Modern Pathology (2020) 33:2058–2066
https://doi.org/10.1038/s41379-020-0551-y

ARTICLE

Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies

Patricia Raciti[1] · Jillian Sue[1] · Rodrigo Ceballos[1] · Ran Godrich[1] · Jeremy D. Kunz[1] · Supriya Kapur[1] · Victor Reuter[2] · Leo Grady[1] · Christopher Kanan[1] · David S. Klimstra[2] · Thomas J. Fuchs[1,2]

Received: 19 December 2019 / Revised: 3 April 2020 / Accepted: 5 April 2020 / Published online: 11 May 2020

Abstract

Prostate cancer (PrCa) is the second most common cancer among men in the United States. The gold standard for detecting PrCa is the examination of prostate needle core biopsies. Diagnosis can be challenging, especially for small, well-differentiated cancers. Recently, machine learning algorithms have been developed for detecting PrCa in whole slide images (WSIs) with high test accuracy. However, the impact of these artificial intelligence systems on pathologic diagnosis is not known. To address this, we investigated how pathologists interact with Paige Prostate Alpha, a state-of-the-art PrCa detection system, in WSIs of prostate needle core biopsies stained with hematoxylin and eosin. Three AP-board certified pathologists assessed 304 anonymized prostate needle core biopsy WSIs in 8 hours. The pathologists classified each WSI as benign or cancerous. After ~4 weeks, pathologists were tasked with re-reviewing each WSI with the aid of Paige Prostate Alpha. For each WSI, Paige Prostate Alpha was used to perform cancer detection and, for WSIs where cancer was detected, the system marked the area where cancer was detected with the highest probability. The original diagnosis for each slide was rendered by genitourinary pathologists and incorporated any ancillary studies requested during the original diagnostic assessment. Against this ground truth, the pathologists and Paige Prostate Alpha were measured. Without Paige Prostate Alpha, pathologists had an average sensitivity of 74% and an average specificity of 97%. With Paige Prostate Alpha, the average sensitivity for pathologists significantly increased to 90% with no statistically significant change in specificity. With Paige Prostate Alpha, pathologists more often correctly classified smaller, lower grade tumors, and spent less time analyzing each WSI. Future studies will investigate if similar benefit is yielded when such a system is used to detect other forms of cancer in a setting that more closely emulates real practice.

Introduction

Prostate cancer (PrCa) is the second most common cancer among men in the United States and, globally, the fifth leading cause of cancer death among males [1]. The late 1980s–1990s saw a dramatic increase in PrCa detection because of the use of prostate specific antigen (PSA) testing, especially in the United States [1]. Many of these previously undetected cancers were of limited clinical stage, leading to the creation of a new clinical-stage classification, T1c, in which PrCa is diagnosed despite a normal digital rectal exam [2]. To counteract "overdiagnosis" and potential overtreatment of biologically indolent/low-risk forms of PrCa that resulted from increased PSA screening as well as saturation biopsy sampling, treatment strategies have evolved [3]. Active surveillance has become the most common management approach for men with localized low-risk PrCa, rather than primary curative therapy (i.e., radical prostatectomy or radiation) [4]. The gold standard diagnosis of PrCa is prostate needle core biopsy, and various parameters obtained from evaluation of these biopsies constitute important inclusion and exclusion criteria for active surveillance in many centers. While the Gleason score is an important parameter in this determination, the number of

These authors contributed equally: Patricia Raciti, Jillian Sue

Supplementary information The online version of this article (https://doi.org/10.1038/s41379-020-0551-y) contains supplementary material, which is available to authorized users.

✉ Patricia Raciti
patricia.raciti@paige.ai

[1] Paige.AI, 11 East Loop Road, FL5, New York, NY 10044, USA

[2] Department of Pathology, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA

cores that harbor PrCa and the percentage of involvement by PrCa, in particular, are critical [4–6]. A recent survey of oncologists and surgeons found that 94% use the number of positive cores to assess tumor extent [7]. Thus, the recognition of small, well-differentiated foci of PrCa is crucial to triaging the patient for appropriate treatment.

However, diagnosis of PrCa in core needle biopsies can be challenging, especially when only small, well-differentiated foci are present. While immunohistochemical stains (IHC) can be used to investigate suspicious foci and can increase PrCa detection, there is no justification for additional investigation if suspicious foci were not detected by the pathologist first [5, 6]. In order to ensure that all suspicious foci are detected by the pathologist, one effective solution is blinded re-review of slides, which has been shown to greatly improve cancer detection and accuracy [8, 9]. However, universal second review is resource- and time-intensive because it requires duplicative efforts by two pathologists. In addition, there is a possibility that the second reviewing pathologist might also fail to detect cancer [8]. Thus, few institutions have incorporated second review into clinical workflow [10]. If utilized, second reads are used to confirm the presence of cancer in cases already recognized as malignant; cases diagnosed as benign generally are not subjected to second reads [11]. Blinded re-review of all prostate needle core biopsies is particularly challenging to implement because the number of biopsies performed worldwide is increasing due to an aging population, improved access to screening, and greater adoption of saturation biopsies, while the number of pathologists is decreasing [12].

