Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study

Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, Geert Litjens

Summary

Background The Gleason score is the strongest correlating predictor of recurrence for prostate cancer, but has substantial inter-observer variability, limiting its usefulness for individual patients. Specialised urological pathologists have greater concordance; however, such expertise is not widely available. Prostate cancer diagnostics could thus benefit from robust, reproducible Gleason grading. We aimed to investigate the potential of deep learning to perform automated Gleason grading of prostate biopsies.

Methods In this retrospective study, we developed a deep-learning system to grade prostate biopsies following the Gleason grading standard. The system was developed using randomly selected biopsies, sampled by the biopsy Gleason score, from patients at the Radboud University Medical Center (pathology report dated between Jan 1, 2012, and Dec 31, 2017). A semi-automatic labelling technique was used to circumvent the need for manual annotations by pathologists, using pathologists' reports as the reference standard during training. The system was developed to delineate individual glands, assign Gleason growth patterns, and determine the biopsy-level grade. For validation of the method, a consensus reference standard was set by three expert urological pathologists on an independent test set of 550 biopsies. Of these 550, 100 were used in an observer experiment, in which the system, 13 pathologists, and two pathologists in training were compared with respect to the reference standard. The system was also compared to an external test dataset of 886 cores, which contained 245 cores from a different centre that were independently graded by two pathologists.

Findings We collected 5759 biopsies from 1243 patients. The developed system achieved a high agreement with the reference standard (quadratic Cohen's kappa 0·918, 95% CI 0·891–0·941) and scored highly at clinical decision thresholds: benign versus malignant (area under the curve 0·990, 95% CI 0·982–0·996), grade group of 2 or more (0·978, 0·966–0·988), and grade group of 3 or more (0·974, 0·962–0·984). In an observer experiment, the deep-learning system scored higher (kappa 0·854) than the panel (median kappa 0·819), outperforming 10 of 15 pathologist observers. On the external test dataset, the system obtained a high agreement with the reference standard set independently by two pathologists (quadratic Cohen's kappa 0·723 and 0·707) and within inter-observer variability (kappa 0·71).

Interpretation Our automated deep-learning system achieved a performance similar to pathologists for Gleason grading and could potentially contribute to prostate cancer diagnosis. The system could potentially assist pathologists by screening biopsies, providing second opinions on grade group, and presenting quantitative measurements of volume percentages.

Funding Dutch Cancer Society.

Lancet Oncol 2020

Published Online
January 8, 2020
https://doi.org/10.1016/S1470-2045(19)30739-9

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

Department of Pathology (W Bulten MSc, H Pinckaers MD, T de Bel MSc, J van der Laak PhD, G Litjens PhD), **and Department of Radiology & Nuclear Medicine** (Prof B van Ginneken PhD), **Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, Netherlands, Department of Pathology, Antoni van Leeuwenhoek Hospital, The Netherlands Cancer Institute, Amsterdam, Netherlands** (H van Boven MD); **and Laboratory of Pathology East Netherlands, Hengelo, Netherlands** (R Vink MD, C Hulsbergen-van de Kaa PhD)

Correspondence to:
Wouter Bulten, Department of Pathology, Radboud University Medical Center, Nijmegen 6500 HB, Netherlands
wouter.bulten@radboudumc.nl

Introduction

With 1·2 million new prostate cancer cases each year worldwide,[1] a high incidence-to-mortality ratio, and risk of overdiagnosis and overtreatment,[2] an accurate assessment of patient prognosis is needed. The Gleason score,[3] assigned by a pathologist after microscopic examination of cancer morphology, is the most powerful prognostic marker for patients with prostate cancer. However, substantial inter-observer and intra-observer variability in grading[4,5] reduces its usefulness for individual patients. Specialised urological pathologists have greater concordance,[6] but such expertise is not widely available. Prostate cancer diagnostics could thus benefit from robust, reproducible Gleason grading.

Treatment planning for prostate cancer is based mainly on the biopsy Gleason score. After the biopsy procedure, tissue specimens are formalin-fixed and paraffin-embedded, cut into thin sections, stained with haematoxylin and eosin, and examined under a microscope by a pathologist. The Gleason system stratifies the architectural patterns of prostate cancer into five types, from 1 (low risk) to 5 (high risk). The Gleason score in biopsies is the sum of the most common pattern and the highest secondary pattern (eg, 3+5). Growth patterns 1 and 2 are not or rarely reported for biopsies.[7]

In the latest revision of the Gleason grading system, five prognostically distinct grade groups were introduced;[8] assigning scores 3+3 and lower to group 1, 3+4 to group 2,

Research in context

Evidence before this study

We searched the online databases Medline and arXiv for the query "(gleason AND prostate) AND (deep learning OR convolutional neural network OR machine learning OR image analysis) NOT (MRI OR CT)". We did not use a date restriction, but we limited the search to English. On June 30, 2019, this search generated in 278 results, of which 14 were on image analysis or machine-learning-based Gleason grading using haematoxylin and eosin histopathology. Most of these studies were done on preselected, smaller subimages or tissue microarrays, with only four covering applications on full whole slide images. Of those, only one focused on a small set of 96 prostate biopsies but did not include comparisons to multiple pathologists. One study did include a large dataset and comparison to multiple pathologists but was done on radical prostatectomy specimens and did not include a consensus reference standard.

Added value of this study

We showed that a fully automated deep-learning system can reliably grade prostate cancer in a large cohort of patients. Furthermore, we proposed a new technique to train such a system without needing detailed manual annotations by pathologists, the key limiting factor within computational pathology.

Validation of our method was done using a representative dataset, and the system was compared with an expert reference standard set by three pathologists with a subspeciality in uropathology and more than 20 years of experience. In a separate observer experiment, we showed that the deep-learning system reaches pathologist-level performance and is able to group patients in relevant risk categories.

