ARTICLE

Artificial intelligence system shows performance at the level of uropathologists for the detection and grading of prostate cancer in core needle biopsy: an independent external validation study

Minsun Jung[1,8], Min-Sun Jin[2,8], Chungyeul Kim[3], Cheol Lee[4], Ilias P. Nikas[5], Jeong Hwan Park[4,6] and Han Suk Ryu[4,7]✉

Accurate diagnosis and grading of needle biopsies are crucial for prostate cancer management. A uropathologist-level artificial intelligence (AI) system could help make unbiased decisions and improve pathologists' efficiency. We previously reported an artificial neural network-based, automated, diagnostic software for prostate biopsy, DeepDx® Prostate (DeepDx). Using an independent external dataset, we aimed to validate the performance of DeepDx at the levels of prostate cancer diagnosis and grading and evaluate its potential value to the general pathologist. A dataset composed of 593 whole-slide images of prostate biopsies (130 normal and 463 adenocarcinomas) was assembled, including their original pathology reports. The Gleason scores (GSs) and grade groups (GGs) determined by three uropathology experts were considered as the reference standard. A general pathologist conducted user validation by scoring the dataset with and without AI assistance. DeepDx was accurate for prostate cancer detection at a similar level to the original pathology report, whereas it was more concordant than the latter with the reference GGs and GSs (kappa/quadratic-weighted kappa = 0.713/0.922 vs. 0.619/0.873 for GGs and 0.654/0.904 vs. 0.576/0.858 for GSs). Notably, it outperformed the original report, especially in the detection of Gleason patterns 4/5, and achieved excellent agreement in quantifying the Gleason pattern 4. When the general pathologist used AI assistance, the concordance of GG between the user and the reference standard increased (kappa/quadratic-weighted kappa, 0.621/0.876 to 0.741/0.925), while the average slide examination time was substantially decreased (55.7 to 36.8 s/case). Overall, DeepDx was capable of making expert-level diagnosis in prostate core biopsies. In addition, its remarkable performance in detecting high-grade Gleason patterns and enhancing the general pathologist's diagnostic performance supports its potential value in routine practice.

Modern Pathology (2022) 35:1449–1457; https://doi.org/10.1038/s41379-022-01077-9

INTRODUCTION

Prostate cancer is the 2nd most common malignancy in males worldwide[1]. Accurate diagnosis and grading of core needle biopsies are crucial for managing patients with prostate cancer[2]. However, the reported high inter-observer variability in prostate cancer Gleason grading may lead to suboptimal therapy decisions[3]. In the 2019 International Society of Urological Pathology (ISUP) Consensus Conference, the ISUP members suggested that artificial intelligence (AI) will play a role in prostate cancer screening by supporting decision-making and improving efficiency of pathologists[4]. It was also emphasized that an AI algorithm first needs to reach the diagnostic level of uropathologists before its future implementation into the routine practice[4].

Several studies for AI detecting and grading prostate cancer have been published since the first article released in 2016[5–7]. However, whether AI can diagnose and grade prostate cancer at the level of a uropathologist is still largely unknown. We previously demonstrated that an AI system for prostate biopsy showed an excellent performance for detecting, grading, and measuring tumor length of prostate cancer, using whole slides images (WSIs) of 700 core biopsies as an internal test set[8]. However, based on the recent ISUP consensus, any developed AI algorithm for prostate core biopsies should exhibit a comparable diagnostic accuracy with expert uropathologists through implementing an independent external validation[4].

In this study, we validated the performance of the AI for detecting and grading prostate cancer, using an independent

[1]Department of Pathology, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea. [2]Department of Pathology, Bucheon St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Bucheon, Republic of Korea. [3]Department of Pathology, Korea University Guro Hospital, Seoul, Republic of Korea. [4]Department of Pathology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Republic of Korea. [5]School of Medicine, European University Cyprus, Nicosia, Cyprus. [6]Department of Pathology, SMG-SNU Boramae Medical Center, Seoul, Republic of Korea. [7]Center for Medical Innovation, Biomedical Research Institute, Seoul National University Hospital, Seoul, Republic of Korea. [8]These authors contributed equally: Minsun Jung, Min-Sun Jin. ✉email: karlnash@naver.com

Received: 11 December 2021 Revised: 27 March 2022 Accepted: 28 March 2022
Published online: 29 April 2022

external dataset. To investigate its potential impact in everyday diagnostic practice, we also compared the performance of a general pathologist with or without AI support. We found that the AI performed at the level of uropathologists for the detection and scoring of prostate cancer. Thus, it could help pathologists to process prostate cancer core biopsies more effectively.

MATERIALS AND METHODS

Dataset assembly

The artificial neural network-based, automated, deep learning system for prostate biopsy diagnosis used in this study (DeepDx® Prostate [DeepDx]; Deep Bio Inc., Seoul, Korea) has been previously presented by our group (Supplementary Fig. 1)[8]. To validate the diagnostic efficacy of DeepDx (clinical trial no. DPB-Prostate-02), we obtained hematoxylin and eosin-stained prostate needle biopsy slides processed between 2018 and 2019 (one slide per patient), including their original pathology reports, from the Department of Pathology of the Seoul National University Hospital (SNUH). Immunohistochemical (IHC) staining, whenever addressed in the original reports, was also retrieved. The cases were randomly selected among the biopsies that all slides were available. This cohort was independent from that used for the development of DeepDx[8]. All glass slides were deidentified and then digitally scanned at ×400 magnification and 0.2535 μm/pixel resolution using the Aperio AT2 scanner (Leica Biosystems, Wetzlar, Germany). Three experts in uropathology (M.-S.J., C.K., and H.S.R.), each having more than 10 years of experience in uropathology, independently reviewed the cases for the presence or absence of cancer and assigned a Gleason score (GS) and grade group (GG) for each malignant case, using an identical hardware platform (Gram 17″ Ultra-Lightweight Laptop, LG Electronics, Seoul, Republic of Korea) to examine the WSIs. The GSs were defined by the sum of the most prevalent and the worst Gleason patterns and ranged from GS 3 + 3 to GS 5 + 5, according to the latest WHO classification and ISUP guidelines[9]. Gleason patterns ranged from 3 (well-differentiated) to 5 (poorly-differentiated). GGs were assigned as follows: GG1 (GS 3 + 3), GG2 (GS 3 + 4), GG3 (GS 4 + 3), GG4 (GS 4 + 4, 3 + 5, or 5 + 3), and GG5 (GS 4 + 5, 5 + 4, or 5 + 5)[9]. The Institutional Review Board of SNUH approved this study (IRB no. D-2006-105-1134).

