An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study

Liron Pantanowitz, Gabriela M Quiroga-Garza, Lilach Bien, Ronen Heled, Daphna Laifenfeld, Chaim Linhart, Judith Sandbank, Anat Albrecht Shach, Varda Shalev, Manuela Vecsler, Pamela Michelow, Scott Hazelhurst, Rajiv Dhir

Summary

Background There is high demand to develop computer-assisted diagnostic tools to evaluate prostate core needle biopsies (CNBs), but little clinical validation and a lack of clinical deployment of such tools. We report here on a blinded clinical validation study and deployment of an artificial intelligence (AI)-based algorithm in a pathology laboratory for routine clinical use to aid prostate diagnosis.

Methods An AI-based algorithm was developed using haematoxylin and eosin (H&E)-stained slides of prostate CNBs digitised with a Philips scanner, which were divided into training (1 357 480 image patches from 549 H&E-stained slides) and internal test (2501 H&E-stained slides) datasets. The algorithm provided slide-level scores for probability of cancer, Gleason score 7–10 (*vs* Gleason score 6 or atypical small acinar proliferation [ASAP]), Gleason pattern 5, and perineural invasion and calculation of cancer percentage present in CNB material. The algorithm was subsequently validated on an external dataset of 100 consecutive cases (1627 H&E-stained slides) digitised on an Aperio AT2 scanner. In addition, the AI tool was implemented in a pathology laboratory within routine clinical workflow as a second read system to review all prostate CNBs. Algorithm performance was assessed with area under the receiver operating characteristic curve (AUC), specificity, and sensitivity, as well as Pearson's correlation coefficient (Pearson's *r*) for cancer percentage.

Findings The algorithm achieved an AUC of 0·997 (95% CI 0·995 to 0·998) for cancer detection in the internal test set and 0·991 (0·979 to 1·00) in the external validation set. The AUC for distinguishing between a low-grade (Gleason score 6 or ASAP) and high-grade (Gleason score 7–10) cancer diagnosis was 0·941 (0·905 to 0·977) and the AUC for detecting Gleason pattern 5 was 0·971 (0·943 to 0·998) in the external validation set. Cancer percentage calculated by pathologists and the algorithm showed good agreement (r=0·882, 95% CI 0·834 to 0·915; p<0·0001) with a mean bias of –4·14% (–6·36 to –1·91). The algorithm achieved an AUC of 0·957 (0·930 to 0·985) for perineural invasion. In routine practice, the algorithm was used to assess 11 429 H&E-stained slides pertaining to 941 cases leading to 90 Gleason score 7–10 alerts and 560 cancer alerts. 51 (9%) cancer alerts led to additional cuts or stains being ordered, two (4%) of which led to a third opinion request. We report on the first case of missed cancer that was detected by the algorithm.

Interpretation This study reports the successful development, external clinical validation, and deployment in clinical practice of an AI-based algorithm to accurately detect, grade, and evaluate clinically relevant findings in digitised slides of prostate CNBs.

Funding Ibex Medical Analytics.

Lancet Digital Health 2020; 2: e407–16

See **Comment** page e383

Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA (L Pantanowitz MD, G M Quiroga-Garza MD, R Dhir MD); **Department of Anatomical Pathology, University of the Witwatersrand and National Health Laboratory Services, Johannesburg, South Africa** (L Pantanowitz, P Michelow MBBCh); **Ibex Medical Analytics, Tel Aviv, Israel** (L Bien, R Heled, D Laifenfeld PhD, C Linhart PhD, J Sandbank MD, M Vecsler PhD); **Institute of Pathology, Maccabi Healthcare Services, Rehovot, Israel** (J Sandbank); **Shamir Medical Center, Beer Yaakov, Israel** (A Albrecht Shach MD); **KSM Research and Innovation institute, Maccabi Healthcare Services, Tel Aviv, Israel** (V Shalev MD); **and School of Electrical & Information Engineering and Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa** (S Hazelhurst PhD)

Correspondence to:
Dr Liron Pantanowitz, Department of Pathology, UPMC Cancer Pavilion, Pittsburgh, PA 15232, USA
lironp@med.umich.edu

Introduction

Adenocarcinoma of the prostate is the second most common cancer diagnosed in men,[1] with more than 1 million newly diagnosed cases of prostate cancer annually. Hence, prostate specimens are frequently encountered in surgical pathology practice. Today, the histopathological assessment of biopsy tissue is the mainstay of diagnosing prostate cancer, which includes core needle biopsy (CNB) and, if warranted, surgical resection. For most pathology laboratories, the current method of rendering a tissue diagnosis involves light microscopic examination of haematoxylin and eosin (H&E)-stained tissue sections.[2] Due to changing guidelines, there has been a dramatic increase in the number of CNBs reviewed per case over the past decade. Coupled with an increase in overall cancer incidence and growing shortage of pathologists worldwide,[3] there is an emerging need to develop automated, artificial intelligence (AI)-based tools to support pathologists. Management of afflicted men hinges on diagnosis and on the Gleason grade of their prostate adenocarcinoma for disease prognosis. Reliably diagnosing and correctly grading prostate adenocarcinoma in CNBs is

Research in context

Evidence before this study

We searched PubMed and Google Scholar on Feb 20, 2020, with the keywords "artificial intelligence" OR "deep learning" OR "machine learning" AND "pathology" AND "prostate cancer". The advanced search was limited to the English language. This search rendered 185 results, but only a few relevant studies retrieved employed image analysis for cancer detection or artificial intelligence (AI)-based Gleason grading performed on whole slide images. Additional relevant articles were reviewed from the publications' references. Despite much discussion in the field, there were few studies actually validating high performance of these algorithms, especially in large independent blinded studies. Most studies report performance on their internal test sets and only on few external validation sets. In the context of supporting clinical decisions, the need for an AI-based tool that combines high accuracy levels validated on large independent cohorts that incorporates clinically meaningful features is still unmet. Furthermore, the technical feasibility of deploying such a system in routine clinical workflow has not been addressed before. To the best of our knowledge, no report regarding clinical deployment of an AI system embedded into routine pathology practice has yet been published.

