Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study

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*

Summary

Background An increasing volume of prostate biopsies and a worldwide shortage of urological pathologists puts a strain on pathology departments. Additionally, the high intra-observer and inter-observer variability in grading can result in overtreatment and undertreatment of prostate cancer. To alleviate these problems, we aimed to develop an artificial intelligence (AI) system with clinically acceptable accuracy for prostate cancer detection, localisation, and Gleason grading.

Methods We digitised 6682 slides from needle core biopsies from 976 randomly selected participants aged 50–69 in the Swedish prospective and population-based STHLM3 diagnostic study done between May 28, 2012, and Dec 30, 2014 (ISRCTN84445406), and another 271 from 93 men from outside the study. The resulting images were used to train deep neural networks for assessment of prostate biopsies. The networks were evaluated by predicting the presence, extent, and Gleason grade of malignant tissue for an independent test dataset comprising 1631 biopsies from 246 men from STHLM3 and an external validation dataset of 330 biopsies from 73 men. We also evaluated grading performance on 87 biopsies individually graded by 23 experienced urological pathologists from the International Society of Urological Pathology. We assessed discriminatory performance by receiver operating characteristics and tumour extent predictions by correlating predicted cancer length against measurements by the reporting pathologist. We quantified the concordance between grades assigned by the AI system and the expert urological pathologists using Cohen's kappa.

Findings The AI achieved an area under the receiver operating characteristics curve of 0·997 (95% CI 0·994–0·999) for distinguishing between benign (n=910) and malignant (n=721) biopsy cores on the independent test dataset and 0·986 (0·972–0·996) on the external validation dataset (benign n=108, malignant n=222). The correlation between cancer length predicted by the AI and assigned by the reporting pathologist was 0·96 (95% CI 0·95–0·97) for the independent test dataset and 0·87 (0·84–0·90) for the external validation dataset. For assigning Gleason grades, the AI achieved a mean pairwise kappa of 0·62, which was within the range of the corresponding values for the expert pathologists (0·60–0·73).

Interpretation An AI system can be trained to detect and grade cancer in prostate needle biopsy samples at a ranking comparable to that of international experts in prostate pathology. Clinical application could reduce pathology workload by reducing the assessment of benign biopsies and by automating the task of measuring cancer length in positive biopsy cores. An AI system with expert-level grading performance might contribute a second opinion, aid in standardising grading, and provide pathology expertise in parts of the world where it does not exist.

Funding Swedish Research Council, Swedish Cancer Society, Swedish eScience Research Center, EIT Health.

Introduction

Histopathological evaluation of prostate biopsies is crucial to the clinical management of men suspected of having prostate cancer. However, the histopathological diagnosis of prostate cancer is associated with several challenges. More than one million men undergo prostate biopsy in the USA annually.[1] With the standard biopsy procedure resulting in 10–12 needle cores per patient, more than 10 million tissue samples need to be examined by pathologists. The increasing incidence of prostate cancer in an ageing population means that the number of biopsies is likely to further increase. Additionally, a global shortage of pathologists exists. For example, China has only one pathologist per 130 000 population, and in many African countries the ratio is in the order of one per million.[2,3] Western countries are facing similar problems, with an expected decline in the number of practicing pathologists due to retirement.[4] Gleason grade is a strong prognostic factor for the survival of patients with prostate cancer and is crucial for treatment decisions.

Lancet Oncol 2020

Published Online January 8, 2020 https://doi.org/10.1016/S1470-2045(19)30738-7

See Online/Comment/ https://doi.org/10.1016/S1470-2045(19)30793-4

*These authors contributed equally

Department of Medical Epidemiology and Biostatistics (P Ström MSc, H Olsson MSc, J Lindberg PhD, Prof H Grönberg MD, M Rantalainen PhD, M Eklund PhD) **and Department of Oncology and Pathology** (Prof L Egevad MD), **Karolinska Institutet, Stockholm, Sweden; Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland** (K Kartasalo MSc, P Ruusuvuori PhD); **Centre for Image Analysis, Department of Information Technology** (L Solorzano MSc, Prof C Wählby PhD) **and Department of Immunology, Genetics, and Pathology** (C Lindskog PhD), **Uppsala University, Uppsala, Sweden; Department of Pathology and Molecular Medicine, Wellington School of Medicine and Health Sciences, University of Otago, Wellington, New Zealand** (Prof B Delahunt MD); **Barts Cancer Institute, Queen Mary University of London, London, UK** (Prof D M Berney MD); **Bostwick Laboratories, Orlando, FL, USA** (Prof D G Bostwick MD); **Laboratory Medicine Program, University Health Network, Toronto General Hospital, Toronto, ON, Canada** (A J Evans MD, Prof T H van der Kwast MD); **Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA**

(Prof D J Grignon MD); **Department of Pathology, Yale University School of Medicine, New Haven, CT, USA** (Prof P A Humphrey MD); **Department of Pathology, Medical College of Wisconsin, Milwaukee, WI, USA** (Prof K A Iczkowski MD); **Department of Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital and Central Clinical School, University of Sydney, Sydney, NSW, Australia** (Prof J G Kench MD); **Institute of Pathology, University Hospital Bonn, Bonn, Germany** (Prof G Kristiansen MD); **Department of Urology, Laboratory of Medical Research, University of São Paulo Medical School, São Paulo, Brazil** (Prof K R M Leite MD); **Pathology and Laboratory Medicine Institute, Cleveland Clinic, Cleveland, OH, USA** (J K McKenney MD); **Department of Cellular Pathology, Southmead Hospital, Bristol, UK** (J Oxley MD); **Department of Pathology, Taipei Veterans General Hospital, Taipei, Taiwan** (C Pan MD); **Aquesta Uropathology and University of Queensland, Brisbane, QLD, Australia** (Prof H Samaratunga MD); **Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada** (Prof J R Srigley MD); **Department of Pathology, Jikei University School of Medicine, Tokyo, Japan** (H Takahashi MD); **Department of Surgical Pathology, School of Medicine, Aichi Medical University, Nagakute, Japan** (Prof T Tsuzuki MD); **Department of Cellular Pathology, University Hospital of Wales, Cardiff, UK** (M Varma MD); **Department of Pathology, UT Southwestern Medical Center, Dallas, TX, USA** (Prof M Zhou MD); **BioImage Informatics Facility of SciLifeLab, Uppsala, Sweden** (Prof C Wählby); **and Department of Oncology, St Göran Hospital, Stockholm, Sweden** (Prof H Grönberg)

Research in context

Evidence before this study

We did a literature search in PubMed, searching the title, abstract, and keywords of peer-reviewed, English-language journal and conference articles published between database inception and May 17, 2019, using the terms "prostate cancer" AND "histo*" AND ("machine learning" OR "deep learning" OR "artificial intelligence"). We also examined the reference lists of relevant publications. Contemporary studies using whole slide imaging of entire histopathological slides and deep learning techniques have shown promising results for detection of prostate cancer, and attempts at grading in prostatectomies and tissue microarrays. These previous studies have not shown experienced urological pathologist-level consistency in grading or investigated grading of needle biopsies, which is the diagnostic sampling method used in routine clinical practice. Moreover, automated estimation of tumour burden in biopsies has not been reported. None of the previous studies have relied on a well defined sample cohort, which allows for clinically meaningful estimation of diagnostic performance metrics, such as sensitivity and specificity.

