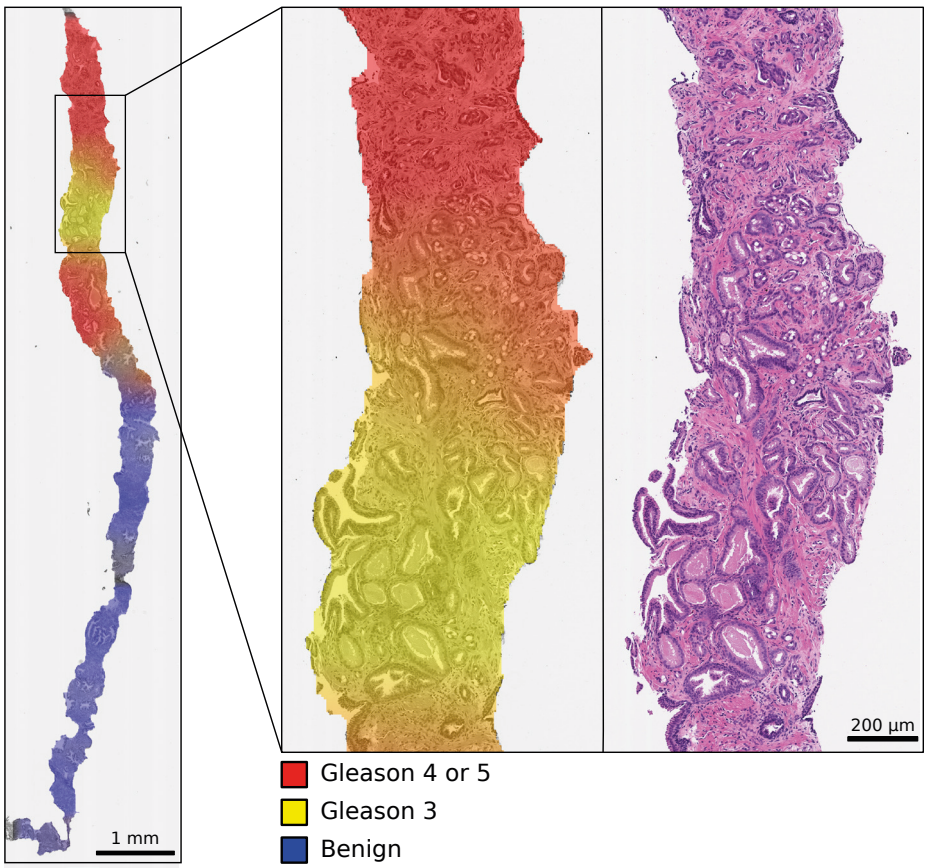
Thesis for doctoral degree (Ph.D.)
2020

Artificial intelligence for streamlining prostate cancer diagnostics



Peter Ström

Karolinska Institutet

From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Artificial intelligence for streamlining prostate cancer diagnostics

Peter Ström

Karolinska Institutet

Stockholm 2020

All published papers reproduced with permission
Published by Karolinska Institutet
Printed by E-Print AB 2020

Typeset by the author using LaTeX 2$_\varepsilon$

ISBN 978-91-7831-795-0

Institutionen för Medicinsk Epidemiologi och Biostatistik

Artificial intelligence for streamlining prostate cancer diagnostics

AKADEMISK AVHANDLING som för avläggande av medicine doktorseksamen vid Karolinska Institutet offentligen försvaras i hörsal Atrium, Nobels väg 12B, Karolinska Institutet, Solna

Fredag 15 Maj 2020, kl 09.00

By
Peter Ström

Principal supervisor:
Associate Professor Martin Eklund
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Co-supervisor:
Doctor Tobias Nordström
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Associate Professor Mattias Rantalainen
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Associate Professor Mark Clements
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Opponent:
Associate Professor Darren Treanor
University of Leeds
Leeds Institute of Molecular Medicine

Examination board:
Professor Fang Fang
Karolinska Institutet
Ïnstitute of Environmental Medicine

Associate Professor Nicocla Orsini
Karolinska Institutet
Department of Global Public Health

Doctor Bjørnar Gilje
Stavanger University Hospital
Department of hematology and oncology

To my family, Sara, Måns, and Svante.

Abstract

With around 1.2 million cases per year, prostate cancer is the second most common cancer among men. It is usually a slow growing disease that affects older men. It is also a cancer that is heterogenous, often multifocal, and rarely show symptoms as long as it is localized. All these things make the disease difficult to detect, diagnose and study. The objective of this thesis is to develop and improve technologies for prostate cancer diagnostics and to acquire knowledge related to these technologies that directly translate to clinical utility.

In Study I, we extended analysis of the multivariable diagnostic prediction model S3M by exploring the relative contribution from the individual predictors and evaluating the model in reflex setting where the test is only given to men positive on a PSA test. We also updated the list of included predictors and refitted the model to more data.

In Study II, we digitized a substantial part of the biopsy cores collected from the men in study I. These images were used to develop and validate an AI for prostate cancer diagnostics by detecting, grading, and measuring the extent of cancer in the biopsies. The AI achieved nearly perfect detection of cancer and expert pathologist level grading of the biopsies. It also well predicted the total tumor burden of the patient.

In Study III, we focused our attention on perineural invasion, a common finding in prostate biopsies. This study has added to the evidence that there is substantial and independent prognostic information in this finding and argued that it should be included as a compulsory part in pathology reporting guidelines for prostate biopsies.

In Study IV, we developed an AI for detection and localization of perineural invasion in biopsies. The AI achieved high discriminative ability on an independent test set. We are currently collecting external data to validate these results in another environment and to compare the results of the AI against expert pathologists.

In conclusion, the technologies developed in this thesis has shown promise in streamlining the clinical workload around prostate cancer detection and diagnostics. The thesis has also contributed to pieces of information related to these technologies.

List of publications

I. **Peter Ström**, Tobias Nordström, Markus Aly, Lars Egevad, Henrik Grönberg, and Martin Eklund
The Stockholm-3 Model for Prostate Cancer Detection: Algorithm Update, Biomarker Contribution, and Reflex Test Potential.
European Urology 2018

II. **Peter Ström***, Kimmo Kartasalo*, Henrik Olsson, Leslie Solorzano, Brett Delahunt, Daniel M Berney, David G Bostwick, Andrew J. Evans , David J Grignon, Peter A Humphrey, Kenneth A Iczkowski, James G Kench, Glen Kristiansen, Theodorus H van der Kwast, Katia RM Leite, Jesse K McKenney, Jon Oxley, Chin-Chen Pan, Hemamali Samaratunga, John R Srigley, Hiroyuki Takahashi, Toyonori Tsuzuki, Murali Varma, Ming Zhou, Johan Lindberg, Cecilia Lindskog, Pekka Ruusuvuori, Carolina Wählby, Henrik Grönberg, Mattias Rantalainen, Lars Egevad, and Martin Eklund
Grading prostate biopsies with artificial intelligence: a diagnostic study
LANCET Oncology 2019

III. **Peter Ström**, Tobias Nordström, Brett Delahunt, Hemamali Samaratunga, Henrik Grönberg, Lars Egevad, and Martin Eklund
Prognostic value of perineural invasion in prostate needle biopsies: a population-based study of patients treated by radical prostatectomy
Journal of Clinical Pathology 2020

IV. **Peter Ström**, Kimmo Kartasalo, Pekka Ruusuvuori, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Lars Egevad, and Martin Eklund
Detection of Perineural Invasion in Prostate Needle Biopsies with Deep Neural Networks
ArXiv pre-print 2020

* Equal contribution
The articles will be referred to in the text by their Roman numerals, and are reproduced in full at the end of the thesis.

Related publications

ORIGINAL RESEARCH

- Anna Wallerstedt*, Peter Strom*, Henrik Gronberg, Tobias Nordstrom, and Martin Eklund
  Risk of Prostate Cancer in Men Treated With 5a-ReductaseInhibitors—A Large Population-Based Prospective Study
  JNCI 2018
- Henrik Grönberg, Martin Eklund, Wolfgang Picker, Markus Aly, Fredrik Jäderling, Jan Adolfsson, Martin Landquist, Erik Skaaheim Haug, Peter Ström, Stefan Carlsson, and Tobias Nordström
  Prostate Cancer Diagnostics Using a Combination of the Stockholm3 Blood Test and Multiparametric Magnetic Resonance Imaging
  Eur Urol 2018

CORRESPONDENCE

- Martin Eklund, Peter Ström, Tobias Nordström, and Henrik Grönberg
  Reply to Ola Bratt and Anna Öfverholm's Letter to the Editor re: Peter Ström, Tobias Nordström, Henrik Grönberg, Martin Eklund. The Stockholm-3 Model for Prostate Cancer Detection: Algorithm Update, Biomarker Contribution, and Reflex Test Potential.
  Eur Urol 2018
- Peter Ström, Anna Wallerstedt Lantz, Henrik Grönberg, Tobias Nordström, and Martin Eklund
  Response to Walsh
  JNCI 2018
- Martin Eklund, Peter Ström, Henrik Grönberg, and Tobias Nordström
  Reply to Erik Rud, Peter Lauritzen, and Eduard Baco's Letter to the Editor re: Henrik Grönberg, Martin Eklund, Wolfgang Picker, et al. Prostate Cancer Diagnostics Using a Combination of the Stockholm3 Blood Test and Multiparametric Magnetic Resonance Imaging.
  Eur Urol 2019
- Martin Eklund, Kimmo Kartasalo, Henrik Olsson, and Peter Ström
  The importance of study design in the application of artificial intelligence methods in medicine.
  NPJ Digit Med. 2019

* Equal contribution.

Contents

List of abbreviations

CI	Confidence Interval
GG	Gleason Grade
IQR	Inter Quartile Range
ISUP	International Society of Urological Pathology
PSA	Prostate-Specific Antigen
HR	Hazard Ratio
RP	Radical Prostatectomy
ROC	Receiver Operating Characteristics
AUC	Area Under Curve (Receiver Operating Characteristic)
IoU	intersection over Union
BCR	Biochemical recurrence (here: prosate cancer relapse)

Chapter 1

Aims of the thesis

Due to the common use of the imprecise PSA screening test and the complexity and subjective nature of the diagnosis of prostate cancer, this disease is known for both high overtreatment and undertreatment leading to unnecessary anxiety, unpleasant and risky operational procedures, and, in the worst case, death. By making use of modern computation, high quality data and advanced prediction models, this thesis aims to improve the screening of patients, develop efficient diagnostic algorithms and improve prognostication.

