















OPEN

Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge


Wouter Bulten^{1,60}  , Kimmo Kartasalo^{2,3,60}  , Po-Hsuan Cameron Chen^{4,60}  , Peter Ström², Hans Pinckaers¹, Kunal Nagpal⁴, Yuannan Cai⁴, David F. Steiner⁴ , Hester van Boven⁵, Robert Vink⁶, Christina Hulsbergen-van de Kaa⁶, Jeroen van der Laak^{1,7} , Mahul B. Amin⁸ , Andrew J. Evans⁹, Theodorus van der Kwast¹⁰ , Robert Allan¹¹, Peter A. Humphrey¹², Henrik Grönberg^{2,13} , Hemamali Samaratunga¹⁴, Brett Delahunt¹⁵, Toyonori Tsuzuki¹⁶ , Tomi Häkkinen³, Lars Egevad¹⁷, Maggie Demkin¹⁸, Sohier Dane¹⁸, Fraser Tan⁴, Masi Valkonen¹⁹, Greg S. Corrado⁴, Lily Peng⁴, Craig H. Mermel⁴ , Pekka Ruusuvoori^{3,19,61}, Geert Litjens^{16,61} , Martin Eklund^{2,61}  and the PANDA challenge consortium*

Artificial intelligence (AI) has shown promise for diagnosing prostate cancer in biopsies. However, results have been limited to individual studies, lacking validation in multinational settings. Competitions have been shown to be accelerators for medical imaging innovations, but their impact is hindered by lack of reproducibility and independent validation. With this in mind, we organized the PANDA challenge—the largest histopathology competition to date, joined by 1,290 developers—to catalyze development of reproducible AI algorithms for Gleason grading using 10,616 digitized prostate biopsies. We validated that a diverse set of submitted algorithms reached pathologist-level performance on independent cross-continental cohorts, fully blinded to the algorithm developers. On United States and European external validation sets, the algorithms achieved agreements of 0.862 (quadratically weighted κ , 95% confidence interval (CI), 0.840–0.884) and 0.868 (95% CI, 0.835–0.900) with expert uropathologists. Successful generalization across different patient populations, laboratories and reference standards, achieved by a variety of algorithmic approaches, warrants evaluating AI-based Gleason grading in prospective clinical trials.

Gleason grading¹ of biopsies yields important prognostic information for prostate cancer patients and is a key element for treatment planning². Pathologists characterize tumors into different Gleason growth patterns based on the histological architecture of the tumor tissue. Based on the distribution of Gleason patterns, biopsy specimens are categorized into one of five groups, commonly referred to as International Society of Urological Pathology (ISUP) grade groups, ISUP grade, Gleason grade groups or simply grade groups (GGs)^{3–6}. This assessment is inherently subjective with considerable inter- and intrapathologist variability^{7,8}, leading to both undergrading and overgrading of prostate cancer^{9–10}.

AI algorithms have shown promise for grading prostate cancer^{11,12}, specifically in prostatectomy samples^{13,14} and biopsies^{15–18},

and by assisting pathologists in the microscopic reviews^{19,20}. However, AI algorithms are susceptible to various biases in their development and validation^{21,22}. This can result in algorithms that perform poorly outside the cohorts used for their development. Moreover, shortcomings in validating the algorithms' performance on additional cohorts may lead to such deficiencies in generalization going unnoticed^{23,24}. Algorithms are also often developed and validated in a siloed manner: the same researchers who develop the algorithms also validate them. This leads to risks of introducing positive bias, because the developing researchers have control over, for example, establishing the validation cohorts and selecting the pathologists providing the reference standard. There has yet to be an independent evaluation of algorithms for prostate cancer

¹Department of Pathology, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands. ²Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. ³Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. ⁴Google Health, Palo Alto, CA, USA. ⁵Department of Pathology, Antoni van Leeuwenhoek Hospital, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ⁶Laboratory of Pathology East Netherlands, Hengelo, The Netherlands. ⁷Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden. ⁸Department of Pathology and Laboratory Medicine, University of Tennessee Health Science Center, Memphis, TN, USA. ⁹Laboratory Medicine, Mackenzie Health, Toronto, Ontario, Canada. ¹⁰Department of Pathology, Laboratory Medicine and Pathology, University Health Network and University of Toronto, Toronto, Ontario, Canada. ¹¹Pathology and Laboratory Medicine Service, North Florida/South Georgia Veterans Health System, Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, FL, USA. ¹²Department of Pathology, Yale School of Medicine, New Haven, CT, USA. ¹³Department of Surgery, Capio St. Göran's Hospital, Stockholm, Sweden. ¹⁴Aquesta Uropathology and University of Queensland, Brisbane, QLD, Australia. ¹⁵Department of Pathology and Molecular Medicine, Wellington School of Medicine and Health Sciences, University of Otago, Wellington, New Zealand. ¹⁶Department of Surgical Pathology, School of Medicine, Aichi Medical University, Nagakute, Japan. ¹⁷Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden. ¹⁸Kaggle Inc, Mountain View, CA, USA. ¹⁹Institute of Biomedicine, Cancer Research Unit and FICAN West Cancer Centre, University of Turku and Turku University Hospital, Turku, Finland. ⁶⁰These authors contributed equally: Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen. ⁶¹These authors jointly supervised this work: Pekka Ruusuvoori, Geert Litjens, Martin Eklund. *A list of authors and their affiliations appears at the end of the paper.  e-mail: wouter@wouterbulten.com; kimmo.kartasalo@ki.se; cameronchen@google.com

diagnosis and grading to assess whether they generalize across different patient populations, pathology labs, digital pathology scanner providers and reference standards derived from intercontinental panels of uropathologists. This represents a key barrier to implementation of algorithms in clinical practice.

AI competitions have been an effective approach to crowd source the development of performant algorithms^{25–27}. Despite their effectiveness in facilitating innovation, competitions still tend to suffer from a set of limitations. Validation of the resulting algorithms has typically not been performed independently of the algorithm developers. In a competitive setup, the incentive for conscious or subconscious introduction of positive bias by the developers is arguably further increased, and a lack of independent validation also means that the technical reproducibility of the proposed solutions is not verified. Moreover, competitions have typically not been followed up by validation of the algorithms on additional international cohorts, casting doubt on whether the resulting solutions possess the generalization capability to truly answer the underlying clinical problem, as opposed to being fine-tuned for a particular competition design and dataset²⁸.

Through the present study, we aimed to advance the methodology for the design and evaluation of medical imaging AI innovations to develop and rigorously validate the next generation of algorithms for prostate cancer diagnostics. We organized a global AI competition, the Prostate cANcer graDe Assessment (PANDA) challenge, by compiling and publicly releasing a European (EU) cohort for AI development, the largest publicly available dataset of prostate biopsies to date. Second, we fully reproduced top-performing algorithms and externally validated their generalization to independent US and EU cohorts and compared them with the reviews of pathologists. The competition setup isolated the developers from the independent evaluation of the algorithms' performance, minimizing the potential for information leakage and offering a true assessment of the diagnostic power of these techniques. Taken together, we show how the combination of AI and innovative study designs, together with prespecified and rigorous validation across diverse cohorts, can be utilized to solve challenging and important medical problems.

Results

Characteristics of the datasets. In total, 12,625 whole-slide images (WSIs) of prostate biopsies were retrospectively collected from 6 different sites for algorithm development, tuning and independent validation (Table 1, Extended Data Fig. 1 and Supplementary Tables 7 and 8). Of these, 10,616 biopsies were available for model development (the development set), 393 for performance evaluation during the competition phase (the tuning set), 545 as the internal validation set in the postcompetition phase and 1,071 for external validation.

Cases for development, tuning and internal validation originated from Radboud University Medical Center, Nijmegen, the Netherlands and Karolinska Institutet, Stockholm, Sweden (Extended Data Fig. 1 and Supplementary Methods 1, 2 and 3). The external validation data consisted of a US and an EU set. The US set contained 741 cases and was obtained from two independent medical laboratories and a tertiary teaching hospital. The EU external validation set contained 330 cases and was obtained from the Karolinska University Hospital, Stockholm, Sweden. The histological preparation and scanning of the external validation samples were performed by different laboratories to those responsible for the development, tuning and internal validation data.

Reference standards of the datasets. The reference standard for the Dutch part of the training set was determined based on the pathology reports from routine clinical practice. For the Swedish part of the training set, the reference standard was set by one uropathologist (L.E.) following routine clinical workflow. The reference

standard for the Dutch part of the internal validation set was determined through consensus of three uropathologists (C.H.v.d.K., R.V. and H.v.B.) from two institutions with 18–28 years of clinical experience after residency (mean of 22 years). For the Swedish subset, four uropathologists (L.E., B.D., H.S. and T.T.) from four institutions, all with >25 years of clinical experience after residency, set the reference standard.