Added value of this study

To the best of our knowledge, we present for the first time an algorithm that reaches a performance comparable to experienced urological pathologists in the detection, tumour burden estimation, and grading of prostate cancer in needle biopsies. The AI system was developed and evaluated on a population-based dataset prospectively collected within a clinical trial, which included standardised biopsy procedures, centralised pathology reporting, and blinding to clinical characteristics, such as PSA. This dataset represents a broad spectrum of malignant morphologies of prostatic tissue encountered in clinical practice.

Implications of all the available evidence

Use of AI to assist pathologists could substantially decrease their workload by pre-screening cases and by automatically estimating tumour burden, improve patient safety by alarming about potentially missed cancers, and reduce variability in grading by providing decision support. Our results warrant prospective validation in clinical trials to confirm the potential benefits of AI-assisted prostate histopathology in routine clinical practice.

Gleason grade is based on morphological examination and is recognised as subjective. This subjectivity is reflected in high intrapathologist and interpathologist variability in reported grades, as well as both underdiagnosis and overdiagnosis of prostate cancer.[5,6]

A possible solution to these challenges is the application of artificial intelligence (AI) to prostate cancer histopathology. The development of an AI system to identify benign biopsies with high accuracy could decrease the workload of pathologists and allow them to focus on difficult cases. Furthermore, an accurate AI could assist the pathologist with the identification, localisation, and grading of prostate cancer among those biopsies not excluded in the initial screening process, thus providing a safety net to protect against potential misclassification of biopsies. AI-assisted pathology assessment could reduce inter-observer variability in grading, leading to more consistent and reliable diagnoses and better treatment decisions.

By use of high resolution scanning, tissue samples can be digitised to whole slide images and used as the input for the training of deep neural networks (DNNs), an AI technique that has achieved state-of-the-art accuracy in many classification problems across various fields, including medical imaging.[7–10] However, little work has been undertaken in prostate diagnostic histopathology.[11–16] Attempts at grading prostate biopsies by DNNs have been limited to small datasets or subsets of Gleason patterns, and they have not analysed the clinical implications of the introduction of AI-assisted prostate pathology. In this study, we aimed to develop an AI system with clinically acceptable accuracy for prostate cancer detection, localisation, and Gleason grading.

Methods

Study design and participants

Between May 28, 2012, and Dec 30, 2014, the prospective, population-based, screening-by-invitation STHLM3 study (ISRCTN84445406) evaluated a diagnostic model for prostate cancer in men aged 50–69 years residing in Stockholm, Sweden.[17,18] STHLM3 participants had 10–12-core transrectal ultrasound-guided systematic biopsies if they had prostate-specific antigen (PSA) concentration of 3 ng/mL or more or a Stockholm3 test score of 10% or more. Urologists who participated in the study and the study pathologist were blinded to the clinical characteristics of the patients. A single pathologist (LE) graded all biopsy cores according to the International Society of Urological Pathology (ISUP) grading classification (where Gleason scores 6, 3+4, 4+3, 8, and 9–10 are reported as ISUP grade 1 to 5, also referred to as Gleason Grade Groups).[19] LE also delineated cancerous areas using a marker pen and measured the linear cancer extent.

The biopsy cores were formalin fixed and stained with haematoxylin and eosin. A random selection of 8571 biopsies from 1289 STHLM3 participants stratified by ISUP grade was digitised (figure 1). The cases were chosen to represent the full range of diagnoses, with an overrepresentation of high-grade disease. To further enrich the data with high-grade cases, 271 slides from 93 men with ISUP 4 and 5 prostate cancers were obtained from outside STHLM3 (figure 1; appendix p 3). These slides were regraded by LE, digitised, and used for training purposes only. We used 1631 cores from a random selection of 246 (19·1%) men to evaluate the performance of the AI (the independent test set), and the rest were used

for model training. All biopsies from a given participant were assigned to either the training or the test dataset.[20]

Because slides from different pathology labs differ in appearance and quality due to differences in slide preparation and because the characteristics and appearance of whole slide images vary by scanner, assessment of the performance of DNN models on external labs and scanners (ie, images of slides from different pathology labs and scanners than the images on which the model was trained) from a real-world clinical setting is crucial. We therefore obtained 330 slides (73 men) from the Karolinska University Hospital and digitised them on the scanner available at the hospital's pathology laboratory to replicate their entire workflow of processing and slide digitisation (the external validation dataset; figure 1). The selection of slides was enriched for higher ISUP grades to permit evaluation of predictions for these uncommon grades. LE graded all biopsies in the external test dataset to avoid confounding from introducing a different reporting pathologist and a different laboratory and scanner workflow simultaneously.

As an additional test dataset, we digitised 87 cores from the Pathology Imagebase, a reference database launched by ISUP to promote the standardisation of reporting of urological pathology (figure 1).[21] These cases were independently reviewed by 23 highly experienced urological pathologists (the ISUP Imagebase panel). The experts were selected on the basis of their international reputation and scientific production. A Medline search informed that they had authored an average of 105 papers on prostate pathology (range 21–321), with an average of 39 first-author or last-author papers (5–190) at the time of

Correspondence to:
Dr Martin Eklund, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm SE-171 77, Sweden
martin.eklund@ki.se