Specifically, the aims were to:

- Further develop the Stockholm-3 model (S3M) for prostate cancer risk prediction, and to evaluate its use as a reflex test to PSA with to avoid unnecessary biopsies and decrease overdiagnosis of indolent disease.
- Develop an artificial intelligence (AI) for pathology assessment of prostate biopsies to automatically detect cancer, grade the cancer and to estimate the amount of cancer in the biopsy.
- Estimate the prognostic value of perineural invasion (PNI) in prostate biopsies among men undergoing radical prostatectomy.
- Develop an AI for prostate biopsies to automatically detect PNI with clinically acceptable diagnostic accuracy.

Chapter 2

Background

2.1 Prostate Epidemiology

Prostate cancer is the most common cancer among men in the Western world, and it is estimated that 1 in 8 men will develop the disease within his lifetime and that 1 in 5 diagnosed cancers originate from the prostate. [1] It is the second most common cause of cancer death among men in Europe and North America. There is not so much known about the causes of prostate cancer, and the main known risk factors are high age (most cancers occur after age 60), ethnicity (higher incidents among African descent and lower among men of Asian descent), and family history of the disease.

2.2 Screening

2.2.1 PSA and other blood markers

For prostate cancer, as for most cancers, it is crucial to diagnose the disease in a certain window of the cancer's natural history - with too early testing a present cancer may too small to be possible to diagnose and with too late testing the disease may no longer be curable. For decades, Prostate-Specific Antigen (PSA) has been the primary marker for early detection of prostate cancer, for assessing the prognosis and to monitor actively treated patients or patients with a diagnosed low-risk cancer who are undergoing active surveillance.

PSA's biological role is believed to involve cleavage of seminal proteins and thereby liquefying the seminal fluid. [2] In a young and healthy prostate, the PSA is largely contained within the prostate, but with older age there may be leakage of PSA into the blood. [3] This can be caused by prostate cancer or non-malignant disease such as benign prostatic hyperplasia (BPH) or inflammation. However, prostate cancer can be present also without elevated PSA in the blood. The fact that PSA can be elevated for other reasons than prostate cancer and that prostate cancer can be present without leading to elevated PSA leads to poor test characteristics when using PSA as a screening

tool. Therefore, no government has introduced a national screening program based on PSA despite its low cost and non-invasive sampling procedure. The PSA test measures the total PSA (tPSA) in the blood, but this can be divided into complexed- and free PSA (cPSA and fPSA). The fraction of fPSA to tPSA has been shown to have significant better ability to discriminate between BPH and prostate cancer than total PSA alone and is therefore sometime used in combination with PSA in evaluating the patients' risk profile. [4] fPSA can be found in various forms, and some of these provide additional value as predictive markers. Four of these that have shown to discriminate between prostate cancer and either healthy tissue or BPH are inactive PSA (ProPSA), intact PSA (iPSA), nicked PSA and BPSA. [5]

2.2.2 S3M

The STHLM3 study was designed to develop and validate a novel prediction model for clinically relevant prostate cancer. [6] It was a prospective study were men in the ages 50-69 from the Stockholm county were invited to participate between May 2012, and December 2014. The diagnostic prediction model, the Stockholm-3 Model (S3M), was developed on the first 11,130 participants, and later evaluated on 47,688 independent men. The S3M is a logistic regression model which predicts the risk of having clinically significant prostate cancer (for S3M defined as ISUP grade 2 or higher, see Section 2.3.4). It uses a wide range of predictors such as age, first-degree family member with history of prostate cancer, and if the patient has had a previous biopsy, findings from digital rectal examination and the prostate volume. It also contains molecular information: blood-based protein biomarkers in the form of PSA and its derivatives, total PSA, free PSA, intact PSA, the ratio of free to total PSA, hK2 (human kallikrein 2), MIC1 (Macrophage inhibitory cytokine-1), MSMB (microseminoprotein-beta), and genetic markers in the form of a genetic score based on 254 single-nucleotide polymorphisms (SNPs). In a later modification of the S3M, intact PSA was removed and instead the HOXB13 SNP was used as an individual marker due to its relatively high prevalence in the population and particularly for its high influence on the risk for developing prostate cancer.

When validating S3M, it was compared against PSA. Specifically, the evaluation was such that both PSA and S3M by design would detect equally many cancer patients (with grade ISUP2 or higher), and the outcome of interest was how many men each test would need to refer to a biopsy to find those cancers. Men with serum PSA 1 ng/mL or higher were assessed for the predictors needed to evaluate S3M. Prostate biopsy referral was then based on positivity on either PSA above 3 ng/mL or S3M score above 10%. The study concluded that S3M could reduce the number of men referred to biopsy by 32 percent, as well as reduce the number of diagnosed indolent cancers

(ISUP 1, which are typically considered overdiagnosed cancers) by 17%.

2.2.3 Other diagnostic prediction models

Several statistical diagnostic tools have demonstrated improvement relative to PSA in discriminating between clinically relevant prostate cancer and healthy men or men with nearly harmless cancer. The four-kallikrein (4Kscore) blood based multivariate prediction model combines measures of tPSA, fPSA, iPSA, and human kallikrein 2 (hK2) with age, digital rectal examination, and indication of previous prostate biopsy. [7] The Prostate Health Index (PHI) is another well known and clinically used risk model. [8] It uses PSA, ProPSA and fPSA to calculate a risk score based on the deterministic formula

$$\frac{\text{ProPSA}}{\text{fPSA}} \times \sqrt{PSA}.$$

Non-blood-based approaches are the urinary based PCA3-test which measures over-expression of the mRNA PCA3, and the RC3-test which make use of clinical pre-biopsy information (PSA, digital rectal examination and prostate volume) to predict clinically relevant cancer. [9, 10] These tests have also shown improvements in discrimination (AUC) and clinical test characteristics (avoided biopsies and reduced number of diagnosed men with indolent cancer) compared to the use of PSA alone.

2.3 Diagnosis

2.3.1 Importance of the diagnosis

Prostate cancer almost exclusively affects older men and often has a very slow development. It is also very heterogeneous, with low grade disease rarely causing any harm to the patient while high-grade often ultimately lead to the death of the patient. With high risk of severe side effects such as incontinence and impotence from radical prostatectomy or radiotherapy, it therefore not obvious to perform treatment with curative intent after diagnosing the cancer. Things to consider are the non-cancer related life expectancy, the preference of the patient (e.g. anxiety of untreated cancer and inconvenience of repeated biopsies to monitor the disease), and not least the pathological grade of the cancer, which is the main prognostic information.

2.3.2 Prostate biopsies

In suspicion of prostate cancer, typically 10 or 12 needles are inserted in the prostate for tissue sampling from the most common locations of cancer lesions in the prostate (the peripheral zone). The histological evaluation of these samples provides the basis for diagnosis. Even though the aim is to get a representative sample of the prostate, it is

possible that the needles miss present cancer. This is more likely to happen with small tumours. Therefore, pre-biopsy magnetic resonance imaging (MRI) of the prostate has become increasingly common. MRI permits the use of targeted biopsies, where needles are directed towards areas identified as suspicious by the MRI. MRI and targeted biopsies have been shown to both increase the sensitivity and specificity of prostate biopsies. [11]

2.3.3 The Gleason and ISUP grading system

In 1966 the first version of the well know Gleason scoring system of histopathological tissue of the prostate was presented. It has since then been the primary prognostic tool for prostate cancer patients. Today, the Gleason grading system is used by 99.5 percent of european uro-pathologist. [12] Each cancerous region is graded 1-5 according to morphological pattern of the glands, where 5 corresponds to the least differentiated cells (i.e. limited or no glandular structure remaining), see Figure 2.1. Originally, the most common and second most common grade was combined to a score (e.g. $4 + 3 = 7$). In 2005, this was changed to the most common and either the highest of the remaining grades (if higher than the dominant grade) or the second most common grade. The second most common grade must either be higher than the dominant or with a prevalence of $>5\%$ to be counted. [13] Further, it was recommended not to include grade 1 and 2. In 2014, the same group (International Society of Urological Pathology) decided on a new score (ISUP grade 1-5) based on the Gleason score. [14] ISUP 1 and 2 are the most common scores and corresponds to Gleason score $3 + 3$ and $3 + 4$, respectively. For prostatectomies it is the most common and second most common (with the 5% rule) that forms the Gleason Score.

2.3.4 Cancer length

Although the ISUP (or Gleason) score is the main information used for prognostication and for deciding on treatment, it is also of value to estimate the extent of the tumor. There is no consensus of how the extent is to be measured, but it is often discussed at pathology conferences (personal communication with Prof. Lars Egevad (L.E.)). All measures of cancer extent in this thesis was assessed by a single pathologist (L.E.) and the linear cancer extent was generally measured from end to end in cases with discontinuous cancer. However, in cases with 1 or 2 cores infiltrated by low grade discontinuous cancer with a benign gap exceeding 3 mm, benign tissue was subtracted in the reporting of total cancer extent. This measure was performed with a ruler under the microscope and measured in the direction of the biopsy core. The smallest reported size was <0.5 mm.

