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

Personalized prostate cancer management: AI-assisted prostate pathology and improved active surveillance

Henrik Olsson

Karolinska Institutet

Stockholm 2023

All published papers reproduced with permission
Published by Karolinska Institutet
Printed by US-AB 2023

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

ISBN 978-91-8016-912-7

Institutionen för Medicinsk Epidemiologi och Biostatistik

Personalized prostate cancer management: AI-assisted prostate pathology and improved active surveillance

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

Fredag 3 februari 2023, kl 09.00

By
Henrik Olsson

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

Co-supervisor:
Associate Professor Anna Lantz
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Co-supervisor:
Associate Professor Mark Clements
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Opponent:
Associate Professor Maarten van Smeden
University Medical Center Utrecht
Department of Epidemiology

Examination board:
Associate Professor Rebecka Godtman
University of Gothenburg
Department of Urology

Associate Professor Rino Bellocco
Karolinska Universitet
Department of Medical Epidemiology and Biostatistics

Adjunct Professor Claes Lundström
Linköping University
Department of Science and Technology

Till mina älsklingar Matilda, Rut och Eva

Abstract

Prostate cancer is a major global health concern and is the most common cancer-related cause of death in Sweden. Prostate cancer screening using PSA has been shown to reduce prostate cancer mortality but also leads to significant overdiagnosis and overtreatment of low-risk cancers. Improved risk stratification and effective active surveillance are crucial to balancing the benefits of screening with the risk of overdiagnosis and overtreatment.

In **Study I**, we studied the uptake and the follow-up of active surveillance using a retrospective cohort of patients who were diagnosed with low-risk prostate cancer between 2008 and 2017 in Stockholm County. Our results showed that only 50% of eligible active surveillance patients received active surveillance as their primary treatment choice at diagnosis. Most men that enrolled in active surveillance remained on surveillance during the first years after diagnosis (82% during a median 3.5 years), but did not receive a follow up according to guidelines with regard to repeat biopsies and PSA tests.

Current clinical practice has seen an increase in the use of magnetic resonance imaging (MRI) and the incorporation of risk prediction models to select men with the highest suspicion of clinically significant prostate cancer for prostate biopsy. However, the effectiveness and how MRI and risk prediction models should be incorporated into active surveillance follow-up have yet to be established. **Study II** evaluated the performance of MRI-targeted biopsies and a blood-based risk prediction model (the Stockholm3 test) for monitoring disease progression in patients on active surveillance and compared this to the conventional follow-up using PSA and systematic biopsies. When MRI-targeted and systematic biopsies were combined, the detection rate of clinically significant prostate cancer increased when compared to conventional systematic biopsies. Biopsies performed in MRI-positive men resulted in a 49% reduction in performed biopsies, at the expense of failing to diagnose 1.4% clinically significant prostate cancer in MRI-negative men. The incorporation of the Stockholm3 test showed a 27% reduction in required MRI investigations and a 57% reduction in performed biopsies compared to performing only systematic biopsies.

In **Study III**, we digitized biopsy cores from STHLM3 participants to develop an artificial intelligence (AI) for prostate cancer diagnostics. The AI system demonstrated clinically useful performance that was comparable to that of the study pathologist for cancer detection (AUC of 0.986) and for predictions of cancer length (correlation of 0.87) and grading performance that was on par with that of expert prostate pathologists.

In **Study IV**, we developed a conformal predictor to estimate the uncertainty of the predictions for the model in Study III. The uncertainty estimates were used to control the error rate so that only predictions with high confidence are accepted and unreliable predictions can be detected. The conformal predictor was able to identify unreliable predictions as a result of variations in

digital pathology scanners, preparation of tissue in different pathology laboratories, and the existence of unusual prostate tissue that the AI model was not exposed to during training.

Little is known about the relationships between prostate cancer genetic risk factors and the morphology of prostate tissue. In **Study V:**, we investigated whether weakly supervised deep learning can learn to detect such possible associations. The findings in this paper imply relationships between prostatic tissue morphology and genetic risk factors for prostate cancer, particularly in young men. These results provide proof of principle for exploring the use of morphological information in multi-modal prostate cancer risk prediction algorithms.

In conclusion, the purpose of this thesis was to describe possible extensions to improve prostate cancer active surveillance management, as well as to develop prediction models for improved prostate cancer diagnostics.

List of publications

I. **Henrik Olsson**, Tobias Nordström, Mark Clements, Henrik Grönberg, Anna Wallerstedth Lantz, Martin Eklund
Intensity of Active Surveillance and Transition to Treatment in Men with Low-risk Prostate Cancer
European Urology Oncology, 2020, Vol 3, 640-647

II. **Henrik Olsson**, Tobias Nordström, Fredrik Jäderling, Lars Egevad, Hari T. Vigneswaran, Magnus Annerstedt, Henrik Grönberg, Martin Eklund, Anna Lantz
Incorporating Magnetic Resonance Imaging and Biomarkers in ActiveSurveillance Protocols - Results From the Prospective Stockholm3Active Surveillance Trial (STHLM3AS)
JNCI J Natl Cancer Inst, 2021, Vol 113, 632-640

III. 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 R M 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, Martin Eklund
Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study
The Lancet Oncology, 2020, Vol 21, 222-232

IV. **Henrik Olsson**, Kimmo Kartasalo, Nita Mulliqi, Marco Capuccini, Pekka Ruusuvuori, Hemamali Samaratunga, Brett Delahunt, Cecilia Lindskog, Lars Egevad, Ola Spjuth, Martin Eklund
Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction
Nature Communications, in print

V. **Henrik Olsson**, Kimmo Kartasalo, Nita Mulliqi, Pekka Ruusuvuori,Anna Plym, Fredrik Wiklund, Hemamali Samaratunga, Brett Delahunt, Cecilia Lindskog, Lars Egevad, Martin Eklund,
Associations between prostate cancer genetic risk factors and prostatic tissue morphology
Manuscript

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

- Axel Möller, Henrik Olsson, Henrik Grönberg, Martin Eklund, Markus Aly, Tobias Nordström
  The Stockholm3 blood-test predicts clinically-significant cancer onbiopsy: independent validation in a multi-center community cohort
  Prostate Cancer and Prostatic Diseases 2019

- Kerri Beckmann, Netty Kinsella, Henrik Olsson, Anna Wallerstedt Lantz, Tobias Nordstrom, Markus Aly, Jan Adolfsson, Martin Eklund and Mieke Van Hemelrijck
  Is there any association between prostate-specific antigen screening frequency and uptake of active surveillance in men withlow or very low risk prostate cancer?
  BMC Urology 2019

- Kerri Beckmann, Danielle Crawley, Tobias Nordström, Markus Aly, Henrik Olsson, Anna Lantz, Noor Binti Abd Jalal, Hans Garmo, Jan Adolfsson, Martin Eklund, Mieke Van Hemelrijck
  Association between antidiabetic medications and prostate-specific antigen levels and biopsy results
  JAMA network open, 2019, Vol 2, e1914689-e1914689

- Martin Eklund, Kimmo Kartasalo, Henrik Olsson, Peter Ström
  The importance of study design in the application of artificial intelligence methods in medicine
  npj Digital Medicine, 2019, Vol 2, Article number 101

- Lars Egevad, Peter Ström, Kimmo Kartasalo, Henrik Olsson, Hemamali Samaratunga, Brett Delahunt, Martin Eklund
  The utility of artificial intelligence in the assessment of prostate pathology
  Histopathology, 2020, Vol 76, 790-792

- Lars Björnebo, Henrik Olsson, Tobias Nordström, Fredrik Jäderling, Henrik Grönberg, Martin Eklund, Anna Lantz
  Predictors of adverse pathology on radical prostatectomy specimen in men initially enrolled in active surveillance for low-risk prostate cancer
  World Journal of Urology 2021

- Hari T Vigneswaran, Thorgerdur Palsdottir, Henrik Olsson, Erik S Haug, Wolfgang Picker, Sven Löffeler, Henrik Grönberg, Martin Eklund, Tobias Nordström
  Biomarker discrimination and calibration with MRI-targeted biopsies: an analysis with the Stockholm3 test
  Prostate Cancer and Prostatic Diseases, 2021, Vol 24, 457-464

Contents

List of abbreviations

CI	Confidence Interval
GG	Gleason Grade
IQR	Inter Quartile Range
ISUP	International Society of Urological Pathology
PSA	Prostate-Specific Antigen
RP	Radical Prostatectomy
MRI	Magnetic Resonance Imaging
ROC	Receiver Operating Characteristics
AUC	Area Under Curve (Receiver Operating Characteristic)
PI-RADS	Prostate Imaging Reporting and Data System

Chapter 1

Introduction

After testing with the Prostate-Specific Antigen (PSA) blood test began, the incidence of known prostate cancer has increased greatly. PSA testing can lead to a reduced risk of dying from prostate cancer, as the disease can be diagnosed and treated at an early stage. Many of the tumors that are detected are small, localized, slow growing, and unlikely to cause symptoms to the diagnosed individual. A major problem today is that the current diagnostic and prognostic methods show a limited capacity to distinguish between indolent tumors and early-stage aggressive cancer. Men with low-risk prostate cancer are managed conservatively with active surveillance (AS), which involves regular PSA testing and systematic prostate biopsies to monitor the course of the disease and provide delayed, targeted treatment with a curative purpose.

The histological diagnosis of prostate biopsies determines the clinical treatment for men suspected of having prostate cancer. However, the current clinical practice is associated with challenges such as high inter-observer variability between pathologists and a worldwide shortage of experienced pathologists. Artificial intelligence (AI) and machine learning technologies in digital pathology have become available due to the possibility of digitizing whole-slide images of tissue. The development of AI has the potential to assist pathologists in diagnosis and aims to reduce inter- and intra-observer variability among pathologists.

This thesis aimed to improve risk stratification through the development of diagnostic prediction models and to assess possible improvements to active surveillance that are less invasive and more cost-effective, through the incorporation of risk prediction models and magnetic resonance imaging (MRI) targeted biopsies in the follow-up.

Chapter 2

Aims of the thesis

This thesis aims to improve patient screening through the development of diagnostic prediction models for improved risk stratification and possible improvements to active surveillance through the incorporation of risk prediction models and MRI-targeted biopsies in the follow-up.

More specifically, the aims were to:

- To study active surveillance in Sweden with respect to the uptake of low-risk prostate cancer patients, describe the follow-up of active surveillance patients, and study the transition from active surveillance to curative treatment during a 10-year follow-up period.
- To study the performance of MRI-targeted biopsies and the Stockholm3 test for monitoring disease progression in patients on active surveillance.
- Develop AI for digital pathology assessment of prostate biopsies for cancer detection and grading and to estimate the cancer length.
- Develop a conformal predictor for AI-assisted prostate pathology and estimate the uncertainty of the predictions for these models, to use the uncertainty estimates to control the error rate of the predictions such that unreliable predictions can be detected and flagged for human assessment.
- To study if weakly supervised deep learning can detect possible associations between genetic risk factors (SNPs) and morphological features in prostate tissue.

Chapter 3

Background

3.1 Epidemiology of Prostate Cancer

Prostate cancer is a major health problem. The lifetime risk for men to be diagnosed with prostate cancer is one out of nine and constitutes 21% of all diagnosed cancers [1]. It is Sweden's most common form of cancer among men and accounts for a third of all cancers in men. In 2017, approximately 10 000 men received a prostate cancer diagnosis and 2,500 died due to the disease. The median age at diagnosis is age 69 and the disease is uncommon before the age of 50. Prostate cancer incidence has rapidly increased with the introduction of PSA testing at the start of the 1990s, particularly with respect to men diagnosed with small localized tumors. The incidence has decreased slightly in later years, which could be due to the stabilization of PSA testing and recommendations against prostate cancer screening (Figure 3.1). The prostate cancer mortality rates vary less between developed regions. The highest mortality rates are reported in parts of Africa, Northern Europe, and South America, and lower mortality rates in Asia (Figure 3.2). The age-standardized mortality rate has decreased by 30 percent since 2000, which can be explained by early detection and improved treatment of the disease. With over 100,000 prevalent cases in Sweden, prostate cancer imposes a substantial health burden and cost on the healthcare system. Older age, family history of the disease, and ethnicity are recognized prostate cancer risk factors (higher risk in African men and lower risk in Asian men). Given the high prevalence and the lack of primary prevention of prostate cancer, improvement of diagnostic performance can help to substantially reduce the health burden of the disease with fewer men undergoing biopsy and less overdiagnosis and overtreatment.

3.2 Current prostate cancer diagnostics

3.2.1 PSA screening

Prostate cancer is usually diagnosed through assessment of PSA levels. The main role of PSA is to liquefy the seminal fluid, and it is found in seminal plasma at one million-

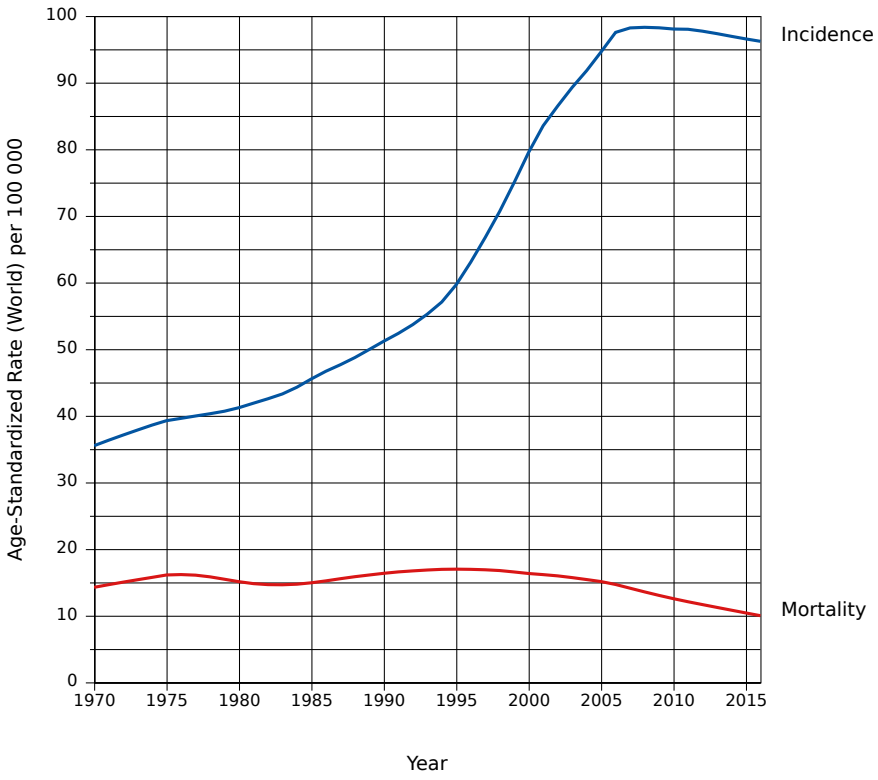
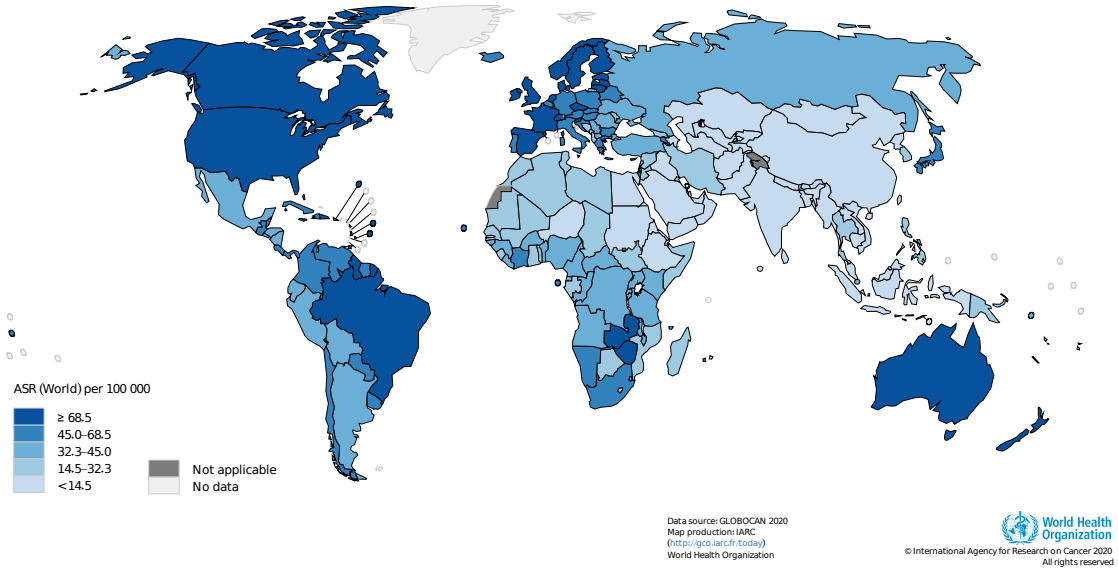


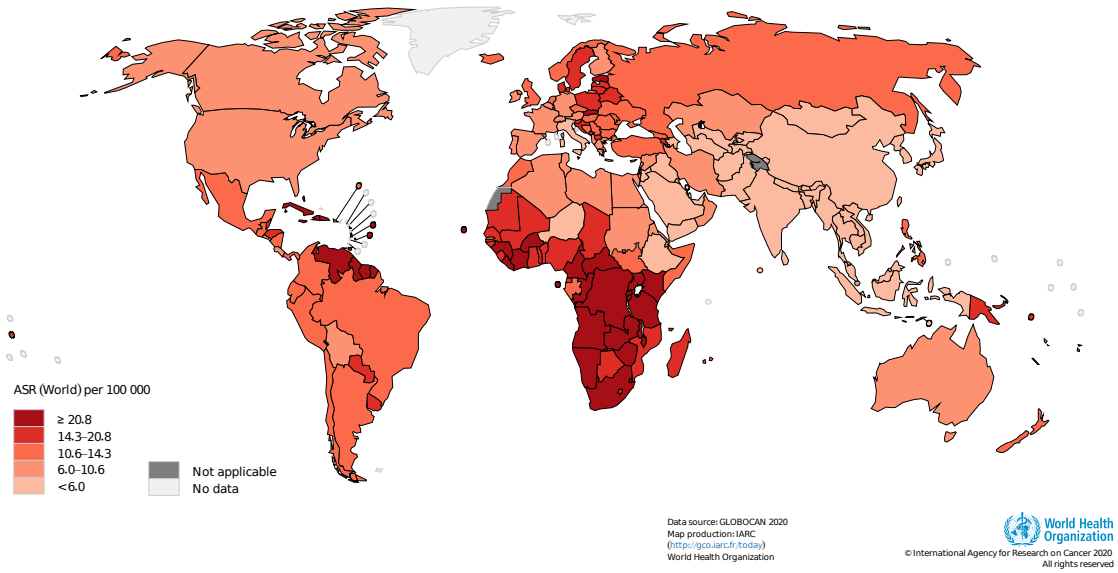
Figure 3.1: Age-standardized prostate cancer incidence and mortality rates per 100.000 men in Sweden, age [0-84]. Age-adjustment according to the world population Data from IARC (https://gco.iarc.fr/, accessed 22.09.2022)

fold times higher concentration than in blood [3]. Men with elevated PSA levels are recommended to undergo a prostate biopsy for further prostate cancer work-up.