Recently, artificial intelligence (AI) systems have been shown to be capable of accurately detecting PrCa from digital whole slide images (WSIs) of core needle biopsies stained with hematoxylin and eosin (H&E) [13, 14]. This technology could provide prospective, universal, objective, and systematic second review of all prostate core biopsy material. We hypothesized that an AI system would be most useful to general pathologists in the detection of small, well-differentiated foci of PrCa. To investigate this hypothesis, we conducted a study to assess how Paige Prostate Alpha, an AI-based PrCa detection system, influenced pathologists during the diagnosis of PrCa.

Materials & methods

Three AP-board certified pathologists participated in this study. All three completed at least one fellowship, none in genitourinary pathology (two cytopathology, one surgical pathology, one gynecologic pathology). All three had practiced general pathology for 1–5 years in community hospitals, and all three rated their comfort level using a web-based software to evaluate a scanned digital slide as a 9 or greater on a scale of 1–10, with 10 representing highest level of comfort. None of the pathologists were using digital pathology in routine clinical practice. Pathologists were compensated for their participation in the study.

The study consisted of two distinct phases separated by ~4 weeks, where the second phase included assistance from Paige Prostate Alpha. In each phase, pathologists were given 8 h to assess 304 anonymized WSIs of H&E-stained prostate needle core biopsies via our web-based viewer, Paige Insight Alpha. All slides were scanned using Leica AT2 scanners at a 20× magnification (0.50 µm/pixel). All WSIs met routine quality standards at the diagnosing institution, but no additional curation to remove slides because of artifacts was performed.

One day prior to each phase, pathologists were presented with a 30-min overview of a web-based viewer, Paige Insight Alpha, and instructions for the study, which concluded with a live demonstration. In both study phases, pathologists used a monitor and a web browser of their choice. In both study phases, they classified each WSI as cancerous or benign and rated their confidence in correct classification on a scale of 1 (least confident) to 100 (most confident). For all images they classified as cancerous, pathologists marked the cancerous focus using a rectangle. For images they classified as benign, pathologists had the option to mark a suspicious focus using a rectangle. The rectangle could be of any size and could be placed anywhere on the WSI. During Phase I, WSIs were presented in a random order to the pathologists.

Phase II was identical to Phase I, except each WSI was pre-screened by Paige Prostate Alpha. Paige Prostate Alpha was used to perform cancer detection on each WSI and, for WSIs where cancer was detected, the system marked the area where cancer was detected with the highest probability (see Fig. 1). The pathologist had the option to toggle off the cancer indicator after it was displayed to better visualize the focus. At the end of each study phase, pathologists took a survey which included various questions about their experience with Paige Prostate Alpha.

Paige Prostate Alpha is based on the weakly-supervised deep learning algorithm in Campanella et al. [13], which we briefly describe here. First, each WSI is broken up into a collection of 224 px × 224 px tiles, with all tiles identified as background removed from analysis. During prediction, a ResNet-34 convolutional neural network outputs the probability of cancer for all nonbackground tiles. Subsequently, a 512-dimensional feature vector (embedding) is extracted from the convolutional neural network for the top tiles with the largest probabilities, and then these are passed into a recurrent neural network that aggregates information across tiles to make the final prediction. Paige Prostate Alpha was trained on 36,644 WSIs (7,514 had cancerous foci).

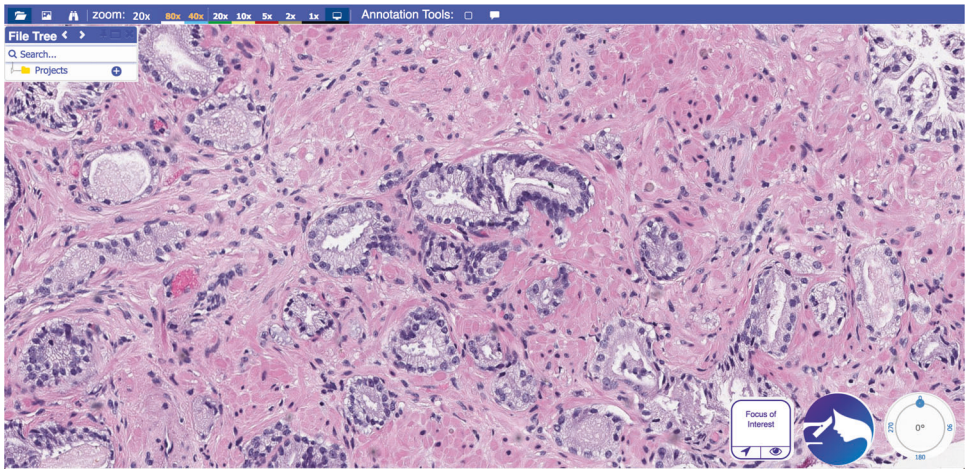

Fig. 1 Illustrative output of Paige Prostate Alpha for the pathologist. Paige Prostate Alpha was run to detect cancer in each WSI. When cancer was detected, the area of strongest signal was displayed by Paige Prostate Alpha.

Table 1 Analyzed dataset consisting of 232 WSIs of prostate needle core biopsies.