Added value of this study

We report the validation and performance of an AI-based prostate histopathology algorithm that extends beyond just cancer detection and grading, but also measures cancer proportion and the detection of perineural invasion to meet clinical reporting needs. Additionally, we describe the clinical deployment of such an algorithm in routine clinical practice, where implementation of a second read system showed early utility in preventing a missed prostate cancer diagnosis.

Implications of all the available evidence

An accurate, robust, and validated AI-based algorithm to detect, grade, and automatically impart clinically relevant diagnostic parameters regarding prostate adenocarcinoma offers an important tool for computer-assisted diagnosis in routine pathology practice. Demonstration of the technical feasibility for deploying such an AI-based system in the routine clinical workflow of a pathology laboratory elevates existing discussion of AI-based tools in pathology to a practical level, revealing how computational pathology can lead to improved efficiency, accuracy, consistent diagnoses, and better patient management.

challenging because of potential cancer mimics and the presence of only small foci of well differentiated adenocarcinoma. Hence, for an automated tool to be adopted into practice, it will need to have clinical grade accuracy.

Recent technological and regulatory advances in whole slide imaging have accelerated the adoption of digital imaging in pathology.[4] Now that pathology departments have several commercial digital pathology platforms available for diagnostic work, there is interest in leveraging AI tools. Several publications have shown the feasibility of developing AI-based algorithms to analyse histopathology images, specifically for prostate cancer.[4–8] However, there is minimal literature clinically validating such algorithms using large, independent blinded studies. Furthermore, routine clinical deployment of such deep learning systems in pathology laboratories has not been reported.

The aim of this study was twofold. The first aim was to do a blinded clinical validation of an AI-based algorithm that extended beyond just prostate adenocarcinoma detection and Gleason grading, by also detecting clinically meaningful features such as tumour extent and perineural invasion. The second aim was to deploy this system in a pathology laboratory for routine clinical use.

Methods

Study design

The study encompasses three key steps: (1) development and testing of an AI-based algorithm for prostate CNBs; (2) blinded algorithm validation in an external, independent dataset; and (3) algorithm deployment in routine clinical use. Each of these steps is described below.

Institutional review board approval was obtained for this study (University of Pittsburgh Institutional Review Board PRO18030439; Maccabi Ethics Helsinki Committee 0153-16-ASMC and 0081-18-BBL; and University of the Witwatersrand, Johannesburg Human Research Ethics Committee [Medical] M191003).

Algorithm development

Training and internal test datasets came from prostate CNBs retrieved from the archive of the Pathology Institute at Maccabi Healthcare Services' centralised laboratory (MegaLab) in Israel. H&E-stained slides were scanned using a Philips IntelliSite Scanner (Philips Digital Pathology Solutions; Best, Netherlands) at 40× magnification (resolution of 0·25 µm/pixel).

The algorithm that we developed, whose core technology is based on multilayered convolutional neural networks (CNNs) that were specifically designed for image classification tasks, analyses a whole slide image in three consecutive steps: tissue detection, classification, and slide-level analysis. Briefly, the first step uses a Gradient Boosting classifier, trained on thousands of image patches, to distinguish between tissue and background areas within the slide. After this, an ensemble of three CNN-based models is run on all tissue areas.

The models were trained on 1 357 480 labelled image patches that were extracted from manual annotations on 549 slides, selected from more than 65 000 slides in the

archive on the basis of various criteria, such as reported Gleason grade and rare findings. Annotations were done by three senior pathologists, each with 20–40 years of experience. The classification step yields predictions (probabilities) for every area in the slide belonging to each of 18 predefined classes, such as Gleason pattern 3 cancer, normal gland, and chronic inflammation (appendix p 6). Finally, the third step combines the 18 probability heatmaps (appendix p 6) and calculates slide-level scores for chosen endpoints.

Low concordance in Gleason scoring between pathologists[9] confounds the ability to obtain robust ground truth. Therefore, we focused on groupings of clinical significance rather than the full range of individual Gleason scores, choosing the following five endpoints: presence of cancer (primary endpoint), Gleason score 7–10 (including scores 3+4, 4+3, 4+4, 3+5, 5+3, 4+5, 5+4, and 5+5) versus Gleason score 6 (3+3) or atypical small acinar proliferation (ASAP), Gleason pattern 5 (including scores 3+5, 5+3, 4+5, 5+4, and 5+5), perineural invasion, and calculation of the percentage of cancer present in CNB material. The Gleason endpoints represent clinically relevant endpoints for disease management (Gleason score 7–10) and aggressive cancers (Gleason pattern 5).[9] More details on algorithm development are provided in the appendix (pp 2–3).

Algorithm testing

To evaluate the accuracy of the prostate algorithm, its output was compared with ground truth on a large collection of whole slide images independent from the training set. The internal test dataset included all 213 consecutive prostate CNBs (2576 H&E-stained slides) received at MegaLab from March 1 to June 30, 2016, which were also digitised with a Philips scanner. Associated immunohistochemistry slides from these cases were used to review cases and establish ground truth. 51 (2·0%) H&E-stained slides were not scanned due to limiting slide physical conditions (eg, broken glass or missing slides), and 24 (0·9%) H&E-stained slides that had been chosen previously for algorithm training as part of the randomly selected CNBs were filtered out. Thus, the internal test was run on 2501 H&E-stained slides from 210 cases. Ground truth was established at the slide level on the basis of the pathologists' diagnosis, either from the original pathology report (for benign cases) or after review by two senior pathologists with 40 years (JS) and 20 years (AAS) of experience for cases where there was a diagnosis of cancer in the report.

External validation

External validation of the algorithm was done at the University of Pittsburgh Medical Center (UPMC) using a validation dataset with different pre-imaging and scanning parameters. Glass slides at UPMC were scanned at 40× magnification (0·25 µm/pixel resolution) using an Aperio AT2 scanner (Leica Biosystems; Buffalo Grove, IL, USA). UPMC diagnoses are reported per part, where a part is one of the three biopsy regions (upper, mid, or base) in one of the bilateral prostate lobes. One part is often represented in multiple whole slide images.

