




## Original Article

# Artificial intelligence for detection of prostate cancer in biopsies during active surveillance

Ida Arvidsson<sup>1</sup>, Edvard Svanemur<sup>3</sup>, Felicia Marginean<sup>2,4</sup>, Athanasios Simoulis<sup>4</sup> , Niels Christian Overgaard<sup>1</sup>, Kalle Åström<sup>1</sup>, Anders Heyden<sup>1</sup>, Agnieszka Krzyzanowska<sup>2</sup>  and Anders Bjartell<sup>2,5</sup> 

<sup>1</sup>Centre for Mathematical Sciences, <sup>2</sup>Division of Urological Cancers, Department of Translational Medicine, Lund University, Lund, <sup>3</sup>Department of Urology, Vrinnevi Hospital, Norrköping, <sup>4</sup>Department of Pathology and Molecular Diagnostics, and <sup>5</sup>Department of Urology, Skåne University Hospital, Malmö, Sweden

A.K. and A.B. co-last author.

## Objectives

To evaluate a cancer detecting artificial intelligence (AI) algorithm on serial biopsies in patients with prostate cancer on active surveillance (AS).

## Patients and methods

A total of 180 patients in the Prostate Cancer Research International Active Surveillance (PRIAS) cohort were prospectively monitored using pre-defined criteria. Diagnostic and re-biopsy slides from 2011 to 2020 ( $n = 4744$ ) were scanned and analysed by an in-house AI-based cancer detection algorithm. The algorithm was analysed for sensitivity, specificity, and for accuracy to predict need for active treatment. Prognostic properties of cancer size, prostate-specific antigen (PSA) level and PSA density at diagnosis were evaluated.

## Results

The sensitivity and specificity of the AI algorithm was 0.96 and 0.73, respectively, for correct detection of cancer areas. Original pathology report diagnosis was used as the reference method. The area of cancer estimated by the pathologists correlated highly with the AI detected cancer size ( $r = 0.83$ ). By using the AI algorithm, 63% of the slides would not need to be read by a pathologist as they were classed as benign, at the risk of missing 0.55% slides containing cancer. Biopsy cancer content and PSA density at diagnosis were found to be prognostic of whether the patient stayed on AS or was discontinued for active treatment.

## Conclusion

The AI-based biopsy cancer detection algorithm could be used to reduce the pathologists' workload in an AS cohort. The detected cancer amount correlated well with the cancer length measured by the pathologist and the algorithm performed well in finding even small areas of cancer. To our knowledge, this is the first report on an AI-based algorithm in digital pathology used to detect cancer in a cohort of patients on AS.

## Keywords

artificial intelligence, prostate cancer, active surveillance, PRIAS, deep learning

## Introduction

Active surveillance (AS) is recommended for monitoring men with low-risk prostate cancer (PCa). The goal of the method is to avoid treatment of tumours that most likely will not cause any problems if left untreated, while still keeping track of any eventual progression of the cancer requiring active treatment. The purpose of the Prostate Cancer Research International Active Surveillance (PRIAS) study was to provide evidence-based recommendations for the most

optimal selection and follow-up of men on AS [1]. The PRIAS guidelines are continuously evaluated, and the protocol updated to further reduce unnecessary treatments while not risk the wellbeing of any patient [2]. The inclusion criteria of the PRIAS study are designed to enrol men with low PSA levels, clinical stage T1c and T2 and only small areas of cancer with Gleason Score 3 + 3 or 3 + 4 (with  $\leq 10\%$  tumour involvement/biopsy in maximum two cores), corresponding to an International Society of Urological Pathology (ISUP) Gleason Grade Group (GG) 1 or GG 2.

Artificial intelligence (AI) has been proven successful for automated grading and diagnosis of prostate biopsies in numerous studies [3–11]. There are several arguments for AI to be a useful tool in a clinical setting, i.e., to facilitate and increase the speed of pathologists' work and to reduce the inter-observer variability between pathologists and the detection of small cancer foci or atypical lesions [12]. Most of the to-date reported AI-based algorithms have focussed on analysis of large areas of cancer, and some have put high cut-off thresholds to ignore smaller cancer areas [3]. However, if AI is to be used in clinical practice, it should also detect smaller cancer areas, including small atypical foci [13].

In this study, we evaluated an AI algorithm for cancer detection on prostate biopsies from an AS cohort, something that previously has not been done. Using an AS cohort for evaluation implies that distinguishing between normal tissue and low-grade cancer is particularly important. It is often suggested that AI could be used to exclude benign cases from the pathologists' work in order to speed up the histopathological examination [14], thus evaluation of AI on an AS cohort is highly relevant. Another purpose of the study was to investigate if early signs of growing cancer can be detected by AI and could be used as additional information to predict if, and when, a patient will need active treatment.

The aim of the present study was to investigate the usefulness of cancer-detecting AI in histological preparations from the Swedish subgroup of patients in the PRIAS cohort. We present here the results of the AI algorithm in detecting cancer on >4000 biopsies from 180 patients followed-up in the AS protocol over a period of >10 years. We furthermore correlate these results with various clinical parameters.

## Patients and Methods

### Cohort – Study Population

This study used samples from patients included in the PRIAS study who were enrolled at the Skåne University Hospital, Malmö, Lund and Kristianstad hospital at Region Skåne, Sweden between 2007 and 2020. The study was approved by the Regional Ethical Committee in Lund (approval number 708/2008 and 11/2018) and amendment approved by the National Ethical Board in Sweden (approval number 2021-00233).

