JAMA Network Open

Original Investigation | Pathology and Laboratory Medicine

Evaluation of the Use of Combined Artificial Intelligence and Pathologist Assessment to Review and Grade Prostate Biopsies

David F. Steiner, MD, PhD; Kunal Nagpal, MS; Rory Sayres, PhD; Davis J. Foote, BS; Benjamin D. Wedin, BA; Adam Pearce, BS; Carrie J. Cai, PhD; Samantha R. Winter, PhD; Matthew Symonds, BS; Liron Yatziv, PhD; Andrei Kapishnikov, MS; Trissia Brown, MBBS; Isabelle Flament-Auvigne, MD; Fraser Tan, PhD; Martin C. Stumpe, PhD; Pan-Pan Jiang, PhD; Yun Liu, PhD; Po-Hsuan Cameron Chen, PhD; Greg S. Corrado, PhD; Michael Terry, PhD; Craig H. Mermel, MD, PhD

Abstract

IMPORTANCE Expert-level artificial intelligence (AI) algorithms for prostate biopsy grading have recently been developed. However, the potential impact of integrating such algorithms into pathologist workflows remains largely unexplored.

OBJECTIVE To evaluate an expert-level AI-based assistive tool when used by pathologists for the grading of prostate biopsies.

DESIGN, SETTING, AND PARTICIPANTS This diagnostic study used a fully crossed multiple-reader, multiple-case design to evaluate an AI-based assistive tool for prostate biopsy grading. Retrospective grading of prostate core needle biopsies from 2 independent medical laboratories in the US was performed between October 2019 and January 2020. A total of 20 general pathologists reviewed 240 prostate core needle biopsies from 240 patients. Each pathologist was randomized to 1 of 2 study cohorts. The 2 cohorts reviewed every case in the opposite modality (with AI assistance vs without AI assistance) to each other, with the modality switching after every 10 cases. After a minimum 4-week washout period for each batch, the pathologists reviewed the cases for a second time using the opposite modality. The pathologist-provided grade group for each biopsy was compared with the majority opinion of urologic pathology subspecialists.

EXPOSURE An AI-based assistive tool for Gleason grading of prostate biopsies.

MAIN OUTCOMES AND MEASURES Agreement between pathologists and subspecialists with and without the use of an AI-based assistive tool for the grading of all prostate biopsies and Gleason grade group 1 biopsies.

RESULTS Biopsies from 240 patients (median age, 67 years; range, 39-91 years) with a median prostate-specific antigen level of 6.5 ng/mL (range, 0.6-97.0 ng/mL) were included in the analyses. Artificial intelligence–assisted review by pathologists was associated with a 5.6% increase (95% CI, 3.2%-7.9%; *P* < .001) in agreement with subspecialists (from 69.7% for unassisted reviews to 75.3% for assisted reviews) across all biopsies and a 6.2% increase (95% CI, 2.7%-9.8%; *P* = .001) in agreement with subspecialists (from 72.3% for unassisted reviews to 78.5% for assisted reviews) for grade group 1 biopsies. A secondary analysis indicated that AI assistance was also associated with improvements in tumor detection, mean review time, mean self-reported confidence, and interpathologist agreement.

Key Points

Question Is the use of an artificial intelligence–based assistive tool associated with improvements in the grading of prostate needle biopsies by pathologists?

Findings In this diagnostic study involving 20 pathologists who reviewed 240 prostate biopsies, the use of an artificial intelligence–based assistive tool was associated with significant increases in grading agreement between pathologists and subspecialists, from 70% to 75% across all biopsies and from 72% to 79% for Gleason grade group 1 biopsies.

Meaning The study's findings indicated that the use of an artificial intelligence tool may help pathologists grade prostate biopsies more consistently with the opinions of subspecialists.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

CONCLUSIONS AND RELEVANCE In this study, the use of an AI-based assistive tool for the review of prostate biopsies was associated with improvements in the quality, efficiency, and consistency of cancer detection and grading.

JAMA Network Open. 2020;3(11):e2023267.
Corrected on December 15, 2020. doi:10.1001/jamanetworkopen.2020.23267

Introduction

For patients with prostate cancer, the Gleason grade represents one of the most important factors in risk stratification and treatment.[1-3] The current Gleason grade group (GG) system involves classification into 1 of 5 prognostic groups (GG1 through GG5, with higher GG indicating greater clinical risk) based on the relative amounts of Gleason patterns (ranging from 3 to 5, with 3 indicating low-grade carcinoma with well-formed glands and 5 indicating undifferentiated, or anaplastic, carcinoma) present. Despite its clinical importance, Gleason grading is highly subjective, with substantial interpathologist variability.[4-9] Although urologic subspecialty–trained pathologists have been reported to have higher rates of concordance with each other as well as higher accuracy than general pathologists with regard to the risk stratification of patients,[10-12] the number of urologic subspecialists is insufficient to review the large volume of prostate biopsies performed each year.

Several deep learning–based algorithms with expert-level performance (ie, high agreement with subspecialist urologic pathologists) for prostate cancer detection and Gleason grading have recently been developed.[13-15] Although it has been suggested that such algorithms may be able to improve the quality or efficiency of biopsy grading by pathologists, this potential has not been formally investigated. In other areas of pathology, studies have indicated the potential for AI-based assistance to improve diagnostic performance on tasks such as cancer detection[16,17] and mitoses quantitation.[18,19] Initial efforts to understand the impact of AI assistance with regard to more complex diagnostic tasks, such as cancer subtype classification, have also been described recently.[20-22] To date, the benefit of such algorithms has been most clear for computer-aided detection, primarily aiding the pathologist in detecting small regions of interest that might otherwise be easily missed or laborious to find. In contrast, computer-aided diagnosis aims to address a more challenging problem involving both detection and interpretation. To improve diagnostic accuracy in the grading of prostate biopsies, an assistive tool must have both high performance and the ability to guide pathologists toward the most accurate interpretation.

In this study, we developed and validated an AI-based assistive tool for prostate biopsy interpretation. This assistive tool was based on a recently developed deep learning model for prostate biopsy grading.[23] We tested the use of the tool in a large fully crossed multiple-reader, multiple-case study by using a diverse set of prostate biopsies, a rigorous reference standard, and integration of human-computer interaction insights.

Methods

Study Data and Design

Deidentified whole slide images of prostate core needle biopsy specimens were obtained using biopsies from the validation set of a previous study,[23] in which the process was described. Biopsies with nongradable prostate cancer variants or quality issues that precluded diagnosis were excluded. A set of 240 biopsies (**Table 1**) was sampled to power for grading performance differences on GG1 biopsies while also approximating clinical distribution among tumor-containing biopsies.[24] Additional details are available in the Study Data section of eMethods in the Supplement. The study was approved by the institutional review board of Quorum (Seattle, Washington) and deemed exempt from informed consent because all data and images were deidentified. This study followed the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline.