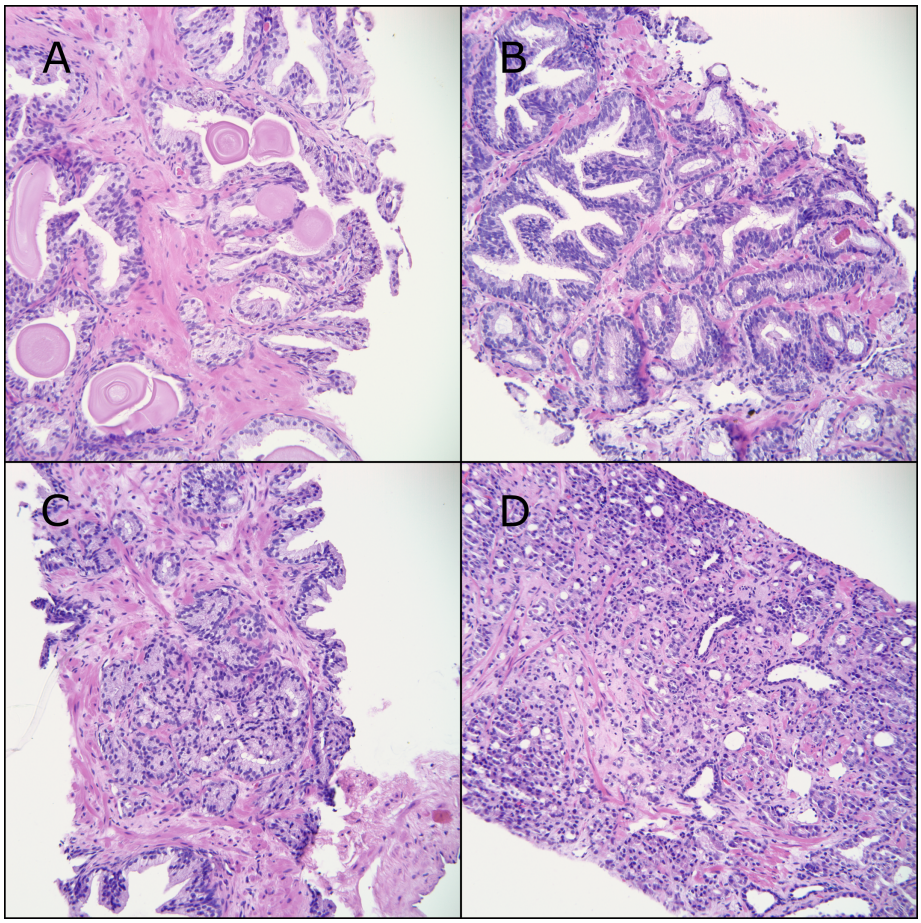


Figure 2.1: Examples of Gleason patterns. **(A)** Benign prostate glands have a distinct folded shape with small nuclei and pale cytoplasm. **(B)** Gleason pattern 3, glands become smaller with darker cytoplasm and larger nuclei. The glands are separated. **(C)** Gleason pattern 4, the glands show loss of differentiation and fuse together. **(D)** Gleason pattern 5 (and 4), even more loss of differentiation and often difficult to distinguish the gland from the connective tissue. These are not unique representations from each Gleason pattern. All examples are eosin and hematoxylin stained biopsy cores from the STHLM3 study taken with microscope at 20X resolution. *Photograph: L. Egevad and P. Ström.*

2.3.5 Staging

Staging is used to quantify how far the tumor has developed. Together with information of PSA and microscopic findings (primarily Gleason/ISUP grade), this is the foundation for primary treatment decisions. The T stage aims to capture the size of the tumor:

- T1: The tumor is not palpable or noticeable by ultrasound or imaging.
- T2: The tumor is palpable but still fully contained within the prostate.

- T3: The tumor has now broken through the capsule and possibly grown into the seminal vesicles.
- T4: The tumor has grown into nearby organs.

In addition to the T staging, there are also N and M stages (referred to as TNM staging). N is for regional spread to the lymph nodes, and M is for metastasis i.e. for spread to distant organs in the body:

- N0: No spread after evacuation of nearby lymph.
- Nx: No assessment.
- N1: Shown regional lymph spread.

and

- M0: No metastasis.
- M1: Metastasis.

2.4 Prognosis

The prognosis after a prostate cancer diagnosis is relatively good. The relative 10-year survival is 88%. Particularly good is the prognosis for low grade diseases, which are slow growing and very seldom shortens the patient's life even without curative treatment. [15] The Swedish national guidelines categorize the patients into risk groups to facilitate treatment decisions, see Table 2.1. [16] In addition to this there are other

Table 2.1: The Swedish national guidelines for categorizing patients into risk groups.

Risk	Definition	Treatment
Low	T1-T2a and Gleason 6 and PSA $<$ 10 ng/mL	Active surveillance
Intermediate	T2b or Gleason $=$ 7 or 10 $\leq$ PSA $\leq$ 19 ng/mL	Possibly radiology or surgery
High	T2c-T3 or extensive Gleason $=$ 3+4 or higher or PSA $\geq$ 20 ng/mL	Recommended radiology or surgery

factors to consider, such as the general health status of the patient and other findings from the microscopic assessment of the biopsy. For example, cribriform, intraductal cancer and perineural invasion has shown to be associated with poorer prognosis. [17][18]

Without curative treatment about 1% to 10% of the low risk patients will die from the cancer within 10 to 15 years.[19][20] In one study, 8 of 545 men on active surveillance died within 10 years. [19] For the intermediate risk group, the corresponding

risk is 20%. It should be noted that these estimates are based on diagnoses from before 2005 and today more than half of the low risk patients would be categorized as intermediate risk. For the high-risk group, the prognosis is a lot worse: between 20% and 30% die within 5 years from diagnosis. [20][21]

2.5 Initial Treatment

There are several options for treatment after prostate cancer diagnosis, depending on the stage at which the prostate cancer is diagnosed.

2.5.1 Active Surveillance

A Gleason score of 3 + 3 (i.e. only containing Gleason pattern 3) is defined as cancer and is therefore malignant. However, prostate cancer is typically slow growing and occurring in relatively old men. Since this Gleason score is associated with the most favorable prognosis, there may not be any need for treatment. Also, these cancers likely lack the potential to metastasize. [22] What makes these patients such a difficult group is that the cancer may potentially be able to migrate to a higher grade, and higher grade lesions could be missed by the biopsy sampling procedure. [23] This means that these patients are recommended active surveillance which mean repeated PSA testing, biopsies and the anxiety of living with cancer.

2.5.2 Radical Prostatectomy and Radiation

In Sweden, radical prostatectomy (i.e. surgical removal of the entire prostate gland) is the most common treatment for clinically relevant prostate cancer. According to guidelines, it is the recommended treatment for mid-risk prostate cancer if life expectancy is at least 10 years. Radical prostatectomy is not without side effects. For example, 70% of the treated men experience erectile dysfunction and 20% are incontinent one year after the surgery. [24] Radiation of the prostate is also used to treat men with intermediate/high-risk prostate cancer. During this procedure, high-energy radiation beams are aimed at the prostate gland with the goal of destroying the cancer cells. If there are signs of metastases, the radical prostatectomy or radiation are typically combined with systemic treatment.

Chapter 3

Material

3.1 Overview of STHLM3

This thesis is centered around data generated from the STHLM3 study conducted between May 28, 2012, and Dec 30, 2014. [6] STHLM3 was a population based and prospective diagnostic trial where a random selection of 145,905 subjects aged 50 to 69 from the Stockholm county were invited to participate, out of which 59,159 subjects accepted. They had an initial PSA test, and subjects with PSA above $>$1ng/mL were also assessed for the variables included in the S3M (see Background). Any subject positive on either PSA $\geq$3ng/mL or an S3M probability of high-grade prostate cancer above 10% were referred to biopsy. The urologist performing the biopsy and the pathologist were blinded to the tests outcomes and other clinical information related to the patient. In total, 7,406 subjects were biopsied.

Later, we initiated a project of digitizing biopsy cores from this study. Since each biopsy in the study consisted of either 10 or 12 needle cores, the total number of cores were about 84,000. The time and cost to digitize them all were not within the scope of this project, so we prioritized a subset stratified on ISUP grade to get a large selection of all ISUP grades, including rare high-grade cases.

3.2 Slide preparation and digitization

The biopsy cores were fixated in formalin and stained with Hematoxylin and Eosin (HE). The staining is used to highlight the interesting parts of the tissue for aiding the pathological assessment. After the fixation, the formalin blocks are cut in thin slices of about 5μm. In the STHLM3 study, two of these sections were placed on a single microscope glass slide. The reason for using two sections is that some regions can be damaged (e.g. the slice of tissue folds when mounted on the glass slide) or that the morphology in a region of interest is difficult to judge and it helps to consider a second slice of that region (see Figure 3.1).

Figure 3.1: A macro image of a digitized slide from a STHLM3 biopsy core. Two sections of the same core was mounted on the glass slide. On top of the cover glass the pathologist has highlighted the cancer foci with with black ink next to the tissue. The tissue was stained with HE and the scanner was Hamamatsu. The original image size was 51,200x27,392 pixels.

In recent years, digital pathology (i.e. assessing digital images of the tissue on a computer rather than using a microscope) has exploded in popularity. It is mainly due to improved quality of high-resolution pathology scanners. There are several advantages with digitalization, it allows for very precise annotations on the images, easy distribution to other labs or countries if the expertise or labor is of shortage, more suitable for collaboration and research, and most importantly for the theme of this thesis, it opens up for algorithmic solutions to aid and possibly improve the clinical workflow and diagnosis. But there is still a need for microscopes as they arguably still have much preferred optical qualities, mainly the possibility to focus deep into the tissue (i.e. 3D instead of a 2D scanned image). Even if there are scanners that to some degree manages this today, the already large image size (often hundreds of megabytes) become difficult to manage.

There are several manufacturers of pathology scanners I this thesis, we have mainly worked with Hamamastu (Hamamatsu, Japan) and Aperio (Leica, Germany). We have also started a project digitizing a much larger set of images than what is presented in this thesis using the IntelliSite Ultra Fast Scanner (Philips, Netherlands). At the timepoint of writing this we have digitized more than 30,000 biopsy cores from prostate biopsies, both from Sweden and other countries.