Establishment of reference standards

Based on the sample size estimation and a margin of 10%[8, 10], we initially collected 594 biopsy cores. The three expert reviewers classified the samples into one of the following eight entities: normal, high-grade prostate intraepithelial neoplasm (HPIN), atypical small acinar proliferation (ASAP), and GG1 to GG5 prostate adenocarcinomas. If adenocarcinoma and HPIN or ASAP were found together in a prostate biopsy, the reviewers classified it as adenocarcinoma, and if HPIN and ASAP were together in a prostate biopsy, they classified it as ASAP. After the initial review, the majority vote from the three experts determined the reference standard among these entities. To determine the reference group for the cases all experts provided a different interpretation, a consensus meeting was held. When available, the experts referred to IHC staining for a basal cell cocktail (34βE12 + p63) (Ventana, Oro Valley, AZ) and AMACR (Agilent, Santa Clara, CA). In detail, the IHC stains used to diagnose two cases, as addressed in their original reports, were provided. Additionally, IHC staining was conducted for six more cases, and results were taken into consideration during the consensus meeting. During this process, one specimen that remained nonconsensual even with IHC stains was withdrawn, finalizing 593 biopsy cores for this study. The ultimate reference standard fell into one of the six groups (normal and GGs 1-5), while no HPIN or ASAP cases were present (Supplementary Table 1).

Analysis of cancer detection and grade concordance

Supplementary Fig. 2 shows the overall study design. We compared the results of DeepDx to the original pathology reports for cancer detection, GGs, and GSs. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated for the detection of cancer (cancer vs. normal) and high-grade Gleason patterns (present vs. absent). The concordance of GG and GS was evaluated using Cohen's kappa coefficient with and without quadratic weighting[11]. The correlation of GS was also evaluated using a Spearman's rank test. The illustrations were made using Prism software ver. 8 (GraphPad, San Diego, CA) or R (R Foundation for Statistical Computing, Vienna, Austria). A two-tailed $P < 0.05$ was considered significant.

Gleason pattern 4 measurement

For all GG2 and GG3 cancers ($n = 202$), one expert reviewer (H.S.R.) semi-quantified Gleason pattern 4, as 0–5%, 5–10%, 10–25%, 25–50%, 50–75%, or 75–100%, after modifying a previous suggestion[12]. The correlation of the stratified proportion of Gleason pattern 4 between the reviewer and DeepDx was examined using kappa and quadratic-weighted kappa.

Area annotation for Gleason patterns

Ten adenocarcinomas, two cores for each GG were randomly selected. One expert reviewer (H.S.R.) annotated the glands as benign or Gleason patterns 3–5. The ratio of each pattern was assessed at 0.5×0.5 mm^2-area patches, which correspond to a ×200 magnification field under light microscopy. The correlation of the pattern annotations was evaluated using the Pearson's correlation coefficient.

User test validation of DeepDx assistance

One board-certified pathologist (M.J.), who had <3 years of experience in uropathology while being a 1st-time user of DeepDx, blindly examined the final dataset twice, following the aforementioned procedure, without referring to the IHC staining. During the 1st session, the examiner completed the dataset without any AI assistance and the case order was subsequently randomized. Four weeks later, the investigator examined the dataset with AI assistance; the AI annotations of distinct Gleason patterns, their proportion, and the finalized GS were accessible for all cases. The time spent to complete every set of 30 cases was recorded for both readings. One case that was classified as ASAP by the user pathologist in the 2nd reading was omitted from the concordance analysis because: (1) ASAP is a borderline category that often needs IHC staining for diagnosis[13], while this was not available to the user pathologist and (2) it is unclear how to determine the concordance between normal/GG1 and ASAP.

RESULTS

DeepDx is highly accurate in the detection and grading of prostate cancer biopsies

We tested DeepDx and the original reports against the reference standard that was established from the review by the three uropathology experts. DeepDx showed excellent sensitivity, specificity, PPV, NPV, and accuracy for cancer detection, which were comparable or superior to those of the original report (Fig. 1A and Supplementary Table 2). In the analysis of the concordance of GG using kappa and quadratic-weighted kappa (Supplementary Table 3), DeepDx (0.713 and 0.922, respectively; Fig. 1B, red) achieved higher agreement to the reference standard than the original report did (0.619 and 0.873, respectively; Fig. 1B, green). The kappa and quadratic-weighted kappa values varied between the initial reviewer results (average, 0.524 and 0.810, respectively) (Fig. 1B, black), suggesting there was discordance in prostate biopsy diagnosis even between uropathology experts. In contrast, DeepDx consistently showed high kappa and quadratic-weighted kappa with the reference standard (average, 0.603 and 0.873, respectively) (Fig. 1B, purple). From these observations, we hypothesized that objective and unbiased decisions made by DeepDx could help to diagnose disputable prostate biopsies. We tested this hypothesis by examining DeepDx within the cases that one ($n = 268$) or all experts ($n = 66$) initially disagreed during the review (Supplementary Table 3), wherein DeepDx maintained excellent concordance of GG with the reference standard, compared to the original pathology report (Fig. 1C, D). These results suggested that DeepDx results are robust and reproducible, even when dealing with "hard" prostate biopsies, offering a potential solution to the high inter-observer variability found even among experts.