To prepare the algorithm for external validation, we used a set of 32 prostate CNB cases (selected from cases occurring between August, 2014, and January, 2018), comprising 159 parts, to calibrate the algorithm for UPMC-specific whole slide image attributes (eg, scanner and staining) and to verify the technical validity of the whole slide images (eg, file format and resolution). This set included diagnoses with cancers of various Gleason scores (appendix p 5), high-grade prostatic intraepithelial neoplasia, inflammation, and atrophy. We divided the set into a calibration set (also known as a tuning set) and an internal test set. The calibration set comprised 44 parts, selected according to criteria used in previous training slides (appendix pp 2–3), that were manually annotated by senior pathologists (JS, AAS). The remaining 115 parts that were not annotated were used for internal validation.

See Online for appendix

A separate archival dataset of whole slide images that included 100 consecutive cases of prostate CNB cases previously received and signed out (ie, formally diagnosed and reported) by the Genitourinary Center of Excellence at UPMC served as an independent, external, blinded validation set. Each case included all associated H&E-stained slides, giving a total of 1627 H&E-stained whole slide images. The 100 cases were organised into 379 parts with an average of four slides per part (typically deeper H&E-stained section levels of the same cores). All study data were anonymised and metadata for each case (eg, patient demographics and immunohistochemistry results) were recorded. Patients' age distribution, percentage of cancer cases, and Gleason score distribution are summarised in the appendix (p 7).

The algorithm was applied to the external, blinded validation set. Data were subsequently unblinded and the results of the algorithm compared with the ground truth diagnoses. Assessment of performance was done on a per-part level. The performance on detection of the five endpoints was assessed.

Ground truth

Ground truth was established on the basis of the original UPMC pathology report using the diagnosis rendered by one of three genitourinary subspecialty pathologists before this study. To address the potential for discordance among pathologists, and to ensure accurate assessment of algorithmic results, a subset of the parts in the external validation set were chosen for ground truth ascertainment. Ground truth ascertainment was done on four of the features assessed in the external validation set (ie, cancer detection, Gleason score 7–10, Gleason pattern 5, and perineural invasion); cancer percentage was not subject to discrepancy analysis owing to the protocol-specific nature of discrepancies. The subset of parts selected for ground truth ascertainment was based

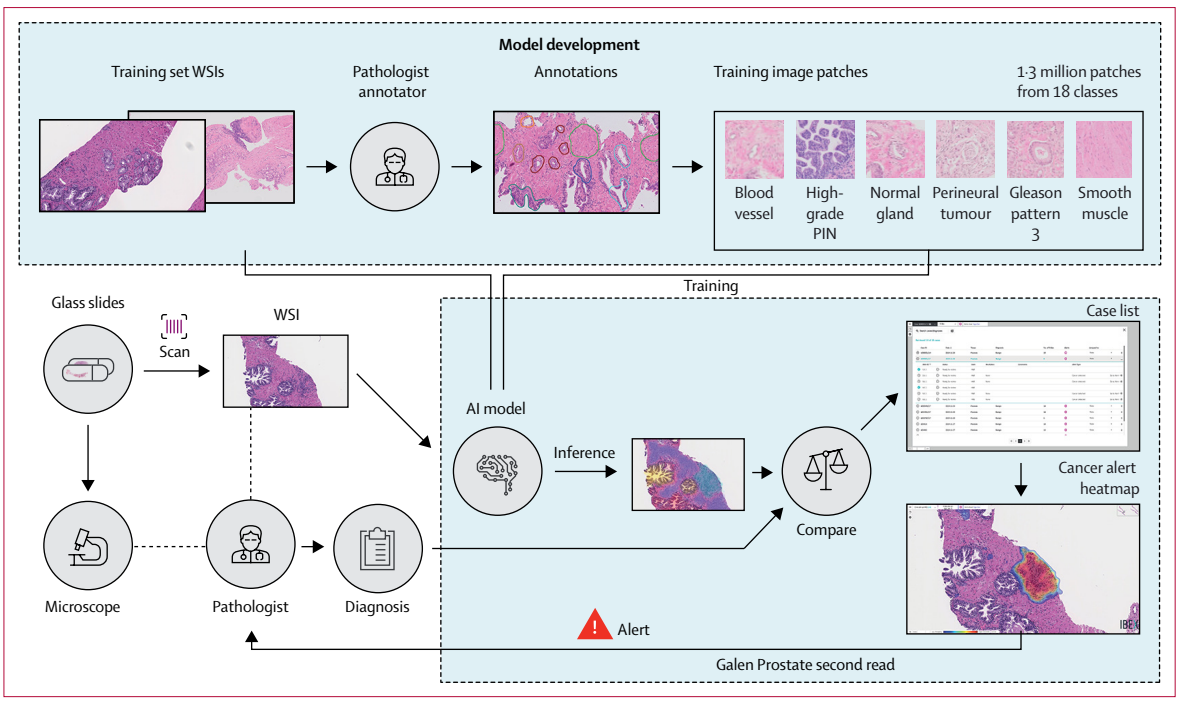

Figure 1: **Overview of the algorithm and clinical deployment of the Galen Prostate second read system**
AI=artificial intelligence. WSI=whole image slide. PIN=prostatic intraepithelial neoplasia.

on likelihood of affecting the performance metrics of the algorithm. Thus, 10% of the highest-scoring negative parts (eg, parts reported as benign in the pathology report) and 5% of the lowest-scoring positive parts (eg, parts reported as positive for cancer in the pathology report) in the external validation set were sent for review by two independent pathologists with genitourinary pathology expertise who did not previously diagnose any of the cases, one from UPMC (GMQ-G), and one from Maccabi (JS). Their review was done in a blinded manner using the same digital slides and image viewer to ensure uniformity between reviewers. Ground truth was based on consensus between the two pathologists. For parts where there was an initial discrepancy between the two pathologists, a third pathologist (RD) with genitourinary pathology expertise blindly reviewed the slides and the majority vote was used as the final consensus diagnosis. 65 parts were reviewed (appendix p 7).