PSA was originally intended for disease monitoring after diagnosis and initial treatment. However, the PSA test also quickly became common around the world as a tool for screening for prostate cancer, leading to a rapidly increasing prostate cancer incidence during the 1990ies, especially for men with localized low-grade cancer. Studies have shown that PSA-driven early detection reduces mortality [4, 5]. The test characteristics of the PSA test have been estimated to have a sensitivity of 72% and a specificity of 93% at the cutoff level of 4 ng/ml [6]. However, in Sweden, a lower cutoff of 3 ng/ml is used for men younger than 70 years, which results in higher sensitivity but at the cost of reduced specificity. The European Randomized study of Screening for Prostate Cancer (ERSPC) is a randomized screening trial evaluating PSA testing in eight European countries. Estimates from the ERSPC trial indicate that PSA screening leads to significant overtreatment, as 48 men would need to be treated to prevent one death from prostate cancer during a ten-year period [4, 7]. An annual 10% drop in prostate cancer incidence was observed in the United States between 2010 and 2014 as a result of the US Preventive Services Task Force recommendations against the use of PSA as a screening test. The 2012 recommendations concluded that the benefits of



(a) Prostate cancer incidence rates



(b) Prostate cancer mortality rates

Figure 3.2: Prostate cancer incidence and mortality rates around the world in 2020. Data from WHO. [2]

using PSA for screening of prostate cancer do not compensate for the harms with regard to overdiagnosis, overtreatment, and treatment-related problems. This was updated in 2018, recommending individual decisions for screening men under the age of 70.

3.2.2 Prostate biopsies

Men with an elevated risk of prostate cancer, commonly estimated using PSA and digital rectal exam, are typically recommended to undergo prostate biopsies. Although the biopsy procedure is currently quickly shifting to using magnetic resonance imaging (MRI) to guide the tissue sampling (see section 3.4.2), systematic biopsies where 10 to 12 biopsy cores are sampled with the use of transrectal ultrasound are still common in many countries around the world. The biopsy cores are then systematically sampled from the peripheral part of the prostate, which is the most common location of cancer in the prostate. Nonetheless, it is still not uncommon that systematic biopsies miss small cancer foci. The risk of non-representative biopsy findings has been shown to lead to high disease reclassification rates following radical prostatectomy [8, 9].

3.2.3 The Gleason grading system

Prostate cancer diagnosis is based on pathological evaluation of the tissue sampled from needle biopsies using the Gleason grading system. The system quantifies cell morphology into five different grades, where higher grades are associated with worse prognosis (Figure 3.3). An important cut-off prediction of prognosis with Gleason score is Gleason patterns 4 and 5, which corresponds to the least differentiated cells, meaning that most of the glandular structure in the tissue has been lost. Tumors graded with Gleason pattern 4 and 5 are much more aggressive and are strongly associated with higher risk of prostate cancer death compared to Gleason pattern 3 [10]. Since 2005 the International Society of Urological Pathology (ISUP) recommends that only grades 3-5 are used; grades 1 and 2 are not defined as cancer [11]. The primary grade, the most prevalent grade, and the secondary, the second most prevalent grade, are combined in a total score (e.g. 3 + 4 =7). This was revised again in 2014 and the updated ISUP grade group system was proposed. ISUP grade group 1 corresponds to Gleason score 3 + 3, ISUP 2 to Gleason score 3 + 4, ISUP 3 to Gleason score 4 + 3, ISUP 4 to Gleason score 4 + 4 and ISUP 5 corresponds to Gleason score 4 + 5 or higher.

3.2.4 Genetics

Familial history of prostate cancer is a strong risk factor for the disease. Men with a diagnosed first-degree relative are 2-3 times more likely to be affected than men without prostate cancer in the family, and the risk also increases with earlier age at onset of the relatives [12, 13]. But the genetics behind hereditary predisposition to prostate

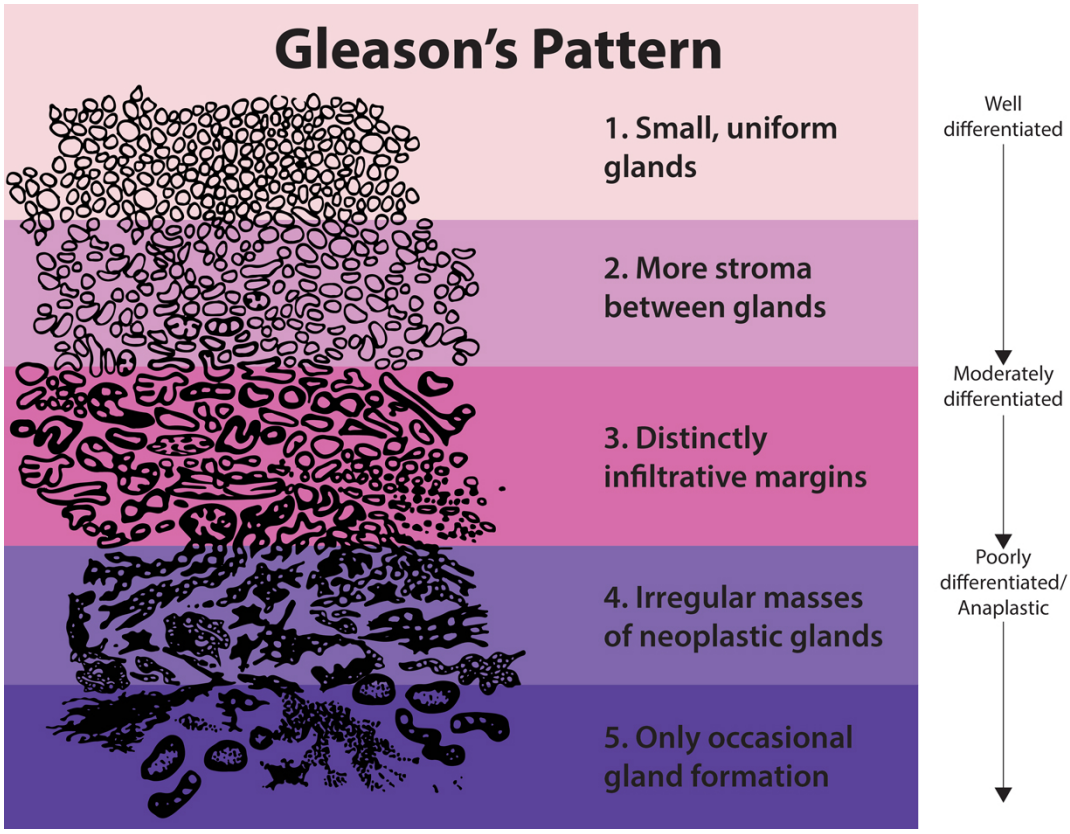


Figure 3.3: Gleason grade patterns. The cancerous regions are graded from 1 to 5, with grade 5 corresponding to the least differentiated cells, with limited to no glandular structure residual in the tissue.

cancer are complex [14]. Only a couple of genes of clinical significance with moderate to significant effect sizes have been identified, BRCA1 and BRCA2, and HOXB13 [15, 16]. BRCA2 gene mutation carriers have about a three-fold increased prostate cancer risk and have been associated with younger disease onset, poorly differentiated prostate cancer, and poorer survival. Many relatively common gene variants, single nucleotide polymorphisms (SNPs), can also affect risk of prostate cancer. To date, more than 200 SNPs have been linked to an increased prostate cancer risk [17]. The known SNPs account for approximately 33% of the inherited prostate cancer risk [18]. Individually, they do not provide clinically relevant information, but the combined effect of many such variants plus the family history of the disease may provide a significant increase or decrease in the estimated prostate cancer risk [19, 20, 21].

3.3 Treatment

Treatment of prostate cancer can be coarsely categorized into three groups: active surveillance (conservative treatment), curative treatment (surgery, radiation), and non-curative, life-prolonging (hormone treatment, chemotherapy, radiation). The initial treatment is determined by the cancer's stage and Gleason score, PSA levels, comor-

bidities, and the patient's age and preferences.

3.3.1 Active surveillance

The goal of active surveillance (AS) is to decrease the overtreatment of low-risk prostate cancer to avoid adverse effects of the treatment without compromising possible future curative treatment in men who show signs of disease progression over time. Studies have shown that most men who are diagnosed with low- or intermediate-risk prostate cancer will die from other causes than prostate cancer 10-15 years after diagnosis [22, 23]. Current prostate cancer guidelines recommend active surveillance for low-risk prostate cancer patients with an expected remaining lifetime of 10 years or more, with a follow-up based on repeat biopsy, serial PSA, DRE, and MRI for patients with a PSA $\geq$ 2 ng/mL increase during a two year period. Prostate cancer guidelines today state the initial selection of patients for AS should be based on MRI prior to a confirmatory biopsy, followed by both systematic and targeted biopsies [24, 25]. AS may also be relevant for intermediate-risk ISUP 2 prostate cancer patients with low PSA $<$ 10 ng/mL, a small amount of Gleason pattern 4, and a small number of positive cores, but the evidence is lower for this group of patients. The presence of ISUP 3 disease, cribriform patterns, or intraductal cancer excludes the men from active surveillance. According to the current Swedish guidelines, patients on active surveillance should undergo a confirmatory biopsy within two to six months following the diagnostic biopsy and thereafter a follow-up biopsy every 2-3 years for 10 years after starting active surveillance. Furthermore, it is recommended that patients on AS undergo PSA testing every four months during the first year and thereafter every six months. The largest AS cohorts now provide long-term follow-up data, recommending AS rather than curative treatment for localized low-risk prostate cancer. In these studies, the prostate cancer-specific risk of death or metastasis was less than 1% over a 10-year period [26, 27, 28]. A randomized controlled trial by Wilt et al. did not show a significant difference in mortality between observation and radical prostatectomy during a 12-year follow-up [29]. The ProtecT trial randomized patients with localized prostate cancer to either active monitoring, surgery, or radiotherapy, and showed very low disease-specific mortality after 10 years of follow-up, irrespective of treatment assignment, although active monitoring is considerably less stringent than active surveillance protocols [30].

A negative consequence of active surveillance, however, is the invasive surveillance associated with repeated tissue sampling. That may significantly reduce the quality of life in these patients, cause infections, and increase the risk of urosepsis, which is the most feared complication of transrectal prostate biopsies (sepsis rates range from 1-3%) [31].

3.3.2 Radical Prostatectomy and Radiation

For intermediate and high-risk localized prostate cancer with a life expectancy of ten years or more, curative treatment is recommended.

The prostate gland and seminal vesicles are surgically removed during a radical prostatectomy, while whenever possible preserving pelvic organ function. Radical prostatectomy is performed either using open surgery, traditional laparoscopic technique, or robot-assisted laparoscopic prostatectomy. The advantages of robot-assisted prostatectomy compared to open prostatectomy include less risk of blood transfusion and shorter hospital stay. However, review reports have shown no significant difference between the two surgical techniques with regard to oncological, urinary, and sexual outcomes [32]. Results from the ProtecT trial, report a 99% 10-year cause-specific survival following curative treatment with radical prostatectomy in a screening population [30]. Even if the goal of radical prostatectomy is to remove the prostate cancer while preserving continence and potency, 20% of the treated experience incontinence, and 70% experience erectile dysfunction one year after surgery [33].

Radiotherapy is also used to treat clinically significant prostate cancer, where gamma radiation beams are focused towards the prostate and the surrounding tissue. Randomized studies have shown that radiotherapy in combination with hormone therapy decrease mortality in high-risk cancers [34, 35, 36]. Radiation therapy is considered to provide the same chance of cure as surgery. The only existing randomized trial that directly compared radiation with radical prostatectomy, showed no difference in 10-year cancer-specific mortality for localized prostate cancer [30]. The most common side effects after radiation therapy are reduced erectile function, rectal problems, and more frequent urination.

3.3.3 Chemo- and hormonal treatment

Prostate cancer cells rely on testosterone for growth. The goal of hormone therapy is to reduce levels of the male hormone testosterone and to prevent them from reaching prostate cancer cells. Hormone therapy is used to shrink and hinder growth of prostate cancer cells, but it does not cure prostate cancer. Hormone therapy includes both drug treatment and surgical procedures to remove one or both testicles to lower testosterone levels. Chemotherapy refers to drug treatments that uses chemicals to hinder the growth of cancer cells. Chemotherapy might be used for metastatic prostate cancer or if the patient is not responding to hormone therapy and is used to relieve symptoms and to improve quality of life.

3.4 Improvements in Prostate Cancer Diagnostics

Due to the insufficient sensitivity and specificity of PSA and the conventional use of systematic biopsies, the current prostate cancer diagnostic approach has been shown to result in high rates of overdiagnosis, overtreatment, and cancer misdiagnosis [7, 8, 37]. Therefore, a better-organized screening with improved methods for prostate cancer is needed. Future clinical workflows for prostate cancer will likely incorporate risk prediction models, targeted biopsies using magnetic resonance imaging (MRI), and AI-assisted digital pathology of biopsy samples (Figure 3.4). We hypothesize that such a pipeline will have better sensitivity to identify clinically significant prostate cancer while also decreasing the harms associated with opportunistic PSA testing.

3.4.1 Multivariate Prediction Models in Prostate Cancer detection

Predictions model that estimates an individual's risk can provide input for clinical decision-making. Risk prediction models, such as the blood tests Prostate Health Index, the 4KScore, the Stockholm3 test, and the urine test Prostate Cancer Gene 3 (PCA3) have improved risk stratification for men in prostate cancer diagnosis [38, 39, 40, 41]. These models have shown improved prostate cancer early detection, in terms of increased sensitivity for identification of significant prostate cancer as well as decreased rates of unnecessary biopsies and overdiagnosis.

The STHLM3 study developed and prospectively evaluated the Stockholm3 prediction model for screening of clinically significant cancer. The Stockholm3 score utilizes a panel of 254 single nucleotide polymorphisms, 6 plasma protein biomarkers (PSA, intact PSA, free PSA, hk2, MIC1, MSMB), and clinical variables to predict the probability of ISUP ≥ 2 cancer. The test has been demonstrated to decrease the overdiagnosis of ISUP 1 tumors and the number of performed biopsies while preserving the sensitivity for clinically relevant prostate cancer detection.

3.4.2 Magnetic resonance imaging and targeted biopsies

Magnetic resonance imaging (MRI)-guided biopsies of suspicious lesions have been proposed to aid in risk stratification and treatment selection in prostate cancer diagnosis. The Prostate Imaging Reporting and Data System (PI-RADS) score was established for reporting MRI data. Each lesion receives a score between 1 and 5, with higher scores indicating lesions more suspicious of malignancy, and scores 3 to 5, are usually used to define a positive MRI. Studies such as the MRI-first study and PRECISION study have shown that MRI improves the detection of clinically significant disease, avoids unnecessary biopsies in men without MRI lesions, and reduces overdiagnosis of clinically insignificant prostate cancer [42, 43, 44, 45].