For the US external validation set, the reference standard was set by a panel of six US or Canadian uropathologists (M.A., A.E., T.v.d.K., M.Z., R.A. and P.H.) from six institutions with 18–34 years of clinical experience after residency (mean of 25 years). Each specimen was first reviewed by two uropathologists from the panel. A third uropathologist reviewed discordant cases to arrive at a majority opinion. For this external dataset, immunohistochemistry was available to aid in tumor identification. The EU external validation set was reviewed by a single uropathologist (L.E.). For details on the uropathologist review protocol, see Supplementary Methods 2. On validation sets, the pathologists who contributed to the reference standards showed high pairwise agreement (0.926 on a subset of the EU internal validation set and 0.907 on the US external validation set, Supplementary Table 6). To ensure consistency across different reference standards, we additionally investigated the agreement between reference standards across continents (Supplementary Table 9). We found high agreement between pathologists across the regions when EU uropathologists reviewed US data and vice versa. Moreover, majority votes of the panels were highly consistent with the reference standard of the other region (quadratically weighted κ 0.939 and 0.943 for, respectively, the EU and US pathologists, Supplementary Table 9).

Overview of the competition. The study design of the PANDA challenge was preregistered²⁹ and consisted of a competition and a validation phase. The competition was open to participants from 21 April until 23 July 2020 and was hosted on the Kaggle platform (Supplementary Methods 5). During the competition phase, 1,010 teams, consisting of 1,290 developers from 65 countries, participated and submitted at least one algorithm (Fig. 1). Throughout the competition, teams could request evaluations of their algorithm on the tuning set (Supplementary Methods 2). The algorithms were then simultaneously blindly validated on the internal validation set (Fig. 2). All teams combined submitted 34,262 versions of their algorithms, resulting in a total of 32,137,756 predictions made by the algorithms.

The first team to achieve an agreement with the uropathologists of >0.90 (quadratically weighted Cohen's κ) on the internal validation set already occurred within the first 10 days of the competition (Fig. 2). In the 33rd day of the competition, the median performance of all teams exceeded a score of 0.85.

Overview of evaluated algorithms. After the competition, teams were invited to join the PANDA consortium. Of all teams, 33 submitted a proposal to join the validation phase of the study. From these, the competition organizers selected 15 teams based on their algorithm's performance on the internal validation set and method description (Supplementary Methods 6). Among the 10 highest ranking teams in the competition, 8 submitted a proposal and were accepted to join the consortium. A further seven teams in the consortium all ranked within the competition's top 30.

All selected algorithms made use of deep learning-based methods^{30,31}. Many of the solutions demonstrated the feasibility of end-to-end training using case-level information only³², that is, using the International Society of Urological Pathology (ISUP) GG of a specimen as the target label for an entire WSI. Most leading teams, including the winner of the competition, adopted an approach in which a sample of smaller images, or patches, is first extracted from the WSI. The patches are then fed to a convolutional

Table 1 | Data characteristics of the development set, tuning set, internal validation set and the two external validation sets

	EU development set	EU development set	EU tuning set	EU tuning set	EU internal validation set	EU internal validation set	US external validation set	EU external validation set	Total
Source	Radboud University Medical Center Netherlands	Karolinska Institutet Sweden	Radboud University Medical Center Netherlands	Karolinska Institutet Sweden	Radboud University Medical Center Netherlands	Karolinska Institutet Sweden	Medical Laboratories, CA/UT, USA; tertiary teaching hospital, CA, USA	Karolinska University Hospital Sweden	–
No. of sites	1	1	1	1	1	1	3	1	6
No. of cases	1,028	1,085	72	33	129	82	741	330	3,500
No. of biopsies	5,160	5,456	195	198	333	212	741	330	12,625
Nontumor	967 (19)	1,925 (35)	95 (49)	58 (29)	155 (47)	66 (31)	254 (34)	108 (33)	3,628 (29)
Tumor-containing (ISUP GG breakdown below)	4,193 (81)	3,531 (65)	100 (51)	140 (71)	178 (53)	146 (69)	487 (66)	222 (67)	8,997 (71)
GG 1	852 (17)	1,814 (33)	24 (12)	48 (24)	48 (14)	53 (25)	247 (33)	65 (20)	3,151 (25)
GG 2	675 (13)	668 (12)	15 (8)	32 (16)	35 (11)	34 (16)	122 (16)	63 (19)	1,644 (13)
GG 3	925 (18)	317 (6)	15 (8)	14 (7)	38 (11)	16 (8)	70 (9)	49 (15)	1,444 (11)
GG 4	768 (15)	481 (9)	19 (10)	30 (15)	16 (5)	22 (10)	21 (3)	19 (6)	1,376 (11)
GG 5	973 (19)	251 (5)	27 (14)	16 (8)	41 (12)	21 (10)	27 (4)	26 (8)	1,382 (11)
No. of cases with general pathologist reviews	–	–	–	–	70	–	237	–	307
No. of pathologist reviews	–	–	–	–	910	–	4,740	–	5,650

The values in parentheses give the percentage. The development set was available to competition teams for algorithm development, and the tuning set for limited algorithm evaluation during the competition. All validation sets were fully independent and blinded to the algorithm developers. Additional details on reference standard protocol can be found in Supplementary Methods 2 and 3.

neural network, the resulting feature responses are concatenated and the final classification layers of the network are applied to these features. This allows training a single model end-to-end in a computationally efficient fashion to directly predict the ISUP GG of a WSI. Such weakly supervised approaches do not require detailed pixel-level annotations as often used in fully supervised training.

Another algorithmic feature adopted by several top-performing teams was to apply automated label cleaning, where samples considered as erroneously graded by the pathologists were either excluded from training or relabeled. Several teams indicated the label noise associated with the subjective grading assigned by pathologists as a key problem, and tackled it by algorithms that detect samples where the reference standard deviates considerably from the predictions of the model. Label denoising was then typically applied iteratively to refine the labels more aggressively as the model's performance improved during training.

A third key feature shared by all teams of the PANDA consortium was the use of ensembles consisting of diverse models, featuring, for example, different data preprocessing approaches or different neural network architectures. Despite the relative diversity in these algorithmic details, by averaging the predictions of the models constituting the ensembles, most teams achieved comparable overall performance.

For a summary and details on the individual algorithms see Supplementary Methods 7 and Supplementary algorithm descriptions. Most of the evaluated algorithms are available freely for research use (please see Supplementary algorithm descriptions for further details).

Classification performance in the internal validation dataset. In the validation phase, all selected algorithms were fully reproduced on two separate computing platforms. The average agreement of the selected algorithms with the uropathologists was high with a quadratically weighted κ of 0.931 (95% CI, 0.918–0.944, Fig. 3). Algorithms showed high sensitivity for tumor detection, with the representative algorithm (selected based on median balanced accuracy, see Statistical analysis) achieving a sensitivity of 99.7% (95% CI of all algorithms, 98.1–99.7, Fig. 4) and a specificity of 92.9% (95% CI of all algorithms, 91.9–96.7). The classification performances of the individual algorithms are presented in Extended Data Figs. 2–4 and Supplementary Tables 2 and 3.

Classification performance in the external validation datasets. The algorithms were independently evaluated on the two external validation sets. The agreements with the reference standards were high with a quadratically weighted κ of 0.862 (95% CI, 0.840–0.884) and 0.868 (95% CI, 0.835–0.900) for the US and EU external validation sets, respectively. The main algorithm error mode was overdiagnosing of benign cases as ISUP GG 1 cancer (Extended Data Figs. 5 and 6).

The representative algorithm identified cases with tumor in the external validation sets, with sensitivities of 98.6% (95% CI of all algorithms, 97.6–99.3) and 97.7% (95% CI of all algorithms, 96.2–99.2) for the US and EU sets, respectively. In comparison to the internal validation set, the algorithms misclassified more benign cases as malignant, resulting in specificities of 75.2% (95% CI of all algorithms, 66.8–80.0) and 84.3% (95% CI of all algorithms, 70.5–87.9) for the representative algorithm.