1A			
Grade group	Grade	#	(%)
1	3 + 3	63	68
2	3 + 4	13	14
3	4 + 3	3	3
4	4 + 4	10	11
5	4 + 5	1	1
Treated	N/A	3	3
Total		93	100

1B		
Measurement	#	(%)
<=0.6 mm	23	25
0.7–1.0 mm	11	12
>=1.1 mm	59	63
Total	93	100

All Gleason grade groups were represented, although not all Gleason grade combinations. Approximately 1/3 of the dataset consisted of cancers <1 mm.

The ground truth diagnosis, which definitively classified all slides as cancerous or noncancerous, was based on all studies (i.e., IHC, recuts, expert consultation) performed at the time the case was first reviewed at the diagnosing institution. Slides harboring high-grade prostatic intraepithelial neoplasia alone ($n = 12$) were assigned to the noncancerous category.

In our analysis, we excluded any slides that were used during development of Paige Prostate Alpha and any slides in which a definitive diagnosis could not be established from the single WSI. This analyzed dataset consisted of 232 anonymized H&E-stained prostate needle core biopsy WSIs. Most WSIs showed benign prostatic tissue (139, 60%), while the remaining cancerous minority (93, 40%) showed prostatic adenocarcinoma, acinar type. Three WSIs showed treated prostatic adenocarcinoma. All Gleason Grade groups were represented (Table 1A). Slightly over one-third (37%) of cancers measured 1 mm or less (Table 1B). Six (6) WSI were from treated patients. Twenty (20) WSIs harbored high-grade prostatic intraepithelial neoplasia; of these, 12 showed high-grade prostatic intraepithelial neoplasia alone and 8 showed high-grade prostatic intraepithelial neoplasia with adenocarcinoma. Two WSIs showed intraductal carcinoma in additional to conventional prostatic acinar adenocarcinoma; WSIs with intraductal carcinoma alone were excluded. In 3% of cases ($n = 7$, all cancerous cases with tumors measuring ≤1.5 mm), a PIN4 IHC was performed at the diagnosing institution which supported the diagnosis. There were no cases diagnosed as atrophy or chronic prostatitis.

In order to provide more information about the performance characteristics of our approach, we have included an analysis of an earlier version of Paige Prostate Alpha on a larger dataset that includes additional high-grade carcinoma cases as well additional benign cases with a subset of those demonstrating either atrophy or hyperplasia (Supplemental Table 1).

Statistical analysis

Statistical analysis was performed with MATLAB 2018 (MathWorks, Natick, MA, USA). We used McNemar's test to compare sensitivity and specificity with and without Paige Prostate Alpha [15]. To analyze if pathologists were faster with Paige Prostate Alpha, two-tailed paired *t* tests were used. *P* values below 0.05 were considered statistically significant.

Results

Overall sensitivity and specificity for Paige Prostate Alpha and pathologists is shown in Fig. 2 and Table 2. Performance is analyzed in greater detail below.

Paige prostate Alpha performance

Evaluated as standalone performance, Paige Prostate Alpha's sensitivity was 96% to detect cancer. Of 93 cancerous slides, Paige Prostate Alpha did not detect cancer on four slides. Of the cancerous slides where no cancer was detected, two slides showed Gleason Grade Group 1 PrCa (in one, a PIN4 IHC was performed at the diagnosing institution, supporting the cancerous diagnosis), one showed treated carcinoma, and the last slide showed only perineural invasion of Gleason Grade Group 4 PrCa (Fig. 3). Average tumor length on misclassified slides was 1.7 mm and the average tumor percentage was 12% (Table 3, Fig. 3).

Evaluated as standalone performance, Paige Prostate Alpha's specificity to detect cancer was 98%. Of 139 benign slides, Paige Prostate Alpha detected cancer on three slides, one of which showed high-grade prostatic intraepithelial neoplasia. An evaluation of these foci shows that smaller, well-formed normal glands somewhat separated from adjacent larger normal glands were misclassified cancerous (Fig. 3).

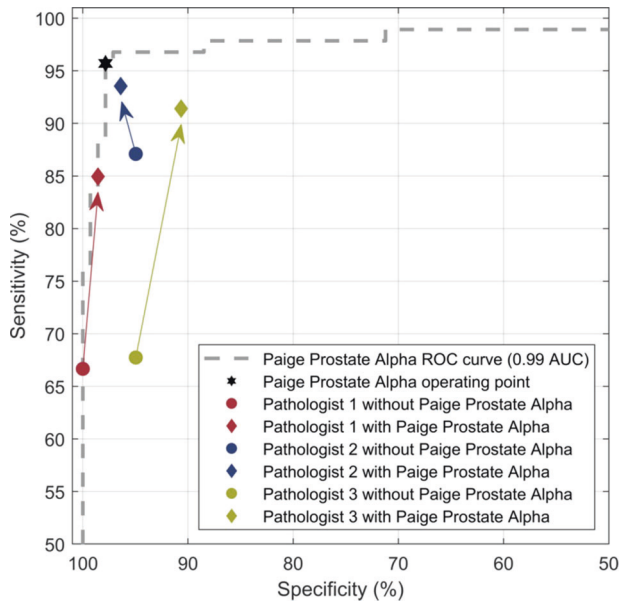

Fig. 2 Performance comparison of Paige Prostate Alpha and pathologists. The Paige Prostate Alpha operating point used for detecting cancer on a WSI is indicated. Sensitivity for all three pathologists increased with Paige Prostate Alpha (average sensitivity without Paige Prostate Alpha: 74% ± 11%; with Paige Prostate Alpha: 90% ± 4%). Specificity decreased for two pathologists, by 1–4 percentage points, and increased for one pathologist by 1 percentage point with Paige Prostate Alpha.