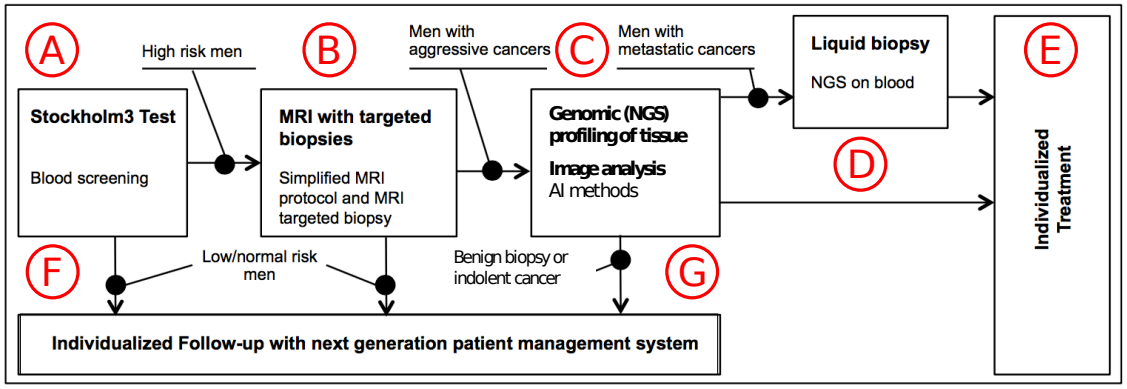


Figure 3.4: A possible future Prostate Cancer Diagnostic Pipeline. (A) Initial screening is performed using the Stockholm3 test. (B) High-risk men are referred to imaging using MRI. Men with a positive MRI (PIRADS $\geq$ 3) are referred to MRI-ultrasound fusion guided targeted biopsies, where 2-4 core biopsies directed to the areas of the prostate where the MRI indicates that the lesions are located. (C) Tumor profiling using next generation sequencing in combination with histological image analysis of the prostate biopsy slides are used to predict the prognosis and optimal treatment. (D, E) The patient gets an individualized treatment pathway based on the results from (C). (F) Men who are negative in step (A) and (B) are followed up according to a screening protocol matching their risk level. (G) Men diagnosed with low-risk prostate cancer are offered active surveillance where they are monitored every year using S3M+MRI.

Data on screening trials that incorporate MRI are emerging and the role of MRI in PSA screening has been studied in two recent randomized clinical trials, the STHLM3-MRI screening trial, and the IP1-PROSTAGRAM study. The STHLM3-MRI screening trial was a non-inferiority trial that compared an experimental biopsy strategy of MRI with targeted and systematic biopsy in men with positive results on MRI compared to standard biopsy in an intention-to-treat analysis of men with PSA 3 ng/ml or higher [46]. The experimental strategy using MRI with targeted and systematic biopsy resulted in non-inferior detection of clinically significant prostate cancer, a 48% reduction in performed biopsies, and a 62% reduction in detection of clinically insignificant disease. The IP1-PROSTAGRAM trial evaluated prostate screening with MRI using a PI-RADS score of 4 to 5 to define a positive test result and compared with screening using PSA levels of 3 ng/ml or higher and ultrasound-guided systematic biopsy. MRI detected nearly twice as many clinically significant cancers and resulted in a similar number of performed biopsies and detection of clinically insignificant cancers [47]. The STHLM3-MRI trial also evaluated the combination of a blood-based risk prediction with MRI-targeted biopsies. The main analysis compared Stockholm3 and PSA tests in the experimental group (MRI-targeted and systematic biopsies) and showed that the Stockholm3 test, provided non-inferior detection of clinically significant cancer, 36% reduction in MRI procedures, and 8% fewer biopsies compared to the PSA test. In a secondary analysis, the novel strategy of Stockholm3, MRI-targeted, and systematic biopsy provided 52%

fewer biopsies and a 69% reduced overdetection of clinically insignificant low-grade cancers compared to the traditional screening using PSA and systematic biopsies [48].

3.4.3 AI-assisted digital pathology of biopsy samples

The number of men undergoing prostate biopsies has markedly increased in the past decades due to opportunistic PSA testing and there is a shortage of expert uro-pathologists world-wide. As a result, pathology departments face a substantial increase in workload and complexity of diagnosis and grading of cancer. Gleason grading is the most important factor for predicting prognosis and for treatment planning of prostate cancer. However, pathological assessment of prostate biopsies using Gleason grade is to some degree subjective and suffers from high intra- and inter-observer variability. [49]. The use of reference databases [50] and AI is hoped to improve the consistency and accuracy of Gleason grading and reduce the pathology workload.

Chapter 4

Data Material

This thesis used three primary data sources: the Stockholm PSA and Biopsy Register, the STHLM3 diagnostic trial, and the STHLM3-AS study that we conducted during the work of this thesis, to evaluate the Stockholm3 test in combination with MRI-guided biopsies for disease monitoring in prostate cancer active surveillance.

4.1 The Stockholm PSA and Biopsy Register (SPBR)

The SPBR is a population-based database that includes data on all men who have resided in Stockholm County since 2003. The register consists of all prostate biopsies and PSA tests that have been performed in Stockholm between 2003 to 2016 [51].

For Study I, a new linkage of the SPBR and a number of Swedish population registers was performed by the Swedish National Board of Health and Welfare (Socialstyrelsen) for the study period during 2008-2017. Data linkage to the National Prostate Cancer Register (NPCR) provided detailed data on diagnoses, tumor characteristics of biopsied individuals (tumor stage and Gleason score) and primary treatment selection of patients (e.g. active surveillance or curative treatment using either radical prostatectomy or radiation). The NPCR has a coverage of 98% compared with the Swedish Cancer registry [52]. Further linkage was performed with the Swedish national population registers for the extraction of patient data, causes of death, drug prescriptions, and population demographics and labor statistics. Finally, we collected additional re-biopsy material from the three different pathology departments in Stockholm County (Karolinska University Laboratories Huddinge, Aleris, Unilabs). Detailed information about Gleason grading and cancer length in re-biopsies during the study period of 2008-2017 was retrieved for the study participants from medical charts and linked to the database.

4.2 STHLM3 trial

The STHLM3 trial was a population-based prospective prostate cancer screening trial including men without prostate cancer aged 50-69 years from Stockholm County during 2012-2014. The study used a paired screening, where the subjects were initially screened using PSA above $\geq$ 1ng/mL, and in the second step men were referred to a biopsy either if they had a PSA $\geq$ 3ng/mL or a Stockholm3 score predicted probability of ISUP $\geq$ 2 cancer above 10%. In total a random selection of 145,905 men were invited to the study, 59,149 men were included in the study and 7,417 men underwent prostate biopsy. The STHLM3 cohort is a unique data source that contains detailed information on all study participants by combining clinical variables, plasma protein biomarkers such as PSA, and Gleason scores on the biopsied individuals [40].

4.3 STHLM3-AS

The STHLM3-AS trial (NCT03956108) was a prospective, cross-sectional, multi-center trial embedded within the STHLM3 trial [53]. The aim of the study was to evaluate the performance of the Stockholm3 test in combination with MRI-targeted biopsies for monitoring disease progression in patients on active surveillance and comparing this to the standard follow-up for active surveillance at the time of the study using (PSA + systematic biopsies).

Participants from the STHLM3 trial who had received a prostate cancer diagnosis, and currently enrolled in active surveillance, i.e. with no history of starting treatment using either surgery, chemotherapy, radiation therapy, or hormone therapy were invited. From the STHLM3 study 1374 men were diagnosed with ISUP 1 cancer and out of these 541 men that were currently on active surveillance were invited to take part in the STHLM3-AS trial.

The inclusion of patients started in August 2018 and ended in December 2019. During the study period, 309 eligible men were registered to the trial. In total we excluded 21 men that declined to undergo the study biopsies and 8 men for which the lab analysis of the blood samples failed. After exclusion criteria's, 290 men were included in the study. At baseline, blood sampling was performed for Stockholm3 and PSA analyses, and the patient was also asked to complete a study-specific questionnaire with questions on quality of life and anxiety. An MRI was performed and evaluated in line with PI-RADS v2 guidelines at baseline. Men who had negative MRI results (PI-RADS $<$ 3) underwent systematic biopsies (10 to 12 cores) and men who had positive MRI results (PI-RADS $\geq$ 3) underwent both systematic and targeted biopsies. The original pathological review was completed centralized for all study sites at Unilabs Stockholm.

A second pathological reevaluation was carried out for study participants that were upgraded from diagnostic ISUP 1 cancer to ISUP $\geq$ 2 cancer in the study biopsy, together with a random sample of ISUP 1 and benign biopsies. The second pathology review was performed to validate the upgraded diagnoses, and was carried out by the same expert uropathologist who completed the diagnostic pathology evaluation in the STHLM3 trial. In total, 509 (13%) of the study biopsies were reevaluated in the second pathology review.

Chapter 5

Methods

5.1 Deep learning

Machine learning methods use algorithms and statistical models to build computer systems that are designed to perform a specific task, without being explicitly programmed how to perform the task. By providing the machine learning algorithm with large volumes of either labeled or unlabeled data, the goal for the learning algorithm is to build a mathematical model based on the training data that can be used to make reliable predictions. Deep learning is a type of machine learning based on neural networks. The capacity of deep learning methods has developed quickly in recent times primarily due to access to large datasets and increasing computational resources.

Much of the development in deep learning dates back to 1940-1980, but it is only in recent years these models have reached new highs in popularity. Some of the earliest work were intended to provide computational models that were inspired by biological learning of the brain. Later work arose in the field of cognitive science with the successful implementation of the back-propagation algorithm for training of deep neural networks, that is still the main approach used today for training these models [54]. However, these models were generally believed to be too computationally costly to fit with the hardware available at the time and did initially not live up to the expectations, causing research in this field to fall out of favor during the 1990s. The improvements in performance can mainly be attributed to advances in computational efficiency, improved algorithms, and the increased availability of large sets of labeled data. Deep learning has been applied with great results in many fields such as natural language processing and computer vision, outperforming many traditional machine learning methods that require hand-crafted features [55]. A representation learning algorithm learns to perform the mapping from input data to output prediction, while also learning the representation of the most important features directly from data. Deep learning uses representation learning to build a nested hierarchy where more complex features are derived from simpler features. In image processing, for instance,

lower layers of the neural network might identify edges, whereas higher layers may identify clear parts of the object such as images, digits, and letters. A traditional neural network consists of three layers, an input layer that takes the input data, hidden layers that extract features from the input space, and an output layer that generates the final conditional prediction over all classes.

Convolutional Neural Networks (CNNs) is a subclass of deep neural networks that have been proven to be very effective in image processing. CNNs work better for images compared to traditional neural networks due to a number of reasons. 1. CNNs preserve the spatial structure by applying small filters on spatially neighboring pixels across all axes of the image. The output of the filter is the degree of activation that was generated in the precise specific spatial position of the input image. 2. Translation invariance, meaning that the elements of interest can appear anywhere in the image, e.g. translated by one pixel and still be detected. This is achieved by parameter sharing of the features across multiple image locations. 3. Parameter sharing enables an efficient parametrization of the neural network, effectively reducing the number of model parameters which simplifies training and reduces the risk of overfitting to training data.

5.1.1 Deep learning in histopathology

The application of CNNs to images has obvious applications in medicine, in particular in radiology and pathology [56]. Recently, deep learning models have been demonstrated to be able to scan retinal images for identifying diabetic retinopathy in retinal fundus photographs [57], to achieve dermatologist level assessment of malignant melanoma [58], and to be able to detect lymph node metastasis in tissue sections from breast cancer patients [59].

The most significant applications of deep learning to prostate diagnostic histopathology have been conducted in recent years. A study by Litjens et al. used 225 pixel-wise annotated whole slide images to train a deep learning model that attained an AUC of 0.99 for cancer detection in prostate biopsies [60]. Two studies published in 2020, one by our research group and the other by Bulten et al., demonstrated that AI could achieve expert-level ISUP grading of prostate biopsies [61, 62]. A recently published challenge for prostate pathology showed that the performance of AI algorithms could generalize across different patient populations, laboratories, and scanners in a large multinational validation setting [63]. Furthermore, studies have also applied deep learning to grading of prostatectomy tissue samples [64, 65].

5.2 Conformal prediction

Conformal prediction is a mathematical framework that uses past experience to determine precise levels of confidence in new predictions. Given a user-specified error probability ϵ, together with a prediction method (such as an AI system), it produces a prediction region which is a set of labels that contains the true label with probability 1 - ϵ. This contrasts to point predictions from conventional prediction models. Conformal prediction is lightweight and can be implemented for all learning algorithms, and is founded on a well-defined mathematical theory that can guarantee that predictions are valid under the assumption of independent and identically distributed data [66].

Given a training set $(x_1, y_1), ..., (x_n, y_n)$, where x_i is a feature vector and $y_i \in (Y_1, .., Y_k)$ is a known class label out of a finite set of possible classes. The goal of the classification task is to predict the class label y_{n+1} for a new patient x_{n+1}. Most standard prediction models output predicted probabilities for each class or a single value classification y_{n+1}. Conformal predictors instead outputs a prediction region $\Gamma^\epsilon = \{y_i, .., y_k\}$ given a user-specified significance level or error probability $\epsilon \in (0, 1)$. Conformal prediction works by trying every potential class label $(Y_1, ..., Y_k)$ as a candidate label for y_{n+1} and evaluating how well the potential class label is conforming to the training instances. $(x_1, y_1), ..., (x_n, y_n)$.

5.2.1 Nonconformity measure

The concept of conformity is represented by a *non-conformity score*, that intuitively quantifies how different the new example is compared to the training examples. The most common non-conformity for classification problems is one minus the predicted probability. However, this score function is completely user-defined and can take any form. The non-conformity measure is used to compute the non-conformity scores a_i for each labeled example $i = 1, \ldots, N$ of the training set together with the score of each test example a_{n+1} (with an unknown label). These sets of non-conformity scores can be used to quantify how different the new example is compared to the training examples.

5.2.2 P-values

To measure how conforming a potential label for a test example N+1 is with previous data, we compute the fraction of non-conformity scores a_i $(i = 1, ..., N + 1)$ that are equal or larger than a_{N+1}: $|i = 1, .., n + 1 : a_i \geq a_{n+1}|/n + 1$. A large p-value for x_{n+1} and a proposed class label Y indicates that the pair is similar or conforming to previous examples and that Y is a likely class label.

5.2.3 Prediction regions

Given a user-specified confidence level 1 - ϵ, the potential labels whose p-value are larger than ϵ will be accepted. Using this strategy, it is guaranteed that the true label will be included in the prediction region with probability 1 - ϵ [66].

Conformal predictors output prediction regions that can contain any number of all possible labels, which contrasts with the single-value classifications that we are used to from regular prediction models. A smaller prediction region is more informative or efficient, and ideally, the prediction set would consist of a single predicted class. A conformal predictor can output multiple classes when the model is unable to differentiate between the classes at the user set confidence level. Empty set predictions are examples where the model was not able to assign any labels, i.e. an example for which no p-value was larger than ϵ and a prediction could thus not be made, usually indicating that the new test example is very different or non-conforming to the training data that the model was developed on.

Similar to how higher intended levels of confidence in parameter estimates result in wider confidence intervals, larger desired levels of confidence in the prediction result in larger prediction regions. Conformal prediction thus permits us to maintain a low error rate by only accepting predictions with high confidence. The clinical cost of generating incorrect predictions can be taken into consideration by adjusting the confidence level. For example, a false positive test could lead to an incorrect diagnosis, which in the worst case could lead to unnecessary treatment, while a false negative test result could lead to missed cancer detection.

5.2.4 Transductive and Inductive conformal prediction

The initial definition of conformal prediction was in an online transductive framework [66, 67]. This online setting uses all available data to compute the conformity score for every new instance that we need to predict, and therefore retraining of the underlying prediction model is required for the calibration and test examples. The online setting is appealing in that it utilizes all available data for the prediction of new examples, but it is frequently computationally too costly. Additionally, not all applications—particularly those in medicine—can be used in an online setting. An inductive offline framework is most commonly employed instead, which uses a fixed model that is only updated between longer intervals.

Conformal prediction was therefore extended to the inductive setting [68], where the model is developed in the training set and then used to make predictions on the held-out test set. When compared to transductive conformal predictors, inductive con-

formal prediction is computationally more efficient. For inductive conformal prediction, the training data must be split into a proper training set, which is used for training the underlying prediction algorithm, and a calibration set which is used for tuning the conformal predictor.

5.2.5 Mondrian conformal prediction

The error rate at the population level is guaranteed by conformal predictors for all observations in a dataset. However, this means that it is possible that the error rate is higher or lower within different substrata of the data, for example, the error rate might be higher for one class and lower for another class label. Mondrian conformal prediction was created in order to guarantee the pre-specified error rate within each substrata of the population. By tuning the conformal predictor within each substratum rather than applying it to the entire population, the desired error rate is mathematically guaranteed within each stratum.

5.3 Model Evaluation

Discrimination refers to how well the prediction model can separate between those with an event and those without the event and can be described in terms of sensitivity (true positive fraction) and specificity (true negative fraction). To classify a test as positive or negative, we apply a cutoff to the predicted probability, and the sensitivity and specificity can be evaluated at each possible threshold.

The Receiver operating characteristics (ROC) curve is a plot of the sensitivity and 1 – specificity (false positive rate) over all possible thresholds for the probability of an outcome. The ROC curve is the primary way to visualize the operating characteristics and the trade-off between the sensitivity and specificity of a test, as well as the clinical consequences of different thresholds. Higher thresholds lead to increased specificity at the cost of lower sensitivity. For example, lowering the threshold for the PSA test for prostate cancer screening leads to overdiagnosis and overtreatment of clinically non-significant prostate cancer.