The design of this fully crossed multiple-reader, multiple-case diagnostic study is illustrated in **Figure 1**A. A total of 240 biopsies were reviewed by 20 pathologists in both AI-assisted and unassisted modes between October 2019 and January 2020. All pathologists were board certified in the US, with a median of 7.5 years (range, 1-27 years) of posttraining clinical experience without urologic subspecialization. The median self-reported prostate biopsy review volume was 1 to 2 cases per week (range, 0 to ≥5 cases per week). Additional details are available in the Study Design section of eMethods and in eTable 1 in the Supplement.

Digital Design and Development

The deep learning system underlying the assistive tool used in this study has been previously described.[23] In brief, an AI model was trained to perform Gleason grading of prostate biopsies using pathologist-annotated digitized histologic slides. Additional details are available in the Digital Assistant Design and Development section of eMethods in the Supplement.

In addition to ensuring an accurate AI model, the development of a useful assistive tool requires an effective (eg, clear, intuitive, and presenting the most salient information without distraction) user interface and an understanding of how to use the tool. For this study, the user interface and training materials were developed via formative user studies and previous research on this topic.[25,26] The final design of the user interface included overall GG classification for the biopsy, Gleason pattern localization, quantitation of Gleason patterns, and total tumor involvement (Figure 1B). An optional visualization of the AI confidence level for Gleason pattern interpretations was also created (eFigure 1 in the Supplement). Training materials were developed to provide all pathologists with working knowledge of the viewer and the assistive tool before reviewing study biopsies. Additional details are available in the Pathologist Training and Onboarding section of eMethods in the Supplement.

Biopsy Review and Classification

All needle biopsies were independently reviewed by urologic subspecialist pathologists to establish reference standard GGs. For each biopsy, subspecialists had access to 3 serial sections of hematoxylin and eosin–stained images as well as a PIN4 (comprising alpha-methylacyl coenzyme A racemase, tumor protein p63, and high-molecular-weight cytokeratin antibodies) immunohistochemistry–

Table 1. Patient Characteristics[a]

	No. (%)		
Characteristic	ML1	ML2	ML1 and ML2
Total participants, No.	85	155	240
Age at biopsy, y			
<65	43 (50.6)	53 (34.2)	96 (40.0)
≥65	37 (43.5)	102 (65.8)	139 (57.9)
Not available	5 (5.9)	NA	5 (2.1)
PSA level at biopsy, ng/mL			
<10	22 (25.9)	102 (65.8)	124 (51.7)
≥10	2 (2.3)	34 (21.9)	36 (15.0)
Not available	61 (71.8)	19 (12.3)	80 (33.3)
Reference standard grade group			
No tumor	20 (23.5)	20 (12.9)	40 (16.7)
Grade group			
1	35 (41.2)	75 (48.4)	110 (45.8)
2	10 (11.8)	40 (25.8)	50 (20.8)
3	10 (11.8)	10 (6.5)	20 (8.3)
4-5	10 (11.8)	10 (6.5)	20 (8.3)

Abbreviations: ML, medical laboratory; NA, not applicable; PSA, prostate-specific antigen.

SI conversion factor: To convert PSA to micrograms per liter, multiply by 1.0.

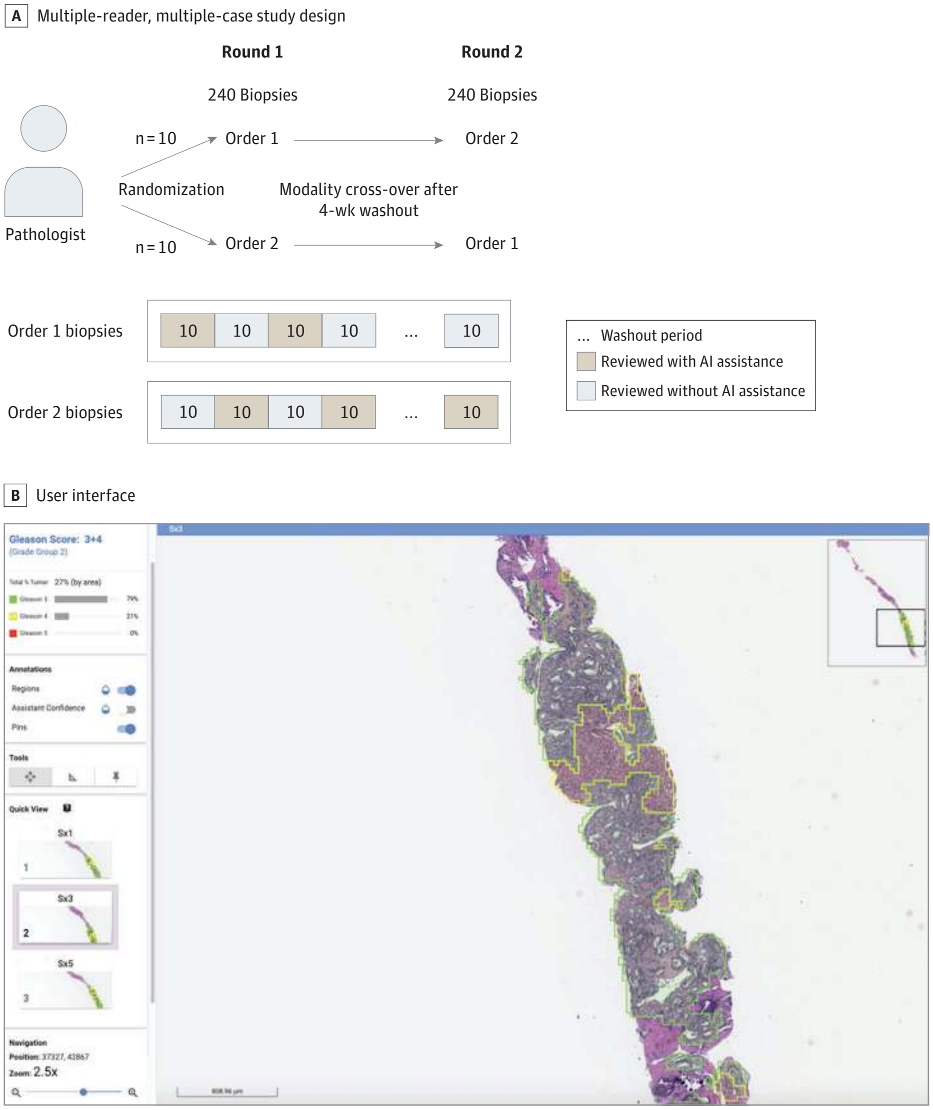
[a] Biopsies were obtained from 2 independent medical laboratories (ML1 and ML2). One core needle biopsy per independent case was included in the study.

stained image. Each biopsy was first reviewed by 2 subspecialists from a cohort of 6 subspecialists with a median of 20 years (range, 18-34 years) of posttraining experience. For instances in which the first 2 subspecialists agreed on the final GG (180 cases [75.0%]), that GG was used. If the 2 subspecialists did not agree on the final GG (60 cases [25.0%]), a third subspecialist independently reviewed the biopsy, and the majority opinion was used.