Chapter 4

Methods

4.1 Evaluation of Prediction models

Receiver operating characteristics (ROC) curves and its Area Under the Curve (AUC) are usually the primary (and often the only) endpoints for evaluating medical diagnostic prediction models. There are good reasons for this. The ROC is a function of TPF (true positive fraction) and FPF (false positive fraction) for all possible positivity thresholds of the predictions, and these two parameters are important for evaluating the usefulness regarding a population. If $\hat{D}$ is indicator for model classification of Disease (D) given a certain threshold for positivity, then:

$$TPF = P(\hat{D}|D)$$
$$FPF = P(\hat{D}|\text{not}D).$$

Since these parameters condition on disease status they give the most obvious information for evaluating the test: the proportion of the diseased subjects correctly identified as such (which we want to be high) and the proportion of healthy subjects who are wrongly identified as diseased (which we want to be low). TPF and FPF are classification probabilities and are used when we want to describe how well a test discriminate between healthy and diseased subjects. [25]

The AUC of the ROC gives a summary measure of TPF(c) and FPF(c) evaluated over all c, where c is the cutoff value for deciding whether a test is positive or negative. Not only does it reduce to a single number, it also has a useful interpretable characteristic. It can be interpreted as the probability that two randomly picked subjects, one healthy and one diseased, are correctly ranked. However, the AUC as an endpoint for model-development and comparisons have been criticized on several grounds. Among others, for being too insensitive for model improvements (i.e. small model improvements may not be reflected in the AUC). It summarizes the model over all probabilities

even though only a relatively small range of probabilities may be of clinical interest. Also, comparing AUC for models evaluated on distinct populations may be misleading, for example, if one population has been previously screened and therefore possibly thinned out of true positives. This latter property is not specific for the AUC but is a general shortcoming when comparing medical tests - some populations are easier to discriminate than others with respect to the disease status of interest.

If the true outcome is continuous, say a survival time, then the Harrell's c can be used to evaluate predictive performance. [26] It is a generalization of the AUC in the sense that if the outcome it dichotomized the two parameters are equivalent, and when continuous it compares the rank of the predictions to the rank of the outcome among all pairs where it is possible to rank the outcome (e.g. it is not possible in survival data if none of the subjects yet have had the outcome).

4.2 Neural networks

Deep learning has been used since the 1960s, but it is only in last few years these models have exploded in popularity. And for good reasons. They have now become the first choice in many fields of prediction and classification due to many examples where these models have outperformed more traditional feature-based machine learning methods. It is also a scalable approach in contrast to method where specific features need to be hand-crafted for each unique task; with deep learning, such features are data driven and typically hidden from the user. The recent improvement in performance are mainly due to computationally efficient hardware, more complex models built by less complex parts, software for efficient fitting and large sets of labeled data.

So, what is a neural network? In a supervised setting, where we feed a model labeled data (e.g. an input matrix X and a corresponding class it belongs to or continuous value y), a neural network ends with a conditional distribution over all classes $P(y|x,\Theta)$. But instead of directly relate the matrix to the output distribution, we make a linear predictor of the matrix and corresponding parameters (called weights), add an intercept (called biases) and apply an non-linear function (called activation function). If several such functions (neurons) are applied to the matrix and the output from these are the input to a predictor for the output distribution, then we have a neural network with one hidden layer. Typically, one uses at least 5 or 10 hidden layers when referring to deep learning, where the depth refers to the numbers of hidden layers of parametric output and input connected in a network.

4.3 CNN

4.3.1 Overview

The most successful type of neural networks so far has been the Convolutional Neural Networks (CNNs), which can preserve the spatial structure in the network of neurons by sliding spatially small filters over the input and at each location calculate the output of the neuron based on the linear predictor of the filter parameters and the input values at that location. These outputs form a feature map of the level of activation that this filter induced in the corresponding spatial location in the input. In the fashion of neural networks, one may stack several of these convolutional layers into a network and possibly combine with the original hidden layers, so called fully connected layers. The main benefit of this is that the convolutional layers are translational invariant so that the model does not need individually specified parameters for a specific feature (i.e. convolutional filter) occurring at different spatial location in the input but can instead share these parameters.

4.3.2 CNN in histopathology

Developing mathematical algorithms for histopathological tasks such as cancer localization and cancer grading is not new, but it has historically proven difficult to extract and use relevant so called 'hand-crafted' features from these images. The recent advances in CNNs has brought new hope for success in these tasks, and some breakthroughs have already been made. In 2017, a research group used one of Google's CNN architectures (Inception V3) to classify skin-cancer subtypes. [27] They convincingly demonstrated dermatologist-level performance by evaluating it against a panel of 21 dermatologists. Another recent success has been in detecting breast cancer lymph node metastases, where deep learning algorithms outperformed a panel of 11 pathologists. [28] In prostate cancer there was no such success prior to the works of this thesis. The most significant contribution was in 2016 when Litjens et al. achieved an AUC of 0.9 discriminating between cancer and noncancer. [29] Deep learning shows great promise, but many research groups do not take their studies far enough according to a recent editorial in Nature. [30] They sacrifice reliable data and sound evaluation for quick and crowd-pleasing announcements. Many in the field, including our own research group, value rigorous study designs and high-quality evaluation, preferable in an external data set against a panel of human experts or head-to-head evaluation against current clinical tools.

4.4 Semantic segmentation

In *segmentation* we are not only concerned with classifying an image (or predicting what is in the image) but want to highlight where in the image an object is located. In the simplest case this can be the pixel coordinates of a bounding box surrounding the object of interest. But sometime this is not enough, and you need to put a label on each pixel in an image. This is called *semantic segmentation*. The output from the model will then be an image of the same width and height as the input image, with pixel values of, say, 1 for humans, 2 for cars, etc. We can go even further to distinguish each occurrence of a class in an image. E.g. instead of simply accepting that all pixels corresponding to humans are one indistinguishable mass (all with pixel value 1), we can localize each unique human in the image, say, by additionally putting a bounding box around each occurrence. This is called *instance segmentation*. In this thesis we are only concerned with semantic segmentation.

The biggest issue in going from classification (a simple CNN) to segmentation is that the spatial resolution keeps getting smaller and smaller due to the typical funnel shape of CNNs (i.e. the relevant information in an image successively loose spatial resolution). This is fine for classification where it is only of interest the know if an object exists in an image or not, but for locating the position of the object all the information that remains may be that a human is somewhere in the bottom half of the image; not at all the resolution needed for pixel-wise classification. Most approaches for not losing the spatial resolution involves a combination upsampling the last non-fully connected layer with so called skip-connections. These are direct connections from a resolution level in the downsampling stage (i.e. the encoder) to the corresponding resolution level in the upsampling stage (i.e. the decoder). These connections are typically applied at each resolution level to aid the network in preserving the special localization of the task it is trained on.

4.5 t-SNE

Sometimes it is desirable to visualize high-dimensional data. A common way of reducing the data is by PCA, but unfortunately it is often not very useful for visualization. A better approach is instead to use t-distributed stochastic neighbor embedding (t-SNE). [31] The idea is to create a similarity score between each pair of points in the high dimensional space. This is done by considering a Gaussian distribution centered over each point so that the density at all other points reflect their similarity scores. The score of a point to itself is defined as zero. These scores are then normalized so each point

has an associated distribution of similarity scores:

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

Now we allocate points in low dimensional space so that a similar approach on those data would result in pointwise similarity distributions the closely resemble those seen in high dimensional space. The problem is how to allocate them in this way. The algorithm stochastically allocates them in low dimensional space and then iteratively move the points using gradient decent with the Kullback-Leibler divergence as the objective function. The Kullback-Leibler divergence is defined as:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

and is a common way to measure the distance between two distributions. The approach described so far is basically the SNE algorithm that was published by Hinton 2002. [32] The t-SNE extends this approach in two meaningful ways. First, it uses a t-distribution with one degree of freedom for the low dimensional space since it has heavier tails than the Gaussian distribution which allow for moderately distant points in high dimensional space to map with larger distances in low dimensional space. This allows truly close point to be better separated from more distant points. Second, it used joint distributions p_{ij} and q_{ij} instead of $p_{i|j}$ and $g_{i|j}$ for high and low dimensional space, respectively. In low dimensional space

$$q_{ij} = \frac{\exp(-||y_i - y_j||^2)}{\sum_{k \neq l} \exp(-||y_k - x_l||^2)}$$

and in high dimensional space p_{ij} is defined as

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$$

With this it is possible to minimize a single Kullback-Leibler divergence between the join distributions P and Q.

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The results from this algorithm typically produces data that are very nicely visualized in 2D or 3D, with clearly distinct islands for data points that clusters in high dimensional space. For an example see the Supplementary of Study 2.

4.6 Boosted trees

4.6.1 Decision tree

Decision trees are, like single neurons, very simple algorithms and often quite poor when used on their own. But when used in abundance with computer intensive methods they often perform very well. An example of this is the XGBoost algorithm which often is a key component in the winning solution of prediction modeling competitions. [33] At its core, a tree is the simplest model you can imagine; it first splits the covariate space (the tree root) with regard to some covariate at the point where it best separate the groups you want to predict (tree depth 1 nodes). Then it continues to grow by again making a split in one or both nodes (tree depth-2 nodes), see Figure 4.1. And it continues to iteratively dichotomize the covariate space as long as it improves on the discrimination of the outcome groups. But since further splitting almost always improves discrimination, we must regularize the tree in some way to avoid overfitting. This can be done in several ways. One way is to pre-specify the maximum depth of the tree, the minimum number of samples in a node to allow it to split, and so on. Another way is to grow a very large tree and then prune it down by only keeping splits that improves discrimination at some predetermined significance level.

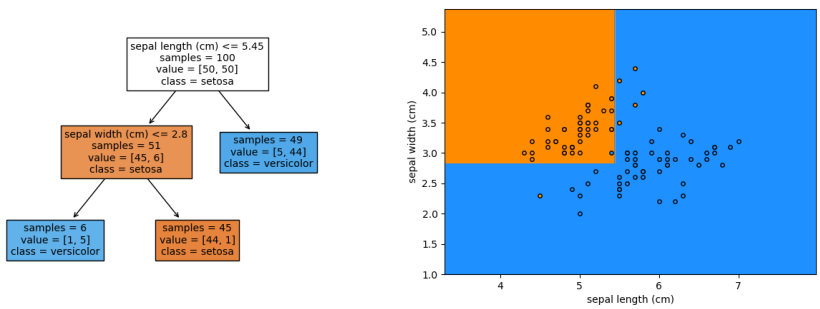


Figure 4.1: Illustration of a tree model for classifying the flowers *Setosa* and *Versicolor* from the public Iris data. **Left:** 50 samples each of the two flowers are classified based on the septal length and width. The first split is on length. If the value is higher than 5.45 cm then the model classifies Versicolor, but if it is lower it will do an additional split on width. Here all flowers with septal width lower than 2.8 cm classifies as Versicolor and the rest as Setosa. **Right:** The background color correspond to how the tree classify in the covariate space and the color of the dots indicate the true value for the species.