DeepDx is useful for the detection of the high-grade Gleason patterns

We next determined the accuracy of DeepDx for detecting each GG. DeepDx showed higher sensitivity than the original report to identify normal and cancer tissues, except for the GG1 (Fig. 2A, B). In a further analysis of GSs, DeepDx demonstrated significant

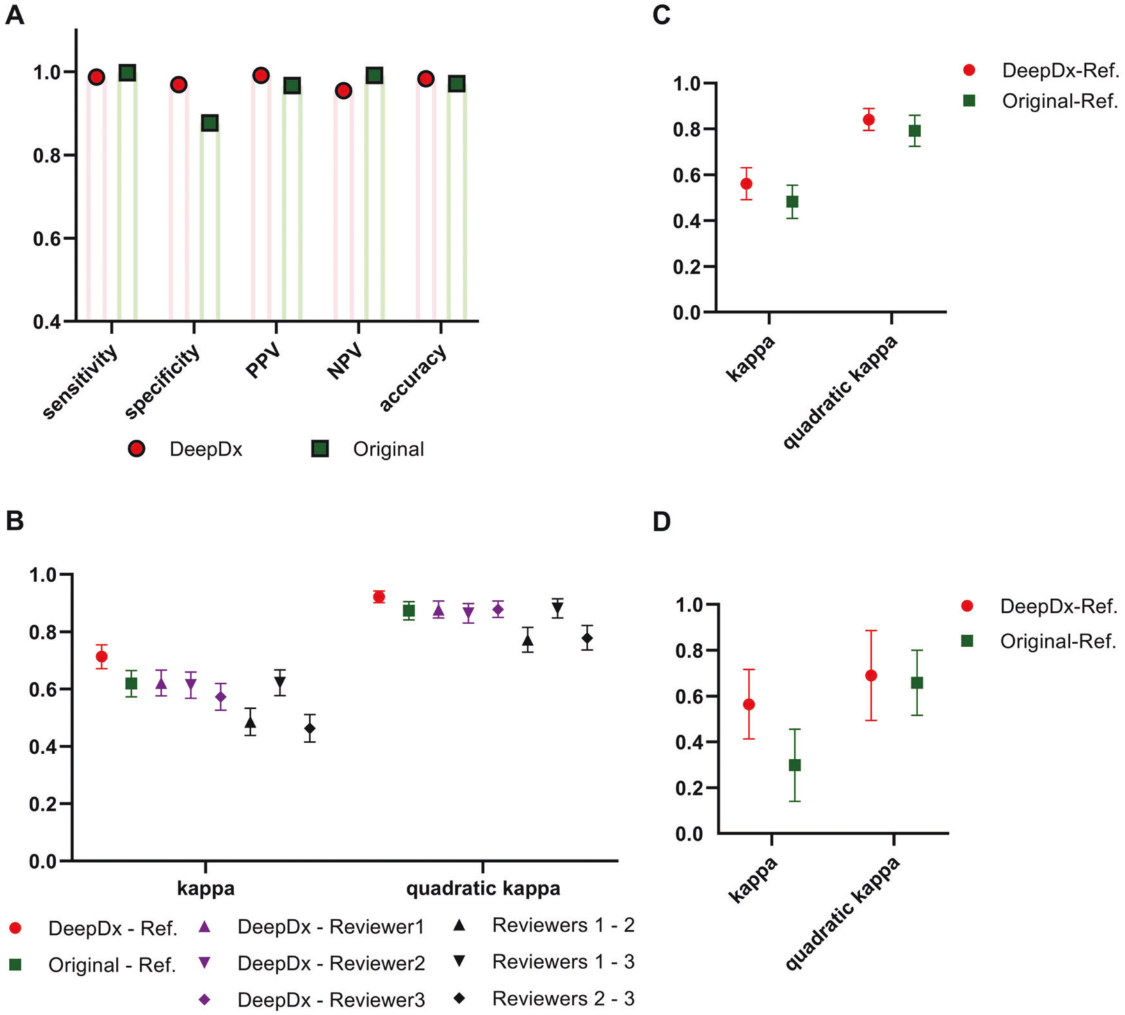


Fig. 1 The overall performance of DeepDx. A The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy for prostate cancer detection by DeepDx and the original hospital report. **B** The concordance of Gleason grade group (GG) by each subject. **C** The concordance within the cases in which GGs assigned differently by one expert reviewer compared to the other two ($n = 258$). **D** The concordance within cases in which GGs were assigned differently by all three expert reviewers ($n = 66$). The error bars indicate 95% confidence intervals. The detailed values are summarized in Supplementary Tables 2, 3.

correlation with the reference standard (kappa/quadratic-weighted kappa = 0.654/0.904; Spearman rho = 0.938, $P < 0.0001$) (Fig. 2C), which was higher than that of the original pathology report (kappa/quadratic-weighted kappa = 0.576/0.858; Spearman rho = 0.879, $P < 0.0001$) (Fig. 2D). Among adenocarcinomas, DeepDx tended to predict Gleason patterns 4/5, resulting in an excellent detection of high GS cancers, yet an exaggeration of low GS cancers (Fig. 2C), compared to the original report (Fig. 2D). Within the normal group, DeepDx was also more accurate than the original report (Fig. 2C, D).

Patients with high-grade prostate adenocarcinoma, or Gleason patterns 4/5, are generally not eligible for active surveillance[14]. We determined the ability of DeepDx to recognize high-grade patterns (Supplementary Table 4). In the detection of GGs 2–5 (GS 3 + 4 or more) among all cases (Fig. 3A, top), GGs 2–5 (GS 3 + 4 or more) among cancers (Fig. 3A, middle), and Gleason pattern 5 among all cases (Fig. 3A, bottom), its sensitivity and NPV exceeded those of the original report. Thus, DeepDx could offer a meticulous screening of high-grade prostate cancer. We further explored how precisely DeepDx quantified the Gleason pattern 4. DeepDx tended to identify it more often than the expert reviewer (H.S.R.) (Fig. 3B); however, the overall concordance was good (kappa/quadratic-weighted kappa = 0.770/0.940). Lastly, we examined Gleason patterns annotated by DeepDx at the gland levels (Fig. 3C). In ten randomly-selected cases, the Gleason patterns recognized by DeepDx were similar to those identified by the reviewer, especially for normal glands and Gleason patterns 4/5 (Fig. 3D and Supplementary Fig. 3). Collectively, the data showed that DeepDx was highly accurate for the detection of high-grade prostate cancer.

DeepDx assistance improved the performance of grading prostate cancer and time efficiency

Supplementary Table 2 summarizes the user pathologist's results without (UT1) and with DeepDx assistance (UT2). In line with the algorithm's credibility for high-grade pattern recognition, the diagnostic accuracy was improved in UT2, especially for identifying the benign cases and GGs 4/5 cancers (Fig. 4A), which resulted in significant enhancement in the concordance of GG in UT2, compared to UT1 (kappa/quadratic-weighted kappa, 0.925/0.741 vs. 0.621/0.876) (Fig. 4B). The average time consumed for diagnosis was substantially reduced as well, from 55.7 s/case without DeepDx assistance to 36.8 s/case with DeepDx assistance (Fig. 4C). From these results, we infer that DeepDx could enable pathologists to diagnose prostate cancer more precisely and efficiently.