A total of 20 general pathologists reviewed 240 prostate needle biopsies from 240 cases. Each pathologist was randomized to 1 of 2 study cohorts. The 2 cohorts reviewed every case in the modality (with AI assistance vs without AI assistance) opposite to each other, with the modality switching after every 10 cases. After a minimum 4-week washout period for each batch, pathologists reviewed the cases for a second time using the modality opposite to what they had previously used. The pathologist-provided grade group for each biopsy was compared with the majority opinion of the urologic subspecialists.

Pathologists interpreted biopsies based on the 2014 International Society of Urological Pathology grading guidelines,[27] providing GGs as well as tumor and Gleason pattern quantitation. Clinical information was not provided during grading. The pathologists were asked to review and

Figure 1. User Interface for Artificial Intelligence (AI)–Based Assistive Tool and Summary of Study Design



A, Each pathologist was randomized to 1 of 2 study cohorts. The 2 cohorts reviewed every case in the opposite assistance modality to each other, with the modality switching after every 10 cases. After a minimum 4-week washout period for each batch, each pathologist reviewed the cases for a second time using the opposite modality. Details of the implementation of case distribution and washout period are available in the Study Design section of eMethods in the Supplement. The order of biopsies within each block was randomized independently for each pathologist and each round of the crossover. B, The interface of the AI-based assistive tool illustrates localized region-level Gleason pattern interpretations as colored outlines overlaid on the tissue image. Green indicates Gleason pattern 3; yellow, Gleason pattern 4; and red, Gleason pattern 5 (not present). In the left toolbar, the AI-provided Gleason score, grade group, and Gleason pattern percentage are summarized, with toggles provided so that users can turn the visibility of several features on or off. Slide thumbnails allow users to quickly switch between multiple sections of the biopsy.

grade the biopsies as they would for a clinical review, without time constraints. Interaction with the AI assistive tool involved the information (eg, overall GG classification, quantitation of Gleason patterns, and Gleason pattern overlays) illustrated in Figure 1B. Overlay Gleason pattern outputs and AI confidence visualization could be toggled on and off, and the opacity could be adjusted. When biopsies were reviewed without AI assistance, the digital viewer continued to include all other tools and information, such as magnification level, serial sections, a marking tool, and a ruler. Additional details are available in the Biopsy Review and Classification section of eMethods in the Supplement.

Statistical Analysis

Prespecified primary analyses included GG agreement with the majority opinion of subspecialists for all cases and for GG1 cases alone. Grading performance was analyzed using the 2-sided Obuchowski-Rockette-Hillis procedure, which is a standard approach for multiple-reader, multiple-case studies that accounts for variance across both readers and cases.[28] Grade group 1 was selected as a focus of this study given the substantial clinical implications of misgrading these cases and the high interpathologist variability reported for these cases.[6]

For the analyses of review time and confidence, linear mixed-effects models were applied, which considered the individual pathologists and biopsies as random effects and the assistance modality and crossover arm as fixed effects. For mixed-effects models, *P* values were obtained using the likelihood ratio test. Agreement between pathologists and subspecialists was also measured using quadratic-weighted κ. Interobserver agreement for assisted vs unassisted reviews was measured by the Krippendorff α, which provides a measure of agreement among observers that is applicable to any number of raters.[29]

Confidence intervals were generated with the bootstrap method using 5000 replications without adjustment for multiple comparisons. Confidence interval calculations and the Obuchowski-Rockette-Hillis procedure were conducted using NumPy and SciPy packages in Python software, version 2.7.15 (Python Software Foundation). Analysis of the mixed-effects model was performed using the lme4 package in R software, version 3.4.1 (R Foundation for Statistical Computing). Additional details are available in the Statistical Analysis section of eMethods in the Supplement. Because we specified 2 primary end points, we conservatively prespecified the statistical significance threshold to .025, using Bonferroni correction, for these primary analyses.