See **Online** for appendix

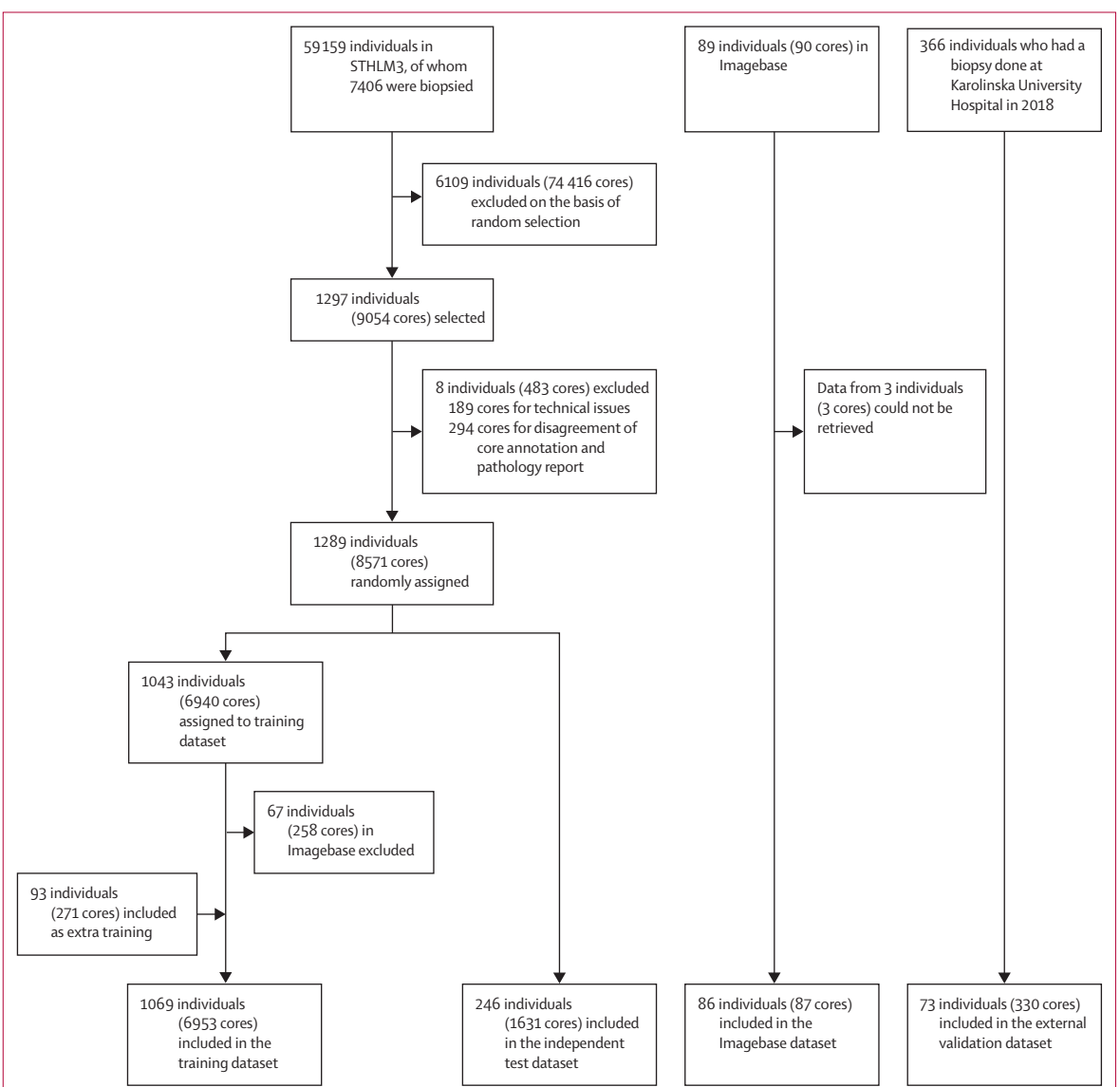

Figure 1: **Study profile**

	STHLM3	Participants (n=1454)				
	Biopsied (n=7406)	Training (n=976)	Extra training (n=93)	Test (n=246)	Imagebase (n=86)	External (n=73)
Age, years						
<49	45 (0·6%)	4 (0·4%)	0	1 (0·4%)	0	2 (2·7%)
50–54	639 (8·6%)	76 (7·8%)	2 (2·2%)	11 (4·5%)	10 (11·6%)	5 (6·8%)
55–59	1221 (16·5%)	136 (13·9%)	4 (4·3%)	44 (17·9%)	8 (9·3%)	10 (13·7%)
60–64	2027 (27·4%)	255 (26·1%)	5 (5·4%)	67 (27·2%)	23 (26·7%)	12 (16·4%)
65–69	3294 (44·5%)	482 (49·4%)	14 (15·1%)	115 (46·7%)	44 (51·2%)	15 (20·5%)
≥70	180 (2·4%)	20 (2·0%)	48 (51·6%)	8 (3·3%)	1 (1·2%)	29 (39·7%)
Missing	0	3 (0·3%)	20 (21·5%)	0	0	0
Previous negative biopsy						
Yes	505 (6·8%)	33 (3·4%)	0	13 (5·28%)	7 (8·1%)	..
No	6901 (93·2%)	940 (96·3%)	0	233 (94·72%)	79 (91·9%)	..
Missing	0	3 (0·3%)	93 (100·0%)	0	0	..
Prostate-specific antigen						
<3 ng/mL	1933 (26·1%)	228 (23·4%)	2 (2·2%)	43 (17·48%)	13 (15·1%)	..
3–<5 ng/mL	3458 (46·7%)	428 (43·9%)	2 (2·2%)	100 (40·65%)	48 (55·8%)	..
5–<10 ng/mL	1612 (21·8%)	213 (21·8%)	13 (14·0%)	73 (29·67%)	16 (18·6%)	..
≥10 ng/mL	403 (5·4%)	104 (10·7%)	47 (50·5%)	30 (12·2%)	9 (10·5%)	..
Missing	0	3 (0·3%)	30 (32·3%)	0	0	..
Digital rectal examination						
Abnormal	680 (9·2%)	133 (13·6%)	46 (49·5%)	39 (15·85%)	12 (14·0%)	..
Normal	6726 (90·8%)	840 (86·1%)	8 (8·6%)	207 (84·15%)	74 (86·0%)	..
Missing	0	3 (0·3%)	39 (41·9%)	0	0	..
Prostate volume						
<35 mL	2701 (36·5%)	425 (43·5%)	19 (20·4%)	92 (37·4%)	42 (48·8%)	..
35–<50 mL	2494 (33·7%)	319 (32·7%)	14 (15·1%)	82 (33·33%)	36 (41·9%)	..
≥50 mL	2211 (29·9%)	229 (23·5%)	19 (20·4%)	72 (29·27%)	8 (9·3%)	..
Missing	0	3 (0·3%)	41 (44·1%)	0	0	..
Cancer length						
No cancer	4605 (62·2%)	142 (14·5%)	0	35 (14·23%)	0	16 (21·9%)
>0–1 mm	545 (7·4%)	133 (13·6%)	2 (2·2%)	35 (14·23%)	4 (4·7%)	1 (1·4%)
>1–5 mm	922 (12·4%)	258 (26·4%)	10 (10·8%)	61 (24·8%)	20 (23·3%)	10 (13·7%)
>5–10 mm	449 (6·1%)	135 (13·8%)	17 (18·3%)	28 (11·38%)	20 (23·3%)	6 (8·2%)
>10 mm	885 (11·9%)	308 (31·6%)	64 (68·8%)	87 (35·37%)	42 (48·8%)	40 (54·8%)
Cancer grade*						
Benign	4605 (62·2%)	142 (14·5%)	0	35 (14·2%)	..	16 (21·9%)
ISUP 1 (3+3)	1558 (21·0%)	413 (42·3%)	1 (1·1%)	104 (42·3%)	..	12 (16·4%)
ISUP 2 (3+4)	761 (10·3%)	200 (20·5%)	1 (1·1%)	53 (21·5%)	..	12 (16·4%)
ISUP 3 (4+3)	253 (3·4%)	96 (9·8%)	1 (1·1%)	16 (6·5%)	..	16 (21·9%)
ISUP 4 (4+4, 3+5, and 5+3)	101 (1·4%)	63 (6·5%)	19 (20·4%)	21 (8·5%)	..	8 (11·0%)
ISUP 5 (4+5, 5+4, and 5+5)	128 (1·7%)	62 (6·4%)	71 (76·3%)	17 (6·9%)	..	9 (12·3%)

Data are n (%). No cancer grade information is shown for Imagebase, because the grading of this set of samples was done independently by multiple observers. Imagebase cancer length was assessed by LE. ISUP=International Society of Urological Pathology. *Numbers in brackets are the Gleason scores associated with the ISUP grades.