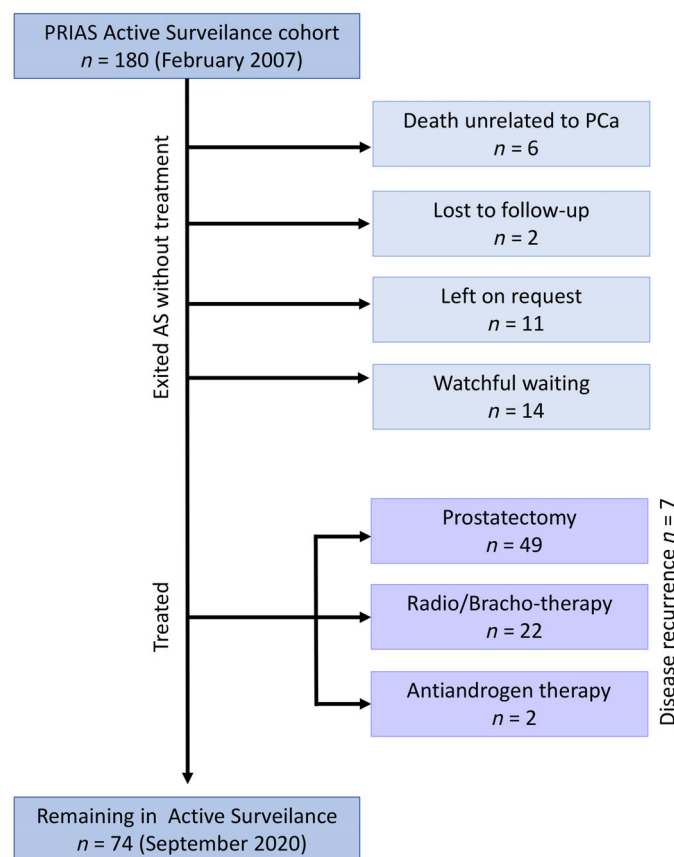
The criteria for inclusion in the study were: men fit for curative therapy, PSA level at diagnosis of  $\leq 10$  ng/mL, PSA density of 0.20 ng/mL/mL, two or fewer biopsy cores bearing PCa, Gleason Score 3 + 3 ( $< 5$  mm maximum length) or 3 + 4 with  $\leq 10\%$  tumour involvement/biopsy in a maximum of two cores, and clinical stage T1c or T2 (<https://www.prias-project.org/>). Since the update in 2016, more than two positive biopsy cores were allowed if a multiparametric MRI with guided biopsies was performed [2].

Patients were monitored according to the PRIAS protocol follow-up schedule. They underwent a PSA test every third month during the first 2 years and every sixth month thereafter. A DRE was offered every sixth month during the first 2 years and every year thereafter. Re-biopsies were scheduled after 1, 4, 7 and 10 years after diagnosis and every fifth year thereafter. The protocol was updated in 2016, where DRE was recommended at the same intervals as re-biopsies.

Biopsy slides and clinical data were retrieved for 180 patients. Of these, 33 were censored due to: death, changing to a different treatment protocol (by request or watchful waiting) or being lost to follow-up (see the Consolidated Standards of Reporting Trials [CONSORT] diagram in Fig. 1). By the time of study conclusion, 73 patients were treated and 74 remained on AS.

Clinical data that were extracted from electronic medical records included pathology information for each biopsy core (Gleason Score, millimetres of cancer, whether additional immunohistochemistry was requested), whether the biopsies were targeted to a specific area or systematic, as well as PSA level, PSA density, and prostate volume. Follow-up information on treatment was also retrieved.

**Fig. 1** The CONSORT diagram of patients included in the study.



## Prostate Biopsies and Scanning

Prostate biopsies were evaluated over the years by 25 different pathologists at three different hospitals within the Skåne region: Malmö, Lund and Kristianstad.

All available biopsy slides for the patients included in the study were retrieved from the Region Skåne Biobank. No slides were available from before 2008. The slides were scanned on an Aperio® CS2 slide scanner (Leica, Newcastle, UK) at  $\times 40$ . The slides from 2008 to 2010 were deemed suboptimal quality due to fading and therefore only material from 2011 onwards was included in the study. One biopsy section per slide was scanned.

A total of 4744 scans were used in the final analysis. The majority of them came from Malmö but slides from two other sites (Lund and Kristianstad) were also included. Table S1 details the number of slides from each site, as well as year of biopsy. We also distinguished whether the biopsy was systematic or MRI targeted, based on the clinical records. MRI-assisted biopsies that were sampled in a systematic way were classified as systematic. The average number of slides per sampling occasion was 10 with a range of three to 16.

## The AI-Based Algorithm

The dataset used for development of the deep learning algorithm consisted of 870 biopsy images scanned in the years 2016–2018. The slides originated from four different sites, (i) Skåne University Hospital in Malmö, Sweden ( $N = 752$ ), (ii) Helsingborg Hospital, Sweden ( $N = 53$ ), (iii) Linköping University Hospital, Sweden ( $N = 16$ ), and (iv) Erasmus University Medical Center in Rotterdam, the Netherlands ( $N = 49$ ). The slides from (i) and (ii) were scanned on an Aperio CS2 scanner (Leica), the slides from (iii) were scanned on an Aperio AT Turbo scanner (Leica), and the slides from (iv) were scanned on a Hamamatsu HT 2.0 scanner (Hamamatsu Photonics K.K, Tokyo, Japan). All images were annotated on gland-level, by two experienced pathologists from Malmö (F.M. and A.S., co-authors), into areas of ISUP Gleason GG 3, GG 4, and GG 5, as well as benign tissue. The annotations were drawn using Sectra ISD7 software, either from scratch ( $N = 652$ ) or by utilising predictions from a previous algorithm and correcting those ( $N = 218$ ), see [11] for details.