So how is discrimination defined in this setting? A common choice is the *Gini impurity*:

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2, \tag{4.1}$$

where $p_{i,k}$ is the fraction of samples from outcome class k in node i.

Example: *The Gini impurity for the leaf (where the branch end) that classify Setosa in Figure 4.1 is* $G_i = 1 - \frac{44}{45}^2 - \frac{1}{45}^2 = 0.04$*, a very low impurity meaning that this subset of the covariate space mostly contain the predicted class.*

To actually grow the tree, we must have a criterion for where to split. One choice that is used in the popular *Classification and Regression Tree* (CART) searches for the covariate k and its threshold t_k that produces the purest split, but it does so by weighing the splits by their size to penalize small pure leafs. [34]

$$J(k, t_k) = \frac{n_Y}{n} G_Y + \frac{n_N}{n} G_N, \tag{4.2}$$

where Y and N indicate if the threshold condition is satisfied or not. If the outcome variable is continuous, as when we predicted cancer length in Study 2, we use the mean value in each leaf for prediction and replace the Gini impurity with MSE.

4.6.2 Gradient Boosting and XGBoost

XGBoost (eXtreme Gradient Boosting) is an exceedingly popular algorithm for prediction modelling, due to its capability to achieve very good results on a wide range of tasks. It is often a key component in winning solutions on prediction competitions. [33] As the name suggests, it is an implementation of gradient boosting. [35] Boosting is a way of sequentially building an ensemble of models, in contrast to Bagging where the ensemble is built in parallel (e.g. with bootstrap samples of the data such as in Random forest). In gradient boosting each sequential model is fitted to the errors of the previous model. In this way, the kth model tries to correct the error that remains after the $(k-1)$ first models have been added to the ensemble. Boosting can be used with any kind of model, but here we will only focus on tree-based models.

Chapter 5

Results

5.1 Brief summary of the results

In **Study I** we extended the analysis of the S3M in the STHLM3 study cohort. We showed that the S3M could be used as a reflex test to PSA and thereby reduce the number of biopsies by 34% with preserved sensitivity for detecting clinically relevant prostate cancer. In **Study II** we digitized many of the biopsy cores from the STHLM3 study and built an AI for automated pathology assessment. The AI achieved world record accuracy for diagnosing cancer by an automated system and we showed for the first-time pathology level grading by evaluating the predictions against a panel of international expert pathologists. In **Study III** and **Study IV** we explored the potential of extending the AI by also flagging for PNI in the biopsies. The former showed the prognostic relevance of reporting PNI and in the latter, we implemented such an AI.

5.2 Study I

In **Study I** we extended the analyses of the S3M in three ways:

- refitted the model with slightly updated covariates and a larger proportion of STHLM3 cohort,
- evaluated the individual contributions from the biomarkers used for S3M, and
- evaluated the model as a reflex test, i.e. a second test for subjects positive on the PSA test ($\geq$3 ng/mL).

This was the second study on the S3M, and this time the training set (n=11,130), validation set (n=47,688) and an additional set of participants that did not enter the validation set due to date of end of study (n=330) were included. We used 10-fold cross-validation to build the models on a much larger data set than the original S3M was fitted on, and to be able to evaluate S3M on the whole data set. The drawback of this approach is that it is not a single model that is evaluated but instead 10 slightly

different models. The result of this should not be interpreted as “this model” but rather “a model fitted by this approach”. Due to technical reasons of the assay used in clinic for measuring PSA derivatives, we had to drop Intact PSA from the model. At the same time, we took the opportunity to single out a SNP from the genetic score (HOXB13) due to its substantial impact on the risk of developing prostate cancer. Since it is only present in a minority of the population, it is not expected to affect population metrics, but it may still have relevance for the HOXB13 carriers. The AUC in this study was only slightly increased from 0.74 to 0.75 when comparing to the preceding study.

A second aim was to do a detailed evaluation of the individual biomarkers' contribution in the S3M. We did this by considering their added value to PSA, and the loss of performance of S3M if a single biomarker was removed. We also evaluated the cumulative increase in performance by including the biomarker one at a time in one (of many possible) specific order. The main conclusions that could be drawn from this were (1) that volume has the largest impact of the biomarkers except for PSA, (2) several of the biomarkers are week predictors but together they make a strong predictor, and (3) some variables do not affect population metrics but can still be of value for a small fraction of exposed patients (such as HOXB13).

The final aim was to evaluate the S3M in a potentially more efficient way. Since refined tests for prostate biopsy referral are much more expensive than the PSA test, and they can also involve more inconvenience for the patient by, for example, measuring the volume of the prostate or performing a digital rectal examination, it is likely beneficial to reserve the refined tests to patients with relatively high risk of having prostate cancer. One way of achieving this is to use the PSA test with a relatively low threshold for positivity as an initial screening and in that way cheaply remove most benign cases and still retain a high sensitivity. In a second step we address the low specificity of the PSA test by using a multivariable prediction model, in this case S3M, on the men positive on the initial PSA test. Since you cannot improve specificity without sacrificing sensitivity, we need an objective strategy to evaluate the reward of such two-step approach. The way we chose to do that was to compare the approach to another natural way increasing the specificity, that is to use a higher threshold for the PSA test. Specifically, we chose a threshold such that both approaches resulted in the same sensitivity and evaluated the benefit in terms of achieved specificity. We found that by allowing a loss in sensitivity of 20% compared to PSA $\geq$ 3ng/mL, the S3M reflex test approach reduced the number of biopsies needed by more than 50%. The corresponding reduction by a high PSA threshold was merely 27%. For the benefit by using alternative values for the accepted loss in sensitivity, see Figure 5.1.

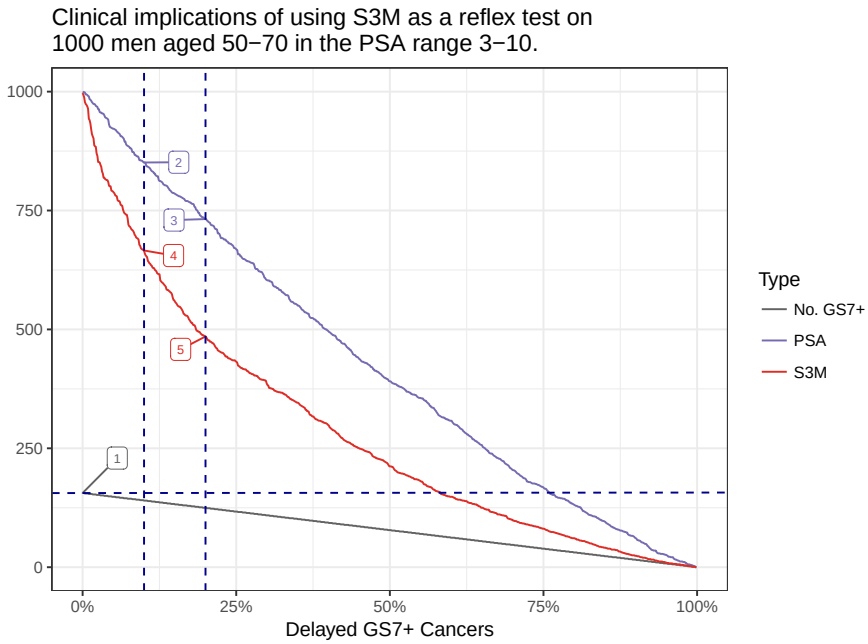


Figure 5.1: The x-axis shows the fraction of missed GS ≥7 cancers among men with PSA between 3 and 10, and the y-axis shows number of biopsies performed (blue and red lines) and the number of GS ≥7 cancers detected (grey line). Biopsy on all men in this PSA range detects (1) 156 GS ≥7 cancers and increasing the PSA threshold to (2) 3.2 and (3) 3.4 corresponds to missing 10% and 20% of these cancers and a total of 851 and 732 performed biopsies, respectively. Using S3M with a probability of GS ≥7 cancer of (6) 8% and (7) 11% also miss 10% and 20% GS ≥7 cancers but with 666 and 485 biopsies, respectively.

5.3 Study II

In **Study II** we digitized a large proportion of the biopsy cores collected in the STHLM3 study, with the overarching aim of building an AI for automatic diagnosis on prostate biopsies. In total, we digitized 8,980 biopsy cores, mostly from the STHLM3 cohort but we included slides from another lab mainly to increase cases with the rare, but very important, Gleason pattern 5. The extra cases were only used for training and not included in the validation data. The validation was performed on two data sets: a random selection of 246 of the STHLM3 subjects and an external data set of 73 subjects.

There are several aspects of histopathological diagnosis of prostate cancer, but arguably the most significant ones are detecting cancer, grade the cancer and measure the extent of the cancer. And it is these three tasks that we have focused on in this study. For detecting cancer in a biopsy core, we achieved an AUC of 0.997 on the STHLM3 test data which means almost perfect discrimination between cancer and benign cores

for any reasonable choice operating point. When we aggregated the detection to a patient level the AUC was 0.999. The corresponding AUCs dropped somewhat when we evaluated the algorithm on external data to 0.986 and 0.979, respectively. Even if these AUCs are unquestionably high, they need to be in this magnitude to not risk patient safety. We can look in more details of the consequences of the missed cancers once we choose an operating point (i.e. a threshold for positivity on the predicted probability for cancer). For this we consider Figure 5.2. Starting with the SHLM3 data we can choose an operating point that result in a sensitivity of, say, 99.3%. With that sensitivity we missed 5 cancer cores of the 721 cancer cores in the test set. Four of these were ISUP 1 (1.1% of all ISUP 1 in the test set) and 1 was ISUP 2 (0.7% of all ISUP 2 in the test set). At this operating point the AI did not fail to detect any subject with cancer. This is because a subject has more than one sampled core at a time of biopsy, and in these cases the men who had a core falsely negative had other cores that were classified positive. The specificity of the AI for this operating point was 88.9% meaning that 809 out of the 910 benign cores was removed from 'human' pathological assessment. Correspondingly for the external data, we can choose an operating point that result in a sensitivity of 98.6% and a specificity of 87.0%. In such case we miss 3 out of the 65 cores which was graded ISUP 1 by the pathologist, and no cores with higher grade. One subject with ISUP 1 disease was falsely classified as healthy.