The area under the ROC curve (AUC) is commonly used to summarize the discriminative ability of a prediction model over all possible thresholds. The AUC describes how well the predicted probabilities can rank-order the outcomes. The AUC can be interpreted as the probability that a randomly selected pair of cases and controls are correctly ranked by the model. The AUC ranges from 100% for a model that discriminates perfectly between cases and controls to 50% for a non-informative model with a discriminative ability equal to a coin flip. Confidence intervals for the AUC are usually

calculated using bootstrap resampling [69]. The DeLong test can be used to test the difference between AUCs [70].

Chapter 6

Results

6.1 Overview of the main findings

In **Study I**, we studied the uptake and the follow-up of active surveillance in patients who were diagnosed with low-risk prostate cancer in Stockholm County between 2008 and 2017. In **Study II**, we evaluated the performance of MRI-targeted biopsies and the Stockholm3 test, for monitoring disease progression in patients on active surveillance. In **Study III**, we constructed an AI system for automated pathology assessment, cancer diagnosis, and Gleason grading, using digitized biopsy core samples from the STHLM3 study. In **Study IV**, we implemented a framework based on conformal prediction that was used to estimate the uncertainty of the prediction for the AI system developed in Study III. The uncertainty estimates were used to control the error rate of the AI system, such that only reliable predictions are accepted and that unreliable predictions can be detected and flagged for human assessment. In **Study V**, we explored if weakly supervised deep learning can be used to learn and detect possible associations between genetic risk factors (SNPs) and morphological features in prostatic tissue.

6.2 Study I

Active surveillance aims to reduce overtreatment and potential treatment-related side effects and is increasingly utilized as an alternative to curative treatment for low-risk prostate cancer. This conservative management includes serial testing for disease progression to provide selective treatment with curative intent. However, an international consensus on the methods and frequency of follow-up testing to be used for screening for disease progression is still lacking.

We performed a retrospective cohort study of 6021 men aged 40-75 years who were diagnosed with low-risk prostate cancer in Stockholm county between 2008 and 2017. The study aim was to describe the uptake and follow-up of active surveillance, as well as to demonstrate the transition for these patients from active surveillance to curative

treatment over a 10-year period based on PSA dynamics and Gleason upgrading in surveillance biopsies.

On one hand, our results showed that the majority of men who enrolled in an active surveillance program remained on active surveillance throughout the follow-up (81.6%, during a median follow-up of 3.5 years), and that initiating curative treatment was usually preceded by an upgrade in Gleason in a repeat biopsy during active surveillance follow up.

On the other hand, we found that active surveillance was underutilized during the study period and that these men to a high degree do not receive a follow-up according to the current guidelines (see Figure 6.1). However, the intensity of active surveillance with regard to PSA testing and surveillance biopsies was improved over the study period. But we also showed that a large proportion of the men who were eligible for active surveillance at diagnosis were not entering AS programs (approximately 50%), but rather opted for a curative treatment that was closely related to the primary diagnosis. Taken together, our results suggest the need for optimized and less invasive protocols to increase adherence and to reduce the problem with overtreatment of low-risk prostate cancer.

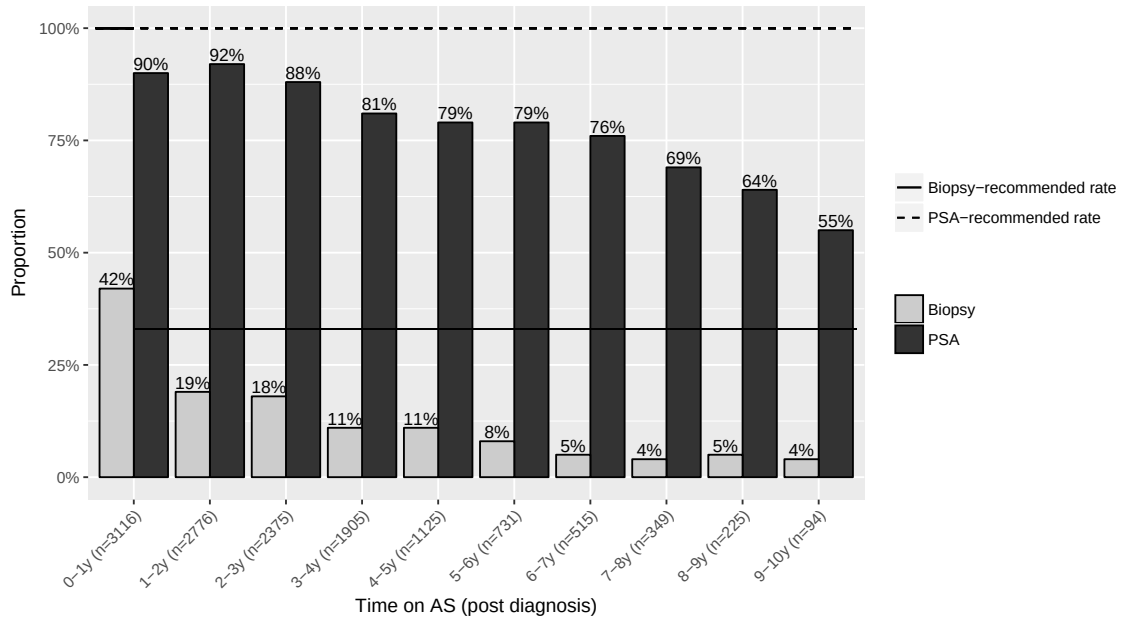


Figure 6.1: The Yearly proportion of individuals with a repeat biopsy and repeat PSA during a follow-up of 10 years (time on active surveillance). The two horizontal lines shows the recommended testing (yearly) by the national guidelines during the study period (2008-2017)

6.3 Study II

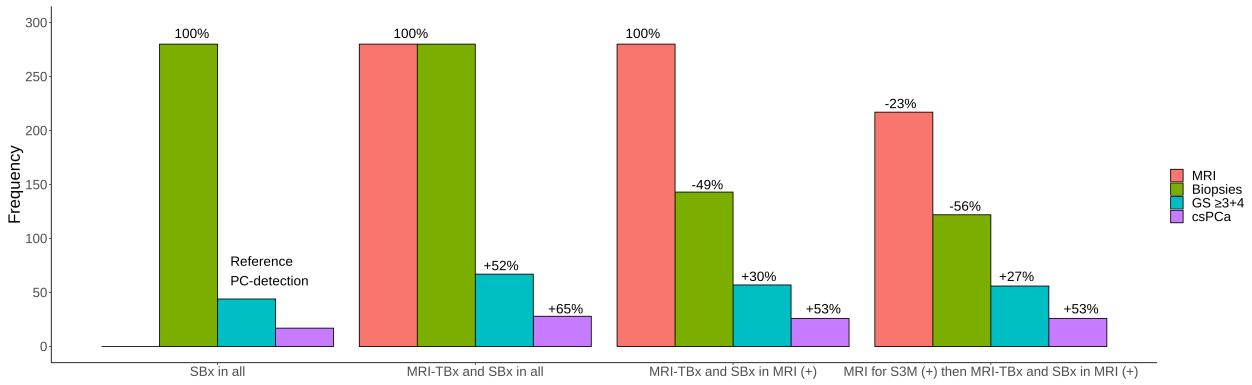
The primary aim was to evaluate cancer detection of MRI-targeted biopsies and compare this to the standard systematic biopsy procedures for men on active surveillance (at the time). The secondary aim was to evaluate the performance of the Stockholm3 test when used as an initial triage instrument to select men needing MRI and subsequently targeted biopsies in an active surveillance cohort, i.e. by filtering out men with a higher predicted probability of ISUP ≥ 2 prostate cancer reclassification. We evaluated two endpoints: 1. Detection of ISUP ≥ 2 prostate cancer. 2. Clinically significant prostate cancer was defined using the National Comprehensive Cancer Network (NCCN) guidelines (ISUP 1 and $> 50\%$ positive cores, ISUP 2 and T2, or ISUP 2 and PSA > 10 ng/mL) [71].

The main results from the study showed that the addition of MRI and targeted biopsies to conventional systematic biopsies during active surveillance showed increased sensitivity to detect both outcomes, 52% more ISUP ≥ 2 prostate cancer and 65% more clinically significant prostate cancer were detected compared with using only systematic biopsies. (see Figure 6.2) Few men with a negative MRI harbored clinically significant prostate cancer (1.4%). Secondly, our results suggested that by incorporating risk prediction models in follow-up to select men for evaluation using MRI, 23% of the MRI investigations could be avoided, and 56% of men could postpone a biopsy. While still detecting 27% more ISUP ≥ 2 prostate cancer and 53% more clinically significant cancer in comparison to standalone systematic biopsies.

In conclusion, the results from our study suggest that combining systematic and MRI-targeted biopsies in active surveillance increase sensitivity to detect prostate cancer reclassification in these men. The inclusion of risk prediction models in active surveillance may decrease the requirement for MRI use in patients with low-risk prostate cancer.

6.4 Study III

The aim of the study was to build a AI system for automated cancer diagnosis and grading of digitized whole-slide images of prostate needle biopsies. To develop the AI system we digitized 6,953 prostate biopsy cores from 1,069 men. The biopsy samples were primarily collected from STHLM3 participants, but an additional 271 slides from men with prostate cancer from another pathology laboratory were included to extend the data for high grade cancer (ISUP 4 and ISUP 5). The held-out test set consisted of a random selection of 1631 digitized biopsy samples from 246 STHLM3 subjects. The data were split by patient level to create entirely independent test data. The external



	SBx in all	MRI-TBx and SBx in all	MRI-TBx and SBx in MRI (+)	MRI for S3M (+) then MRI-TBx and SBx in MRI (+)
Biopsies	280	280	142	121
% Biopsies (95% CI)	100	100	50.7 (45.0, 56.4)	43.2 (37.1, 48.9)
MRI	0	280	280	217
% MRI (95% CI)	0	100	100	77.5 (72.8, 82.1)
Detected No. of GS ≥3+4	44	67	57	56
RS: GS ≥3+4 vs SBx (95% CI)	Reference	1.52 (1.28, 1.85)	1.30 (1.04, 1.67)	1.27 (1.02, 1.65)
Detected No. of csPCa	17	28	26	26
RS: csPCa vs SBx (95% CI)	Reference	1.65 (1.27, 2.45)	1.53 (1.13, 2.36)	1.53 (1.13, 2.36)

Figure 6.2: Evaluation of biopsy methods for detection of ISUP ≥ 2 cancers and clinically significant prostate cancer using the definition from the NCCN guidelines.

test set consisted of 330 biopsy samples from 73 subjects, from the Karolinska University Hospital. Two ensembles of convolutional deep neural networks (DNNs) were used for training the AI system. In the first ensemble, image patches were binary classified into benign or cancer tissue, whereas the second ensemble classified the image patches into Gleason score 3 to 5 patterns. Both ensembles contained thirty Inception V3 models with pre-training on ImageNet.

For cancer detection in prostate biopsies, the AI system achieved an AUC of 0.997 (95% CI: 0.994-0.999) on the held-out test set and an AUC of 0.986 (95% CI: 0.972-0.996) on the external test set. With a sensitivity of 99.3%, the AI system achieved a specificity of 88.9%. At this sensitivity level, the AI system failed to detect five biopsy cores with cancer across 721 malignant biopsy cores - four ISUP 1 grade and one ISUP 2 grade - in the held-out test set. At this operating point, no man was misdiagnosed, since the remaining malignant cores from these men were correctly classified. At a specificity of 88.9%, 809 out of 910 benign cores were spared from “human” pathological evaluation. The correlation between the cancer length measurements by the study pathologists and estimates of the AI system was 0.96 (95% CI: 0.95-0.97) on the held-out test set and 0.87 (95% CI: 0.84-0.90) on the external test set. The Cohen’s linear kappa statistic between the grade assigned by the AI system and the study pathologist was 0.83 on the held-out test set and 0.70 on the external test set. The grading performance was also evaluated on 86 cancer cases from the ImageBase dataset, a reference dataset developed by the ISUP ImageBase expert panel with gradings that were performed independently by 23 international expert pathologists. The mean pairwise kappa achieved by the AI system was 0.62, whereas the kappa values of the patholo-

gists in the expert panel ranged from 0.60 to 0.73.

These results demonstrate the possible clinical utility of automated digital pathology of prostate biopsies in a number of different ways. In a scenario where the pathologists are assisted to pre-screen all biopsy samples, the AI system has the potential to identify and remove benign cases, and then only those predicted positive would need to be assessed by expert pathologists. Furthermore, it appears feasible that measurements of cancer length can be automated by an AI, which is an important task to quantify tumor burden and provide accurate recommendations for where to focus attention when assessing the core, which could lead to time savings for pathologists on this specific task. Lastly, the grading performance is on par with that of leading prostate pathology experts motivating the potential clinical utility of these models in prostate cancer pathology. For example, either as decision support for inexperienced pathologists or to provide expertise to regions where there is a shortage of uro-pathologists.

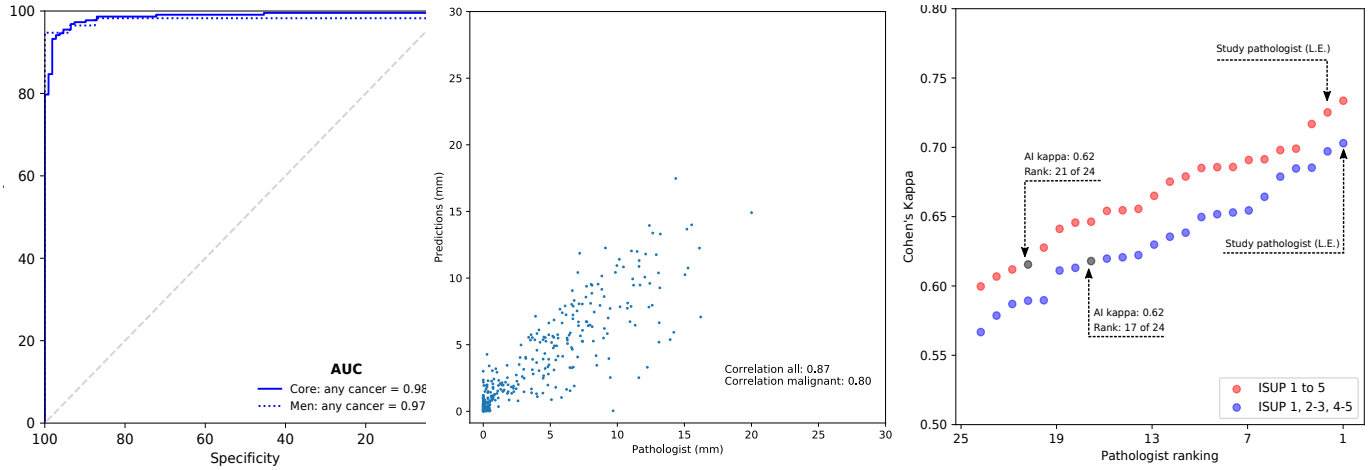


Figure 6.3: **(Left:)** ROC-curves and AUC for cancer detection by men and individual prostate biopsy cores on the external test set. **(Middle:)** Correlation of mm cancer length assessed by the pathologist and predicted by the AI on the external test set. **(Right:)** Cohen's linear kappa statistic for Gleason grading by the AI system and each pathologist on the ISUP ImageBase reference dataset ranked from lowest to highest. Average pairwise kappa for the AI system and each of the pathologists respectively in comparison with the other pathologist. The kappa values of the study pathologist and the AI system are highlighted with arrows.

6.5 Study IV

Studies have shown that AI can provide diagnostic and grading for prostate pathology with equivalent performance to that of expert pathologists. The widespread use of AI systems will unavoidably expose these systems to data that is outside of the domain of training data. Being able to detect unreliable predictions in order to identify and flag them for human assessment will therefore be key to ensuring patient safety. How to detect such unreliable predictions is however a question that so far has received

very little attention in the medical literature. Conformal prediction is a mathematical framework that provides well-calibrated prediction, where the accuracy of the predictions is equal to or greater than a confidence level specified by the user (e.g. a required confidence level of 90% results in at least 90% accurate predictions). Conformal prediction is based on the exchangeability assumption, that all data (training and test data) comes from the same distribution but occurs in a random order, which is a slightly weaker assumption than i.i.d. (usually imposed by most standard statistical models). The validity property of conformal prediction can be used to detect data drift over time, potential variations in the appearance of new data compared to training data that could result in a decline in model performance (an invalid model), and warn that re-calibration is needed to ensure accuracy.

The aim of the study was to develop conformal prediction for AI-assisted prostate pathology and describe how the system can be used to control the error rate of the AI model: such that it only accepts predictions with high confidence. Unreliable prediction can occur due to several reasons: We evaluated the conformal predictor's ability to detect: 1. Systematic variations between the test data and the training data leading to a decline in model performance (poor generalizability due to data drift caused by variations in the preparation of tissue between laboratories and digital scanning with a variety of scanners). 2. Individual atypical prostate tissue, such as rare malignant subtypes and benign mimickers that the model did not encounter during training.