The development dataset was used for training of the algorithm. To confirm the performance and choose the model, 25% of the slides were set aside and used for validation. Patches to train on were cropped from within annotated areas of the whole-slide images. Resolutions between 0.912 and 1.0  $\mu\text{m}$  per pixel (variation caused by different scanners used at the different hospitals) and patches of size  $299 \times 299$  pixels were used. All areas labelled with a ISUP Gleason GG  $\geq 3$  were assigned the label 'malignant',

while patches from the annotated benign areas kept the label 'benign'. After training, the algorithm was fixed and no retraining was done, meaning that the same algorithm was used for evaluation throughout the study. During inference, each patch was classified as benign or malignant based on the highest predicted probability, resulting in a segmentation of predicted cancer areas. The amount of cancer (percentage of tissue area) was determined from the segmentation.

## Evaluation of Performance and Statistical Analysis

The outcome of 'treated' or 'stayed on AS' was used to separate the patients into groups. Treatments were based on protocol advice and included radical prostatectomy (RP), radiation, brachytherapy, or antiandrogen therapy (Fig. 1). The treated group was further subdivided based on biochemical recurrence (BCR) to 'treated with no BCR' (BCR−) and 'treated with BCR' (BCR+). BCR was defined as a PSA level of  $\geq 0.2$  ng/mL after RP or PSA level rise of  $\geq 2$  ng/mL above the nadir ('Phoenix criteria') after curative radiotherapy/brachytherapy. Paired comparisons within groups at different time points were performed with the aid of the Wilcoxon signed-rank test.

Correlation between the algorithm's and pathologists' estimations was performed using the Pearson correlation. The prognostic value of different variables was analysed using univariable Cox regression, using 'active treatment' as the primary endpoint. The variables were dichotomised using the decision tree regression model before performing the Cox regression and Kaplan–Meier survival analysis with log-rank test statistic. Average 'cancer size AI' and 'cancer size pathology' values for each patient and occasion were calculated based on all biopsy cores from that sampling occasion as the average of the AI predicted percentage of cancer and the total millimetres of cancer found divided by total length of the biopsy cores respectively. This was done to normalise the extent of cancer in each sampling occasion. A  $P < 0.05$  was assumed to be statistically significant. The statistical software SPSS® version 27 (IBM Corp., Armonk, NY, USA) was used for the analysis.

To evaluate the algorithm's performance on the slide-level, receiver operating characteristic (ROC) curves based on predicted percentage of cancer and area under the ROC curve (AUC) values were used, determined using the Scikit-learn library (version 0.22) in Python.

## Results

### Cohort Description

The characteristics of the cohort are shown in Table 1. Of the 180 patients, 33 were censored due to death, loss to follow-up, exiting study on request, or being changed to a watchful waiting protocol (see the CONSORT diagram in Fig. 1). Of