Table 1: **Baseline characteristics**

recruitment to Imagebase.[21] Cores from the men in the three test datasets were not part of model development and were excluded from any analysis until the final evaluation.

The study protocol was approved by Stockholm regional ethics committee (permits 2012/572-31/1, 2012/438-31/3, and 2018/845-32). Additional details concerning data collection are in the appendix (p 3).

Test methods

We processed the whole slide images with a segmentation algorithm based on Laplacian filtering to identify the regions corresponding to tissue sections and annotations drawn adjacent to the tissue. We then extracted digital pixel-wise annotations, indicating the locations of cancerous tissue of any grade, by identifying the tissue

	STHLM3	Digitised biopsy slides (n=8980)				
	Biopsied (n=83 470)	Training (n=6682)	Extra Training (n=271)	Test (n=1631)	Imagebase (n=87)	External (n=330)
Cancer length						
No cancer	73595 (88·2%)	3724 (55·7%)	1 (0·4%)	910 (55·8%)	0	108 (32·7%)
>0–1 mm	3307 (4·0%)	915 (13·7%)	7 (2·6%)	203 (12·4%)	8 (9·2%)	33 (10·0%)
>1–5 mm	4135 (5·0%)	1239 (18·5%)	41 (15·1%)	295 (18·1%)	44 (50·6%)	77 (23·3%)
>5–10 mm	1822 (2·2%)	591 (8·8%)	85 (31·4%)	150 (9·2%)	24 (27·6%)	75 (22·7%)
>10 mm	611 (0·7%)	213 (3·2%)	111 (41·0%)	73 (4·5%)	11 (12·6%)	37 (11·2%)
Missing	0	0	26 (9·6%)	0	0	0
Cancer grade						
Benign	73595 (88·2%)	3724 (55·7%)	1 (0·4%)	910 (55·8%)	..	108 (32·7%)
ISUP 1 (3+3)	5664 (6·8%)	1530 (22·9%)	1 (0·4%)	349 (21·4%)	..	65 (19·7%)
ISUP 2 (3+4)	2051 (2·5%)	538 (8·1%)	1 (0·4%)	142 (8·7%)	..	63 (19·1%)
ISUP 3 (4+3)	903 (1·1%)	261 (3·9%)	2 (0·7%)	66 (4·0%)	..	49 (14·8%)
ISUP 4 (4+4, 3+5, and 5+3)	689 (0·8%)	424 (6·3%)	45 (16·6%)	92 (5·6%)	..	19 (5·8%)
ISUP 5 (4+5, 5+4, and 5+5)	568 (0·7%)	205 (3·1%)	221 (81·5%)	72 (4·4%)	..	26 (7·9%)

Data are n (%). No cancer grade information is shown for Imagebase, because the grading of this set of samples was done independently by multiple observers. Imagebase cancer length was assessed by LE. ISUP=International Society of Urological Pathology. *Numbers in brackets are the Gleason scores associated with the ISUP grades.

Table 2: **Baseline characteristics of included biopsy cores**

region corresponding to each annotation. To obtain training data representing the morphological characteristics of Gleason patterns 3, 4, and 5, we extracted multiple partially overlapping smaller images, or patches, from each whole slide image. We used patch dimensions of 598×598 pixels (around 540×540 µm) at a resolution corresponding to 10× magnification (pixel size around 0·90 µm). The process resulted in around 5·1 million patches usable for training a DNN (appendix p 24).

We used two convolutional DNN ensembles, each consisting of 30 Inception V3 models pretrained on ImageNet, with classification layers adapted to our outcome.[22,23] The first ensemble performed binary classification of image patches into benign or malignant, while the second ensemble classified patches into Gleason patterns 3–5. To reduce label noise in the second ensemble, we trained it on patches extracted from cores containing only one Gleason pattern (ie, cores with Gleason score 3+3, 4+4, or 5+5). The test data still contained cores of all grades to provide a real-world scenario for evaluation. Each DNN in the first and the second ensemble thus predicted the probability of each patch being malignant, and whether it represented Gleason pattern 3, 4, or 5 (appendix p 25).

Once the probabilities for the Gleason pattern at each location of the biopsy core were obtained from the DNN ensembles, we mapped them to core-specific characteristics (ISUP grade and cancer length) using boosted trees, a machine learning algorithm based on decision tree models and gradient boosting.[24] All cores in the training data were used for training the boosted trees. Specifically, aggregated features from the patch-wise probabilities predicted by each DNN for each core were used as input to the boosted trees, and the clinical assessment of ISUP score and cancer length were used as outcomes. The ISUP grade group was assigned based on a Bayesian decision rule of the core-level classifier to obtain ISUP predictions at a clinically relevant operating point (appendix p 14).

Statistical analysis

No formal sample size calculation was done. We summarised the operating characteristics of the AI system in a receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC), both on core-level and patient-level. We then specified a range of acceptable sensitivities for potential clinical use and evaluated achieved specificity when compared to the pathology report. The enrichment of high-grade disease in the independent test data and the external validation data might inflate the estimated AUC values, because high grades might be easier to discriminate from benign cases compared with ISUP 1 and 2. Therefore, we also estimated the AUC when ISUP 3–5 cases were removed from the independent test and the external validation datasets.

We predicted cancer length in each core and compared it with the cancer length described in the pathology report. The comparison was done with individual and aggregated cores (ie, total cancer length) for each participant. Linear correlation was assessed in all cores and participants, as well as limited to positive cores and men.

Cohen's kappa with linear weights was used for evaluation of the AI's performance against the 23 experienced urological pathologists on the Imagebase test dataset. Linear weights emphasise a higher level of disagreement of ratings further away from each other on the ordinal ISUP scale, in accordance with previous publications on the Imagebase study.[21] Each of the 87 slides in Imagebase