For cancer detection, the conformal predictor's error rate was 0.1% as opposed to an error rate of 2% by the AI model without conformal prediction when provided with new data drawn from the same distribution as the training data. Our results on the dataset consisting of atypical prostate tissue showed that the conformal predictor could detect such unusual cases. Using conformal prediction the error rate was decreased from 25% (AI system when not using conformal prediction) to 2% while flagging 80% uncertain predictions for human review. The conformal predictor was able to detect systematic variations in external data, causing a decline in predictive performance using relatively small samples of 49 observations for an external scanner and 10 observations for both an external pathology laboratory and scanner. These results show how conformal prediction can be used to facilitate a responsible implementation of AI systems in clinics, promoting patient safety by keeping the error rate low and providing ways to detect unreliable predictions.

6.6 Study V

Several genetic markers have been established as prostate cancer risk factors in recent years. Molecular profiling to detect genetic risk factors is important in cancer diag-

nosis and prognosis, but it is expensive and thus not widely used in prostate cancer screening. About 5% of the male population aged 50 to 70 has undergone a negative prostate biopsy. Associations between genetic alterations and morphology can be used for triaging patients for genetic testing after undergoing a prostate biopsy, which can impact individualized prostate cancer screening strategies. The aim was to study if weakly supervised deep learning can be used to learn associations between genetic risk factors and morphological features in prostatic tissue. The genetic risk factors that were investigated were a polygenic risk score (PRS) derived from a panel of 254 single nucleotide polymorphisms (SNPs) and mutations in higher penetrant genes (HOXB13). This manuscript will in its final form also include associations between mutations in higher penetrance genes (e.g. BRCA1 and BRCA2) and prostate tissue morphology, but sequencing results have not yet been finalized at the time of writing this.

Tissue morphology changes radically with cancer development and could potentially confound any association between genetic risk markers and morphology. Therefore, the evaluation was performed only on benign men. We digitized 9,752 benign biopsy cores from 961 STHLM3 participants. The polygenic risk score was categorized into two risk groups: low genetic risk load ($\leq$20th PRS percentile) and high genetic risk load ($\geq$80th PRS percentile). From the benign men in the STHLM3 trial, we identified 120 men with mutations in the HOXB13 gene and included 999 benign biopsies in the training set. Weakly supervised DNNs were trained to predict genetic risk load class-wise probabilities as well as HOXB13 mutation status for each biopsy core, and five-fold cross-validation was used to evaluate the performance.

The AUC for predicting low vs. high polygenic risk was 0.58, and the AUC for the prediction of HOXB13 mutation carriership was 0.65. PRS load was predicted more accurately at younger ages (AUC = 0.64) and lower PSA levels (AUC = 0.65). These results provide proof of principle for studying the use of morphological information in multi-modal prostate cancer risk prediction models. That has the potential to improve risk stratification among men following a negative biopsy and would offer a cost-effective way to detect genetic risk factors without molecular profiling.

Chapter 7

Discussion

7.1 Improved active surveillance

7.1.1 Current active surveillance

Active surveillance is the recommended primary treatment choice for low-risk prostate cancer patients (tumor stage T1-T2a and ISUP grade 1 and PSA < 10 ng/mL) with an expected remaining lifetime of 10 years or more. Studies now provide long-term follow-up of these patients showing strong evidence that active surveillance is a safe management strategy for men with low-risk prostate cancer, with a risk of death and metastasis of less than 1% over a 10-year period. The follow-up of patients on active surveillance with an expected remaining lifetime of 10 years or more currently consists of repeat biopsy, serial PSA, and DRE. The invasive follow-up that follows from repeated tissue samples, however, is a drawback of active surveillance since it increases the risk of urosepsis and infections and lowers the patient quality of life. The analysis of the data from the SPBR in **Study I**, showed that a majority of men did not receive a follow-up according to the guidelines at the time, and that only 50% of eligible men received active surveillance as their primary treatment choice during the study period. These findings suggest the need for improved and less invasive methods for active surveillance to improve adherence and reduce the issue with overtreatment of low-risk prostate cancer. These results provided motivation for **Study II**, which investigated the potential use of MRI and risk prediction models in the follow-up of active surveillance.

7.1.2 Incorporating MRI and biomarkers in active surveillance

Improved risk stratification is important in the active surveillance setting to be able to filter out significant prostate cancer that would benefit from curative treatment, which is crucial to reduce the overtreatment of low-risk disease. Several studies have demonstrated the potential benefits of MRI in prostate cancer screening. Multiple studies have also shown that MRI is useful for the selection of men for AS and is recommended by the national and European prostate cancer guidelines due to its improved sensitivity

and specificity. MRI for AS follow-up is also beginning to appear in the guidelines. However, strategies for incorporating MRI into follow-up and assessing its efficacy for disease monitoring in AS have yet to be fully established [24, 25, 72].

One important question is whether men should undergo MRI-targeted biopsies in addition to systematic biopsies during follow-up of active surveillance. Our results from **Study II** support the use of a combined approach using both MRI-targeted and systematic biopsies, which showed an increased detection of 52% more ISUP $\geq$ 2 prostate cancer compared to performing systematic biopsies in all men. A finding that is consistent with previous observations in a screening setting [73]. Although retrospective data on active surveillance cohorts have suggested similar findings, this has not been fully established for active surveillance [74, 75]. There is an ongoing debate if MRI-negative men need to undergo routine biopsies during follow-up. An alternative approach was evaluated in **Study II** that only performed targeted and systematic biopsies in MRI-positive men showed a 30% increase in the detection of ISUP $\geq$ 2, reducing the number of performed biopsies by 49%, while failing to diagnose 1.4% of clinically significant prostate cancers in MRI-negative men. Data from other prospective trials have reported rates of upgrading to ISUP $\geq$ 2 cancer in MRI-negative men ranging between 1.8%, 4%, and 11% [76, 77, 78]. Two recent meta-analyses have evaluated the value of serial MRI to predict upgrading in active surveillance follow-up. In summary, the findings of these two independent studies concluded that evaluation using MRI alone is currently not an option to provide safe monitoring of disease progression in active surveillance. The need for additional biomarkers and clinical variables that can be combined with MRI evaluation, as well as further standardization of reporting of MRI is identified as key points to address [79, 80].

The introduction of MRI in prostate cancer diagnosis brings a couple of new challenges. The MRI evaluation is known to be associated with a high inter-observer variability making the accuracy highly dependent on the experience of the reporting radiologist. The large natural variability in the presentation of MRI lesions on serial MRI complicates the monitoring of disease development with MRI. Another consequence of the introduction of MRI in prostate cancer diagnostics is that it is believed to cause a grade shift that might lead to overtreatment. MRI is more sensitive to detecting smaller, high-grade cancers that, on average, have a better prognosis than those detected by systematic biopsies [81]. This could potentially lead to unnecessary reclassification and overtreatment in active surveillance cohorts. It has been suggested that MRI findings may be combined with PSA density, PSA velocity, or biomarker risk prediction to improve the selection of men for follow-up biopsy. Implementing MRI in the follow-up of AS would also induce a substantial cost and pose a risk of overwhelming healthcare resources. The inclusion of additional biomarkers or risk prediction models might help

to mitigate some of these problems – by providing additional risk stratification before MRI evaluation and targeted biopsies. Incorporating biomarkers might decrease the number of MRI investigations needed. Our results showed that adding the Stockholm3 test to initially select men at increased risk led to a reduction of MRI investigations by 23% and the number of performed biopsies by 56%, while only missing the diagnosis of 1.3% clinically significant disease in men with a negative MRI and a negative Stockholm3 test.

7.2 AI-assisted prostate pathology

7.2.1 Clinical adoption of AI in prostate pathology and main challenges

The work by our research group and others has demonstrated that the performance of these AI systems is equivalent to that of leading experts in prostate pathology with respect to sensitivity, specificity, and grading concordance with expert uro-pathologists. Pathologists are also demonstrating growing interest and optimism in the clinical implementation of AI methods. In a recent survey of ISUP members, 71% of the participants agreed that machine learning will play an increasingly important role in screening and decision support in prostate cancer histopathology [82]. Despite these successes in advancing AI for prostate pathology, more work needs to be done before we will have mature AI systems handling both diagnosis and grading implemented clinically. As far as we are aware, there is currently no high-level evidence demonstrating that AI systems improve the quality of prostate pathology in a prospective clinical setting, and we do not know of any ongoing prospective multi-site clinical trials. Similarly, to the best of our knowledge, only one study exists to this date that independently assesses the validity of multiple algorithms for the problem in a multinational setting (the PANDA challenge [63]). In particular, the problem with the generalizability of AI systems is currently unsolved. We expect that widespread clinical implementation of AI systems will unavoidably expose these models to data that is outside of the domain of training data, with data originating from different laboratories, different pathology scanners (or even stain variation or changing processes within a lab [83], a different patient population, wear and tear of scanners, etc. These are challenging problems already with simple clinical risk calculators or nomograms used today [84], and the challenges will only become larger with the implementation of complex AI systems [85].

7.2.2 Improving generalizability of AI for digital prostate pathology

In order to develop a robust AI system for diagnosing and Gleason grading prostate cancer in biopsies, our research group is currently aiming to incorporate a couple of key areas into the model architecture: scanner calibration, scaling up the amount of

training data, improved algorithms, modeling of morphological heterogeneity, and implementation of automated quality control into the prediction algorithm to make sure that the model only makes predictions on the data it was designed to handle.

The use of Generative Adversarial Networks and other computational methods for staining and data augmentation can potentially account for some of the variations in whole slide images across different digital scanners [86, 87]. However, these methods are known to be hard to design to accommodate all possible sources of variations. It has been shown that the use of physical calibration slides can normalize color variations between different scanners. These methods allow the colors of whole slide images to be calibrated according to the International Color Consortium (ICC) profile of scanners [88].

The appearance of digitized biopsies varies greatly in terms of different tissue preparation techniques and digital scanners used in different clinics. To train robust AI models, we have retrospectively collected biopsy samples across nine European laboratories. This data is considerably larger and includes various scanners compared to previous studies, and will be used both to enrich the training data as well as to facilitate validation of the algorithm in an international multisite setting. We are also collecting additional data on atypical prostate tissue, such as rare cancer subtypes and benign mimics of cancer. These subtypes are rare and typically difficult to diagnose and grade for pathologists and might be accidentally mistaken for cancer during pathology assessment. These datasets will be an important data source for quality control and to assess the robustness of the AI system.

We are also improving the generalization of our AI algorithms for prostate pathology by incorporating the main findings of the PANDA challenge into the development of novel AI methods for improved algorithm robustness. Some of the key findings were that weakly supervised AI algorithms are sufficient for obtaining pathologist-level performance in Gleason grading, and that various techniques for controlling label noise are important to improve performance.

Explicit modeling of other features that are important for pathological assessment and prognosis of patients, such as perineural invasion (PNI) and cribriform morphologies, might be another way to improve AI-assisted prostate pathology [89].

7.2.3 AI safety for clinical adoption

The most straightforward way to address the problem of the generalizability of AI systems to external data is to collect more heterogeneous training data. However, accommodating all possible sources of variation will be nearly impossible, and the performance of the models may diverge over time with the inclusion of new scanners or even a drift in the performance of a specific machine (e.g., an aging scanner in a

specific laboratory, or due to scanner software updates). An additional strategy for dealing with the generalizability problem is to incorporate automated quality control into the prediction algorithm to make sure that the model only makes predictions on the data it was designed to handle. As a result, more AI research is emphasizing the importance of including confidence estimates with the predictions. This, in turn, may help pathologists build trust in AI systems and facilitate clinical adoption [90, 91].

The primary objective of **Study IV** was not to directly improve the accuracy of the performance of the underlying deep learning algorithm per se but to construct a framework based on conformal prediction that can assess the reliability and estimate the uncertainty of the predictions for AI systems in digital pathology, such that unreliable predictions can be identified for human pathology assessment. Conformal prediction therefore provides a way to control the error rate of the AI system so that it only accepts predictions with high confidence. Conformal predictors output multi-label predictions when the AI system is unable to assign reliable single predictions. So, for example, in the case of classifying cases as either benign or malignant, the conformal predictor would classify an unreliable prediction as both classes (i.e., both malignant and benign). Such a prediction is not incorrect per definition but would require human assessment since it is inconclusive. Therefore, it can be argued that conformal prediction provides more informative predictions in the sense that we get a prediction interval around the prediction that enables us to assess how confident we are in the prediction and to distinguish between more certain and less certain predictions. We also believe that the prediction regions better reflect the uncertainty in Gleason grading [49]. This also opens up an interesting discussion about the synergies of humans and machines working together to enhance the precision of prostate pathology, where the conformal predictor flags unreliable predictions for human assessment. We can then achieve expert uro-pathologist-level diagnostic accuracy on the flagged biopsies. Expert pathologists would then have more time to concentrate on difficult and unreliable predictions. Conformal prediction provides well-calibrated predictions where the accuracy of the predictions is equal to a user-set confidence level. To achieve this, the conformal predictor produces more conservative prediction sets (classification) compared to the point predictions made by conventional prediction models. The calibration of prediction models is particularly important in addition to the discriminative performance of the model. Calibrated risk estimates are essential for informing patients about reliable risk estimates that are not over- or underestimated (e.g., not giving patients false hope). A well-calibrated model with a lower AUC might be clinically preferable compared to a model with better discrimination that provides poorly calibrated output [92]. The use of more diversified training data will lead to improved generalizability of the performance. Nonetheless, regardless of how good the AI models are, there is always a chance that the models may be exposed to data outside its reliable prediction

domain. This is where we believe that the use of conformal prediction can function as a quality control step to facilitate the safe clinical implementation of AI systems so that unreliable predictions can be detected.

Chapter 8

Ethical considerations

We have ethical approval for all studies from the Ethical Review Board in Stockholm, and all study participants need to provide informed consent, except for Study I, which is a registry-based study. Information that can be used to identify a participant is always kept separate from other data. We use individual-level data in accordance with the General Data Protection Regulation (GDPR), the Swedish law, and KI-guidelines to ensure that information is managed securely. Personal information was handled in compliance with the Swedish Personal Data Act (1998:204). The study participant's blood samples are treated in accordance with the Swedish Biobank Act.

The paired design in Study II enabled us to compare all possible combinations of interventions in an efficient way within each patient. However, it also imposes additional testing on these individuals, with blood tests, an MRI, and two different biopsy schemes. This additional testing raises some ethical questions. The participants were invited to the study by their urologist at their yearly follow-up. Men with a systematic biopsy within 12 months did not undergo systematic biopsies, and the previous biopsy was analyzed. Men with a previous MRI within the past 12 months did not undergo a new MRI within the study. This synchronization minimized the need for additional biopsies and MRIs, and we used the investigations that were already planned within the AS surveillance program for each individual. Due to the paired design used, the risk of missing aggressive cancers is small. Data from the STHLM3 study show that the Stockholm3-test can identify as many or more aggressive prostate cancer compared to PSA. According to existing evidence, targeted biopsies find more high-grade cancers than systematic biopsies. There might be instances where small clinically significant tumors are missed since they are not visible on MRI. However, systematic biopsies were also performed on these men, with the possibility of finding small tumors not visible on MRI. We argue that this design leads to fewer ethical dilemmas than if we had randomly assigned participants to separate surveillance arms. With randomization, each participant would receive only one surveillance scheme, where one might be more beneficial than the other.

Current prostate pathology is associated with a number of challenges. Such as an

increase in the number of performed biopsies and analyzed globally, a global shortage of experienced uro-pathologists, and high inter-reader variability among pathologists. AI in medicine has the potential to alleviate some of these problems by providing decision support to pathologists and diagnostic support in parts of the world where there might be a lack of experienced pathologists. However, the implementation of artificial intelligence in medicine will undoubtedly raise a number of ethical concerns in the future. Such as the question about responsibility and how these systems should be able to account for patients' unique characteristics and conditions. As a result, there is current resistance to medical AI implementations among patients and physicians. But the research in this thesis is still in the proof-of-principle phase and does not currently raise any direct ethical issues about AI in medicine.

Chapter 9

Conclusions

Prostate cancer screening, whether unorganized or organized, will lead to the detection of many localized tumors that are unlikely to cause symptoms in the diagnosed man or be the primary cause of death. AS plays a key role in balancing overdiagnosis and overtreatment of such cancers. The pathological evaluation of prostate biopsies determines the therapeutic course of treatment for prostate cancer patients. However, current clinical practice faces challenges such as a high level of inter-observer variability among pathologists and a global shortage of uro-pathologists. Prostate pathology may benefit from the development of AI technology in digital pathology, which also aims to lessen inter- and intra-observer variability.

More specifically we conclude the following:

- Study I described active surveillance for low-risk prostate cancer in Stockholm between 2008 and 2017. The results point to the need for less invasive methods for active surveillance to improve uptake and follow-up.
- In Study II, we performed the STHLM3-AS trial to investigate the potential use of MRI and risk prediction models in the follow-up of AS. The results show that the Stockholm 3 test and MRI have the potential to provide less invasive and more cost-effective monitoring for AS.
- In Study III, an AI model was developed for cancer detection and grading of prostate biopsies. The results show that computational pathology can assist pathologists in prostate cancer diagnostics with high accuracy, which has the potential to reduce inter-observer variability and provide high-accuracy diagnostics in parts of the world where there is a shortage of prostate pathology specialists.
- Even though computational pathology shows promise in prostate cancer diagnostics, the clinical implementation of these models is not trivial. The performance of the models is highly sensitive to variations in the underlying data, and the generalizability of AI models to external data is currently unsolved. In Study IV,

we implemented and used conformal prediction to assess the reliability of the predictions of the AI model from Study 3. The conformal predictor to detect unreliable predictions introduced by variations in tissue preparation, digital scanners, and the existence of unusual prostate tissue that the model was not exposed to during training.