**Table 1** The patients' characteristics.

	All in study (N = 180)	AS (n = 74)	Treated (n = 73)	Treated BCR- (n = 66)	Treated BCR+ (n = 7)
Age at diagnosis, years, median (IQR)	66 (61–69)	65 (61–68)	66 (61–69)	66 (62–69)	62 (59–65)
Year of diagnosis, n (%)					
2007–2009	26 (14)	7 (9.5)	13 (17)	11 (17)	2 (29)
2010	14 (7.8)	4 (5.4)	6 (8.2)	5 (7.6)	1 (14)
2011	23 (13)	6 (8.1)	11 (15)	10 (15)	1 (14)
2012	15 (8.3)	3 (4.1)	11 (15)	11 (17)	0
2013	26 (14)	10 (14)	9 (12)	7 (11)	2 (29)
2014	40 (22)	19 (26)	13 (18)	13 (20)	0
2015	19 (11)	10 (14)	8 (11)	7 (11)	1 (14)
2016	10 (5.6)	9 (12)	1 (1.4)	1 (1.5)	0
2017	7 (3.9)	6 (8.1)	1 (1.4)	1 (1.5)	0
PSA at diagnosis, ng/mL, median (IQR)	4.9 (3.9–6.3)	5.1 (4.0–6.8)	4.7 (3.8–5.8)	4.6 (3.8–5.9)	5.2 (4.1–5.3)
PSA density at diagnosis, ng/mL/mL, median (IQR)	0.12 (0.10–0.16)	0.11 (0.09–0.15)	0.14 (0.11–0.16)	0.14 (0.10–0.16)	0.16 (0.13–0.20)
Prostate volume at diagnosis, mL, median (IQR)	42 (33–57)	44 (35–56)	38 (27–49)	38 (28–53)	33 (21–42)
Gleason GG (ISUP) at diagnosis, n (%)					
1	161 (89)	66 (89)	64 (87)	57 (86)	7 (100)
2	18 (10)	8 (11)	8 (11)	8 (12)	0
3	1 (0.55)	0	1 (1.4)	1 (1.5)	0
Number of cores with PCa at diagnosis, n (%)					
1	120 (67)	53 (72)	43 (59)	38 (58)	5 (71)
2	55 (31)	20 (27)	26 (36)	24 (36)	2 (29)
3	4 (2.2)	1 (1.4)	3 (4.1)	3 (4.5)	0
4	1 (0.55)	0	1 (1.4)	1 (1.5)	0
% of cores with cancer, median (IQR)	10 (10–17)	10 (8–17)	10 (10–20)	10 (10–20)	10 (10–20)
Clinical T-stage at diagnosis, n (%)					
T1c	151 (84)	66 (89)	59 (81)	55 (83)	4 (57)
T2a	25 (14)	6 (8.1)	12 (16)	9 (14)	3 (43)
T2b	2 (1.1)	1 (1.4)	1 (1.4)	1 (1.5)	0
T2c	2 (1.1)	1 (1.4)	1 (1.4)	1 (1.5)	0
Total follow-up time, months: median (IQR)	78 (59–103)	72 (58–97)	97 (76–111)	97 (76–111)	96 (90–117)
Number of re-biopsies, n (%)					
0	7 (3.9)	2 (2.7)	0	0	0
1	59 (33)	18 (24)	30 (41)	26 (39)	4 (57)
2	68 (38)	36 (49)	26 (36)	24 (36)	2 (29)
3	35 (20)	14 (19)	14 (19)	14 (21)	0
4	8 (4.5)	3 (4.1)	2 (2.7)	2 (3.0)	0
5	2 (1.1)	0	1 (1.4)	0	1 (14)
6	1 (0.6)	1 (1.4)	0	0	0
Last known Gleason GG (ISUP), n (%)					
0	54 (30)	40 (54)	0	0	0
1	60 (33)	27 (36)	16 (22)	14 (21)	2 (29)
2	44 (24)	7 (9.5)	35 (48)	33 (50)	2 (29)
3	15 (8.3)	0	15 (21)	15 (23)	0
4	5 (2.8)	0	5 (6.8)	4 (6.1)	1 (14)
5	2 (1.1)	0	2 (2.7)	0	2 (29)
Last biopsy % of PCa, median (IQR)	17 (0–37)	0 (0–15)	38 (25–54)	36 (24–50)	42 (25–83)
Last PSA level, ng/mL, median (IQR)	5.8 (4.3–9.3)	5.3 (3.5–8.9)	6.8 (5.2–9.6)	6.6 (5.3–9.4)	9.7 (4.7–12.0)
Last PSA density, ng/mL/mL, median (IQR)	0.14 (0.09–0.19)	0.10 (0.08–0.16)	0.17 (0.13–0.24)	0.17 (0.13–0.22)	0.22 (0.15–0.46)
Metastasis, n (%)	5 (2.8)	0	4 (5.5)	1 (1.5)	3 (43)
Death from PCa, n (%)	1 (0.6)	0	1 (1.4)	0	1 (14)

Data represents all patients in the cohort, those who remained on active surveillance at end of study and those who were treated. The treated group is further subdivided into those who did or did not have BCR (BCR+ /BCR-).

the remaining 147 patients, 74 remained on AS at the study's end point (30 September 2020) and 73 had received active treatment. Seven of the treated patients had BCR by the study's end point.

Most of the characteristics at diagnosis, including age, PSA, Gleason score, number of biopsy cores with cancer, and clinical T-stage were similar in the two groups. Prostate

volume at diagnosis was higher in the remain on AS group (47.5 vs 40.3 mL), and the PSA density at diagnosis was higher in the treated group (0.14 vs 0.12 ng/mL/mL; Table 1). The biopsy characteristics at the end of the study were different in the two groups: last-known Gleason score was higher in the treated group, with 22 patients being in ISUP GG 3, 4 or 5. Only seven patients who remained on AS



group had ISUP Gleason GG 2 patterns, compared to 35 in the treated group. The average percentage of cores with cancer on last biopsy was 41% for the treated group, compared to 9.0% in the remain on AS group.

No increase in PSA or PSA density levels at last measurement compared to diagnosis was observed for the AS patients (Table 1). The PSA and PSA density levels at last assessment in the treated group were increased in both the BCR– group and in the BCR+ group compared to diagnosis.

Metastases and death from PCa were observed only in the treated group; however, the numbers were low (2.7% and 0.68% of the 147 patients, Table 1).

Performance of the Algorithm

The algorithm was tested on a total of 4744 slides. Of those, 671 (14%) were originally diagnosed as having an area of cancer (data obtained from the pathology reports, Table 2). On a per slide basis and defining an AI-generated cancer prediction of ≥50% as positive, 1094 (27%) of 4073 benign slides were called as positive, 444 (95%) of 469 slides with Gleason Score 3 + 3 (ISUP GG 1), 118 (100%) of 118 slides with Gleason Score 3 + 4 (ISUP GG 2), 46 (98%) of 47 slides with Gleason Score 4 + 3 (ISUP GG 3), 37 (100%) of 37 slides with ISUP Gleason GG 4 and 5 (Table 2). Given a sample of 1000 slides from AS patients, using our AI algorithm, 633 (63%) slides would not need to be read by a pathologist as they were classed as benign, at the risk of missing five (0.55%) slides harbouring cancer, of which less than one (0.021%) slide would be Gleason Score 3 + 4 (ISUP GG 2) or above. A true positive was defined if the algorithm found cancer on a slide that was diagnosed as having cancer by the pathologist, true negative was when algorithm found no cancer on a slide considered ‘benign’ by the pathologist. The algorithm was highly sensitive (0.961), missing cancer in only 26 (0.55%) of the slides. However, the algorithm over-detected cancer: the specificity was 0.731 (Table 2). The accuracy measured as the AUC for all the slides included in

the study was 0.922 (Table S1). The AUC for the algorithm was consistent on slides from different years and from different locations (Table S1), suggesting it deals well with colour variations. The algorithm also had a similar agreement with different pathologists’ diagnoses (Fig. S1A,B). There was high correlation between the size of cancer estimated by the pathologist and by the algorithm ( $r = 0.83$ , Pearson correlation, Fig. 2, left). The correlation was even higher if only slides from MRI guided biopsies were included ( $r=0.89$ , Pearson correlation, Fig. 2, right).