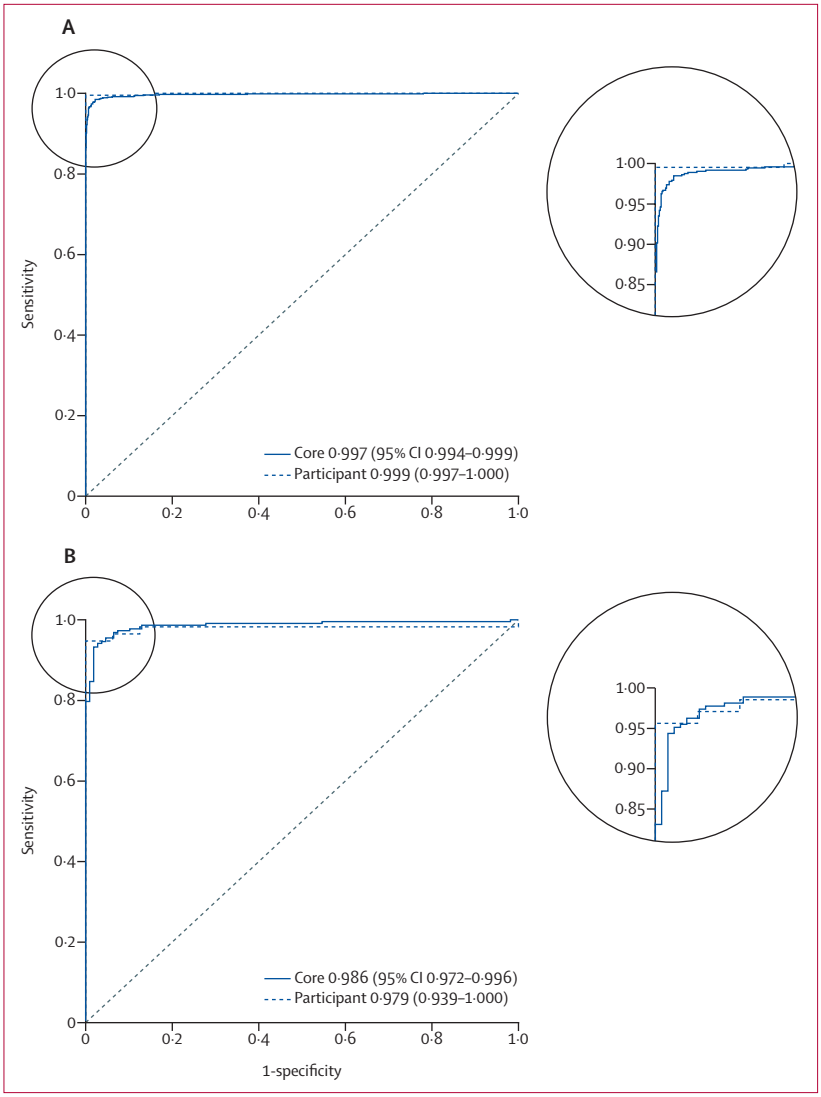

Figure 2: **Receiver operating characteristic curves and AUC for cancer detection in individual cores and individual participants**

(A) Independent test dataset. (B) External validation dataset. Dashed grey lines represent the baseline curve corresponding to random guessing. AUC=area under the curve.

was graded by each of the 23 Imagebase panel pathologists and by the AI system. To evaluate how well the AI system agreed with the pathologists, we calculated all pairwise kappas and summarised the mean for each of the 23 raters. Additionally, we estimated the kappa with a grouping of the Gleason scores in ISUP grades (grade groups) 1, 2–3, and 4–5. We further estimated Cohen's kappa against the study pathologist's ISUP grading of the independent test dataset and the external validation dataset. For the external validation dataset, we also estimated Cohen's kappa after calibrating the probabilities (ie, scaling the ISUP probabilities before assigning the predicted class).

We used t-distributed stochastic neighbour embedding and the deep Taylor decomposition to interpret the representation of the image data learned by the DNN models (appendix p 17).[25]

We excluded cores in which the on-slide annotations did not match the pathology report and cores with technical issues. Participants with missing patient characteristic data were not excluded, because these variables were not used in the statistical analysis.

All CIs are two-sided with 95% confidence and calculated from 1000 bootstrap samples. DNNs were implemented in Python (version 3.6.4) using TensorFlow (version 1.11), and all boosted trees using the Python interface for XGBoost (version 0.72; appendix p 5).

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Among the 59159 STHLM3 participants, 7406 (12·5%) underwent systematic biopsy according to a standardised protocol consisting of 10 or 12 needle cores, with 12 cores being taken from prostates larger than 35 mL (figure 1; tables 1, 2). Among the biopsied participants, we randomly selected 1297, stratified by ISUP score, to be included in this study. After excluding slides with mismatched annotations or technical issues, we randomly split the remaining participants into training and test datasets, resulting in 6682 STHLM3 cores to be used for training of the AI system. We added another 271 cores from outside the study to the training dataset. The data are representative for a screening by invitation setting and include various diagnostically challenging cancer variants encountered in clinical practice (appendix p 35).

The AUC representing the ability of the AI system to distinguish malignant from benign cores was 0·997 (95% CI 0·994–0·999) for the independent test dataset (benign=910, malignant=721) and 0·986 (0·972–0·996) for the external validation dataset (benign=108, malignant=222; figure 2). When ISUP 3–5 cases were removed, AUC values were 0·996 (0·992–0·999) for the independent test dataset and 0·980 (0·959–0·995) for the external validation dataset (appendix p 27). The performance of the AI system for cancer detection is summarised in table 3.

A visualisation of the estimated localisation of malignant tissue for an example biopsy is presented in the appendix (p 33) and the correlation between the cancer length estimates of the AI system and the measurements of the pathologist is presented in figure 3. The correlation between cancer length predicted by the AI and assigned by the reporting pathologist was 0·96 (95% CI 0·95–0·97) for the independent test dataset and 0·87 (0·84–0·90) for the external validation dataset. Further randomly selected example biopsies can be

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

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

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

inspected using TissUUmaps, an online tool for interactive examination of predictions alongside the core tissue. Results of model interpretation are shown in the appendix (pp 31–32).

For Gleason grading, the mean pairwise kappa achieved by the AI system on the 87 Imagebase cases was 0·62. The pathologists had values ranging from 0·60 to 0·73, and the study pathologist (LE) had a kappa of 0·73. When considering a narrower grouping of ISUP grades (ISUP 1, 2–3, and 4–5), which often forms the basis for primary treatment selection, the AI system scored higher than when considering all ISUP grades (figure 4A). The grades assigned by the panel and the AI to each Imagebase case are shown in the appendix (p 26).

The kappa obtained by the AI system relative to the pathology report in the independent test dataset of 1631 cores was 0·83 (figure 4B). The kappa on the external validation dataset was 0·70 (figure 4C). By scaling the ISUP probabilities before assigning the predicted class (calibrating to the new site), the kappa increased to 0·76 on the external validation data (figure 4D). Moreover, we compared the predictions of the AI system and the pathologist in terms of PSA relapses among the participants in the test dataset who underwent radical prostatectomy (appendix pp 22,36)

Discussion

We have shown that an AI system based on DNNs can achieve excellent discrimination between benign biopsy cores versus cores containing cancer and that the time-consuming task of measuring cancer length can be automated with high precision. Moreover, we have shown that an AI system can grade prostate biopsies within the performance range of highly experienced urological pathologists.

For **TissUUmaps** see https://tissuumaps.research.it.uu.se/sthlm3/

Owing to the poor discriminative ability of the PSA test and the systematic biopsy protocol of 10–12 needle cores, which is still in common use, most biopsies encountered in clinical practice are of benign tissue. To reduce the workload of assessing these samples, we evaluated the AI system's potential to assist the pathologist by prescreening benign from malignant cores. Because the pathology report was used as gold standard for this evaluation, the AI system, by design, cannot achieve a higher sensitivity than the reporting pathologist. However, the sensitivity of the AI system could in fact be higher, because some malignant cores might be overlooked by the pathologist but detected by the AI. For example, Ozkan and colleagues[5] evaluated the agreement of two pathologists in the assessment of cancer in biopsy cores. Following examination of 407 cases, one pathologist found cancer in 231 cases, and the other found cancer in 202 cases. This finding suggests that an AI system could not only streamline the workflow, but also improve sensitivity by detecting cancer foci that would otherwise be accidentally overlooked.