- In study V, we investigated whether deep learning models could learn and detect possible relationships between tissue morphology and prostate genetic risk factors. The findings offer proof of principle for investigating the potential of incorporating morphological information in multi-modal prostate cancer risk prediction models.

Acknowledgements

I would like to thank all the amazing people that contributed to this thesis, and that supported me throughout the years

I would like to express my sincere gratitude to the following:

To my main supervisor, **Martin Eklund**. Working with you has truly been inspiring, and I have learned a lot from you. Always encouraging me to be creative and explore my ideas, while providing your brilliant instincts about when to do what, how to push on to new levels, and guiding us in the right direction if necessary. I also want to thank you for being such a kind, goodhearted, and supportive supervisor. Having had two children during my studies, I appreciate our shared value that family always comes first.

Anna Lantz, my co-supervisor, you have taught me so many things about prostate cancer and how to perform good science. I also want to thank you for always being there, always being the first to answer my emails, and to offer support when needed. Thank you for letting me sit in on a prostate cancer surgery, which provided insight into the clinical work and context for our research. I am deeply impressed by how you manage to combine clinical and research work so efficiently.

To my co-supervisor **Mark Clements**. When I first arrived at KI many years ago, you were one of the first to introduce not only yourself but also everyone else in the room in a 5-minute crash course. This person knows this and that, is a brilliant statistician, wrote a book, is super smart, etc. I instantly felt the positive and kind energy that you spread around you. Hearing your g'day in the morning always brings joy! You are also one of the most knowledgeable statisticians and scientists that I know, and you always raise the quality of the research.

To **Henrik Grönberg**, for being an inspiration as a researcher and leader, and thank you for the fantastic research environment that you provide.

To **Lars Egevad**, thank you for the immense work and expertise that you provide

us with, and for always taking the time to explain and educate us on prostate pathology.

Thank you to the wonderful co-workers and friends in the prostate cancer research group. **Kimmo Kartasalo, Peter Ström, Thorgerdur Palsdottir, Andreas Karlsson, Nita Mulliqi, Xiaoyi Ji, Lars Björnebo, Alessio Crippa, Andrea Discacciati, Hari Vigneswaran, Venkatesh Chellappa, Rebecca Bergström, Berit Larsson, Chiara Micoli, Fredrik Wiklund, Tobias Nordström**. You are all fantastic researchers who are also kind and generous. The achievements of our group are truly the results of the combined amazing efforts of each individual.

To **Astrid Björklund, Carin Cavalli-Björkman, Britt-marie Hune**, thank you for your amazing work, your friendship, and your interest in our work.

To **Alessandra Nanni, Anna Berglund, Gunilla Nilsson Roos, Marie Jansson**, for your guidance with all the administrative proceedings at KI. You made my life a lot simpler. Thank you!

Thank you to current and previous **Heads of Department** for making MEB such a great place to do research.

I would like to thank present and former members of the fantastic biostatistics group at MEB. **Cecilia Lundholm, Robert Karlsson, Therese Andersson, Henrik Winell, Ralf Kuja-Halkola, Sven Sandin, Anna Johansson, Alexander Ploner, Arvid Sjölander, Paul Dickman, Keith Humphreys, Yudi Pawitan, Marie Reilly and Li Yin**. I have learned a lot from you!

I also want to thank my present and former fellow PhD students, friends, and colleagues at MEB, for all the fikas, lunches, and social activities and for making this a fantastic workplace! A special thanks to, **Elisabeth Dahlqwist, Shuang Hao, Caroline Weibull, Gabriel Isheden, Robert Szulkin, Hannah Bower, Rickard Strandberg, Flaminia Chiesa, Johan Zetterquist, Xingrong Liu, Anna Plym, Johanna Holm, Andreas Jangmo, Emilio Ugalde Morales, Tong Gong, Phillipe Weitz, Yinxi Wang, Mattias Rantalainen**.

To all my lovely friends that I met during my student life at Umeå University: **Oskar Törnmarck, Albin Hortell, Mikael Johansson, Gustaf Malmström, Johannes Wikström, Erik Säterborg, Sebastian Geijer, Eva Hagel, and David Hägg**. One of the best decisions that I have ever made was to move to Umeå and get to meet all of you lovely people! Thank you! A special thanks to **Oskar Törnmarck**! I believe it is safe to say, on behalf of all of us, that you are such a big reason for how we managed

to stay together after all these years. You are the glue! I also want to thank you for keeping me "somewhat" physically fit this past year with all the tennis matches that provided some much-needed breaks from writing.

I want to thank my old childhood friends back home from Trollhättan, **Mathias Lindell, Viktor Lundberg, Robert Carlsson, Per Sandblom**. We do not see each other as often as I would like. But when we do it is like no time has passed and it always feels like coming home again. Love you guys!

To my beloved parents, **Lena** and **Mikael**, and dear sister **Karin**. We have been through a lot together in life, and whatever it is, I know we'll figure it out together. You gave me the best of childhoods and now you give Rut and Eva all that love too!

Finally, thanks to my beloved family. **Matilda** and our wonderful daughters, **Rut** and **Eva**, you are the most important part of my life, and everything else is secondary. Matilda, I wouldn't have been able to do this without you. I love our life together and I love you!

References

[1] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70:7–30, 1 2020.

[2] F Lam, M Colombet, L Mery, M Pineros, A Znaor, I Soerjomataram, F Bray, J Ferlay, and M Ervik. Global cancer observatory: Cancer today. http://gco.iarc.fr/today. Lyon, France: International Agency for Research on Cancer. Online accessed: 22.09.2022.

[3] H. Lilja. A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein. *Journal of Clinical Investigation*, 76:1899–1903, 1985.

[4] Fritz H Schröder, Jonas Hugosson, Monique J Roobol, Teuvo L J Tammela, Marco Zappa, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Liisa Määttänen, Hans Lilja, Louis J Denis, Franz Recker, Alvaro Paez, Chris H Bangma, Sigrid Carlsson, Donella Puliti, Arnauld Villers, Xavier Rebillard, Matti Hakama, Ulf-Hakan Stenman, Paula Kujala, Kimmo Taari, Gunnar Aus, Andreas Huber, Theo H van der Kwast, Ron H N van Schaik, Harry J de Koning, Sue M Moss, and Anssi Auvinen. Screening and prostate cancer mortality: results of the european randomised study of screening for prostate cancer (erspc) at 13 years of follow-up. *The Lancet*, 384:2027–2035, 12 2014.

[5] P. Stattin, S. Carlsson, B. Holmstrom, A. Vickers, J. Hugosson, H. Lilja, and H. Jonsson. Prostate cancer mortality in areas with high and low prostate cancer incidence. *JNCI Journal of the National Cancer Institute*, 106:dju007–dju007, 3 2014.

[6] K. Mistry and G. Cable. Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. *The Journal of the American Board of Family Medicine*, 16:95–101, 3 2003.

[7] Stacy Loeb, Marc A. Bjurlin, Joseph Nicholson, Teuvo L. Tammela, David F. Penson, H. Ballentine Carter, Peter Carroll, and Ruth Etzioni. Overdiagnosis and overtreatment of prostate cancer. *European Urology*, 65:1046–1055, 2014.

[8] Jonathan I. Epstein, Zhaoyong Feng, Bruce J. Trock, and Phillip M. Pierorazio. Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: Incidence and predictive factors using the modified gleason grading system and factoring in tertiary grades. *European Urology*, 61:1019–1024, 2012.

[9] Ricardo Soares, Antonina Di Benedetto, Zach Dovey, Simon Bott, Roy G. McGregor, and Christopher G. Eden. Minimum 5-year follow-up of 1138 consecutive laparoscopic radical prostatectomies. *BJU International*, 115:546–553, 4 2015.

[10] Jonathan I. Epstein, Mahul B. Amin, Victor E. Reuter, and Peter A. Humphrey. Contemporary gleason grading of prostatic carcinoma. *American Journal of Surgical Pathology*, 41:e1–e7, 4 2017.

[11] Jonathan I Epstein, William C Allsbrook, Mahul B Amin, and Lars L Egevad. The 2005 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *American Journal of Surgical Pathology*, 29:1228–1242, 9 2005.

[12] L. E. Johns and R. S. Houlston. A systematic review and meta-analysis of familial prostate cancer risk, 6 2003.

[13] Andreas Brandt, Justo Lorenzo Bermejo, Jan Sundquist, and Kari Hemminki. Age-specific risk of incident prostate cancer and risk of death from prostate cancer defined by the number of affected family members. *European Urology*, 58:275–280, 8 2010.

[14] Rosalind Eeles, Chee Goh, Elena Castro, Elizabeth Bancroft, Michelle Guy, Ali Amin Al Olama, Douglas Easton, and Zsofia Kote-Jarai. The genetic epidemiology of prostate cancer and its clinical implications, 1 2014.

[15] Mok Oh, Nasser Alkhushaym, Saad Fallatah, Abdulhamid Althagafi, Rana Aljadeed, Yazed Alsowaida, Joanne Jeter, Jennifer R. Martin, Hani M. Babiker, Ali McBride, and Ivo Abraham. The association of brca1 and brca2 mutations with prostate cancer risk, frequency, and mortality: A meta-analysis. *Prostate*, 79:880–895, 6 2019.

[16] Tommy Nyberg, Koveela Govindasami, Goska Leslie, Tokhir Dadaev, Elizabeth Bancroft, Holly Ni Raghallaigh, Mark N. Brook, Nafisa Hussain, Diana Keating, Andrew Lee, Romayne McMahon, Angela Morgan, Andrea Mullen, Andrea Osborne, Reshma Rageevakumar, Zsofia Kote-Jarai, Rosalind Eeles, and Antonis C. Antoniou. Homeobox b13 g84e mutation and prostate cancer risk. *European Urology*, 75:834–845, 5 2019.

[17] David V. Conti, Burcu F. Darst, Lilit C. Moss, Edward J. Saunders, Xin Sheng, Alisha Chou, Fredrick R. Schumacher, Ali Amin Al Olama, Sara Benlloch, Tokhir Dadaev,

Mark N. Brook, Ali Sahimi, Thomas J. Hoffmann, Atushi Takahashi, Koichi Matsuda, Yukihide Momozawa, Masashi Fujita, Kenneth Muir, Artitaya Lophatananon, Peggy Wan, Loic Le Marchand, Lynne R. Wilkens, Victoria L. Stevens, Susan M. Gapstur, Brian D. Carter, Johanna Schleutker, Teuvo L.J. Tammela, Csilla Sipeky, Anssi Auvinen, Graham G. Giles, Melissa C. Southey, Robert J. MacInnis, Cezary Cybulski, Dominika Wokołorczyk, Jan Lubiński, David E. Neal, Jenny L. Donovan, Freddie C. Hamdy, Richard M. Martin, Børge G. Nordestgaard, Sune F. Nielsen, Maren Weischer, Stig E. Bojesen, Martin Andreas Røder, Peter Iversen, Jyotsna Batra, Suzanne Chambers, Leire Moya, Lisa Horvath, Judith A. Clements, Wayne Tilley, Gail P. Risbridger, Henrik Gronberg, Markus Aly, Robert Szulkin, Martin Eklund, Tobias Nordström, Nora Pashayan, Alison M. Dunning, Maya Ghoussaini, Ruth C. Travis, Tim J. Key, Elio Riboli, Jong Y. Park, Thomas A. Sellers, Hui Yi Lin, Demetrius Albanes, Stephanie J. Weinstein, Lorelei A. Mucci, Edward Giovannucci, Sara Lindstrom, Peter Kraft, David J. Hunter, Kathryn L. Penney, Constance Turman, Catherine M. Tangen, Phyllis J. Goodman, Ian M. Thompson, Robert J. Hamilton, Neil E. Fleshner, Antonio Finelli, Marie Élise Parent, Janet L. Stanford, Elaine A. Ostrander, Milan S. Geybels, Stella Koutros, Laura E.Beane Freeman, Meir Stampfer, Alicja Wolk, Niclas Håkansson, Gerald L. Andriole, Robert N. Hoover, Mitchell J. Machiela, Karina Dalsgaard Sørensen, Michael Borre, William J. Blot, Wei Zheng, Edward D. Yeboah, James E. Mensah, Yong Jie Lu, Hong Wei Zhang, Ninghan Feng, Xueying Mao, Yudong Wu, Shan Chao Zhao, Zan Sun, Stephen N. Thibodeau, Shannon K. McDonnell, Daniel J. Schaid, Catharine M.L. West, Neil Burnet, Gill Barnett, Christiane Maier, Thomas Schnoeller, Manuel Luedeke, Adam S. Kibel, Bettina F. Drake, Olivier Cussenot, Géraldine Cancel-Tassin, Florence Menegaux, Thérèse Truong, Yves Akoli Koudou, Esther M. John, Eli Marie Grindedal, Lovise Maehle, Kay Tee Khaw, Sue A. Ingles, Mariana C. Stern, Ana Vega, Antonio Gómez-Caamaño, Laura Fachal, Barry S. Rosenstein, Sarah L. Kerns, Harry Ostrer, Manuel R. Teixeira, Paula Paulo, Andreia Brandão, Stephen Watya, Alexander Lubwama, Jeannette T. Bensen, Elizabeth T.H. Fontham, James Mohler, Jack A. Taylor, Manolis Kogevinas, Javier Llorca, Gemma Castaño-Vinyals, Lisa Cannon-Albright, Craig C. Teerlink, Chad D. Huff, Sara S. Strom, Luc Multigner, Pascal Blanchet, Laurent Brureau, Radka Kaneva, Chavdar Slavov, Vanio Mitev, Robin J. Leach, Brandi Weaver, Hermann Brenner, Katarina Cuk, Bernd Holleczek, Kai Uwe Saum, Eric A. Klein, Ann W. Hsing, Rick A. Kittles, Adam B. Murphy, Christopher J. Logothetis, Jeri Kim, Susan L. Neuhausen, Linda Steele, Yuan Chun Ding, William B. Isaacs, Barbara Nemesure, Anselm J.M. Hennis, John Carpten, Hardev Pandha, Agnieszka Michael, Kim De Ruyck, Gert De Meerleer, Piet Ost, Jianfeng Xu, Azad Razack, Jasmine Lim, Soo Hwang Teo, Lisa F. Newcomb, Daniel W. Lin, Jay H. Fowke, Christine Neslund-Dudas, Benjamin A. Rybicki, Marija Gamulin, Davor Lessel, Tomislav Kulis, Nawaid Usmani, Sandeep

Singhal, Matthew Parliament, Frank Claessens, Steven Joniau, Thomas Van den Broeck, Manuela Gago-Dominguez, Jose Esteban Castelao, Maria Elena Martinez, Samantha Larkin, Paul A. Townsend, Claire Aukim-Hastie, William S. Bush, Melinda C. Aldrich, Dana C. Crawford, Shiv Srivastava, Jennifer C. Cullen, Gyorgy Petrovics, Graham Casey, Monique J. Roobol, Guido Jenster, Ron H.N. van Schaik, Jennifer J. Hu, Maureen Sanderson, Rohit Varma, Roberta McKean-Cowdin, Mina Torres, Nicholas Mancuso, Sonja I. Berndt, Stephen K. Van Den Eeden, Douglas F. Easton, Stephen J. Chanock, Michael B. Cook, Fredrik Wiklund, Hidewaki Nakagawa, John S. Witte, Rosalind A. Eeles, Zsofia Kote-Jarai, and Christopher A. Haiman. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nature Genetics*, 53:65–75, 1 2021.