The algorithm found cancer in 1739 slides (37% of the total). Of these, 1094 were false positives (Table 2). On further analysis, the average size of false positives was 3.3% of a biopsy, whereas true positives had an average algorithm-estimated size of 13%, suggesting that the false positives were pointing to small suspicious-looking areas (Table S2). The 26 cases of cancer that were missed by the algorithm were also small areas (average size: 6.2% of biopsy length), which were significantly smaller from the average cancer size found on positive slides as estimated by a pathologist (20%). Among missed cancers, 17 (65%) also had an immunohistochemistry slide requested by the diagnosing pathologist, suggesting that they were small and particularly difficult areas to detect (Table S2). On further analysis, the missed cancer on the 26 slides would not have changed the decision for treatment or not of the patient (data not shown), thus the AI model was able to identify all clinically significant cancers.

The size of cancer areas reported on the first available biopsies and the last available biopsies were compared. In the group that remained on AS there were no statistical differences between cancer size on the first and last biopsies, estimated by pathologist or AI (Table 3). In the BCR– treated group, there was an increase in cancer size at the last biopsy, estimated by both pathologist and AI ( $P < 0.001$ , Wilcoxon test). This increase was even higher in the BCR+ group. As would be expected per PRIAS protocol, at the last biopsy, both BCR– and BCR+ groups had more cancer than the AS groups, estimated both by pathologist and AI. Interestingly, the patients who eventually got treatment, already presented with higher average cancer sizes at first biopsy occasions compared to AS.

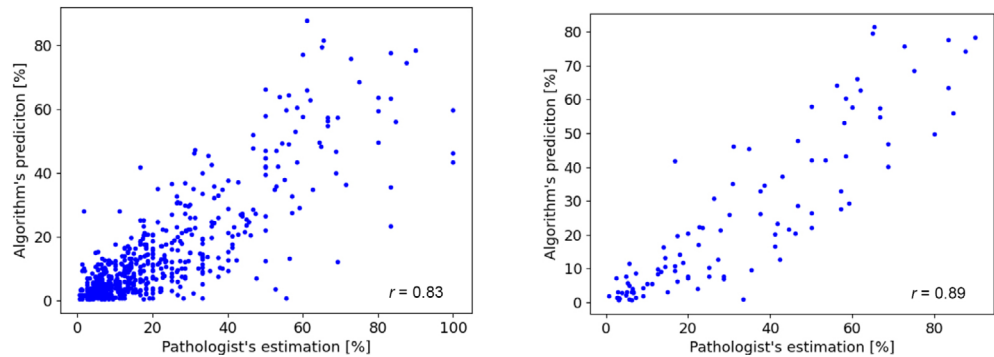
**Table 2** Performance of the AI tool, showing the number of slides scanned and numbers correctly and incorrectly identified to contain cancer, including different ISUP Gleason grades as well as benign slides.

Variable	Value
Slides scanned, <i>n</i>	4744
Cancer diagnosed by pathologist, <i>n</i> (%)	671 (14)
AI correctly detects cancer (true positives), <i>n</i> (%)	645 (14)
AI incorrectly detects cancer (false positives), <i>n</i> (%)	1094 (23)
AI incorrectly misses cancer (false negatives), <i>n</i> (%)	26 (0.55)
AI detect cancer in benign slides ( <i>n</i> = 4073), <i>n</i> (%)	1094 (27)
AI detect cancer in ISUP GG 1 slides ( <i>n</i> = 469), <i>n</i> (%)	444 (95)
AI detect cancer in ISUP GG 2 slides ( <i>n</i> = 118), <i>n</i> (%)	118 (100)
AI detect cancer in ISUP GG 3 slides ( <i>n</i> = 47), <i>n</i> (%)	46 (98)
AI detect cancer in ISUP GG 4 and GG 5 slides ( <i>n</i> = 37), <i>n</i> (%)	37 (100)
Sensitivity	0.961
Specificity	0.731

Factors Predictive of Need for Active Treatment

The size of cancer on the first biopsy was highly predictive of whether the patient would stay on AS (Fig. 3, univariable Cox regression and Kaplan–Meier curves with log-rank test statistic). This was observed for both the pathologists’ cancer length estimate (hazard ratio [HR] 2.7, 95% CI 1.6–4.6) and the AI percentage of cancer calculation (HR 2.5, 95% CI 1.4–4.5). The univariable Cox regression analysis also identified the PSA density at diagnosis as a significant prognostic factor for active treatment (HR 2.2, 95% CI 1.4–3.5).

**Fig. 2** Pearson correlation between the cancer percentage predicted by algorithm and that estimated by a pathologist for all slides (left) and MRI guided slides (right). The percentage from the clinical records was calculated using the formula cancer length divided by whole biopsy length. False positives and false negatives are not included.



**Table 3** The average cancer length (mm) estimated by pathologist and average cancer size (%) detected by AI for the first available biopsies (at diagnosis or soon thereafter) and last available biopsies (last before treatment or last before the end of the study).