Implications of all the available evidence

This automated deep-learning system could improve prostate cancer diagnostics, especially in areas where expertise is not readily available or where a higher efficiency is desired. The developed system can be used as a first reader (eg, as a prescreening tool) or as a second reader to support pathologists in their diagnosis. The results add evidence of the merits of automated grading systems and could increase the acceptance of such systems within clinical practice.

4+3 to group 3, 3+5, 5+3, and 4+4 to group 4, and higher scores to group 5. Although clinically relevant, initial research shows that this transition has not reduced the observer variability of the grading system.[9,10]

Artificial intelligence, particularly deep learning, has the potential to increase the quality of Gleason grading by improving consistency and offering expert-level grading independent of location. Deep learning has already been investigated and shown promising use in diagnostics in several medical fields,[11] with examples in radiology,[12] ophthalmology,[13] dermatology,[14] and pathology.[15] For prostate cancer, previous studies have applied feature-engineering approaches to address Gleason grading.[16–18] Eventually, the field transitioned to applications of deep learning for detecting cancer,[19,20] and later Gleason grading of tissue microarrays,[21] prostatectomies,[19] and biopsies.[22] Studies of biopsies have focused solely on Gleason 3 versus Gleason 4 in small datasets.

We aimed to produce a fully automated cancer detection and Gleason grading system for entire prostate biopsies, trained without the need for manual pixel-level annotations, focusing on the full range of Gleason grades, and evaluated on a large cohort of patients with an expert consensus reference standard, a separate observer study, and an external tissue microarray test dataset.

Methods

Study design and participants

For method development and validation, we retrospectively built several distinct datasets: the internal training, tuning, test, and observer datasets and the external training and test datasets.

From digital patient records of the Radboud University Medical Center, all pathologist reports dated between Jan 1, 2012, and Dec 31, 2017, for patients who underwent a prostate biopsy owing to a suspicion of prostate cancer were retrieved. The need for informed consent was waived by the local ethics review board (2016–2275). The reports were anonymised, and a text search was used to establish the highest mentioned Gleason score in each report. Patient reports were then randomly sampled using the train_test_split function of the scikit-learn Python package (version 0.20.2), stratifying by the Gleason score, resulting in an equal distribution of Gleason scores. Each pathology report was read, and for each patient, a single haematoxylin and eosin stained glass slide containing the most aggressive or prevalent part of the tumour was selected for scanning. Additional reports mentioning only benign biopsies were selected. Patients who had neoadjuvant or adjuvant therapy were excluded. The resulting dataset is further referenced to as the internal dataset.

The selected glass slides were scanned using a 3DHistech Pannoramic Flash II 250 (3DHistech, Hungary) scanner at 20× magnification (pixel resolution 0·24 µm). Each scan contained one to six unique biopsies, commonly with two sections per biopsy. After scanning, trained non-experts assessed all slides and coarsely outlined each biopsy, assigning each with a Gleason score or labelling negative on the basis of the pathology report. A fixed number of slides were randomly assigned into datasets for testing or tuning, and the remainder were assigned to the training dataset. Randomisation was stratified by patient and highest Gleason grade.

From the internal test dataset, a subset of 100 biopsies was selected to be presented to a group of pathologists in an observer experiment, further referenced to as the observer dataset. The size of the observer dataset was decided in consultation with experts (HvB, RV, and CHvdK). One of the expert pathologists (CHvdK) selected 20 benign cases manually, controlling for a broad range of tissue patterns, including inflammation and (partial) atrophy. The remaining 80 biopsies were randomly selected, stratified for Gleason grade group on the basis of the reported values of the same pathologist.

The 100 biopsies were made available through an online viewer, PMA.view (Pathomation, Berchem, Belgium), and distributed to an external panel. Panel members were invited to participate in this study at the United States and Canadian Academy of Pathology 2019 annual meeting in Washington, DC, USA (March 16–21, 2019). Interested pathologists were asked to report their current affiliation, their experience with Gleason grading, and the number of cases they viewed annually, and were subsequently asked to invite colleagues in their network who had experience in Gleason grading. All pathologists who graded all 100 biopsies were included. All panel members had experience with Gleason grading, but with a varying amount of experience. No time restriction was given, although we asked that they complete the grading within 6 weeks.

We also evaluated the system on an external, independent, public dataset of tissue microarrays[21] to assess the robustness of the system to data from a different centre (Department of Pathology and Molecular Pathology, University Hospital Zurich, Switzerland).[23] One tissue core of a representative tumour area per patient was taken from an online database, and every sample that was assigned to the test dataset of Arvaniti and colleagues[21] was used for validation, a total of 245 cores. The complete dataset consisted of 886 tissue cores, each corresponding to a single patient. The cases were prepared and stained in an independent lab and scanned using a different scanner. We had no influence on the composition of and made no changes to the external dataset.

Test methods

The data acquisition of the internal dataset resulted in outlined biopsies with a single label per biopsy. More detailed annotations were required to train the deep-learning system to segment individual glands. We preprocessed the biopsies of the training and tuning set in four steps (appendix p 2). First, tissue was automatically distinguished from background using a tissue segmentation network.[24] Second, within tissue areas, a trained tumour detection system[20] was applied to define a rough outline of the tumour. The outlined tumour regions still contained large areas of stroma, inflammation, or other non-epithelial tissue. Third, to refine the tumour masks, each biopsy was processed by an epithelial tissue detection system,[25] after which tissue that was detected as non-epithelial tissue was removed from the tumour mask. Finally, detected tumour tissue was assigned a label on the basis of the Gleason score retrieved from the pathology report. A description of the individual systems is provided in the appendix (p 17).

