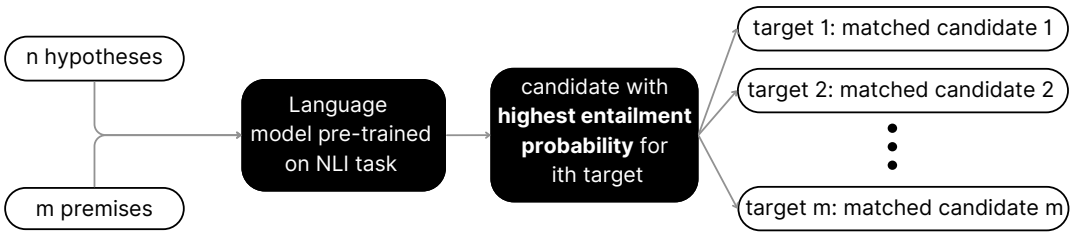


Figure 4.8: A list of hypotheses (candidates) and one or multiple premises (targets) are inputted into a language model pre-trained on a Natural Language Inference task. A dictionary with premises as keys and the respective hypothesis with the highest entailment probability is returned.

ogy and radiology reports, series description extraction, pi-rads score extraction and lesion size extraction), by simple function calls without modifying a single line of code, or need for additional fine-tuning. For the "es_keyword_extractor()" we use:

- BERT base model (case-sensitive) [**Devlin2019BERTPO**]
- 109 million parameters
- Twice fine-tuned by first [**btz682**] ("BioBERT"), and [**alsentzer-etal-2019-publicly**] on multiple medical text datasets, including the MIMIC-III v1.4 database [**Johnson2022**]
- pre-trained model id on huggingface: 'emilyalsentzer/Bio_ClinicalBERT'
- after multiple runs a similarity score = 0.7 has proven best as a threshold for keeping all correct matches and discarding as many wrong matches as possible

For the "nli_keyword_extractor()" we use:

- BART large model [**DBLP:journals/corr/abs-1910-13461**]
- 406 million parameters
- Trained on "Multi-Genre Natural Language Inference (MultiNLI) corpus" [**N18-1101**], with two layer classification head with 1 million additional parameters (total size 407m)
- specifically trained for for zero-shot text classification (0SHOT-TC)
- pre-trained model id on huggingface: 'facebook/bart-large-mnli'

- after multiple runs an entailment probability threshold = 0.5 has proven best for keeping all correct matches and discarding wrong ones (less important for NLI model, as in general substantial score differences between correct and wrong candidates, see chapter 6)

Due to the time constraints of this thesis only a very limited benchmarking and performance evaluation of all used approaches for medical keyword extraction and matching could be performed. More details about the performance, reasoning behind the chosen approaches and models, and limitations are discussed in chapter 6. A more in depth argumentation for "automated medical data curation" in general, and why ensuring an easy maintainability and updatability of the data curation pipeline and dataset might be useful, especially in the medical domain, is also discussed.

Extracting T2W, HBV, DWI from series descriptions

For the extraction of the three required image modalities, the only point of reference available is the series description. However, as visualized in figure 4.2, series descriptions are not standardized and vary significantly in their naming conventions and structure. This makes even evolved "syntactic" approaches like regular expressions very hard to use, as first, a high number of series descriptions (originally 31671 image series) would need to be manually sifted through to then try to derive expressions that include all the possible variations of the respective image modality, and secondly the approach would not necessarily generalize very well to other imaging cohorts with potentially again substantially different naming conventions, resulting in a poor return on investment in terms of time invested and usefulness of the method. Another intuitive option to extract the correct series description would be to just "do it manually", by using only the already matched ip-pp studies (a few hundred at most). This approach would ensure decent matching accuracy (series descriptions are in general despite a varying syntax still clearly identifiable by a human with only little domain knowledge needed). The main downside to manual extraction is - similar to automated syntactic approaches - related to a complex and time consuming maintainability of the data curation pipeline and final dataset. For every alteration or update of any previous step of the data curation pipeline up to the lesion-level matching, the manual image modality extraction would need to be repeated from scratch, to verify which imaging studies still exist, which were deleted and what new ones were added.

As depicted on figure 4.9, by using the es_keyword_extractor() function, we are able to retrieve the correct series description for every modality, by providing the target strings:

- 'AX_T2'
- 'AX_DWI_1400'
- 'Apparent_D_nt_mm2_s_'

The target strings, together with all current series descriptions per image study are provided as lists of strings to the keyword extractor function, that vectorizes every element, finds the semantically closest candidate series description for every modality and returns a dictionary with the matched image modalities. All matched series descriptions are stored in a new dataframe "df_images", together with the absolute path to the actual image study on the external SSD drive. For an overview of how many

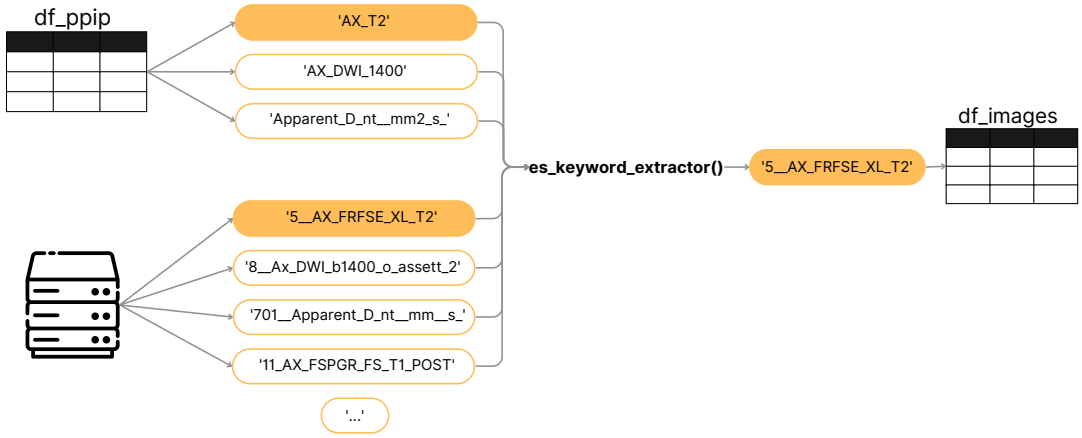


Figure 4.9: semantic extraction of T2W, HBV, ADC image modalities

correct matches were produced by the used model and method, see 5.

Matching diagnosis sections with biopsy coordinates via anatomical zones

For constructing a useful point annotation, consisting of a biopsy coordinate and a matching pathologic diagnosis (e.g Gleason Score), it is necessary to assemble the correct biopsy coordinates with the right single diagnosis section of the pathology dataframe. For constructing a (potentially) useful bounding box annotation, in addition, the lesion diameter and PI-RADS score need to be matched from the radiology dataframe. As all three modalities contain a description of the anatomical region of the prostate either the biopsy was taken from, the pathologic diagnosis is available for, or a suspected lesion was detected during preprocedural diagnosis, it will be used as unique identifier to "sort" all lesion candidates for a specific patient. As in general there

are about seven diagnosis sections in a pathology report, usually at least two suspicious lesions identified in the prostate before a biopsy and coordinates for multiple biopsy targets, the amount of possible combinations amounts to about 30 (which can increase to more than a 100 in some cases, where for example 20 biopsy probes are listed in a pathology report). Conducting a matching manually results in an extremely tedious and error prone task of looking up MRN and study dates in both pathology and radiology txt files, reading through the impression and pathologic diagnosis sections, looking up the biopsy case in the biopsy case archive (sifting through the folder structure to find the preprocedural biopsy target coordinates), and manually writing out the matching biopsy coordinates, preprocedural and intraprocedural diagnostics into an excel sheet. The "syntactic" automation approach is similarly inefficient and error prone due to the high number of edge cases that need to be taken into consideration while processing the respective diagnosis sections (an example of an original code snippet of around 1000 code lines, written for only processing of anatomical zones and gleason score extraction in pathology reports, is visualized in the attachments). In our semantic matching

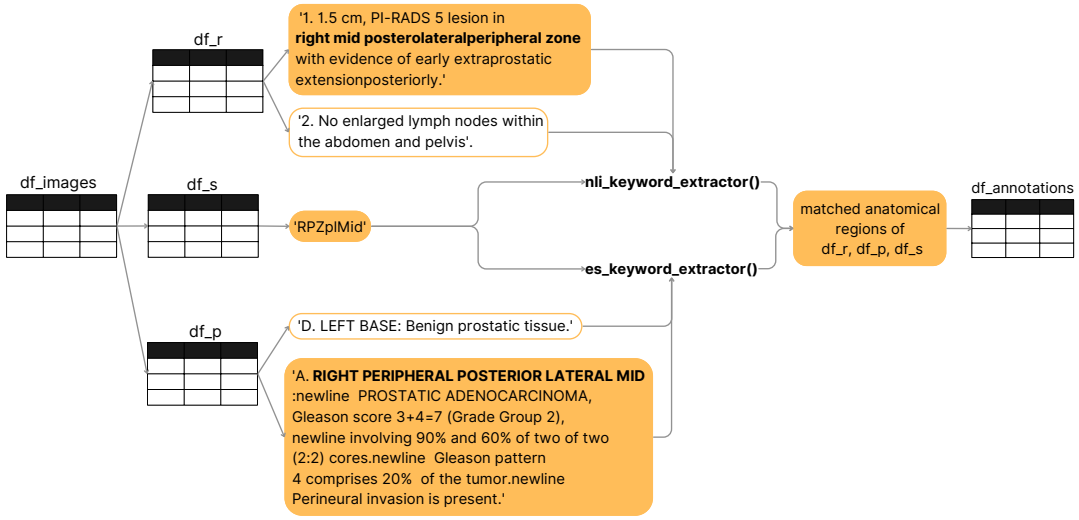


Figure 4.10: semantic anatomical matching of pathology & radiology diagnosis sections with biopsy coordinates, in this case for the "Right peripheral posterior lateral mid" prostate zone

approach the target biopsy label (simply reading in fcsv file of preoperative target candidates), representing an abbreviated form of the anatomical region it was taken from (in figure 4.9 "RPZplMid") is first converted into a full natural text description by applying a static mapping ("RPZplMid" -> "right peripheral zone posterior lateral

mid"). The anatomical region description is then used as target string for both semantic keyword extractors. As anatomical region candidates all available single diagnosis sections from pathology and radiology reports for the specific study and patient are provided as lists of strings (as visualized in figure 4.9, the diagnosis sections vary substantially in terms of content and structure). The single diagnosis sections with the highest scores (cosine similarity or entailment probability) are stored together with the corresponding biopsy target coordinates in a new dataframe df_annotations.

The "nli_keyowrd_extractor()" performed significantly better for classifying and selecting the correct radiology single diagnosis sections, than the "es_keyword_extractor()". As the "es_keyword_extractor()" is less computationally expensive with a significantly smaller parameter size (approximately 4-times smaller, see above section) and performs sufficiently good on the pathology single diagnosis section, both models were used. No extensive quantification or clear interpretation of why these model performance differences exist, can be given at this point, due to time constraints. See chapter 6 for a minimal discussion and an overview of how many correct matches were produced by the used model and method.

The lesion-level matching conclused the two-step matching process, as second step of the automated curation pipeline. As output we get the df_annotations dataframe, including all studies for which all three image modalities could be matched, and the matching single diagnosis section from pathology and radiology reports, together with the biopsy target coordinates for the respective study. A significant amount of targets cannot be matched via their anatomical zone, because of incorrect labeling of the biopsy target coordinate by the human operator (e.g. instead of "LPZaMid", "F-1" was inputted, making it impossible to find the correct pathologic diagnosis and radiology assessment) and hence are discarded. Before manually inspecting the correctness of "df_annotations", all preprocedural and intraprocedural diagnostic scores are extracted from the matched single diagnosis sections (see next section).

4.5 Step 3: extracting diagnostic characteristics

Four main diagnostic measurements can be extracted from pathology and radiology reports:

- Gleason score
- Gleason Grade Group
- PI-RADS score

- lesion diameter (mm/cm)

A combination of syntactic and semantic extraction approaches are used, and described in the following subsections.

4.5.1 Extracting pre-procedural diagnostics

As preprocedural diagnostics the PI-RADS score and the estimated lesion diameter of the respective lesion are extracted from the matched single diagnosis section of the radiology impression section. Due to the high variability in syntax and position of both, semantic extraction yields the best performance with the least effort. An example for two single diagnosis sections of an impression section are visualized in figure 4.9. Figure 4.10 depicts the extraction approach, using the nli_keyword_extractor() function. First a list of all existing PI-RADS scores, provided as strings including the naming (e.g. "PI-RADS 2"), is provided, followed by a list of possible lesion diameters. While the hypothesis space for the PI-RADS score is small (only five scores possible), the lesion size can have a wide range of sizes, sometimes documented in millimeters, sometimes in centimeters by the respective radiologist. Going through all diagnosis sections to define a set range would defeat the purpose of the entire approach, and stepping through a range from 1mm to 20mm would lead to a highly inefficient and computationally expensive operation. A simple solution for getting all "relevant" lesion size candidates for a respective diagnosis section, is to use a simple regular expression to extract all numbers in the provided diagnosis and construct a lesion size candidate by attaching both a "mm" and a "cm" string (see figure 4.11). Using this approach the matching lesion size is guaranteed to be amongst the candidates with a reasonable amount of overall candidates. It shall be noted however - further discussed as a general downside to 0SHOT-TC with models based on NLI in chapter 6 - that with this approach candidates are introduced, that are not at all present in the diagnosis (e.g. '2020mm', if the year of the diagnosis is mentioned). By turning the extraction task into a simpler classification task we introduce additional noise, that can lead to miss-classifications. It turns out however, that the number of wrong extractions is reasonably low for this use case (see chapter 6 for model performance for this task). Both identified "labels" for the single diagnosis section are added to the df_annotations dataframe, as additional rows. Should no match be found for either of the diagnostic scores (based on pre-defined entailement probability threshold), a string "no_match" is stored in the final dataframe. No matches can occur mainly because of two reasons. First, because of an error in the initial partitioning of the impression section of the radiology report, leading to a nonsense single diagnosis section. Second, because of an incomplete documentation of the radiologist (for example referring a previous study for diagnostic score if unchanged, instead of explicitly writing them down again). Both

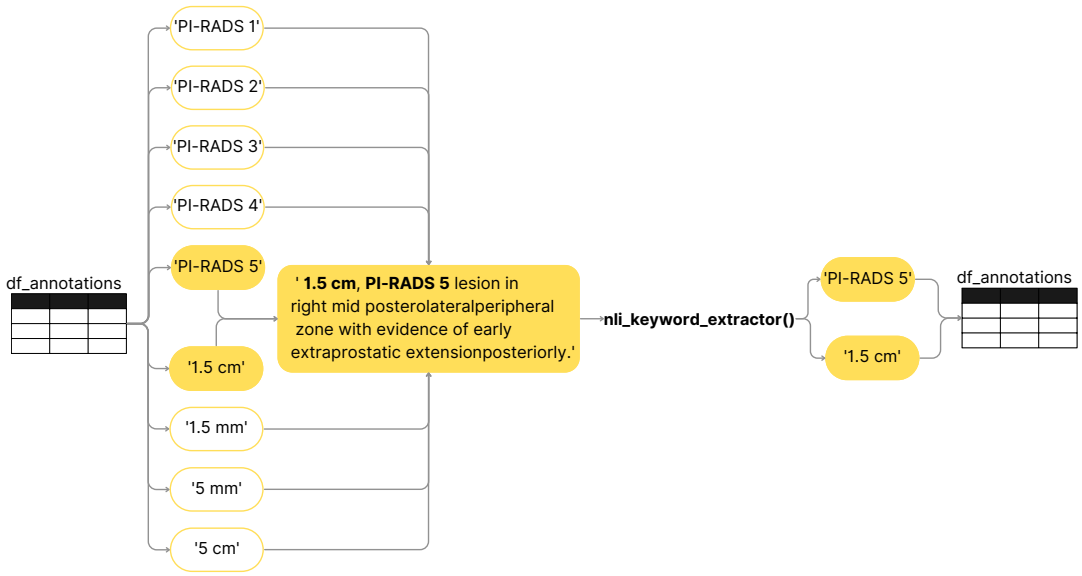


Figure 4.11: semantic-based extraction of PI-RADS score & lesion size

error cases are in the majority of cases correctly recognized as "no_match" by the model (see chapter 6 for performance review).

4.5.2 Extracting intra-procedural diagnostics

The extraction of the intra-procedural diagnostics (Gleason score and Gleason Grade Group) can be conducted using simple regular expressions and does not necessitate the employment of semantic approaches. This is partly due to the more structured way the pathologic diagnosis section is described, in contrast to the impression section of a radiology report, but also because of the characteristic syntax in which especially the Gleason score is usually documented (e.g. 3+3=6). The regular expressions used for extraction are documented in the methods marksheet in the attachment section. In addition to the Gleason Score and Gleason Grade Group extracted from the pathology single diagnosi sections, the biopsy coordinates can also be extracted from the matched preoperative ".fcsv"-files, by simply reading in the file as dataframe and selecting the x-,y- and z-coordinates. This step can be done at any stage of the diagnostics extraction and was only assigned to this subsection, because of the method used (similar to extracting intra-procedural scores in a non-semantic way).

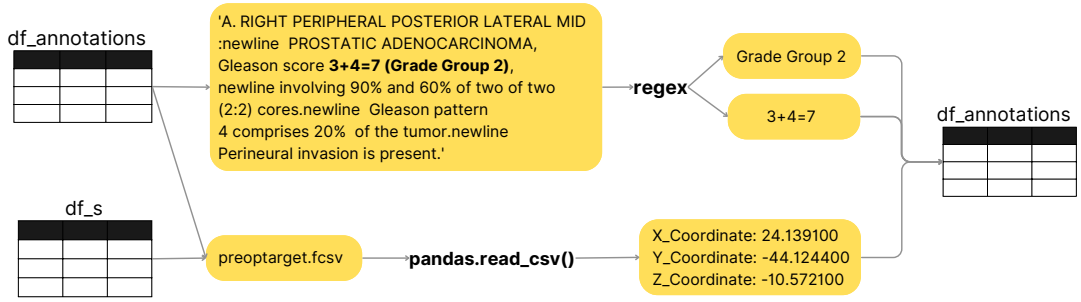


Figure 4.12: regex-based extraction gleason score & gleason grade group

4.6 Step 4: post-processing & dataset export

In steps one to three we have first imported all relevant data modalities, converted them into dataframes, preprocessed them, matched them first on patient-level and then on lesion-level and finally extracted all relevant intraprocedural and preprocedural diagnostics to construct labels for our selected image studies. At this point, two main dataframes are available, consisting of first all relevant image studies with absolute paths to their image series in "df_images" and second of all annotations (diagnostics) that could be extracted in "df_annotations".