Deployment for routine clinical use: second read application

Maccabi Healthcare Services is a large health-care provider in Israel with a centralised pathology institute with approximately 120 000 surgical pathology cases annually, including around 700 prostate CNB cases reviewed by three genitourinary subspecialty pathologists out of a team of 12 senior pathologists. Around 40% of these CNBs are diagnosed with cancer. Galen Prostate (Ibex Medical Analytics), the product based on the prostate algorithm, has been implemented in Maccabi Pathology Institute since March, 2018, as a second read system—namely, a quality control application that reviews whole slide images of all prostate CNBs (figure 1). Slides were scanned using a Philips IntelliSite Scanner and processed via a server with four graphics processing units. With this setup, the algorithm was run in parallel to the pathologists' routine workflow. Triggers were used to alert pathologists in the event there was a case with a discrepancy between their diagnosis and the output of the algorithm. The system generated alerts for slides with a high algorithmic cancer score that were diagnosed as benign by the pathologist, and for slides with a high Gleason score (ie, 7–10) that were diagnosed as Gleason score 6 by the pathologist. The alert threshold was set to correspond to a specificity of 90% (see appendix pp 3–4 for threshold rationale). Pathologists were able to view the case list with alerts and do a second review, specifically focused on the region that triggered the alert outlined by heatmaps in the second read slide viewer. The system took 1 month to set up, including automatic export of slides from the Philips Image Management System to Galen Prostate. A training session was held for each user and for the technician scanning the slides.

Statistical analysis

Sample size for 80% power with two-sided 5% level of significance was calculated on the basis of test performance (appendix p 4). Area under the receiver operating characteristic curve (AUC) was calculated with 95% Wald CIs using the continuous score. Specificity

	Number of slides (internal test) and parts (external validation)	AUC (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	PPV	NPV
Internal test (Maccabi Healthcare Services)						
Benign *vs* cancer	2501: 1957 benign and 490 cancer*	0·997 (0·995–0·998)	90·14% (87·76–92·09)†	99·59% (98·39–99·90)	71·7%	99·9%
External validation (UPMC)						
Benign *vs* cancer	355: 225 benign and 130 cancer‡	0·991 (0·979–1·00)	97·33% (94·43–98·74)	98·46% (94·06–99·61)	95·5%	99·1%
Gleason score 6 or ASAP *vs* Gleason score 7–10§	151: 73 Gleason score 6 or ASAP and 78 Gleason score 7–10	0·941 (0·905–0·977)	90·41% (78·92–95·96)	85·9% (75·72–92·25)	90·5%	85·7%
ASAP or Gleason pattern 3 or 4 *vs* Gleason pattern 5¶	151: 131 ASAP or Gleason pattern 3 or 4 and 20 Gleason pattern 5	0·971 (0·943–0·998)	90·84% (84·18–94·87)	85% (51·24–96·83)	58·6%	97·5%
Cancer without *vs* with perineural invasion	154: 108 without perineural invasion and 46 with perineural invasion‖	0·957 (0·930–0·985)	90·74% (83·10–95·13)	86·96% (74·47–93·84%)	80%	94·2%

AUC=area under the receiver operating characteristic curve. PPV=positive predictive value. NPV=negative predictive value. UPMC=University of Pittsburg Medical Center. ASAP=atypical small acinar proliferation. *34 slides with ASAP and 20 slides for which pathologists did not reach a conclusion were excluded. †Selected to reflect pathologists' review using 10% of slides for quality control. ‡Parts that were diagnosed as ASAP by the pathologists were excluded. §Gleason score 7–10 includes Gleason scores 3+4, 4+3, 4+4, 3+5, 5+3, 4+5, 5+4, and 5+5. ¶ASAP or Gleason pattern 3–4 refers to parts not having Gleason pattern 5—ie, diagnosed as ASAP or Gleason score 3+3, 3+4, 4+3, or 4+4. Gleason pattern 5 includes Gleason scores 3+5, 5+3, 4+5, 5+4, and 5+5. ‖154 parts include: 151 parts with adenocarcinoma and Gleason score, two parts with other cancer, and one part where a consensus was not reached about the Gleason score during ground truth ascertainment.

Table 1: **Algorithm performance**

and sensitivity at multiple cutoff points were calculated for each feature and are presented with two-sided 95% CIs. The CIs were calculated from a generalised estimating equation model using the GENMOD procedure in SAS (dist=bin) using the ilink option in the Lsmeans statement. This was done to accommodate the within-subject correlation due to repeated measurements. Pearson's correlation coefficient is presented for cancer percentage concordance between the algorithm and pathology report. The mean bias (difference between cancer percentage) and its SD are presented as further measures of concordance, as well as the Bland-Altman 95% limits of agreement (presented with 95% CIs). We used SAS version 9.4 for all analyses.

Role of the funding source

The funder of the study supported study design and writing of the report. LP had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Patient characteristics in the study cohorts are shown in the appendix (pp 7, 11) and reflect the representative population undergoing prostate CNBs in Israel and the USA (appendix p 3).

The AUC for cancer detection was 0·997 (95% CI 0·995–0·998) on the internal test set (appendix p 7) and 0·991 (0·979–1·00) on the UPMC external validation set (appendix pp 8, 11; table 1). Additional performance metrics in internal test and external validation sets are shown in table 1.

Ground truth ascertainment for cancer versus benign diagnosis resulted in algorithm-driven diagnostic corrections for seven parts in the external validation dataset: four parts were corrected from benign to cancer (two parts) or ASAP (two parts; appendix p 13) and three parts were corrected from cancer to benign (one part) or ASAP (two parts). At the case level, one case that was originally diagnosed as ASAP was changed to cancer (figure 2B).

Cancer percentage correlation was computed on 126 parts diagnosed as cancer in UPMC pathology reports. Figure 2A shows an example of the algorithm results. There was high correlation between the cancer percentage reported by pathologists with the percentage computed by the algorithm (Pearson's r=0·882, 95% CI 0·834 to 0·915; p<0·0001; appendix p 12). The mean bias was –4·14% (95% CI –6·36 to –1·91), showing that, on average, the algorithm underestimates the cancer percentage by 1·9–6·3% (appendix p 8). The Bland-Altman 95% limits of agreement, which show the difference between reported and calculated percentage, are –28·6% (95% CI –33·9 to –23·3; algorithm lower) to 20·3% (15·0 to 25·7; algorithm higher). Examining discrepancies between the algorithm and the pathologist in cancer percentage reveals that these variances stem largely from specific calculation protocols, rather than algorithmic inaccuracies—eg, the decision to include small benign areas between cancer foci in the calculation (appendix p 14).