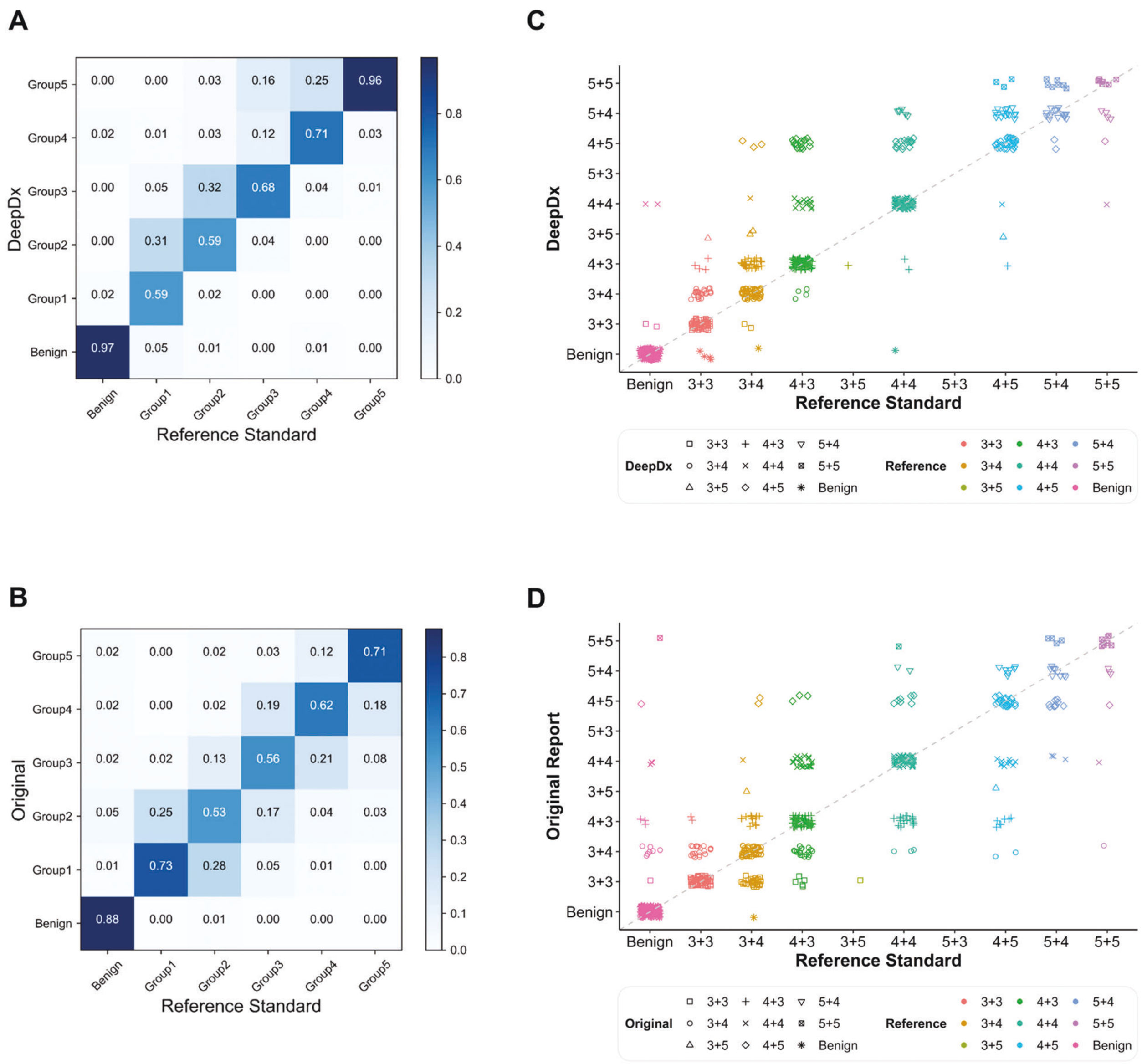


Fig. 2 **Grading and scoring of prostate cancer. A** The sensitivity of DeepDx for each Gleason grade group (GG). **B** The sensitivity of the original report for each GG. **C** The concordance of Gleason scores (GSs) between DeepDx and the reference standard (kappa/quadratic-weighted kappa = 0.654/0.904; Spearman rho = 0.938, $P < 0.0001$). **D** The concordance of GSs between the original report and the reference standard (kappa/quadratic-weighted kappa = 0.576/0.858; Spearman rho = 0.879, $P < 0.0001$).

We further investigated how DeepDx influenced the scoring of prostate cancer by recognizing GS results being: (1) UT1 discordant/both UT2 and DeepDx concordant to the reference standard and 2) UT1 concordant/UT2 and DeepDx same but discordant to the reference standard. From the total of 593 cases, 98 (16.5%) and 52 (8.8%) cases met these criteria, respectively (Fig. 5A). Along with the aforementioned results, DeepDx appeared to be particularly useful to identify benign cases and high-grade cancers (Fig. 5A). Among the 52 misinterpreted cases both in UT2 and DeepDx, three (5.8%) were malignancies (Fig. 5B), which included two in GS 3 + 3 (Fig. 5B, yellow) and one in GS 4 + 4 (Fig. 5B, green). These accounted for the increased false-negative counts, from two in UT1 to six in UT2, leading to a slight decrease in the sensitivity of cancer detection (0.9957 in UT1 to 0.9849 in UT2, Supplementary Table 2). The detailed information for the diagnosis of these three cases are presented in Supplementary Table 5. IHC was used for the case with a GS 4 + 4, during the reference standard establishment. Upon microscopic review, the former cases showed deceptively bland morphology besides patchy (<1%) cancer areas (Fig. 5C, left), while the latter showed a HPIN-like structure with partial stromal infiltration (Fig. 5C, right).