We first trained a deep-learning system only on biopsies with a pure Gleason score (3+3, 4+4, or 5+5).[26] After training, this initial system was applied to the internal training dataset to set the reference standard. By use of the pathologist reports, the output was automatically refined by removing clearly incorrect label assignments, such as cancerous glands in benign biopsies. Any tissue originating from benign biopsies detected as malignant was relabelled as hard negative (ie, a sample of benign tissue that was difficult for the system to correctly classify) to be oversampled during training. A connected components algorithm, based on the ndimage.label function from the Python SciPy package (version 1.2.1), was applied to ensure that each gland was assigned to a single class.

The patients included in the internal test dataset were independent of the patients in the internal training and tuning datasets. To create a strong reference standard, we asked three pathologists with a subspecialty in urological pathology (CHvdK, HvB, and RV) to grade the biopsies individually through the online viewer, PMA.view, following the International Society of Urological Pathology 2014 guidelines.[27] Clinical information of the patients was not available for the experts.

The reference standard for the internal test dataset was determined in three rounds. In the first round, each pathologist reviewed the biopsies individually. For positive biopsies, each pathologist was asked to report: primary, secondary, and tertiary Gleason grade (if present), total tumour volume, tumour volumes for the growth patterns, and the Gleason grade group. In the second round, each biopsy without consensus was regraded by the pathologist whose score differed from the other two. Additional to the pathologist's initial examination, the Gleason scores of the other pathologists were appended anonymously. Biopsies without consensus after round two were discussed in a consensus meeting.

Our deep-learning system consisted of an extended U-Net[28] that was trained on patches extracted from the internal training dataset. After the aforementioned semi-automatic labelling process, the system was trained on the complete training dataset, including biopsies with mixed Gleason growth patterns. The tuning dataset was used to monitor performance during training and to prevent overfitting. Training of the system was halted if no performance increase was measured on the tuning dataset.

See Online for appendix

The label quality of the internal training dataset was determined by labelling the test cases using the same automated method. We compared the retrieved Gleason scores of the test dataset with the final consensus score

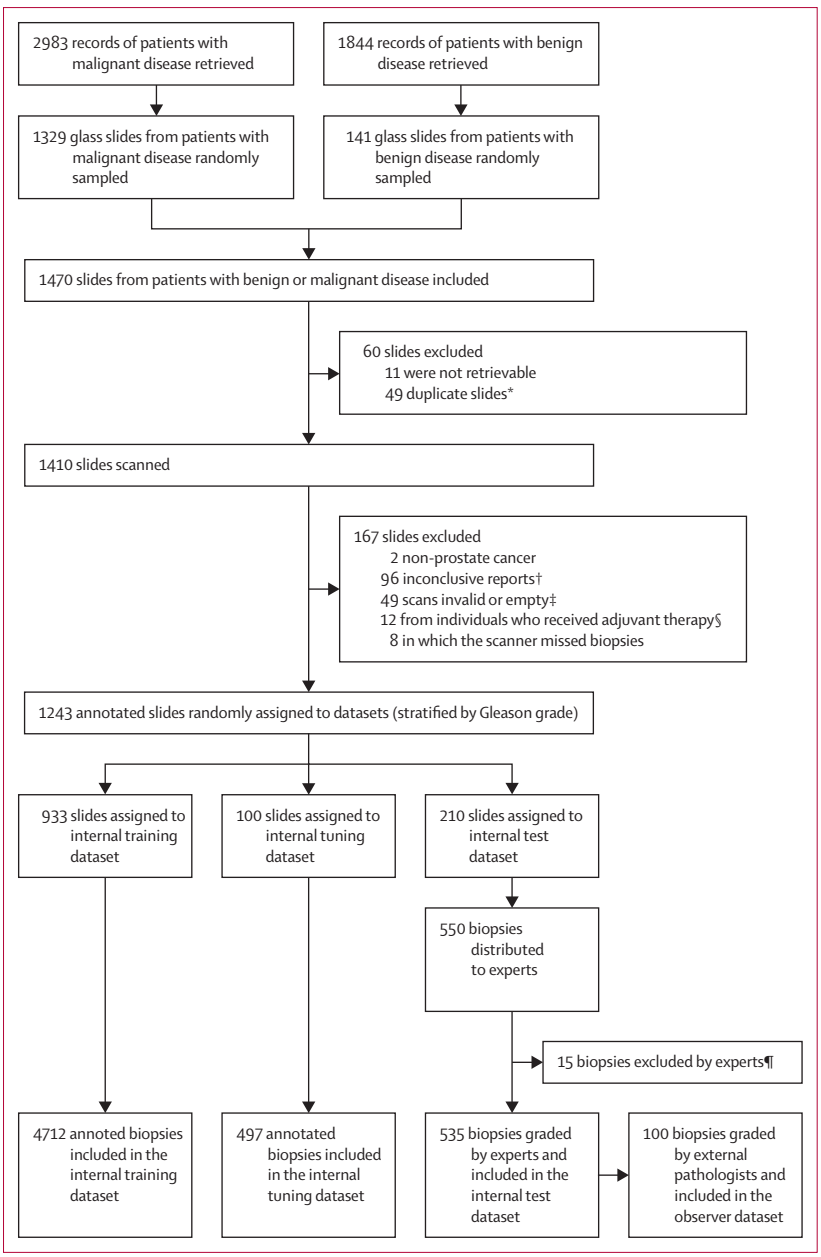

Figure 1: **Study profile**
*For some patients, glass slides were accidentally scanned twice; duplicates were discarded, resulting in one included glass slide per patient. †For some cases, the trained non-experts were unable to match the scanned biopsies to the description from the pathologist report; most commonly, if the report did not explicitly describe the individual biopsies on the glass slide. ‡Scans were excluded if the scanner failed to scan all or the majority of the tissue. §Some patients who had adjuvant therapy were erroneously included the automated text search. ¶Slides were excluded if at least one of the experts determined that the biopsy could not be reliably graded; reasons included biopsy out of focus or not sharp (n=5), biopsy mechanically damaged (n=1), immunohistochemistry needed (n=3), error in loading file (n=1), tumour area too small to grade (n=3), serial section needed (n=1),and image quality too low (n=1).

of the experts. The kappa values of this comparison acted as a measure of label quality.