Before moving to the final step of postprocessing the two dataframes and creating the actual final data cohort, a manual "technical validation" of the dataframes is conducted (by exporting the dataframes as csv-files for example). This is a necessary quality assurance step, where all rows in both dataframes need to carefully be inspected and validated for errors. It is expected that all rows with existing errors are either manually corrected or for the sake of simplicity discarded completely. This task, although being time consuming, should take a fraction of the time necessary to conduct an entirely manual curation of the dataset (more in depth discussion in chapter 6). After the manual validation of the two dataframes is completed the corrected dataframes can easily be re-loaded into the python environment as pandas dataframes.

Two basic final cleaning steps of the two corrected dataframes df_images and df_annotations are implemented in the current pipeline. First, both dataframes are "synced", based on the preprocedural studies that remain after all matching steps. Additionally, we define as minimal requirement for a "case" to be kept for it to at least have all correct image modalities and at least one set of biopsy target coordinates with a correctly matched pathologic diagnosis (either Gleason score or Gleason Grade Group or both).

This definition can easily be altered if required. All "cases" - that is image studies and extracted corresponding diagnostics - not fulfilling this requirement are discarded.

Data "export" refers to the creation of actual data folders that define the final prostate cancer image dataset. As described in chapter 3, two subfolder structures are created in the current version of the curation pipeline, consisting of identical images and labels, just in different data formats - one data cohort "dicom_insp_dir" is created in pure DICOM format for data inspection and visualization purposes, while another data cohort "nifti_dev_dir" is created for machine learning model development. The reason for this distinction is a simplified interaction with the final dataset for on one side easy data visualization and inspection and on the other side reduce required preprocessing for machine learning model development.

First, the DICOM format enables an easy web-based visualization via the "Open Health Imaging Foundation (OHIF)" viewer, needing no specialized local software, but just an easy to share browser link (after adding the necessary permissions). The current implementation includes uploading the final DICOM data to a cloud bucket in GCP, and importing the data to a DICOM data store, that can then be visualized via the OHIF viewer. Second, the NIFTI format enables a fast ingestion of the data into existing machine learning pipelines or the development of new ones, as the DICOM data format is not widely used for algorithm development (this might change in the future).

Important additional post-processing steps include creating DICOM Structured Reports for the point annotations and bounding boxes, "cleaning" the DICOM images (e.g. reducing sequences stored as multi volume to only single volumes, which is often the case for diffusion weighted image sequences, including multiple 3D volumes for acquired with varying b-values - here a filtering for high b-values is implemented) and converting the DICOM images to NIFTI format.

The bounding boxes are constructed by using the respective biopsy coordinate as center point and taking half the extracted lesion diameter extracted in each coordinate direction. The resulting 3D cube is then either stored in a DICOM Structured Report or as a NIFTI segmentation volume. The quality of the bounding box as approximation of a lesion segmentation is discussed in chapters 3 and 6.

The described post-processing and data export pipeline is visualized in figure 4.13.

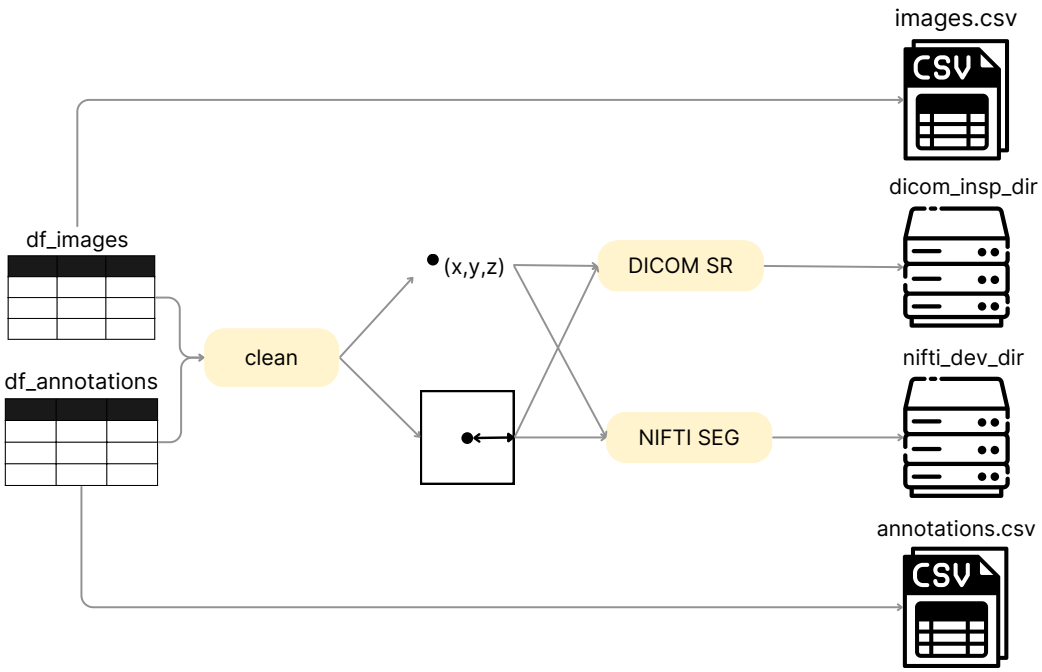


Figure 4.13: "Sync" image and annotation dataframes and use "cleaned" dataframes to create a DICOM folder for simple dataset inspection via OHIF webviewer and a NIFTI folder for convenient model development

4.7 Performance review

A brief quantification of the data reduction / "loss" during data curation and the performance of the "semantic" curation methods is given in the following section. A more detailed discussion of the automated data curation pipeline - including specific failures cases - is described in chapter 6.

4.7.1 Data reduction

Figure 7.5 quantifies the data reduction over the entire data curation process per "modality" and data curation phase. "Modality" in this context does not refer to the data type used, but should relate to different "units" the final created dataset can be characterized into. The phase 0-5 translate to the data curation phase described in the above chapter, as follows:

- **0-1:** import & pre-processing

- **1-2:** patient-level matching
- **2-3:** lesion-level matching
- **3-4:** post-processing & export
- **4-5:** manual inspection

The different data "modalities" visualised, are again depicted in table 4.1, quantifying the reduction or "data loss" for every modality over the entire curation process. As visualized only a fraction of all raw data inputs, can actually be used in the final dataset (all modalities showing a total data loss / reduction of > 80%). Looking at the number of image series for instance - starting at 31,671 thousand image series from the original requested image data - only 258 image series (making < 1%) are relevant for our use-case and are included in the BWH-MRIGPB dataset as the three required MRI sequences (T2-weighted, high b-value diffusion-weighted and apparent coefficient maps). A more detailed discussion and interpretation of the observed data reduction is conducted in chapter 6.

data modality	start size*	final size	total data "loss"
image series	31671	258	99.19%
rad diag sections	11399	99	92.92%
rad reports	10788	86	99.20%
path diag sections	7844	105	98.66%
path reports	3286	86	97.38%
image studies	1916	86	95.51%
biopsy cases	875	86	90.17%
biopsy targets	591	105	82.23%
patients	579	84	85.49%

Table 4.1: Relative data reduction per data "modality" during entire data curation process. See figure 7.5 for full visualization.*Counts for radiology diagnosis sections, pathology diagnosis sections and biopsy targets start in phase 1 (after import and pre-processing), in contrast to the other modalities starting at 0. As in step 0 no pre-processing and filtering has taken place, including counts for the above mentioned modalities would skew the graphs in 7.5 unnecessarily due to very high numbers.

4.7.2 Methods evaluation

Throughout the outlined data curation pipeline multiple methods related to clinical text mining have been used. A complete list of all methods used and some implementation details are listed in figures 7.6, 7.7 in the attachments. Due to time-constraints not all used methods can be properly evaluated in the frame of this thesis. In table 4.2 a brief performance review of all "semantic"-based approaches is given, being of particular interest due to their "probabilistic" nature. All used "semantic"-based approaches show

method	task	#error/ #prediction ratio	accuracy	tot data loss	runtime
ES	extracting correct image series	8/267	97%	0%	42min27sec
ES	matching correct rad diagnosis section	5/108	95.4%	83.9%	27min55sec*
NLI	matching correct path diagnosis section	8/108	92.6%	75.5%	27min55sec*
NLI	extracting PI-RADS score	0/108	100%	0%	55min52sec*
NLI	extracting lesion size	11/108	89.8%	0%	55min52sec*

Table 4.2: Brief performance evaluation of "semantic"-based approaches used for finding the correct imaging modality (T2W, HBV, ADC) for every imaging study, matching correct single diagnosis sections from pathology and radiology reports, and extracting pre-procedural diagnostic scores from radiology single diagnoses. "ES" refers to the "es_keyword_extractor()" function described above, using similarity of word embeddings and "NLI" refers to the "nli_keyword_extractor()", based on entailment probability of models trained in a natural language inference scheme. All methods were run on a Mac M1 Pro with 10-core CPU and 14-core GPU.*the runtime for these tasks refers to the time for running both methods together.

an accuracy of around and above 90%. The total number of errors for the annotations and image dataframes are as follows:

- df_images: 8 errors / 267 predictions, **error rate 3%**
- df_annotations: 43 errors / 4 tasks * 108 predictions, **error rate 10%**
- total "positive" predictions made (df_images + df_annotations): 267 + 4*108 = 699,

total "false positives": 8 + 43 = 51

"Positive predictions" in this context refers to the fact that we are focusing only on the model predictions that actually lead to a match or extraction. We are neglecting all "no_match" predictions for time-reasons, as they are not directly visible by inspecting and verifying rows in the created df_images and df_annotations dataframes. For the imaging series the number of predictions corresponds to the number of rows of the df_images dataframe (each containing a different series description). The annotations dataframe df_annotations consists of 108 rows (105 after manual correction), but for every row four different "semantic" prediction tasks are performed (matching single diagnosis section of pathology and radiology reports and extracting two pre-procedural diagnostics from the radiology reports), explaining the factor four in the above formula.

From all tasks the lesion extraction performs worse (while also taking the longest to run), and the best performing method is the PI-RADS score extraction, classifying all radiology single diagnosis sections into the correct corresponding PI-RADS score class. All accuracy scores were calculated by manually counting the errors made by every method, divided by the total amount of predictions / classifications made. The accuracy should only be considered as a "first indicator" of the performance of each method for the respective task (see chapter 6 for further discussion).

Important to note is the data reduction that comes with every above described method. The data reduction describes by how much the amount of data points has "shrinked" during the lesion level matching in which the above methods were applied. It is of high relevance for the evaluation of each method, as it is important to understand the ratio between the reduction of data points because of "bad data" (corresponding to a working curation pipeline) and data reduction because of method errors in the curation pipeline. As there is a significant data reduction for the "single diagnosis" modality in the lesion level matching phase (see also 7.5), both used methods can potentially contribute significantly to "losing data" by producing false negative predictions (i.e. not finding a matching single diagnosis for a specific anatomical region of the prostate at all). The accuracy of the methods in this case - calculated based on the number of rows of the final outputted df_annotations dataframe - is not sufficient for capturing the performance of the algorithms. To get at least an estimate of the number of false negatives, the number of "no_matches" during execution was counted and amounts to < 10 cases for single diagnosis sections form both pathology and radiology reports. As the extraction of the correct image series and the extraction of the pre-procedural diagnostics do not contribute to any data reduction, the data reduction does not influence the provided accuracy scores.

The total runtime of the data curation pipeline, including data pre-processing and export on a Mac M1 Pro with 10-core CPU and 14-core GPU is of about 2h19min. In contrast, manually inspecting and correcting the resulting image dataframe (df_images) and annotation dataframe (df_annotations) took approximately 5h40min for both tables. Despite relatively high accuracy scores of the methods outlined above (resulting in relatively few rows in both dataframes that needed manual correction), the manual inspection took roughly twice the amount of time of the actual curation and be further reduced significantly on dedicated hardware.

A more elaborate discussion of the performance of the used methods and specific failure cases is given in chapter 6.

4.8 Usage notes and code availability

The entire data curation pipeline is implemented in modularized python scripts with one main function "generate_dataset()" taking a YAML configuration as input. Via the configuration file the following parameters are easy modifiable with the objective to making the data curation script more flexible and reusable:

- **absolute paths to all raw input files and directories:** pathology report txt, radiology report txt, absolute path to biopsy case archive, absolute path to anonymized image study csv, absolute path to conversion table csv for anonymized image study
- **pre-processing parameters:** minimum date considered for clinical reports, to be extracted modalities for image studies
- **deep learning model hyperparameters** model used (hugging face ids), distance metrics used for similarity-based matching and extraction, thresholds used for both models, type of label augmentation (for series descriptions and anatomical matching), target strings for diagnostics extractions
- **data export options:** type of DICOM Structured Report that should be created (based on annotation type, either point or bounding box), image export datatype (dcm, nifti, mha), annotation export datatype (dcm, nifti), absolute paths to output directories for images and annotations

After carefully verifying that no sensitive health information is explicitly or implicitly included, the code will be made publicly available via GitHub. The hope would be that

either portions of the code are useful for trying out various syntactic or semantic clinical keyword extraction methods for similar use cases, the automated medical data curation process can serve as an example for automating data curation in other hospitals / medical cohorts, or the code could even serve as boilerplate code to similar medical data curation projects, simplifying and accelarating the curation process to increase the amount of publicly available medical datasets for AI development.

5 Discussion

The discussion section is divided into a first "Dataset" section, referring to and dicussing the BWH-MRIGPB datset introduced in chapter 3, and a second "Curation" section, corresponding to the data curation approach described in chapter 4. An additional section on the applicability of publicly-available prostate cancer detection models on the BWH-MRIGPB dataset is provided.

5.1 Dataset

In the following the newly introduced first version of the BWH-MRIGPB dataset will be discussed, including quality considerations, limitations and potentials. This section corresponds to chapter 3.

5.1.1 Quality evaluation of provided images and annotations

Bbox annotations - limitations

As mentioned in chapter 3, the lesion bounding boxes are only approximations of actual expert-created lesion delineations / annotations. Although, "relaxing" the initial task of "prostate cancer detection" slightly to "zonal prostate cancer detection", meaning we focus on getting the anatomical region the tumor is located in right, and try to contain it in a bounding box, there are still significant limitations and potential for error in the current approach. We define lesion bounding boxes as "good enough" for the clinical diagnosis workflow, if:

- **They are located in the correct anatomical zone:** this can be verified even by non-radiologists after some training and basic knowledge about the prostate anatomical regions.
- **They contain the majority of a single lesion inside them:** the "majority" again leaves room for setting a specific threshold, and is only verifiable against explicitly derived expert-derived lesion annotations.
- **They are "precise enough":** meaning they are close enough to the actual lesion boundary to still be of clinical use. The correct trade-off for specific clinical tasks

(for example MRI-guided prostate biopsies) needs to be validated with practicing clinicians and radiologists to determine accuracy and requirements.

In figures 5.1 and 5.2, two limitations and potential errors sources of the constructed lesion bounding boxes are highlighted. The size of the extracted lesion diameter plays an important role for either significant overestimation of the actual lesion size or "missing" a significant portion. With a high extracted lesion diameter the lesion bounding box can become quite large, covering a large portion of the MRI scan (even extending beyond the prostate gland), and making the lesion annotation less precise. Especially for an elongated lesion morphology, the amount of non-lesion tissue that is located inside the bounding box increases, as lesion bounding boxes are constructed with a constant size in all coordinate directions, usually using the longest lesion diameter (noted down by radiologist in pre-procedural radiology report impression section). A different failure case can arise, with small estimated lesion sizes, as a significant portion of the lesion tissue might be located outside of the lesion bounding box in case of non-centered biopsy target coordinates. Both failures cases need to be further investigated to assess the influence on the accuracy and correctness of the lesion bounding boxes.

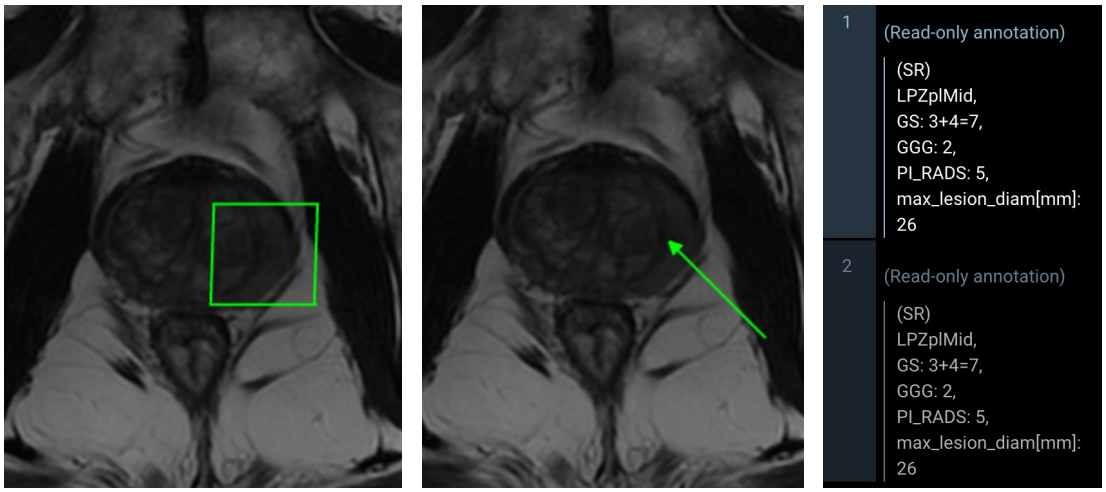


Figure 5.1: Example of rather big estimated lesion diameter by expert radiologist (26mm), leading to large bounding box, making it less accurate and deviate more from actual lesion segmentation. The more elongated / stretched a lesion is in one direction, the less accurate a lesion bounding box can become.

