CASE STUDY

Paige®

Clinical Validation of Artificial Intelligence–Augmented Pathology Diagnosis Demonstrates Significant Gains in Diagnostic Accuracy in Prostate Cancer Detection

Patricia Raciti, Jillian Sue, Juan A. Retamero, Rodrigo Ceballos, Ran Godrich, Jeremy D. Kunz, Adam Casson, Dilip Thiagarajan, Zahra Ebrahimzadeh, Julian Viret, Donghun Lee, Peter J. Schüffler, George DeMuth, Emre Gulturk, Christopher Kanan, Brandon Rothrock, Jorge Reis-Filho, David S. Klimstra, Victor Reuter, Thomas J. Fuchs

Paige designed this ground-breaking study to understand how assistance from Paige Prostate Detect, an AI that aids in the detection of prostate cancers, could impact pathologist performance.

A group of 16 pathologists reviewed 610 prostate needle biopsy whole-slide images first without support from the AI, and then after a washout period, assisted by it. Changes in their sensitivity, specificity, and efficiency were assessed.

FDA

The study's outcomes were pivotal in Paige Prostate Detect ultimately receiving the first FDA-approval for an AI in pathology.

Paige Prostate Detect was found to:

Reduce cancer detection **errors** by

70%

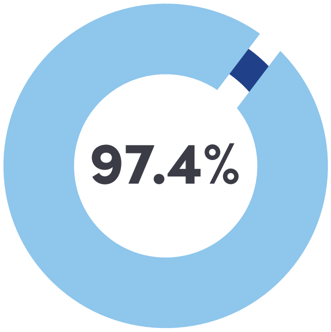
Increase pathologist **specificity** by

24%

Deliver high performance on a challenging data set composed of cases from

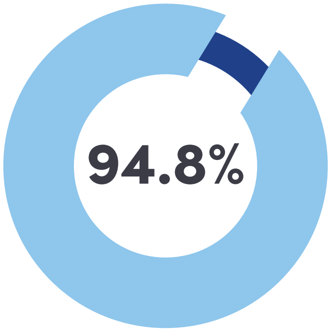
218

unique institutions worldwide



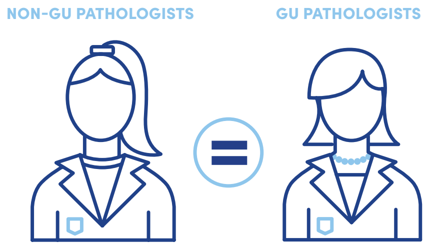
SENSITIVITY

for detecting prostate cancers



SPECIFICITY

for detecting prostate cancers



Statistically significant sensitivity gains were seen among **non-GU pathologists** and **GU pathologists**

Clinical Validation of Artificial Intelligence–Augmented Pathology Diagnosis Demonstrates Significant Gains in Diagnostic Accuracy in Prostate Cancer Detection

Patricia Raciti, MD; Jillian Sue, MS; Juan A. Retamero, MD; Rodrigo Ceballos, MSc; Ran Godrich, MS; Jeremy D. Kunz, MSc; Adam Casson, BS; Dilip Thiagarajan, MS; Zahra Ebrahimzadeh, MSc; Julian Viret, MEng; Donghun Lee, MEng; Peter J. Schüffler, DrSc; George DeMuth, MS; Emre Gulturk, MSc; Christopher Kanan, PhD; Brandon Rothrock, PhD; Jorge Reis-Filho, MD, PhD, FRCPath; David S. Klimstra, MD; Victor Reuter, MD; Thomas J. Fuchs, DrSc

Context.—**Prostate cancer diagnosis rests on accurate assessment of tissue by a pathologist. The application of artificial intelligence (AI) to digitized whole slide images (WSIs) can aid pathologists in cancer diagnosis, but robust, diverse evidence in a simulated clinical setting is lacking.**

Objective.—**To compare the diagnostic accuracy of pathologists reading WSIs of prostatic biopsy specimens with and without AI assistance.**

Design.—**Eighteen pathologists, 2 of whom were genitourinary subspecialists, evaluated 610 prostate needle core biopsy WSIs prepared at 218 institutions, with the option for deferral. Two evaluations were performed sequentially for each WSI: initially without assistance, and immediately thereafter aided by Paige Prostate (PaPr), a deep learning–based system that provides a WSI-level binary classification of suspicious for cancer or benign and pinpoints the location that has the greatest probability of harboring cancer on suspicious WSIs. Pathologists' changes in sensitivity and specificity between the assisted and unassisted modalities were assessed, together with the impact of PaPr output on the assisted reads.**

Results.—**Using PaPr, pathologists improved their sensitivity and specificity across all histologic grades and tumor sizes. Accuracy gains on both benign and cancerous WSIs could be attributed to PaPr, which correctly classified 100% of the WSIs showing corrected diagnoses in the PaPr-assisted phase.**

Conclusions.—**This study demonstrates the effectiveness and safety of an AI tool for pathologists in simulated diagnostic practice, bridging the gap between computational pathology research and its clinical application, and resulted in the first US Food and Drug Administration authorization of an AI system in pathology.**

(*Arch Pathol Lab Med.* doi: 10.5858/arpa.2022-0066-OA)

Prostate cancer (PrCa) is the second most common cancer among men and one of the leading causes of cancer death globally.[1] Pathologic examination of prostate biopsy tissue by light microscopy remains the gold standard in PrCa diagnosis. Cancer identification and the reporting of associated cancer parameters by pathologists allow clinicians to undertake a crucial treatment decision: differentiating between patients at risk of clinically significant PrCa associated with higher mortality and requiring definitive treatment versus patients with clinically indolent PrCa that can be closely followed and does not require immediate treatment.[2] Evidence demonstrates that pathologists have suboptimal sensitivity at detecting cancer in core needle biopsy specimens, particularly when cancerous foci are small or well differentiated.[3–5] Detecting all positive cores, even those containing minimal cancer foci, is crucial for the clinician to decide between definitive therapy and watchful waiting.[2,6] Cocktail immunohistochemical (IHC) analysis can mitigate false-negative diagnoses, but cannot be applied as a universal screening tool on tissue because of cost and diagnostic delay.[5] Additional tools that facilitate the screening of prostate tissue in all patients are needed.

With the approval of the first digital pathology system for routine diagnostic use by the US Food and Drug Admin-

Accepted for publication July 28, 2022.

Supplemental digital content is available for this article. See text for hyperlink.