Pathologist performance without and with Paige Prostate Alpha

Using McNemar's test, we observed a statistically significant improvement in pathologist sensitivity when Paige

Table 3 False negative results across all pathologists & Paige Prostate Alpha.

WSI	Phase	Gleason grade	Tumor quantity % (mm)
WSIs Misclassified by Paige Prostate Alpha and pathologists			
1	I&II	3 + 3 = 6	1 (0.2)
2*	I	3 + 3 = 6	5 (0.5)
WSIs Misclassified by Paige Prostate Alpha only			
3	N/A	4 + 4 = 8	20 (3)
4	N/A	N/A-Treated	20 (3)
WSIs Misclassified by Pathologists only			
5	I	4 + 4 = 8	1 (0.2)
6	I	3 + 3 = 6	3 (0.3)
7	I	3 + 3 = 6	5 (1)
8	I	3 + 3 = 6	20 (2.7)
9	II	3 + 3 = 6	4 (0.6)
10*	II	3 + 3 = 6	8 (1.5)
11	I	3 + 3 = 6	15 (3)
12*	I	3 + 3 = 6	10 (1.5)

An asterisk (*) indicates that a PIN4 IHC was performed at the diagnosing institution, supporting the final diagnosis.

Table 2 Performance by pathologists with and without Paige Prostate Alpha.

	Pathologist 1		Pathologist 2		Pathologist 3		Average	
	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)
−Paige Prostate Alpha	66.7	100.0	87.1	95.0	67.7	95.0	73.8*	96.6
+Paige Prostate Alpha	84.9	98.6	93.5	96.4	91.4	90.6	90.0*	95.2
Change	+18.3	−1.4	+6.5	+1.4	+23.7	−4.3	+16.1	−1.4

*$P < 0.001$

Without Paige Prostate Alpha, pathologists had an average sensitivity of 74% and an average specificity of 97%. However, with Paige Prostate Alpha, the average sensitivity for pathologists increased to 90% while their specificity was 95%. Using McNemar's test, changes in sensitivity for cancer detection by the pathologists were found to be statistically significant between the two phases ($P < 0.001$); however, changes in specificity were not found to be statistically significant ($P = 0.33$).

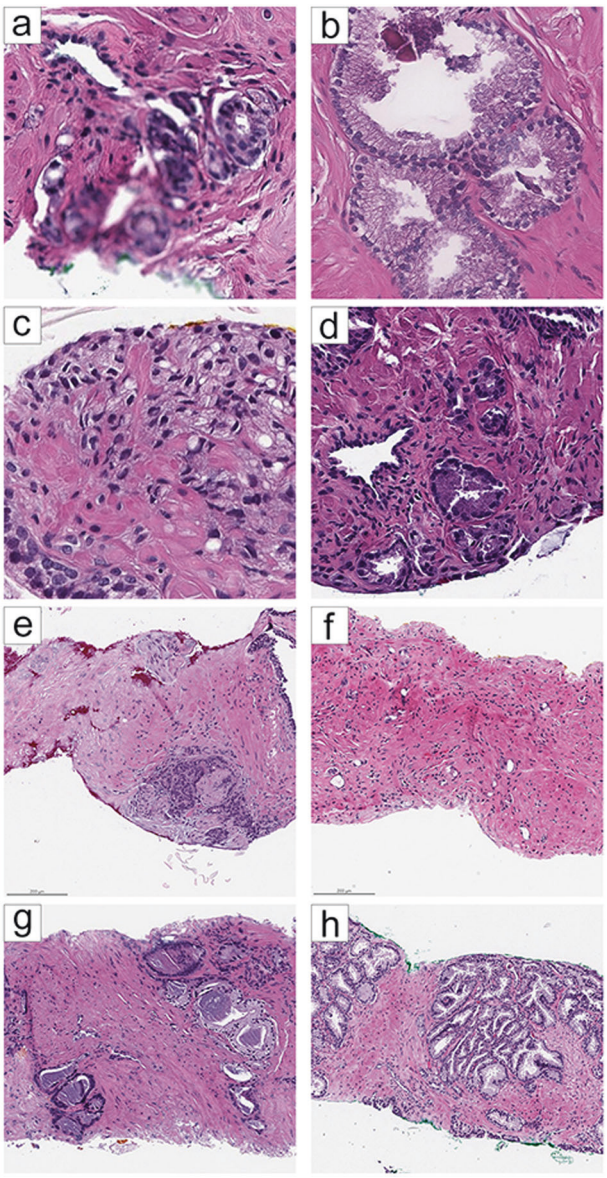

Fig. 3 Representative images of misclassified WSIs. a, **b** False negative WSIs by all pathologists in a slide where Paige Prostate Alpha also detected no cancer. **c**, **d** False negative WSIs by all pathologists when Paige Prostate Alpha correctly detected cancer. **e**, **f** True positive WSIs by most pathologists when Paige Prostate Alpha did not detect cancer. **g**, **h** True negative WSIs by most pathologists when Paige Prostate Alpha inappropriately detected cancer.