Results

Study Data

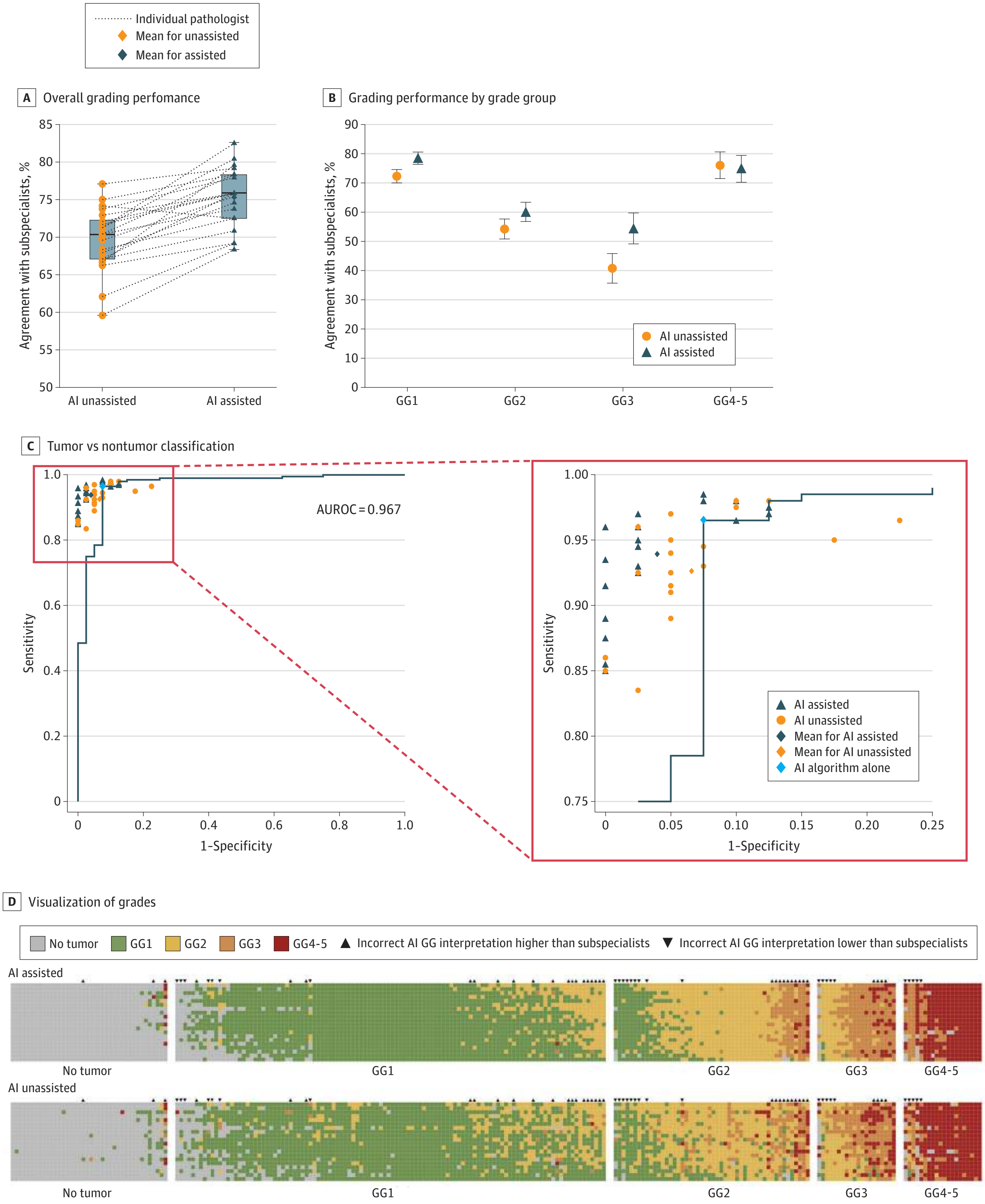
The study included 240 biopsies from 240 patients. At the time of biopsy, the median patient age was 67 years (range, 39-91 years), and the median prostate-specific antigen level was 6.5 ng/mL (range, 0.6-97.0 ng/mL [to convert to micrograms per liter, multiply by 1.0]) (Table 1). Based on the majority opinion of subspecialists for these biopsies, the data set included 40 biopsies with no tumors, 110 biopsies with GG1 tumors, 50 biopsies with GG2 tumors, 20 biopsies with GG3 tumors, and 20 biopsies with GG4-5 tumors.

Grading agreement among the urologic subspecialists for these cases is summarized in eTable 2 in the Supplement. Across 200 tumor-containing biopsies, 60 biopsies (30.0%) required a third review, and 140 (70.0%) did not require a third review.

Biopsy Grading

The use of the AI assistive tool was associated with increases in grading agreement between general pathologists and the majority opinion of subspecialists. The absolute increase in agreement for all 240 biopsies was 5.6% (95% CI, 3.2%-7.9%; *P* < .001), from 69.7% for unassisted reviews to 75.3% for assisted reviews (**Figure 2**A). The absolute increase in agreement for 110 GG1 biopsies was 6.2% (95% CI, 2.7%-9.8%; *P* = .001), from 72.3% for unassisted reviews to 78.5% for assisted reviews (Figure 2B). Among GG1 cases, this finding represents a relative 28.6% reduction in overgrading

Figure 2. Impact of Artificial Intelligence (AI)-Based Assistance for Prostate Biopsy Grading and Tumor Detection



A, Individual pathologist agreement with the majority opinion of subspecialists across all 240 biopsies. Dotted lines connect the points representing the same pathologist for each modality (assisted vs unassisted), and box-plot edges represent quartiles. B, Error bars represent 95% CIs. C, Circles and triangles represent sensitivities and specificities for each pathologist. The black line represents the receiver operating characteristic curve of the underlying deep learning system. D, Visualization of grades provided by all pathologists for all biopsies. Each colored box represents a grade for a single biopsy by a single pathologist. Each column represents a biopsy, and each row represents a pathologist. The greater number of solid-colored blocks in the AI-assisted plot illustrate the assistance-associated increases in interpathologist agreement and accuracy. AUROC indicates area under the receiver operating characteristic curve; GG, grade group.

(16.8% overgrading for unassisted reviews and 12.0% overgrading for assisted reviews). The full comparison of assisted and unassisted responses vs the majority opinion of subspecialists and the AI algorithm alone are presented in **Table 2** and eTable 3 in the Supplement; grading across all biopsies for all pathologists is also represented visually in Figure 2D. We did not observe an association between years of experience and the extent of the benefits provided by AI assistance (eFigure 2 in the Supplement). Analysis of the biopsies from ML2 alone (data source not used in development of the algorithm) had similar results as for the primary analysis across both data sources (eFigure 3 in the Supplement).

Assistance from the AI tool was also associated with increases in agreement for all biopsies when measured by quadratic-weighted κ (for unassisted reviews, κ = 0.80; 95% CI, 0.78-0.82; for assisted reviews, κ = 0.86; 95% CI, 0.84-0.87). For tumor-containing biopsies, the quadratic-weighted κ was 0.74 (95% CI, 0.71-0.76) for unassisted reviews and 0.81 (95% CI, 0.79-0.82) for assisted reviews (eTable 4 in the Supplement).

Artificial intelligence assistance was also associated with substantial improvement in interpathologist Gleason pattern quantitation agreement. For example, the standard deviation of pathologist Gleason pattern 4 quantitation for pattern 4–containing biopsies was 17.7% (95% CI, 15.7%-19.7%) for unassisted reviews and 8.1% (95% CI, 6.7%-9.4%) for assisted reviews (eTable 5 in the Supplement).

Correct vs Incorrect AI Interpretations

To evaluate the association between the performance of the underlying algorithm and the AI-assisted reviews, we also performed an analysis stratified by the correctness of the AI interpretations. We first analyzed the baseline GG classification performance of the unassisted pathologists. Unassisted pathologist performance was substantially lower on biopsies with incorrect AI interpretations (45.1%; 95% CI, 42.3%-47.9%) compared with biopsies with correct AI interpretations (78.1%; 95% CI, 76.7%-79.5%), suggesting that the biopsies with incorrect model interpretations were also challenging for the pathologists to interpret.