An alternative approach to assessing the quality of the constructed lesion bound-

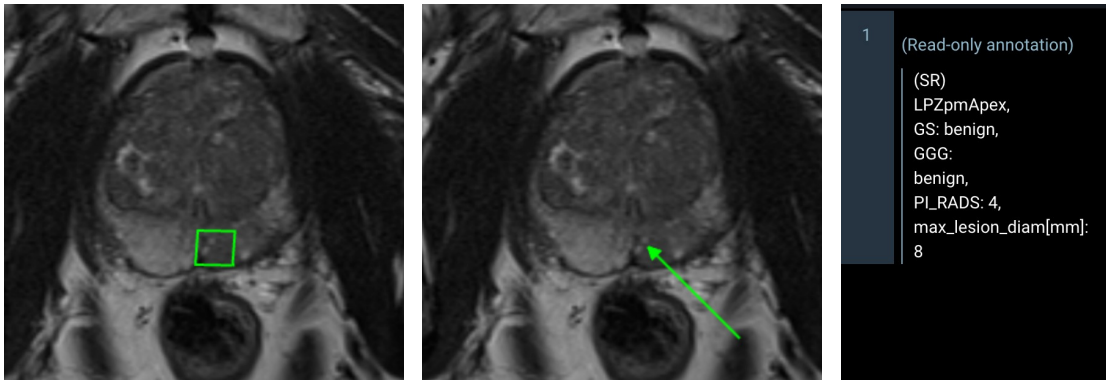


Figure 5.2: Example of relative small estimated lesion diameter by expert radiologist (8mm), leading to precise area in the prostate on one side, but also increasing sensitivity to biopsy target locations that are not approximately in the middle of the predicted lesion. Here, it appears biopsy coordinate is located to the side of a lesion (darker area), making constructed lesion bounding box less accurate.

ing boxes - instead of only focusing on direct clinical applicability - is their ability to help algorithms become better at various prostate-related medical imaging tasks (and by that eventually becoming more relevant for the clinical workflow). The hope here is, that despite their approximate nature, the lesion bounding boxes can help to train specifically "data hungry" deep learning models on more and diverse data to become better at the task of detecting clinically-significant prostate cancer. Considering that some algorithms in any case output predicted rectangular lesion volumes (see subsection on "Model benchmarking on the BWH-MRIGPB dataset"), also using rectangular lesion bounding boxes as ground truth might be less problematic. Approaches like the "weak labels" used for training a portion of the baseline models in the frame of the PI-CAI challenge [**bosma2022annotationefficient**] on "AI-generated" lesion delineations, together with a very low Intersection over Union threshold of =0.1, used within the PI-CAI challenge as "hit criterion" for a correctly predicted lesion (see chapter 2), indicate the complexity of prostate cancer detection and the potential of increasing model performance based on "easy-to-get" but slightly less accurate annotations. No conclusive evidence for the quality of lesion bounding boxes can be demonstrated at this stage and further investigation will be necessary.

Image quality - limitations

All MRI series represented in the final BWH-MRIGPB dataset were acquired during actual pre-procedural MRI exams in the clinical workflow and do not stem from a specifically designed scientific study. As a consequence - even after a successful structuring and matching of all data elements (see chapter 4) - image quality can vary significantly between images. Two example of lacking image quality are visualized in figures 5.4 and 5.3.

5.1.2 General pros & cons of the BWH-MRIGPB dataset (v1)

Main disadvantages of the BWH-MRIGPB dataset v1 include:

- **Small size:** compared to the currently used prostate cancer related publicly-available datasets (see attachments), the BWH-MRIGPB dataset has with 86 curated imaging studies and 105 corresponding biopsy coordinates, a relatively small number of cases. Especially, in the context of deep learning model training the dataset size significantly influences the performance of trained algorithms. As the BWH-MRIGPB dataset was curated as similar as possible to the PI-CAI "Public Training and Development" dataset, the hopes are that it can be submitted to the challenge as additional training data.
- **No expert-derived lesion segmentations:** The BWH-MRIGPB dataset does not include expert lesion annotations at this stage. Therefore, it remains to be conclusively validated whether training models on our annotation approximations are useful in the clinical workflow, and if the dataset can be used for training and evaluating classical prostate cancer detection models.

Main advantages of the BWH-MRIGPB dataset v1 include:

- **Extensive clinical metadata:** the BWH-MRIGPB dataset comes with histopathology-validated lesion point annotations for every case and with ten different clinical variables which are extracted from pre-procedural radiology or patholgy reports. The hopes are, that this metadata can be used by algorithms in addition in addition to the provided image data, and potentially increase performance for prostate cancer detection, or even open up new tasks related to the diagnosis of prostate cancer.
- **Easy extendability:** a main strong point of the BWH-MRIGPB dataset is the way it has been created. The hopes are that by using the automated structuring and labeling pipeline presented in chapter 4, it will require only minimal effort and modification to add more raw input data and "continuously automatically

grow" the number of image studies and labels with every new MRI-guided biopsy case that is performed at the Brigham and Women's Hospital. Assuming approximately three biopsy cases per week, this could potentially double to triple the current dataset size in one year (3x52=156) - only by re-aggregating all raw data sources and re-running the dataset generation code (see estimated curation time in chapter 4 and "Curation" section in chapter 6.

5.2 Curation

This section corresponds to the "automated medical data curation" pipeline described in chapter 4 4.7. Specifically, the approach to data curation taken is discussed, as well as data quality and the performance of applied methods.

5.2.1 Failure cases

While the accuracy scores in 4.2 provide only limited insight into the actual workings of the language models used for the lesion-level matching, in the following all error sources in the matching and extraction are grouped into specific categories, that should provide further insight into advantages and limitations of the used methods, as well as information about in-consistencies in the aggregated medical text data. Syntactic approaches don't contain any "probabilistic" inference and are deliberately used as coarse filters in our pipeline, hence, they will not be part of this discussion. It should be noted however, that the syntactic approaches are of equal importance, especially since they are responsible for a significant data reduction during the patient-level matching (see 7.5) and should be evaluated in future work. All failure cases related to "semantic"-approaches are assigned to one of three categories:

- **Category 1 - model errors:** describing the subset of all errors, that can be directly traced back to the failure of one of the two used large language models (see chapter 4).
- **Category 2 - documentation inconsistencies in clinical workflow:** describing the subset of all errors, that can be traced back to data inconsistencies in the process clinical reports, biopsy targets or imaging series descriptions.
- **Category 3 - vague terminology & radiologist inter-reader disagreement:** describing the subset of all errors, that can be traced back to different characterizations of lesions (in terms of aggressiveness, location and size), and vague terminology (for example varying ways of describing the same anatomical prostate region).

All quantifications are computed with respect to the total number of model "predictions" (referring to all tasks and model types listed in 4.2) computed for getting the final two dataframes. To get an estimate of the relative fraction of each failure case in comparison to all failure cases, we use the total number of "observable" predictions from chapter 3, amounting to 699 total predictions (for df_images and df_annotations), with 51 "false positives" counted during manual inspection.

FC1: similar semantics - category 1

19.60% (10/51) of total error cases, relating to classifying / extracting imaging series descriptions as well as anatomical prostate regions, lead back to a "lack" of semantic difference between candidates. Especially series descriptions and anatomical region descriptions in pathology reports are usually not longer than 30-50 characters, not representing a lot of "context" for the BERT-based word embedding approach. It is in general observable that for both tasks using the "ES" approach - i.e. computing the cosine similarity for distinguishing vectorized short (30-50 characters) medical text sequences - results in rather similar similarity scores. As depicted in figure 5.5, specifically if candidate strings are only distinguishable by a few characters, this can lead to false classifications (in figure 5.5 the identical model failure is demonstrated on the two different tasks of the "es_keyword_extractor()" for pathology diagnosis sections and image series descriptions). In subsection 5.2.2 augmentation techniques are explored to increase semantic similarity and reduce "noise" from imperfectly created candidate strings.

FC2: model hallucinations - category 1

3.92% (2/51) of total error cases. "Hallucinations" in this context refers to semantic-based extraction tasks (specifically extracting lesion size and pi-rads score), where the model "extracts" a value that does not exist in the original diagnosis section. For example for a single radiology diagnosis section, "Changes of BPH in the central gland. Prostate volume 87 ml." - not actually containing a lesion size, the nli_keyword_extractor() extracted a lesion size of "87" instead of returning a "no_match". This error was most likely caused by the semantic similarity of the construction lesion size targets (see chapter 4), e.g. "4mm", an the volume indication in the single diagnosis. A very short target label could be an influencing factor for the miss-classification, or even the fact that actual numbers are the main focus. As potential remedy, a similar augmentation approach to failure case 1 could by tried.

FC3: not meaningful series descriptions - category 2

3.92 % (2/51) of total error cases. This failure case is explicitly separately listed from FC1, as it does not refer to "missing" or "similar" semantics of candidates but to image series descriptions that actually do not contain necessary semantic information for clear classification. As a general consideration, referring to all analyzed image series descriptions, it becomes apparent that current used terminology is not consistent over different imaging studies, not meaningful concerning the content of the respective image series and very hard to interpret. For one it is - even for humans (and to a certain extend medical experts) - impossible to infer from the analyzed image series description whether an endorectal coil was used for the acquisition or not (see a list of all series description of the final dataset in figure 4.3). For the count of endorectal coil acquired images provided in 7.2, every DICOM study had to be loaded manually into Slicer and verified, as in this case even the DICOM attributes across different MRI vendors were not standardized making them unreliable for endorectal coil identification. In addition, the image series descriptions do not allow for a clear correspondence to what anatomical region of the body is analyzed. In figure 5.6 an example case is visualized of an abdominal scan that was wrongly matched during the patient-level matching, and not recognized until the final visual validation step.

FC4: faulty manual data input - category 2

15.68% (8/51) of total error cases. **Manual** documentation and digitization is a main cause for data inconsistencies in clinical reports, as well as in biopsy target labels and imaging series descriptions. Figure 5.7 depicts a human-made error while manually inputting the abbreviated anatomical region of the prostate the targeted biopsy is located in. As a general consideration, we advocate for more automated - machine-inputted - documentation in clinical workflows. A significant amount of "unnecessary" data inconsistencies (including wrongly spelled words, spaces in report headers, wrongly inputted series descriptions (not matching with DICOM attributes), etc.) can clearly be traced back to human-made mistakes.

FC5: ambiguous lesion characterization - category 3

19.60% (10/51) of total error cases. Especially the description of anatomical prostate zones is subject to significant variation in the way they are referred to from radiologist to radiologist and between pathology vs radiology reports. It should be noted that the pathologist is in general copying the regional prostate information provided by the interventional radiologist and hence all variance in the anatomical region description of the prostate is due to radiologists using different formulations to describe the

same anatomical region of the prostate. Example inconsistencies leading to incorrect anatomical matching results, include describing the same lesion as being located in the "right transition zone at the apex" or "in the peripheral zone, located in right, apex, posterolateral region (PZpl)". For this specific example the anatomical region is described in two slightly different ways in the same pre-procedural report, leading to errors when matching histopathology and biopsy target coordinates.

FC6: inter-reader disagreement - category 3

5.88% (3/51) of total error cases. In these cases the interventional radiologist (conducting the MRI-guided prostate biopsy) did not agree with the interpretations made by the previous Radiologist on clinically-significant lesion candidates, see figure 5.8. As a result two different anatomical locations in the prostate are provided, with additionally differing attributed PI-RADS score / lesion aggressiveness. This case was manually corrected and set to the final judgement of the interventional radiologist.

FC7: incomplete radiology impression sections - category 2

31.37% (16/51) of total error cases. Data generated in clinical workflows is rarely generated with downstream analysis tasks in mind. This is a main contributor for making medical data curation a very tedious and time-consuming task. A third of all recorded errors made in the automated lesion-level matching are due to lesion size, PI-RADS score and / or anatomical region not being mentioned in the impression section at all. In some cases these "no_matches" turn out to not relate to actual extraction errors but are due to for example lacking imaging quality, or previous therapy and hence incomplete diagnostic scores given by the Radiologist. In the majority of cases the impression section however, is incomplete because a previous radiology report is reference or the "FINDINGS" section is referenced and includes the complete pre-procedural diagnostics (manually extracted and corrected), as visualized in 5.9. In other cases a clear matching failed due to multiple observations documented in a single diagnosis section, see 5.10. It should be noted here, that simple modifications in the clinical workflow might be possible that make a significant difference for downstream data curation and analysis tasks. For example, by simply introducing "best practices" for writing impression sections with the same structure, the error rate in our case could be reduced by about 30%.

5.2.2 (L)LMs for "out-of-the-box" clinical keyword extraction

Large Language Models are not an "intuitive" choice for medical data curation, as they are probabilistic models with limited explainability and a to a certain extend a "black

box" character. In the frame of this work we demonstrate the use of two language models for medical keyword extraction, as a way to "automate" structuring and labelling in the data curation process. The hope is, that by automating the otherwise tedious and error-prone task of extracting diagnostic scores from unstructured clinical reports we can significantly accelerate the curation process and make it more transparent. In the following a brief deep dive into the reasoning behind using LLMs and how they are used is provided.

Converting clinical Named Entity Recognition into simple text classification

Clinical Named-Entity-Recognition (clinical NER) is still a very difficult task, even for advanced large language models like BERT. The task of named entity recognition from a machine learning perspective consists on a high-level in assigning a pre-determined label to each word in a sentence. After having "recognized" different "entities" these can easily be extracted from unstructured clinical narratives.

As it turns out, from a practical perspective, machine learning-based clinical NER models still depend to a significant extend on the text corpus they were trained on, and even models pre-trained or fine-tuned on clinical text data (e.g. MedSpaCy's pre-trained algorithms) perform poorly when trying to extract PI-RADS scores or anatomical prostate regions with models that were trained on generic clinical texts (due to time-constraints we are not able to provide any extensive benchmarks or reviews of various tested approaches at this stage).

Motivated by pure practical "necessity" - in order to find a feasible approach for automated semantic matching and diagnostics extraction and avoid an extremely time-consuming and error prone manual matching process of anatomical prostate zones (see chapter 4) - we propose "simplifying" the complicated task of clinical named entity recognition to simple "text classification". The reasoning behind this "task re-phrasing" is that from a practical perspective in most cases when we are trying to automatically extract a keyword from unstructured clinical text **we know what we are looking for**. In our case we are specifically looking for anatomical regions of the prostate and therefore do not require labeling of every word in the clinical text (but need only to recognize a specific keyword). Hence, if there was a way to use the fact that we want to only extract a very specific keyword from a vast amount of unstructured text without needing clinical NER, this might simplify the task substantially.

As described in chapter 4, we propose a simple method that is based on finding meaningful word embeddings for both unstructured clinical text snippets, as well as

a provided target text snippets (that we can provide, as we have a specific keyword we would like to extract) and finding the target that best "classifies" or "labels" the respective text snippet. A significant caveat in this approach is that usually we need to first make sure that we have only a single occurrence of the required keyword to extract in the clinical text, as otherwise a simple text classification might lead to wrong results (precisely what happends in 5.10 described in the failure case section). To however, create the sections with single occurrences we still rely on mostly "syntactic" NLP methods to create candidates that can then be classified (see pre-processing section for creating single diagnosis sections for pathology and radiology reports in chapter 4).

For vectorization we use two pre-trained language models. It is important that the language models work "out-of-the-box" - meaning no training or fine-tuning is required - as again from a practical perspective, wanting to apply the language models for automated data curation, we need an efficient and guaranteed working approach. Fine-tuning, in general entailing labelling a subset manually, then training and evaluating performance, to finally hoping that the model actually does work better than manual or syntactic approaches, does not meet this requirement.

Why ES and NLI?

Due to time-constraints no extensive research and/or benchmarking has been conducted for chosing specific algorithms or specific techniques. The two language models described above have been identified based on the following "empirical" considerations and observations:

- **BERT-based word embeddings similarity approach:** The BERT base model is with 109 million parameters significantly smaller than BART, performs "good enough" for our data curation use-case and provides more interpretability of resulting scores - the embeddings can be visualized and compared. The downside of this approach is that word embeddings are still all rather similar, resulting in similar cosine similarity scores for multiple candidates, and making the approach "prone to noise" in the clinical text snippet. By applying augmentation techniques (see "Limitations" below) the "weight" of "noise" in the single diagnosis sections can be reduced slightly, but the approach still does not perform very well for processing radiology impression sections, being a signifcantly closer to unstructured text, than pathology report diagnosis sections.
- **BART-based natural language inference approach:** The BART large model has four times as much parameters, as the BERT-base model and is specifically trained

for natural language inference. Both an increased model size, as well as the fine-tuning on natural language inference might be reasons for the better performance of the model, when applied on radiology impression sections. Interestingly, the used BART model was not trained at all on clinical texts. The "entailment-contradiction" framing of Natural Language Inference tasks fits our use-case of extracting clinical keywords from clinical text snippets perfectly, which manifests itself also in the output entailment probabilities. While the BERT-based approach, as mentioned, usually outputs similar cosine similarity scores for all candidates, the NLI-based approach in general shows a remarkable differentiation between the first "entailed" candidate, and the non-matched candidates. A disadvantage of the NLI-bases model however is precisely that a more sophisticated "Neural Network-based" process is used for classification (instead of cosine similarity), hence making the model less interpret-able.

Limitations

As described in failure case 1, a difficulty arising when applying BERT-based models to the classification / extraction of clinical keywords is the limited "context" that is provided, especially for our use cases of classifying short text snippets between 30-50 characters corresponding to prostate anatomical regions described in pathology reports and image series descriptions. Specialized medical vocabulary and abbreviations further increase the difficulty of finding meaningful embeddings by using "out-of-the-box" pre-trained models without fine-tuning. To increase the semantic difference between candidates, we applied the following simple augmentation technique for image series descriptions:

1. Identify character sequences that are "characteristic" for a specific image modality, for example "t2" for T2-weighted image series or "1400" for high b-value diffusion weighted imaging.

2. Define a static dictionary mapping all specified "characteristic" character sequences to significantly longer descriptions. For example we convert "t2" into its official definition of "t two produced by using longer TE and TR times" and "t1": "t one produced by using short TE and TR times".

3. The dictionary created in step 2 is then applied to both the target string (e.g. "AX_T2" for retrieving axial t2-weighted imaging series) and the candidate strings together with the standard pre-processing (lowercase, discard underscors etc.)

The augmentation significantly helped increasing the number of correctly matched series descriptions however, no quantitative analysis was conducted so far to underline

the approach.