	Avoided benign biopsy cores, n (specificity)	Detected cancer biopsy cores, n (sensitivity)	Missed cores with cancer by ISUP score, n(%)					Missed men with cancer, n (%)
			ISUP 1	ISUP 2	ISUP 3	ISUP 4	ISUP 5	
Independent test dataset								
Example operating point 1—sensitivity ≥99·9	570 (62·6%)	720 (99·9%)	0	1 (0·7%)	0	0	0	0
Example operating point 2—sensitivity ≥99·6	788 (86·6%)	718 (99·6%)	2 (0·6%)	1 (0·7%)	0	0	0	0
Example operating point 3—sensitivity ≥99·3	809 (88·9%)	716 (99·3%)	4 (1·1%)	1 (0·7%)	0	0	0	0
Example operating point 4—sensitivity ≥99·0	864 (94·9%)	714 (99·0%)	4 (1·1%)	2 (1·4%)	0	0	1 (1·4%)	1 (0·5%)
External validation								
Example operating point 1—sensitivity ≥99·5	49 (45·4%)	221 (99·5%)	1 (1·5%)	0	0	0	0	1 (1·8%)
Example operating point 2—sensitivity ≥99·1	78 (72·2%)	220 (99·1%)	2 (3·1%)	0	0	0	0	1 (1·8%)
Example operating point 3—sensitivity ≥98·6	94 (87·0%)	219 (98·6%)	3 (4·6%)	0	0	0	0	1 (1·8%)
Example operating point 4—sensitivity ≥97·7	97 (89·8%)	217 (97·7%)	3 (4·6%)	1 (1·6%)	1 (2·0%)	0	0	1 (1·8%)

Presented for each operating point are the number of benign biopsy cores that could be discarded from further consideration (specificity), the number of correctly detected malignant biopsy cores needing pathological evaluation (sensitivity), the number of missed malignant cores by ISUP score (percentage of all cores with the given ISUP score), and the number of missed men (percentage of all men with cancer). ISUP=International Society of Urological Pathology.

Table 3: **Sensitivity and specificity at selected points on the receiver operating characteristic curves for cancer detection**

Figure 5.2: A table of the implications of choosing various operating points as positivity criteria

It is important to report the total cancer burden, measured as the cancer length in the

biopsies along the length of the cores. On this task we achieved a correlation with the study pathologist measurements of 0.98 on the STHLM3 data and 0.94 on the external data, results that should be well within the accepted accuracy for clinical utility.

The hardest task was to learn and evaluate the ISUP grading. Since there is a very large disagreement between pathologists in this task, it is very difficult to define a ground truth and to define what is good enough results. For these reasons we evaluated the AI grading on Imagebase, a reference database with gradings by 23 internationally recognized pathologists and related the AI performance to the performance of the individual pathologists. The measure for evaluation was average agreement with the other pathologists measured with Cohan's Kappa with linear weights. The AI had an average kappa of 0.62 which was within the range of the pathologists (0.60-0.73).

5.4 Study III

Perineural invasion (PNI) is a major pathway for cancer to migrate from its organ of origin. However, there has been a debate of the clinical relevance of PNI for prostate cancer, and whether the presence of PNI in prostate biopsies is associated with poorer prognosis. A possible explanation for this confusion is the difficulty to define a large enough study cohort with power to reject the null hypothesis, in combination with a common misconception that not rejecting the null is evidence for the null. Loeb et al. conducted a study with 1,256 subjects by which multivariable adjusted PNI failed to reach statistical significance for predicting biochemical recurrence. [36] In an editorial comment one of the authors state "...we were able to show that although patients with perineural invasion were more likely to have extraprostatic extension, this was not an independent predictor of biochemical failure". [37] Even though the confidence covered unity, the point estimate suggested at least a 50% increase in the rate of relapse. In our study with we combined the evidence from four studies with similar design, methods and research question, including the study above and an analysis on the STHLM3 data. All four studies where roughly of the same size and with similar point estimate; two which could reject the null and two which could not. However, the combined estimate showed very strong support for an independent prognostic value of PNI in prostate biopsies.

5.5 Study IV

In **Study IV** we developed an AI for detecting PNI in prostate biopsies. We used the slides from STHLM3 that we had digitized in the **study II** and searched the pathology reports for cases with PNI that had not been digitized in the previous study. These were then digitized and all cases of PNI from the pathology reports were re-assessed for PNI

by the study pathologist L.E. In addition to verify the presence of PNI he also outlined the regions of each lesion of PNI in the slides. In total we used 8,803 slides of which 485 where positive for PNI based on the re-assessment of the slides.

The AI achieved an AUC of 0.98 (95% CI: 0.97-0.99) for discriminating between PNI positive and negative slides. For the chosen operating point, this corresponded to a sensitivity of 0.87 and a specificity of 0.97, see Figure 5.3. This is somewhat lower sensitivity than what we would like, but the operating point was decided on prior to the evaluation on the test set, to facilitate a truly independent validation. The positive and negative predicted values were 0.67 and 0.99, respectively. Since the negative predictive value is high, we can accept a somewhat lower positive predictive value with the caveat that the positive predicted cases should be verified by a pathologist. With the relatively rare case of PNI in a slide, this approach has potential to substantially reduce the workload of the pathologist for detecting and diagnosing PNI.

	Operating point	Sensitivity	Specificity	PPV	NPV	Accuracy
Cores	0.99	0.82	0.98	0.78	0.99	0.97
Postive n = 106	0.95 (index test)	0.87	0.97	0.67	0.99	0.97
Negative n = 1652	0.90	0.92	0.96	0.60	0.99	0.96
	0.85	0.92	0.95	0.54	0.99	0.95

Figure 5.3: Diagnostic properties of the model. PPV = Positive predictive value, NPV = Negative predictive value

For the task of highlighting each unique focus of PNI, we used IoU as the target parameter. Specifically, we calculated IoU for each slide positive for PNI and averaged the results across the slides. The average IoU was 0.5 (95% CI: 0.46-0.55).

Chapter 6

Discussion

In this thesis we have mainly focused on diagnostics of prostate cancer, even though **study I** focused on screening for prostate cancer by a using a multivariate diagnostic prediction model and **study III** estimated the prognostic value of PNI to assess the usefulness of reporting it with the diagnosis. In particular, the thesis relates to prediction models and AI solutions to some of the challenges of prostate cancer in the clinic: poor discriminative ability of the PSA test, too many unnecessary biopsies, high subjectivity and uncertainty in the grading, and increasing work load for a decreasing work force of uro-pathologists. We are still very early in the AI era for pathology, but with the rapid digitization of pathology labs and the recent improvements with AI techniques, software and hardware, we are bound to see many improvements in the coming years.

Despite promising results – not only by our AI but also the AI developed by Wouter *et. al.* which was published in the same issue of Lancet Oncology – there are still much to be done before we will see a fully autonomous diagnostic AI in the clinic. [38] The main challenges now are to ensure that such an AI generalizes across different labs, countries, and scanners, and that we can ensure quality control to avoid erroneous predictions if the AI encounter something new or unexpected, say, a new component in the stain or pen marks on top of the tissue. For generalizability we have initiated a project, the OncoWatch project, where we collect thousands of samples from 10 European countries. Some of these samples will be used to further train the AI and others to validate the performance in novel environments. For quality control we have initiated several studies, such as collecting samples with rare morphologies, providing uncertainty estimates along with the predictions, and evaluating the impact of the choice of scanner. To evaluate the latter, we are currently scanning a large sample of slides on three different scanners, not only to evaluate the loss in performance on a novel scanner but also estimating what measures are needed to ensure generalizability. Is it enough with data augmentation and color matching, or do we need to include the scanner in the training? And if so, to what degree?

Besides these crucial challenges, we also need to continue to strive for improvements of the AI. Two ways to do this is to explicitly target more features relevant for pathological assessment and to further train the AI on the tasks it already performs. An example of the former is the explicit training of detecting PNI in study IV. Another task that could be useful is to detect high grade prostate intraepithelial neoplasm (HGPIN), which is a pre-cancer and therefore treated as benign by the AI. Even if this is correct, it would be useful to provide the information of a pre-cancer in the assessment. Then there are other important morphologies such as cribriform and comodonecrosis, but since these are implicit in the Gleason grading system (grade 4 and 5, respectively), they are arguably not need to be targeted explicitly by the AI. For the improvement of computer assisted grading, we are at the time of this writing about to launch a competition together with Wouter *et. al.* on the prediction competition platform Kaggle, where we have combined the data sets used in both the studies published in Lancet Oncology 2020. We anticipate that hundreds of teams will participate from a wide range of backgrounds. This may prove fruitful, not only since the size of the data is doubled, but also since there are so much room for creativity in creating these models and the computational resources prohibit a single research group to explore more than a fraction of all approaches that may or may not lead to improvements.