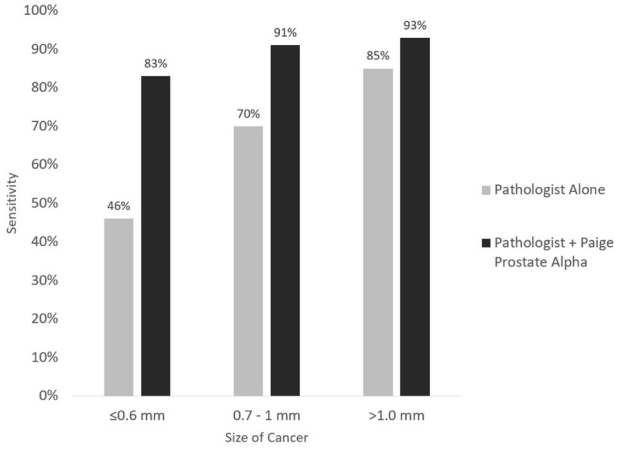

Fig. 4 Average sensitivity stratified by tumor size. With Paige Prostate Alpha, pathologists were more likely to correctly identify smaller cancers, with greatest gains in sensitivity seen in the smallest tumors. Average sensitivity in detection of tumors under 0.6 mm increased from 46% without Paige Prostate Alpha to 83% with Paige Prostate Alpha.

Prostate Alpha was used ($P < 0.001$) (Table 2). Without Paige Prostate Alpha, pathologists had an average sensitivity of 74% ± 11%. With Paige Prostate Alpha, average pathologist sensitivity increased to 90% ± 4% (Table 2). While Paige Prostate Alpha increased sensitivity for tumors of all sizes, the gains were greatest for the smallest tumors. Average sensitivity in detection of tumors under 0.6 mm increased from 46% without Paige Prostate Alpha to 83% with Paige Prostate Alpha (Fig. 4). With Paige Prostate Alpha, pathologists were more likely to correctly recognize lower grade (Grade Group 1, 2, and 3) cancers (Table 4). Average sensitivity in detection of Gleason Grade Group 1 tumors increased from 69 to 89% with Paige Prostate Alpha (Table 4). Of cancerous WSIs on which a PIN4 IHC was performed at the diagnosing institution ($n = 7$), the use of Paige Prostate Alpha increased pathologist average sensitivity from 38 to 67%.

Changes in pathologist specificity when Paige Prostate Alpha was used were not found to be statistically significant ($P = 0.327$). Without Paige Prostate Alpha, pathologists had an average specificity of 97% ± 3%. With Paige Prostate Alpha, average specificity was 95% ± 4% (Table 2). Of the benign cases with high-grade prostatic intraepithelial neoplasia ($n = 12$), specificity was unchanged for all pathologists in Phase I and Phase II, except one pathologist whose specificity decreased from 100 to 93% from Phase I versus Phase II.

Ten cancerous slides were classified as benign by all pathologists in Phase I and/or Phase II. Most of these slides were Gleason Grade Group 1 (9 WSIs), while one case showed Gleason Grade Group 4. Average tumor size was 1.2 mm and average tumor percentage was 7%. One slide showed a larger volume of tumor (2.7 mm) with tumor at the edge and partially surrounded by inflammation. A PIN4 IHC was performed at the diagnosing institution on three of these slides. Paige Prostate Alpha did not detect cancer in two of these 10 slides (WSIs 1, 2) harboring small (≤0.5 mm), low grade (Gleason 3 + 3) tumors. In WSI 1, the focus of tumor was present in a tissue fold and was out of focus in the WSI (Table 3, Fig. 2).

We observed an increase in the false negative rate among treated cancers. Of the three treated cancers evaluated by pathologists in both phases (total possible false negative

Table 4 Average sensitivity by Gleason grade groups with and without Paige Prostate Alpha.

	Grade Group 1	Grade Group 2	Grade Group 3	Grade Group 4	Grade Group 5
−Paige Prostate Alpha	69%	85%	89%	90%	100%
+Paige Prostate Alpha	89%	97%	100%	90%	100%
Change	+20%	+13%	+11%	NC	NC

With Paige Prostate Alpha, pathologists were more likely to correctly classify lower grade (Grade Group 1, 2 and 3) cancers. Average sensitivity in detection increased with Paige Prostate Alpha for all Gleason Grade Groups <4. *NC* no change.

calls = 18), 9 false negative calls were made by pathologists; Paige Prostate Alpha misclassified one slide as benign. We observed an increase in the false positive rate among treated benign slides. Of three treated benign slides evaluated by pathologists in both phases (total possible false positive calls = 18), 7 false positive calls were made by pathologists, while Paige Prostate Alpha correctly detected no cancer in all treated benign slides.