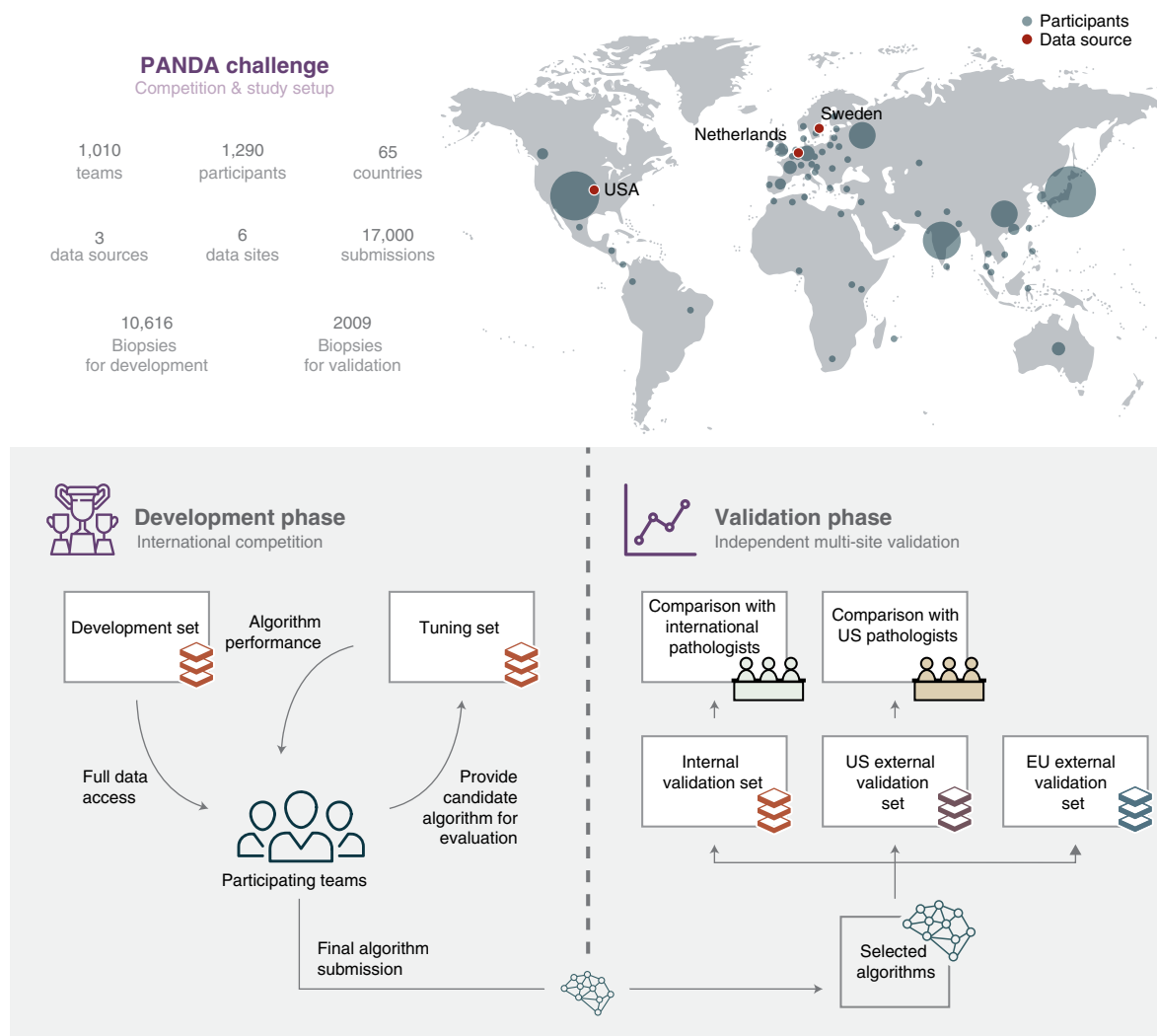


Fig. 1 | Overview of the PANDA challenge and study setup. The global competition attracted participants from 65 countries (top: size of the circle for each country illustrates the number of participants). The study was split into two phases. First, in the development phase (bottom left), teams competed in building the best-performing Gleason grading algorithm, having full access to a development set for algorithm training and limited access to a tuning set for estimating algorithm performance. In the validation phase (bottom right), a selection of algorithms was independently evaluated on internal and external datasets against reference grading obtained through consensus across expert uropathologist panels, and compared with groups of international and US general pathologists on subsets of the data.

Classification performance compared with pathologists. To compare algorithms' performances with those of general pathologists, we obtained reviews from two panels of pathologists on subsets of the internal and US external validation sets. For the Dutch part of the internal validation set, 13 pathologists from 8 countries (7 from Europe and 6 outside of Europe) reviewed 70 cases. For the US external validation set, 20 US board-certified pathologists reviewed 237 cases. For details on the pathologist review protocol, see Supplementary Methods 3.

The algorithms scored significantly ($P < 0.001$) higher in agreement with the uropathologists (0.876, 95% CI, 0.797–0.927; Fig. 3) than the international general pathologists did (0.765, 95% CI, 0.645–0.852) on the 70 cases from the Dutch part of the internal validation set. The representative algorithm had higher sensitivity for tumor (98.2%, 95% CI of all algorithms 97.4–100.0) than the representative pathologist (96.5%, 95% CI of all pathologists 95.4–100.0) and higher specificity (100.0%, 95% CI of all algorithms 90.6–100.0, versus 92.3%, 95% CI of all pathologists 77.8–97.8). On average, the algorithms missed 1.0% of cancers, whereas the pathologists missed

1.8%. Differences in grade assignments between the algorithms and pathologists are visualized in Fig. 5.

On the subset of the US external validation set with pathologist reviews, the algorithms exhibited a similar level of agreement with the uropathologists as the US general pathologists did (0.828, 95% CI, 0.781–0.869 versus 0.820, 95% CI, 0.760–0.865; $P = 0.53$). The representative algorithm had higher sensitivity for tumor (96.4%, 95% CI of all algorithms, 96.6–99.5) than the representative pathologist (91.9%, 95% CI of all pathologists, 89.3–95.5) but lower specificity (75.0%, 95% CI of all algorithms, 61.2–82.7 versus 95.0%, 95% CI of all pathologists, 87.4–98.1). On average, the algorithms missed 1.9% of cancers, whereas the pathologists missed 7.3%.

Discussion

AI has shown promise for diagnosis and grading of prostate cancer, but these results have been restricted to siloed studies with limited proof for generalization across diverse multinational cohorts, representing one of the central barriers to implementation of AI algorithms in clinical practice. The objective of the present study

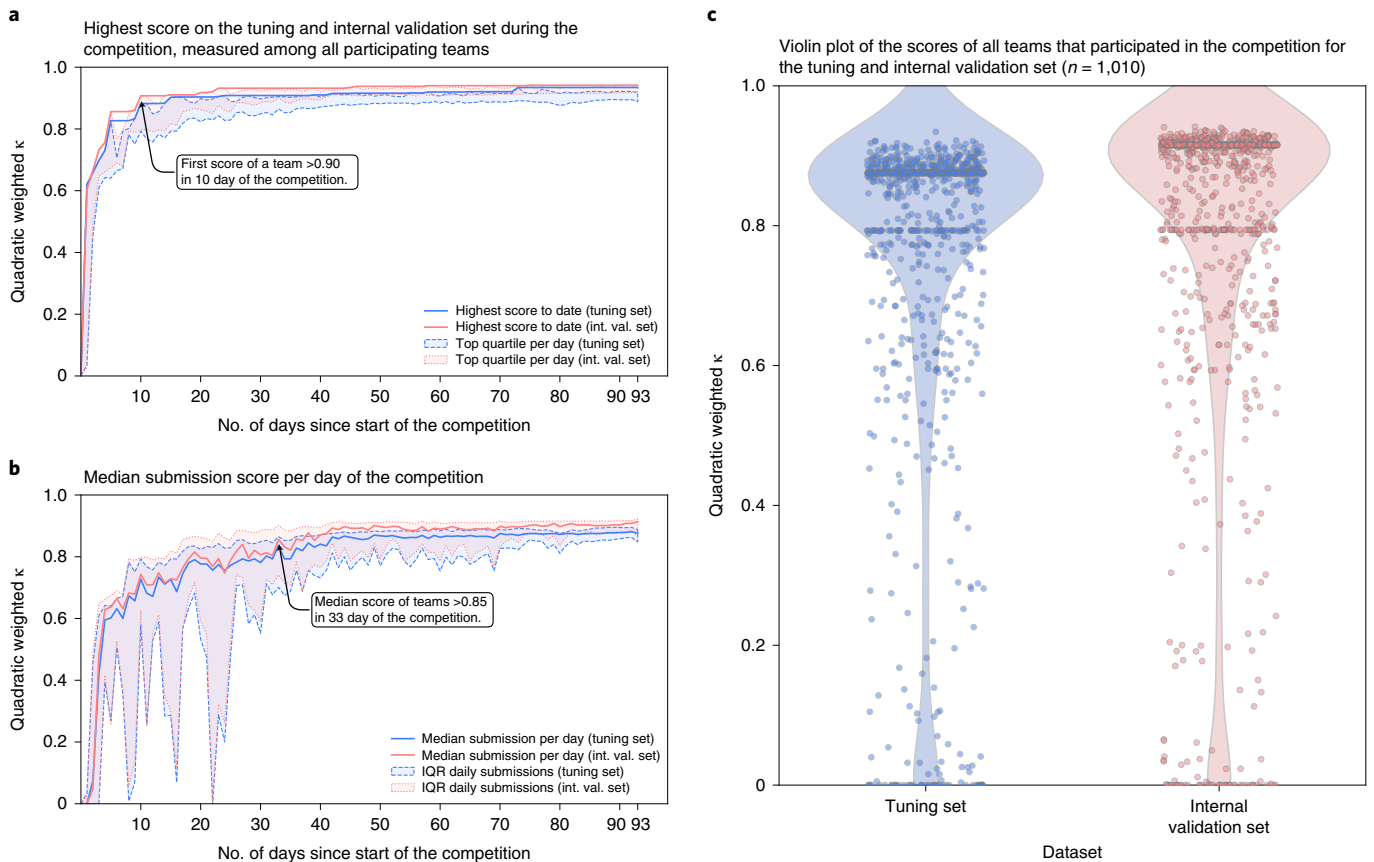


Fig. 2 | Progression of algorithms' performances throughout the competition. During the competition, teams could submit their algorithm for evaluation on the tuning set, after which they received their score. At the same time, algorithms were evaluated on the internal validation set, without disclosing these results to the participating teams. **a,b**, The development of the top score obtained by any team (**a**) and the median score over all daily submissions (**b**) throughout the timeline of the competition showing the rapid improvement of the algorithms. **c**, A large fraction of teams reached high scores in the range 0.80–0.90, and retained their performance on the internal validation set.