	First biopsy cancer length, mm, (pathologist)			Last biopsy Cancer length, mm, (pathologist)		
	AS	T BCR–	T BCR+	AS	T BCR–	T BCR+
Mean	0.14	0.28	0.23	0.22	1.76	4.26
Median	0.10	0.20	0.13	0.00	1.11	5.42
SD	0.16	0.29	0.24	0.54	2.12	2.36

	First biopsy cancer size, % of total area, (AI)			Last biopsy cancer size, % of total area, (AI)		
	AS	T BCR–	T BCR+	AS	T BCR–	T BCR+
Mean	1.20	2.13	1.08	1.93	8.59	20.23
Median	0.83	1.27	0.94	0.97	5.43	21.70
SD	1.21	2.69	0.78	2.97	12.17	12.87

Patients are grouped into those who remained on AS ( $n = 73$ ) at the end of the study, those who underwent active treatment and had no BCR (T BCR–,  $n = 66$ ) and those who had active treatment and had BCR during the study duration (T BCR+,  $n = seven$ ).

Discussion

In this study, we evaluated the characteristics of a subset of the PRIAS cohort, which consisted of men on AS monitored using a specific protocol. Furthermore, we evaluated the performance of our in-house AI algorithm to identify cancer areas in diagnostic and repeat biopsies from the same cohort. Our AI algorithm is a modification of the algorithm we recently validated in PCa biopsies [11]. Different AI PCa-detecting algorithms have been described in recent years [3–6,8–10]; however, none have specifically been used on an AS cohort. Prostate biopsies from men on AS differ from other cohorts, having an exceptionally high ratio of slides without cancer or with small areas of low-grade cancer. The latter part constitutes cases that are hard for a cancer detection system, which also makes them extra important to include when evaluating the performance of a cancer-detecting algorithm. The AS cohort studied here

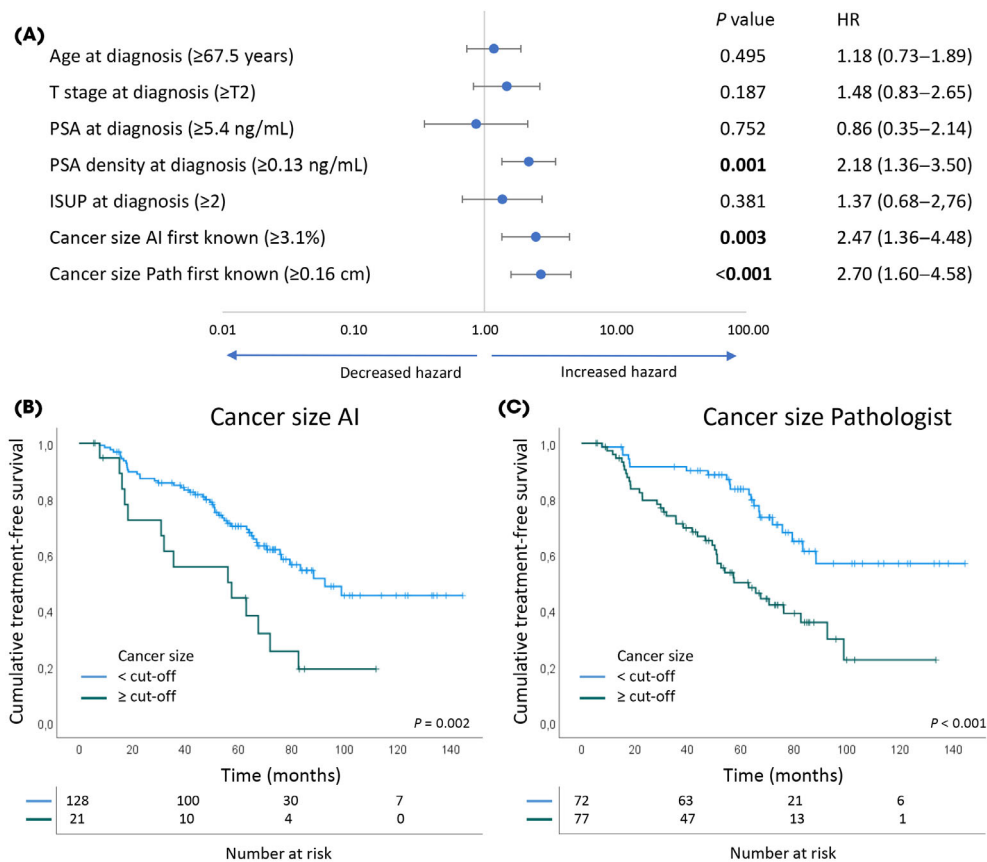
included 180 patients, of whom 73 were converted to active treatment by the end of the study. The decision to initiate active treatment was based on pathologists’ diagnoses and clinical parameters such as PSA levels, DRE, and MRI results. Patients who received treatment due to anxiety were censored and were not included in the ‘treated’ group. Seven of the treated patients had a disease progression in terms of BCR by the end of the study. While most of the characteristics at diagnosis were the same for the ‘remain on AS’ group and the ‘treated’ group, we observed an increase in prostate volume and PSA density in the group that eventually was treated, as would be expected as these factors have also been suggested as independent predictors of PCa outcome [15,16]. Similarly to what has been published previously [17], we found PSA density at diagnosis to be a strong prognostic marker in the likelihood of receiving treatment.

Our AI algorithm performed well on individual scanned slides (sensitivity: 96%, specificity: 73%). Its performance was consistent over different years, locations, and pathologists’ diagnoses. Overall, the cancer areas missed by the AI algorithm and the areas incorrectly detected as cancer were minimal. Most importantly, the missed cancer areas were from clinically insignificant cancers and would not have affected any diagnoses, meaning that the sensitivity for detecting patients who should get treatment was 100%.