When assessing the algorithm's performance on Gleason grading, the AUC for distinguishing between ASAP or Gleason score 6 versus higher Gleason scores was 0·941 (95% CI 0·905–0·977) and 0·971 (0·943–0·998) for detecting any Gleason pattern 5 in a CNB (table 1). Additional performance metrics on graing performance are summarised in table 1 and in the appendix (p 9).

Ground truth ascertainment resulted in algorithm-driven corrections of four parts from Gleason score 6 to Gleason score 7–10 (figure 2C), and two parts from Gleason

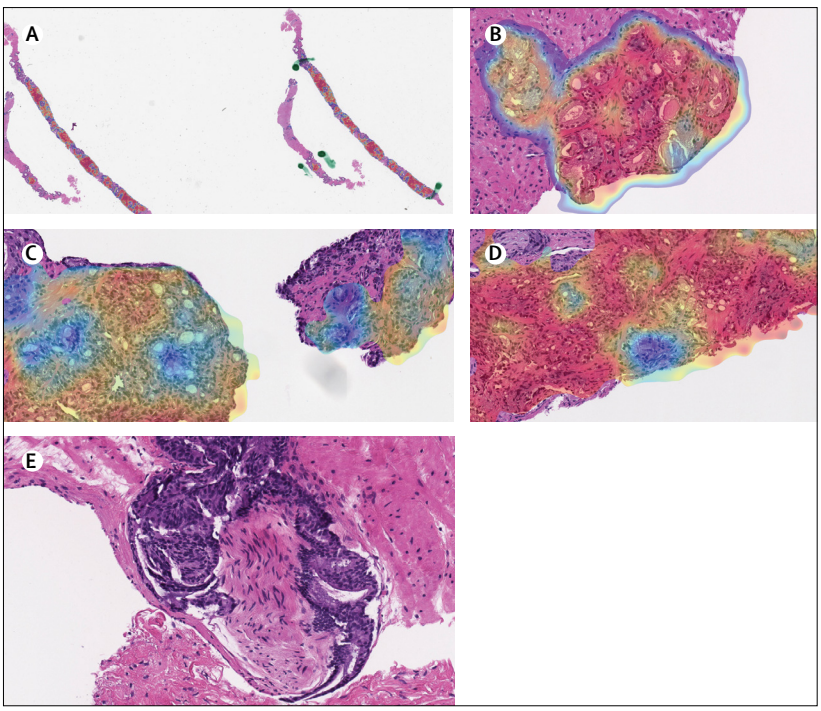

***Figure 2*: Examples of diagnoses after review**
All images are stained with haematoxylin and eosin and displayed at 0·6× (16·67 µm/pixel; panel A) and 20× (0·50 µm/pixel; panels B–E) magnification. (A) Example of cancer proportion reported by pathologists versus calculations performed by the algorithm in the external validation dataset. Prostate CNBs (part 662_6) with the cancer heatmap (where blue shows low probability and red shows high probability) displayed. Tumour proportion calculated by the algorithm was 33% compared with 40% reported by a pathologist. Panels B–E provide examples of revised diagnoses. (B) Prostate CNB (part 665_1) with cancer heatmap (where blue shows low probability and red shows high probability) originally diagnosed as benign that was subsequently changed to cancer with Gleason score 6=3+3 after review. The case diagnosis was updated from ASAP to cancer. (C) Prostate biopsy (part 598_5) with a Gleason pattern heatmap (where blue shows Gleason pattern 3, yellow shows Gleason pattern 4, and red shows Gleason pattern 5) that was updated from Gleason score 6=3+3 to Gleason score 7 (there was no concordance between reviewers as to 3+4 *vs* 4+3). (D) Prostate biopsy (part 606_1) with a Gleason pattern heatmap (where blue shows Gleason pattern 3, yellow shows Gleason pattern 4, and red shows Gleason pattern 5). This biopsy part and case were updated from Gleason score 7=3+4 to Gleason score 8=3+5. (E) Example of a prostate CNB (part 686_4) that was updated to positive for perineural invasion after review. The colours represent the original staining. CNB=core needle biopsy. ASAP=atypical small acinar proliferation.

	False positive	False negative
Adenocarcinoma	1 benign, 2 ASAP	2 cancer, 2 ASAP
Gleason score 7–10	2	4
Gleason pattern 5	0	1
Perineural invasion	1	2

Data are number of parts. ASAP=atypical small acinar proliferation.

***Table 2*: Pathologists' misdiagnoses identified by the algorithm**

score 7–10 to Gleason score 6. One part was corrected from Gleason score 7 to Gleason grading with pattern 5 present (figure 2D).

The algorithm detected perineural invasion with an AUC of 0·957 (95% CI 0·930–0·985), further summarised in table 1 and in the appendix (p 9).

Ground truth ascertainment resulted in algorithm-driven corrections for two parts from negative to positive for perineural invasion (figure 2E). For one part that was positive for perineural invasion in the report, it was determined that perineural invasion was absent.

Discrepancies between the two reviewing pathologists for 65 parts that were sent for ground truth ascertainment were observed for all endpoints analysed, with major discrepancies in Gleason grading (nine [14%] parts for Gleason grade group difference >1 and a total of 24 [37%] parts with differences in Gleason grading) and perineural invasion diagnosis (nine [14%] parts), followed by distinguishing between benign tissue and cancer (six [9%] parts) and between ASAP and cancer (two [3%] parts). Misdiagnoses by pathologists occurred for 17 parts and six cases, including one case in which cancer was misdiagnosed as ASAP and five cases where the revised diagnosis changed the Gleason score, with potential clinical significance (table 2).