From Paige (Raciti, Sue, Retamero, Ceballos, Godrich, Kunz, Casson, Thiagarajan, Ebrahimzadeh, Viret, Lee, Schüffler, Gulturk, Kanan, Rothrock, Klimstra, Fuchs), New York, New York; Stat One (DeMuth), Morrisville, North Carolina; and the Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, New York (Reis-Filho, Reuter). Raciti and Sue contributed equally to this work.

This study was funded by Paige. Paige supported the study design; the collection, analysis, and interpretation of data; the writing of the report; and the decision to submit the paper for publication.

Raciti, Sue, Retamero, Ceballos, Godrich, Kunz, Casson, Thiagarajan, Ebrahimzadeh, Viret, Lee, Gulturk, Rothrock, Klimstra, and Fuchs are employees at Paige and are equity holders in Paige. Kanan is a paid consultant to Paige and a member of Paige's Scientific Advisory Board. Schüffler is a consultant for Paige and is equity holder in Paige. Retamero is also a consultant for Sakura-Finetek. DeMuth is owner of Stat One LLC, which provided analysis support as a paid consulting activity. Reis-Filho is a paid consultant for Paige and a member of the Scientific Advisory Board of Paige. 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. Reuter has no relevant financial interest in the products or companies described in this article.

Corresponding author: Thomas J. Fuchs, DrSc, Paige, 11 Times Sq, FL 37, New York, NY 10036 (email: thomas.fuchs@paige.ai).

istration (FDA) in 2017,[7,8] the field of pathology is undergoing a digital revolution whereby glass slides of tissue samples are digitized as whole slide images (WSIs) and examined using computer monitors rather than traditional light microscopes. Not only has digital pathology enhanced ease of consultation and viewing,[9,10] improved efficiency,[11,12] and boosted pathologist satisfaction,[13,14] but crucially, digital pathology permits the application of artificial intelligence (AI) systems.[15] Recent advances in machine learning, particularly in the design and training of deep neural networks, have accelerated the development of computational pathology, in which state-of-the-art deep learning solutions can power decision support systems applicable to diagnostic pathology as well as tools to aid biomarker discovery.[16,17]

Prostate pathology in particular has become an area of strong interest in the application of machine learning algorithms.[18–21] Machine learning algorithms applied to PrCa diagnosis have demonstrated high diagnostic standalone performance, with areas under the curve above 0.98.[18–21] An alpha version of the Paige Prostate (PaPr) system for cancer diagnosis, trained on vast data sets using sophisticated machine learning approaches that did not require pixel-level manual annotation of cancer, resulted in clinical-grade outputs, invariant to stain and preparation differences across laboratories, and importantly resulted in breakthrough designation by the FDA in 2019.[3,18] Akin to previous tools designed as diagnostic aids, a comprehensive assessment of such tools in the hands of diagnosticians is paramount to understand the impact on diagnostic accuracy. An early, small-scale study of PaPr Alpha was undertaken in a simulated sign-out environment to test how pathologists interact with AI tools, and it showed that pathologists using PaPr Alpha had a statistically significant increase in sensitivity, from 74% to 90%.[3]

Encouraged by these initial results, we conducted a more comprehensive study to robustly establish the impact of using PaPr, a more mature version of PaPr Alpha, as a diagnostic aid to pathologists, assisting them in the detection of prostatic acinar adenocarcinoma in digitized core needle biopsy specimens in a simulated clinical workflow. This study was designed to demonstrate sufficient efficacy and safety of the software to justify its use in routine diagnostic practice.

METHODS

Paige Prostate

The details about the training of PaPr have been reported elsewhere.[4,18] In brief, PaPR is a deep learning–based system that was trained using 32 341 prostate biopsy WSIs from 6775 patients, scanned at ×20 magnification using multiple-instance learning,[18] enabling it to be trained without any pixel-level manual annotations. The ground truth for training consisted solely of the binary classification of each WSI as benign or cancerous, based on the information within the corresponding pathology report. All glass slides were prepared and diagnosed by genitourinary (GU) subspecialist pathologists at Memorial Sloan Kettering Cancer Center (MSKCC), New York, New York.

PaPr outputs a binary WSI-level classification for suspicion of cancer based on applying a cutoff value to the continuous score, and if a WSI is suspicious, it additionally outputs a focus location on the WSI with the greatest statistical evidence for suspicion of cancer. See the supplemental digital content (containing 8 tables) for additional details on the design, training, and optimization of PaPr.

Study Setup

The objective of the study was to evaluate pathologist performance in detecting invasive cancer and foci suspicious for invasive cancer on WSIs of hematoxylin-eosin (H&E)–stained prostate core needle biopsies unaided and with PaPr assistance.

The data set included 610 de-identified prostate needle biopsy WSIs stained with H&E. Fifty percent of the slides were originally prepared at MSKCC and the other 50% were prepared at 217 external institutions; all slides were reviewed and diagnosed using conventional microscopy at MSKCC. For slides originating from MSKCC, a single glass slide with 3 levels was digitized; for slides originating outside of MSKCC, the glass slide felt to be most representative of the diagnostic findings was digitized. A Philips Ultra-Fast Scanner Series 2 (In Vitro Diagnostics version 3.2), which scans only at ×40 and is part of an FDA-approved digital pathology solution, was used for digitization. The ground truth diagnosis, which definitively classified all slides as cancerous or noncancerous, was based on all studies (ie, IHC, recuts, expert consultation) performed at the time the case was reported by GU subspecialized pathologists at MSKCC; PaPr was not used as an aid in establishing the original diagnosis. Four hundred twenty (of 610; 69%) WSIs were benign (Supplemental Table 1), and 190 (of 610; 31%) WSIs, selected both consecutively and based on tumor size, harbored invasive acinar adenocarcinoma, intraductal carcinoma, or atypical small acinar proliferation suspicious for cancer (ASAP). The main inclusion criterion was tumor size, so in cases with multiple parts and Gleason patterns, the part with the lowest tumor volume was selected, regardless of Gleason grade. Ninety WSIs harbored invasive adenocarcinoma measuring 0.5 mm or less (defined as hard cases) or intraductal carcinoma, 90 WSIs contained invasive adenocarcinoma measuring more than 0.5 mm (defined as easy cases), and 10 WSIs were classified as ASAP. This enrichment of the data for small tumor foci was done to challenge the system with lesions that can be overlooked by pathologists. Benign slides contained a representative spectrum of pathologic findings, including basal cell hyperplasia, high-grade prostatic intraepithelial neoplasia (HGPIN), atrophy, and inflammation, encountered among a consecutively benign biopsy cohort, without data set enrichment. In particular, atrophy was present in 4 of 420 benign WSIs (1%), whereas 29 WSIs were of previously treated prostatic tissue. WSIs with significant scanning issues that compromised the ability to render a definitive diagnosis (eg, out-of-focus areas, tissue folds, or missing tissue), were excluded. WSIs used for development of PaPr were excluded. The protocol and all study materials were reviewed and approved by an institutional review board of WIRB-Copernicus Group.