[18] Ali Amin Al Olama, Zsofia Kote-Jarai, Sonja I. Berndt, David V. Conti, Fredrick Schumacher, Ying Han, Sara Benlloch, Dennis J. Hazelett, Zhaoming Wang, Ed Saunders, Daniel Leongamornlert, Sara Lindstrom, Sara Jugurnauth-Little, Tokhir Dadaev, Malgorzata Tymrakiewicz, Daniel O. Stram, Kristin Rand, Peggy Wan, Alex Stram, Xin Sheng, Loreall C. Pooler, Karen Park, Lucy Xia, Jonathan Tyrer, Laurence N. Kolonel, Loic Le Marchand, Robert N. Hoover, Mitchell J. Machiela, Merideth Yeager, Laurie Burdette, Charles C. Chung, Amy Hutchinson, Kai Yu, Chee Goh, Mahbubl Ahmed, Koveela Govindasami, Michelle Guy, Teuvo L.J. Tammela, Anssi Auvinen, Tiina Wahlfors, Johanna Schleutker, Tapio Visakorpi, Katri A. Leinonen, Jianfeng Xu, Markus Aly, Jenny Donovan, Ruth C. Travis, Tim J. Key, Afshan Siddiq, Federico Canzian, Kay Tee Khaw, Atsushi Takahashi, Michiaki Kubo, Paul Pharoah, Nora Pashayan, Maren Weischer, Borge G. Nordestgaard, Sune F. Nielsen, Peter Klarskov, Martin Andreas Røder, Peter Iversen, Stephen N. Thibodeau, Shannon K. McDonnell, Daniel J. Schaid, Janet L. Stanford, Suzanne Kolb, Sarah Holt, Beatrice Knudsen, Antonio Hurtado Coll, Susan M. Gapstur, W. Ryan Diver, Victoria L. Stevens, Christiane Maier, Manuel Luedeke, Kathleen Herkommer, Antje E. Rinckleb, Sara S. Strom, Curtis Pettaway, Edward D. Yeboah, Yao Tettey, Richard B. Biritwum, Andrew A. Adjei, Evelyn Tay, Ann Truelove, Shelley Niwa, Anand P. Chokkalingam, Lisa Cannon-Albright, Cezary Cybulski, Dominika Wokołorczyk, Wojciech Kluźniak, Jong Park, Thomas Sellers, Hui Yi Lin, William B. Isaacs, Alan W. Partin, Hermann Brenner, Aida Karina Dieffenbach, Christa Stegmaier, Constance Chen, Edward L. Giovannucci, Jing Ma, Meir Stampfer, Kathryn L. Penney, Lorelei Mucci, Esther M. John, Sue A. Ingles, Rick A. Kittles, Adam B. Murphy, Hardev Pandha, Agnieszka Michael, Andrzej M. Kierzek, William Blot, Lisa B. Signorello, Wei Zheng, Demetrius Albanes, Jarmo Virtamo, Stephanie Weinstein, Barbara Nemesure, John Carpten, Cristina Leske, Suh Yuh Wu, Anselm Hennis, Adam S. Kibel, Benjamin A. Rybicki, Chris-

tine Neslund-Dudas, Ann W. Hsing, Lisa Chu, Phyllis J. Goodman, Eric A. Klein, S. Lilly Zheng, Jyotsna Batra, Judith Clements, Amanda Spurdle, Manuel R. Teixeira, Paula Paulo, Sofia Maia, Chavdar Slavov, Radka Kaneva, Vanio Mitev, John S. Witte, Graham Casey, Elizabeth M. Gillanders, Daniella Seminara, Elio Riboli, Freddie C. Hamdy, Gerhard A. Coetzee, Qiyuan Li, Matthew L. Freedman, David J. Hunter, Kenneth Muir, Henrik Gronberg, David E. Neal, Melissa Southey, Graham G. Giles, Gianluca Severi, Michael B. Cook, Hidewaki Nakagawa, Fredrik Wiklund, Peter Kraft, Stephen J. Chanock, Brian E. Henderson, Douglas F. Easton, Rosalind A. Eeles, and Christopher A. Haiman. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nature Genetics*, 46:1103–1109, 9 2014.

[19] S. Lilly Zheng, Jielin Sun, Fredrik Wiklund, Shelly Smith, Pär Stattin, Ge Li, Hans-Olov Adami, Fang-Chi Hsu, Yi Zhu, Katarina Bälter, A. Karim Kader, Aubrey R. Turner, Wennuan Liu, Eugene R. Bleecker, Deborah A. Meyers, David Duggan, John D. Carpten, Bao-Li Chang, William B. Isaacs, Jianfeng Xu, and Henrik Grönberg. Cumulative association of five genetic variants with prostate cancer. *New England Journal of Medicine*, 358:910–919, 2 2008.

[20] Markus Aly, Fredrik Wiklund, Jianfeng Xu, William B. Isaacs, Martin Eklund, Mauro D'Amato, Jan Adolfsson, and Henrik Grönberg. Polygenic risk score improves prostate cancer risk prediction: Results from the stockholm-1 cohort study. *European Urology*, 60:21–28, 7 2011.

[21] Minh-Phuong Huynh-Le, Roshan Karunamuni, Chun Chieh Fan, Lui Asona, Wesley K. Thompson, Maria Elena Martinez, Rosalind A. Eeles, Zsofia Kote-Jarai, Kenneth R. Muir, Artitaya Lophatananon, Johanna Schleutker, Nora Pashayan, Jyotsna Batra, Henrik Grönberg, David E. Neal, Børge G. Nordestgaard, Catherine M. Tangen, Robert J. MacInnis, Alicja Wolk, Demetrius Albanes, Christopher A. Haiman, Ruth C. Travis, William J. Blot, Janet L. Stanford, Lorelei A. Mucci, Catharine M. L. West, Sune F. Nielsen, Adam S. Kibel, Olivier Cussenot, Sonja I. Berndt, Stella Koutros, Karina Dalsgaard Sørensen, Cezary Cybulski, Eli Marie Grindedal, Florence Menegaux, Jong Y. Park, Sue A. Ingles, Christiane Maier, Robert J. Hamilton, Barry S. Rosenstein, Yong-Jie Lu, Stephen Watya, Ana Vega, Manolis Kogevinas, Fredrik Wiklund, Kathryn L. Penney, Chad D. Huff, Manuel R. Teixeira, Luc Multigner, Robin J. Leach, Hermann Brenner, Esther M. John, Radka Kaneva, Christopher J. Logothetis, Susan L. Neuhausen, Kim De Ruyck, Piet Ost, Azad Razack, Lisa F. Newcomb, Jay H. Fowke, Marija Gamulin, Aswin Abraham, Frank Claessens, Jose Esteban Castelao, Paul A. Townsend, Dana C. Crawford, Gyorgy Petrovics, Ron H. N. van Schaik, Marie Élise Parent, Jennifer J. Hu, Wei Zheng, Ian G. Mills, Ole A. Andreassen, Anders M. Dale, and Tyler M. Seibert.

Prostate cancer risk stratification improvement across multiple ancestries with new polygenic hazard score. *Prostate Cancer and Prostatic Diseases*, 2 2022.

[22] Jan-Erik Johansson, Ove Andren, Swen-Olof Andersson, Paul W Dickman, Lars Holmberg, Anders Magnuson, Hans-Olov Adami, Ove Andre, and Paul W Dickman. Natural history of early , localized. *Jama*, 291:2713–2719, 2014.

[23] Brian Kessler and Peter Albertsen. The natural history of prostate cancer. *Urologic Clinics of North America*, 30:219–226, 2003.

[24] Prostatacancer, Nationelltvårdprogram. Version: 7.0. https://cancercentrum.se/samverkan/cancerdiagnoser/prostata/vardprogram/. Online accessed: 10.10.2022, 2022.

[25] Eau guidelines. edn. presented at the EAU annual congress amsterdam 2022. isbn 978-94-92671-16-5. https://uroweb.org/guidelines/prostate-cancer. Online accessed: 10.10.2022, 2022.

[26] Jeffrey J. Tosoian, Mufaddal Mamawala, Jonathan I. Epstein, Patricia Landis, Katarzyna J. Macura, Demetrios N. Simopoulos, H. Ballentine Carter, and Michael A. Gorin. Active surveillance of grade group 1 prostate cancer: Long-term outcomes from a large prospective cohort. *European Urology*, 77:675–682, 6 2020.

[27] Sigrid Carlsson, Nicole Benfante, Ricardo Alvim, Daniel D. Sjoberg, Andrew Vickers, Victor E. Reuter, Samson W. Fine, Hebert Alberto Vargas, Michal Wiseman, Maha Mamoor, Behfar Ehdaie, Vincent Laudone, Peter Scardino, James Eastham, and Karim Touijer. Long-term outcomes of active surveillance for prostate cancer: The memorial sloan kettering cancer center experience. *The Journal of urology*, 203:1122–1127, 6 2020.

[28] Rebecka Arnsrud Godtman, Erik Holmberg, Ali Khatami, Carl Gustaf Pihl, Johan Stranne, and Jonas Hugosson. Long-term results of active surveillance in the göteborg randomized, population-based prostate cancer screening trial. *European Urology*, 70:760–766, 2016.

[29] Timothy J. Wilt, Michael K. Brawer, Karen M. Jones, Michael J. Barry, William J. Aronson, Steven Fox, Jeffrey R. Gingrich, John T. Wei, Patricia Gilhooly, B. Mayer Grob, Imad Nsouli, Padmini Iyer, Ruben Cartagena, Glenn Snider, Claus Roehrborn, Roohollah Sharifi, William Blank, Parikshit Pandya, Gerald L. Andriole, Daniel Culkin, and Thomas Wheeler. Radical prostatectomy versus observation for localized prostate cancer. *New England Journal of Medicine*, 367:203–213, 7 2012.

[30] Freddie C Hamdy, Jenny L Donovan, J Athene Lane, Malcolm Mason, Chris Metcalfe, Peter Holding, Michael Davis, Tim J Peters, Emma L Turner, Richard M

Martin, Jon Oxley, Mary Robinson, John Staffurth, Eleanor Walsh, Prasad Bollina, James Catto, Andrew Doble, Alan Doherty, David Gillatt, Roger Kockelbergh, Howard Kynaston, Alan Paul, Philip Powell, Stephen Prescott, Derek J Rosario, Edward Rowe, and David E Neal. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer, 2016.

[31] Marco Borghesi, Hashim Ahmed, Robert Nam, Edward Schaeffer, Riccardo Schiavina, Samir Taneja, Wolfgang Weidner, and Stacy Loeb. Complications after systematic, random, and image-guided prostate biopsy [figure presented]. *European Urology*, 71:353–365, 2017.

[32] Dragan Ilic, Sue M Evans, Christie Ann Allan, Jae Hung Jung, Declan Murphy, and Mark Frydenberg. Laparoscopic and robotic-assisted versus open radical prostatectomy for the treatment of localised prostate cancer. *The Cochrane database of systematic reviews*, 9:CD009625, 2017.

[33] Eva Haglind, Stefan Carlsson, Johan Stranne, Anna Wallerstedt, Ulrica WilderÃ€ng, Thordis Thorsteinsdottir, Mikael Lagerkvist, Jan-Erik Damber, Anders Bjartell, Jonas Hugosson, Peter Wiklund, Gunnar Steineck, and LAPPRO steering committee. Urinary incontinence and erectile dysfunction after robotic versus open radical prostatectomy: A prospective, controlled, nonrandomised trial. *European urology*, 68:216–25, 2015.

[34] Anders Widmark, OlbjÃžrn Klepp, Arne Solberg, Jan-Erik Damber, Anders Angelsen, Per Fransson, Jo Ãsmund Lund, Ilker Tasdemir, Morten Hoyer, Fredrik Wiklund, and Sophie D FossÃ¥. Endocrine treatment, with or without radiotherapy, in locally advanced prostate cancer (spcg-7/sfuo-3): an open randomised phase iii trial. *The Lancet*, 373:301–308, 1 2009.

[35] James W Denham, Allison Steigler, David S Lamb, David Joseph, Sandra Turner, John Matthews, Chris Atkinson, John North, David Christie, Nigel A Spry, Keen-Hun Tai, Chris Wynne, and Catherine D'Este. Short-term neoadjuvant androgen deprivation and radiotherapy for locally advanced prostate cancer: 10-year data from the trog 96.01 randomised trial. *The Lancet Oncology*, 12:451–459, 5 2011.

[36] Malcolm D. Mason, Wendy R. Parulekar, Matthew R. Sydes, Michael Brundage, Peter Kirkbride, Mary Gospodarowicz, Richard Cowan, Edmund C. Kostashuk, John Anderson, Gregory Swanson, Mahesh K.B. Parmar, Charles Hayter, Gordana Jovic, Andrea Hiltz, John Hetherington, Jinka Sathya, James B.P. Barber, Michael McKenzie, Salah El-Sharkawi, Luis Souhami, P.D. John Hardman, Bingshu E. Chen, and Padraig Warde. Final report of the intergroup randomized study of combined androgen-deprivation therapy plus radiotherapy versus androgen-deprivation

therapy alone in locally advanced prostate cancer. *Journal of Clinical Oncology*, 33:2143–2150, 7 2015.

[37] Fritz H. Schröder, Jonas Hugosson, Monique J. Roobol, Teuvo L.J. Tammela, Stefano Ciatto, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Hans Lilja, Marco Zappa, Louis J. Denis, Franz Recker, Antonio Berenguer, Liisa Määttänen, Chris H. Bangma, Gunnar Aus, Arnauld Villers, Xavier Rebillard, Theodorus van der Kwast, Bert G. Blijenberg, Sue M. Moss, Harry J. de Koning, and Anssi Auvinen. Screening and prostate-cancer mortality in a randomized european study. *New England Journal of Medicine*, 360:1320–1328, 3 2009.

[38] Jeffrey J. Tosoian, Sasha C. Druskin, Darian Andreas, Patrick Mullane, Meera Chappidi, Sarah Joo, Kamyar Ghabili, Mufaddal Mamawala, Joseph Agostino, Herbert B. Carter, Alan W. Partin, Lori J. Sokoll, and Ashley E. Ross. Prostate health index density improves detection of clinically significant prostate cancer. *BJU International*, 120:793–798, 12 2017.

[39] 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*, 68:464–470, 9 2015.

[40] Henrik Grönberg, Jan Adolfsson, Markus Aly, Tobias Nordström, Peter Wiklund, Yvonne Brandberg, James Thompson, Fredrik Wiklund, Johan Lindberg, Mark Clements, Lars Egevad, and Martin Eklund. Prostate cancer screening in men aged 50-69 years (sthlm3): A prospective population-based diagnostic study. *The Lancet Oncology*, 16:1667–1676, 2015.

[41] 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*, 179:1587–1592, 4 2008.

[42] Veeru Kasivisvanathan, Antti S. Rannikko, Marcelo Borghi, Valeria Panebianco, Lance A. Mynderse, Markku H. Vaarala, Alberto Briganti, Lars BudÃ€us, Giles Hellawell, Richard G. Hindley, Monique J. Roobol, Scott Eggener, Maneesh Ghei, Arnauld Villers, Franck Bladou, Geert M. Villeirs, Jaspal Virdi, Silvan Boxler, GrÃ©goire Robert, Paras B. Singh, Wulphert Venderink, Boris A. Hadaschik,

Alain Ruffion, Jim C. Hu, Daniel Margolis, SÃ©bastien Crouzet, Laurence Klotz, Samir S. Taneja, Peter Pinto, Inderbir Gill, Clare Allen, Francesco Giganti, Alex Freeman, Stephen Morris, Shonit Punwani, Norman R. Williams, Chris Brew-Graves, Jonathan Deeks, Yemisi Takwoingi, Mark Emberton, and Caroline M. Moore. Mri-targeted or standard biopsy for prostate-cancer diagnosis. *New England Journal of Medicine*, page NEJMoa1801993, 2018.

[43] Olivier RouviÃšre, Philippe Puech, RaphaÃ«le Renard-Penna, Michel Claudon, Catherine Roy, Florence MÃšge-Lechevallier, Myriam Decaussin-Petrucci, Marine Dubreuil-Chambardel, Laurent Magaud, Laurent Remontet, Alain Ruffion, Marc Colombel, SÃ©bastien Crouzet, Anne Marie Schott, Laurent Lemaitre, Muriel Rabilloud, Nicolas Grenier, Nicolas Barry Delongchamps, Romain Boutier, Flavie Bratan, Serge Brunelle, Philippe Camparo, Pierre Colin, Jean Michel CorrÃ©as, FranÃ§ois CornÃ©lis, FranÃ§ois Cornud, Fanny Cros, Jean Luc Descotes, Pascal Eschwege, Gaelle Fiard, Jean Philippe Fendler, Hocine Habchi, Philippe Hallouin, Ahmed Khairoune, HervÃ© Lang, Yann Lebras, FrÃ©dÃ©ric LefÃšvre, Bernard Malavaud, Paul Cezar Moldovan, Nicolas Mottet, Pierre Mozer, Pierre Nevoux, Gaele Pagnoux, Gilles Pasticier, Daniel Portalez, Eric Potiron, Athivada Soto Thammavong, Marc Olivier Timsit, Arnault Viller, and Jochen Walz. Use of prostate systematic and targeted biopsy on the basis of multiparametric mri in biopsy-naive patients (mri-first): a prospective, multicentre, paired diagnostic study. *The Lancet Oncology*, 20:100–109, 2019.

[44] Michael Ahdoot, Andrew R. Wilbur, Sarah E. Reese, Amir H. Lebastchi, Sherif Mehralivand, Patrick T. Gomella, Jonathan Bloom, Sandeep Gurram, Minhaj Siddiqui, Paul Pinsky, Howard Parnes, W. Marston Linehan, Maria Merino, Peter L. Choyke, Joanna H. Shih, Baris Turkbey, Bradford J. Wood, and Peter A. Pinto. Mri-targeted, systematic, and combined biopsy for prostate cancer diagnosis. *New England Journal of Medicine*, 382:917–928, 2020.

[45] Frank-Jan H Drost, DaniÃ«l F Osses, Daan Nieboer, Ewout W Steyerberg, Chris H Bangma, Monique J Roobol, and Ivo G Schoots. Prostate mri, with or without mri-targeted biopsy, and systematic biopsy for detecting prostate cancer. *Cochrane Database of Systematic Reviews*, 4 2019.

[46] Martin Eklund, Fredrik JÃ€derling, Andrea Discacciati, Martin Bergman, Magnus Annerstedt, Markus Aly, Axel Glaessgen, Stefan Carlsson, Henrik GrÃ¶nberg, and Tobias NordstrÃ¶m. Mri-targeted or standard biopsy in prostate cancer screening. *New England Journal of Medicine*, 385:908–920, 9 2021.