Using zero-shot text classification for clinical keyword extraction, can lead to what we define "model hallucinations" (see failure case 2). Hallucinations should refer to extractions of for example lesion sizes or PI-RADS scores from single diagnosis sections that are not present at all in the original text snippet. Turning the extraction task into a "simpler" classification task - enabling us to use LLMs for medical entity exraction without fine-tuning - can lead to the model coming up with "invented" values. We do not provide any guarantees that extracted keywords are actually "included" in the original diagnosis section, but emphasize the need for manual inspection after the extraction has taken place. This - while still requiring some manual work in the data curation process - should still be faster than extracting all diagnostics manually form the start. In addition, considering the low percentage of error cases related to "invented" values (failure cases 2), it appears that language models pre-trained on Natural Language Inference tasks are rather well at identifying "wrong labels".

In general, a more extensive benchmarking is needed, evaluating each approach on a larger dataset, as well as comparing the performance of each approach to derive more specific pro and cons for each method (interesting would be to evaluate the importance of the natural language inference approach by testing the same model with and without fine-tuning on an NLI dataset).

5.2.3 "Code-based" curation

In the beginning of chapter 4 three main requirements for any (medical) data curation task are presented: transparency, efficiency and maintainability. In chapter 4 we outline an entirely code-based curation pipeline with following advantages:

- **short curation time:** with the currently implemented data curation and the necessary "up-to-date" raw data sources locally available, a complete labelled dataset can be automatically be created in under three hours on commodity hardware.
- **easy "debuggable":** inconsistencies in the final dataset can be "debugged" in a regular software development workflow. As the final dataset is created by just a single function call "generate_dataset()" in a modulraized python script, all intermediate stages can be visualized and easily be modified.
- **easy "maintainable":** Instead of having to start from scratch every time an updated "data aggregation" is performed (which would be the case when manually

curating the dataset, as all data points in the updated raw data elements would need to be verified and compared), a code-based approach enables easy updatability and slight modification in the data curation pipeline.

Disadvantages of the proposed approach include:

- **missing guarantees for correctness:** the usage of "probabilistic" text mining approaches prevents giving hard guarantees for the correctness of the annotations.
- **questionable generalizability:** different hospitals may have significantly changing clinical workflows, resulting in different data structures and formats. Hence, the practicality of the proposed curation pipeline in slightly different environments would need to be firther investigated.

We conclude that significant more time is needed to properly benchmark all used methods, validate the correctness of the created dataset and the foremost the usefulness of the approach.

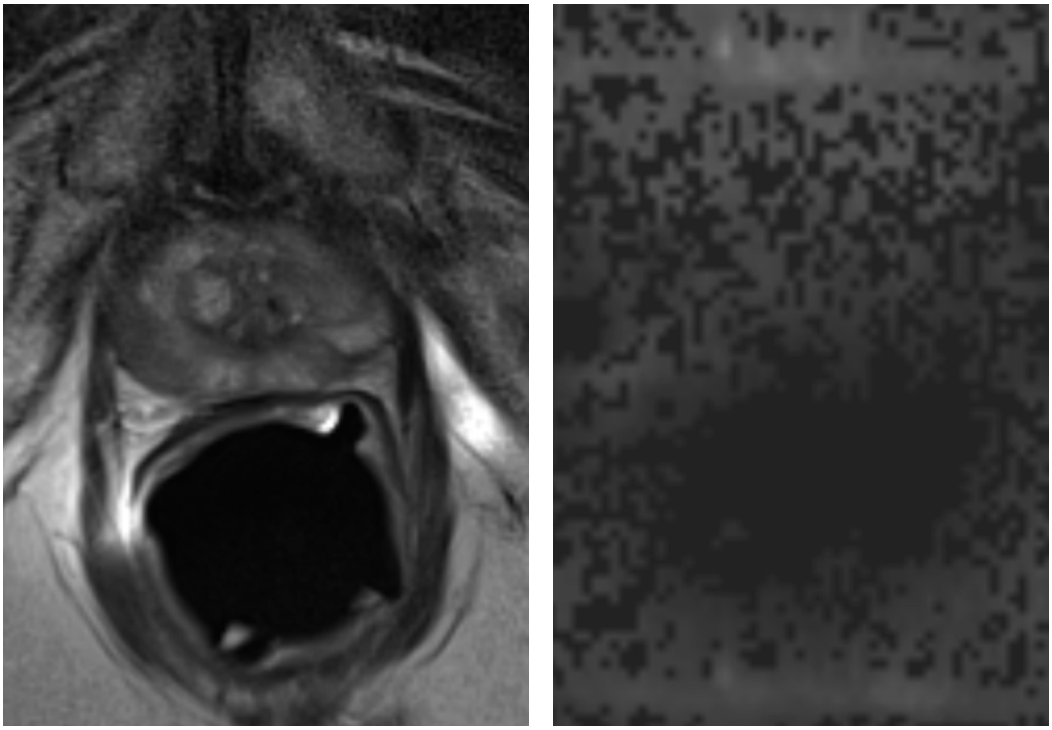
5.3 Model benchmarking on the BWH-MRIGPB dataset

Due to time constraints and the non-validated nature of the constructed lesion bounding boxes (needed for proper evaluation, see chapter 3), no complete quantitative benchmarking of various prostate cancer detection models was performed on the current baseline version of the BWH-MRIGPB dataset. However, the pre-trained semi-supervised nnDetection model, introduced as baseline model as part of the PI-CAI challenge (see chapter 2), has been applied to a subset of the dataset, showing some interesting qualitative results, as depicted on figure 5.11. This particular version of the nnDetection model was selected, based on two reasons:

- From challenge organizers designated as currently best performing baseline model on PI-CAI challenge public training and development dataset (no official benchmarking provided).
- Model output suits lesion bounding box annotations, as the nnDetection framework inherently produces rectangular lesion candidates, and not pixel-level segmentations.

It shall be noted, that these examples are at most a proof that publicly available models can indeed be applied to the BWH-MRIGPB dataset. Tentative interpretations about the different nature of the input imaging sequences between the PI-CAI challenge public training and development dataset and the BWH-MRIGPB dataset can be drawn (as

done in the figures description), but no conclusive information should be derived at this stage. For a proper benchmarking (see future work section), the quality of the derived lesion bounding boxes would first need to be assured by for example direct comparison with lesion bounding boxes drawn by an expert radiologist. Only after making sure that the automatically created bounding boxes represent an acceptable ground truth, are quantitative evaluations of different models and architectures reasonable.



IMPRESSION
1. Suboptimal evaluation due to nondiagnostic
diffusion-weighted sequences owing to artifact from the patient's
bilateral hip prostheses. However, within these limitations,
there are two PI-RADS 4 lesions in the right mid gland (1.4 cm)
and left mid/apex (0.9 cm) peripheral zones. The left mid/apex
lesion demonstrates focal angular bulging, raising the
possibility of minimal focal extraprostatic extension.
2. Sigmoid diverticulosis.

Figure 5.3: Example of endorectal coil acquired MRI sequence with, showing the T2-weighted image series on the left, and a diffusion-weighted series with very bad quality on the right. Having additional access to and information from pre-procedural radiology reports accompanying the imaging studies can help finding series with lacking image quality (see impression section on figure).

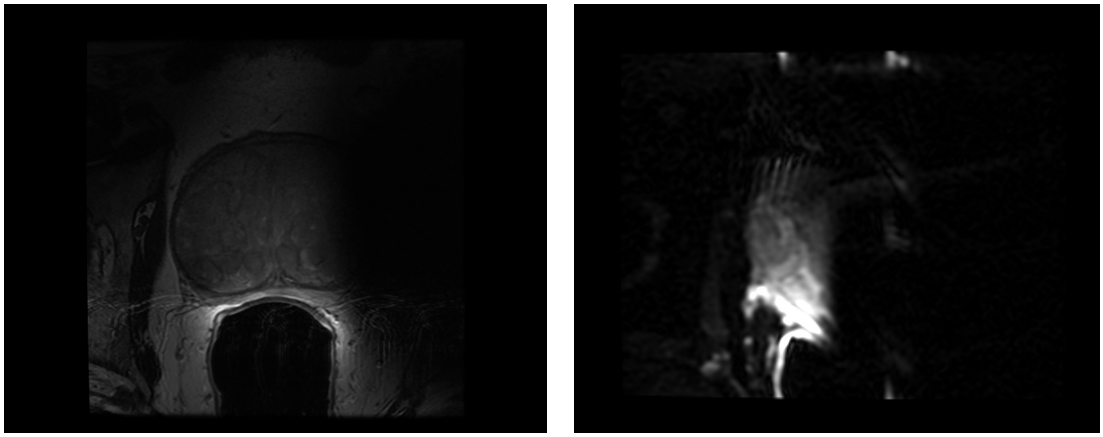


Figure 5.4: Example of axial T2-weighted and axial difusion-weighted MRI scans, acquired using an endorectal coil. Clear artifact on the right side, almost covering half of the prostate and making DWI scan almost impossible to read. To be determined if exam is too bad to be kept in final dataset, or if it might help AI models to be able to recognize artifacts and handle "messy" real-life data.

15_AX_T1_DYNAMIC
1__3PL_LOC
2__SAG_T2
3__AX_T2
4__COR_T2
5__AX_3D_T1_FS
6__AX_DIFFUSION_TRACEW_DFC
7__AX_DIFFUSION_00_ADC_DFC
8__AX_DIFFUSION_C_BVAL_DFC
9__AX_DIFFUSION_EW_DFC_MIX
10_AX_DIFFUSION_DC_DFC_MIX
11_AX_DIFFUSION_AL_DFC_MIX
12_AX_T1_DYNAMIC_DRY_RUN
13_AX_T1_DYNAMIC
43_POST_AX_T1_FS_VIBE
44_AX_3D_T1_FS_POST_pelvis
45_SUB_S20_S13_1

PATHOLOGIC DIAGNOSIS:

A. PROSTATE, LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID:
Benign prostatic tissue.

B. PROSTATE, RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID:
Benign prostatic tissue.

Figure 5.5: Right: full diagnosis section of pathology report with two targeted biopsies with very similar anatomical region description leading to almost identical word embedding, Left: subset of image series descriptions of image study with very similar semantic and structure for "ADC" and "HBV/DWI" image modality, leading to wrong classification of high b-value diffusion-weighted image series.

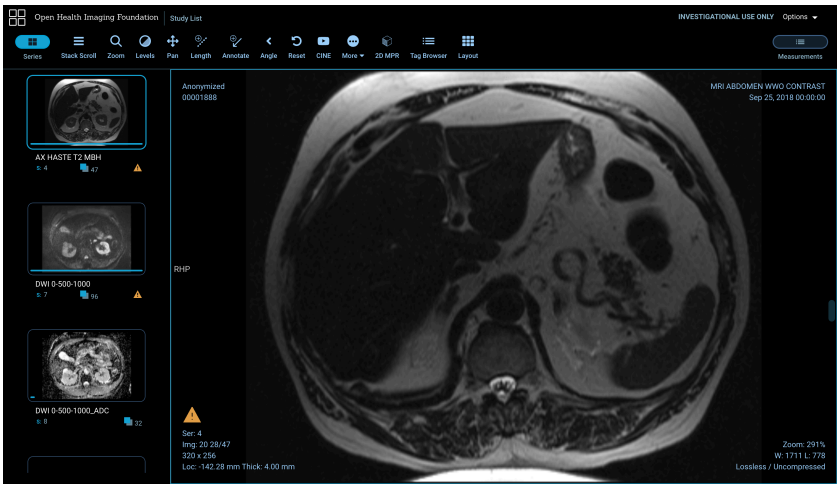


Figure 5.6: Non-prostate MRI in final dataset, due to patient-level matching error - only recognized during visual validation of final matched dataset. Reason: Temporal matching failed ("finding most recent pre-procedural MRI scan before biopsy procedure"), because of wrongly classified radiology report. Regex for classifying prostate reports "(?i)prostat" and regex for filtering out intra-procedural reports "(?i)biopsy" both failed in this case, as report contains the word "prostate" and also does not include the word "biopsy" in the header, but still is the wrong report. Error was manually corrected by picking the next prostate related pre-procedural radiology report.

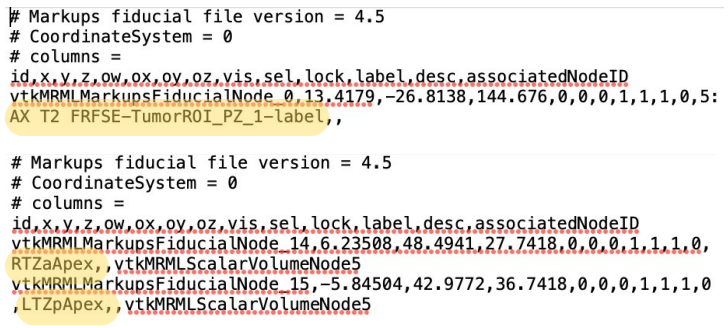


Figure 5.7: Correctly matched pre-procedural target .fcsv files. Left: anatomical region was not correctly inputted by Slicer operator during prostate biopsy, having as consequence that no clear link between histopathology and pre-procedural diagnostics can be created and target needs to be discarded, Right: correctly labeled biopsy target with anatomical zone abbreviation that can be matched "semantically" to single diagnosis sections from radiology and pathology reports.

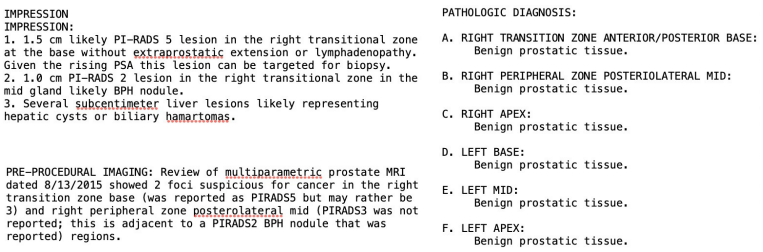


Figure 5.8: Inter-reader disagreement between Radiologists in assessing severity and location of lesion on pre-procedural MRI. Left upper text snippet: pre-procedural impression section given by first Radiologist before biopsy procedure, Left lower text snippet: intra-procedural report, written by interventional radiologist during biopsy procedure, right: resulting pathology diagnosis from extracted tissue during biopsy (corresponds to anatomical regions given by interventional radiologist)

FINDINGS:
PROSTATE GLAND SIZE: 3.4 x 5.0 x 4.8 cm, volume 42 mL.

FOCAL LESION(S): Stable 5 mm focal lesion in the peripheral zone, located in right, apex, posterolateral region (PZPL) abutting the capsule, previously designated as a PI-RADS 4 lesion. MRI guided biopsy of this yielded Gleason 6 prostate cancer. A tiny new focus of hemorrhage in the right posterolateral peripheral zone near the apex is likely related to the biopsy.

No new lesions in the prostate.

TRANSITIONAL ZONE: Changes of glandular and stromal hyperplasia.

PERIPHERAL ZONE: Normal.

EXTRA-PROSTATIC EXTENSION: None

NEUROVASCULAR BUNDLE: Normal.

SEMINAL VESICLES: Normal.

URINARY BLADDER: Normal.

PELVIC LYMPH NODES: Small external iliac lymph nodes are of doubtful clinical significance.

RETROPERITONEUM: .

BONES: No suspicious lesions.

IMPRESSION
IMPRESSION:

1. No convincing evidence of clinically significant prostate cancer.
2. Stable previously designated PI-RAD 4 lesion in the right apex peripheral zone, biopsied under MRI guidance in the interim yielding Gleason 6 (3+3) disease. No new lesions in the prostate.

Figure 5.9: The estimated lesion size is missing from the IMPRESSION section of the pre-proceddural radiology report but mentioned in the FINDINGS section.

```
IMPRESSION
1. Two PI-RADS 4 lesions in the mid gland peripheral zone, one on
the right and the left. Neither of these has changed since MRI
January 20, 2015.
2. Changes of BPH in the central gland. Prostate volume 87 ml.
3. Bilateral renal cysts.
```

Figure 5.10: Radiology impression section can deviate substantially from standard form of having one lesion observation per bullet point, including a dedicated PI-RADS score and estimated lesion size. In this case one bullet contains two lesion descriptions without containing the lesion size (reference to an earlier report).

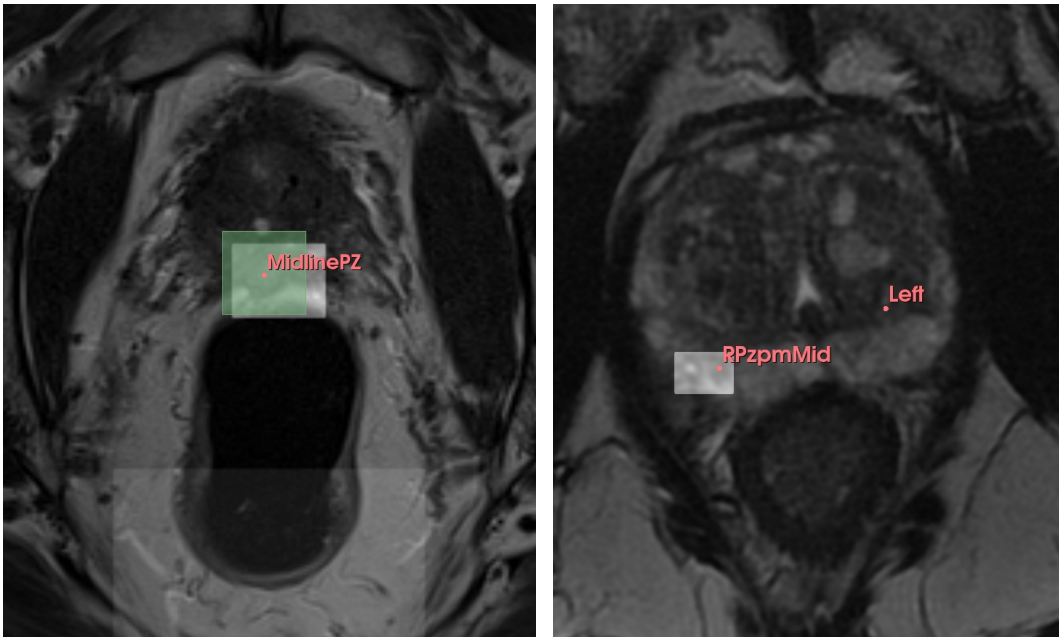


Figure 5.11: Two randomly drawn example MRI exams (visualized in Slicer) of the BWH-MRIGPB dataset with predicted csPCa lesion bounding boxes (visualized as white bounding boxes). Predictions were generated by the semi-supervised nnDetection framework, provided as one of three baseline models in the frame of the PI-CAI challenge, and pre-trained on the "Public Training and Development" dataset. On the left, in addition to the marked pre-procedural biopsy target coordinate (labeled "MidlinePZ" for its location in the Midline Peripheral Zone), a lesion bounding box can be used as ground truth to evaluate the position and approximate size of the predicted lesion bouding box by the nnDetection model. On the right, the point annotation (labeled RPZpmMid, for Right Peripheral Zone posterior medial Mid), can at least validate the approximate position of the predicted lesion bounding box, for example by verifying if the point annotation is located within the predicted bounding box.