DISCUSSION

We performed an independent external validation study to assess the diagnostic ability of DeepDx, using the WSIs of 593 prostate core biopsies. The AI algorithm showed higher concordance with the experts than the original pathology report to grade, also similar sensitivity and superior specificity than the original report to detect prostate cancer. We also demonstrated that DeepDx was robust and reproducible, devoid of the inter-observer variability found among pathologists; thus, it could be a step towards a standardized diagnosis of prostate needle biopsy. In the Gleason pattern analysis, DeepDx exceled in the recognition of Gleason patterns 4/5, highlighting the value of this algorithm in automatic scoring of prostate cancer and its clinical relevance. In addition,

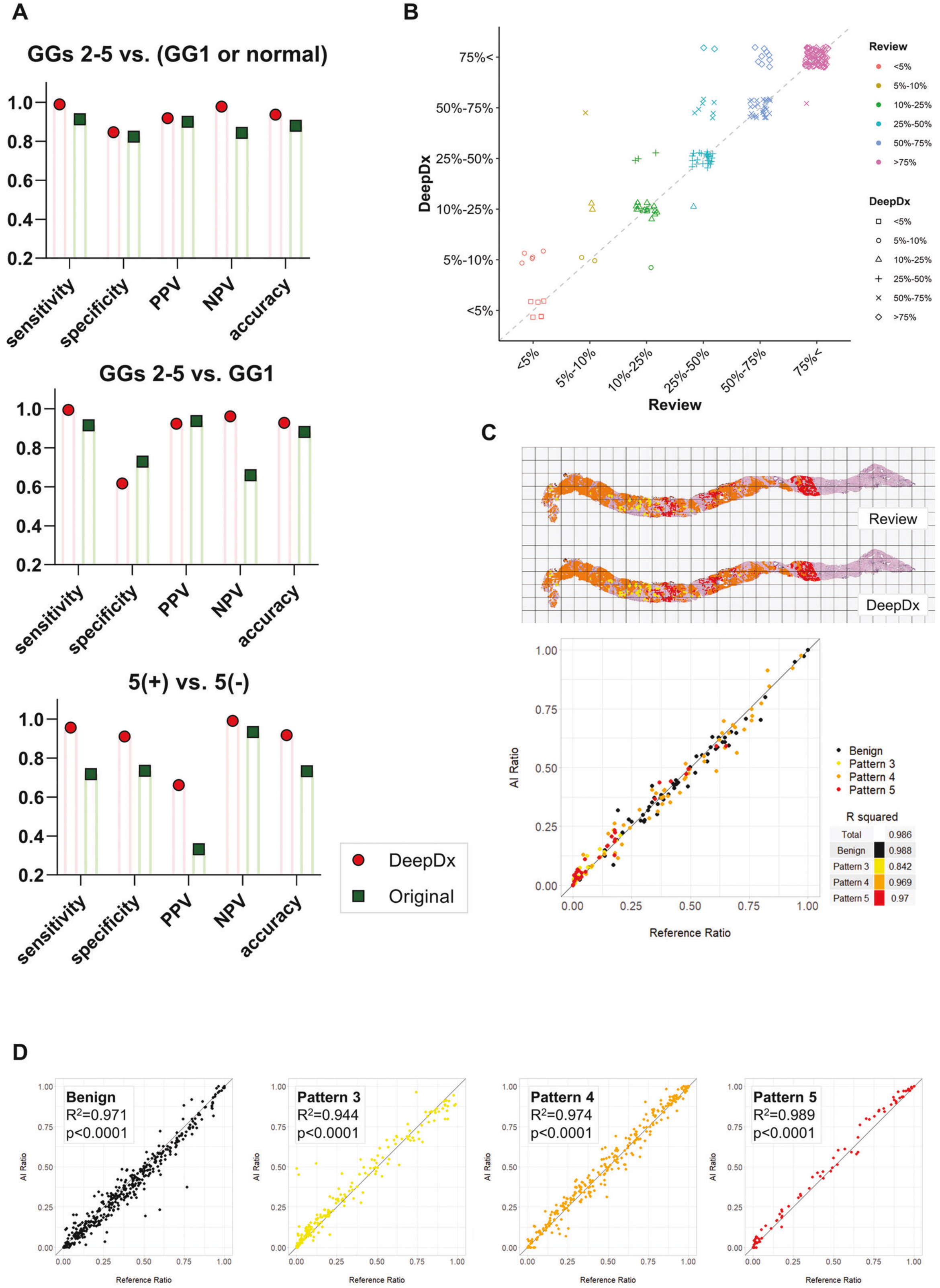


Fig. 3 Detection of Gleason patterns 4 and 5 by DeepDx. A Detection of Gleason grade groups (GG) 2–5 among all cases (top), GGs 2–5 among cancers (middle), Gleason pattern 5 among all cases (bottom). **B** The concordance of the % Gleason pattern 4 between DeepDx and the expert reviewer (kappa = 0.770, quadratic-weighted kappa = 0.940). **C** Evaluation of the correlation of Gleason pattern annotation in a representative case. **D** In ten examined cases, the gland-level patterns of DeepDx was significantly correlated to those of the expert reviewer (Pearson's correlation $P < 0.0001$).