Between March, 2018, and November, 2019, 941 cases (11 429 H&E-stained slides) were processed by the second read system at the Maccabi Pathology Institute. Cancer alerts were raised for an average of 10·9% of the slides belonging to cases diagnosed by pathologists as benign, with 1·35 cancer alerts (slides) per such case. Of 1451 slides from 115 cases diagnosed as Gleason score 6=3+3, the system issued 90 (6·2%) Gleason score 7–10 alerts and 560 (38·6%) cancer alerts. Upon pathologist review of the cancer alerts, 509 (90·9%) alerts required no substantial effort: most areas that triggered the alerts were identified as either atrophic glands or crushed glands, consistent with an irregular feature that mimics malignant glands. 51 (9·1%) alerts led to additional cuts or stains being ordered, two (4%) of which led to a third opinion request. Alerts were focused on specific areas, meaning that review time was minimal, resulting overall in approximately 1% of the pathologist's time.

We report on the first case with missed cancer in a CNB detected by the system, which occurred immediately following deployment. The biopsy from a 55-year-old patient was diagnosed by the pathologist as benign with foci of acute and chronic inflammation and atrophic changes in both prostate gland lobes. The second read system raised alerts for the presence of cancer in three slides within the case with a high probability score for cancer (>0·99; figure 3). Subsequent re-examination by the pathologist, additional recut sections, and an immunohistochemistry stain for CK903 were done, and the diagnosis was revised to minute foci of adenocarcinoma of acinar type with Gleason score 6 in the right lobe core biopsy. Three cores were reported as involved, and the estimated percentage of tumour in prostatic tissue was less than 5%. The left core biopsy diagnosis remained unchanged (ie, negative). As a result, this patient was included in an active surveillance protocol and his prostate-specific antigen level taken 3 months after this biopsy was above normal at 5·56 ng/mL and continued to rise during the surveillance to 7·93 ng/mL. Additional missed cancers were identified during the clinical deployment, details of which were not provided by the laboratory.

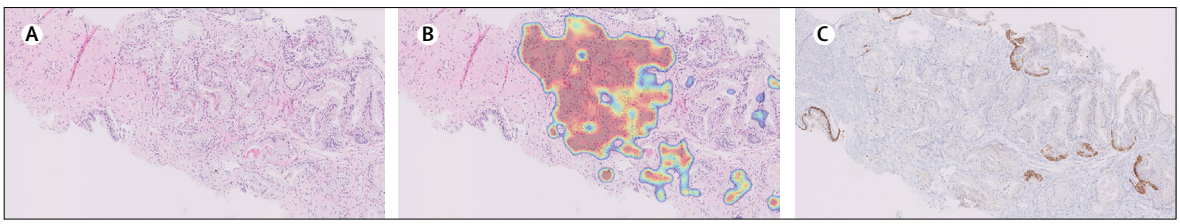

Figure 3: **Missed cancer case originally diagnosed as benign**
Images in panels A and B are stained with haematoxylin and eosin. All images are displayed at 10× (1 µm/pixel) magnification. (A) Prostate biopsy showing abnormal glandular focus. (B) Algorithm heatmap detecting cancer with high probability (red areas) in the same small focus. (C) Corresponding area on the immunohistochemistry slide (CK903) with absent basal cells confirming the diagnosis of adenocarcinoma.

Discussion

We report the development of a medical-grade AI-based algorithm for evaluating digitised prostate CNB slides, and successful deployment of this AI tool in routine clinical practice. We show high accuracy of the algorithm, based on a large, blinded, external validation dataset, to identify and quantify prostate cancer, differentiate between low-grade and high-grade tumours, and detect perineural invasion in CNBs. To the best of our knowledge, this is the first report of an AI-based algorithm that extends beyond cancer detection and grading of prostatic cancer in histopathological images, and one of the first instances of clinical use of an AI-based algorithm in routine pathology practice. We are aware of one other study where implementation of an AI-based algorithm for the automatic detection of prostate cancer in a pathology laboratory was reported,[10] but no outcomes were published.

AI algorithms and models are typically developed from data that are assumed to be representative. Overfitting is a common problem, where despite good performance on training data or on test data similar to the training data, performance deteriorates on novel data. Multiple strategies exist to avoid overfitting (eg, cross-validation or bootstrapping), but true performance can only be determined through blinded studies with external datasets. The demonstration herein of high performance of the algorithm using an external, geographically distinct, blinded dataset is crucial for true evaluation of the performance and utility of the algorithm. To our knowledge, very few studies have attempted such validation, with narrower applications and significantly lower performance.

Deployment of such an AI tool in clinical practice is timely, not only because prostate adenocarcinoma is one of the most common cancers seen in men, but also due to the substantial increase in pathologists' workload as cancer cases rise, combined with increased complexity of histopathological assessment with changes in guideline recommendations. Unfortunately, there is a concomitant decline in the pathology workforce. When adjusted by new cancer cases per year, the workload per US pathologist has risen by 41·73%[3] and this gap is growing, potentially resulting in delayed cancer diagnoses and diagnostic errors. Reports of missed prostate cancer and a lack of concordance with Gleason grading have been well documented,[11] especially when these diagnoses are rendered by general pathologists instead of subspecialist urological pathologists.[12–15] Raciti and colleagues[16] showed that in a non-clinical setting, AI can increase the detection of prostate cancer in whole slide images of CNBs. In our study, the two pathologists who participated disagreed on cancer detection in 9% of biopsy parts and on Gleason grade grouping in 37% of biopsy parts (14% major disagreement). Overall, 17 misdiagnosed parts and six misdiagnosed cases in the UPMC dataset were identified in this study, including misdiagnoses in cancer detection and grading, as well as detection of perineural invasion, some of which might have affected treatment. In a pathology laboratory in France, the same second read system identified 12 misdiagnosed cases, including missed high-grade cancers.[17] Several pathology practices have established a second review process for quality control purposes on a portion (eg, 10%) of cases to thwart misdiagnoses.[18] Although useful, this additional quality control step, when done manually, further adds to a pathologist's workload, and therefore is typically practised rarely. AI tools, as shown in this study, can be leveraged to help to automate this safety net task.[19,20] Indeed, the deployment described here of Galen Prostate as a second read application is the first instance of 100% quality control of prostate CNBs in a laboratory, drastically diminishing the chances for misdiagnoses with negligible impact on a pathologist's workload.