After training, the deep-learning system was applied to all biopsies from the test dataset and compared with the reference standard. Test positivity cutoffs were determined before the analysis of the test dataset using the tuning dataset. The system determined the grade group of a biopsy in two steps. In the first step, the whole biopsy was segmented by assigning Gleason growth patterns to tumorous glands, and benign glands are classified as benign. From this segmentation, a normalised ratio of epithelial tissue could be calculated as percentages of benign, grade 3, grade 4, or grade 5. Based on the tuning dataset, we classified a biopsy as malignant if at least 10% of the epithelial tissue was predicted as cancer by the system (appendix p 17). In the second step, the grade group was determined on the basis of the normalised volume percentages of each growth pattern.

To apply the system to the external dataset and to account for stain and scanner variations, we applied an unsupervised normalisation algorithm based on CycleGANs.[29] After normalisation of the external test images, our deep-learning system, without any modification, was applied to the normalised test images. The reference standard for the external test dataset was based on the Gleason score. To account for this difference in test metrics, we determined test positive cutoffs on the external training data (appendix p 18). For the external test dataset, no consensus score was available for the two pathologists who graded all cases; instead, we evaluated our method using both pathologists in turn as the reference standard.

Statistical analysis

After consultation with experts (HvB, RV, and CHvdK), 550 cases for the test dataset was established as a good balance between time investment and case diversity.

We defined the main metric as the agreement with the consensus reference standard, measured using quadratic Cohen's kappa. To compare the performance of the system with the external panel of pathologists, we did multiple permutation tests. The test statistic was defined as the difference between the kappa of the deep-learning system and the median kappa of the pathologists. The analysis of the receiver operating characteristic curves was done using the difference in F1 score as test metric. Statistical analysis was done using Python 3.6 with the NumPy (1.16.3), pandas (0.25.1), scipy (1.2.1), scikit-learn (0.20.2) and matplotlib (2.2.4) packages. Further details are provided in the appendix (p 18).

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

After screening of 1329 slides from patients with malignant disease and 141 slides from patients with benign disease, 1410 slides were scanned. After exclusion of 167 slides, the internal dataset consisted of 1243 patients

and 5759 biopsies. The training dataset consisted of 933 (75%) of 1243 slides (4712 biopsies), the tuning dataset of 100 (8%) of 1243 slides (497 biopsies), and the test dataset of 210 (17%) of 1243 slides (550 biopsies; figure 1). The observer dataset was sampled from the test dataset and consisted of 100 biopsies from 78 patients.

After the first round of grading of biopsies in the internal test dataset, 333 (61%) of 550 cases had complete consensus. The three experts' inter-rater agreement was high (quadratic Cohen's kappa of 0·925). The majority vote was taken for some slides: 11 (2%) of 550 cases with an agreement on grade group, but a difference in Gleason pattern order (eg, 5+4 versus 4+5); 11 (2%) cases with an equal grade group, but a disagreement on Gleason score; and 110 (21%) cases for which two pathologists agreed and the third had a maximum deviation of one grade group. Cases with a disagreement on malignancy were always flagged for a second read. Immunohistochemistry was used in seven (1%) of 550 cases to determine the benign or malignant label as it was present in the original report. 15 (3%) of the 550 cases were excluded by the experts because they could not be reliably graded (appendix p 14). In the second round, 63 (11%) of 550 cases were regraded. 27 (5%) of 550 cases of biopsies without consensus after round two were discussed in a consensus meeting (appendix p 15). Grade group distribution and confusion matrices are presented in the appendix (p 4).

To get an estimate of label noise, the reference standard was compared with labels generated by the semi-automatic method; the accuracy of the retrieved labels versus the reference was 0·675 (kappa 0·819) for Gleason score and 0·720 (kappa 0·853) for grade group.

On the 535 biopsies of the internal test dataset, our deep-learning system achieved an agreement of 0·918 (quadratic Cohen's kappa, 95% CI 0·891–0·941) with the consensus grade group. Most errors by the deep-learning system are made in distinguishing between grade group 2 and 3, and grade group 4 and 5 (figure 2, table 1).

In the internal test dataset, the deep-learning system missed 13 malignant cases, of which all but one were determined as grade group 1 by the experts (figure 2). In 12 of these cases, the system detected a tumour, but the predicted volume was below our threshold for malignancy.

Receiver operating characteristic curve analysis on three clinically relevant cutoffs showed the ability of the system to group cases in risk categories with high accuracy (figure 3; table 2). The decision threshold of the system can be tuned to correctly predict 99% of biopsies containing tumour with a specificity of 82%.

For the observer study, 13 pathologists and two pathologists in training from 14 independent labs and ten countries individually graded all 100 biopsies following the International Society of Urological Pathology 2014 guidelines. This external panel showed a median inter-rater agreement of 0·819 (quadratic kappa, 95% CI 0·726–0·869) on Gleason grade group with the consensus