6 Conclusion

In the frame of this work, a new prostate image dataset, comprising T2-weighted images, diffusion-weighted (b-value >= 1400) images, and apparent diffusion coefficient maps was presented. Two types of annotations - target biopsy coordinates and lesion bounding boxes - were presented in chapter 3 and automatically generated as described in chapter 4. 13 unique clinical variables were curated together with detailed information about image acquisition parameters. A detailed comparison between the BWH-MRIGPB dataset and three existing, publicly-available prostate cancer image datasets was conducted in chapter 3. The BWH-MRIGPB dataset was explicitly curated for the newly introduced task of zonal detection of clinically-significant prostate cancer lesions on biparametric MRI. Finally, a quality assessement of both curated images and annotations was conducted and discussed in chapter 5.

A detailed dataset curation framework was presented for the curation of the BWH-MRIGPB dataset. A complete process overview of the automatic structuring and labeling of locally-available raw data modalities is given in chapter 4. A brief quantitative performance review of the applied methods is conducted in chapter 4 and discussed in chapter 5.

6.0.1 Future work

For the BWH-MRIGPB dataset the following future work is envisioned:

- **image quality assessment by expert:** an expert radiologist should validate, that all automatically curated images allow for complete analysis and interpretation of usual pre-procedural diagnostics, chapter 1
- **annotation quality and "clinical usefulness" assessment by expert:** first the "clinical usefulness" of the algorithms that can conceptionally be developed with the BWH-MRIGPB dataset should be validated by discussion with an interventional radiologist. If the "clinical usefulness" is given, the quality of the automatically derived labels (especially lesion bounding boxes) could be benchmarked against human-derived bounding boxes or even lesion segmentations.

- **Making BWH-MRIGPB dataset publicly available:** after the quality assessment is completed, the dataset can be submitted to public repositories, such as TCIA and zenodo.
- **Model benchmarking and development:** a first straight forward model use-case would be to apply all pre-trained baseline model from the PI-CAI challenge and benchmark their performance on the BWH-MRIGPB dataset in comparison to other publicly-available datasets, like for example Prostate158.

For the curation pipeline presented in this work, a number of limitations and disadvantages have been described in 5. Future work, could focus on the following:

- Further reducing the extraction and matching error discussed in chapter 5 for both "syntactic" and "semantic" methods.
- Publishing the de-identified code on GitHub, to get feedback on "generallizability" of the developed dataset curation framework and further understand conceptual and method limitations.
- In the mid- to longterm the "automated" nature of the curation pipeline could be tested to "continuously grow" the BWH-MRIGPB dataset with every new MRI-guided prostate biopsy case at the Brigham and Women's Hospital with minimal effort. Especially, the first "data aggregation" step of the curation pipeline that still has to be done manually, would need to be evaluated further.

7 Appendix

Path report (.txt)

```
Report Status:  Final
Type:  Surgical Pathology
Pathology Report: [redacted]

CASE: [redacted]
PATIENT: [redacted]
Birth Date: [redacted]
Sex: [redacted]

Brigham and Women's Hospital
Department of Pathology
75 Francis Street, Boston, MA 02115

CLIA License No.: [redacted]
Laboratory Director: [redacted]

Physician: [redacted]
Procedure Date: [redacted]

Pathologist: [redacted]

PATHOLOGIC DIAGNOSIS:

A. LEFT PERIPHERAL ZONE ANTERIOR APEX LESION:
      PROSTATIC ADENOCARCINOMA, Gleason score 3+3=6 (Grade Group 1),
        involving 70% of the total tissue.
      No perineural invasion.

B. LEFT APEX:
      Benign prostatic tissue.

C. LEFT MID:
      Benign prostatic tissue.

D. LEFT BASE:
      Benign prostatic tissue.

E. RIGHT APEX:
      PROSTATIC ADENOCARCINOMA, Gleason score 3+3=6 (Grade Group 1),
        involving 10% of one (1) fragmented core.
      No perineural invasion.

F. RIGHT MID:
      Benign prostatic tissue.

G. RIGHT BASE:
      Benign prostatic tissue.

CLINICAL DATA:
History: None provided.
Operation: None provided.
Operative Findings: None provided.
Clinical Diagnosis: Elevated PSA.

TISSUE SUBMITTED:
A/1. Left peripheral zone anterior apex lesion
B/2. Left apex
C/3. Left mid
D/4. Left base
E/5. Right apex
F/6. Right mid
G/7. Right base
```

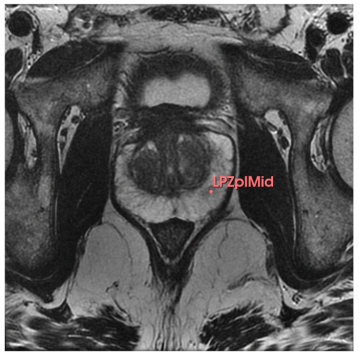
Rad report (.txt)

```
BRRADMR.AB.PROSTATE|MRI PROSTATE WITH AND WITHOUT CONTRAST

TECHNIQUE: Multiplanar MR imaging of the pelvis was performed using T1, T2, fat saturated, and diffusion
weighted techniques. Dynamic multiphase imaging was also performed after administration of an intravenous
gadolinium contrast agent.

COMPARISON: MRI PROSTATE WITH AND WITHOUT CONTRAST [redacted]; MRI PROSTATE WITH AND WITHOUT CONTRAST [redacted]

Additional Information:
PSA: 10.34 ng/ml
PSA density: 0.19 ng/ml/cc
Biopsy Date: [redacted]
Biopsy results: Gleason pattern 3 + 3, left peripheral zone posterior medial mid
Prior treatment or Active Surveillance: Active surveillance.

FINDINGS:
Prostate Gland Size: 4.1 x 5.0 x 5.2 cm
Prostate Volume: 56 mL

PERIPHERAL ZONE:
Patchy areas of low T2 weighted signal. Symmetric area of T2 hypointensity within the bilateral
posteromedial regions (4:13), unchanged compared to prior examinations and likely represents the prostate
central zone.

Focal Lesion(s):
1. There is a 0.9 cm focal lesion (4 : 19) in the left peripheral zone, located at the mid, posterolateral
region (PZpl). The lesion is focally markedly hypointense on ADC and markedly hyperintense on high b-value
DWI; < 1.5 cm in greatest dimension, dynamic contrast enhancement is positive, and is circumscribed,
homogenous moderate hypointense focus/mass confined to prostate and < 1.5 cm in greatest dimension on T2WI.
There is no evidence of extraprostatic extension. The lesion is PI-RADS 4.

2. There is a 0.9 cm focal lesion (4 : 19) in the right peripheral zone, located at the mid, anterior region
(PZa). The lesion is focally markedly hypointense on ADC and markedly hyperintense on high b-value DWI; <
1.5 cm in greatest dimension, dynamic contrast enhancement is negative, and is circumscribed, homogenous
moderate hypointense focus/mass confined to prostate and < 1.5 cm in greatest dimension on T2WI. There is no
evidence of extraprostatic extension. The lesion is PI-RADS 4.

TRANSITION ZONE:
Changes of glandular and stromal hyperplasia (BPH).

The Membranous Urethra Length (MUL) is: 0.8 cm

Seminal Vesicles: Normal.

Bladder: Trabeculated.

Rectum: Normal. No wall thickening.

Lymph Nodes: Normal, no pelvic lymphadenopathy.

Vessels: Normal.

Bones/Soft Tissues: No destructive osseous lesions. Unchanged T1 hypointense lesions in the left iliac bone
(11:119), likely benign

IMPRESSION:
Since [redacted]
1.  0.9 cm PI-RADS 4 lesion in the left mid peripheral zone, anterior region, slightly increased in size.
2.  0.9 cm PI-RADS 4 lesion in the right mid peripheral zone, anterior region, slightly increased in size.
3.  No lymphadenopathy or extraprostatic extension.
```

Biopsy target coordinates (.fcsv)

```
# Markups fiducial file version = 4.7
# CoordinateSystem = 0
# columns = id,x,y,z,ow,ox,oy,oz,vis,sel,lock,label,desc,associatedNodeID
vtkMRMLMarkupsFiducialNode_0,10.1031,-26.0372,135.82,0,0,0,1,1,1,0,LPZplMid,,vtkMRMLScalarVolumeNode1
vtkMRMLMarkupsFiducialNode_1,41.6779,-25.1796,143.032,0,0,0,1,1,1,0,Right,,vtkMRMLScalarVolumeNode1
```

Diagnostic MRI (DICOM)



Regional anatomical information

Figure 7.1: 4 modalities as raw data input

Dataset	PI-CAI Public Training and Development Dataset	SPIE-AAPM-NCI PROSTATEx Challenges	prostate158	BWH-MRIGPB
Purpose	Cancer detection, disease classification	Cancer detection, disease classification	Cancer detection, zones segmentation	Cancer detection, disease classification
Data source	RUMC, UMCG, ZGT	RUMC	CUB	BWH
Acquisition years	2012-2021	2012	-	2015-2018
Publication year	2022	2017	2022	TBD
Number of patients	1476	346	-	84
Number of studies	1500	349	158	86
Benign or indolent Pca cases	1075 (71.7%)	-	-	61 (70.9%)
csPCa (ISUP>=2) cases	425 (28.3%)	-	-	25 (29.1%)
Median Age (years)	67 {IQR: 61-71}	-	-	66 {IQR: 61-70}
Median PSA (ng/mL)	8.5 {IQR: 6-13}	-	-	7.5 {IQR: 4.83-10.75}
Median Prostate Volume (mL)	57{IQR: 40-80}	-	-	52.35 {IQR: 36.18-79.5}
Number of positive MRI lesions	1087 [PI-RADS 3 - 246 (23%), PI-RADS 4 - 438 (40%), PI-RADS 5 - 403 (37%)]	-	-	93 [PI-RADS 3 - 22 (21%), PI-RADS 4 - 51 (48.6%), PI-RADS 5 - 20 (19%)]*
Number of ISUP-based lesions	775 [ISUP 1 - 310 (40%), ISUP 2 - 260 (34%), ISUP 3 - 109 (14%), ISUP 4 - 41 (5%), ISUP 5 - 55 (7%)]	-	-	47 [ISUP 1 - 21 (44.7%), ISUP 2 - 15 (31.9%), ISUP 3 - 4 (8.5%), ISUP 4 - 3 (6.4%), ISUP 5 - 4 (8.5%)]
Train/Test split	1500/100 (7,500)	-	139/19	TBD
MRI sequences	axial, sagittal and coronal T2W, axial high b-value DWI, axial ADC	axial, sagittal and coronal T2W, DWI, ADC, DCE, PDW	axial T2W, axial DWI, axial ADC	axial T2W, axial high b-value DWI, axial ADC
Scanner type and model	Siemens n=1,221 [Skyra 3T, TrioTim 3T, Aera 1.5T, Avanto 1.5T, Espree 1.5T], Phillips n=279 [Ingenia 3T, Achieva 1.5T, Intera 1.5]	Siemens Trio (n=57), Siemens Skyra (n=289)	-	Siemens n=8 [Prisma, Verio, Aera], GE n=58 [Discovery MR750w]
Magnetic field strength	3T (n=1,418), 1.5T (n=82)	3T	3T	3T (n=84), 1.5T (n=2)
Coil type	Surface	Surface	Surface	Surface (n=31), Endorectal (n=55)
Data format	.mha (scans), .nii.gz (annotations)	.dcm (scans), .mhd (ktrans images), bmp (thumbnail images)	.nii.gz (scans and annotations)	.dcm (scans and annotations), .nii.gz (scans and annotations)
Files size	32.5GB	15.4GB	2.6GB	~1GB
Histopathologic confirmation	YES	YES	YES	YES
expert-derived lesion delineations	YES	NO	YES	NO
biopsy location coordinates	NO	YES	NO	YES
annotations	lesion segmentation (n=1,295)	-	zone segmentations (n=139), lesion segmentations (n=83)	lesion bounding boxes (n=99)
clinical metadata	age (n=1,500), psa (n=1,460), psa density (n=1,047), gleason grade group (n=1,001), prostate volume (n=1,473), histopathology type (n=1,001)	lesion location (n=344), significant/insignificant cancer (n=204), gleason grade group (n=99), number of lesions per patient (n=204)	-	age (n=86), race (n=86), psa (n=102), gleason grade group (n=105), gleason score (n=105), prostate volume (n=104), histopathology type (n=105), lesion location (n=105), pi-rads (n=93), recent prior biopsies (n=73), prior therapy (n=18), anatomical region (n=105), lesion size (n=99)
Data hosted on (accessible via)	zenodo (scans), github (annotations)	TCIA	zenodo (scans), github (annotations)	TBD

Figure 7.2: Prostate cancer public datasets comparison.

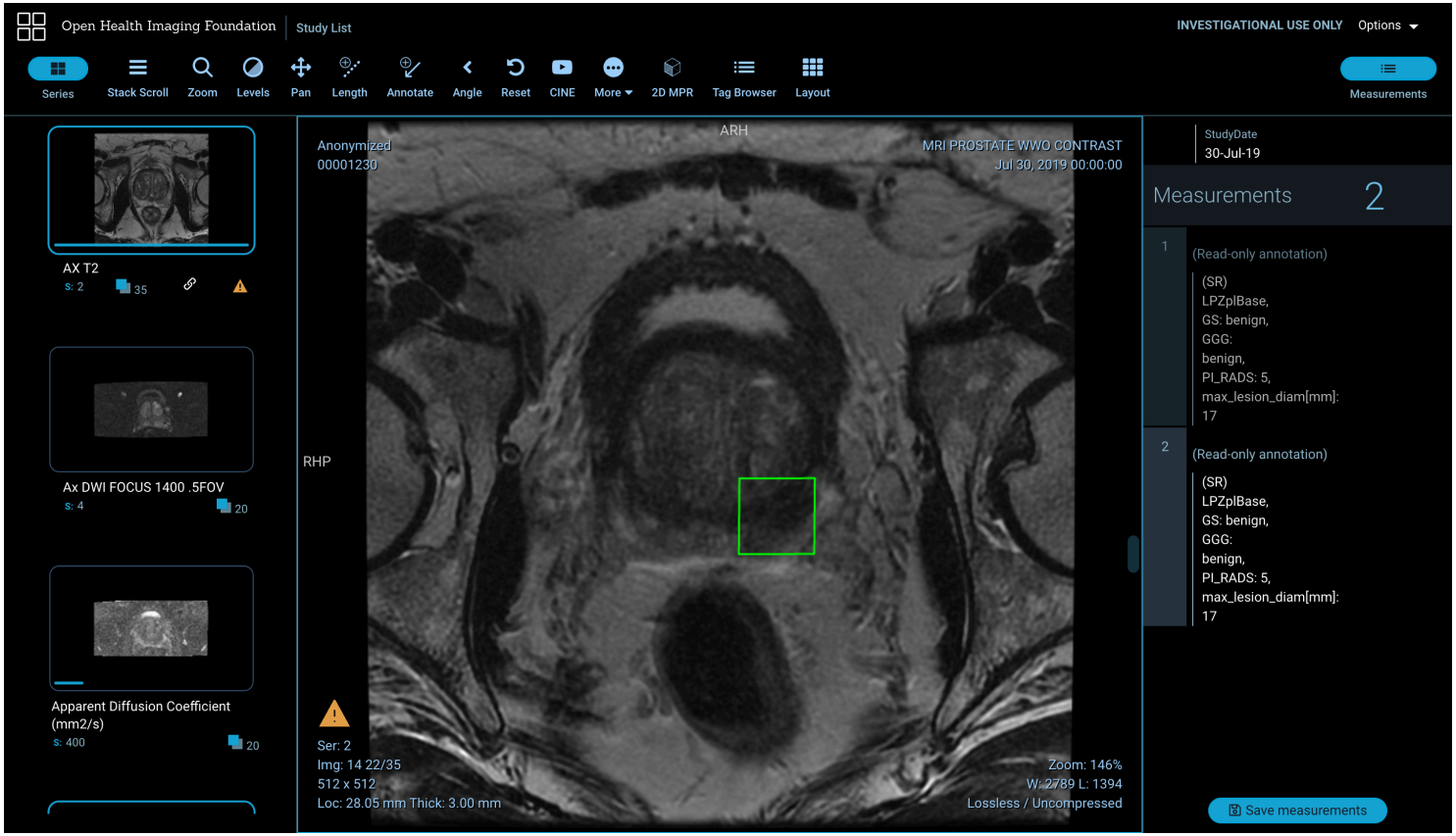


Figure 7.3: Patient 1230, study 3960: A full view of the lesion bounding box in the web-based OHIF viewer, constructed around the biopsy coordinate with the lesion diameter given by the expert radiologist as estimate. On the right panel, the anatomical region, post-procedurally assigned Gleason Grade Group, and PI-RADS score with estimated lesion diameter, given pre-procedurally by an expert radiology, are visualized. A dark region around the biopsy arrow, indicates potentially the prostate cancer lesion. Note: The study date visible on the top right does not correspond to the original study date, but is de-identified (the same applies to patient and study id).