Interaction between pathologists and Paige Prostate Alpha

We sought to investigate how the detection by Paige Prostate Alpha influenced pathologists' classification of WSIs. The aggregate number of slides classified correctly (true negative or true positive) by pathologists without Paige Prostate Alpha was 609. In Phase II, 587 of those slides remained correct, all of which were slides in which Paige Prostate Alpha correctly detected cancer, while 22 became incorrect. Of the 22 WSI that became incorrect in Phase II, Paige Prostate Alpha correctly detected cancer in 17 cases and incorrectly detected cancer in five cases (Fig. 5).

The aggregate number of WSIs classified incorrectly (false negative or false positive) by pathologists without Paige Prostate Alpha was 87. In Phase II, 61 of those slides were correctly classified, while 26 remained incorrect. Of the 61 slides that were corrected in Phase II, Paige Prostate Alpha correctly detected cancer in 59 cases and of the 26 cases that pathologists still classified incorrectly in Phase II, Paige Prostate Alpha correctly detected cancer in 19 cases (Fig. 5).

All pathologists showed high confidence scores without Paige Prostate Alpha, and thus, the mean confidence score increased slightly between Phase I and Phase II, from 91 to 93.

Finally, we analyzed whether pathologists were faster at reviewing slides when using Paige Prostate Alpha with two-tailed, paired *t* tests. Due to technical factors related to our time tracking, we excluded from analysis any WSI that took a pathologist longer than 5 min to evaluate from both phases (of the 232 WSIs reviewed, total of excluded WSIs: 34; from Phase I: 23, from Phase II: 11). Overall, pathologists were significantly faster with Paige Prostate Alpha (paired *t* test, $P < 0.001$), with pathologists taking an average of 63 ± 39 seconds per slide without Paige Prostate Alpha and 55 ± 43 seconds per slide with Paige Prostate Alpha. While Paige Prostate Alpha reduced the average amount of time taken for both benign WSI and cancerous WSI, the improvement was larger for WSI with cancer. For WSI showing cancer, there was a significant improvement in speed between phases ($P < 0.001$): pathologists were 13 s faster with Paige Prostate Alpha, with a mean time per WSI of 61 ± 34 s without Paige Prostate Alpha and 48 ± 41 s with Paige Prostate Alpha. This significant improvement in speed was maintained even for cancers ≤1 mm ($P = 0.026$), where the mean time was 64 ± 33 s without Paige Prostate Alpha and 52 ± 42 s with Paige Prostate Alpha. For benign WSI, pathologists were faster by 5 s, with a mean time per WSI of 64 ± 43 s without Paige Prostate Alpha and 59 ± 44 s with Paige Prostate Alpha; however, this did not rise to the level of statistical significance ($P = 0.086$). The greater improvement in speed for cancerous slides than benign slides with Paige Prostate Alpha may be because pathologists were directly presented with visual evidence when cancer was found so less exhaustive checking was done.

In the survey given to pathologists after Phase I, all the pathologists reported they would consider digitally reviewing WSIs for primary diagnosis. After Phase II, all the pathologists reported they would consider digitally reviewing WSIs for primary diagnosis if such a system included Paige Prostate Alpha.

Discussion

This study demonstrates that Paige Prostate Alpha is new technology that has the potential to help general pathologists more accurately, efficiently diagnose PrCa in core needle biopsies, providing evidence that such an AI-enabled digital workflow offers significant benefits. To our knowledge, there have been no studies that have analyzed how the use of cancer detection technology by pathologists to interpret prostate needle biopsy slides impacts sensitivity and specificity, our main endpoints. This study showed that the use of Paige Prostate Alpha can increase diagnostic sensitivity of PrCa with statistical significance, especially small, low grade lesions which are difficult to detect, with no statistically significant impact on specificity. In addition, we showed that Paige Prostate Alpha helped pathologists review slides faster.