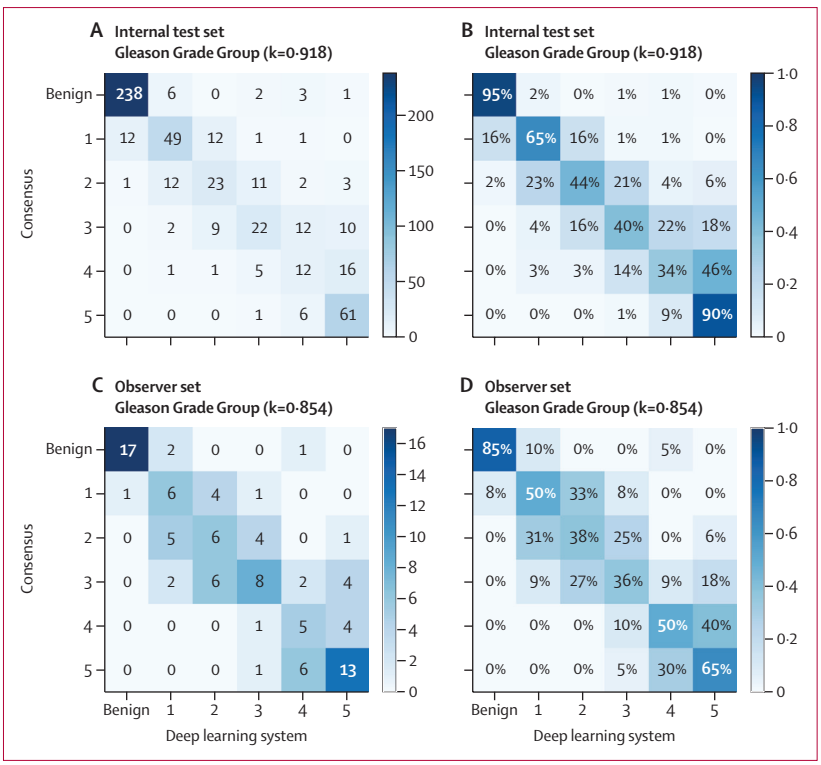

Figure 2: **Confusion matrices on Gleason grade group**
Data are shown for the whole internal test dataset (A, B) and the observer dataset (C, D). Agreement between the system's predictions and the reference standard is shown in quadratic Cohen's kappa above each matrix. Absolute frequency (A, C) and relative frequency (B, D) are shown. For the relative frequency, the number of cases in each cell is divided by the total cases in each row.

	Cases, n (%)	Accuracy	Precision	Recall	Specificity	Negative predictive value
Internal test dataset						
Negative	250 (45%)	0·953	0·948	0·952	0·954	0·958
Grade group 1	75 (14%)	0·912	0·700	0·653	0·954	0·944
Grade group 2	52 (9%)	0·905	0·511	0·442	0·954	0·941
Grade group 3	55 (10%)	0·901	0·524	0·400	0·958	0·933
Grade group 4	35 (6%)	0·912	0·333	0·343	0·952	0·954
Grade group 5	68 (12%)	0·931	0·670	0·897	0·936	0·984
Observer dataset						
Negative	20 (20%)	0·960	0·944	0·850	0·988	0·963
Grade group 1	12 (12%)	0·850	0·400	0·500	0·898	0·929
Grade group 2	16 (16%)	0·800	0·375	0·375	0·881	0·881
Grade group 3	22 (22%)	0·790	0·533	0·364	0·910	0·835
Grade group 4	10 (10%)	0·860	0·357	0·500	0·900	0·942
Grade group 5	20 (20%)	0·840	0·591	0·650	0·887	0·910

For each grade group, the metrics are calculated individually, using cases from that group as positive and the other cases as negative samples. The deep-learning system was not optimised for specific Gleason grade groups but was optimised over the whole range of Gleason grades. Confusion matrices on the deep-learning system's prediction versus the reference standard are shown in figure 2.

Table 1: **Class-wise classification metrics of the deep-learning system for the internal test and observer datasets**