Sixteen pathologists, board certified in anatomic pathology and practicing for 2 to 34 years (mean, 11.6 ± 10.8 years), completed this study. Two (12.5%) had completed a fellowship in GU pathology; the other 14 (87.5%) had completed at least one fellowship, but none in GU pathology. Three (18.7%) completed the study at a Clinical Laboratory Improvement Amendments–certified site and 13 (81.3%) completed the study remotely. Pathologists were compensated for their participation in the study.

All participants were provided with a Philips PP27QHD monitor. At review sites, upload speed ranged from 0.1 to 244.25 Mbps and download speed ranged from 5.72 to 498.56 Mbps (Supplemental Table 2). All participants were asked to submit workstation photographs to document connection of the monitor in a functional state. Prior to study initiation, participants were trained on the use of Paige's FDA-cleared pathology viewer, FullFocus; PaPr; and the data capture tool via a presentation during a live videoconference session; participants had to demonstrate competency at the end of training. All study participants were assigned to review 610 WSIs in a randomized order (Figure 1). Using FullFocus, each pathologist reviewed each WSI twice, sequentially (paired read), resulting in 19 520 unique reads and 9760 paired reads. During the initial (unassisted) read, the participant was presented with the WSI without PaPr assistance; during the second (assisted) read, immediately following the first, the participant evaluated the

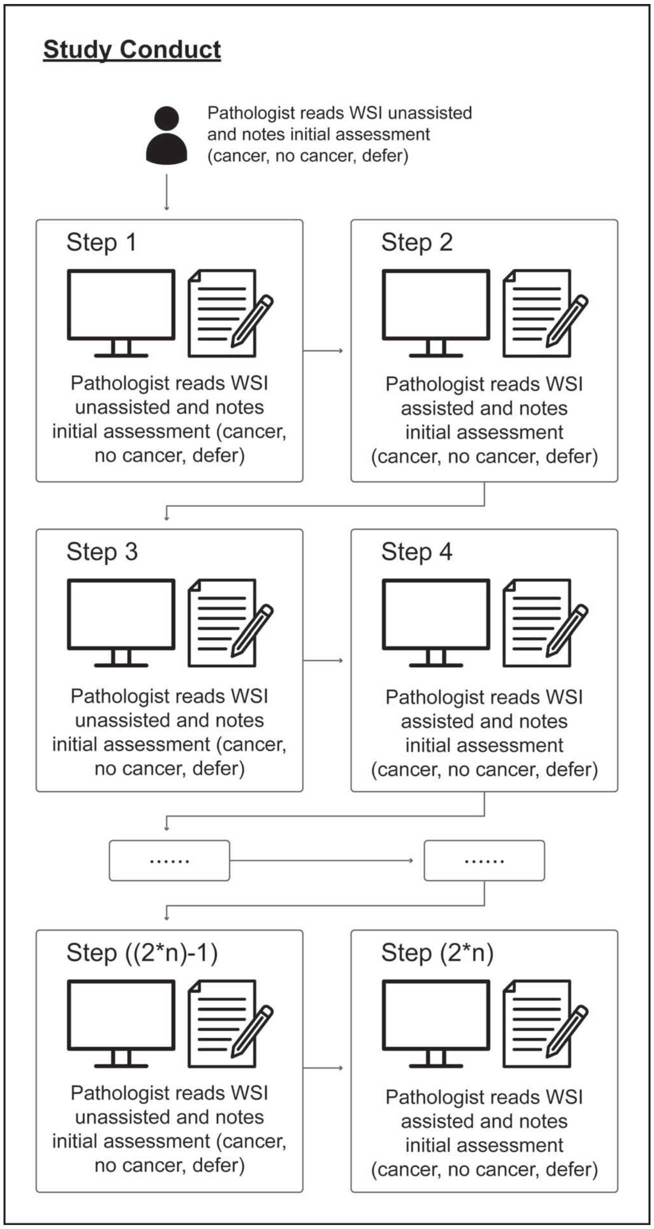

Study Conduct
Pathologist reads WSI unassisted and notes initial assessment (cancer, no cancer, defer)
Step 1
Pathologist reads WSI unassisted and notes initial assessment (cancer, no cancer, defer)
Step 2
Pathologist reads WSI assisted and notes initial assessment (cancer, no cancer, defer)
Step 3
Pathologist reads WSI unassisted and notes initial assessment (cancer, no cancer, defer)
Step 4
Pathologist reads WSI assisted and notes initial assessment (cancer, no cancer, defer)
......
......
Step ((2*n)-1)
Pathologist reads WSI unassisted and notes initial assessment (cancer, no cancer, defer)
Step (2*n)
Pathologist reads WSI assisted and notes initial assessment (cancer, no cancer, defer)

Figure 1. *Study protocol: sequential paired-read study design to assess potential benefit to pathologists of using Paige Prostate. Abbreviation: WSI, whole slide image.*

same WSI with the results of PaPr. The study was thus designed to assess the use of PaPr as a second-read device. For each read, the participant was instructed to classify the slide as (1) harboring invasive cancer or suspicious for cancer (ie, ASAP), (2) not harboring cancer, or (3) defer for more information, which could include an intent to request more histologic levels, seek consultation, or undertake IHC analysis, or another documented reason. The classification of the first read could not be altered after the assisted mode was displayed. All participants were given up to 7 days to complete the study.

Statistical Analysis

Results were reported using descriptive statistics. Primary analyses were based on coprimary endpoints of sensitivity and specificity obtained from the reader evaluations compared with the diagnostic categorization of each WSI. The stand-alone performance of the PaPr was also summarized. For all analyses, "defer for more information" was considered a correct assessment in either modality, because the various steps taken after a deferral (ie, IHC analysis, additional levels, second opinion) most likely would have led to a correct diagnosis.

The primary analyses were performed using the multireader multicase (MRMC) method outlined in Gallas et al[22] for binary data from a fully crossed study. The overall average of reader responses was obtained giving all readers equal weight. Variance estimates for each type of read and the differences in reads were obtained and used to calculate 2-sided 95% CIs and evaluate the primary study hypotheses. Other methods used in the analysis of the study included bootstrapping and random effect logistic models.