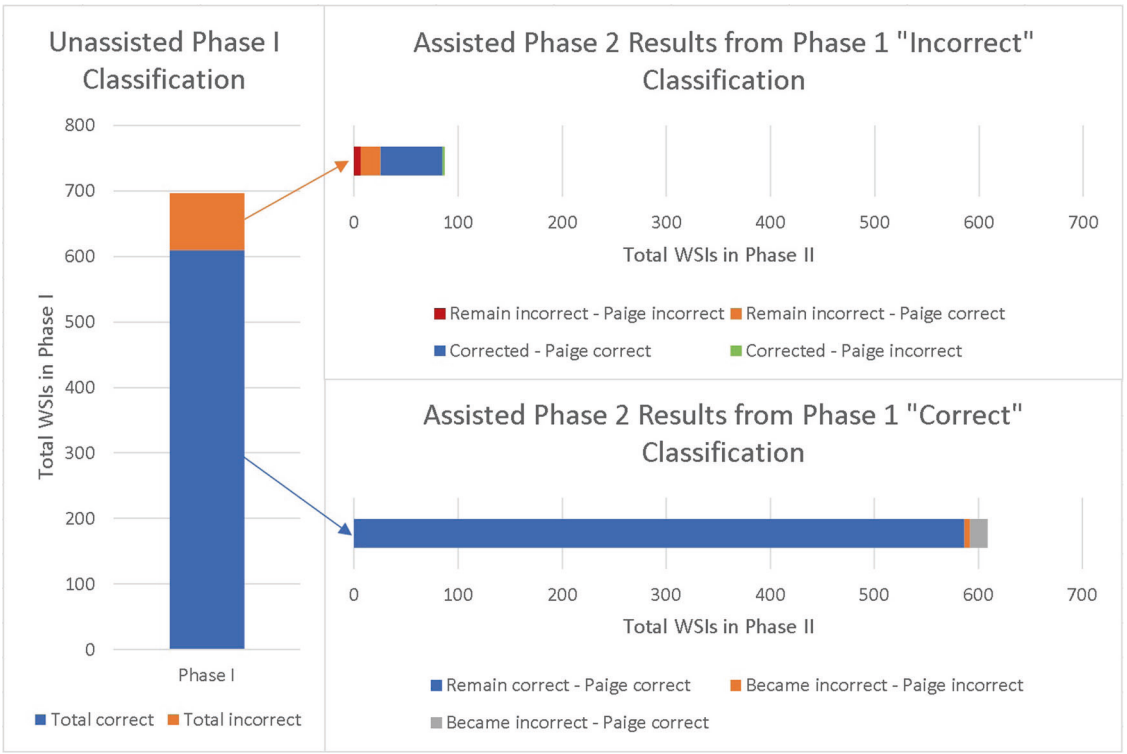

Fig. 5 Investigation of interaction between Paige Prostate Alpha and pathologists with effect on the final diagnosis. Paige Prostate Alpha likely contributed to most correct diagnoses and was unlikely to have contributed to incorrect diagnoses. Of 87 WSIs that were incorrectly diagnosed (false negative or false positive) without Paige Prostate Alpha, 61 became correctly diagnosed with Paige Prostate Alpha and, in 59 of those WSIs, Paige Prostate Alpha appropriately detected cancer or no cancer. Of 609 WSIs that were correctly diagnosed (true negative or true positive) without Paige Prostate Alpha, 22 became incorrectly diagnosed with Paige Prostate Alpha. Of those 22, Paige Prostate Alpha correctly detected cancer in 17 WSIs and incorrectly detected cancer in 5 WSIs.

The true incidence of false-negative prostate biopsy rates is unknown. The few studies published false-negative diagnosis rates between 1 and 3% although rates as high as 10% are reported [16, 17]. The variability is likely a result of a combination of the varying threshold of each pathologist to make a diagnosis of adenocarcinoma, the experience level of the pathologist, and the way that false negative diagnosis rates are measured (i.e., slide level or case level).

Our study showed that sensitivity of pathology diagnosis increased in a statistically significant manner with the use of Paige Prostate Alpha, with no statistically significant decrease in specificity. Without Paige Prostate Alpha, pathologists most frequently missed small, well-differentiated cancers, which most closely mimic benign prostate. With Paige Prostate Alpha, pathologists correctly classified smaller tumors and well-differentiated tumors, that are most likely to be missed in practice. In assessing the interaction between Paige Prostate Alpha and the pathologists, we determined that the use of Paige Prostate Alpha was likely responsible for most correct diagnoses and was unlikely to be responsible for incorrect diagnoses. We conclude that the small decrease in specificity seen in Phase II was not entirely a direct result of Paige Prostate Alpha since pathologists incorrectly diagnosed benign WSIs even in cases where Paige Prostate Alpha appropriately detected no cancer. False positive calls in Phase II might instead be a result of heightened awareness of small, well-differentiated tumors shown to the pathologist by Paige Prostate Alpha on other slides that biased pathologists on benign slides, or reflect the variability in criteria employed by pathologists in establishing diagnoses for small, well-differentiated lesions.

Steiner et al. measured the improvement in sensitivity and specificity of pathologists as well as their efficiency with and without the use of an AI system with respect to detection of breast cancer metastasis in lymph nodes [18]. Similarly, they found that the greatest gains with AI usage were seen in sensitivity, especially in the detection of small metastatic foci (micrometastasis). They also observed efficiency gains in the time to review slides. Consistent with our results, the Steiner et al. study reinforces that AI systems can improve sensitivity and efficiency, particularly in cases where tumor burden is low.

Our study has some important limitations. Although the skill of the pathologists varied despite similar backgrounds, the number of study participants is small and limited to pathologists with less genitourinary subspecialty experience, which might have contributed to the sensitivity without Paige Prostate Alpha. Further studies would be

needed to determine if our results generalize to the general pathology community or to pathologists with more genitourinary pathology experience.

We did not provide pathologists with data regarding the performance of Paige Prostate Alpha for detecting cancer in a standalone setting. A follow-up study could assess how knowledge of the efficacy of this technology would influence the pathologists' behavior.

Finally, our dataset was limited and could have expanded to include more benign mimickers of malignancy, a greater variety in Gleason Grade, and additional, rare variants of prostatic adenocarcinoma.