was to overcome these critical issues. First, we aimed to facilitate community-driven development of AI algorithms for cancer detection and grading on prostate biopsies. Second, we sought to transcend isolated assessment of the diagnostic performance of individual AI solutions by focusing on reproducibility and fully blinded validation of a diverse group of algorithms on intercontinental and multinational cohorts.

The resulting PANDA challenge was, to the best of our knowledge, the largest competition in pathology organized to date, in terms of both the number of participants and the size of the datasets, and the first study to analyze a variety of AI algorithms for computational pathology on this scale³³. The datasets included variability in biopsy sampling procedure, specimen preparation process and whole-slide scanning equipment, and had different and multinational sets of pathologists contributing to the reference standard of the validation sets. Our main finding was that AI algorithms obtained from a competition setup could successfully detect and grade tumors, reaching pathologist-level concordance with expert reference standards. We further compared the algorithms with previously published results (Supplementary Table 5 and Extended Data Fig. 7)^{15–17}. The algorithms outperformed earlier works on subsets of the EU validation sets. On the US external validation set, the algorithms reached similar performance without any fine-tuning, demonstrating a successful generalization to an unseen independent validation set and beyond any current state of the art. Last, groups of international and US pathologists also reviewed subsets of the internal and external validation datasets. The algorithms had

a concordance with the reference standard that was similar to or higher than that of these pathologists.

In the external validation sets, the main algorithm error mode was overdiagnosing benign cases as ISUP GG 1. This is probably due to the data distribution shift between training data and external validation data³⁴, in combination with the study design of independent validation, where the teams did not have any access to the validation sets, potentially leading to suboptimal selection of operating thresholds based only on the tuning set. We observed this in the US external validation set (Fig. 4), where the algorithms appear to be shifted toward higher sensitivity but lower specificity compared with the general pathologists. A potential solution to address the natural data distribution shift is to calibrate the models' predictions using sampled data from the target sites. In addition, we showed high consistency between reference standards (Supplementary Table 9), adding additional proof that the performance drop was not caused by a difference in grading characteristics.

In the US external validation set, tumor identification was confirmed by immunohistochemistry, supporting the finding that the algorithms missed fewer cancers than the pathologists. This higher sensitivity shows promise for reducing pathologist workload by automated identification and exclusion of most benign biopsies from review. Analysis of an ensemble constructed from the algorithms suggests that combining existing algorithms could improve specificity (Supplementary Table 4).

Further analysis of the grade assignments by the algorithms and general pathologists showed that the algorithms tended to assign

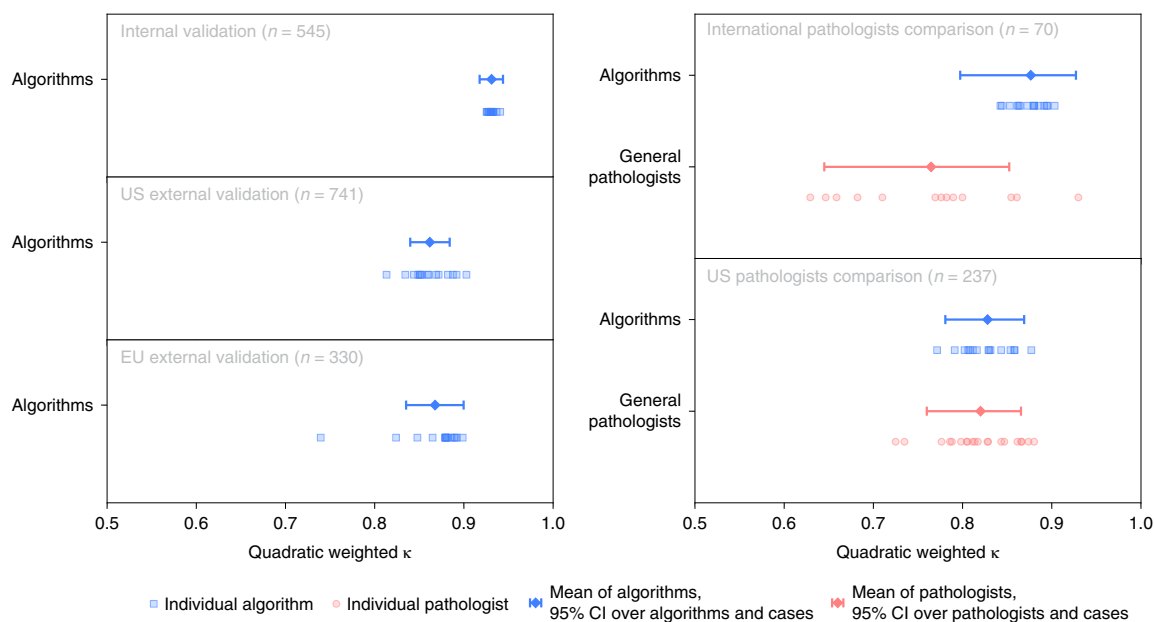


Fig. 3 | Algorithm agreement with reference standards and comparison to pathologists. Algorithms' agreement (quadratically weighted κ) with reference standards established by urologists shown for the internal and external validation sets (left). On subsets of the internal and US external validation sets, agreement of general pathologists with the reference standards is additionally shown for comparison (right).

higher grades than the pathologists (Fig. 5). For example, in the US external validation set, algorithms overgraded a substantial portion of ISUP GG 3 cases as GG 4. The general pathologists, in contrast, tended to undergrade cases, most notably in the high-grade cases. These differences suggest that general pathologists supported by AI could reach higher agreements with urologists, potentially alleviating some of the rater variability associated with Gleason grading^{19,20}. It should be noted that the algorithms' operating points were selected solely based on the EU tuning set. For clinical usage, the operating points can be adjusted based on the needs and the intended use cases. For example, for a prescreening use case aimed to reduce pathologist workload, one could select an operating point favoring high sensitivity to minimize false negatives. Alternatively, if AI was used as a stand-alone tool, increasing the algorithms' specificity to tumors, while retaining a high sensitivity, could be an important prerequisite for clinical implementation to prevent overdiagnosis.

We aimed to lower the entry barrier to medical AI development by providing access to a large, curated dataset, typically attainable only through large research consortia, and by organizing this competition to facilitate joint development with experience sharing among the teams. The results show that the publication of such datasets can lead to rapid development of high-performing AI algorithms. Dissemination and fast iteration of new ideas resulted in the first team achieving pathologist-level performance in the first 10 days of the challenge (Fig. 2). These results show the important role data play in the development of medical AI algorithms, given the short lead-time of top-performing solutions by various teams. At the same time, often raised criticisms of medical AI challenges are the lack of detailed reporting, and limited interpretation and reproducibility of results²⁸. Typically algorithms are evaluated only on internal competition data and by participants themselves, which introduces a risk of overfitting and reduced likelihood of reproducibility. We addressed these limitations in our challenge design by using preregistration, blinded evaluation, full reproduction of algorithm results, independent validation of algorithms on external data and comparison with pathologists.

This study has limitations. First, for the validation phase we were limited to including 15 teams from the pool of 1,010. To ensure transparent selection of teams and minimizing potential bias in the external validation, we disclosed the selection criteria and process beforehand to all participating teams, included both score and algorithm descriptions as criteria, and performed the selection before running the analyses (Supplementary Methods 6).