Next, we evaluated the association of incorrect assistance with grading. Among 179 biopsies for which the AI interpretation was correct, AI assistance was associated with increases in reader performance across all GGs. For 61 biopsies with incorrect AI interpretations, AI assistance was associated with decreases in reader agreement between pathologists and the majority opinion of subspecialists, from 45.1% (95% CI, 42.3%-47.9%) for unassisted reviews to 38.0% (95% CI, 35.4%-40.8%) for assisted reviews (eTable 6 in the Supplement). Among the subset of cases with incorrect

Table 2. Confusion Matrices (Contingency Tables) for Unassisted and Assisted Reviews Relative to the Subspecialist Reference Standard

	Pathologist grade group, %[a]				
Subspecialist reference standard grade group	No tumor	GG1	GG2	GG3	GG4-5
Pathologists with AI assistance					
No tumor	748 (15.6)	36 (0.8)	5 (0.1)	4 (0.1)	7 (0.1)
GG1	241 (5.0)	1590 (33.1)	347 (7.2)	19 (0.4)	3 (0.1)
GG2	14 (0.3)	260 (5.4)	542 (11.3)	143 (3.0)	41 (0.9)
GG3	14 (0.3)	15 (0.3)	128 (2.7)	163 (3.4)	80 (1.7)
GG4-5	28 (0.6)	2 (0.0004)	17 (0.4)	49 (1.0)	304 (6.3)
Pathologists without AI assistance					
No tumor	769 (16.0)	19 (0.4)	4 (0.1)	1 (0.0002)	7 (0.1)
GG1	207 (4.3)	1727 (36.0)	261 (5.4)	4 (0.1)	1 (0.0002)
GG2	10 (0.2)	601 (5.0)	601 (12.5)	129 (2.7)	20 (0.4)
GG3	12 (0.3)	123 (0.1)	123 (2.6)	218 (4.5)	42 (0.9)
GG4-5	16 (0.3)	0	2 (0.0004)	83 (1.7)	299 (6.2)

Abbreviations: AI, artificial intelligence; GG, grade group.

[a] Values are percentage of total readings across all biopsies for the indicated assistance modality (n = 4800 per assistance modality; 20 pathologists multiplied by 240 biopsies).

AI interpretations, AI assistance was associated with increases in interobserver agreement (for unassisted reviews, Krippendorff α = 0.56; for assisted reviews, Krippendorff α = 0.69).

Tumor Detection

For the binary task of tumor detection, performance was higher with AI assistance. The absolute increase in accuracy was 1.5% (95% CI, 0.6%-2.4%; *P* = .002), with an accuracy of 92.7% (95% CI, 92.0%-93.4%) for unassisted reviews, 94.2% (95% CI, 93.6%-94.9%) for assisted reviews, and 95.8% (95% CI, 93.3%-97.9%) for the AI algorithm alone.

Increases in both sensitivity and specificity were also observed (Figure 2C). The specificity for tumor detection was higher for assisted reviews (96.1%; 95% CI, 94.8%-97.4%) than for either unassisted reviews (93.5%; 95% CI, 91.8%-95.1%) or the AI algorithm alone (92.5%; 95% CI, 86.6%-97.8%). The AI algorithm generated false-positive final tumor interpretations for 3 biopsies; of those, 1 biopsy was associated with a small assistance-associated decrease in specificity, and 2 biopsies were associated with small assistance-associated increases in specificity (eTable 7 in the Supplement).

The highest sensitivity observed was for the algorithm alone (96.5%; 95% CI, 94.4%-98.3%), followed by assisted reviews (93.9%; 95% CI, 93.1%-94.7%) and unassisted reviews (92.6%; 95% CI, 91.8%-93.4%) (Figure 2C). Additional details and discussion are included in the Tumor Detection section of eResults in the Supplement. Examples of biopsies with AI assistance–associated changes in sensitivity or specificity are shown in eFigure 4 in the Supplement.

Review Time and Additional Analyses

The analysis of review time across GGs is summarized in **Table 3**. Overall, 13.5% less time was spent on assisted reviews (3.2 minutes; 95% CI, 3.2-3.3 minutes) vs unassisted reviews (3.7 minutes; 95% CI, 3.6-3.8 minutes; *P* = .006).

Additional analyses are summarized in eResults in the Supplement. These summaries include analyses of confidence (eFigure 5 and eTable 9 in the Supplement), interpathologist agreement among the 20 study pathologists (eTable 10 in the Supplement), tumor quantitation (eFigure 6 in the Supplement), and pathologist feedback (eFigure 7 in the Supplement).

Discussion

Several deep learning applications for Gleason grading of prostate biopsies have recently been described.[13-15] However, the evaluation of AI-based tools in the context of clinical workflows remains a largely unaddressed component in the translation of algorithms from code to clinical utility. In the present analysis, we evaluated an AI-based assistive tool via a fully crossed multiple-reader, multiple-case study. Use of the AI-based tool was associated with increases in the agreement between general pathologists and urologic subspecialists for Gleason grading and tumor detection. In addition, AI assistance was associated with increases in efficiency, interpathologist consistency, and pathologist confidence. To our knowledge, this work represents one of the largest studies to date with the aim of

Table 3. Mean Review Time per Biopsy With and Without Artificial Intelligence Assistance

		Mean time per biopsy (95% CI), min	
Category	Biopsies, No.	Unassisted	Assisted
All	240	3.7 (3.6-3.8)	3.2 (3.2-3.3)
No tumor	40	2.6 (2.5-2.8)	2.2 (2.1-2.3)
Grade group			
1	110	3.6 (3.5-3.7)	3.0 (2.9-3.1)
2	50	4.3 (4.1-4.5)	3.8 (3.7-4.0)
3	20	4.4 (4.1-4.7)	4.0 (3.8-4.3)
4-5	20	4.3 (4.0-4.5)	4.1 (3.8-4.4)

understanding the use of AI-based tools for concurrent review and interpretation of histopathologic images.

The observed benefit for patients with GG1 tumors has particular clinical relevance, as overgrading of these cases can result in overtreatment (eg, radical prostatectomy) rather than active surveillance.[2,30] Furthermore, most tumor-positive biopsy results in clinical practice are categorized as GG1 cases and represent a substantial portion of the more than 1 million total biopsies performed each year in the US alone.[31,32] Thus, improving the grading accuracy and consistency for this large number of biopsies has substantial implications for informing clinical decisions among patients with prostate cancer.

In this study, AI assistance was also associated with significant decreases in interobserver variability for Gleason pattern quantitation. Most notably, on GG2 biopsies in which Gleason pattern 4 quantitation has been reported to be prognostic in increments as small as 5%,[33] AI assistance was associated with substantial improvement in interpathologist quantitation agreement (eTable 5 in the Supplement). Such improvement in interobserver consistency may facilitate more reliable clinical decision-making and enable studies to more precisely define relevant quantitation thresholds for clinical management.