For sensitivity, the primary analysis was based on demonstrating the difference in assisted minus unassisted average reader performance exceeded a superiority margin. A 1-sided *P* value was generated based on the Z-score calculated with the estimated reader average $\text{Sensitivity}_{\text{assisted}} - \text{Sensitivity}_{\text{unassisted}} - \text{SM}$ (superiority margin) divided by the standard error for the difference obtained from the MRMC model. A 1-sided *P* value of less than or equal to 0.025 was considered statistically significant. The 2-sided 95% CI for the difference in sensitivity rates (assisted minus unassisted) was also provided based on the reader averages and variance of the difference from the MRMC model.

For specificity, the same approach was used to evaluate if the difference in assisted minus assisted average reader performance exceed a noninferiority margin. A 1-sided *P* value was calculated based on the Z-score calculated with the estimated reader average $\text{Specificity}_{\text{assisted}} - \text{Specificity}_{\text{unassisted}} + \text{NIM}$ (noninferiority margin) divided by the standard error for the difference from the MRMC model. A 1-sided *P* value less than or equal to .025 was considered statistically significant. The 2-sided 95% CI for the average reader specificities for each method and difference was calculated using the same method.

FINDINGS

The stand-alone diagnostic accuracy performance of PaPr on this data set had a sensitivity of 97.4% (95% CI, 94.0%–99.1%) and specificity of 94.8% (95% CI, 92.2%–96.7%) at the WSI level. The AUC was 0.99 (95% CI, 0.98–0.99). Performance was evaluated across patient age, race, and ethnicity, and differences in performance across these variables were insignificant (Supplemental Table 3). Of the 5 false-negative WSIs, 4 were diagnosed as ASAP and 1 showed 0.2 mm of Gleason 3 + 3 cancer present only perineurally. PaPr showed a specificity on previously treated (ie, status post radiation or hormonal therapy) prostatic tissue of 92%. False-positive classifications (22) included WSIs with atrophy (1), treated benign tissue (2), HGPIN (2), and aggregates of small, crowded benign glands.

The differences between sensitivity and specificity of unassisted pathologist reads and PaPr-assisted reads were evaluated. The average sensitivity of PaPr-assisted pathologist reads increased significantly, by 8 percentage points, from 88.7% to 96.6% (95% MRMC CI, 4.5%–11.5%; $P <$ .001), reducing detection errors by 70%. The average specificity of PaPr-assisted pathologist reads also significantly increased, by 0.7 percentage points, from 97.3% to 98.0% (95% MRMC CI, 0.1%–1.2%; $P =$ .02) (Figure 2, A and B; Supplemental Table 4), increasing specificity overall by 24%. Statistically significant sensitivity gains were seen among non-GU pathologists (8.5% gain; 95% MRMC CI, 4.8%–12.6%; $P <$ .001) and GU pathologists (3.9% gain; 95% MRMC CI, 0.5%–7.9%; $P =$ .02) (Figure 2, A and B); gains in specificity were also seen in both non-GU pathologists and specialists, but reached statistical significance only among non-GU pathologists (0.7% gain; 95% MRMC CI, 0.0%–1.6%; $P =$ .04). However, there was no

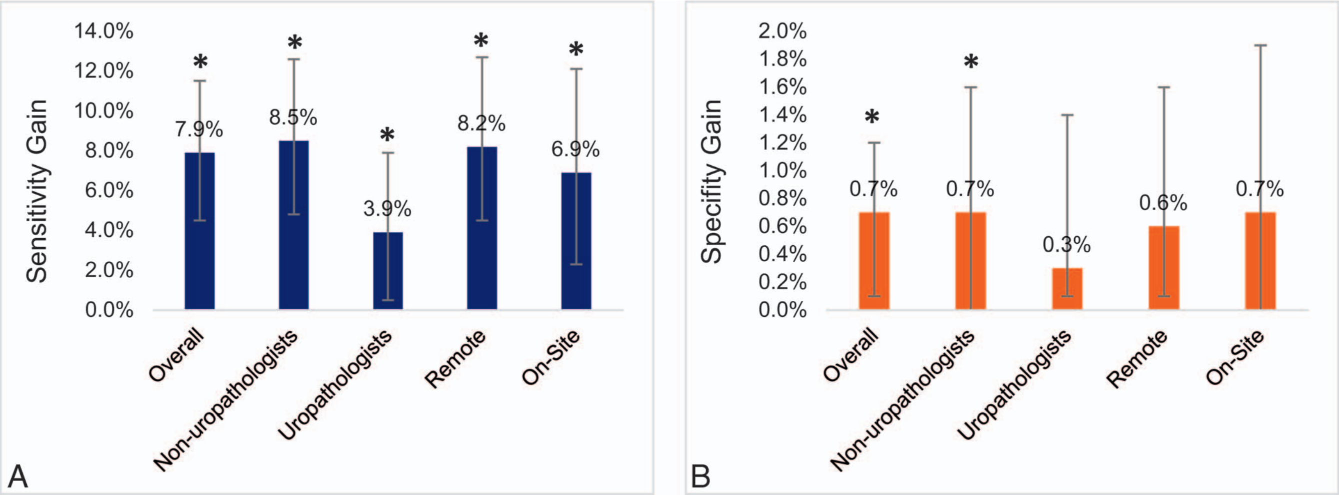


Figure 2. *Gains in performance by pathologists, stratified by experience and location. Asterisks (*) indicate statistically significant changes (*P *< .05). A, Statistically significant gains in sensitivity were seen regardless of pathologist type and location of slide review. B, Statistically significant gains in specificity were seen among all pathologists overall and among nonuropathologists.*

statistically significant difference in these improvements between these 2 groups (see Supplemental Table 5). There were nonsignificant differences in unassisted sensitivity and specificity between reader groups with 10 or less and more than 10 years of experience.

Gains in sensitivity were observed across all histologic grades and tumor sizes (see Supplemental Table 6). Similarly, the presence of atrophy or HGPIN in benign slides was not correlated with reader performance (see Supplemental Table 7). Statistically significant gains in sensitivity were seen among pathologists reviewing WSIs on-site (6.8%; 95% MRMC CI, 2.3%–12.1%; $P = .004$; 0.7%) and remotely (8.2% sensitivity gain; 95% MRMC CI, 4.5%–12.7%; $P < .001$) (Figure 2, A and B; Supplemental Table 6); gains in specificity were observed for on-site and remote participants, but did not reach statistical significance. There was no statistically significant difference in the observed improvements between the remote and on-site pathologists.