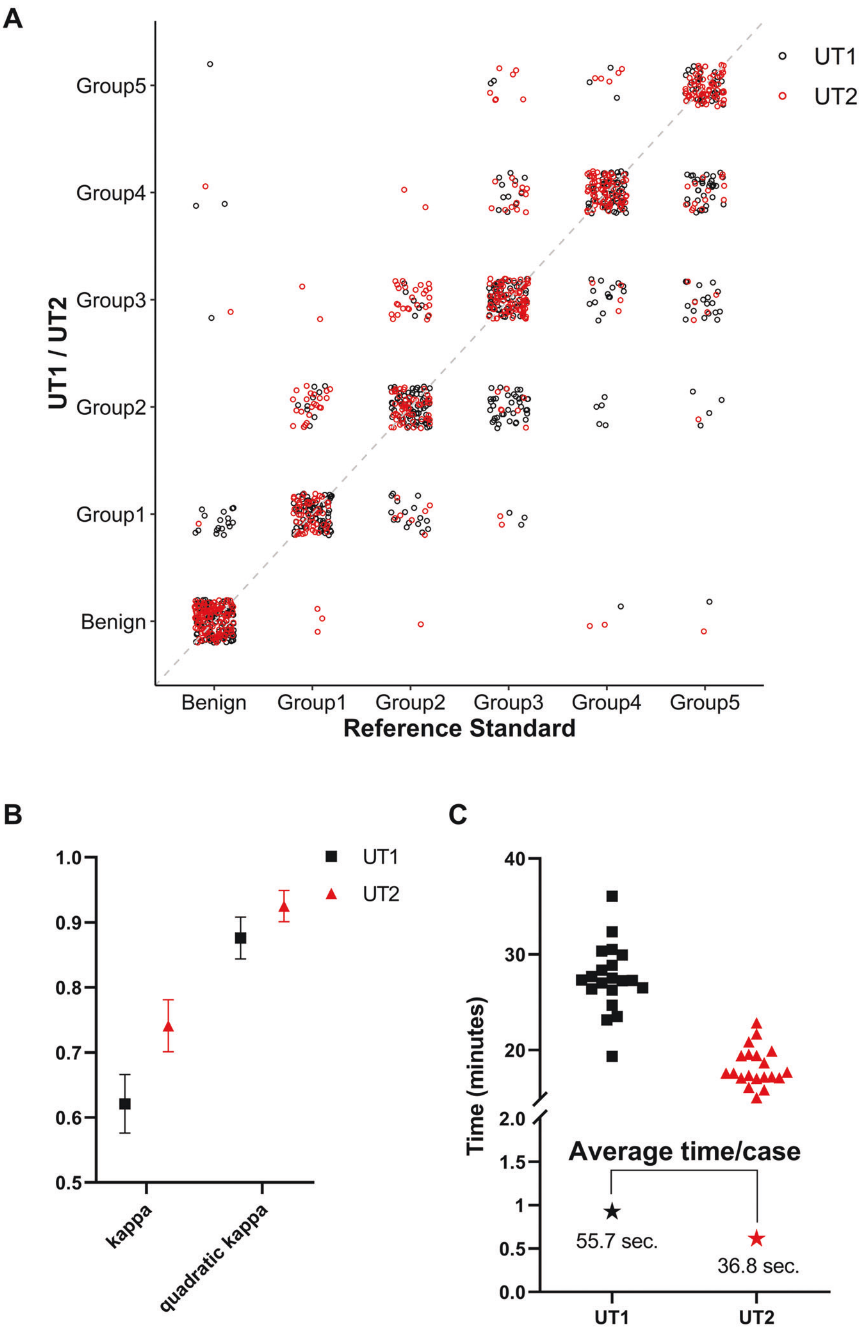
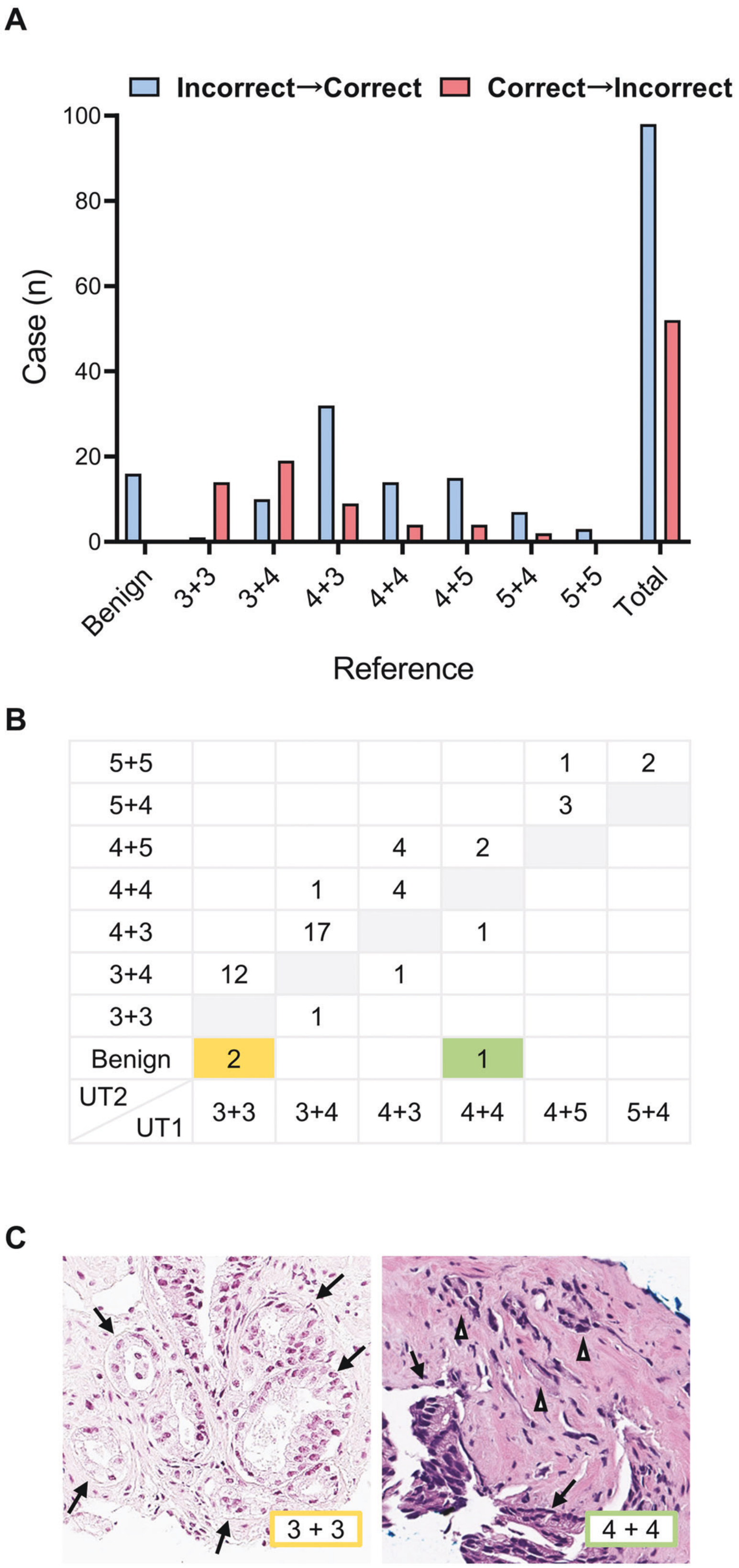


Fig. 4 User validation test of DeepDx assistance. **A** The results of individual cases in UT1 (without AI assistance) and UT2 (with AI assistance). **B** Kappa and quadratic-weighted kappa values of the grouping against the reference standard in UT1 and UT2. The error bars indicate 95% confidence intervals. **C** The time spent for UT1 and UT2. Squares and triangles indicate the records for every 30 cases. Stars denote the average time consumed for one case.

the user pathologist showed higher concordance with the reference GGs when assisted by the AI, while the time spent to examine the cases of the dataset was shortened.