The first attempt to use DNNs for the detection of cancer on prostate biopsies was reported by Litjens and colleagues.[15] Using an approach similar to ours, but based on a small dataset, they could safely exclude 32% of benign

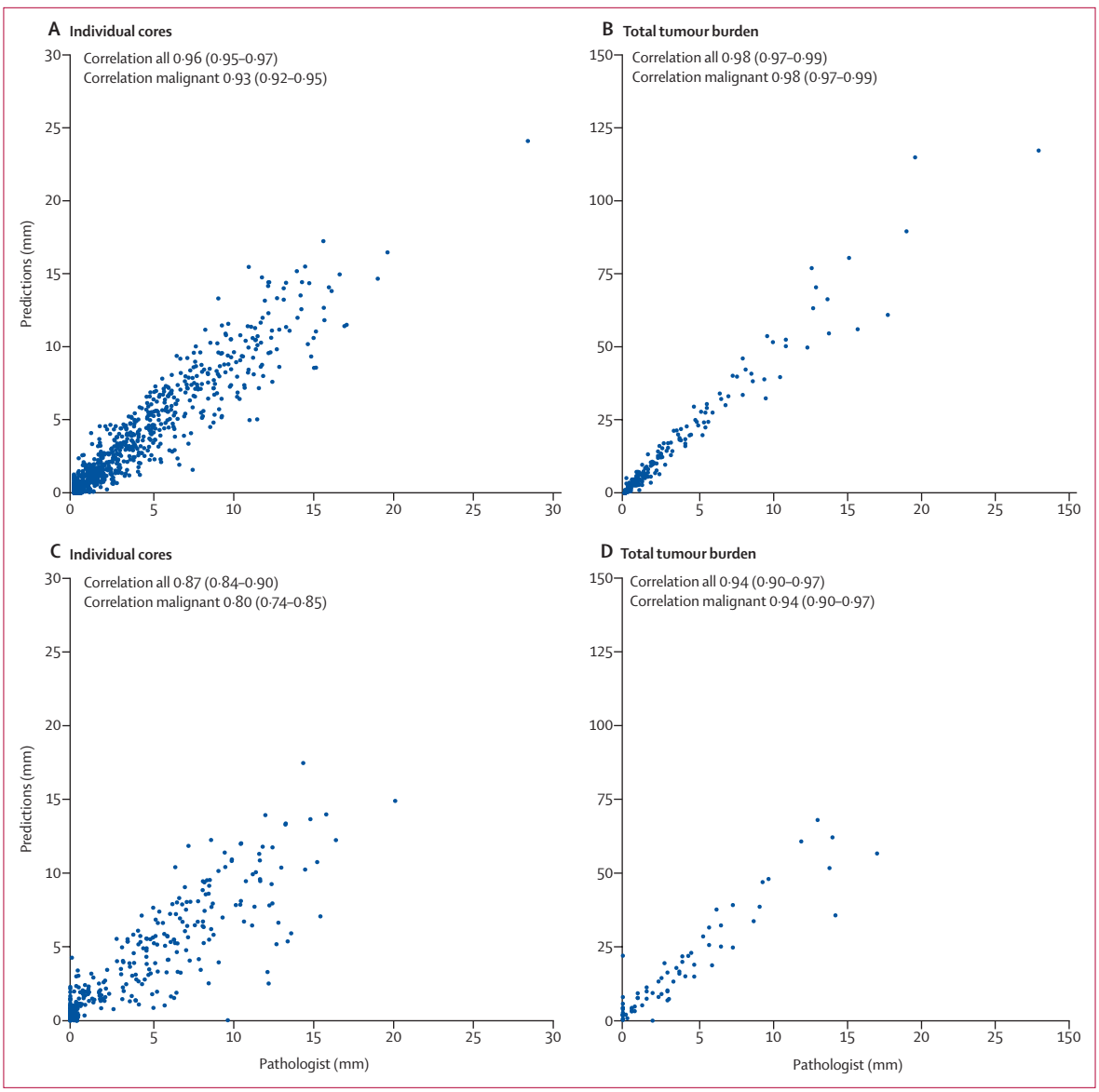

Figure 3: **Concordance between cancer lengths estimated by the AI system and the pathologist**
(A) Individual cores in the independent test dataset. (B) Total tumour burden (per participant) in the independent test dataset. (C) Individual cores in the external validation dataset. (D) Total tumour burden (per participant) in the external validation dataset. Corresponding linear correlation coefficients computed for all cores and malignant cores only are shown in each plot. Datapoints in the left plot are scattered along the x-axis for clarity.

cores. Campanella and colleagues[16] showed an AUC of 0·991 for cancer detection on an independent test dataset and 0·943 on external validation data. Attempts at grading of prostate tissue derived from prostatectomy or based on tissue microarrays have also been made.[14,26] None of these studies achieved expert urological pathologist-level consistency in Gleason grading, estimated tumour burden, or investigated grading on needle biopsies, which is notable because this type of sampling is used for diagnosis and grading in virtually every pathology laboratory worldwide. To the best of our knowledge, no previous study has used a well defined cohort of samples to estimate the clinical implications, with respect to key medical operating characteristic metrics, such as sensitivity and specificity.[27]

The strengths of our study include the use of well controlled data collected within the STHLM3 trial, which included standardised biopsy procedures, centralised pathology reporting, and blinding of both the urologists and the pathologist to clinical characteristics, such as PSA. The prospectively collected, population-based data cover a large random sample of men. Prostate cancers diagnosed in STHLM3 are representative for a screening-by-invitation setting, and the data include cancer variants that are difficult to diagnose (pseudohyperplastic and atrophic carcinoma), slides that required immunohistochemistry,