To ensure that the enhanced human performance in both the detection of cancerous foci and the recognition of benign tissue was not simply a result of reexamination of the WSIs, we assessed the number of paired reads with Paige-driven changes, defined as those in which the PaPr classification was correct and matched the PaPr-assisted pathologist reads. Overall, PaPr-assisted pathologist reads differed from PaPr-unassisted pathologist reads in 797 reads (8.2% of total 9760 paired reads). All 341 paired reads (100%) that became correct (initially incorrect in PaPr-unassisted pathologist reads and either correct or deferred in PaPr-assisted pathologist reads) were Paige driven; 85.2% of 54 paired reads that became incorrect (initially correct or deferred in PaPr-unassisted pathologist reads and incorrect in PaPr-assisted pathologist reads) were Paige driven (Figure 3; Supplemental Table 8).

We sought to understand how PaPr might have influenced the assisted read in cases where it differed from the unassisted read to evaluate the assay's impact on sensitivity and specificity (Figures 3 through 5, A through H; Supplemental Table 8).

Assessing improvements in specificity, in 15 paired reads, a benign WSI was incorrectly read as cancer in the unassisted read and correctly classified as benign in the PaPr-assisted read; in 100% of these paired reads, PaPr correctly identified the slide as benign.

Assessing improvements in sensitivity, in 59 paired reads, a cancerous WSI was incorrectly read as benign in the unassisted read and correctly read as cancerous in the PaPr-assisted read; in 100% of these paired reads, PaPr correctly identified the slide as cancerous.

In 15 paired reads, a benign WSI was correctly read as benign in the unassisted read and incorrectly read as cancer in the assisted read; in 20% (3 of 15) of these paired reads, PaPr correctly identified the slide as benign. Review of the 11 WSIs corresponding to these reads demonstrated that the most commonly misclassified slides contained small foci of small glands (mimicking ASAP), florid HGPIN, and a benign prostatic gland around a nerve. For the majority of these WSIs (8 of these WSIs with 39 paired reads), however, the more common paired read pattern was true-negative in the unassisted read and defer in the assisted read.

In 4 paired reads, a cancerous WSI was correctly read as cancerous in the unassisted read and incorrectly read as benign in the PaPr-assisted read; in 100% of these paired reads, PaPr correctly identified the slide as cancerous. Each of these shifts came from a different participating pathologist, and showed International Society of Urological Pathology 1, 2, and 4 cancers involving from 2% to 80% of the tissue. It is likely that these were a result of transcription errors by the participants; however, in a WSI, PaPr correctly classified the WSI as cancerous, but the focus of interest misidentified the cancerous focus.

Finally, we assessed potential efficiency gains and efficiency losses in the use of PaPr. We defined paired reads resulting in efficiency gains as those that were initially deferred in PaPr-unassisted pathologist reads and correct in the PaPr-assisted pathologist reads. Of 288 paired reads showing an efficiency gain, 99.7% (287) were Paige driven. We defined paired reads resulting in efficiency losses as those that initially correct in PaPr-unassisted pathologist reads and deferred in the PaPr-assisted pathologist reads. Of 114 paired reads showing an efficiency loss, 98.2% (112) were Paige driven (Figure 3; Supplemental Table 8).

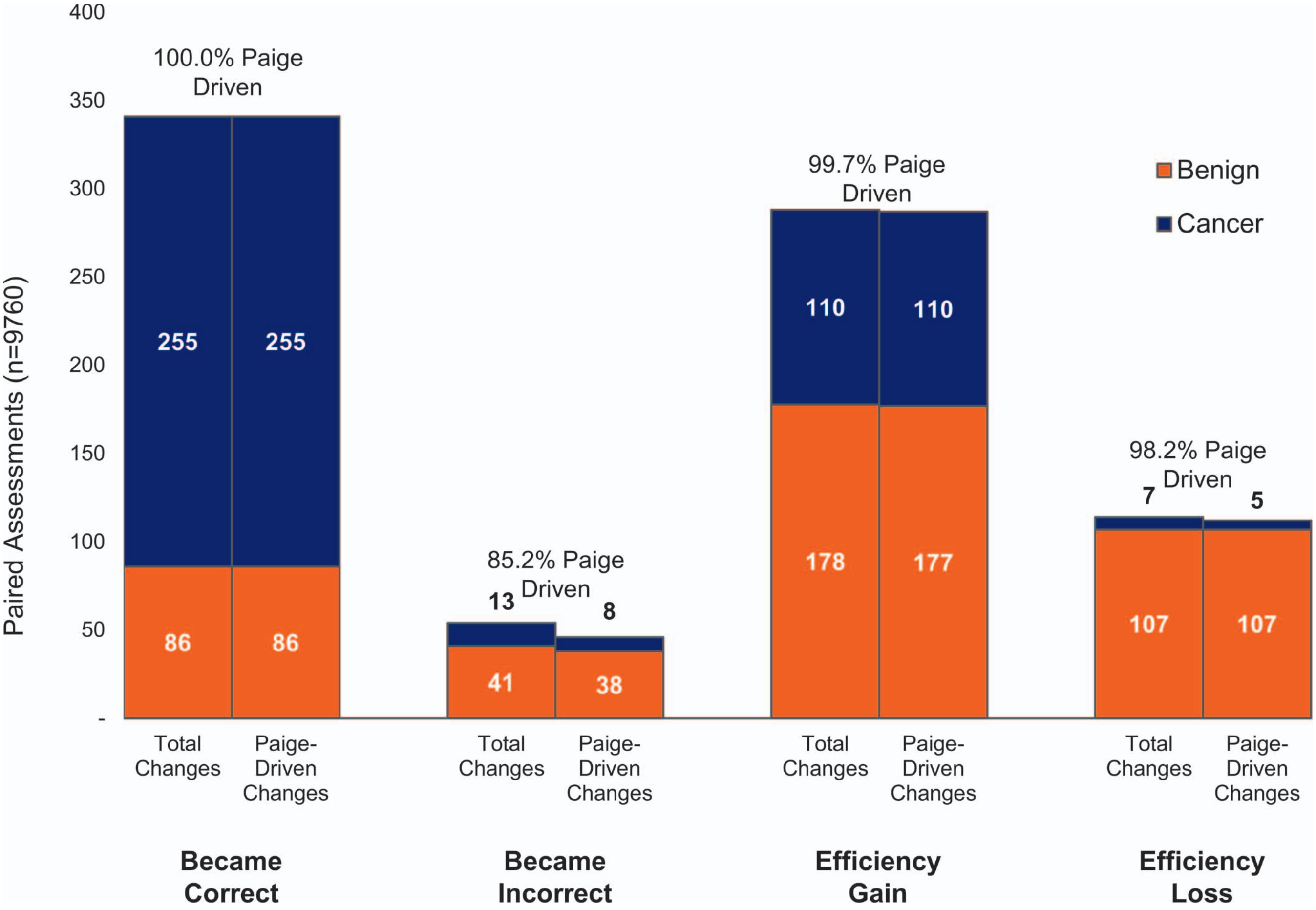


Figure 3. *Change in pathologists' assessments driven by Paige Prostate. Paige Prostate–assisted pathologist reads differed from Paige Prostate–unassisted pathologist reads in 797 reads (8.2% of total 9760 paired reads). Paige-driven changes are defined as those in which the Paige Prostate classification was correct and matched the Paige Prostate–assisted pathologist reads. All reads that became correct were Paige driven. Many reads that became incorrect were also Paige driven. Paige-driven reads that resulted in efficiency gains slightly outnumbered those that resulted in efficiency losses, but overall rates were similar.*