We previously performed the internal test study for DeepDx and found that DeepDx demonstrated higher concordance of GG with the reference standard set by pathologists than that of the original pathology report.[8] In this current independent external validation study, DeepDx showed higher concordance with the reference standard for grading prostate adenocarcinomas than the original pathology reports. Even for the individual GSs, DeepDx performance was significantly correlated with the reference standard, and was higher than that of the original report. Therefore, we suggest that DeepDx has a diagnostic capacity that corresponds to the level of uropathologists. To date, there are several published papers on AI showing good performance in detecting and grading prostate cancer[5–7, 15–18]. However, to the best of our knowledge, only five studies have performed external validation for grading prostate cancer in core biopsy[7, 15–18]; of them, only two research teams have established reference standard provided by expert uropathologists[7, 16]. In the studies by Nagpal et al.[16] and Ström et al.[7] the authors performed external validation of their AI systems using reference standards provided by experts in ~330 cases in both studies, and revealed that the AIs were concordant to the experts. Using 593 biopsy cores, the largest independent



UT2 \ UT1	3+3	3+4	4+3	4+4	4+5	5+4
5+5					1	2
5+4					3	
4+5			4	2		
4+4		1	4			
4+3		17		1		
3+4	12		1			
3+3		1				
Benign	2			1		

Fig. 5 The effect of DeepDx on the user test results. A The numbers of cases that the general pathologist's diagnosis aligned with the reference standard (blue, $n = 98$ in total) or not (red, $n = 52$ in total) in the second round by following DeepDx. **B** The Gleason score (GS) distribution of the 52 discordant cases. Highlighted in yellow (GS 3 + 3) and green (GS 4 + 4) are the carcinomas recognized at first (no AI assistance) but missed at the second round (AI assistance). **C** One of the missed cancers with GS 3 + 3 (yellow in **B**) showed deceptively bland tumor glands (arrows; left). The GS 4 + 4 cancer (green in **B**) showed high-grade prostatic intraepithelial neoplasia-like architecture (arrows; right) with patchy stromal infiltration of poorly-formed glands (arrowheads; right) (original magnification, ×200; hematoxylin-eosin staining).

external dataset made by experts, we validated that DeepDx had the ability to diagnose prostate cancer at the level of uropathologists. Furthermore, even in difficult cases where diagnosis was inconsistent among the experts, the AI system exhibited better concordance with the reference standard than the original reports. Collectively, the results suggest that the introduction of a uropathologist-level AI system such as DeepDx could help grade prostate cancer accurately and offer a solution to the reported high inter-observer variability among pathologists.

The presence of Gleason patterns 4/5 is crucial to assess prognosis and manage prostate cancer patients[19]. Patients with Gleason patterns 4/5 generally need radical surgery[14]. In addition, patients with Gleason pattern 5 are classified into the prognostic stage group IIIC according to the American Joint Committee on Cancer (AJCC)[20] or high/very high risk groups according to the NCCN guideline[21], regardless of TNM stage or PSA; therefore, the ability to detect Gleason patterns 4/5 is important to estimate the clinical utility of an AI system. According to our analysis, DeepDx showed better accuracy than the original pathology reports to detect Gleason pattern 4/5. DeepDx tended to predict high-grade Gleason patterns 4/5 rather than Gleason pattern 3. GG 1 should be composed only of Gleason pattern 3 and is very liable to any amount of Gleason patterns 4/5 in needle biopsy[9]. We suspect that the high sensitivity and accuracy of DeepDx for high-grade Gleason patterns, even when a small amount of them was present in each slide, may have resulted in the system's underperformance for GG 1. Two other studies have performed a similar analysis to ours, including the ability to distinguish benign/ASAP/GG1 from GGs 2–5 or GGs 1-2 from GGs 3–5, using external validation cohorts[15, 17]. In the 2019 ISUP guideline, it was also recommended to report the percentage of Gleason pattern 4 for GGs 2/3 cancers in needle biopsies[4], because the amount of Gleason pattern 4 predicts biochemical recurrence of GGs 2/3 prostate cancers. By comparing the % of Gleason pattern 4 and pattern annotations, we identified that DeepDx was highly accurate in the measurement of Gleason patterns 4/5. To our knowledge, this is the first study to externally validate the measurement ability of AI for Gleason pattern 4. This robust performance of AI to detect Gleason patterns 4/5 meets the indispensable threshold to apply DeepDx into the future pathology practice.

In addition, we conducted a user test with and without AI support to investigate the system's actual impact on the pathologist's prostate cancer diagnostic performance. AI support improved the concordance of GG between the user and the reference standard and decreased the average slide examination time. To the best of our knowledge, only three previous studies performed a user test[22–24]. Raciti et al. conducted a user test to detect prostate cancer in 304 cores by three general pathologists, reporting that sensitivity for cancer detection was increased, while the time to reach diagnosis was decreased when the users were supported by AI[24]. However, their study differs from ours, as they designed it to detect small and low-grade tumors; furthermore, they did not study the ability of their AI system to grade prostate cancer[24]. Steiner et al. performed a user test where 20 general pathologists graded prostate cancer using 240 cores; when aided by AI, concordance with experts was significantly enhanced and spending time for grading was significantly reduced[23]. Bulten et al. also revealed that the agreement with the reference standard provided by experts was significantly increased, and the time required for diagnosis was decreased when supported by AI using 160 cores[22]. Reducing the workload of pathologists and their time spent to reach a diagnosis is critical for the potential adoption of AI in the future, due to the reported shortage of pathologists[25]. The user tests conducted in ours and previous studies suggest that the adoption of AI could enable general pathologists to detect and grade prostate cancer accurately and efficiently. In the future, prospective user tests should be performed to reinforce the potential use of AI in the daily clinical practice.

There are limitations in our study. We used the same scanner for our dataset and did not employ slides from diverse patient populations[26]. Concerning the user time efficiency test, we measured only the overall time spent to examine every 30 cases with or without AI assistance, not per slide separately. In addition, we tested the user performance of only one general pathologist. Our future plans regarding the clinical use of DeepDx for prostatic biopsies include implementing an extended user test.