Second, algorithm validation was restricted to the assessment of individual biopsies whereas, in clinical practice, pathologists examine multiple biopsies per patient. Future studies can focus on patient-level evaluation of tissue samples, taking multiple cores and sections into account for the final diagnosis. Third, this study focused on grading acinar adenocarcinoma of the prostate, and algorithm responses to other variants and subtypes of cancer, precancerous lesions or nonprostatic tissue were not specifically assessed. Although cases with potential pitfalls were not excluded, it is of interest to further examine algorithm performance on such cases (for example, benign mimickers, severe inflammation, high-grade prostatic intraepithelial neoplasia, partial atrophy) and to investigate which patterns consistently result in classification errors. Although not quantitatively assessed, an analysis of cases with frequent miscalls showed that these cases often contained patterns such as cutting artifacts, and different inflammatory and other biological processes—all common occurrences within pathology—that could have resulted in the algorithms' miscalls. A comprehensive understanding of potential error modes is especially important when these algorithms leave controlled research settings and are used in clinical settings. Therefore, future research should more extensively assess what common tissue patterns in pathology routinely affect algorithm performance, whether they are the same patterns that are notoriously difficult for pathologists and how we can build safeguards to prevent such errors.

Fourth, algorithms were compared against reference standards set by various panels of pathologists. Although the gold standard in the field, relying on pathologists' gradings introduces a risk of bias because algorithms could learn the grading habits of specific pathologists and not generalize well to other populations. To remedy this effect, panels of urologists established the reference

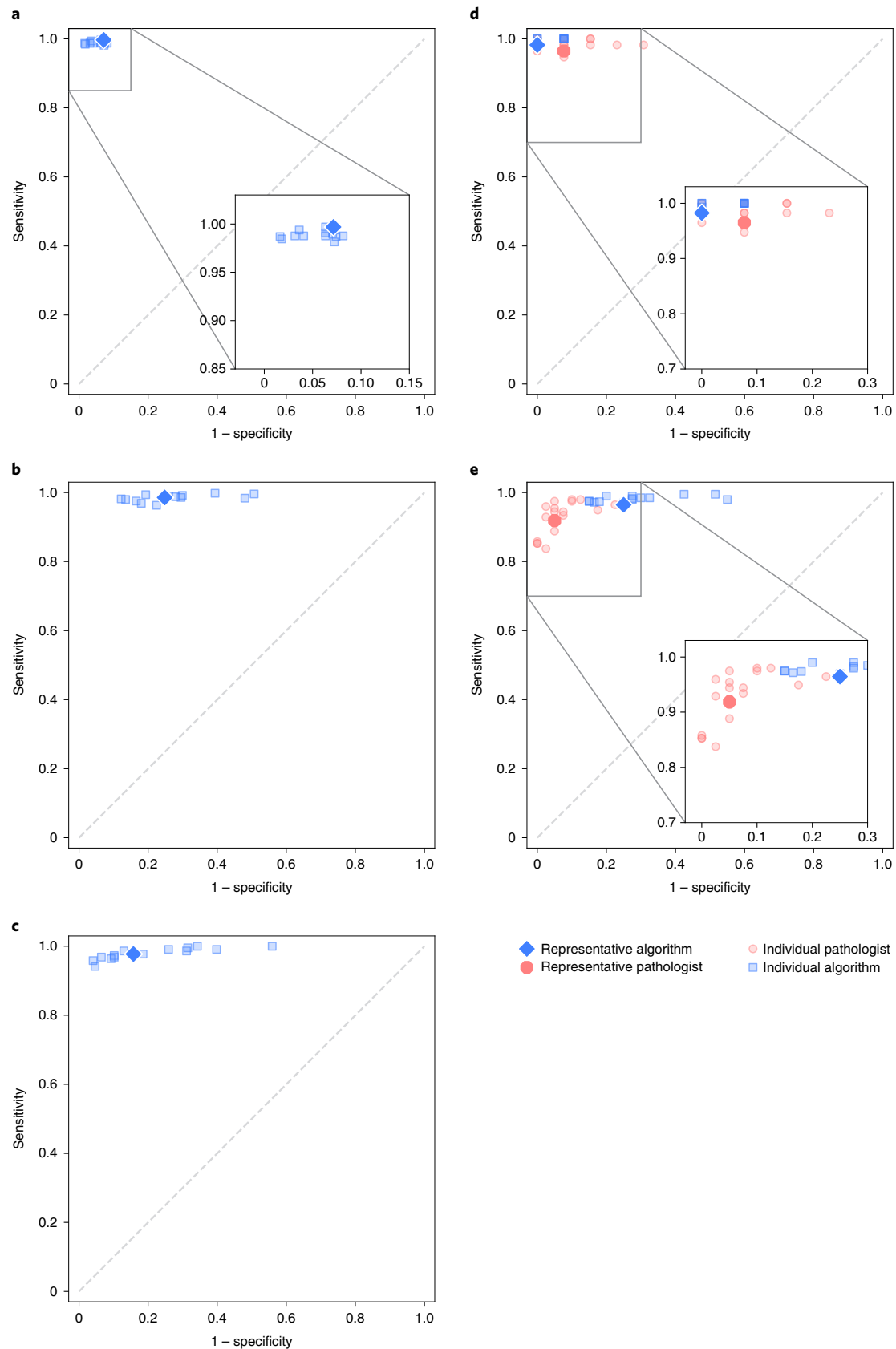


Fig. 4 | Algorithm performance in detecting prostate tumors on validation sets and comparison to pathologists. a,b,c, The sensitivity and specificity of the algorithms relative to reference standards established by urologists shown for the internal (**a**) and external validation sets (**b, c**). **b,** US external validation. **c,** EU external validation. **d,e,** On subsets of the internal and US external validation sets, the sensitivity and specificity of general pathologists are also shown for comparison. **d,** International pathologists' comparison. **e,** US pathologists' comparison.

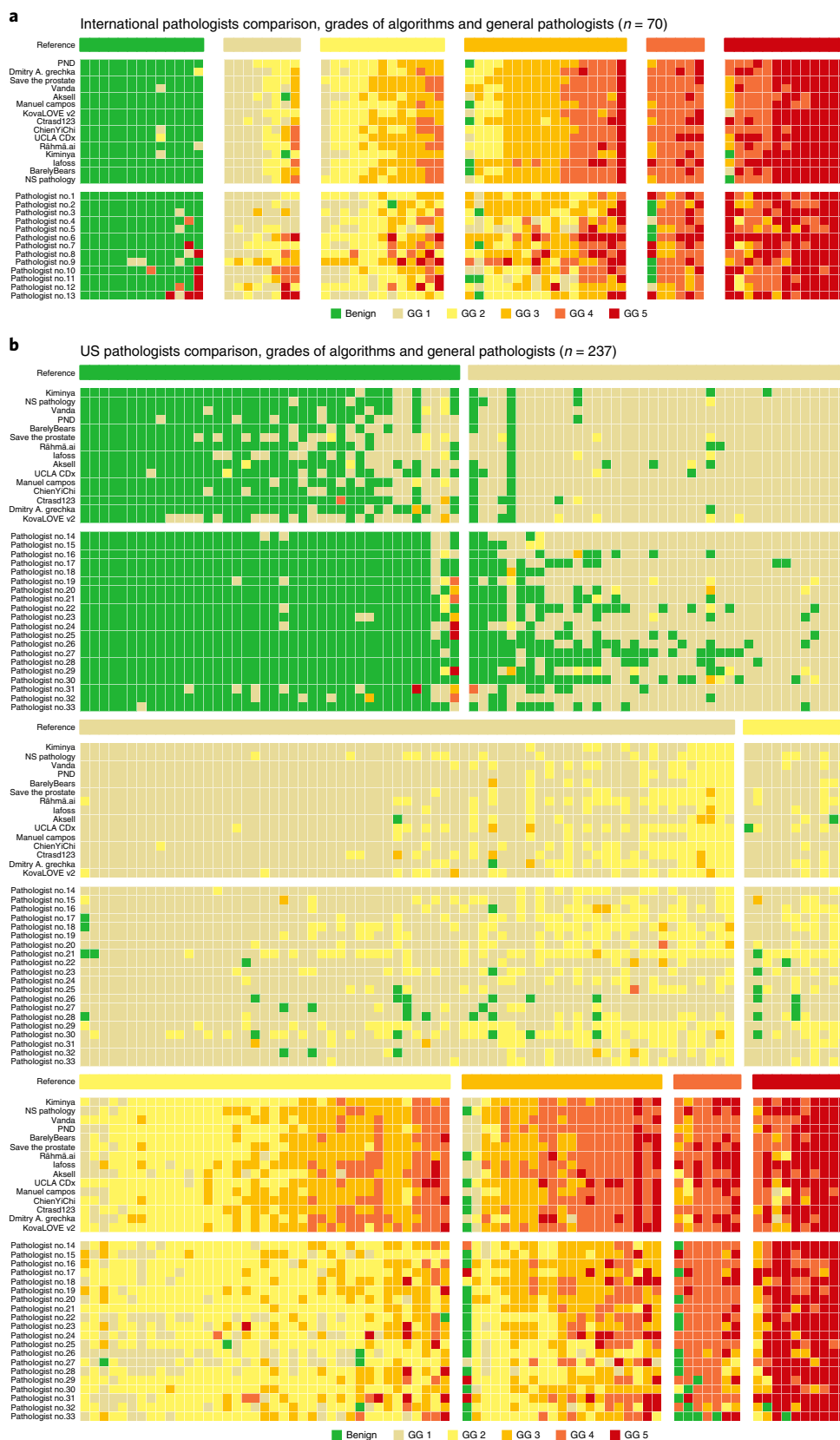


Fig. 5 | ISUP GG assignment by algorithms and pathologists. **a, b**, Algorithms compared with international general pathologists on a subset of the internal validation set (**a**) and US general pathologists on a subset of the US external validation set (**b**). Cases are ordered primarily by the reference ISUP GG and secondarily by the average GG of the algorithms and pathologists. Algorithms and pathologists are ordered by their agreement (quadratically weighted κ) with the reference standard on the respective sets. The comparison between pathologists and algorithms gives insight into the difference in their operating points and for which GGs most miscalls are made. The algorithms are less likely to miss a biopsy containing cancer, but at the same time more likely to overgrade benign cases.