INTERPRETATION

This study robustly demonstrates the impact that clinical-grade AI tools can have on patient diagnosis when applied to prostatic core needle biopsy interpretation. This work demonstrates the efficacy and safety of PaPr, and it was the basis for the first marketing authorization by the FDA for an AI system in pathology, which showed objective and systematic analysis of the system being used within a digital pathology workflow. The FDA authorization paves the way for the use of AI in routine clinical work; Tim Stenzel, MD, PhD, director of the FDA's Office of In Vitro Diagnostics and Radiological Health, stated that "[t]he authorization of this AI-based software can help increase the number of identified prostate biopsy samples with cancerous tissue, which can ultimately save lives."[7]

This study demonstrates the robust stand-alone performance of PaPr on a challenging data set composed of cases from 218 unique institutions worldwide, a testament to the robustness of the system, its insensitivity to H&E staining and tissue preparation variabilities, and its overall generalizability. Thus, a high level of performance of PaPr can be expected without the need for on-site calibration to maintain the algorithm's performance characteristics, a crucial distinguishing feature from other similar tools.[20] Furthermore, performance was evaluated across patient age, race, and ethnicity, and differences in performance across these variables were insignificant, a particularly important analysis that ensures there is no unknown bias in output and that such tools do not contribute to health care disparities.[23]

The greatest contribution of the present study is that it examines, in a simulated clinical environment, how the diagnostic precision of pathologists improves when aided by clinical-grade AI, specifically resulting in more true positives and more true negatives.

Accuracy gains on both benign and cancerous WSIs can be attributed to PaPr, which correctly classified 100% of the WSIs showing correct diagnoses in the PaPr-assisted phase.

Diagnostic aid tools may be viewed as having utility limited to certain diagnosticians (ie, physicians with less experience) or certain disease states (ie, low-volume or well-differentiated carcinomas), calling into question the benefit to the patient of broad adoption. Here, we show that both non–GU-specialist pathologists and GU-specialist pathologists showed statistically significant improvements in sensitivity based on using PaPr. Despite the smaller number of GU-specialist participants, the differences between the reader types were not statistically significant, suggesting that the use of PaPr may bring the performance of non-GU pathologists closer to that of GU specialists, democratizing

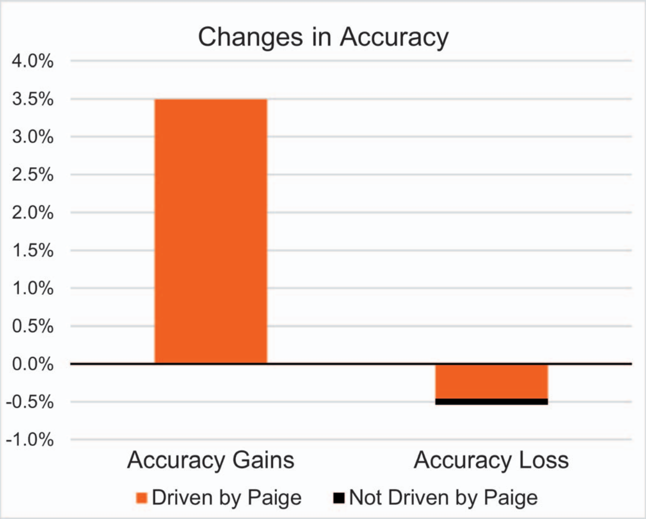


Figure 4. *Overall rates for assessment shifts between unassisted and assisted reads. Gains in accuracy of 3.5% were observed, as well as losses of 0.5%. All gains were attributed to Paige Prostate results, whereas only 0.08% of the losses could be attributed to Paige Prostate results. Accuracy gains are defined as incorrect Paige Prostate–unassisted pathologist reads that became correct or were deferred in the Paige Prostate–assisted pathologist reads. Accuracy losses are defined as correct or deferred Paige Prostate–unassisted pathologist reads that became incorrect in the Paige Prostate–assisted pathologist reads.*

expertise. Given that non-GU pathologists diagnose the majority of prostate biopsies, it is expected that the broad use of a tool such as PaPr would bring benefits to a wide portion of the patient population. Improvements in cancer detection were also not limited to specific grades or sizes, as gains in sensitivity with the use of PaPr were observed across all histologic grades and tumor sizes, even in WSIs where the grade and volume of cancer might affect the decision about the need for definitive therapy. Though PaPr was designed to maximize sensitivity, our results show a concomitant modest improvement in specificity, demonstrating how PaPr could aid pathologists in correctly classifying benign mimickers of cancer, such as treatment effect and atrophy.

Previous work on the human-AI interaction in pathology is limited. Steiner et al[24] demonstrated that a tool for the detection of breast cancer metastasis in lymph nodes can improve sensitivity in pathologists' detection of metastases, particularly for small metastatic foci. Pantanowitz et al[20] described the use of a PrCa detection algorithm as a postdiagnostic second-read tool. When used in practice, the tool issued cancer alerts on 509 of the 11 429 H&E-stained slides. The majority (90.9%) of the cancer alerts shown to the pathologist resulted in no additional action; a minority of alerts were found to trigger additional IHC analysis, levels, or consultation, and, in one case, the system led to the detection of cancer that was initially overlooked, allowing the patient to be enrolled in an active surveillance protocol.[20] These studies, however, did not take into account pathologists' stand-alone performance with the cases, nor how their interaction with AI affected diagnostic reads in simulated clinical practice.