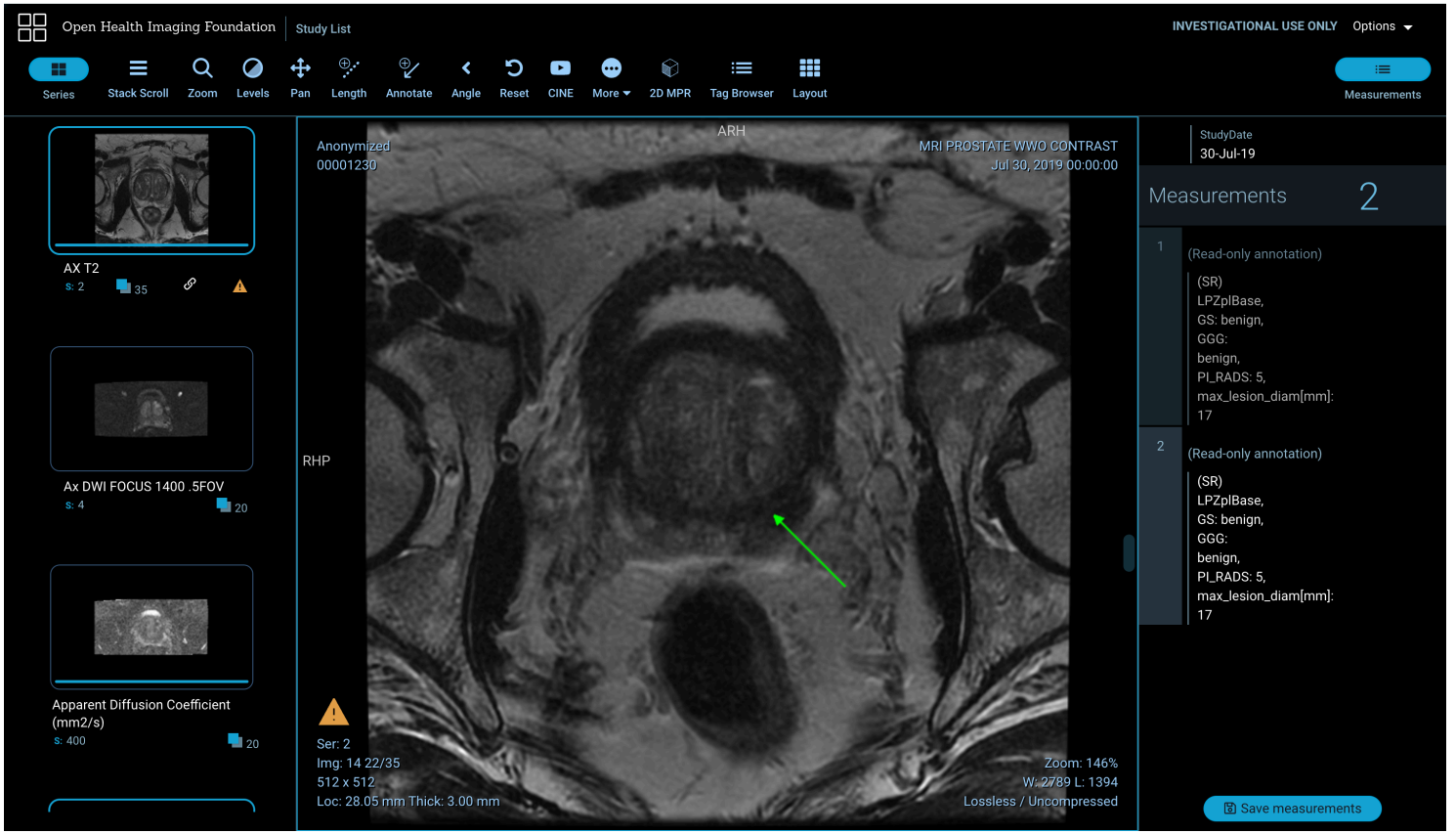


Figure 7.4: Patient 1230, study 3960: A full view of the biopsy point coordinate visualized in the web-based OHIF viewer. On the right panel, the anatomical region, post-procedurally assigned Gleason Grade Group, and PI-RADS score with estimated lesion diameter, given pre-procedurally by an expert radiology, are visualized. A dark region around the biopsy arrow, indicates potentially the prostate cancer lesion. Note: The study date visible on the top right does not correspond to the original study date, but is de-identified (the same applies to patient and study id).

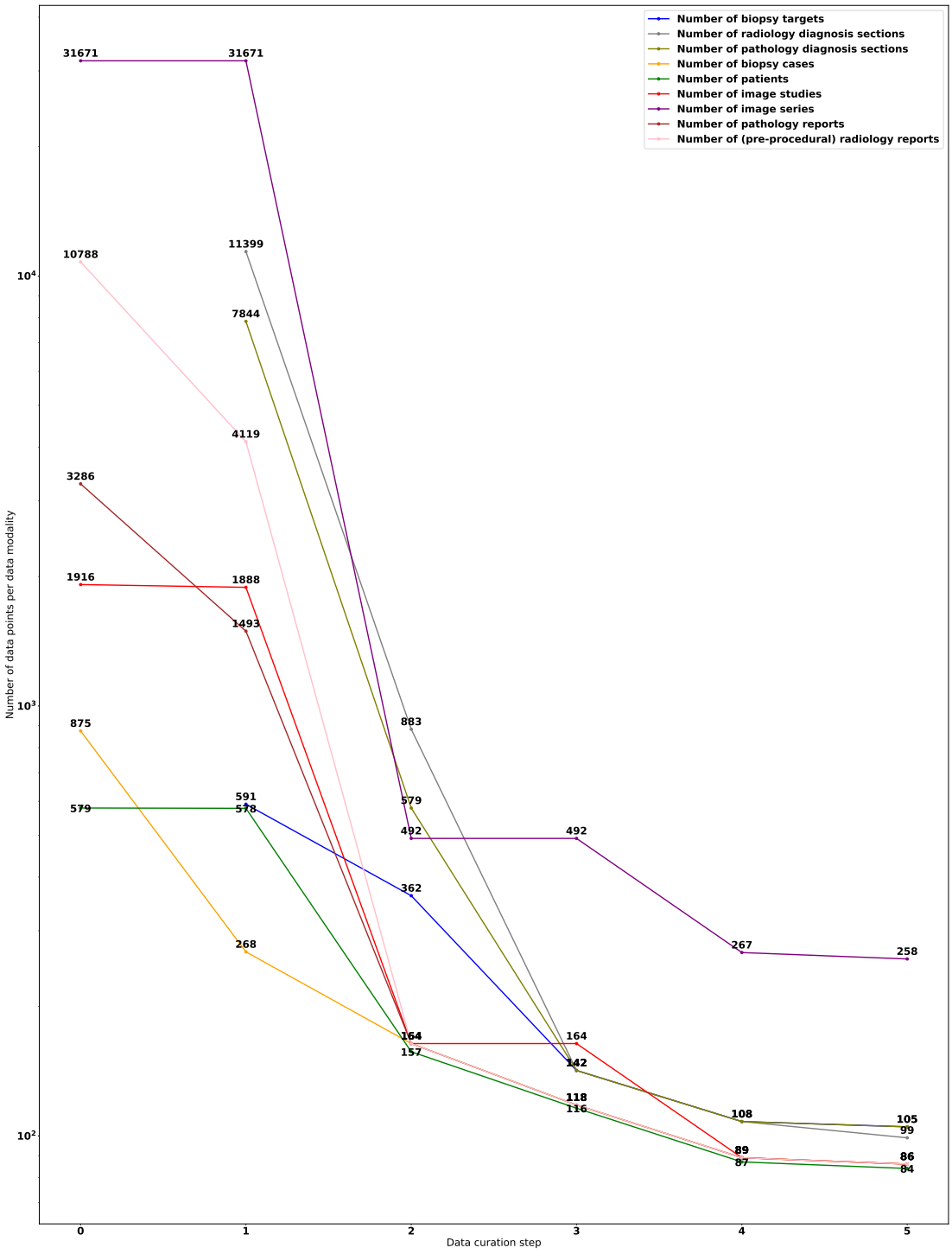


Figure 7.5: Using an automated "code-based" curation approach the data reduction after every step and for all modalities can be visualized and used for debugging.

Task Category	Task	Task Type	Runtime	Performance**	Phase*	Expression(s) used	Main function in code	Python package(s)
syntactic	aggregating image studies, clinical reports and biopsy cases	data transferring	TBD	TBD	0	n/a	n/a	gsutil, s5cmd
syntactic	split reports	regex pattern matching	TBD	TBD	**1**	r'\[report_end\]'	main	re, pandas
syntactic	filter for prostate reports	regex pattern matching	TBD	TBD	1	r'(?i)prostat'	process_txt_report	re, pandas
syntactic	extract insitution from report header	string comparison	TBD	TBD	1	MGH', 'BWH'	process_txt_report	stdlib
syntactic	extract mrn from report header	string partitioning	TBD	TBD	1	'| ' and ' | H'	process_txt_report	stdlib
syntactic	extract an from report header	string partitioning	TBD	TBD	1	| '	process_txt_report	stdlib
syntactic	extract ad from report header	regex pattern matching	TBD	TBD	1	r'(\d{1,2}/\d{1,2}/\d{2,4})'	process_txt_report	re, datetime, pandas
syntactic	filter for reports >= 01/01/2005	string comparison	TBD	TBD	1		process_txt_report	datetime, pandas
syntactic	extract diagnosis section (path)	regex pattern matching	TBD	TBD	1	DIAGNOSIS', 'CLINICAL', 'newlinenewline'	process_diagnosis_section	re, pandas
syntactic	extract procedure date (path)	regex pattern matching	TBD	TBD	1	r'(\d{1,2}/\d{1,2}/\d{2,4})', r'(?i)(Procedure Date:|Date of Operation:|Date Taken:|Date of Procedure:)\s+(\d{1,2}/\d{1,2}/\d{2,4})'	process_diagnosis_section	re, pandas
syntactic	extract impression section (rad)	regex pattern matching	TBD	TBD	1	IMPRESSION', 'CLINICAL', 'newline'	process_impression_section	
syntactic	extract diagnosis section (rad)	regex pattern matching	TBD	TBD	1	r"(NOTE:newlinePI-RADS \(Prostate Imaging Reporting and Data System\)|newlinePI-RADS 1:)", r"(?<!PI-RADS\s)(?<=\b)\d\.\s"	process_impression_section	re, pandas
syntactic	extract all fcsv files in directory	string comparison	TBD	TBD	1	filepath + '.fcsv'	process_slicer_dir	os
syntactic	extract all intraop dicom files	string comparison	TBD	TBD	1	filepath + 'DICOM', 'Intraop', 'Case'	process_slicer_dir	os
syntactic	extract biopsy case number	string partitioning	TBD	TBD	1	filepath + '/Case'	process_slicer_dir	pandas
syntactic	extract biopsy case date	string partitioning	TBD	TBD	1		process_slicer_dir	datetime, pandas
syntactic	extract patient MRN in case	reading dcm object	TBD	TBD	1	dicom_file.PatientID	process_slicer_dir	pydicom
syntactic	filter for only biopsy targets related to a preprocedural imaging study	regex pattern matching	TBD	TBD	1	r'(?i)pre'	process_slicer_dir	re, pandas
syntactic	remove all columns with no date extracted	string comparison	TBD	TBD	1	'No date!'	process_slicer_dir	pandas
syntactic	merge dfs to get MRN (extracted from intraop dcm file) for every fcsv file	dataframe merge	TBD	TBD	1	on=['Case_Number', 'Date']	process_slicer_dir	pandas
syntactic	merge df_rp (processed rad reports) and df_cvbq (identified image studies), df_rcvq	dataframe merge	TBD	TBD	**2**	left_on=['ORIG_MRN', 'ORIG_StudyDate'], right_on=['MRN', 'Accession_Date']	patient_level_matching	pandas
syntactic	create all preprocedural candidates (filter out all intraprocedural studies)	regex pattern matching	TBD	TBD	2	r"(?i)biopsy"	filter_out_ips	re, pandas
syntactic	merge df_rp, df_pp and df_s to df_spr to create all intraprocedural candidates	dataframe merge	TBD	TBD	2	left_on=['Procedure_Date', 'MRN'], right_on=['Date', 'MRN'], left_on=['Accession_Date', 'MRN'], right_on=['Date', 'MRN'],	patient_level_matching	re, pandas
syntactic	df_ppip, match up intraprocedural study with most recent preprocedural study, storing MRNs, ANs, ADs	dataframe creation	TBD	TBD	2	groupby('MRN'), pp_rows['Accession_Date'].idxmax()	patient_ip_pp	pandas
syntactic	drop unconclusive first biopsies in case of repeat biopsies (2 ips 1 pp)	dataframe operation	TBD	TBD	2	sort_values(by=['Accession_Date_pp'], ascending=False), drop_duplicates(subset=['Accession_Number_pp'], keep='first')	patient_level_matching	pandas
semantic	**extract required image modalities (AXT2, HBV, ADC) from every image study, based on Seriesdescription and store in df_images**	**semantic similarity matching**	**TBD**	**TBD**	**3**	**'emilyalsentzer/Bio_ClinicalBERT'**	**t2w_hbv_adc_extraction**	**transformers, pandas**

Figure 7.6: Task-Matching-Table, p. 1

Legend: * refers to phase in dataset curation cycle (see figure 4.2), ** see performance metric used in table (metrics for performance evaluation of method-task type), MRN = medical record number, AN = accession number, AD = accession date

Task Category	Task	Task Type	Runtime	Performance**	Phase*	Expression(s) used	Main function in code	Python package(s)
syntactic	select all biopsy cases with target coordinates and anatomical region description	reading structured csv file	TBD	TBD	3	ar_mapping = {'RPZplMid': 'RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID' , ...}	lesion_level_matching	pandas
semantic	**find correct pathology diagnosis section to biopsy target based on anatomical prostate region**	**semantic similarity matching**	**TBD**	**TBD**	**3**	**'emilyalsentzer/Bio_ClinicalBERT'**	**lesion_level_matching**	**transformers, pandas**
semantic	**find correct radiology single diagnosis to biopsy target based on anatomical prostate region**	**0-shot text classification**	**TBD**	**TBD**	**3**	**'facebook/bart-large-mnli'**	**lesion_level_matching**	**transformers, pandas**
syntactic	store rows of all matched pathology diagnosis sections and radiology single diagnoses in df_annotations dataframe	dataframe creation	TBD	TBD	3		lesion_level_matching	pandas
semantic	**extract pirads score from matched radiology single diagnosis**	**0-shot text classification**	**TBD**	**TBD**	**4**	**'facebook/bart-large-mnli'**	**pp_diagnostics_extraction**	**transformers, pandas**
semantic	**extract lesion diameter from matched radiology single diagnosis**	**0-shot text classification**	**TBD**	**TBD**	**4**	**'facebook/bart-large-mnli'**	**pp_diagnostics_extraction**	**transformers, pandas**
syntactic	extract gleason score from matched pathology diagnosis section	regex pattern matching	TBD	TBD	4	r'(\D|^)(\d+\s*\+\s*\d+\s*=\s*\d+)(?=\D|$)', r'(?i)benign'	ip_diagnostics_extraction	re, pandas
syntactic	extract grade group from matched pathology diagnosis section	regex pattern matching	TBD	TBD	4	r'(?i)Grade\s*Group\s*\d', r'(?i)benign'	ip_diagnostics_extraction	re, pandas
manual inspection	**manually verify correctness of rows in df_images and df_annotations**	**numerical validation**	**TBD**	**n/a**	**5**	**n/a**	**n/a**	**n/a**
syntactic	turn pirads score into single number	regex pattern matching	TBD	TBD	6	r'(\d+\.?\d*)'	convert_pirads	re, pandas
syntactic	turn lesion size into single number [mm]	regex pattern matching	TBD	TBD	6	r'(\d+\.?\d*)', 'cm', 'mm'	convert_lesion	re, pandas
syntactic	select all studies from final dfs (df_images, df_annotations) that have at least an identified biopsy target with matching histopathology	string comparison	TBD	TBD	7	'no_match'	dfs_cleaning	pandas
syntactic	processing for visualization of images and annotations in the OHIF webviewer for easy inspection	regex pattern matching, point projection	TBD	TBD	7	r"\ ?(-?[0-9]*\.[0-9]*)\ +(-?[0-9]*\.[0-9]*)\ +(-?[0-9]*\.[0-9]*)", DICOM attributes: ImageOrientationPatient, ImagePositionPatient	create_ohif_labels	pydicom, re, pandas, numpy
syntactic	export image studies after cleaning as DICOM, MHA, NIFTI	data conversion and export	TBD	TBD	7	DICOM attributes: DiffusionBValue, SequenceName, ImagesInAcquisition, InstanceNumber, etc.	images_export	pydicom, os, pandas, re, picai_prep
syntactic	export annotations as points, and constructed bounding boxes as DICOM Structured Reports and NIFIT segmentation objects	data conversion and export	TBD	TBD	7	Comprehensive3DSR(), PlanarROIMeasurementsAndQualitativeEvaluations(), etc.	annotations_export	numpy, nibabel, scipy, os, highdicom, pydicom, re, pandas
manual inspection	**manually look at all images and annotations in OHIF webviewer**	**visual validation**	**TBD**	**n/a**	**5**	**n/a**	**n/a**	**n/a**

Figure 7.7: Task-Matching-Table, p. 2

Legend: * refers to phase in dataset curation cycle (see figure 4.2), ** see performance metric used in table (metrics for performance evaluation of method-task type), MRN = medical record number, AN = accession number, AD = accession date

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	pca3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cspca	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
653	37654	51	n/a	n/a	6	181	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	NO	NO	48.569700 -31.992100 -38.743800	RPZplApex	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL APEX	n/a	n/a	benign	benign
663	4081	65	white	non hispanic	7.96	91	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-67.32397	RPzplApex	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL APEX	4	7	benign	benign
670	4415	76	white	non hispanic	8.82	61.9	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-19.643100 -41.957700 82.611100	LPZpmMid	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL MID	4	6	3+3=6	Grade Group 1
670	4415	76	white	non hispanic	8.82	61.9	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-18.701000 -43.574700 75.811200	LPZpmApex	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL APEX	4	6	benign	benign
829	4478	63	white	non hispanic	4.75	33	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-26.481200 3.355030 -53.076000	LPZplBase	LEFT PERIPHERAL ZONE POSTERIOR LATERAL BASE	4	4	benign	benign
829	4478	63	white	non hispanic	4.75	33	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-89.34436	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	3	3	benign	benign
941	4119	67	white	n/a	15.77	66.3	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-22.965648	RTZaBase	RIGHT TRANSITION ZONE ANTERIOR BASE	4	14	benign	benign
1157	4245	62	white	non hispanic	7.88	86	3+4	n/a	n/a	Not reported as deceased	MRBx	YES	NO	NO	-20.651200 -32.602800 131.409000	LPZpmMid	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL MID	n/a	n/a	benign	benign
1211	4510	69	white	non hispanic	10.2	73.5	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	27.278500 -31.262600 -65.308800	RPZplApex	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL APEX	4	13	benign	benign
1217	4204	58	white	non hispanic	1.22	19	3+4	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-2.050500 -60.050700 18.717100	LPZplBase	LEFT PERIPHERAL ZONE POSTERIOR LATERAL BASE	4	9	3+3=6	Grade group 1