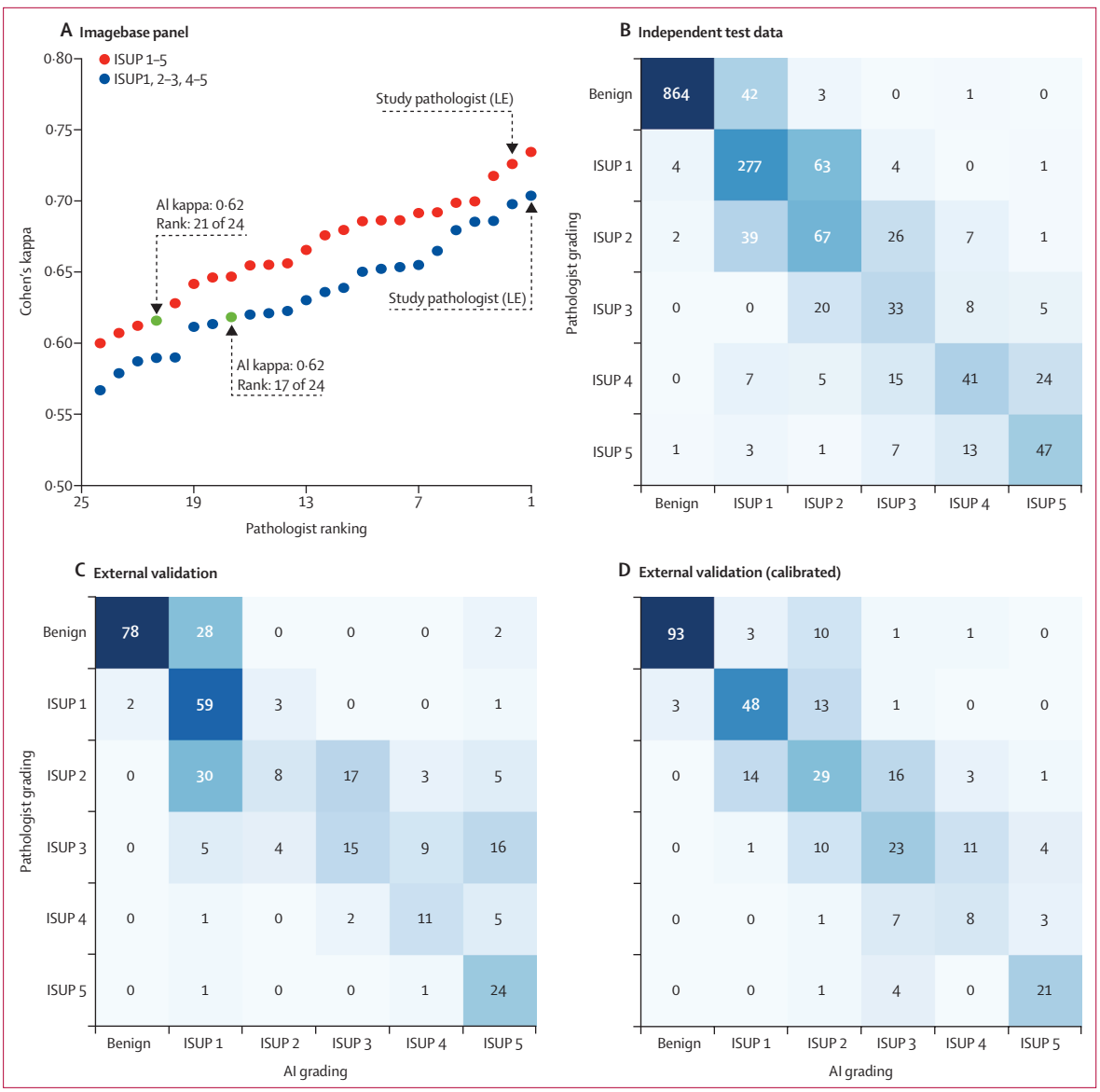

Figure 4: **Gleason grading performance on test data**
(A) Cohen's kappa for each pathologist ranked from lowest to the highest. Each kappa value is the average pairwise kappa for each of the pathologists compared with the others. To account for the natural order of the ISUP scores, we used linear weights. The AI is highlighted with a green dot and an arrow. The study pathologist (LE) is highlighted with an arrow. Values computed based on all five ISUP scores are plotted in red, whereas values based on a grouping of ISUP scores commonly used for treatment decision are shown in blue. (B) A confusion matrix on the independent test data of 1631 slides. (C) A confusion matrix on the external validation data of 330 slides. (D) Results on external validation data following calibration of the slide-level model. The blue shading represents the number of cores in each cell of the matrix. This procedure did not involve any model retraining. The results are presented for an operating point achieving a minimum cancer detection sensitivity of 99%. AI=artificial intelligence. ISUP=International Society of Urological Pathology.

benign mimickers of cancer, slides with thick cuts, and fragmented cores and poor staining. Despite these difficult cases, the AI system achieved excellent diagnostic concordance with the study pathologist. Furthermore, we confirmed that the enrichment of high-grade cases in our datasets did not result in optimistic estimates of discriminative performance. The study was subjected to a strict protocol, in which the splitting of cases into training and test datasets was performed at a patient level and all analyses were prespecified before the evaluation of the independent test dataset, including code for producing tables, figures, and result statistics. A further strength is the use of Imagebase, which is a unique dataset for testing the performance of the AI against highly experienced urological pathologists.

We trained the AI system using annotations from a single, highly experienced urological pathologist (LE). The decision to rely on a single pathologist for model

training was done to avoid presenting the system with conflicting labels for the same morphological patterns and to thereby achieve more consistent predictions. The study pathologist has shown high concordance with other experienced urological pathologists in several studies,[28,29] and therefore represents a good reference for model training. For model evaluation, however, it is crucial to assess performance against multiple pathologists.

Technical variability is introduced during slide preparation and scanning, which might affect the predictions of the AI system. Given the sensitivity of DNNs to differences in input data, differences across labs and scanners could invalidate any discriminatory capacity of a DNN.[30] Here, we showed that the capacity of the AI in discriminating between benign and malignant biopsies decreased, but remained excellent, in the external validation data compared with the independent test dataset. We did, however, observe some reduction in performance with respect to cancer length predictions and overall Gleason grading. By contrast with cancer detection, in which only a handful of correctly predicted patches might be sufficient, cancer length estimation relies on all patches being correctly predicted. Thus, imperfect generalisation is likely to first manifest itself in the length estimates. The reduction in grading performance was most notable for ISUP 2 grades. However, by scaling the AI's predictions for the different classes (ie, calibrating five scalar parameters to the new site), the results were more similar to the results achieved on the independent test data. This is a key observation, because it suggests that although some fine tuning to a new site or scanner is likely required to achieve optimal performance, this tuning is lightweight and can be done using little data. Notably, it does not require redevelopment or retraining of either the DNN models or the slide-level models, which would be infeasible both from a practical and regulatory perspective. Albeit a limitation of the method, requirement for such calibration is not uncommon when using a diagnostic test at a new site (eg, calibrants are routinely used in laboratory diagnostics to diagnose and prevent site-specific differences and variation in test results over time) and is unlikely to present a major hurdle for the clinical application of AI-based diagnostics.

A limitation of this study is the absence of exact pixel-wise annotations, because the annotations might highlight regions that include a mixture of benign and malignant glands of different grades. To address this issue, we trained the algorithm on slides with pure Gleason grades, used a patch size large enough to cover glandular structures, but small enough to minimise the presence of mixed grades within a patch, and we focused our attention on core and patient performance metrics, which avoids caveats of patch-level evaluation and is clinically more meaningful. Another limitation is the difficulty of using a subjective measure like ISUP grade as ground truth for AI models. We approached this problem by evaluating the ISUP grade assigned by the AI against a panel of experienced pathologists. We also confirmed that the classifications of the AI did not substantially differ from the pathologist's when evaluating PSA relapses among the operated men in the trial.

We believe that the use of an AI system like the one presented in this Article could increase sensitivity and promote patient safety by focusing the attention of the pathologist on regions of interest, reduce pathology workload by automated culling of benign biopsies, and reduce the high intra-observer variability in the reporting of prostate histopathology by producing reproducible decision support for grading. A further benefit is that AI can provide diagnostic expertise in regions where it is unavailable.