Although our study was designed to simulate some of tasks a pathologist must complete when analyzing a slide, the study design asked the pathologists to determine a diagnosis by reviewing the H&E alone, without ancillary studies or consultation. Our finding that the use of Paige Prostate Alpha increased pathologist average sensitivity in assessing cancerous WSIs on which a PIN4 IHC was performed at the diagnosing institution might suggest that Paige Prostate Alpha could be used as an alternate form of evidence of cancer, in the same way an IHC or an internal consultation might be used, potentially reducing costs and turnaround time. However, further studies are needed to assess how the use of Paige Prostate Alpha might impact this use of these additional tools.

Importantly, in our study, WSIs were not filtered or manually reviewed for overall image quality after scanning beyond the standard, clinical scanning workflow and the equipment used for viewing was not standardized across participants, supporting the clinical utility of a cancer detection tool such as this one as an effective universal, unbiased second review tool for digitized prostate needle core biopsies. Second review of slides has shown to be effective in improving diagnosis. However, it is rarely employed and, even when it is used, only cancerous cases are assessed, potentially allowing for improvements in specificity, but unlikely in sensitivity [8–11]. Furthermore, re-review of cancerous cases by a second pathologist can still result in incorrect classification; the second reviewer might make the same interpretative error as the diagnosing pathologist and, because rereview increases workload, might devote less time to the review [8].

In summary, this study provides evidence that Paige Prostate Alpha can improve sensitivity of diagnosis with statistical significance in comparison to current methods in pathology, and that it could serve as an effective universal second read tool for every prostate needle core biopsy case. Because Paige Prostate Alpha was built on the expert knowledge of genitourinary pathologists, it also demonstrates that this approach can democratize expert knowledge so that it can be used in locations where subspecialists are not present, including countries with large scale healthcare disparities.

Acknowledgements We would like to thank Carla Leibowitz and Helen Melville for their help in creating the figures.

Funding Funding for this study was provided by Paige.AI, Inc.

Compliance with ethical standards

Conflict of interest PR, JS, CK, RC, RG, JDK, LG, and TJF are employees at Paige and are equity holders in Paige. SK is an equity holder in Paige. DSK and VR are consultants for Paige. VR is a consultant for Cepheid. DSK has received speaking/consulting compensation from Merck and is equity holder in Paige. TJF has intellectual property interests relevant to the work that is the subject of this paper. Memorial Sloan Kettering Cancer Center has financial interests in Paige and intellectual property interests relevant to the work that is the subject of this paper.

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. Catalona WJ. Prostate cancer screening. Med Clin North Am. 2018;102:199–214.
3. Matoso A, Epstein JI. Defining clinically significant prostate cancer on the basis of pathological findings. Histopathology. 2019;74:135–45.
4. Mahal BA, Butler S, Franco I, Spratt DE, Rebbeck TR, D'Amico AV, et al. Use of active surveillance or watchful waiting for low-risk prostate cancer and management trends across risk groups in the United States, 2010–2015. JAMA. 2019;321:704.
5. Amin MB, Lin DW, Gore JL, Srigley JR, Samaratunga H, Egevad L, et al. The critical role of the pathologist in determining eligibility for active surveillance as a management option in patients with prostate cancer: consensus statement with recommendations supported by the College of American Pathologists, International Society of Urological Pathology, Association of Directors of Anatomic and Surgical Pathology, the New Zealand Society of Pathologists, and the Prostate Cancer Foundation. Arch Pathol Lab Med. 2014;138:1387–405.
6. Montironi R, Hammond EH, Lin DW, Gore JL, Srigley JR, Samaratunga H, et al. Consensus statement with recommendations on active surveillance inclusion criteria and definition of progression in men with localized prostate cancer: the critical role of the pathologist. Virchows Arch. 2014;465:623–8.
7. Varma M, Narahari K, Mason M, Oxley JD, Berney DM. Contemporary prostate biopsy reporting: insights from a survey of clinicians' use of pathology data. J Clin Pathol. 2018;71:874–8.
8. 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;54:310–4.
9. Brimo F, Schultz L, Epstein JI. The value of mandatory second opinion pathology review of prostate needle biopsy interpretation before radical prostatectomy. J Urol. 2010;184:126–30.
10. Renshaw AA, Cartagena N, Granter SR, Gould EW. Agreement and error rates using blinded review to evaluate surgical pathology of biopsy material. Am J Clin Pathol. 2003;119:797–800.

11. Renshaw AA, Gould EW. Measuring the value of review of pathology material by a second pathologist. Am J Clin Pathol. 2006;125:737–9.
12. 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.
13. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25:1301–9.
14. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep. 2016;6. https://doi.org/10.1038/srep26286.
15. Trajman A, Luiz RR. McNemar χ^2 test revisited: comparing sensitivity and specificity of diagnostic examinations. Scand J Clin Lab Invest. 2008;68:77–80.
16. Yang C, Humphrey PA. False-negative histopathologic diagnosis of prostatic adenocarcinoma. Arch Pathol Lab Med 2019. https://doi.org/10.5858/arpa.2019-0456-RA.
17. van der Kwast TH, Lopes C, Martikainen PM, Pihl C-G, Santonja C, Neetens I, et al. Report of the Pathology Committee: false-positive and false-negative diagnoses of prostate cancer. BJU Int. 2003;92:62–65.
18. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol. 2018;42:1636–46.