Figure 7.8: Clinical Metadata, p. 1

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	pca3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cspca	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
1230	3960	60	white	non hispanic	21.5	91	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-18.217500 -29.719100 27.922700	LPZplBase	LEFT PERIPHERAL ZONE POSTERIOR LATERAL BASE	5	17	benign	benign
1230	3232	60	white	non hispanic	17.3	101	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-34.186100 -38.184200 34.459600	LPZpBase	LEFT PERIPHERAL ZONE POSTERIOR BASE	5	18	benign	benign
1265	3996	77	n/a	n/a	15.4	76	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	23.161500 - 22.005600 42.281600	RTZpMid	RIGHT TRANSITION ZONE POSTERIOR MID	4	14	3+3=6	Grade Group 1
1366	4573	64	white	non hispanic	5.05	72	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-20.267300 -22.674800 105.416000	LTZaMid	LEFT TRANSITION ZONE ANTERIOR MID	4	12	benign	benign
1366	4573	64	white	non hispanic	5.05	72	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	1.471200 - 35.777900 107.207000	RPZpmApex	RIGHT PERIPHERAL ZONE POSTERIOR MEDIAL APEX	4	6	3+3=6	Grade Group 1
1382	4422	57	white	non hispanic	10.7	71.5	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	50.883900 - 32.872600 34.140200	RPZplApex	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL APEX	3	15	benign	benign
1535	4359	57	white	non hispanic	3.4	43.2	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-3.107290 - 4.767820 29.371400	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	3	9	3+3=6	Grade Group 1
1695	4234	56	white	non hispanic	4.9	63	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-18.307300 -49.242600 47.178800	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	11	benign	benign
1706	4745	66	white	non hispanic	6.44	63	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	17.332100 - 35.607400 94.081700	RPZpmMid	RIGHT PERIPHERAL ZONE POSTERIOR MEDIAL MID	4	6	benign	benign
1716	4465	64	white	non hispanic	n/a	36.6	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	22.955300 1.848560 71.117500	LTZaMid	LEFT TRANSITION ZONE ANTERIOR MID	4	9	3+3=6	Grade Group 1

Figure 7.9: Clinical Metadata, p. 2

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	psa3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cohort	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
1768	4187	63	black	non hispanic	7.7	81.2	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-32.268000 -39.590500 38.907900	LPZpMBase	LEFT PERIPHERAL ZONE POSTERIOR LATERAL BASE	4	11	benign	benign
1779	4446	53	white	non hispanic	1.26	40	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-3.960450 - 44.201500 52.293100	LPZpMMid	LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	15	benign	benign
1787	3955	67	white	non hispanic	2.9	42	3+3	101	n/a	Not reported as deceased	MRBx	YES	YES	NO	16.639700 - 44.638800 - 82.350100	RPZpLApex	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL APEX	4	5	benign	benign
1787	4381	67	white	non hispanic	3.9	42.5	n/a	101	n/a	Not reported as deceased	MRBx	YES	YES	NO	0.477423 - 40.172400 - 32.679800	RPZpLMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	4	benign	benign
1888	37386	74	white	non hispanic	5.3	34	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-60.1925	LTZpMid	LEFT TRANSITION ZONE POSTERIOR MID	4	11	3+3=6	Grade Group 1
1918	4499	66	white	non hispanic	4.08	73	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	12.094700 - 37.831100 60.951500	RPZpmMid	RIGHT PERIPHERAL ZONE POSTERIOR MEDIAL MID	3	6	benign	benign
2010	3144	59	white	non hispanic	9.15	52.7	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	16.221000 - 38.289000 16.898000	RPZpLBase	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL BASE	4	15	3+4=7	Grade Group 2
2035	4564	62	white	n/a	9.99	87	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-4.436380 - 53.409600 9.623170	LPZpLMid	LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	11	benign	benign
2035	4564	62	white	n/a	9.99	87	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	34.117900 - 49.527400 9.536380	RPZpLMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	12	benign	benign
2063	4576	55	white	non hispanic	6.1	42	3+3	n/a	active surveillance	Not reported as deceased	MRBx	YES	YES	NO	-76.8219	LPZpLMid	LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	14	benign	benign

Figure 7.10: Clinical Metadata, p. 3

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	pca3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cspca	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
2141	37701	67	white	non hispanic	13.16	70	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	31.819300 - 21.727200 - 50.815000	RTZaMid	RIGHT TRANSITION ZONE ANTERIOR MID	n/a	17	benign	benign
2141	37701	67	white	non hispanic	13.16	70	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	8.218270 - 43.526600 - 36.828700	LPZpmBase	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL BASE	n/a	13	benign	benign
2143	4108	63	white	non hispanic	9.54	114	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	34.536200 - 43.166900 31.345600	RPZplBase	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL BASE	5	24	4+3=7	Grade Group 3
2171	4158	72	white	non hispanic	22.1	74	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-10.997700 -9.730180 74.688400	LTZaMid	LEFT TRANSITION ZONE ANTERIOR MID	4	13	3+3=6	Grade Group 1
2203	4195	58	white	non hispanic	7.33	81	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	27.158000 - 0.935067 37.177200	LTZaMid	LEFT TRANSITION ZONE ANTERIOR MID	4	16	benign	benign
2327	21507	62	white	non hispanic	7.12	30.7	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	NO	NO	20.074700 - 22.618300 130.089000	RTZaMid	RIGHT TRANSITION ZONE ANTERIOR MID	n/a	n/a	3+3=6	Grade Group 1
2437	3802	70	white	non hispanic	2.65	10.8	3+4	n/a	brachytherapy	Not reported as deceased	MRBx	YES	YES	YES	-6.354020 4.171990 142.281000	MidlinePZ	MIDLINE PERIPHERAL ZONE	n/a	8	3+4=7	Grade Group 2
2498	4516	70	white	non hispanic	11.1	211.7	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	30.421700 - 27.232300 92.533400	LTZpBase	LEFT TRANSITION ZONE POSTERIOR BASE	3	14	benign	benign
2532	4459	69	asian	non hispanic	2.7	13.9	3+3	n/a	radiation therapy	Not reported as deceased	MRBx	YES	YES	NO	-1.718270 - 38.183100 113.633000	RPZpmApex	RIGHT PERIPHERAL ZONE POSTERIOR MEDIAL APEX	3	9	benign	benign
2547	4634	53	white	non hispanic	5.44	36.6	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	10.797700 - 16.800500 - 43.353800	RPZaMid	RIGHT PERIPHERAL ZONE ANTERIOR MID	4	7	benign	benign
2550	4765	64	white	non hispanic	7.99	24	3+3	n/a	active surveillance	Not reported as deceased	MRBx	YES	YES	NO	38.922700 - 33.457100 41.948500	LPZIApex	LEFT PERIPHERAL ZONE LATERAL APEX	3	6	benign	benign

Figure 7.11: Clinical Metadata, p. 4

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	pca3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cspca	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
2724	4166	64	white	non hispanic	1.86	14	4+4	n/a	brachytherapy	Not reported as deceased	MRBx	YES	YES	NO	-2.030400 -17.967600 76.782400	AFSmid	ANTERIOR FIBROMUSCULAR STROMA MID	n/a	11	benign	benign
2901	3531	73	white	non hispanic	19.39	28.7	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	NO	YES	-43.319400 -23.600800 12.576200	LPZbase	LEFT PERIPHERAL ZONE BASE	n/a	n/a	3+4=7	Grade Group 2
2912	4227	56	white	non hispanic	8	41	3+3	n/a	active surveillance	Not reported as deceased	MRBx	YES	YES	YES	12.451300 -5.666770 32.461900	LTZaMid	LEFT TRANSITION ZONE ANTERIOR MID	5	22	3+4=7	Grade Group 2
2912	4227	56	white	non hispanic	8	41	3+3	n/a	active surveillance	Not reported as deceased	MRBx	YES	YES	NO	3.498950 -28.743900 25.230400	LPZpmApex	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL APEX	4	11	3+3=6	Grade Group 1
2928	3578	85	white	non hispanic	22.28	55	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	-11.335900 9.925480 -47.606000	RPZpApex	RIGHT PERIPHERAL ZONE POSTERIOR APEX	5	40	4+4=8	Grade Group 4
2947	3606	67	white	non hispanic	14.2	158	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-2.988990 -36.333900 124.724000	RTZaApex	RIGHT TRANSITION ZONE ANTERIOR APEX	3	19	benign	benign
2966	4549	68	white	non hispanic	3.29	65	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-1.718270 -38.183100 113.633000	LPZplApex	LEFT PERIPHERAL ZONE POSTERIOR LATERAL APEX	5	22	3+3=6	Grade Group 1
2983	3675	68	white	non hispanic	19.37	72	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-9.890590 -17.770600 20.778600	LPZpmApex	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL APEX	4	8	benign	benign
2991	4202	64	white	non hispanic	27.9	51	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	-6.456290 -18.415700 0.997143	LPZplApex	LEFT PERIPHERAL ZONE POSTERIOR LATERAL APEX	4	14	3+4=7	Grade Group 2
3012	3729	66	white	non hispanic	7.76	42	negative	n/a	radiation proctitis	Not reported as deceased	MRBx	YES	YES	YES	26.953800 -39.074200 -88.763200	RTZpMid	RIGHT TRANSITION ZONE POSTERIOR MID	5	20	4+5=9	Grade group 5
3024	4441	71	white	non hispanic	16.25	40	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-33.551200 3.647390 47.088100	LPZplMid	LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	15	benign	benign

Figure 7.12: Clinical Metadata, p. 5

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	pca3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cspca	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
3035	3772	68	white	non hispanic	19.19	40	negative	n/a	n/a	Date of Death reported from a Partners Hospital	MRBx	YES	YES	YES	-39.19187	LPZplMid	LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID	5	26	3+4=7	Grade Group 2
3071	4751	70	white	non hispanic	5.2	104	n/a	n/a	laser vaporization of the prostate for BPH	Not reported as deceased	MRBx	YES	YES	NO	14.943800 -34.738700 -59.465500	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	11	3+3=6	Grade Group 1
3071	4751	70	white	non hispanic	5.2	104	n/a	n/a	laser vaporization of the prostate for BPH	Not reported as deceased	MRBx	YES	YES	YES	-108.75183	LPZpmMid	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL MID	3	7	4+4=8	Grade Group 4
3071	4751	70	white	non hispanic	5.2	104	n/a	n/a	laser vaporization of the prostate for BPH	Not reported as deceased	MRBx	YES	YES	NO	-130.7636	LPZplMid	LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID	3	5	benign	benign
3104	21269	51	white	non hispanic	17.44	37	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	39.452300 -20.019500 64.800300	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	5	17	4+3=7	Grade Group 3
3105	3895	69	white	n/a	6.2	31	3+4	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	-58.709300 -10.058100 26.039800	LPZplApex	LEFT PERIPHERAL ZONE POSTERIOR LATERAL APEX	5	18	4+3=7	Grade Group 3
3121	4520	59	white	non hispanic	10.3	34.9	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	49.359600 -14.787500 86.550100	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	14	benign	benign
3139	3957	70	white	n/a	7.21	48.7	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	23.942900 -21.568200 76.285900	RPZaMid	RIGHT PERIPHERAL ZONE ANTERIOR MID	4	11	benign	benign
3148	3975	68	white	non hispanic	8.27	41	3+4	n/a	high intensity focus ultrasound	Not reported as deceased	MRBx	YES	YES	YES	24.191700 -42.874600 160.577000	RPZplBase	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL BASE	4	6	4+5=9	Grade Group 5
3153	3985	62	white	non hispanic	n/a	60.7	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	0.284682 5.244830 39.222600	LTZpBase	LEFT TRANSITION ZONE POSTERIOR BASE	4	13	3+3=6	Grade Group 1

Figure 7.13: Clinical Metadata, p. 6

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	pca3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cspca	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
3192	4065	62	white	non hispanic	2.4	22	3+3	n/a	brachytherapy	Not reported as deceased	MRBx	YES	YES	YES	19.255900 - 22.261600 17.071800	RTZaMid	RIGHT TRANSITION ZONE ANTERIOR MID	5	19	4+5=9	Grade Group 5
3197	4074	79	white	non hispanic	2.78	81	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	38.231300 - 9.653510 73.793600	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	5	6	4+3=7	Grade Group 3
3202	4087	68	asian	non hispanic	7.27	92	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-28.47016	RTZaMid	RIGHT TRANSITION ZONE ANTERIOR MID	5	25	3+3=6	Grade Group 1
3207	4094	54	black	non hispanic	4.1	31	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	4.699060 - 9.532520 - 62.294900	RTZaMid	RIGHT TRANSITION ZONE ANTERIOR MID	5	16	3+3=6	Grade Group 1
3207	4094	54	black	non hispanic	4.1	31	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	10.078500 - 26.961300 - 61.578100	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	4	benign	benign
3219	4115	77	white	n/a	7.5	71	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	11.607000 - 34.921300 53.301800	LPZplMid	LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	14	3+4=7	Grade Group 2
3219	4115	77	white	n/a	7.5	71	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	55.880200 - 31.048000 53.527400	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	5	3+4=7	Grade Group 2
3228	4130	81	white	non hispanic	13.78	27	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-158.5451	LPZplApex	LEFT PERIPHERAL ZONE POSTERIOR LATERAL APEX	4	10	benign	benign
3230	4132	56	white	non hispanic	3.49	46.2	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	35.375100 - 32.320400 32.075100	RPZpmApex	RIGHT PERIPHERAL ZONE POSTERIOR MEDIAL APEX	4	10	3+4=7	Grade Group 2
3234	4138	58	white	non hispanic	5.13	56.4	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-4.041520 0.492446 21.100000	LPZaApex	LEFT PERIPHERAL ZONE ANTERIOR APEX	3	8	benign	benign

Figure 7.14: Clinical Metadata, p. 7

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	pca3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cspca	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
3242	4149	59	white	non hispanic	4.1	79	n/a	n/a	transurethral microwave thermotherapy	Not reported as deceased	MRBx	YES	YES	NO	-75.4907	LPZpmApex	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL APEX	4	13	benign	benign
3243	4150	58	white	non hispanic	1.42	n/a	n/a	n/a	RP	Not reported as deceased	MRBx	YES	NO	YES	7.955710 -9.328740 87.572600	Nodule	NODULE	n/a	n/a	3+4=7	Grade Group 2
3246	4154	66	white	non hispanic	10.91	24	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	3.929410 -10.270300 75.600400	LTZaMid	LEFT TRANSITION ZONE ANTERIOR MID	5	23	4+5=9	Grade Group 5
3249	4161	73	black	non hispanic	6.1	33.4	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-110.55667	LPZpmMid	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL MID	5	35	benign	benign
3250	4163	75	white	non hispanic	5.9	99	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	28.757200 -40.513300 195.915000	LPZplApex	LEFT PERIPHERAL ZONE POSTERIOR LATERAL APEX	5	17	3+3=6	Grade Group 1
3262	4190	76	white	non hispanic	10.75	88	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	0.344011 -31.063900 78.889600	LPZplBase	LEFT PERIPHERAL ZONE POSTERIOR LATERAL BASE	4	17	benign	benign
3262	4190	76	white	non hispanic	10.75	88	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	45.301900 -34.877000 69.228200	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	3	9	benign	benign
3264	4194	77	white	non hispanic	8.87	44	3+4	n/a	brachytherapy	Not reported as deceased	MRBx	YES	YES	NO	-5.196720 -7.422940 38.650400	LTZaMid	LEFT TRANSITION ZONE ANTERIOR MID	n/a	15	benign	benign
3264	4194	77	white	non hispanic	8.87	44	3+4	n/a	brachytherapy	Not reported as deceased	MRBx	YES	YES	YES	-18.351400 -65.959900 98.920500	LeftLN	LEFT LATERAL NODULE	n/a	11	metastatic pca	metastatic pca
3268	4210	70	white	non hispanic	6.26	49.7	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-47.96918	LTZaMid	LEFT TRANSITION ZONE ANTERIOR MID	4	14	3+3=6	Grade Group 1
3269	4211	70	white	n/a	5.61	67	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	45.269200 -39.285500 15.844500	RTZpMid	RIGHT TRANSITION ZONE POSTERIOR MID	3	9	benign	benign

Figure 7.15: Clinical Metadata, p. 8

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	pca3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cspca	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
3280	4226	70	white	non hispanic	0.82	29	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	35.574700 -29.012900 18.115300	RPZplApex	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL APEX	4	11	3+3=6	Grade Group 1
3280	4226	70	white	non hispanic	0.82	29	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	10.294500 -28.501400 25.404500	LPZpmMid	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL MID	3	10	benign	benign
3287	4240	73	n/a	non hispanic	12.7	143	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-15.235100 -42.376400 108.606000	LPZplMid	LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	14	benign	benign
3403	4464	61	white	non hispanic	13.98	26	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	-8.103870 -24.071700 72.288000	RTZaApex	RIGHT TRANSITION ZONE ANTERIOR APEX	4	10	3+4=7	Grade Group 2
3403	4464	61	white	non hispanic	13.98	26	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-26.576100 -30.671000 85.626100	LPZplMid	LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID	3	9	benign	benign
3446	4560	61	black	n/a	38.3	110	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	31.427300 -12.677800 46.263500	RTZpMid	RIGHT TRANSITION ZONE POSTERIOR MID	3	10	3+4=7	Grade Group 2
3446	4560	61	black	n/a	38.3	110	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-0.046527 -17.603700 28.913300	LPZpmApex	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL APEX	3	8	benign	benign
3448	4563	70	asian	non hispanic	7.59	14	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	24.139100 -44.124400 -10.572100	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	5	15	3+4=7	Grade Group 2
3468	4599	73	black	non hispanic	9.9	63.3	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	-1.803370 -26.900000 5.383960	LTZaMidApex	LEFT TRANSITION ZONE ANTERIOR MID APEX	5	29	3+4=7	Grade Group 2
3473	4612	72	white	non hispanic	20.25	50	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	30.050500 -25.993700 -25.699300	LPZaMid	LEFT PERIPHERAL ZONE ANTERIOR MID	4	14	4+4=8	Grade Group 4