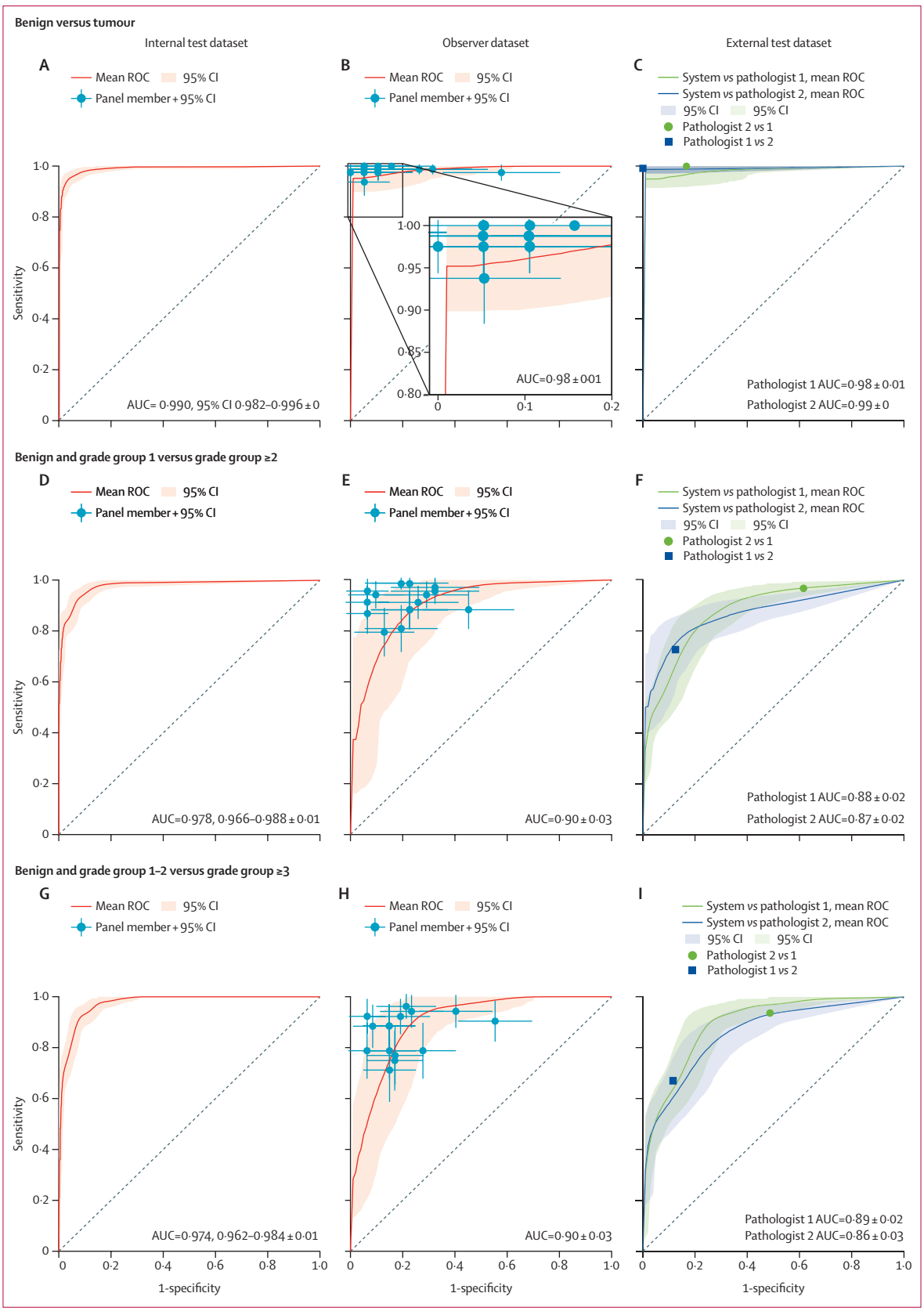

Figure 3: **Bootstrapped ROC analysis on three clinically relevant cutoffs**
Bootstrapped ROC analysis on three clinically relevant cutoffs: tumour versus benign (A–C), benign and grade group 1 versus grade group 2 or more (D–F), and low (benign and grade group 1–2) versus high grade (grade group 3–5) cancer (G–I). A, D, and G show the internal test dataset (535 cases). B, E, and H show the observer dataset (100 cases); the values for each panel member have been added to the graphs. C, F, and I show the external test dataset, comparing the predictions of the deep-learning system with the two pathologists that set the reference standard. ROC=receiver operator characteristic. AUC=area under the curve.

	Number of cases	Area under the curve (95% CI)	F1 score	Accuracy	Precision	Recall	Specificity	Negative predictive value
Internal test set								
Benign *vs* malignant	250/285	0·990 (0·982–0·996)	0·956	0·953	0·958	0·954	0·952	0·948
Benign and grade group 1 *vs* grade group ≥2	325/210	0·978 (0·966–0·988)	0·915	0·933	0·907	0·924	0·938	0·950
Benign and grade group 1–2 *vs* grade group ≥3	377/158	0·974 (0·962–0·984)	0·887	0·931	0·858	0·918	0·936	0·964
Observer set								
Benign *vs* malignant	20/80	0·984 (0·971–1·000)	0·975	0·960	0·963	0·988	0·850	0·944
Benign and grade group 1 *vs* grade group ≥2	32/68	0·904 (0·831–0·964)	0·904	0·870	0·910	0·897	0·812	0·788
Benign and grade group 1–2 *vs* grade group ≥3	48/52	0·899 (0·833–0·956)	0·854	0·850	0·863	0·846	0·854	0·837
External test set (pathologist 1 as reference standard)								
Benign *vs* malignant	12/233	0·980 (0·967–0·997)	0·983	0·967	0·991	0·974	0·833	0·625
Benign and grade group 1 *vs* grade group ≥2	91/154	0·878 (0·834–0·920)	0·851	0·792	0·772	0·948	0·527	0·857
Benign and grade group 1–2 *vs* grade group ≥3	119/126	0·892 (0·851–0·930)	0·844	0·824	0·779	0·921	0·723	0·896
External test set (pathologist 2 as reference standard)								
Benign *vs* malignant	10/235	0·988 (0·984–1·000)	0·987	0·976	1·000	0·974	1·000	0·625
Benign and grade group 1 *vs* grade group ≥2	40/205	0·869 (0·821–0·916)	0·898	0·837	0·937	0·863	0·700	0·500
Benign and grade group 1–2 *vs* grade group ≥3	69/176	0·855 (0·806–0·903)	0·825	0·767	0·899	0·761	0·783	0·562

For this analysis, the threshold of the deep-learning system was not optimised for the different decision cutoffs.

Table 2: **Classification performance metrics for the three datasets using the default decision threshold of the deep-learning system**

(appendix p 5). The system achieved a kappa value of 0·854 (quadratic kappa, 0·777–0·914) on the cases of the observer dataset, scoring higher than the median value of the panel and outperforming 10 of the 15 panel members (appendix p 5). The performance of the deep-learning system was better than that of pathologists with less than 15 years of experience (two-sided permutation test, p=0·036) and scores not significantly different than pathologists with more than 15 years of experience (two-sided permutation test, p=0·96; appendix p 5). To exclude bias towards the experts in our results, we computed the inter-rater agreement between all panel members independently of the reference standard. We then computed the agreement of the system with all members of the panel (appendix p 7). Sorted by the median kappa value, the deep-learning system had the third highest inter-rater agreement score (appendix p 7).

The deep-learning system scores better than three pathologists of the panel but lower than most on accuracy (appendix p 6). The lower accuracy is mostly caused by one-off errors between grade groups 2 versus 3 and 4 versus 5 (figure 2). A two-sided permutation test on the difference between system accuracy and the median of the panel showed no significant difference (p=0·15). See the appendix (p 10) for example cases.

After receiver operating characteristic analysis of the observer dataset, a two-sided permutation test on the median F1 score showed no difference between the deep-learning system and the panel for both malignant versus benign (p=0·70), grade group 2 as a cutoff (p=0·84), and grade group 3 as a cutoff (p=0·65; figure 3; table 2).