Previous publications of prostate cancer algorithms devoted to analysing histopathology images have reported lower or similar performance characteristics for cancer detection to those obtained herein (table 3).[6,8,21–25] However, many previous publications on this topic report on algorithms limited to just one task (ie, narrow AI) and provide performance metrics verified mostly on internal test sets. For example, several researchers have explored the use of machine and deep learning techniques to provide just Gleason grading.[23,24,26,27] By contrast, our study reports high performance characteristics of a multifunction algorithm for prostate CNB interpretation (ie, cancer detection, grading, evaluation of tumour extent, and perineural invasion), assessed on an independent external dataset. The algorithm in our study was able to simultaneously evaluate CNBs for

	Dataset dimension	Cancer *vs* benign accuracy	Gleason grading accuracy
Performance assessed on external validation datasets			
Campanella et al (2019)[6]	12 727 slides	AUC 0·932	Not available
Bulten et al (2020)[8]	245 tissue microarray cores*	AUC 0·98	AUC 0·87 for benign or Gleason grade group 1 *vs* Gleason grade group 2–5 AUC 0·86 for benign or Gleason grade group 1–2 *vs* Gleason grade group 3–5
Nir et al (2019)[7]	230 slides from 56 prostatectomy cases	AUC 0·75	Not available
Ström et al (2020)[21]	330 cores from 73 cases	AUC 0·986	Cohen's κ 0·7
Current study	1627 slides from 100 cases	AUC 0·991 overall AUC 0·957 for perineural invasion†	AUC 0·941 for ASAP or Gleason score 6 *vs* Gleason score 7–10 AUC 0·971 for ASAP or Gleason pattern 3–4 *vs* Gleason pattern 5
Performance assessed only on internal test sets			
Arvaniti et al (2018)[22]	245 tissue microarray cores*	Not available	Cohen's κ 0·55
Nagpal et al (2019)[23]	1226 slides	Not available	AUC 0·7
Nir et al (2018)[24]	86 tissue microarray cores from 60 cases	AUC 0·85; sensitivity 91·3%; specificity 84%	Accuracy 79%; sensitivity 75·9%; specificity 77·9%
Lucas et al (2019)[25]	96 sections from 38 cases	Accuracy 92%; sensitivity 90%; specificity 93%	Accuracy 90%; sensitivity 77%; specificity 94%

Details on Gleason patterns, scores, and grading groups can be found in Rice-Stitt et al.[9] AUC=area under the receiver operating characteristic curve. ASAP=atypical small acinar proliferation. *The same tissue microarray was used in these studies. †The current study is the only study to assess perineural invasion.

Table 3: **Performance of algorithms in detection and grading of prostate cancer**

perineural invasion, a feature not reported by other researchers but one that is well known to have clinical and prognostic significance.[28] Notably, perineural invasion does not lend itself to some of the published AI-based algorithms, as it is typically very small (unlike adenocarcinoma, which sometimes covers a large fraction of the tissue core) and relatively uncommon. Furthermore, cancer percentage in CNBs calculated by our algorithm highly correlated with the proportion that was estimated by pathologists and can be further tailored to mimic the protocol for calculating cancer percentage for each laboratory. Ström and colleagues[21] reported slightly lower correlation between their algorithm and pathologists when calculating tumour length in prostate biopsies compared with our algorithm for cancer percentage. This is not surprising, as the quantitative determination of prostate cancer (eg, tumour volume) in CNBs is controversial and fraught with subjectivity between pathologists.[29] Algorithms that can take these measurements offer laboratories an opportunity to harmonise clinical practice among pathologists.

Our study has several limitations. Gleason grading for prostate cancer is a well established grading system globally adopted for the treatment and prognosis of prostate cancer, and is considered the gold standard in the field. Still, discordance between pathologists in Gleason grading is widely acknowledged,[11] and thus a true and agreed-upon ground truth is sometimes difficult to reach. The AI-based algorithm employed in this study was trained using annotations from multiple independent pathologists and thus the computed Gleason score of the deep-learning algorithm is expected to be generalisable, consistent with the high performance shown here. However, true and objective performance metrics can ultimately only be determined using a combination of a committee of expert pathologists for grading and long-term outcome data, which would also allow for individual grade assessment rather than the groupings used here. Ongoing studies are assessing these questions and are also focused on assessing performance characteristics of additional features computed by the algorithm, including areas corresponding to certain Gleason scores or percentage of certain Gleason scores within cancerous tissue, detection of high-grade prostatic intraepithelial neoplasia, and inflammation. These, as well as the performance of the algorithm in detection of rare variants of adenocarcinoma, are important features of the algorithm that were beyond the scope of the current study. Still, the utility of the study for wider deployment is evident, as we report here on six patient-level misdiagnoses at UPMC, including clinically significant misdiagnoses such as involvement of higher Gleason patterns.

Implementation of AI-based tools for routine clinical work carries practical considerations of generalisability, infrastructure requirements, and throughput. Although algorithms had been shown to have low performance for whole slide images of different file formats acquired by different scanners,[30] we were able to show equally high performance when slides were scanned using a Philips (clinical deployment dataset) and Leica (external validation dataset) scanner, although more work, such as

establishing a standardised file format—eg, Digital Imaging and Communications in Medicine—is warranted. We also show here that the algorithm calibration to a new laboratory requires a set of only around 30 cases. In fact, future versions of the algorithm incorporating additional training sets are underway, including a total of four different scanners and ten different laboratories, ultimately eliminating the need for calibration altogether. Finally, our approach to annotations enabled us to reach high accuracy with only three CNNs, as opposed to other published studies using a large number of CNNs, rendering their deployment impractical in terms of hardware and run-time requirements. From a throughput perspective, the system deployed in Maccabi analyses many slides per hour, keeping up with the rate at which slides are scanned.

In summary, we report the development, external clinical validation, and deployment in routine practice of an AI-based algorithm to detect, grade, and evaluate additional clinically relevant tumour features in digitised slides of prostate CNBs. These data suggest that this AI-based algorithm could be used as a tool to automate screening of prostate CNBs for primary diagnosis, assess signed-out cases for quality control purposes, and standardise reporting to improve patient management. Studies reporting on deployment in additional laboratories and associated clinical utility are underway.