[47] David Eldred-Evans, Paula Burak, Martin J. Connor, Emily Day, Martin Evans, Francesca Fiorentino, Martin Gammon, Feargus Hosking-Jervis, Natalia

Klimowska-Nassar, William McGuire, Anwar R. Padhani, A. Toby Prevost, Derek Price, Heminder Sokhi, Henry Tam, Mathias Winkler, and Hashim U. Ahmed. Population-based prostate cancer screening with magnetic resonance imaging or ultrasonography: The ip1-prostagram study. *JAMA Oncology*, 7:395–402, 3 2021.

[48] Tobias NordstrÃ¶m, Andrea Discacciati, Martin Bergman, Mark Clements, Markus Aly, Magnus Annerstedt, Axel Glaessgen, Stefan Carlsson, Fredrik JÃ€derling, Martin Eklund, Henrik GrÃ¶nberg, Carin Cavalli-BjÃ¶rkman, Astrid BjÃ¶rklund, Britt Marie Hune, Shuang Hao, Mats WalldÃ©n, Ola Steinberg, Karl Andersson, Fredrik Wimmercranz, Edward Meurling, Axel Gleassgen, Nada Majeed, Rihab Awadelkarim, Ing Marie Fyhr, Dag SandstrÃ¶m, Linda Waage, Otabek Imamov, Rafael Lantz, Andreas Thorstensson, Carl Stiernstedt, Dushaid Wande, Gunnar Trygg, Harald SÃ¶derbÃ€ck, Jerzy Michajlowski, Lars Leykamm, Nils Erik Svedberg, Tommy Bergman, Raimundas Sabockis, and Sirvan Akrawi. Prostate cancer screening using a combination of risk-prediction, mri, and targeted prostate biopsies (sthlm3-mri): a prospective, population-based, randomised, open-label, non-inferiority trial. *The Lancet Oncology*, 22:1240–1249, 9 2021.

[49] Lars Egevad, Amar S. Ahmad, Ferran Algaba, Daniel M. Berney, Liliane Boccon-Gibod, Eva Compérat, Andrew J. Evans, David Griffiths, Rainer Grobholz, Glen Kristiansen, Cord Langner, Antonio Lopez-Beltran, Rodolfo Montironi, Sue Moss, Pedro Oliveira, Ben Vainer, Murali Varma, and Philippe Camparo. Standardization of gleason grading among 337 european pathologists. *Histopathology*, 62:247–256, 2013.

[50] Lars Egevad, John Cheville, Andrew J. Evans, Jonas Hörnblad, James G. Kench, Glen Kristiansen, Katia R.M. Leite, Cristina Magi-Galluzzi, Chin Chen Pan, Hemamali Samaratunga, John R. Srigley, Lawrence True, Ming Zhou, Mark Clements, Brett Delahunt, P. Argani, D. M. Berney, D. G. Bostwick, M. Brunelli, Y. Chen, L. Cheng, E. Comperat, J. Eble, S. W. Fine, S. Fleming, D. J. Grignon, D. Hansel, A. Hartmann, O. Hes, P. Humphrey, K. A. Iczkowski, L. Kunju, K. Leite, A. Lopez-Beltran, J. McKenney, G. Martignoni, H. Moch, G. Netto, J. Oxley, G. P. Paner, I. Sesterhenn, J. H. Shanks, H. Takahashi, F. Tavora, X. Teng, S. Tickoo, T. Tsuzuki, T. van der Kwast, and M. Varma. Pathology imagebase—a reference image database for standardization of pathology. *Histopathology*, 71:677–685, 2017.

[51] Tobias Nordström, Markus Aly, Mark S. Clements, Caroline E. Weibull, Jan Adolfsson, and Henrik Grönberg. Prostate-specific antigen (psa) testing is prevalent and increasing in stockholm county, sweden, despite no recommendations for psa screening: Results from a population-based study, 2003-2011. *European Urology*, 63:419–425, 2013.

[52] Nationella prostatacancerregistret (NPCR). Prostatacancer nationell kvalitetsrapport för 2020. https://npcr.se/rapporter/nationella-arsrapporter/. Online accessed: 19.09.2022, 2021.

[53] Karolinska Institutet. Anna Lantz (PI) Martin Eklund, Henrik Olsson. Active surveillance for low-risk prostate cancer - evaluation of the efficacy of minimally invasive active surveillance. https://ichgcp.net/clinical-trials-registry/nct03956108, 2020.

[54] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 10 1986.

[55] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[56] Saurabh Jha and Eric J. Topol. Adapting to artificial intelligence: Radiologists and pathologists as information specialists. *JAMA - Journal of the American Medical Association*, 316:2353–2354, 2016.

[57] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316:2402, 12 2016.

[58] 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*, 542:115–118, 2 2017.

[59] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, Oscar Geessink, Nikolaos Stathonikos, Marcory CRF van Dijk, Peter Bult, Francisco Beca, Andrew H Beck, Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Aoxiao Zhong, Qi Dou, Quanzheng Li, Hao Chen, Huang-Jing Lin, Pheng-Ann Heng, Christian Haß, Elia Bruni, Quincy Wong, Ugur Halici, Mustafa Ümit Öner, Rengul Cetin-Atalay, Matt Berseth, Vitali Khvatkov, Alexei Vylegzhanin, Oren Kraus, Muhammad Shaban, Nasir Rajpoot, Ruqayya Awan, Korsuk Sirinukunwattana, Talha Qaiser, Yee-Wah Tsang, David Tellez, Jonas Annuscheit, Peter Hufnagl, Mira Valkonen, Kimmo Kartasalo, Leena Latonen, Pekka Ruusuvuori, Kaisa Liimatainen,

Shadi Albarqouni, Bharti Mungal, Ami George, Stefanie Demirci, Nassir Navab, Seiryo Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Hady Ahmady Phoulady, Vassili Kovalev, Alexander Kalinovsky, Vitali Liauchuk, Gloria Bueno, M. Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Daniel Racoceanu, and Rui Venâncio. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318:2199, 12 2017.

[60] 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.

[61] 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 R M 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. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology*, 21, 2020.

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

[63] Wouter Bulten, Kimmo Kartasalo, Po Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, Christina Hulsbergen van de Kaa, Jeroen van der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus van der Kwast, Robert Allan, Peter A. Humphrey, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Tomi Häkkinen, Lars Egevad, Maggie Demkin, Sohier Dane, Fraser Tan, Masi Valkonen, Greg S. Corrado, Lily Peng, Craig H. Mermel, Pekka Ruusuvuori, Geert Litjens, Martin Eklund, Américo Brilhante, Aslı Çakır, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, Guilherme Pereira, Paromita Roy, Günter Saile, Paulo G.O. Salles, Ewout Schaafsma, Joëlle Tschui, Jorge Billoch-Lima, Emíio M. Pereira, Ming Zhou, Shujun He, Sejun Song, Qing Sun, Hiroshi Yoshihara, Taiki Yamaguchi, Kosaku Ono, Tao Shen, Jianyi Ji, Arnaud Roussel, Kairong Zhou, Tianrui Chai, Nina Weng, Dmitry Grechka, Maxim V. Shugaev, Raphael Kiminya, Vassili Kovalev, Dmitry Voynov, Valery Malyshev, Elizabeth Lapo, Manuel Campos, Noriaki Ota,

Shinsuke Yamaoka, Yusuke Fujimoto, Kentaro Yoshioka, Joni Juvonen, Mikko Tukiainen, Antti Karlsson, Rui Guo, Chia Lun Hsieh, Igor Zubarev, Habib S.T. Bukhar, Wenyuan Li, Jiayun Li, William Speier, Corey Arnold, Kyungdoc Kim, Byeonguk Bae, Yeong Won Kim, Hong Seok Lee, and Jeonghyuk Park. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 2022.

[64] Eirini Arvaniti, Kim S. Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter J. Wild, Jan H. Rüschoff, and Manfred Claassen. Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific Reports*, 8:12054, 12 2018.

[65] Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L. Smith, Arash Mohtashamian, James H. Wren, Greg S. Corrado, Robert MacDonald, Lily H. Peng, Mahul B. Amin, Andrew J. Evans, Ankur R. Sangoi, Craig H. Mermel, Jason D. Hipp, Martin C. Stumpe, Po-Hsuan, Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L. Smith, Arash Mohtashamian, James H. Wren, Greg S. Corrado, Robert MacDonald, Lily H. Peng, Mahul B. Amin, Andrew J. Evans, Ankur R. Sangoi, Craig H. Mermel, Jason D. Hipp, and Martin C. Stumpe. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *npj Digital Medicine*, 2, 2019.

[66] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. *Algorithmic Learning in a Random World*, pages 1–324, 2005.

[67] Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *Computer Journal*, 50:151–163, 2007.

[68] Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. *Tools in Artificial Intelligence*, 2008.

[69] Ewout W Steyerberg. *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media, 2008.

[70] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845, 1988.

[71] James L. Mohler, Emmanuel S. Antonarakis, Andrew J. Armstrong, Anthony V. D'Amico, Brian J. Davis, Tanya Dorff, James A. Eastham, Charles A. Enke, Thomas A. Farrington, Celestia S. Higano, Eric Mark Horwitz, Michael Hurwitz, Joseph E. Ippolito, Christopher J. Kane, Michael R. Kuettel, Joshua M. Lang, Jesse

McKenney, George Netto, David F. Penson, Elizabeth R. Plimack, Julio M. Pow-Sang, Thomas J. Pugh, Sylvia Richey, Mack Roach, Stan Rosenfeld, Edward Schaeffer, Ahmad Shabsigh, Eric J. Small, Daniel E. Spratt, Sandy Srinivas, Jonathan Tward, Dorothy A. Shead, and Deborah A. Freedman-Cass. Prostate cancer, version 2.2019, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 17:479–505, 5 2019.

[72] Vasilis Stavrinides, Francesco Giganti, Mark Emberton, and Caroline M. Moore. Mri in active surveillance: a critical review. *Prostate Cancer and Prostatic Diseases*, 22:5–15, 3 2019.

[73] Paolo Dell'Oglio, Armando Stabile, Matteo Soligo, Giorgio Brembilla, Antonio Esposito, Giorgio Gandaglia, Nicola Fossati, Carlo Andrea Bravi, Federico Dehò, Francesco De Cobelli, Francesco Montorsi, R. Jeffrey Karnes, and Alberto Briganti. There is no way to avoid systematic prostate biopsies in addition to multiparametric magnetic resonance imaging targeted biopsies. *European urology oncology*, 3:112–118, 2 2020.

[74] Pedro Recabal, Melissa Assel, Daniel D. Sjoberg, Daniel Lee, Vincent P. Laudone, Karim Touijer, James A. Eastham, Hebert A. Vargas, Jonathan Coleman, and Behfar Ehdaie. The efficacy of multiparametric magnetic resonance imaging and magnetic resonance imaging targeted biopsy in risk classification for patients with prostate cancer on active surveillance. *Journal of Urology*, 196:374–381, 8 2016.

[75] Gregory T Chesnut, Emily A Vertosick, Nicole Benfante, Daniel D Sjoberg, Jonathan Fainberg, Taehyoung Lee, James Eastham, Vincent Laudone, Peter Scardino, Karim Touijer, Andrew Vickers, and Behfar Ehdaie. Role of changes in magnetic resonance imaging or clinical stage in evaluation of disease progression for men with prostate cancer on active surveillance. *Eur. Urol.*, 2019.

[76] Kevin Michael Gallagher, Edward Christopher, Andrew James Cameron, Scott Little, Alasdair Innes, Gill Davis, Julian Keanie, Prasad Bollina, and Alan McNeill. Four-year outcomes from a multiparametric magnetic resonance imaging (mri)-based active surveillance programme: Psa dynamics and serial mri scans allow omission of protocol biopsies. *BJU International*, 123:429–438, 2019.

[77] Michael A. Liss, Lisa F. Newcomb, Yingye Zheng, Michael P. Garcia, Christopher P. Filson, Hilary Boyer, James D. Brooks, Peter R. Carroll, Matthew R. Cooperberg, William J. Ellis, Martin E. Gleave, Frances M. Martin, Todd Morgan, Peter S. Nelson, Andrew A. Wagner, Ian M. Thompson, and Daniel W. Lin. Magnetic resonance imaging for the detection of high grade cancer in the canary prostate active surveillance study. *Journal of Urology*, 204:701–706, 10 2020.

[78] Amer Amin, Matthijs J Scheltema, Ron Shnier, Alexandar Blazevski, Daniel Moses, Thomas Cusick, Amila Siriwardena, Brian Yuen, Pim J van Leeuwen, Anne Maree Haynes, Jayne Matthews, Phillip Brenner, Gordon O'Neill, Carlo Yuen, Warick Delprado, Phillip Stricker, and James Thompson. The mri in active surveillance "mrias" trial: Use of baseline mpmri and saturation biopsy to reduce the frequency of surveillance prostate biopsies. *J. Urol.*, page 101097JU0000000000000693, 2019.

[79] Dinul Hettiarachchi, Rob Geraghty, Patrick Rice, Ashwin Sachdeva, Arjun Nambiar, Mark Johnson, Rahul Gujadhur, Michael Mcneill, Phillip Haslam, Naeem Soomro, B. M. Zeeshan Hameed, Bhaskar Somani, Rajan Veeratterapillay, and Bhavan Prasad Rai. Can the use of serial multiparametric magnetic resonance imaging during active surveillance of prostate cancer avoid the need for prostate biopsies?-a systematic diagnostic test accuracy review. *European urology oncology*, 4:426–436, 6 2021.

[80] Pawel Rajwa, Benjamin Pradere, Fahad Quhal, Keiichiro Mori, Ekaterina Laukhtina, Nicolai A. Huebner, David D'Andrea, Aleksandra Krzywon, Sung Ryul Shim, Pascal A. Baltzer, Raphaële Renard-Penna, Michael S. Leapman, Shahrokh F. Shariat, and Guillaume Ploussard. Reliability of serial prostate magnetic resonance imaging to detect prostate cancer progression during active surveillance: A systematic review and meta-analysis, 11 2021.

[81] Andrew Vickers, Sigrid V. Carlsson, and Matthew Cooperberg. Routine use of magnetic resonance imaging for early detection of prostate cancer is not justified by the clinical trial evidence, 9 2020.

[82] Geert JLH van Leenders, Theodorus H van der Kwast, David J Grignon, Andrew J Evans, Glen Kristiansen, Charlotte F Kweldam, Geert Litjens, Jesse K McKenney, Jonathan Melamed, Nicholas Mottet, Gladell P Paner, Hemamali Samaratunga, Ivo G Schoots, Jeffry P Simko, Toyonori Tsuzuki, Murali Varma, Anne Y Warren, Thomas M Wheeler, Sean R Williamson, Isup Grading Workshop Panel Members, and Kenneth A Iczkowski. The 2019 international society of urological pathology (isup) consensus conference on grading of prostatic carcinoma, 2020.

[83] Cong Cong, Sidong Liu, Antonio Di Ieva, Maurice Pagnucco, Shlomo Berkovsky, and Yang Song. Colour adaptive generative networks for stain normalisation of histopathology images. *Medical Image Analysis*, 82, 11 2022.

[84] Jan Chandra Engel, Thorgerdur Palsdottir, Donna Ankerst, Sebastiaan Remmers, Ashkan Mortezavi, Venkatesh Chellappa, Lars Egevad, Henrik Grönberg, Martin Eklund, and Tobias Nordström. External validation of the prostate biopsy collaborative group risk calculator and the rotterdam prostate cancer risk calculator

in a swedish population-based screening cohort. *European Urology Open Science*, 41:1–7, 7 2022.

[85] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363:1287–1290, 2019.

[86] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:2242–2251, 2017.

[87] M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. 4 2018.

[88] PinkyA Bautista, Noriaki Hashimoto, and Yukako Yagi. Color standardization in whole slide imaging using a color calibration slide. *Journal of Pathology Informatics*, 5:4, 2014.

[89] Kimmo Kartasalo, Peter Ström, Pekka Ruusuvuori, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Martin Eklund, and Lars Egevad. Detection of perineural invasion in prostate needle biopsies with deep neural networks. *Virchows Archiv*, 481:73–82, 7 2022.

[90] Richard Colling, Helen Pitman, Karin Oien, Nasir Rajpoot, Philip Macklin, Velicia Bachtiar, Richard Booth, Alyson Bryant, Joshua Bull, Jonathan Bury, Fiona Carragher, Graeme Collins, Clare Craig, Maria Freitas da Silva, Daniel Gosling, Jaco Jacobs, Lena Kajland-Wilén, Johanna Karling, Darragh Lawler, Stephen Lee, Keith Miller, Guy Mozolowski, Richard Nicholson, Daniel O'Connor, Mikkel Rahbek, Alan Sumner, Dirk Vossen, Kieron White, Charlotte Wing, Corrina Wright, David Snead, Tony Sackville, and Clare Verrill. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. *Journal of Pathology*, 249:143–150, 10 2019.

[91] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380:1347–1358, 4 2019.

[92] Ben Van Calster and Andrew J. Vickers. Calibration of risk prediction models: Impact on decision-analytic performance. *Medical Decision Making*, 35:162–169, 2 2015.