The article from study IV has not yet been submitted for peer review, but there is a pre-print on ArXiv. Before submission, we will address a few shortcomings. The main extension is to show that the AI generalizes to external data. For this we are currently collecting one hundred images from Dr Toyonori Tsuzuki, Japan, (new lab and new scanner) of which approximately half are positive for PNI. Another question is how the results relate to human pathologists' performance, and whether the concordance of the AI with expert uro-pathologists is within the range of the concordance between the pathologists themselves. Prostate PNI diagnostic accuracy or agreement has to the best of our knowledge never been evaluated, and for this we have selected all positive cases in the test set and a random selection of equally many benign cores (in total over 200 samples) which will be independently assessed for PNI by four expert pathologists: Brett Delahunt (New Zeeland), Hemamali Samaratunga (Australia), Toyonori Tsuzuki (Japan), and Lars Egevad (Sweden). Another shortcoming of the study is the difficulty of assessing the performance of pixel-wise localization of PNI. For this we have used average IoU across positive biopsy cores, but this metric has some disadvantages. First, it does not address the false positive pixels in negative cores. Second, it does not consider that each PNI in a core is a unique lesion, and that they very much vary in size. The latter feature has implications such that the IoU benefit from focusing on large PNI lesions. If a core contains, say, one small and one large PNI, the metric is higher if most of the larger PNI is located and the small PNI is overlooked, compared to if both lesions have about half predicted localization each. Despite these shortcomings, we

argue that this metric is a good balance of relevant qualities and interpretability.

Finally, why do we train our models against Gleason labels – a subjective proxy endpoint for prostate cancer prognosis with large uncertainty – when we could train directly against survival (prostate specific cancer death or a surrogate such as biochemical relapse or metastasis)? There are several reasons for this! The successes of deep convolutional neural networks have to a large extent depended on huge collections of labeled samples. For pixel-labeled (or core labeled) data we can extract hundreds or even thousands or labeled images per core and with 10 to 12 cores per subject we have plenty of labeled images in a single subject. But since a subject only dies once, all images belonging to this subject will have to be attributed to the same outcome. Also, prostate cancer is a slow growing disease and it can take 20 years or longer for death to occur. This is of course coupled with many challenges of competing causes of death and censored data. And not least that the technology of scanners, labs, and treatments changes over time so the AI will be trained on samples that may not be so relevant 20 years later. Only relying on retrospective data has disadvantages such that the tissue may decade over time, and prospectively collected data will only be useful very long time from now. Another challenge to overcome is that all patients today are treated based on the current Gleason system, with the radicality of the treatments depending heavily on the ISUP grade. This is not only a challenge for building and training an AI but also for evaluating it. To do so properly, we would need to treat hundreds, or thousands, of patents based on diagnoses assigned by the AI system. This would have severe ethical implication and would be very time consuming, since it takes at least ten to fifteen years to evaluate prostate cancer. Despite all these challenges, there are enormous potential in such approaches due to the shortcomings of the Gleason system. And with prostate cancer being one of the major cancers with millions of men effected, there are certainly strong reasons to work in this direction.

This is certainly an exciting development with AI specialists and pathologists working together to ensure safe diagnosis and better prognostication, with the goal of better choice of treatment and, with that, decreased mortality from prostate cancer.

Chapter 7

Ethical considerations

In the first study, the update of S3M, we explicitly included HOXB13, a relatively common mutation in a nucleotide which is strongly associated with prostate cancer. This led to a Letter to the Editor in European Urology with some questions regarding genetic consulting prior to the test and of who should receive the information and by whom. We replied that we welcome such a discussion and until there are national guidelines regarding this, we do not report back carriership information to the index person, we only use it to improve his risk assessment. However, this information will likely play an important role in individualizing the recommended time until future screenings among men who not yet developed (detectable) cancer. There are guidelines for conveying genetic information in general, but these include sensitive genetic information where stigmatization, discrimination or depression may follow.[39] Here the starting point is that genetic information should be conveyed to the index person by medical professionals if wished for, and the index person in turn decides if relatives who also may be exposed should be informed. This is a discussion we will likely see more of in the future and is relevant with the increase of data intense and automatic procedures in medicine.

Regarding artificial intelligence in medicine, there are many ethical issues that needs to be addressed. Likely, we have not yet formulated all questions that concerns ethical issues with this new and powerful technology. Even though our research is translational in its nature, we are still at a proof-of-principle stage, and there are no direct ethical issues related to artificial intelligence in medicine for these projects. There are of course the usual ethical issues when working with sensitive data.

As researchers in this new field, it is important to reflect on the ethical issues that may spring for a success in these proof-of-principle studies. Among these is the question of responsibility. For example, if we implement an end-to-end algorithm for diagnosis in an open source pathology software and a hospital in a country with no available pathologists use it and it leads to a false put diagnosis, who is responsible?

Acknowledgements

To my main supervisor, **Martin Eklund**. It has truly been a pleasure working with you for these four years. You have been an inspiration, not only as a researcher, but as a college, leader and friend! I hope you continue the hard and thorough work you do. Not only will research benefit greatly, but you will give so many students an inspiring and rewarding journey. We have had many interesting discussions about anything and everything – and I look forward to many more in the years to come.

To co-supervisor **Tobias Nordström**, for the ton of positive energy you bring, the dedicated and important work you do for research and for all the fun times!

To co-supervisor **Matias Rantalainen**, for being an ambitious, talented and hard working researcher. You make the hardest tasks appear achievable and I have really enjoyed all the discussions we have had.

To co-supervisor **Mark Clements**, for always raising the bar of statistical quality no matter what the cost in terms of comprehensibility. You are a truly great statistician and researcher who cares a lot about others and always makes us smile.

To my mentor **Sven Sandin**, for knowing that you always got my back! I have very much enjoyed our talks and you always have good advice.

To **Henrik Grönberg**, for the fantastic research environment you have created and the inspiration you are for what is achievable!

To **Lars Egevad**, for the fruitful collaboration and all the interesting conversations we had. All I know about pathology I have learnt from you. We would never have achieved what we did without your astonishing competence and dedicated work.

To **Brett Delahunt, Hemamali Samaratunga and Toyonori Tsuzuki**, for without hesitation offering to participate in our projects. Your high quality assessment and deep knowledge has generated very valuable data for research.

To **Kimmo Kartasalo**, I could not have whished for a better research partner and friend! Never have I worked so closely, fruitfully and enjoyable with anyone before. We still have more work to do!

To **Henrik Olsson**, for being the best of friends! You are hardworking, helpful and you make it a pleasure coming to the office. I wish you and your new family the best, and I hope to see you often in the coming years.

To **Andreas Karlsson**, you are an amazingly good friend who always make time to listen and make a serious effort to understand the core of a research problem or a technical problem whenever I have asked you. I learnt a lot form you and it has always been a pleasure being your friend.

To **Alessio Crippa, Andrea Discacciati, Thorgerdur Palsdottir, Robert Karlsson, Elisabeth Dahlqwist, Jonas Ludvigsson, Alexander Ploner, Henric Winell and Therese Andersson** – you are fantastic researchers, kind and sharing!

To **Ola and Martin Steinberg, Karl Andersson, Erik Berner, Astrid Björklund, Carin Cavalli-Björkman and Britt-Marie Hune** for believing in our research!

To **Pekka Ruusuvuori and Carolina Wählby** for your sharing your knowledge in image analysis.

To the biostatistics professors **Paul Dickman, Keith Humphreys, Yudi Pawitan, Marie Reilly, and Juni Palmgren** for creating such a wonderful environment!

To the applied biostatistics group **Cecilia Lundholm, Mikael Andersson Franko, Andrea Discacciati, Andreas Karlsson, Robert Karlsson, Ralf Kuja-Halkola, Michael Sachs, Sven Sandin, Agnieszka Szwajda, Henric Winell, and Li Yin** – I learnt a lot from you. The best things have been how to solve real world problems with real world data. It prepared me greatly for my PhD studies.

To my friends and colleges **Mortezavi Ashkan, Venkatesh Chellappa, Anna Lantz, Johan Lindberg, Rebecca Bergström, Berit Larsson, Bram De Laere, Yinxi Wang, Bojing Liu, Philippe Weitz, Paul Lambert, Michael Crowther, Rino Bellocco, Arvid Sjölander, Erin Gabriel, Nghia Trung Vu, Maya Alsheh Ali, Anna Johansson, Bénédicte Delcoigne, Caroline Weibull, Cecilia Radkiewicz, Gabriel Isheden, Kathleen Bokenberger, Johan Zetterquist, Daniela Mariosa, Xingrong Liu, Hannah Bower, Robert Szulkin, Linda Abrahamsson, Nurgul Batyrbekova, Wenjiang Deng, Shuang**

Hao, Zheng Ning, Ninoa Malki, Frida Lundberg, Tong Gong, Rickard Strandberg, Rikard Öberg, Marie Janson, Alessandra Nanni, Jacqueline Knight, Anna Berglund, Gunilla Sonnebring, Flaminia Chiesa, Tor-Arne Hegvik, Alessandra Grotta, Pär Sparrén, and Jiangrong Wang, for all the various ways you have brought joy and knowledge during my time at MEB.

To my parents **Karin and Ingemar Ström** for always having a hot stew ready whenever we retreat to the beautiful Dalsland. You gave me the best of childhoods and now you give Måns and Svante all that love too!

To **Eva Ström, Aina Ström, Anders Nödtveidt, Tony Ström, Hjördis Karlsson and Arne Karlsson**, I know you would have come to the defense if it was possible. Your support is much appreciated!

Last but not least, to my family **Sara, Måns, and Svante**, without whom I would be lost. With you every day is a joy. Thank you for being the beautiful people you are for the love you give – this one is for you!

This work was supported **Cancerfonden** (the Swedish Cancer Society), **VR** (Vetenskaps Rådet) and **FORTE** (Forskningsrådet för hälsa, arbetsliv och välfärd).

References

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

[2] H. Lilja. A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein. *Journal of Clinical Investigation*, 1985.

[3] Thomas A. Stamey, Norman Yang, Alan R. Hay, John E. McNeal, Fuad S. Freiha, and Elise Redwine. Prostate-Specific Antigen as a Serum Marker for Adenocarcinoma of the Prostate. *New England Journal of Medicine*, 1987.

[4] William J. Catalona, Alan W. Partin, Kevin M. Slawin, Michael K. Brawer, Robert C. Flanigan, Anup Patel, Jerome P. Richie, Jean B. DeKernion, Patrick C. Walsh, Peter T. Scardino, Paul H. Lange, Eric N.P. Subong, Robert E. Parson, Gail H. Gasior, Kathleen G. Loveland, and Paula C. Southwick. Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: A prospective multicenter clinical trial. *Journal of the American Medical Association*, 1998.