The external test dataset contained 245 cores that were independently graded by two pathologists (inter-rater agreement quadratic Cohen's kappa 0·71). Concerning the two pathologists, the system obtained a 0·723 and 0·707 quadratic kappa on Gleason score. A receiver operating characteristic analysis for relevant decision thresholds was also done for the external test set (figure 3; table 2). Overall, the deep learning system performed comparably to the two pathologists who set the reference standard.

Discussion

We have developed a fully automated method to grade prostate biopsies and have shown that this method can achieve a performance similar to pathologists on both the internal and external test datasets. The performance of the deep-learning system could only reliably be assessed by use of an expert reference standard. We asked three expert urological pathologists to grade the complete test dataset, which resulted in a minimum of three independent reads for every case. The deep-learning system achieved a high agreement (quadratic kappa of 0·918) with the reference standard. We also compared the system with a panel of independent pathologists and pathologists in training. In this observer dataset, the deep-learning system outperformed ten of 15 panel members. On the external test dataset, the system showed it could generalise to external and unseen data. The system scored comparably to the results attained by Arvaniti and colleagues[21] (quadratic kappa, 0·723 and 0·707 *vs* 0·71 and 0·75) and within inter-observer variability of the pathologists who set the reference standard (kappa 0·71), although our system was not trained on data from that set.

The training data was labelled in a semi-supervised way, saving resources that would otherwise have been

spent in manually labelling slides. Moreover, it is often practically unfeasible to precisely annotate the vast amounts of data required for deep learning, even though unannotated or sparsely annotated data is often readily available in pathology archives. One limitation of this method is that it can introduce label noise in the training dataset. However, the ability of deep-learning systems to handle substantial amounts of label noise is well known.[30]

The black-box characteristic of deep learning is often mentioned as a drawback of such systems, especially for medical decision making. We addressed this drawback by having our system show predictions at multiple abstraction levels, instead of using an additional learned model on top of the deep-learning system. The precise gland-level segmentations of the developed system made it possible to use a simple ruleset on grade volume percentages to obtain the biopsy-level grade, similar to the one described in the Gleason grading system. This approach allows pathologists to assess whether the epithelium was correctly classified, whether the system missed certain glands, and the grades assigned to individual or groups of glands. As such, our system provides a higher level of interpretability compared with competing approaches.[10]

Given the high prevalence of prostate cancer, reducing workload for pathologists is of clinical value. In the test dataset, our deep-learning system achieved an AUC of 0·990 on determining the malignancy of a biopsy, on the observer set an AUC of 0·984. Furthermore, the system can be tuned to achieve a sensitivity of 99%. As such, our system could be implemented as a prescreening triage tool within pathology labs, giving priority to high-grade biopsies and filtering out low-risk benign biopsies.

More work is needed to increase the discrimination power of the system between Gleason grade groups 2 and 3 and groups 4 and 5. The boundaries between group 2 and 3 are defined by the relative volume percentages of the Gleason growth patterns, which makes an accurate estimate of those volumes essential for correct classification. Group 4 versus 5 is complicated by a wide range of Gleason scores that fall under these two groups (3+5, 5+3, 4+4, 4+5, 5+4, and 5+5). Discrepancies can also be related to the way the system and pathologists differ in estimating the relative volume of growth patterns. The system counts the exact area of the individual glands, whereas a pathologist assesses the volume more qualitatively.

Our results extend previous work on prostate cancer detection[19,20] and automated Gleason grading.[10,21,22] We extend on these works by focusing on automated Gleason grading for prostate biopsies, the strongest histological correlating predictor of recurrence for patients with prostate cancer, of which the grading system differs from prostatectomies. Furthermore, by including both benign biopsies and biopsies from the full spectrum of Gleason grades, we created a system that is usable as a prescreening tool and as a second reader.

For **Automated Gleason Grading** see https://www.computationalpathologygroup.eu/software/automated-gleason-grading/

Grading of the biopsies, both by the experts and the external panel, was done through digital viewing of the slides. For the external panel, not all members had previous experience with digital viewing or with the digital viewer used in this study. Owing to this inexperience, we cannot know whether it affected their grading. Nonetheless, research has shown that digital viewing is non-inferior to microscopy.[31]

Before our system can be used in clinical practice, some limitations must be addressed. First, the data that were used to develop the deep-learning system originated from a single centre. Although the performance on the external test dataset is within the range of inter-observer variability, including data from multiple centres, with different staining protocols and whole slide scanners, could further increase the robustness of the system. Second, we focused on the grading of acinar adenocarcinoma in prostate biopsies, although other tumour types and foreign tissue can be present in prostate biopsies (eg, colon glands, which should be identified and excluded for grading). Additionally, other prognostic information could be present in the biopsies that we did not extract (eg, the detection of intraductal carcinoma).[32] Finally, in this study, each biopsy is treated independently, both by the pathologists and by the deep-learning system. In clinical practice, multiple biopsies are sampled from different regions of the prostate. An update to the deep-learning system could take multiple biopsies into account and give a grade group prediction at the patient level.

The developed system will be made available for scientific and non-commercial use, through the Radboudumc Computational Pathology Group website. Furthermore, through a grand challenge on Gleason grading, we will publish a part of our data for others to use in developing new methods for automated Gleason grading.

Our automated deep-learning system achieved a performance similar to pathologists in terms of Gleason grading. With further evaluation, the system could assist pathologists by screening biopsies, providing second opinions on grade group, and presenting quantitative measurements of volume percentages.

Contributors

WB selected data, did the experiments, analysed the results, and wrote the manuscript. HP was involved with the data collection and experiments. HvB, RV, and CH-vdK graded all cases in the test dataset. TdB was involved with the application of the method to the external data. GL, CH-vdK., JvdL, and BvG supervised the work and were involved in setting up the experimental design. All authors reviewed the manuscript and agreed with its contents.