Our group previously evaluated an alpha version of PaPr in the hands of non–GU-specialized pathologists, assessing the interaction of the pathologist and the AI tool in PrCa detection in a simulated clinical practice.[3] This study involved unassisted and assisted reads separated by a 4-week washout period. There was a statistically significant increase in sensitivity with the use of PaPr Alpha, which was maintained for easy-to-miss small and low-grade cancers.[3] This study was limited by its small participant size, small data set, and lack of a deferral option, and thus, alone, could not be used to justify routine use of PaPr Alpha.

This current study differed from and improved on the prior studies in many important ways. First, the larger participant pool included subspecialized GU pathologists. Second, by including the option for deferral, it more closely approximated the real-world pathology workflow. Third, the data set, although still enriched for small, well-differentiated cancers, included a greater breadth of carcinoma (in both volume and Gleason scores) as well as benign conditions known to mimic PrCa. Fourth, the function of PaPr was optimized compared with the prior alpha version. Last, the study design, which lacked a washout period and consisted of a prior PaPr-independent evaluation stage, was purposeful in its intent to evaluate PaPr as a second-read software device.

Further insight into the interplay between pathologist and PaPr can be seen in the rare cases that were originally correct without PaPr that were then deferred or became incorrect in the PaPr-assisted read. In benign WSIs, when the unassisted diagnosis was correct but changed to defer or cancerous in assisted mode, PaPr had identified false-positive foci, some of which showed benign mimics of carcinoma, in most shifts. In cancerous WSIs, when the unassisted diagnosis was correct but changed to defer in the assisted mode, PaPr had incorrectly classified the WSI as benign in the majority of shifts (Figures 3 and 4; Supplemental Table 8). This analysis highlights the need for pathologists to understand the scenarios where machine learning tools can underperform and the need for pathologists to maintain a high level of active participation in the diagnostic process. It further underscores the importance of studies such as this one in enhancing our understanding of the human-machine interaction and its potential impact on the final diagnosis.

Our study has limitations. The entire range of rare cancer subtypes and benign mimics was not included in the data set, and the performance of PaPr in such cases has not been assessed. Atrophy, for example, can be mistakenly diagnosed as malignancy or ASAP, and the atrophic variant of acinar adenocarcinoma may pose a diagnostic challenge. In our study's data set, atrophy was present in 1% (4 of 420) of benign slides, but the performance of AI-assisted pathologists in discriminating benign mimickers and the variety of cancer subtypes, although of great relevance, was beyond the scope of this study. Pathologists reviewed only one H&E-stained WSI devoid of clinical or radiologic context; it is possible that the number of deferrals would have been less had the participants had access to levels and clinical-radiologic information. Future studies should include complete cases, ideally in a prospective manner. The study was not designed to measure the difference in time spent reviewing WSIs unassisted compared with assisted, and future studies might be designed to capture these data and understand the impact of user design in result display, connection speeds, and other factors on reading times. Finally, in a small subset of WSIs, pathologists changed an initially correct interpretation to an incorrect interpretation after PaPr displayed incorrect results. Future studies are warranted to ascertain the factors that can optimize the AI-