[5] Hans Lilja, David Ulmert, and Andrew J. Vickers. Prostate-specific antigen and prostate cancer: Prediction, detection and monitoring, 2008.

[6] H. Grönberg, J. Adolfsson, M. Aly, T. Nordström, P. Wiklund, Y. Brandberg, J. Thompson, F. Wiklund, J. Lindberg, M. Clements, L. Egevad, and M. Eklund. Prostate cancer screening in men aged 50-69 years (STHLM3): a prospective population-based diagnostic study. *Lancet Oncol.*, 16(16):1667–1676, Dec 2015.

[7] Dipen J. Parekh, Sanoj Punnen, Daniel D. Sjoberg, Scott W. Asroff, James L. Bailen, James S. Cochran, Raoul Concepcion, Richard D. David, Kenneth B. Deck, Igor Dumbadze, Michael Gambla, Michael S. Grable, Ralph J. Henderson, Lawrence Karsh, Evan B. Krisch, Timothy D. Langford, Daniel W. Lin, Shawn M. McGee, John J. Munoz, Christopher M. Pieczonka, Kimberley Rieger-Christ, Daniel R. Saltzstein, John W. Scott, Neal D. Shore, Paul R. Sieber, Todd M. Waldmann, Fredrick N. Wolk, and Stephen M. Zappala. A Multi-institutional Prospective Trial in the USA Confirms that the 4Kscore Accurately Identifies Men with High-grade Prostate Cancer. *European Urology*, 2015.

[8] Claire De La Calle, Dattatraya Patil, John T. Wei, Douglas S. Scherr, Lori Sokoll, Daniel W. Chan, Javed Siddiqui, Juan Miguel Mosquera, Mark A. Rubin, and Martin G. Sanda. Multicenter evaluation of the prostate health index to detect aggressive prostate cancer in biopsy Naïve men. *Journal of Urology*, 2015.

[9] Ina L. Deras, Sheila M.J. Aubin, Amy Blase, John R. Day, Seongjoon Koo, Alan W. Partin, William J. Ellis, Leonard S. Marks, Yves Fradet, Harry Rittenhouse, and Jack Groskopf. PCA3: A Molecular Urine Assay for Predicting Prostate Biopsy Outcome. *Journal of Urology*, 2008.

[10] Monique J. Roobol, Jan F.M. Verbeek, Theo van der Kwast, Intan P. Kümmerlin, Charlotte F. Kweldam, and Geert J.L.H. van Leenders. Improving the Rotterdam European Randomized Study of Screening for Prostate Cancer Risk Calculator for Initial Prostate Biopsy by Incorporating the 2014 International Society of Urological Pathology Gleason Grading and Cribriform growth. *European Urology*, 2017.

[11] V. Kasivisvanathan, M. Emberton, and C. M. Moore. MRI-Targeted Biopsy for Prostate-Cancer Diagnosis. *N. Engl. J. Med.*, 379(6):589–590, 08 2018.

[12] L. Egevad, F. Algaba, D. M. Berney, L. Boccon-Gibod, D. F. Griffiths, A. Lopez-Beltran, G. Mikuz, M. Varma, and R. Montironi. Handling and reporting of radical prostatectomy specimens in Europe: A web-based survey by the European Network of Uropathology (ENUP). *Histopathology*, 2008.

[13] Jonathan I. Epstein, William C. Allsbrook, Mahul B. Amin, Lars L. Egevad, Sheldon Bastacky, Antonio López Beltrán, Aasmund Berner, Athanase Billis, Liliane Boccon-Gibod, Liang Cheng, Francisco Civantos, Cynthia Cohen, Michael B. Cohen, Milton Datta, Charles Davis, Brett Delahunt, Warick Delprado, John N. Eble, Christopher S. Foster, Masakuni Furusato, Paul B. Gaudin, David J. Grignon, Peter A. Humphrey, Kenneth A. Iczkowski, Edward C. Jones, Scott Lucia, Peter A. McCue, Tipu Nazeer, Esther Oliva, Chin Chen Pan, Galina Pizov, Victor Reuter, Hemamali Samaratunga, Thomas Sebo, Isabell Sesterhenn, Maria Shevchuk, John R. Srigley, Sueli Suzigan, Hiroyuki Takahashi, Pheroze Tamboli, Puay Hoon Tan, Bernard Tètu, Satish Tickoo, John E. Tomaszewski, Patricia Troncoso, Toyonori Tsuzuki, Lawrence D. True, Theo Van Der Kwast, Thomas M. Wheeler, Kirk J. Wojno, and Robert H. Young. The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *American Journal of Surgical Pathology*, 29(9):1228–1242, 2005.

[14] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. The 2014 international society of urological pathology

(isup) consensus conference on gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, 40(2):244–252, 2016.

[15] Cancer i siffror 2018. http://socialstyrelsen.se. Online accessed: 24.03.2020.

[16] Regionala Cancercentrum i Samverkan. Prostatacancer. *Nationellt vårdprogram*, 2017.

[17] Eva Hollemans, Esther I Verhoef, Chris H Bangma, John Rietbergen, Monique J Roobol, Jozien Helleman, and Geert J L H van Leenders. Clinical outcome comparison of Grade Group 1 and Grade Group 2 prostate cancer with and without cribriform architecture at radical prostatectomy. *Histopathology*, jan 2020.

[18] P. Ström, T. Nordström, B. Delahunt, H. Samaratunga, H. Grönberg, L. Egevad, and M. Eklund. Prognostic value of perineural invasion in prostate needle biopsies: a population-based study of patients treated by radical prostatectomy. *J. Clin. Pathol.*, Feb 2020.

[19] F. C. Hamdy, J. L. Donovan, J. A. Lane, and et. al. 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. *N. Engl. J. Med.*, 375(15):1415–1424, 10 2016.

[20] Primär behandling av prostatacancer utan spridning. https://kunskapsbanken.cancercentrum.se/diagnoser/prostatacancer/vardprogram/primar-behandling-av-prostatacancer-utan-spridning/. Online accessed: 25.03.2020.

[21] J. R. Rider, F. Sandin, O. Andreasson, P. Wiklund, J. Hugosson, and P. Stattin. Long-term outcomes among noncuratively treated men according to prostate cancer risk category in a nationwide, population-based study. *Eur. Urol.*, 63(1):88–96, Jan 2013.

[22] Hillary M. Ross, Oleksandr N. Kryvenko, Janet E. Cowan, Jeffry P. Simko, Thomas M. Wheeler, and Jonathan I. Epstein. Do adenocarcinomas of the prostate with gleason score (GS)6 have the potential to metastasize to lymph nodes?, 2012.

[23] Melissa Assel, Anders Dahlin, David Ulmert, Anders Bergh, Pär Stattin, Hans Lilja, Andrew J Vickers, and Giacomo Novara. Association Between Lead Time and Prostate Cancer Grade: Evidence of Grade Progression from Long-term Follow-up of Large Population-based Cohorts Not Subject to Prostate-specific Antigen Screening. 2017.

[24] Eva Haglind, Stefan Carlsson, Johan Stranne, Anna Wallerstedt, Ulrica Wilderäng, Thordis Thorsteinsdottir, Mikael Lagerkvist, Jan Erik Damber, Anders Bjartell, Jonas Hugosson, Peter Wiklund, and Gunnar Steineck. Urinary Incontinence

and Erectile Dysfunction after Robotic Versus Open Radical Prostatectomy: A Prospective, Controlled, Nonrandomised Trial. *European Urology*, 2015.

[25] Peihua Qiu. The Statistical Evaluation of Medical Tests for Classification and Prediction. *Journal of the American Statistical Association*, 2005.

[26] Frank E. Harrell, Kerry L. Lee, Robert M. Califf, David B. Pryor, and Robert A. Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 1984.

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

[28] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A W M van der Laak, the CAMELYON16 Consortium, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, Oscar Geessink, Nikolaos Stathonikos, Marcory Crf van Dijk, Peter Bult, Francisco Beca, Andrew H Beck, Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Aoxiao Zhong, Qi Dou, Quanzheng Li, Hao Chen, Huang-Jing Lin, Pheng-Ann Heng, Christian Haß, Elia Bruni, Quincy Wong, Ugur Halici, Mustafa Ümit Öner, Rengul Cetin-Atalay, Matt Berseth, Vitali Khvatkov, Alexei Vylegzhanin, Oren Kraus, Muhammad Shaban, Nasir Rajpoot, Ruqayya Awan, Korsuk Sirinukunwattana, Talha Qaiser, Yee-Wah Tsang, David Tellez, Jonas Annuscheit, Peter Hufnagl, Mira Valkonen, Kimmo Kartasalo, Leena Latonen, Pekka Ruusuvuori, Kaisa Liimatainen, Shadi Albarqouni, Bharti Mungal, Ami George, Stefanie Demirci, Nassir Navab, Seiryo Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Hady Ahmady Phoulady, Vassili Kovalev, Alexander Kalinovsky, Vitali Liauchuk, Gloria Bueno, M Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Daniel Racoceanu, and Rui Venâncio. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 2017.

[29] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-van de Kaa, Peter Bult, Bram van Ginneken, and Jeroen van der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6:26286, 2016.

[30] AI diagnostics need attention. *Nature*, 555(7696):285, 2018.

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[32] Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2003.

[33] Tianqi Chen and Carlos Guestrin. XGBoost, 2016.

[34] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.

[35] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[36] Stacy Loeb, Jonathan I. Epstein, Elizabeth B. Humphreys, and Patrick C. Walsh. Does perineural invasion on prostate biopsy predict adverse prostatectomy outcomes? *BJU International*, 2010.

[37] Patrick C Walsh. Re: Does Perineural Invasion on Prostate Biopsy Predict Adverse Prostatectomy Outcomes? *The Journal of Urology*, 185(2):515–516, 2011.

[38] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. Hulsbergen-van de Kaa, and G. Litjens. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.*, 21(2):233–241, Feb 2020.

[39] Niklas Juth. The Right Not to Know and the Duty to Tell: The Case of Relatives. *Journal of Law, Medicine and Ethics*, 2014.