The use of AI assistance was also associated with decreases in the mean review time per case, with approximately 13% less time spent per biopsy. Possible explanations for the decreases in mean review time include more efficient quantitation, reduced time spent on Gleason pattern grading, and faster localization of small regions of interest. Notably, the increase in efficiency was not simply associated with overreliance, as the pathologists appeared able to disregard the AI interpretations in many cases, and performance was higher for AI-assisted reviews (75.3%) than for the AI algorithm alone (74.6%). Taken together, these results suggest that pathologists incorporated the interpretations from the AI assistive tool into their own diagnostic expertise, highlighting the potential of AI-assisted prostate biopsy grading to improve both the quality and efficiency of biopsy review without extensive overreliance.

Regarding the possibility of overreliance, the evaluation of incorrect AI interpretations provides additional insights. For biopsies with incorrect AI predictions, AI assistance was associated with decreased GG agreement with subspecialists (45.1% without assistance vs 38.0% with assistance; eTable 6 in the Supplement). The performance of unassisted pathologists was notably low for these cases, indicating that these particular biopsies were challenging to interpret for both the pathologists and the AI algorithm. For these difficult biopsies, AI assistance was associated with increases in interobserver agreement (the Krippendorff α was 0.56 for unassisted reviews vs 0.69 for assisted reviews), supporting the potential of AI assistance to improve interpathologist consistency, particularly with regard to the interpretation of challenging biopsies that otherwise have high grading variability. For tumor detection, a modest decrease in sensitivity was observed with AI assistance for the small number of biopsies with false-negative AI interpretations (eTable 8 in the Supplement). For specificity, among the 3 biopsies with false-positive AI results indicating the presence of tumor, 1 biopsy had a small assistance-associated decrease in specificity (eTable 7 in the Supplement). For the other 2 false-positive biopsy interpretations, the AI assistance was appropriately disregarded by the pathologists. Although the mean impact of AI assistance across all biopsies was positive, these findings do suggest the important possibility that incorrect AI interpretations may result in incorrect tumor identification in some cases. Understanding error modes and designing clinical applications to mitigate potential overreliance remain important challenges to address.

Providing information to inform decisions about when to rely on AI (and when not to) has important implications for maximizing benefit and minimizing automation bias. We conducted extensive human-computer interaction research to incorporate the information that was most important and useful to pathologists. Notable insights included the need to establish sufficient trust in the AI assistive tool, the desire for an explanation of the AI interpretations (eg, why the AI algorithm made the interpretation it did), and requests for information about how the AI assistive tool was developed and tested. Results of these efforts informed the final user interface as well as the development of visualizations for the AI interpretations and the training materials used in the study.

A recent article from Bulten et al[21] also described a study of an AI-based assistive tool for prostate biopsy review. In their study, the researchers similarly observed that AI assistance was associated with increases in agreement between general pathologists and subspecialists. Both the Bulten et al[21] study and the present study provide important and distinct insights. For example, Bulten et al[21] described an interesting association between pathologist experience and the benefit of AI assistance, and our study provides analysis stratified by GG as well as data regarding review time, confidence, and interpathologist agreement. Taken together, these studies complement each other and may initiate useful discussions about implementation and design considerations, such as the benefits of AI to provide second readings vs concurrent reviews or the importance of different user interface elements to maximize the usefulness of the AI interpretations.

Optimal AI integration into pathologic clinical practice will depend on several factors, including the strengths of the specific tool, the needs of the practice, and the clinical workflow. For example, a highly sensitive algorithm for cancer detection might be best used for triage or as a second reading tool to avoid missing a tumor. Concurrent review instruments such as the present AI assistive tool, which provides GG interpretation and quantitation interpretations, might be optimal for use in community practice settings in which second opinions and the expertise of urologic specialists may be less readily available for challenging cases. This value may extend to improved calibration of pathologists across diverse practice settings, especially if the underlying models can be kept accurate and representative of current grading guidelines.

Limitations

This study has several limitations. First, although multiple serial hematoxylin and eosin sections were provided for review, only 1 core biopsy per case was available for review. Thus, the impact of AI assistance in the context of multiple cores per case was not addressed. Second, this study is a retrospective review of biopsies in a nonclinical setting, without additional clinical information available at the time of review. In addition, population demographic characteristics were not available for this study. Future validation among diverse patient populations is an important consideration to address the risk of unintended population biases. Validation in clinical settings that represent the real-world distribution of GGs, tumor-containing cases, and preanalytical variability will also be important to further inform our understanding of potential diagnostic benefits. Third, the reference standard used in this study was based on the majority opinion of multiple urologic subspecialists with extensive experience in the grading of prostate biopsies; however, even among subspecialists, Gleason grading remains a task with considerable interobserver disagreement. Future evaluation of deep learning systems and AI-based assistance for cancer grading will benefit from reference standards that are based on both clinical outcomes and expert review.

Conclusions

This diagnostic study indicated the potential ability of an AI-based assistive tool to improve the accuracy, efficiency, and consistency of prostate biopsy review by pathologists. The relatively large number of biopsies and pathologists included in the study allowed for a robust analysis of the benefits of an AI-based tool for the concurrent review of prostate biopsies and provided insights into caveats regarding overreliance, which may only have been apparent owing to the opportunity to observe infrequent occurrences in a large study. Additional efforts to optimize clinical workflow integration and to conduct prospective evaluation of AI-based tools in clinical settings remain important future directions.

ARTICLE INFORMATION

Accepted for Publication: August 10, 2020.

Published: November 12, 2020. doi:10.1001/jamanetworkopen.2020.23267

Correction: This article was corrected on December 15, 2020, to fix Dr Stumpe's affiliations in the Author Affiliations section.

Corresponding Authors: David F. Steiner, MD, PhD (davesteiner@google.com), and Craig H. Mermel, MD, PhD (cmermel@google.com), Google Health, 3400 Hillview Ave, Palo Alto, CA 94304.

Author Affiliations: Google Health, Palo Alto, California (Steiner, Nagpal, Sayres, Foote, Wedin, Pearce, Cai, Winter, Symonds, Yatziv, Kapishnikov, Tan, Stumpe, Jiang, Liu, Chen, Corrado, Terry, Mermel); Google Health via Advanced Clinical, Deerfield, Illinois (Brown, Flament-Auvigne); Now with Tempus Labs, Chicago, Illinois (Stumpe).