Declaration of interests

WB and HP report grants from the Dutch Cancer Society, during the conduct of the study. BvG reports that they are co-founder of, shareholder of, and earn royalties from Thirona, grants and royalties from Delft Imaging Systems, and grants from MeVis Medical Solutions, outside the submitted work. JvdL reports personal fees from Philips, ContextVision, and AbbVie and grants from Philips and Sectra, outside the submitted work. GL reports grants from the Dutch Cancer Society,

during the conduct of the study, and grants from Philips Digital Pathology Solutions and personal fees from Novartis, outside the submitted work. All other authors declare no competing interests.

Acknowledgments

This study was funded by a grant from the Dutch Cancer Society (KWF), grant number KUN 2015-7970. We thank the following pathologists and pathologists in training for participating in our study as part of the panel: Paulo G. O. Salles (Instituto Mário Penna, Belo Horizonte, Brazil); Vincent Molinié (CHU de Martinique, Université des Antilles, Fort de France, Martinique); Jorge Billoch-Lima, (HRP Labs, San Juan, Puerto Rico); Ewout Schaafsma (Radboud University Medical Center, Nijmegen, The Netherlands); Anne-Marie Vos (Radboud University Medical Center, Nijmegen, The Netherlands); Xavier Farré (Department of Health, Public Health Agency of Catalonia. Lleida, Catalonia, Spain); Awoumou Belinga Jean-Joël (Department of Morphological Sciences and Anatomic Pathology, Faculty of Medicine and Biomedical Sciences, University of Yaounde 1, Cameroon); Joëlle Tschui (Medics Pathologie, Bern, Switzerland); Paromita Roy (Tata Medical Center, Kolkata, India); Emílio Marcelo Pereira (Oncoclínicas group, Brazil); Asli Cakir (Istanbul Medipol University School of Medicine, Pathology Department, Istanbul, Turkey); Katerina Geronatsiou (Centre de Pathologie, Hopital Diaconat Mulhouse, France); Günter Saile, (Histo- and Cytopathology, labor team w ag, Goldach SG, Switzerland); Américo Brilhante, (SalomãoZoppi Diagnostics, São Paulo, Brazil); Guilherme Costa Guedes Pereira (Laboratory Histo Patologia Cirúrgica e Citologia, João Pessoa-PB, Brazil). We also thank Jeffrey Hoven for assisting with the data collection and scanning, and Milly van de Warenburg, Nikki Wissink, and Frederike Haverkamp for their help making the manual annotations.

References

1 Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018; **68:** 394–424.

2 Schröder FH, Hugosson J, Roobol MJ, et al. Prostate-cancer mortality at 11 years of follow-up. *N Engl J Med* 2012; **366:** 981–90.

3 Epstein JI. An update of the Gleason grading system. *J Urol* 2010; **183:** 433–40.

4 Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Hum Pathol* 2001; **32:** 81–88.

5 Egevad L, Ahmad AS, Algaba F, et al. Standardization of Gleason grading among 337 European pathologists. *Histopathology* 2013; **62:** 247–56.

6 Allsbrook WC Jr, Mangold KA, Johnson MH, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. *Hum Pathol* 2001; **32:** 74–80.

7 Epstein JI, Allsbrook WC Jr, Amin MB, Egevad LL. The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *Am J Surg Pathol* 2005; **29:** 1228–42.

8 Epstein JI, Zelefsky MJ, Sjoberg DD, et al. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. *Eur Urol* 2016; **69:** 428–35.

9 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.

10 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.

11 Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; **42:** 60–88.

12 Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* 2019; **25:** 954–61.

13 De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018; **24:** 1342–50.

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

15 Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; **318:** 2199–210.

16 Naik S, Doyle S, Feldman M, Tomaszewski J, Madabhushi A. Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information. Proceedings of 2nd Workshop on Microsopic Image Analysis with Applications in Biology; Piscataway, NJ, USA; 2007.

17 Gertych A, Ing N, Ma Z, et al. Machine learning approaches to analyze histological images of tissues from radical prostatectomies. *Comput Med Imaging Graph* 2015; **46:** 197–208.

18 Nguyen TH, Sridharan S, Macias V, et al. Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning. *J Biomed Opt* 2017; **22:** 36015.

19 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.

20 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.

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

22 Lucas M, Jansen I, Savci-Heijink CD, et al. Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. *Virchows Arch* 2019; **475:** 77–83.

23 Zhong Q, Guo T, Rechsteiner M, et al. A curated collection of tissue microarray images and clinical outcome data of prostate cancer patients. *Sci Data* 2017; **4:** 170014.

24 Bándi P, van de Loo R, Intezar M, et al. Comparison of different methods for tissue segmentation in histopathological whole-slide images. 2017 IEEE 14th International Symposium on Biomedical Imaging; Melbourne; April 18–21, 2017 (591–95).

25 Bulten W, Bándi P, Hoven J, et al. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Sci Rep* 2019; **9:** 864.

26 Bulten W, Pinckaers H, Hulsbergen-van de Kaa C, Litjens G. Automated gleason grading of prostate biopsies using deep learning. United States and Canadian Academy of Pathology 108th Annual Meeting; Washington, DC; 2019 (abstr 1467).

267 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.

28 Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Cham: Springer International Publishing, 2015.

289 de Bel T, Hermsen M, Kers J, van der Laak J, Litjens G. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In: Cardoso MJ, Aasa F, Ben G, et al, eds. Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning; Proceedings of Machine Learning Research. 2019. 151–63.

30 Rolnick D, Veit A, Belongie S, Shavit N. Deep learning is robust to massive label noise. *arXiv* 2018; published online Feb 26. 1705.10694 (preprint).

31 Mukhopadhyay S, Feldman MD, Abels E, et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). *Am J Surg Pathol* 2018; **42:** 39–52.

32 Kweldam CF, Kümmerlin IP, Nieboer D, et al. Disease-specific survival of patients with invasive cribriform and intraductal prostate cancer at diagnostic biopsy. *Mod Pathol* 2016; **29:** 630–36.