standards of the EU internal and US external validation sets. Although these sets were graded in silo by different panels, we have shown that a majority-vote reference standard is highly consistent, even in a cross-continent setting (Supplementary Table 9). The EU external validation set was an exception, because a single uropathologist established the reference standard for that set. However, we observed high concordance between the grading by this pathologist and other pathologists when evaluated on the internal validation set (Supplementary Methods 2). For the training sets, we relied on reference standards extracted from clinical diagnostics, typically set by a single pathologist. Although unfeasible due to the high number of cases, a training reference standard based on multiple pathologists' reviews could have potentially further increased algorithm performance.

Fifth, all the data were collected retrospectively across the institutions and the general pathologist reviews were conducted in a nonclinical setting, without additional clinical information available at the time of review. Sixth, despite the international nature of our evaluation (in terms of both pathologists' practice and data sources), the countries involved were predominantly white, and demographic characteristics were not available for all datasets in the present study. Further investigation is required to validate the use of AI algorithms in more diverse settings^{35,36}. Last, this study did not evaluate the algorithm grading's association directly with radical prostatectomy or clinical outcomes.

We found that a group of AI Gleason grading algorithms developed during a global competition generalized well to intercontinental and multinational cohorts with pathologist-level performance. On all external validation sets, the algorithms achieved high agreement with uropathologists and high sensitivity for malignant biopsies. The performance exhibited by this group of algorithms adds evidence of the maturity of AI for this task and warrants evaluation of AI for prostate cancer diagnosis and grading in prospective clinical trials. We foresee a future where pathologists can be assisted by algorithms such as these in the form of a digital colleague. To stimulate further advancement of the field, the full development set of 10,616 biopsies has been made publicly available for noncommercial research use <https://panda.grand-challenge.org/>.

Online content


Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-021-01620-2>.

Received: 5 March 2021; Accepted: 8 November 2021;
Published online: 13 January 2022

References

- Epstein, J. I. An update of the gleason grading system. *J. Urol.* **183**, 433–440 (2010).
- Mohler, J. L. et al. Prostate cancer, version 2.2019, NCCN clinical practice guidelines in oncology. *J. Natl Compr. Canc. Netw.* **17**, 479–505 (2019).
- van Leenders, G. J. L. H. et al. The 2019 international society of urological pathology (ISUP) consensus conference on grading of prostatic carcinoma. *Am. J. Surg. Pathol.* **44**, e87 (2020).
- Epstein, J. I. et al. 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.* **40**, 244–252 (2016).
- Pierorazio, P. M., Walsh, P. C., Partin, A. W. & Epstein, J. I. Prognostic Gleason grade grouping: data based on the modified Gleason scoring system. *BJU Int.* **111**, 753–760 (2013).
- Epstein, J. I. et al. A contemporary prostate cancer grading system: a validated alternative to the gleason score. *Eur. Urol.* **69**, 428–435 (2016).
- Allsbrook, W. C. Jr et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. *Hum. Pathol.* **32**, 74–80 (2001).
- Melia, J. et al. A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology* **48**, 644–654 (2006).
- Ozkan, T. A. et al. Interobserver variability in Gleason histological grading of prostate cancer. *Scand. J. Urol.* **50**, 420–424 (2016).
- Egevad, L. et al. Standardization of Gleason grading among 337 European pathologists. *Histopathology* **62**, 247–256 (2013).
- Goldenberg, S. L., Nir, G. & Salcudean, S. E. A new era: artificial intelligence and machine learning in prostate cancer. *Nat. Rev. Urol.* **16**, 391–403 (2019).
- Nir, G. et al. Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. *JAMA Netw. Open* **2**, e190442 (2019).
- Han, W. et al. Histologic tissue components provide major cues for machine learning-based prostate cancer detection and grading on prostatectomy specimens. *Sci. Rep.* **10**, 9911 (2020).
- Naggal, K. et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit. Med.* **2**, 48 (2019).
- Naggal, K. 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).
- Bulten, W. et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* **21**, 233–241 (2020).
- Ström, P. et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol.* **21**, 222–232 (2020).
- Kott, O. et al. Development of a deep learning algorithm for the histopathologic diagnosis and gleason grading of prostate cancer biopsies: a pilot study. *Eur. Urol. Focus* **7**, 347–351 (2021).
- Steiner, D. F. 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).
- Bulten, W. et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod. Pathol.* <https://doi.org/10.1038/s41379-020-0640-y> (2020).
- Challen, R. et al. Artificial intelligence, bias and clinical safety. *BMJ Qual. Saf.* **28**, 231–237 (2019).
- Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
- AI diagnostics need attention. *Nature* **555**, 285–285 (2018).
- Yasaka, K. & Abe, O. Deep learning and artificial intelligence in radiology: current applications and future directions. *PLoS Med.* **15**, e1002707 (2018).
- Ehteshami Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
- Bandi, P. et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Trans. Med. Imaging* **38**, 550–560 (2019).
- Caicedo, J. C. et al. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nat. Methods* **16**, 1247–1253 (2019).
- Maier-Hein, L. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **9**, 5217 (2018).
- Bulten, W. et al. The PANDA challenge: Prostate cANcer graDe Assessment using the Gleason grading system. *Zenodo* <https://doi.org/10.5281/zenodo.3715938> (2020).
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
- Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
- van der Laak, J., Ciompi, F. & Litjens, G. No pixel-level annotations needed. *Nat. Biomed. Eng.* **3**, 855–856 (2019).
- Hartman, D. J., Van Der Laak, J. A. W. M., Gurcan, M. N. & Pantanowitz, L. Value of public challenges for the development of pathology deep learning algorithms. *J. Pathol. Inform.* **11**, 7 (2020).
- Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
- Rajkumar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **169**, 866–872 (2018).
- Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.* **178**, 1544–1547 (2018).

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

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adap-

tation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the

article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

the PANDA challenge consortium

Américo Brilhante²⁰, Aslı Çakır²¹, Xavier Farré²², Katerina Geronatsiou²³, Vincent Molinié²⁴, Guilherme Pereira²⁵, Paromita Roy²⁶, Günter Saile²⁷, Paulo G. O. Salles²⁸, Ewout Schaafsma¹, Joëlle Tschui²⁹, Jorge Billoch-Lima³⁰, Emílio M. Pereira³¹, Ming Zhou³², Shujun He³³, Sejun Song³⁴, Qing Sun³³, Hiroshi Yoshihara³⁵, Taiki Yamaguchi³⁶, Kosaku Ono³⁷, Tao Shen³⁸, Jianyi Ji³⁹, Arnaud Roussel⁴⁰, Kairong Zhou⁴¹, Tianrui Chai⁴², Nina Weng⁴³, Dmitry Grechka⁴⁴, Maxim V. Shugaev⁴⁵, Raphael Kiminya⁴⁶, Vassili Kovalev⁴⁷, Dmitry Voynov⁴⁷, Valery Malyshev⁴⁷, Elizabeth Lapo⁴⁷, Manuel Campos⁴⁸, Noriaki Ota⁴⁹, Shinsuke Yamaoka⁴⁹, Yusuke Fujimoto⁵⁰, Kentaro Yoshioka⁵¹, Joni Juvonen⁵², Mikko Tukiainen⁵², Antti Karlsson⁵³, Rui Guo⁵⁴, Chia-Lun Hsieh⁵⁵, Igor Zubarev⁵⁶, Habib S. T. Bukhar⁵⁷, Wenyuan Li⁵⁸, Jiayun Li⁵⁸, William Speier⁵⁸, Corey Arnold⁵⁸, Kyungdoc Kim⁵⁹, Byeonguk Bae⁵⁹, Yeong Won Kim⁵⁹, Hong-Seok Lee⁵⁹ and Jeonghyuk Park⁵⁹

²⁰Salomão Zoppi Diagnostics/DASA, São Paulo, Brazil. ²¹Pathology Department, School of Medicine, Istanbul Medipol University, Istanbul, Turkey.