Author Contributions: Drs Terry and Mermel contributed equally and are considered co-senior authors. Drs Steiner and Mermel had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Steiner, Nagpal, Sayres, Foote, Wedin, Pearce, Cai, Winter, Kapishnikov, Brown, Flament-Auvigne, Stumpe, Liu, Chen, Corrado, Terry, Mermel.

Acquisition, analysis, or interpretation of data: Steiner, Nagpal, Sayres, Foote, Wedin, Winter, Symonds, Yatziv, Kapishnikov, Tan, Jiang, Liu, Chen.

Drafting of the manuscript: Steiner, Nagpal, Sayres, Foote, Cai, Winter, Symonds, Yatziv, Liu, Chen, Mermel.

Critical revision of the manuscript for important intellectual content: Steiner, Nagpal, Sayres, Wedin, Pearce, Winter, Kapishnikov, Brown, Flament-Auvigne, Tan, Stumpe, Jiang, Liu, Chen, Corrado, Terry.

Statistical analysis: Steiner, Nagpal, Sayres, Foote, Wedin, Chen.

Obtained funding: Corrado, Mermel.

Administrative, technical, or material support: Foote, Wedin, Pearce, Symonds, Yatziv, Brown, Flament-Auvigne, Tan, Stumpe, Jiang, Liu, Chen, Terry, Mermel.

Supervision: Steiner, Stumpe, Liu, Corrado, Mermel.

Conflict of Interest Disclosures: Dr Steiner, Mr Nagpal, and Dr Sayres reported owning shares of Alphabet and having a patent pending for artificial intelligence–assistance components of this research during the conduct of the study. Mr Wedin reported owning shares of Alphabet and having a patent pending for artificial intelligence–assistance components of this research during the conduct of the study. Mr Pearce reported owning shares of Alphabet and having a patent pending for artificial intelligence–assistance components of this research during the conduct of the study. Dr Cai reported owning shares of Alphabet and having a patent pending for artificial intelligence–assistance components of this research during the conduct of the study. Mr Symonds reported owning shares of Alphabet and having a patent pending for an artificial intelligence–based assistive tool for concurrent review of core needle prostate biopsies during the conduct of the study. Dr Yatziv reported owning shares of Alphabet and having a patent pending for artificial intelligence–assistance components of this research during the conduct of the study. Mr Kapishnikov reported owning shares of Alphabet and having a patent pending for an artificial intelligence–based assistive tool for concurrent review of core needle prostate biopsies during the conduct of the study. Dr Flament-Auvigne reported receiving personal fees from Advanced Clinical during the conduct of the study. Dr Tan reported owning shares of Google outside the submitted work. Dr Stumpe reported receiving personal fees from Google and Tempus Labs; owning shares of Alphabet and Tempus Labs; and submitting and publishing several patents related to machine learning, particularly digital pathology, during the conduct of the study. Dr Liu reported owning shares of Alphabet and having multiple patents in various stages for machine learning for medical images outside the submitted work. Dr Chen reported owning shares of Alphabet and having a patent pending for an artificial intelligence–based assistive tool for concurrent review of core needle prostate biopsies during the conduct of the study. Dr Corrado reported having a patent pending for artificial intelligence assistance on prostate biopsy review and Gleason grading. Dr Terry reported owning shares of Alphabet and having a patent pending for artificial intelligence–assistance components of this research during the conduct of the study. Dr Mermel reported owning shares of Alphabet and having a patent pending for an artificial intelligence–based assistive tool for concurrent review of core needle prostate biopsies during the conduct of the study. No other disclosures were reported.

Funding/Support: This study was supported by Google Health.

Role of the Funder/Sponsor: Google Health was involved in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: Yuannan Cai, Can Kirmizibayrak, Hongwu Wang, Allen Chai, Melissa Moran, Angela Lin, and Robert MacDonald of the Google Health pathology team provided software infrastructure, slide digitization, and data collection support. Andy Coenen, Jimbo Wilson, and Qian Yang provided important insights and contributions on user experience research and design. Scott McKinney provided analysis review and guidance. Paul Gamble, Naama Hammel, and Michael Howell reviewed the manuscript. None of these individuals (all Google employees) received financial compensation outside of their normal salary. This work would not have been possible without the contributions and expertise of the many pathologists who participated in the various components of this work.

REFERENCES

1. Epstein JI, Allsbrook WC Jr, Amin MB, Egevad LL; ISUP Grading Committee. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *Am J Surg Pathol*. 2005;29(9):1228-1242. doi:10.1097/01.pas.0000173646.99337.b1

2. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA; Grading Committee. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: definition of grading patterns and proposal for a new grading system. *Am J Surg Pathol*. 2016;40(2): 244-252. doi:10.1097/PAS.0000000000000530

3. Mohler JL, Antonarakis ES, Armstrong AJ, et al. Prostate cancer, version 2.2019, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw*. 2019;17(5):479-505. doi:10.6004/jnccn.2019.0023

4. Veloso SG, Lima MF, Salles PG, Berenstein CK, Scalon JD, Bambirra EA. Interobserver agreement of Gleason score and modified Gleason score in needle biopsy and in surgical specimen of prostate cancer. *Int Braz J Urol*. 2007;33(5):639-646. doi:10.1590/S1677-55382007000500005

5. Ozdamar SO, Sarikaya S, Yildiz L, Atilla MK, Kandemir B, Yildiz S. Intraobserver and interobserver reproducibility of WHO and Gleason histologic grading systems in prostatic adenocarcinomas. *Int Urol Nephrol*. 1996;28(1):73-77. doi:10.1007/BF02550141

6. Egevad L, Ahmad AS, Algaba F, et al. Standardization of Gleason grading among 337 European pathologists. *Histopathology*. 2013;62(2):247-256. doi:10.1111/his.12008

7. Allsbrook WC Jr, Mangold KA, Johnson MH, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. *Hum Pathol*. 2001;32(1):74-80. doi:10.1053/hupa.2001.21134

8. Melia J, Moseley R, Ball RY, et al. A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology*. 2006;48(6):644-654. doi:10.1111/j.1365-2559.2006.02393.x

9. Abdollahi A, Meysamie A, Sheikhbahaei S, et al. Inter/intra-observer reproducibility of Gleason scoring in prostate adenocarcinoma in Iranian pathologists. *Urol J*. 2012;9(2):486-490.