The AI algorithm found 37% of the slides as containing cancer. According to pathology reports, only 14% slides had cancer in the whole cohort. While this shows an over-estimation by the algorithm, it could be used to narrow-down slides of interest in order for the pathologist to examine only a subset of a huge volume of slides. This would reduce the work-load by approximately two-thirds.

The estimate of the total amount of cancer on the first biopsy was predictive of whether the patient would need active treatment. This was observed both for the pathologists estimate as well as the AI estimate. Thus, in a reasonably homogenous population of low-risk patients, the AI was able

**Fig. 3 (A)** A forest plot showing the HR and 95% CIs associated with variables considered in the univariable analyses with time to the primary endpoint (active treatment) as the dependent variable. Circles represent the HR and the horizontal bars extend from the lower limit to the upper limit of the 95% CI of the estimate of the HR. Data for each variable was dichotomised based on classification/regression tree (dichotomisation values in brackets). Bold values statistically significant at  $P < 0.05$ . **(B, C)** Kaplan–Meier curves showing treatment-free survival for the AI **(B)** and pathologists’ estimate **(C)**. Log-rank test statistic was used to determine the  $P$  value. Cut-off value was determined with classification/regression tree. AI, artificial intelligence algorithm; Path, pathologists’ estimation.



to indicate at the time of diagnosis, which patients would be more likely to eventually fail AS. The Gleason score at first biopsy was not predictive, as all patients enrolled had ISUP Gleason GG 1 or 2.

In this study, we aimed at developing a highly sensitive cancer-detection algorithm, as a means to scan through high volumes of biopsy slides and highlight the areas of cancer for further analysis by a pathologist. The next step would be to develop such an algorithm that also assesses Gleason grading with high performance. Such a tool would be of high interest as the criteria for the PRIAS exclusion include number of biopsies with cancer, and the amount of Gleason pattern  $\geq 4$ . Variability in cancer assessment between pathologists is well documented [14,18,19] and results in different treatment choices [20]. An AI system assisting with both cancer detection and grading could reduce these differences and may guide urologists to better understand which patients are likely to fail AS. Our suggested tool could be one key part of that

system, indicating AS failure based on the extent of tumour in the biopsies. However, it cannot yet be a standalone tool. The present study has some limitations. First, the reference method for cancer detection on each biopsy constitutes of a single pathologist’s assessment and variations in grading small cancer areas are known [12,13], meaning that the reference method used in this study certainly is a limitation. Another limiting factor is that very few slides with high-grade cancer were included, due to the PRIAS study’s inclusion criteria. As ISUP Gleason GG 5 can appear very different from the lower grades, this could be a problem. However, the algorithm was trained using a dataset including high-grade cancers [11] and had no problem detecting high-grade cancer in the few cases available in the PRIAS cohort. Also, as mentioned above, Gleason grading was not assessed but would be a valuable feature of the AI tool. It should also be kept in mind, that the PRIAS guidelines evolve to match the cumulative knowledge and evolving technology [2]. This study spans over 15 years

and the criteria for the decision of whether to offer a patient active treatment have changed over the years, e.g., due to increased use of MRI.

In recent years, there have been numerous studies looking into AI-assisted detection of cancer on prostate histological preparations [3–6,8–10]. Most of these studies look at the AI algorithm output in comparison with the pathologist's result. Few have also applied the algorithms to routine pathology clinical practice [7] or as screening tools [21]. AI algorithms can help in diagnosis, by finding and staging cancer. However, for added clinical value, AI should also be able to help to predict the patient's outcome and guide treatment decisions [12]. For this purpose, cohorts with long follow-up and many events are needed. Several studies have found that AI can predict the survival after RP [22] and BCR [8,23] as well or even better than the pathologist-assigned Gleason grade. In our AS cohort, the number of BCR/death events was too low to use for any prognostic analysis; however, we could observe that cancer detected on the early biopsies was predictive of later treatment.

In conclusion, in this study we developed and evaluated a highly sensitive AI algorithm that can be used in examination of biopsies from patients with PCa on AS. The algorithm shows high sensitivity and specificity, although biopsies from the AS cohort often have very limited areas of cancer. We found that the area of cancer on the initial biopsies may be predictive of the need for active treatment and thus disease progression. It can therefore be proposed as a screening tool to speed up the diagnosis of early-stage PCa, such as patients with PCa undergoing repeated biopsies for AS. Moreover, PSA density at diagnosis was found to be highly prognostic of future disease progression and should be considered when deciding upon the best treatment protocols for patients with low-stage PCa. For the future, Gleason grading should be included in the AI system. Furthermore, a prospective study investigating the usage of the AI tool in a clinical setting would be needed.

## Acknowledgements

The authors would like to thank Macarena Palominos for help with scanning and the clinical information collection, Kristina Ekström-Holka for help with the scanning of the slides, Anna Stiehm, Margareta Persson, Annica Löfgren, Jennifer Amidi and Åsa Andersson for help with patient clinical information collection. We are grateful for scanned biopsy images from Rotterdam Erasmus Hospital provided by Geert van Leenders and Linköping University Hospital provided by Claes Lundström.

## Disclosure of Interests

There are no conflicts of interest for any of the authors.

## Funding

This study was funded by grants from The Swedish Cancer Society (Cancerfonden #2018/522 and #2021/1629), Swedish Research Council (Vetenskapsrådet, #2020-02017), The Research Funds at Skåne University Hospital, The Governmental Funding (ALF) through The Faculty of Medicine, Lund University (contract number F2018:810), The Swedish Prostate Cancer Foundation (Prostatancerförbundet), Vinnova-Swelife and Vinnova-Medtech4Life programmes, strategic research project eSENCE.