In conclusion, DeepDx showed promising results to accurately detect and grade prostate cancer at the uropathologist level and reduce inter-observer variability among pathologists. This could help clinicians assess prognosis and facilitate the management of their patients. In order to promote the potential use of prostate biopsy AI algorithms in the future daily clinical practice, more validation studies using diverse scanner types and patient populations will be needed.

DATA AVAILABILITY

The data are available upon reasonable request by contacting the corresponding author.

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer. J. Clin*. **71**, 209–249 (2021).
2. Epstein, J. I. An update of the Gleason grading system. *J. Urol*. **183**, 433–440 (2010).
3. Allsbrook Jr., W. C., Mangold, K. A., Johnson, M. H., Lane, R. B., Lane, C. G. & Epstein, J. I. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Hum. Pathol*. **32**, 81–88 (2001).
4. van Leenders, G., van der Kwast, T. H., Grignon, D. J., Evans, A. J., Kristiansen, G., Kweldam, C. F. et al. The 2019 International Society of Urological Pathology (ISUP) consensus conference on grading of prostatic carcinoma. *Am. J. Surg. Pathol*. **44**, e87–e99 (2020).
5. Litjens, G., Sanchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I. et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep*. **6**, 26286 (2016).
6. Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med*. **25**, 1301–1309 (2019).
7. Strom, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D. M. et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet. Oncol*. **21**, 222–232 (2020).
8. Ryu, H. S., Jin, M. S., Park, J. H., Lee, S., Cho, J., Oh, S. et al. Automated gleason scoring and tumor quantification in prostate core needle biopsy images using deep neural networks and its comparison with pathologist-based assessment. *Cancers* **11**, 1860 (2019).
9. Humphrey, P. A., Moch, H., Cubilla, A. L., Ulbright, T. M. & Reuter, V. E. The 2016 WHO classification of tumours of the urinary system and male genital organs-Part B: prostate and bladder tumours. *Eur. Urol*. **70**, 106–119 (2016).
10. Hajian-Tilaki, K. Sample size estimation in diagnostic test studies of biomedical informatics. *J. Biomed. Inform*. **48**, 193–204 (2014).
11. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas*. **20**, 37–46 (1960).
12. Srigley, J. R., Delahunt, B., Samaratunga, H., Billis, A., Cheng, L., Clouston, D. et al. Controversial issues in Gleason and International Society of Urological Pathology (ISUP) prostate cancer grading: proposed recommendations for international implementation. *Pathology* **51**, 463–473 (2019).
13. Montironi, R., Scattoni, V., Mazzucchelli, R., Lopez-Beltran, A., Bostwick, D. G. & Montorsi, F. Atypical foci suspicious but not diagnostic of malignancy in prostate needle biopsies (also referred to as "atypical small acinar proliferation suspicious for but not diagnostic of malignancy"). *Eur. Urol*. **50**, 666–674 (2006).
14. Amin, M. B., Lin, D. W., Gore, J. L., Srigley, J. R., Samaratunga, H., Egevad, L. et al. The critical role of the pathologist in determining eligibility for active surveillance as a management option in patients with prostate cancer: consensus statement with recommendations supported by the College of American Pathologists, International Society of Urological Pathology, Association of Directors of

Anatomic and Surgical Pathology, the New Zealand Society of Pathologists, and the Prostate Cancer Foundation. *Arch. Pathol. Lab. Med.* **138**, 1387–1405 (2014).

15. Pantanowitz, L., Quiroga-Garza, G. M., Bien, L., Heled, R., Laifenfeld, D., Linhart, C. 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* **2**, e407–e416 (2020).
16. Nagpal, K., Foote, D., Tan, F., Liu, Y., Chen, P. C., Steiner, D. F. et al. Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. *JAMA Oncol.* **6**, 1372–1380 (2020).
17. Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B. et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* **21**, 233–241 (2020).
18. Mun, Y., Paik, I., Shin, S. J., Kwak, T. Y. & Chang, H. Yet another automated Gleason grading system (YAAGGS) by weakly supervised deep learning. *NPJ Digit. Med.* **4**, 99 (2021).
19. Chen, R. C., Rumble, R. B. & 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*. **12**, 267–269 (2016).
20. Buyyounouski, M. K., Choyke, P. L., McKenney, J. K., Sartor, O., Sandler, H. M., Amin, M. B. et al. Prostate cancer—major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J. Clin*. **67**, 245–253 (2017).
21. Mohler, J. L., Antonarakis, E. S., Armstrong, A. J., D'Amico, A. V., Davis, B. J., Dorff, T. et al. Prostate cancer, version 2.2019, NCCN clinical practice guidelines in oncology. *J. Natl Compr. Canc. Netw*. **17**, 479–505 (2019).
22. Bulten, W., Balkenhol, M., Belinga, J. A., Brilhante, A., Cakir, A., Egevad, L. et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod. Pathol.* **34**, 660–671 (2021).
23. Steiner, D. F., Nagpal, K., Sayres, R., Foote, D. J., Wedin, B. D., Pearce, A. et al. Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies. *JAMA Netw. Open* **3**, e2023267 (2020).
24. Raciti, P., Sue, J., Ceballos, R., Godrich, R., Kunz, J. D., Kapur, S. et al. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Mod. Pathol*. **33**, 2058–2066 (2020).
25. Robboy, S. J., Weintraub, S., Horvath, A. E., Jensen, B. W., Alexander, C. B., Fody, E. P. et al. Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. *Arch. Pathol. Lab. Med.* **137**, 1723–1732 (2013).
26. Kartasalo, K., Bulten, W., Delahunt, B., Chen, P. C., Pinckaers, H., Olsson, H. et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer in biopsies-current status and next steps. *Eur. Urol. Focus* **7**, 687–691 (2021).

AUTHOR CONTRIBUTIONS

Concept and design: H.S.R.; Acquisition, analysis, or interpretation of data: All authors; Drafting of the paper: M.J., M.S.J.; Language editing: I.P.N.; Critical revision of the paper for important intellectual content: H.S.R.; Statistical analysis: M.J., M.S.J.; Administrative, technical, or material support: M.J., H.S.R.; Supervision: H.S.R.; Final approval of paper: All authors.

COMPETING INTERESTS

The authors declare no competing interests.

ETHICS APPROVAL

The Institutional Review Board of SNUH approved this study (IRB no. D-2006-105-1134). This study was performed in accordance with the Declaration of Helsinki.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41379-022-01077-9.

Correspondence and requests for materials should be addressed to Han Suk Ryu.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.