Contributors

LP conceived the study. DL, CL, PM, and SH contributed to study design. LP, Dl, CL, and RD contributed to the methodology. LP contributed to the project administration. LP and RD contributed to study coordination. LP provided informatics support. LP, LB, DL, and MV contributed to the literature search. LP, LB, RH, CL, JS, VS, and RD contributed to data collection. LP and RD contributed to data curation. LP, GMQ-G, LB, RH, and CL contributed to data analysis. LB, RH, DL, CL, JS, and MV contributed to data interpretation. LB, CL, and MV contributed to the figure design. GMQ-G, JS, AAS, VS, and RD contributed to expert review. LP, GMQ-G, LB, DL, CL, MV, PM, SH, and RD contributed to manuscript preparation and review.

Declaration of interests

LP and VS serve on the medical advisory board of Ibex and report personal fees from Ibex, outside of the submitted work. LB and CL are authors on pending patents US 62/743,559 and US 62/981,925 (including System & Methods for Personalization and Optimization of Digital Pathology Analysis and System and Method of Managing Workflow of Examination of Pathology Slides). AAS reports personal fees for expert consultancy from Ibex, outside of the submitted work. All other authors declare no competing interests.

Data sharing

The software used in the current study is Ibex proprietary IP and cannot be shared. The data collected during this study is patient data obtained under Ethical Committees approval and cannot be shared. De-identified raw whole slide imaging data described in this study are available for viewing online and for download upon request.

Acknowledgments

This study was funded by Ibex Medical Analytics.

References

1 Rawla P. Epidemiology of prostate cancer. *World J Oncol* 2019; **10:** 63–89.
2 Matoso A, Epstein JI. Defining clinically significant prostate cancer on the basis of pathological findings. *Histopathology* 2019; **74:** 135–45.
3 Metter DM, Colgan TJ, Leung ST, Timmons CF, Park JY. Trends in the US and Canadian pathologist workforces from 2007 to 2017. *JAMA Netw Open* 2019; **2:** e194337.
4 Evans AJ, Bauer TW, Bui MM, et al. US Food and Drug Administration approval of whole slide imaging for primary diagnosis: a key milestone is reached and new questions are raised. *Arch Pathol Lab Med* 2018; **142:** 1383–87.
5 Hipp J, Monaco J, Kunju LP, et al. Integration of architectural and cytologic driven image algorithms for prostate adenocarcinoma identification. *Anal Cell Pathol (Amst)* 2012; **35:** 251–65.
6 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.
7 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.
8 Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020; **21:** 233–41.
9 Rice-Stitt T, Valencia-Guerrero A, Cornejo KM, Wu CL. Updates in histologic grading of urologic neoplasms. *Arch Pathol Lab Med* 2020; **144:** 335–43.
10 Fraggetta F. Clinical-grade computational pathology: alea iacta est. *J Pathol Inform* 2019; **10:** 38.
11 Yang C, Humphrey PA. False-negative histopathologic diagnosis of prostatic adenocarcinoma. *Arch Pathol Lab Med* 2020; **144:** 326–34.
12 Steinberg DM, Sauvageot J, Piantadosi S, Epstein JI. Correlation of prostate needle biopsy and radical prostatectomy Gleason grade in academic and community settings. *Am J Surg Pathol* 1997; **21:** 566–76.
13 Kronz JD, Milord R, Wilentz R, Weir EG, Schreiner SR, Epstein JI. Lesions missed on prostate biopsies in cases sent in for consultation. *Prostate* 2003; **54:** 310–14.
14 Truesdale MD, Cheetham PJ, Turk AT, et al. Gleason score concordance on biopsy-confirmed prostate cancer: is pathological re-evaluation necessary prior to radical prostatectomy? *BJU Int* 2011; **107:** 749–54.
15 Goodman M, Ward KC, Osunkoya AO, et al. Frequency and determinants of disagreement and error in Gleason scores: a population-based study of prostate cancer. *Prostate* 2012; **72:** 1389–98.
16 Raciti P, Sue J, Ceballos R, et al. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Mod Pathol* 2020; published online May 11. https://doi.org/10.1038/s41379-020-0551-y.
17 Laifenfeld D, Sandbank J, Linhart C, et al. Performance of an AI-based cancer diagnosis system in France's largest network of pathology institutes. 31st European Congress of Pathology; Sept 7–11, 2019; Nice, France (abstr 043).
18 Owens SR, Wiehagen LT, Kelly SM, et al. Initial experience with a novel pre-sign-out quality assurance tool for review of random surgical pathology diagnoses in a subspecialty-based university practice. *Am J Surg Pathol* 2010; **34:** 1319–23.
19 Parwani AV. Automated diagnosis and Gleason grading of prostate cancer—are artificial intelligence systems ready for prime time? *J Pathol Inform* 2019; **10:** 41.
20 Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. *Nat Rev Urol* 2019; **16:** 391–403.
21 Ström P, Kartasalo K, Olsson H, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol* 2020; **21:** 222–32.
22 Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep* 2018; **8:** 12054.
23 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.
24 Nir G, Hor S, Karimi D, et al. Automatic grading of prostate cancer in digitized histopathology images: learning from multiple experts. *Med Image Anal* 2018; **50:** 167–80.
25 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.

For **de-identified raw whole slide imaging data** see http://image.upmc.edu:8080/Ibex/view.apml

26 Sparks R, Madabhushi A. Statistical Shape Model for Manifold Regularization: Gleason grading of prostate histology. *Comput Vis Image Underst* 2013; **117:** 1138–46.

27 Hossain A, Arimura H, Kinoshita F, et al. Automated approach for estimation of grade groups for prostate cancer based on histological image feature analysis. *Prostate* 2020; **80:** 291–302.

28 Harnden P, Shelley MD, Clements H, et al. The prognostic significance of perineural invasion in prostatic cancer biopsies: a systematic review. *Cancer* 2007; **109:** 13–24.

29 Paner GP, Gandhi J, Choy B, Amin MB. Essential updates in grading, morphotyping, reporting, and staging of prostate carcinoma for general surgical pathologists. *Arch Pathol Lab Med* 2019; **143:** 550–64.

30 Leo P, Lee G, Shih NN, Elliott R, Feldman MD, Madabhushi A. Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images. *J Med Imaging (Bellingham)* 2016; **3:** 047502.