Figure 7.16: Clinical Metadata, p. 9

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	pca3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cspca	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
3473	4612	72	white	non hispanic	20.25	50	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	61.416300 -0.873613 -17.952000	RTZaBase	RIGHT TRANSITION ZONE ANTERIOR BASE	4	6	3+3=6	Grade Group 1
3476	4616	74	white	non hispanic	n/a	28	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	14.690300 -37.622700 -14.542900	LPZplMid	LEFT PERIPHERAL ZONE POSTERIOR LATERAL MID	4	11	benign	benign
3479	4622	66	white	non hispanic	9.26	52	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-18.762700 -28.256000 13.916400	LPZplBase	LEFT PERIPHERAL ZONE POSTERIOR LATERAL BASE	4	6	3+3=6	Grade Group 1
3479	4622	66	white	non hispanic	9.26	52	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	-20.648100 -27.187400 2.913760	LPZaMid	LEFT PERIPHERAL ZONE ANTERIOR MID	3	4	benign	benign
3563	4798	64	white	non hispanic	6.5	34	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	21.725500 -23.631100 33.140400	RTZaMid	RIGHT TRANSITION ZONE ANTERIOR MID	3	10	3+3=6	Grade Group 1
3592	4854	68	white	non hispanic	4.5	47.3	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	YES	14.229800 -35.845800 -51.540300	LPZpmApex	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL APEX	4	9	3+4=7	Grade Group 2
3592	4854	68	white	non hispanic	4.5	47.3	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	NO	NO	3.393940 -34.351300 -37.531400	LPZpmMid	LEFT PERIPHERAL ZONE POSTERIOR MEDIAL MID	n/a	n/a	benign	benign
3605	4876	53	white	non hispanic	5.58	84	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	6.597420 14.741500 71.989100	RTZBase	RIGHT TRANSITION ZONE BASE	5	15	benign	benign
3605	4876	53	white	non hispanic	5.58	84	negative	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	19.078900 7.427500 62.989100	RPZplMid	RIGHT PERIPHERAL ZONE POSTERIOR LATERAL MID	3	10	benign	benign
3624	4910	65	asian	non hispanic	2.57	86	n/a	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	20.333500 -39.970900 -2.934320	LCGmid	LEFT CENTRAL GLAND MID	3	13	benign	benign
12223	37671	60	n/a	non hispanic	2.77	60	3+3	n/a	n/a	Not reported as deceased	MRBx	YES	YES	NO	47.161100 -34.696700 87.454900	RPZpmBase	RIGHT PERIPHERAL ZONE POSTERIOR MEDIAL BASE	3	13	benign	benign

Figure 7.17: Clinical Metadata, p. 10

patient_id	study_id	patient_age	race	ethnicity	recent_psa	prostate_volume	recent_prior_biopsies	pca3	prior_therapy	vital_status	histopath_type	point_annotation	bbox_annotation	case_cspca	pos	a_region	a_description	pi_rads	lesion_size	lesion_gs	lesion_isup
12224	37674	60	white	non hispanic	4.8	32	3+3	n/a	active surveillance	Not reported as deceased	MRBx	YES	YES	NO	1.526530 - 27.396300 - 92.259600	LPZaMid	LEFT PERIPHERAL ZONE ANTERIOR MID	4	12	benign	benign

Figure 7.18: Clinical Metadata, p. 11

patient_id	study_id	scanner_manufacturer	scanner_model	magnetic_field_strength	endorectal_coil	b_value_dwi	mod	seriesdescription
653	37654	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	7__AX_T2_FRFSE
653	37654	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	10_Ax_DWI_1400_Bvalue
653	37654	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_
663	4081	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	3__AX_T2
663	4081	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	4__Ax_DWI_FOCUS_1400__5FOV
663	4081	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	400__Apparent_D_nt__mm2_s_
670	4415	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
670	4415	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
670	4415	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
829	4478	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
829	4478	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_1400_Bvalue
829	4478	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_
941	4119	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	2__AX_T2
941	4119	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	3__Ax_DWI_FOCUS_1400__5FOV
941	4119	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	302__Apparent_D_nt__mm2_s_
1157	4245	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	3__AX_T2_FRFSE
1157	4245	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	4__Ax_DWI_FOCUS_1400_18FOV
1157	4245	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	1000_Apparent_D_nt__mm2_s_
1211	4510	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
1211	4510	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
1211	4510	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
1217	4204	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	4__AX_T2_FRFSE
1217	4204	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_FOCUS_1400_18FOV
1217	4204	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	800__Apparent_D_nt__mm2_s_
1230	3960	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	2__AX_T2
1230	3960	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	4__Ax_DWI_FOCUS_1400__5FOV
1230	3960	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	400__Apparent_D_nt__mm2_s_
1230	3232	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	2__AX_T2
1230	3232	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	3__Ax_DWI_FOCUS_1400__5FOV
1230	3232	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	300__Apparent_D_nt__mm2_s_
1265	3996	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	2__AX_T2
1265	3996	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	4__Ax_DWI_FOCUS_1400__5FOV
1265	3996	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	400__Apparent_D_nt__mm2_s_

Figure 7.19: Image Acquisition Parameters, p. 1

patient_id	study_id	scanner_manufacturer	scanner_model	magnetic_field_strength	endorectal_coil	b_value_dwi	mod	seriesdescription
1366	4573	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
1366	4573	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
1366	4573	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
1382	4422	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
1382	4422	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
1382	4422	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
1535	4359	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
1535	4359	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
1535	4359	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
1695	4234	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	3__AX_T2_FRFSE
1695	4234	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	4__Ax_DWI_FOCUS_1400_18FOV
1695	4234	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	900__Apparent_D_nt__mm2_s_
1706	4745	SIEMENS	Verio	3	YES	n/a	T2W	7__AX_T2_TSE
1706	4745	SIEMENS	Verio	3	YES	1400	DWI	13_DWI_Resolve_WIP0_1400
1706	4745	SIEMENS	Verio	3	YES	n/a	ADC	14_DWI_Resolve__0_1400_ADC
1716	4465	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
1716	4465	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
1716	4465	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
1763	4187	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	2__AX_T2_FRFSE
1763	4187	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	3__Ax_DWI_FOCUS_1400__5FOV
1763	4187	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	1100_Apparent_D_nt__mm2_s_
1779	4446	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
1779	4446	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_1400_Bvalue
1779	4446	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_
1787	3955	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
1787	3955	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	7__Ax_DWI_1400_Bvalue
1787	3955	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	600__Apparent_D_nt__mm2_s_
1787	4381	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
1787	4381	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_1400_Bvalue
1787	4381	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_
1888	4784	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
1888	4784	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
1888	4784	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_

Figure 7.20: Image Acquisition Parameters, p. 2

patient_id	study_id	scanner_manufacturer	scanner_model	magnetic_field_strength	endorectal_coil	b_value_dwi	mod	seriesdescription
1918	4499	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
1918	4499	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
1918	4499	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
2010	3144	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	2__AX_T2
2010	3144	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	3__Ax_DWI_FOCUS_1400__5FOV
2010	3144	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	300__Apparent_D_nt__mm2_s_
2035	4564	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
2035	4564	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
2035	4564	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	800__Apparent_D_nt__mm2_s_
2063	4576	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	4__AX_T2_FRFSE
2063	4576	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	7__Ax_DWI_1400_Bvalue
2063	4576	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
2141	37701	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
2141	37701	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_1400_Bvalue
2141	37701	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	600__Apparent_D_nt__mm2_s_
2143	4108	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	2__AX_T2
2143	4108	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	3__Ax_DWI_FOCUS_1400__5FOV
2143	4108	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	300__Apparent_D_nt__mm2_s_
2171	4158	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	5__AX_T2_FRFSE
2171	4158	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	9__Ax_DWI_FOCUS_1400_18FOV
2171	4158	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	900__Apparent_D_nt__mm2_s_
2203	4195	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	7__AX_T2_FRFSE
2203	4195	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	9__Ax_DWI_FOCUS_1400_18FOV
2203	4195	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	501__Apparent_D_nt__mm2_s_
2327	21507	SIEMENS	Verio	3	YES	n/a	T2W	8__AX_T2
2327	21507	SIEMENS	Verio	3	YES	1400	DWI	12_DWI_Resolve_WIP0_1400
2327	21507	SIEMENS	Verio	3	YES	n/a	ADC	13_DWI_Resolve__0_1400_ADC
2437	3802	SIEMENS	Prisma	3	NO	n/a	T2W	4__AX_T2
2437	3802	SIEMENS	Prisma	3	NO	1400	DWI	7__DWI_1400__ORIG
2437	3802	SIEMENS	Prisma	3	NO	n/a	ADC	6__DWI_1400__ADC_DFC_MIX
2498	4516	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	7__AX_T2_FRFSE
2498	4516	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	10_Ax_DWI_1400_Bvalue
2498	4516	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_

Figure 7.21: Image Acquisition Parameters, p. 3

patient_id	study_id	scanner_manufacturer	scanner_model	magnetic_field_strength	endorectal_coil	b_value_dwi	mod	seriesdescription
2532	4459	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
2532	4459	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
2532	4459	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
2547	4634	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
2547	4634	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_1400_Bvalue
2547	4634	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_
2550	4765	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
2550	4765	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
2550	4765	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	704__Apparent_D_nt__mm2_s_
2724	4166	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
2724	4166	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	10_Ax_DWI_FOCUS_1400_18FOV
2724	4166	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	1000_Apparent_D_nt__mm2_s_
2901	3531	SIEMENS	Aera	1.5	NO	n/a	T2W	3__AX_T2
2901	3531	SIEMENS	Aera	1.5	NO	1000	DWI	8__AX_DIFFUSION_C_BVAL_DFC
2901	3531	SIEMENS	Aera	1.5	NO	n/a	ADC	7__AX_DIFFUSION_00_ADC_DFC
2912	4227	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	4__AX_T2_FRFSE
2912	4227	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_FOCUS_1400_18FOV
2912	4227	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	800__Apparent_D_nt__mm2_s_
2928	3578	SIEMENS	Prisma	3	NO	n/a	T2W	4__AX_T2
2928	3578	SIEMENS	Prisma	3	NO	1400	DWI	7__DWI_1400__ORIG
2928	3578	SIEMENS	Prisma	3	NO	n/a	ADC	6__DWI_1400__ADC_DFC_MIX
2947	3606	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	6__AX_T2
2947	3606	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	7__Ax_DWI_FOCUS_1400__5FOV
2947	3606	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	700__Apparent_D_nt__mm2_s_
2966	4549	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
2966	4549	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
2966	4549	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
2983	3675	SIEMENS	Prisma	3	NO	n/a	T2W	3__AX_T2
2983	3675	SIEMENS	Prisma	3	NO	1400	DWI	6__DWI_1400__ORIG
2983	3675	SIEMENS	Prisma	3	NO	n/a	ADC	5__DWI_1400__ADC_DFC_MIX
2991	4202	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	4__AX_T2_FRFSE
2991	4202	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_FOCUS_1400_18FOV
2991	4202	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	800__Apparent_D_nt__mm2_s_

Figure 7.22: Image Acquisition Parameters, p. 4

patient_id	study_id	scanner_manufacturer	scanner_model	magnetic_field_strength	endorectal_coil	b_value_dwi	mod	seriesdescription
3012	3729	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	3__AX_T2
3012	3729	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	4__Ax_DWI_FOCUS_1400__5FOV
3012	3729	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	400__Apparent_D_nt__mm2_s_
3024	4441	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3024	4441	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
3024	4441	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
3035	3772	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	6__AX_T2
3035	3772	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	7__Ax_DWI_FOCUS_1400__5FOV
3035	3772	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	900__Apparent_D_nt__mm2_s_
3071	4751	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3071	4751	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_1400_Bvalue
3071	4751	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_
3104	21269	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
3104	21269	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_1400_Bvalue
3104	21269	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_
3105	3895	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
3105	3895	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
3105	3895	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	700__Apparent_D_nt__mm2_s_
3121	4520	SIEMENS	Aera	1.5	YES	n/a	T2W	6__AXIAL_T2
3121	4520	SIEMENS	Aera	1.5	YES	1400	DWI	13_AX_DIFFUSION_AL_DFC_MIX
3121	4520	SIEMENS	Aera	1.5	YES	n/a	ADC	12_AX_DIFFUSION_DC_DFC_MIX
3139	3957	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	7__AX_T2
3139	3957	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	8__Ax_DWI_FOCUS_1400__5FOV
3139	3957	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	1000_Apparent_D_nt__mm2_s_
3148	3975	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3148	3975	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	7__Ax_DWI_1400_Bvalue
3148	3975	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	600__Apparent_D_nt__mm2_s_
3153	3985	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3153	3985	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_FOCUS_1400_18FOV
3153	3985	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	600__Apparent_D_nt__mm2_s_
3192	4065	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	12_AX_T2_FRFSE
3192	4065	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	11_Ax_DWI_FOCUS_1400_18FOV
3192	4065	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	1100_Apparent_D_nt__mm2_s_

Figure 7.23: Image Acquisition Parameters, p. 5

patient_id	study_id	scanner_manufacturer	scanner_model	magnetic_field_strength	endorectal_coil	b_value_dwi	mod	seriesdescription
3197	4074	SIEMENS	Prisma	3	NO	n/a	T2W	3__AX_T2_3MM_3
3197	4074	SIEMENS	Prisma	3	NO	1400	DWI	6__DWI_1400_500_ORIG
3197	4074	SIEMENS	Prisma	3	NO	n/a	ADC	4__DWI_1400_500_EW_DFC_MIX
3202	4087	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	4__AX_T2_FRFSE
3202	4087	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	7__Ax_DWI_FOCUS_1400_18FOV
3202	4087	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	600__Apparent_D_nt__mm2_s_
3207	4094	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	2__AX_T2
3207	4094	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	3__Ax_DWI_FOCUS_1400__5FOV
3207	4094	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	300__Apparent_D_nt__mm2_s_
3219	4115	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	4__AX_T2
3219	4115	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	5__Ax_DWI_FOCUS_1400__5FOV
3219	4115	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	800__Apparent_D_nt__mm2_s_
3228	4130	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	4__AX_T2
3228	4130	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	6__Ax_DWI_FOCUS_1400__5FOV
3228	4130	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	600__Apparent_D_nt__mm2_s_
3230	4132	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	3__AX_T2
3230	4132	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	4__Ax_DWI_FOCUS_1400__5FOV
3230	4132	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	400__Apparent_D_nt__mm2_s_
3234	4138	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	3__AX_T2
3234	4138	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	4__Ax_DWI_FOCUS_1400__5FOV
3234	4138	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	1000_Apparent_D_nt__mm2_s_
3242	4149	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	2__AX_T2
3242	4149	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	3__Ax_DWI_FOCUS_1400__5FOV
3242	4149	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	300__Apparent_D_nt__mm2_s_
3243	4150	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	2__AX_T2
3243	4150	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	3__Ax_DWI_FOCUS_1400__5FOV
3243	4150	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	300__Apparent_D_nt__mm2_s_
3246	4154	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	2__AX_T2
3246	4154	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	4__Ax_DWI_FOCUS_1400__5FOV
3246	4154	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	400__Apparent_D_nt__mm2_s_
3249	4161	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3249	4161	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_FOCUS_1400_18FOV
3249	4161	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	900__Apparent_D_nt__mm2_s_

Figure 7.24: Image Acquisition Parameters, p. 6

patient_id	study_id	scanner_manufacturer	scanner_model	magnetic_field_strength	endorectal_coil	b_value_dwi	mod	seriesdescription
3250	4163	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	4__AX_T2
3250	4163	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	5__Ax_DWI_FOCUS_1400__5FOV
3250	4163	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	800__Apparent_D_nt__mm2_s_
3262	4190	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
3262	4190	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	10_Ax_DWI_FOCUS_1400_18FOV
3262	4190	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	1000_Apparent_D_nt__mm2_s_
3264	4194	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	7__AX_T2_FRFSE
3264	4194	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	12_Ax_DWI_FOCUS_1400_18FOV
3264	4194	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	1200_Apparent_D_nt__mm2_s_
3268	4210	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	4__AX_T2_FRFSE
3268	4210	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	5__Ax_DWI_FOCUS_1400_18FOV
3268	4210	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
3269	4211	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	3__AX_T2_FRFSE
3269	4211	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	5__Ax_DWI_FOCUS_1400_18FOV
3269	4211	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	700__Apparent_D_nt__mm2_s_
3280	4226	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	4__AX_T2_FRFSE
3280	4226	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_FOCUS_1400_18FOV
3280	4226	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	800__Apparent_D_nt__mm2_s_
3287	4240	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
3287	4240	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	10_Ax_DWI_FOCUS_1400_18FOV
3287	4240	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	1000_Apparent_D_nt__mm2_s_
3403	4464	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3403	4464	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	10_Ax_DWI_1400_Bvalue
3403	4464	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
3446	4560	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
3446	4560	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_1400_Bvalue
3446	4560	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_
3448	4563	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3448	4563	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
3448	4563	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
3468	4599	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	T2W	6__AX_T2_PROPELLER
3468	4599	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	1400	DWI	11_Ax_DWI_1400
3468	4599	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	NO	n/a	ADC	1100_Apparent_D_nt__mm2_s_

Figure 7.25: Image Acquisition Parameters, p. 7

patient_id	study_id	scanner_manufacturer	scanner_model	magnetic_field_strength	endorectal_coil	b_value_dwi	mod	seriesdescription
3473	4612	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	7__AX_T2_FRFSE
3473	4612	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	10_Ax_DWI_1400_Bvalue
3473	4612	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_
3476	4616	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3476	4616	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
3476	4616	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
3479	4622	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	4__AX_T2_FRFSE
3479	4622	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	11_Ax_DWI_1400_Bvalue
3479	4622	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
3563	4798	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3563	4798	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
3563	4798	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	700__Apparent_D_nt__mm2_s_
3592	4854	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3592	4854	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
3592	4854	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
3605	4876	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
3605	4876	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	8__Ax_DWI_1400_Bvalue
3605	4876	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	701__Apparent_D_nt__mm2_s_
3624	4910	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	6__AX_T2_FRFSE
3624	4910	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	10_Ax_DWI_1400_Bvalue
3624	4910	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	901__Apparent_D_nt__mm2_s_
12223	37671	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
12223	37671	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_1400_Bvalue
12223	37671	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	600__Apparent_D_nt__mm2_s_
12224	37674	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	T2W	5__AX_T2_FRFSE
12224	37674	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	1400	DWI	9__Ax_DWI_1400_Bvalue
12224	37674	GE MEDICAL SYSTEMS	DISCOVERY MR750w	3	YES	n/a	ADC	600__Apparent_D_nt__mm2_s_

Figure 7.26: Image Acquisition Parameters, p. 8

List of Figures

List of Tables