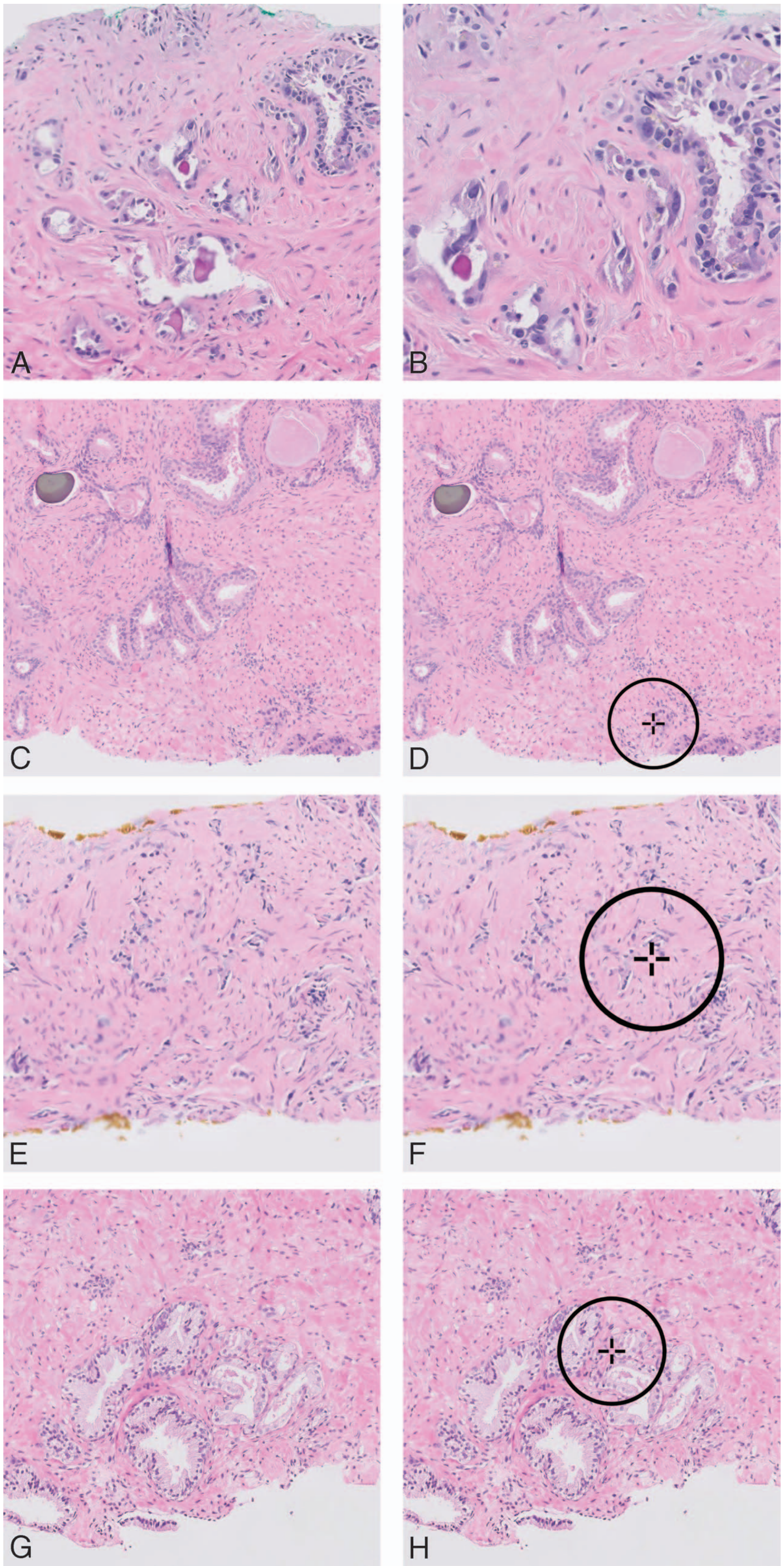


Figure 5. *Representative images of assessment shifts in whole slide images (WSIs). A, Benign seminal vesicle misclassified as cancer by 5 pathologists and deferred by 3 pathologists. Paige correctly classified the WSI as benign. With Paige, 5 pathologists correctly classified the slide as benign and 3 pathologists deferred. B, High-power view showing enlarged cells with cytoplasmic pigment, characteristic of seminal vesicle. C, International Society of Urological Pathology (ISUP) 4 cancer involving 2% of the tissue misclassified as benign by 12 pathologists and deferred by 1 pathologist. Paige correctly classified the WSI as cancerous. With Paige, 9 pathologists correctly classified the slide as cancerous and 4 pathologists deferred. D, Indication of focus of interest in C, by Paige. E, ISUP 5 cancer involving 95% of the tissue misclassified as benign by 1 pathologist and deferred by 1 pathologist. Paige correctly classified the WSI as cancerous. With Paige, both pathologists correctly classified the slide as cancerous. F, Indication of focus of interest in E, by Paige. G, Benign WSI misclassified by Paige as cancerous. Seven pathologists correctly classified the WSI as benign and 1 pathologist deferred. With Paige, 6 pathologists deferred and 1 incorrectly classified the WSI as cancerous. H, Indication of focus of interest in G, by Paige (hematoxylin-eosin, original magnifications ×10 [A and C through H] and ×40 [B]).*

pathologist interaction, whether it be further confidence in the algorithm, more experience in digital diagnostics by the user, or a better user interface.

In summary, our study indicates that PaPr, a clinical-grade, robust AI, is mature enough to be applied broadly in the clinical setting. Its use by GU-specialized and non–GU-specialized pathologists during prostate core needle biopsy assessment results in greater sensitivity and specificity. For clinicians and patients, it means higher confidence in the resulting pathology diagnosis, regardless of access to specialist expertise. For pathologists, it heralds a revolution in the way pathology is practiced and reinforces the central role pathologists play in patient care.

The authors would like to thank Caitlin Chambers, BA, and Rebecca Kalant, BFA, for their assistance in data structuring and figure creation.

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(6):394–424. doi:10.3322/caac.21492

2. Amin MB, Lin DW, Gore JL, 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(10):1387–1405. doi:10.5858/arpa.2014-0219-SA

3. 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;33(10):2058–2066. doi:10.1038/s41379-020-0551-y

4. da Silva LM, Pereira EM, Salles PGO, et al. Independent real-world application of a clinical-grade automated prostate cancer detection system. *J Pathol*. 2021;254(2):147–158. doi:10.1002/path.5662

5. Yang C, Humphrey PA. False-negative histopathologic diagnosis of prostatic adenocarcinoma [published online November 15, 2019]. *Arch Pathol Lab Med*. doi:10.5858/arpa.2019-0456-RA

6. 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(10):874–878. doi:10.1136/jclinpath-2018-205093

7. FDA allows marketing of first whole slide imaging system for digital pathology. FDA Web site. https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-whole-slide-imaging-system-digital-pathology. Accessed May 10, 2021.

8. 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(11):1383–1387. doi:10.5858/arpa.2017-0496-CP

9. Baidoshvili A, Stathonikos N, Freling G, et al. Validation of a whole-slide image-based teleconsultation network. *Histopathology*. 2018;73(5):777–783. doi:10.1111/his.13673

10. Parwani AV. Next generation diagnostic pathology: use of digital pathology and artificial intelligence tools to augment a pathological diagnosis. *Diagn Pathol*. 2019;14(1):1–3. doi:10.1186/s13000-019-0921-2

11. Baidoshvili A, Bucur A, van Leeuwen J, van der Laak J, Kluin P, van Diest PJ. Evaluating the benefits of digital pathology implementation: time savings in laboratory logistics. *Histopathology*. 2018;73(5):784–794. doi:10.1111/his.13691

12. Retamero JA, Aneiros-Fernandez J, del Moral RG. Complete digital pathology for routine histopathology diagnosis in a multicenter hospital network. *Arch Pathol Lab Med*. 2020;144(2):221–228. doi:10.5858/arpa.2018-0541-OA

13. Retamero JA, Aneiros-Fernandez J, Del Moral RG. Microscope?: no, thanks: user experience with complete digital pathology for routine diagnosis. *Arch Pathol Lab Med*. 2020;144(6):672–673. doi:10.5858/arpa.2019-0355-LE

14. Schüffler PJ, Geneslaw L, Yarlagadda DVK, et al. Integrated digital pathology at scale: a solution for clinical diagnostics and cancer research at a large academic medical center. *J Am Med Inform Assoc*. 2021;28(9):1874–1884. doi:10.1093/jamia/ocab085

15. Salto-Tellez M, Maxwell P, Hamilton P. Artificial intelligence—the third revolution in pathology. *Histopathology*. 2019;74(3):372–376. doi:10.1111/his.13760

16. Fuchs TJ, Buhmann JM. Computational pathology: challenges and promises for tissue analysis. *Comput Med Imaging Graph*. 2011;35(7–8):515–530. doi:10.1016/j.compmedimag.2011.02.006

17. Campanella G, Silva VWK, Fuchs TJ. Terabyte-scale deep multiple instance learning for classification and localization in pathology. *ArXiv*. http://arxiv.org/abs/1805.06983. Published May 17, 2018. Accessed January 11, 2019.

18. 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(8):1301–1309. doi:10.1038/s41591-019-0508-1

19. 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. doi:10.1038/srep26286

20. Pantanowitz L, Quiroga-Garza GM, Bien L, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet Digit Health*. 2020;2(8):e407–e416. doi:10.1016/S2589-7500(20)30159-X

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(2):222–232. doi:10.1016/S1470-2045(19)30738-7

22. Gallas BD, Pennello GA, Myers KJ. Multireader multicase variance analysis for binary data. *J Opt Soc Am A*. 2007;24(12):B70. doi:10.1364/JOSAA.24.000B70

23. Jackson BR, Ye Y, Crawford JM, et al. The ethics of artificial intelligence in pathology and laboratory medicine: principles and practice. *Acad Pathol*. 2021;8:237428952199078. doi:10.1177/2374289521990784

24. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol*. 2018;42(12):1636–1646. doi:10.1097/PAS.0000000000001151