²²Department of Health, Public Health Agency of Catalonia, Lleida, Spain. ²³Centre de Pathologie 68, Hopital Diaconat Mulhouse, Groupe Hospitalier de la Région Mulhouse Sud Alsace, Mulhouse, France. ²⁴Aix en Provence Hospital, Aix en Provence, France. ²⁵Histo Patologia Cirúrgica e Citologia, João Pessoa-PB, Brazil. ²⁶Department of Pathology, Tata Medical Center, Kolkata, India. ²⁷Abteilung für Histopathologie und Zytologie, Goldach, Switzerland.

²⁸Instituto Mário Penna, Belo Horizonte, Brazil. ²⁹Medics Pathologie, Bern, Switzerland. ³⁰HRP Labs, San Juan, PR, USA. ³¹Department of Pathology, Oncoclínicas group, São Paulo, São Paulo, Brazil. ³²Department of Pathology and Laboratory Medicine, Tufts Medical Center, Boston, MA, USA. ³³Texas A&M University, College Station, TX, USA. ³⁴Sungnam-ci, Republic of Korea. ³⁵Department of Health Informatics, Kyoto University, Kyoto, Japan.

³⁶Preferred Networks Inc., Tokyo, Japan. ³⁷Nowcast Inc., Tokyo, Japan. ³⁸School of Biological Science and Medical Engineering, Southeast University, Nanjing, China. ³⁹CTAccel Ltd., ShenZhen, China. ⁴⁰Jumio Corp., Montréal, Canada. ⁴¹ELEME Inc., Shanghai, China. ⁴²School of Computer Science and Engineering, Beihang University, Beijing, China. ⁴³DTU Compute, Technical University of Denmark, Lyngby, Denmark. ⁴⁴Moscow, Russia. ⁴⁵Department of Materials Science and Engineering, University of Virginia, Charlottesville, VA, USA. ⁴⁶Nairobi, Kenya. ⁴⁷Biomedical Image Analysis Department, The United Institute of Informatics Problems, Minsk, Belarus. ⁴⁸Madrid, Spain. ⁴⁹Systems Research & Development Center, Technology Bureau, NS Solutions Corp., Kanagawa, Japan. ⁵⁰Rist Inc., Tokyo, Japan. ⁵¹Wireless System Lab., Toshiba Corp., Kawasaki, Japan. ⁵²Silo AI, Turku, Finland. ⁵³University of Turku, Turku, Finland. ⁵⁴University of Michigan, Ann Arbor, MI, USA. ⁵⁵Taipei City, Taiwan. ⁵⁶Tula, Russia. ⁵⁷Janelia Research Campus, Ashburn, VA, USA.

⁵⁸Computational Diagnostics Lab, University of California, Los Angeles, Los Angeles, CA, USA. ⁵⁹VUNO Inc., Seoul, Republic of Korea.

Methods

Study design. The study design of the PANDA challenge was preregistered³⁹. We retrospectively obtained and de-identified digitized prostate biopsies with associated diagnosis from pathology reports from Radboud University Medical Center, Nijmegen, the Netherlands and Karolinska Institutet, Stockholm, Sweden (Extended Data Fig. 1 and Supplementary Methods 1, 2 and 3). At the start of the competition, participating teams gained access to this EU development set of 10,616 biopsies from 2,113 patients for training of the AI algorithms (Table 1 and Supplementary Methods 4). During the course of the competition, the teams could upload their algorithms to the Kaggle platform (Supplementary Methods 5) and receive performance estimates on a tuning set of 393 biopsies. Processing time was limited to 6 h and the maximum graphics processing unit (GPU) memory available was 16 GB.

By the competition closing date, each team picked two algorithms of their choice for their final submission, and the higher scoring of the two determined the team's final ranking. The final evaluation was performed on an internal validation dataset of 545 biopsies, collected from the same sites as the development and tuning sets and fully blinded to the participating teams. Moreover, to obtain an independent internal validation set, all samples from a given patient were used for either development or validation.

After the competition on the Kaggle platform ended, all teams were invited to send in a proposal to join the validation phase of the study as members of the PANDA consortium. Joining the validation phase was fully voluntary and not a prerequisite for partaking in the competition. As a result, 15 teams were selected for further evaluation on two external validation datasets consisting of 741 and 330 biopsies, also fully blinded to the participating teams (Supplementary Methods 6 and 7). The first external validation set was obtained from two independent medical laboratories and a tertiary teaching hospital in the USA. The second external validation set was obtained from the Karolinska University Hospital, Stockholm, Sweden. All datasets consisted of both benign biopsies and biopsies with various ISUP GGs. For details on the inclusion and exclusion criteria, see Supplementary Methods 1 and Extended Data Fig. 1.

WSIs of the biopsies were obtained using four different scanner models from three vendors: 3DHISTECH, Hamamatsu Photonics and Leica Biosystems (Supplementary Table 1). The open source ASAP software (v.1.9: <https://github.com/computationalpathologygroup/ASAP>) was used to export the slides before uploading to the Kaggle platform.

The present study was approved by the institutional review board of Radboud University Medical Center (IRB 2016-2275), Stockholm regional ethics committee (permits 2012/572-31/1, 2012/438-31/3 and 2018/845-32) and Advarra (Columbia, MD; Pro00038251). Informed consent was provided by the participants in the Swedish dataset. For the other datasets, informed consent was waived due to the usage of de-identified prostate specimens in a retrospective setting.

Reproducing algorithms and application to validation sets. All teams selected for the PANDA consortium were asked to provide all data and code necessary for reproducing the exact version of their algorithm that resulted in the final competition submission. For each algorithm, we collected the main Jupyter notebook or python script for running the inference, the specific Kaggle Docker³⁷ image (<https://github.com/Kaggle/docker-python>) used by the team during the competition and any necessary associated files, including model weights and auxiliary code.

We replicated the computational setup of the competition platform and ran the algorithms on two different computational systems: Google Cloud and Puhti compute cluster (CSC—IT Center for Science, Espoo, Finland). On the Google Cloud platform, all algorithms were run using the original Docker images. On Puhti, the Docker images were automatically converted for use with Singularity³⁸ (v.3.8.3). The algorithms and scripts provided by the teams were not modified except for minor adjustments required for successful run-time installation of dependencies on our computational systems. On Puhti, the algorithms had access to 8 central processing unit (CPU) cores, 32 GB of memory, 1 Tesla V100 32GB GPU (Nvidia) and 500 GB of SSD storage. On the Google Cloud platform, the algorithms had access to 8 CPU cores, 30 GB of memory, 1 Tesla V100 GPU 32GB and 10,000 GB of hard disk drive storage.

Before applying the algorithms on the external validation sets, we first validated that the Kaggle computational environment had been correctly replicated and the algorithms' performance on our systems remained identical. To this end, we ran all algorithms on the tuning and internal validation sets on the two systems to reproduce the output generated during the competition on the Kaggle platform. By crosschecking the new results with the competition leaderboard, we additionally assured that the algorithms supplied by the teams were not altered after the competition or tuned to perform better on the external validation sets. The verification runs we performed on the Puhti cluster were used as the basis for all results reported on the internal validation set.

Some algorithms were nondeterministic, for example, because of test time augmentations with nonfrozen random seeds. We ran each of these algorithms five times and averaged the computed metrics.

After verification, we ran all algorithms on the external validation sets. For the US external validation set, we used the Google Cloud platform. For the EU external

validation set, we used the Puhti cluster. This process was done independently of the teams and no prior information about the external datasets was supplied to the teams. The ISUP GG predictions of the algorithms on the cases were saved and used as input for the analysis.

Statistical analysis. We defined the main metric as the agreement on ISUP GG with the reference standard of each particular validation set, measured using quadratic Cohen's κ . To compare the performance of the algorithms with that of the general pathologists, we performed a two-sided permutation test per pathologist panel. The average agreement was calculated as the mean of the κ values across the algorithms and the pathologists, respectively. The test statistic was defined as the difference between the average algorithm agreement and the average pathologist agreement.

We calculated sensitivity and specificity on benign versus cancer-containing biopsies for all algorithms and individual general pathologists, based on the reference standard set by the uropathologists. To further understand how a representative pathologist and algorithm performed, we selected the pathologist and the algorithm with the median balanced accuracy (the average of sensitivity and specificity) as the representative pathologist and the representative algorithm, and reported the associated sensitivity and specificity. A representative pathologist or algorithm was used in favor of averaging across algorithms and pathologists for better estimates of performance. For the 95% CIs of the algorithms' and pathologists' performance metrics, we used bootstrapping across all algorithms or pathologists, with both the algorithm or pathologist and case as the resampling unit.