Contributors

ME had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. PS and KK contributed equally to algorithmic design, implementation, and drafting the manuscript. Additionally, PS was mainly responsible for statistical analysis of results and KK was mainly responsible for high-performance computing. HO was mainly responsible for data management and participated in algorithmic design and implementation, and in drafting the manuscript. LS developed the online viewer application allowing visual examination of results. BD was involved in drafting the manuscript. BD, DMB, DGB, LE, AJE, DJG, PAH, KAI, JGK, GK, THVDK, KRML, JKMK, JO, C-CP, HS, JRS, HT, TT, MV, and MZ did grading of the Imagebase dataset and provided pathology expertise and feedback. CL was involved in data collection. JL was involved in study design. PR and CW contributed to design and supervision of the study and to algorithmic design. Additionally, PR contributed to high-performance computing and CW contributed to designing the online viewer. HG contributed to the conception, design and supervision of the study. MR contributed to the conception, design and supervision of the study and to algorithmic design. LE graded and annotated all the data used in the study, contributed to the conception, design, and supervision of the study, and helped draft the manuscript. ME was responsible for the conception, design and supervision of the study, and contributed to algorithmic design, analysis of results and drafting the manuscript. All authors participated in the critical revision and approval of the manuscript.

Declaration of interests

PS and KK are named on a pending patent (1900061-1) related to cancer diagnostics quality control. HG has five patents (WO2013EP74259 20131120, WO2013EP74270 20131120, WO2018EP52473 20180201, WO2015SE50272 20150311, and WO2013SE50554 20130516) related to prostate cancer diagnostics pending, and has patent applications licensed to Thermo Fisher Scientific. ME has four patents (WO2013EP74259 20131120, WO2013EP74270 20131120, WO2018EP52473 20180201, and WO2013SE50554 20130516) related to prostate cancer diagnostics pending, and has patent applications licensed to Thermo Fisher Scientific, and is named on a pending patent (1900061-1) related to cancer diagnostics quality control. Karolinska Institutet collaborates with Thermo Fisher Scientific in developing the technology for the STHLM3 study. All other authors declare no competing interests.

Acknowledgments

Funding was provided by the Swedish Research Council, Swedish Cancer Society, Swedish Research Council for Health, Working Life, and Welfare, Swedish eScience Research Center, Walter Ahlström Foundation, Tutkijat maailmalle programme, Academy of Finland (313921), Cancer Society of Finland, Emil Aaltonen Foundation, Finnish Foundation for Technology Promotion, Industrial Research Fund of Tampere University of Technology, KAUTE Foundation, Orion Research Foundation, Svenska Tekniska Vetenskapsakademien i Finland, Tampere University Foundation, Tampere University graduate school, The Finnish Society of Information Technology and Electronics, TUT

on World Tour programme, the European Research Council (grant ERC-2015-CoG 682810), and EIT Health. The Tampere Center for Scientific Computing and CSC-IT Center for Science, Finland are acknowledged for providing computational resources. ME and MR report funding from the Swedish Research Council and Swedish Cancer Society. ME reports funding from the Swedish Research Council for Health, Working Life, and Welfare, and Swedish eScience Research Center. The Saint Göran Hospital, Stockholm, is acknowledged for providing additional high-grade slides as training data. Carin Cavalli-Björkman, Britt-Marie Hune, Astrid Björklund, and Olof Cavalli-Björkman have been instrumental in logistical handling of the glass slides. Hannu Hakkola, Tomi Häkkinen, Leena Latonen, Kaisa Liimatainen, Teemu Tolonen, Masi Valkonen, and Mira Valkonen are acknowledged for their helpful advice. We thank the participants in the Stockholm-3 study for their participation.

References

1 Loeb S, Carter HB, Berndt SI, Ricker W, Schaeffer EM. Complications after prostate biopsy: data from SEER-Medicare. *J Urol* 2011; **186:** 1830–34.

2 Egevad L, Delahunt B, Samaratunga H, et al. The International Society of Urological Pathology Education web—a web-based system for training and testing of pathologists. *Virchows Arch* 2019; **474:** 577–84.

3 Adesina A, Chumba D, Nelson AM, et al. Improvement of pathology in sub-Saharan Africa. *Lancet Oncol* 2013; **14:** e152–57.

4 Robboy SJ, Weintraub S, Horvath AE, et al. Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. *Arch Pathol Lab Med* 2013; **137:** 1723–32.

5 Ozkan TA, Eruyar AT, Cebeci OO, Memik O, Ozcan L, Kuskonmaz I. Interobserver variability in Gleason histological grading of prostate cancer. *Scand J Urol* 2016; **50:** 420–24.

6 Melia J, Moseley R, Ball RY, et al. A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology* 2006; **48:** 644–54.

7 Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; **318:** 2199–210.

8 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–18.

9 Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature* 2016; **529:** 484–89.

10 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316:** 2402–10.

11 Gummeson A, Arvidsson I, Ohlsson M, et al. Automatic Gleason grading of H and E stained microscopic prostate images using deep convolutional neural networks. In: Gurcan MN, Tomaszewski JE, eds. Proceedings of SPIE. Volume 10140. Medical Imaging 2017: Digital Pathology. Bellingham, WA: SPIE, 2017.

12 Källén H, Molin J, Heyden A, Lundström C, Åström K. Towards grading gleason score using generically trained deep convolutional neural networks. 2016 IEEE 13th International Symposium on Biomedical Imaging; Prague; 2016 (1163–167).

13 Jiménez del Toro O, Atzori M, Otálora S, et al. Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score. In: Gurcan MN, Tomaszewski JE, eds. Proceedings of SPIE. Volume 10140. Medical Imaging 2017: Digital Pathology. Bellingham, WA: SPIE, 2017.

14 Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep* 2018; **8:** 12054.

15 Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep* 2016; **6:** 26286.

16 Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019; **25:** 1301–09.

17 Grönberg H, Adolfsson J, Aly M, et al. Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study. *Lancet Oncol* 2015; **16:** 1667–76.

18 Ström P, Nordström T, Aly M, Egevad L, Grönberg H, Eklund M. The Stockholm-3 model for prostate cancer detection: algorithm update, biomarker contribution, and reflex test potential. *Eur Urol* 2018; **74:** 204–10.

19 Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 International Society of Urological Pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. *Am J Surg Pathol* 2016; **40:** 244–52.

20 Nir G, Karimi D, Goldenberg SL, et al. Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. *JAMA Netw Open* 2019; **2:** e190442.

21 Egevad L, Delahunt B, Berney DM, et al. Utility of Pathology Imagebase for standardisation of prostate cancer grading. *Histopathology* 2018; **73:** 8–18.

22 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition; Las Vegas; 2016 (2818–26).

23 Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition; Miami; 2009 (248–55).

24 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco; 2016 (785–94).

25 van der Maaten L, Hinton GE. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res* 2008; **9:** 2579–605.

26 Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med* 2019; **2:** 48.

27 Nature. AI diagnostics need attention. *Nature* 2018; **555:** 285.

28 Kweldam CF, Nieboer D, Algaba F, et al. Gleason grade 4 prostate adenocarcinoma patterns: an interobserver agreement study among genitourinary pathologists. *Histopathology* 2016; **69:** 441–49.

29 Egevad L, Cheville J, Evans AJ, et al. Pathology Imagebase—a reference image database for standardization of pathology. *Histopathology* 2017; **71:** 677–85.

30 Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv* 2015; published online March 20. 1412.6572 (preprint).