## References

- 1 Bul M, Zhu X, Valdagni R et al. Active surveillance for low-risk prostate cancer worldwide: the PRIAS study. *Eur Urol* 2013; 63: 597–603
- 2 Bokhorst LP, Valdagni R, Rannikko A et al. A decade of active surveillance in the PRIAS study: an update and evaluation of the criteria used to recommend a switch to active treatment. *Eur Urol* 2016; 70: 954–60
- 3 Bulten W, Pinckaers H, van Boven H et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020; 21: 233–41
- 4 Bulten W, Balkenhol M, Belinga JJA et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod Pathol* 2021; 34: 660–71
- 5 da Silva LM, Pereira EM, Salles PGO et al. Independent real-world application of a clinical-grade automated prostate cancer detection system. *J Pathol* 2021; 254: 147–58
- 6 Huang W, Randhawa R, Jain P et al. Development and validation of an artificial intelligence-powered platform for prostate cancer grading and quantification. *JAMA Netw Open* 2021; 4: e2132554
- 7 Pantanowitz L, Quiroga-Garza GM, Bien L et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet Digit Health* 2020; 2: e407–16
- 8 Pinckaers H, van Ipenburg J, Melamed J et al. Predicting biochemical recurrence of prostate cancer with artificial intelligence. *Commun Med* 2022; 2: 64
- 9 Ström P, Kartasalo K, Olsson H et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol* 2020; 21: 222–32
- 10 Singhal N, Soni S, Bonthu S et al. A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. *Sci Rep* 2022; 12: 3383
- 11 Marginean F, Arvidsson I, Simoulis A et al. An artificial intelligence-based support tool for automation and standardisation of Gleason grading in prostate biopsies. *Eur Urol Focus* 2021; 7: 995–1001
- 12 Egevad L, Delahunt B, Samaratunga H et al. The emerging role of artificial intelligence in the reporting of prostate pathology. *Pathology* 2021; 53: 565–7
- 13 van der Kwast TH, Evans A, Lockwood G et al. Variability in diagnostic opinion among pathologists for single small atypical foci in prostate biopsies. *Am J Surg Pathol* 2010; 34: 169–77
- 14 Egevad L, Ström P, Kartasalo K et al. The utility of artificial intelligence in the assessment of prostate pathology. *Histopathology* 2020; 76: 790–2
- 15 Tang P, Jin XL, Uhlman M et al. Prostate volume as an independent predictor of prostate cancer in men with PSA of 10–50 ng ml<sup>-1</sup>. *Asian J Androl* 2013; 15: 409–12
- 16 Sfoungaristos S, Perimenis P. PSA density is superior than PSA and Gleason score for adverse pathologic features prediction in patients with clinically localized prostate cancer. *J Can Urol Assoc* 2012; 6: 46–50



- 17 Yusim I, Krenawi M, Mazor E, Novack V, Mabjeesh NJ. The use of prostate specific antigen density to predict clinically significant prostate cancer. *Sci Rep* 2020; 10: 20015
- 18 Flach RN, Willemse PPM, Suelmann BBM et al. Significant inter- and intralaboratory variation in gleason grading of prostate cancer: a nationwide study of 35,258 patients in The Netherlands. *Cancers (Basel)* 2021; 13: 5378
- 19 Bulten W, Kartasalo K, Chen PHC et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med* 2022; 28: 154–63
- 20 Flach RN, van Dooijeweert C, Aben KKH et al. Interlaboratory Gleason grading variation affects treatment: a Dutch historic cohort study in 30 509 patients with prostate cancer. *J Clin Pathol* 2022; 76: 690–7
- 21 Perincheri S, Levi AW, Celli R et al. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Mod Pathol* 2021; 34: 1588–95
- 22 Wulczyn E, Nagpal K, Symonds M et al. Predicting prostate cancer specific-mortality with artificial intelligence-based Gleason grading. *Commun Med* 2021; 1: 10
- 23 Sandeman K, Blom S, Koponen V et al. AI model for prostate biopsies predicts cancer survival. *Diagnostics* 2022; 12: 1031

Correspondence: Agnieszka Krzyzanowska, Division of Urological Cancers, Department of Translational Medicine, Lund University, Scheelevägen 8, Building 404, 223 63 Lund, Sweden.

e-mail: [agnieszka.krzyzanowska@med.lu.se](mailto:agnieszka.krzyzanowska@med.lu.se)

Abbreviations: AI, artificial intelligence; AS, active surveillance; AUC, area under the ROC curve; BCR, biochemical recurrence; BCR–, treated with no BCR; BCR+, treated with BCR; CONSORT, Consolidated Standards of Reporting Trials; GG, Grade Group; HR, hazard ratio; ISUP, International Society of Urological Pathology; PCa, prostate cancer; PRIAS, Prostate Cancer Research International Active Surveillance; ROC, receiver operating characteristic; RP, radical prostatectomy.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Table S1.** List of slides included in the study together with the year and the site of origin.

**Table S2.** Average cancer size per slide.

**Fig. S1.** (A) Number of slides diagnosed by each pathologist and the algorithm's AUC result on these slides. Pathologists 9–23 diagnosed <100 slides each of the slides included in the study – the AUC results were averaged for this group. (B) ROC curves for pathologists 1–8 and their corresponding AUC values. The black curve represents the average of all 23 pathologists.