10. Kvale R, Moller B, Wahlqvist R, et al. Concordance between Gleason scores of needle biopsies and radical prostatectomy specimens: a population-based study. *BJU Int*. 2009;103(12):1647-1654. doi:10.1111/j.1464-410X.2008.08255.x

11. Bottke D, Golz R, Storkel S, et al. Phase 3 study of adjuvant radiotherapy versus wait and see in pT3 prostate cancer: impact of pathology review on analysis. *Eur Urol*. 2013;64(2):193-198. doi:10.1016/j.eururo.2013.03.029

12. van der Kwast TH, Collette L, Van Poppel H, et al; European Organisation for Research and Treatment of Cancer Radiotherapy and Genito-Urinary Cancer Groups. Impact of pathology review of stage and margin status of radical prostatectomy specimens (EORTC trial 22911). *Virchows Arch*. 2006;449(4):428-434. doi:10.1007/s00428-006-0254-x

13. 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(2):233-241. doi:10.1016/S1470-2045(19)30739-9

14. Strom 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

15. 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(1):48. doi:10.1038/s41746-019-0112-2

16. Litjens G, Sanchez 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

17. 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

18. Veta M, van Diest PJ, Willems SM, et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med Image Anal*. 2015;20(1):237-248. doi:10.1016/j.media.2014.11.010

19. Tellez D, Balkenhol M, Otte-Holler I, et al. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Trans Med Imaging*. 2018;37(9):2126-2136. doi:10.1109/TMI.2018.2820199

20. Gavrielides MA, Miller M, Hagemann IS, et al. Clinical decision support for ovarian carcinoma subtype classification: a pilot observer study with pathology trainees. *Arch Pathol Lab Med*. 2020;144(7):869-877. doi:10.5858/arpa.2019-0390-OA

21. Bulten W, Balkenhol M, Belinga JA, et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod Pathol*. Published online August 5, 2020. https://arxiv.org/abs/2002.04500

22. Kiani A, Uyumazturk B, Rajpurkar P, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit Med*. 2020;3:23. doi:10.1038/s41746-020-0232-8

23. Nagpal K, Foote D, Tan F, et al. Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. *JAMA Oncol*. 2020;6(9):1372-1380. doi:10.1001/jamaoncol.2020.2485

24. Epstein JI, Zelefsky MJ, Sjoberg DD, et al. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. *Eur Urol*. 2016;69(3):428-435. doi:10.1016/j.eururo.2015.06.046

25. Molin J, Wozniak PW, Lundstrom C, Treanor D, Fjeld M. Understanding design for automated image analysis in digital pathology. In: *NordiCHI '16: Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. Association for Computing Machinery; 2016;58:1-10. https://dl.acm.org/doi/10.1145/2971485.2971561

26. Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. In: *Proceedings of the ACM on Human-Computer Interaction*. Vol 2, No. CSCW. Association for Computing Machinery; 2019;104:1-24. doi:10.1145/3359206

27. Gordetsky J, Epstein J. Grading of prostatic adenocarcinoma: current state and prognostic implications. *Diagn Pathol*. 2016;11:25. doi:10.1186/s13000-016-0478-2

28. Obuchowski NA Jr, Rockette HE Jr. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an ANOVA approach with dependent observations. *Commun Stat Simul Comput*. 1995;24(2):285-308. doi:10.1080/03610919508813243

29. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas*. 2007;1(1):77-89. doi:10.1080/19312450709336664

30. Chen RC, Rumble RB, Jain S. Active surveillance for the management of localized prostate cancer (Cancer Care Ontario guideline): American Society of Clinical Oncology Clinical Practice Guideline endorsement summary. *J Oncol Pract*. 2016;12(3):267-269. doi:10.1200/JOP.2015.010017

31. Shah N, Ioffe V, Cherone S. Prostate biopsy features: a comparison between the pre- and post-2012 United States Preventive Services Task Force prostate cancer screening guidelines with emphasis on African American and septuagenarian men. *Rev Urol*. 2019;21(1):1-7.

32. Kearns JT, Holt SK, Wright JL, Lin DW, Lange PH, Gore JL. PSA screening, prostate biopsy, and treatment of prostate cancer in the years surrounding the USPSTF recommendation against prostate cancer screening. *Cancer*. 2018;124(13):2733-2739. doi:10.1002/cncr.31337

33. Sauter G, Steurer S, Clauditz TS, et al. Clinical utility of quantitative Gleason grading in prostate biopsies and prostatectomy specimens. *Eur Urol*. 2016;69(4):592-598. doi:10.1016/j.eururo.2015.10.029

SUPPLEMENT.

eMethods. Study Data, Digital Assistant Design and Development, Pathologist Training and Onboarding, Study Design, Biopsy Review and Classification, and Statistical Analysis
eResults. ML2-Only Subset Analysis; Tumor Detection–Supplemental Results; Algorithm-Only Performance; Impact of Correct vs Incorrect Assistant Prediction; Biopsy Review Time; Grading Confidence; Interpathologist Agreement; Consistency of Tumor and Gleason Pattern Quantitation; Pathologist Experience; and Biopsies With the Largest Assistance-Associated Impact on Accuracy, Confidence, and Review Time
eTable 1. Self-reported Prostate Biopsy Review Volume of Pathologists Participating in This Study
eTable 2. Subspecialist Concordance on the Biopsies Used in This Study
eTable 3. Complete Performance Results by Grade Group for Unassisted and Assisted Reviews as Well as the Stand-Alone AI Assistant Interpretation
eTable 4. Kappa Agreement With Subspecialist Majority by Assistance Modality

eTable 5. Interpathologist Variability for Quantitation of Gleason Pattern 4 (GP4) and Total Tumor Involvement
eTable 6. Pathologist Review Performance Stratified by AI Assistant Performance
eTable 7. Individual Biopsies With Assistance-Associated Changes in Tumor Detection Specificity
eTable 8. Individual Biopsies With Assistance-Associated Changes in Tumor Detection Sensitivity
eTable 9. Average Pathologist-Reported Confidence per Biopsy (on a 10-Point Scale) Stratified by Biopsies With Correct vs Incorrect Assistant Interpretations
eTable 10. Interpathologist Agreement for Assisted vs Unassisted Reviews
eFigure 1. Assistant Confidence Visualization
eFigure 2. Analysis of Effect Size and Pathologist Experience
eFigure 3. Agreement Analysis Across Grade Groups for Assisted vs Unassisted Reviews for Only the ML2 Data Source Biopsies
eFigure 4. Example Biopsies With Assistance-Associated Changes in Sensitivity or Specificity for Detecting Prostatic Adenocarcinoma
eFigure 5. Average Review Times and Confidence for Biopsies Stratified by Grade Group (GG) and Assistance Modality
eFigure 6. Assistance Increases Agreement for Quantitation of Gleason Pattern 4 and Percentage of Tissue Involved by Tumor
eFigure 7. Pathologist Perceptions and Feedback Regarding the Assistant Tool
eReferences