Analysis was performed using scripts³⁹ written in Python (v.3.8) in combination with the following software packages: scipy (1.5.4), pandas (1.1.4), mlxtend (0.18.0), numpy (1.19.4), scikit-learn (0.23.2), matplotlib (3.3.2), jupyterlab (2.2.9) and notebook (6.1.5).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The full development set, from here on named the PANDA challenge dataset, of 10,616 digitized de-identified hematoxylin and eosin-stained prostate biopsies, will be made publicly available for further research. The data can be used under a Creative Commons BY-SA-NC 4.0 license. To adhere to the 'Attribution' part of the license, we ask anyone who uses the data to cite the current article. The most up-to-date information regarding the dataset is available at the challenge website at <https://panda.grand-challenge.org>. Source data are provided with this paper.

Code availability

Code that was used to generate the results of the various algorithms, and example code on how to load the images in the PANDA dataset is available at <https://github.com/DIAGNijmegen/panda-challenge> and <https://doi.org/10.5281/zenodo.5592578>. Algorithms were built using open source deep learning frameworks, including Pytorch (<https://pytorch.org>) and TensorFlow (<https://www.tensorflow.org>). The Docker image that all the algorithms were based on is available online at <https://github.com/Kaggle/docker-python>. Details on the availability of specific models and the code of the contributed algorithms can be found in the Supplementary algorithm descriptions.

References

- Merkel, D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J.* **2014**, 2 (2014).
- Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: scientific containers for mobility of compute. *PLoS ONE* **12**, e0177459 (2017).
- Bulten, W. et al. PANDA challenge analysis code. *Zenodo* <https://doi.org/10.5281/zenodo.5592578> (2020).

Acknowledgements

We were supported by the Dutch Cancer Society (grant no. KUN 2015-7970, to W.B., H.P. and G.L.); Netherlands Organization for Scientific Research (grant no. 016.186.152, to G.L.); Google LLC, Verily Life Sciences, Swedish Research Council (grant nos. 2019-01466 and 2020-00692, to M.E.); Swedish Cancer Society (CAN, grant no. 2018/741, to M.E.); Swedish eScience Research Center, EIT Health, Karolinska Institutet, Åke Wiberg Foundation and Prostatacancerförbundet (all to M.E.); Academy of Finland (grant nos. 341967 and 335976, to P.Ruusuvuori), Cancer Foundation Finland (project 'Computational pathology for enhanced cancer grading and patient stratification', to P.Ruusuvuori) and ERAPerMed (grant no. 334782, 2020-22, to P.Ruusuvuori). Google LLC approved the publication of the manuscript, and the remaining funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the MICCAI board challenge working group, the MICCAI 2020 satellite event team and the MICCAI challenge reviewers for their support in the challenge workshop and review of the challenge design. We thank Kaggle for hosting the competition, providing compute resources and the competition prizes. We thank

CSC—IT Center for Science, Finland, for providing computational resources. We thank E. Wulczyn, A. Um'rani, Y. Liu and D. Webster for their feedback on the manuscript and guidance of the project. We thank our collaborators at NMCS, particularly N. Olson, for internal reuse of de-identified data, which contributed to the US external validation set.

Author contributions

W.B., K.Kartasalo and P.-H.C.C. 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. P.S., H.P. and K.N. contributed equally to this work. W.B., K.Kartasalo, P.-H.C.C., P.S., P.Ruusuvuori, M.E. and G.L. conceived the project. W.B., K.Kartasalo, P.-H.C.C., P.S., H.P., K.N., Y.C., D.F.S., H.v.B., R.V., C.H.-v.d.K., J.v.d.L., H.S., B.D., T.T., T.H., H.G., L.E., P.Ruusuvuori, G.L., M.E., A.B., A.Ç., X.F., K.G., V.M., G.P., P.Roy, G.S., P.G.O.S., E.S., J.T., J.B.-L., E.M.P., M.B.A., A.J.E., T.v.d.K., M.Z., R.A., P.A.H., M.V. and F.T. curated the data. W.B., K.Kartasalo, P.-H.C.C., K.N., P.Ruusuvuori, M.E. and G.L. did the formal analysis. P.-H.C.C., G.S.C., L.P., C.H.M., G.L., J.v.d.L., P.Ruusuvuori and M.E. acquired the funding. W.B., K.Kartasalo, P.-H.C.C., K.N., P.Ruusuvuori, M.E. and G.L. did the investigations. W.B., K.Kartasalo, P.-H.C.C., P.S., K.N., P.Ruusuvuori, M.E. and G.L. provided the methodology. W.B., K.Kartasalo and M.D. administered the project. W.B., D.F.S., S.D. and P.Ruusuvuori provided the resources. W.B., K.Kartasalo, H.P., P.-H.C.C., K.N., Y.C., S.D., M.V.S., M.Z., J.J., A.R., Y.F., K.Y., W.L., J.L., W.S., C.A., R.K., R.G., C.-L.H., Y.Z., H.S.T.B., V.K., D.V., V.M., E.L., H.Y., T.Y., K.O., T.S., T.C., N.W., N.O., S.Y., K.Kim, B.B., Y.W.K., H.-S.L., J.P., M.C., D.G., S.H., S.S., Q.S., J.J., M.T. and A.K. provided the software. P.-H.C.C., L.P., C.H.M., P.Ruusuvuori, M.E., G.L. and J.v.d.L. supervised the project. W.B., K.Kartasalo, P.-H.C.C., H.P. and K.N. validated the study. W.B. visualized the study. W.B., K.Kartasalo, P.-H.C.C., P.Ruusuvuori, M.E. and G.L. wrote the original draft of the manuscript. W.B., K.Kartasalo, P.-H.C.C., K.N., H.P., P.S., K.N., Y.C., D.F.S., H.v.B., R.V., C.H.-v.d.K., J.v.d.L., H.G., H.S., B.D., T.T., T.H., L.E., M.D., S.D., L.P., C.H.M., P.Ruusuvuori, M.E., G.L., A.B., A.Ç., X.F., K.G., V.M., G.P., P.Roy, G.S., P.G.O.S., E.S., J.T., J.B.-L., E.M.P., M.B.A., A.J.E., T.v.d.K., M.Z., R.A., P.A.H., M.V.S., J.J., A.R., K.Z., Y.F., K.Y., W.L., J.L., W.S., C.A., R.K., R.G., C.-L.H., I.Z., H.S.T.B., V.K., D.V., V.M., E.L., H.Y., T.Y., K.O., T.S., T.C., N.W., N.O., S.Y., K.Kim, B.B., Y.W.K., H.-S.L., J.P., M.C., D.G., S.H., S.S., Q.S., J.J., M.T. and A.K. reviewed and edited the manuscript.

Competing interests

W.B. and H.P. report grants from the Dutch Cancer Society, during the conduct of the present study. J.v.d.L. reports consulting fees from Philips, ContextVision and AbbVie, and grants from Philips, ContextVision and Sectra, outside the submitted work. G.L. reports grants from the Dutch Cancer Society and the NWO, during the conduct of the present study, and grants from Philips Digital Pathology Solutions and personal fees from Novartis, outside the submitted work. M.E. reports grants from Swedish Research Council, Swedish Cancer Society, Swedish eScience Research Center, EIT

Health, Karolinska Institutet, Åke Wiberg Foundation and Prostatancerförbundet. P.Ruusuvuori reports grants from Academy of Finland, Cancer Foundation Finland and ERAPerMed. H.G. has five patents (WO2013EP7425920131120, WO2013EP74270 20131120, WO2018EP52473 20180201, WO2015SE50272 20150311 and WO2013SE50554 20130516) related to prostate cancer diagnostics pending, and has patent applications licensed to A3P Biomedical. M.E. has four patents (WO2013EP74259 20131120, WO2013EP74270 20131120, WO2018EP52473 20180201 and WO2013SE50554 20130516) related to prostate cancer diagnostics pending, and has patent applications licensed to A3P Biomedical. P.-H.C.C., K.N., Y.C., D.F.S., M.D., S.D., F.T., G.S.C., L.P. and C.H.M. are employees of Google LLC and own Alphabet stock, and report several patents granted or pending on machine-learning models for medical images. M.B.A. reported receiving personal fees from Google LLC during the conduct of the present study and receiving personal fees from Precipio Diagnostics, CellMax Life and IBEX outside the submitted work. A.E. is employed by Mackenzie Health, Toronto. T.v.d.K. is employed by University Health Network, Toronto; the time spent on the project was supported by a research agreement with financial support from Google LLC. R.A. and P.A.H. were compensated by Google LLC for their consultation and annotations as expert uropathologists. H.Y. reports nonfinancial support from Aillix Inc. during the conduction of the present study. W.L., J.L., W.S. and C.A. have a patent (US 62/852,625) pending. K.Kim, B.B., Y.W.K., H.-S.L. and J.P. are employees of VUNO Inc. M.B.A. reported receiving personal fees from Google LLC during the conduct of the present study and receiving personal fees from Precipio Diagnostics, CellMax Life and IBEX outside the submitted work. A.E. is employed by Mackenzie Health, Toronto. T.v.d.K. is employed by University Health Network, Toronto; the time spent on the project was supported by a research agreement with financial support from Google LLC. M.Z., R.A. and P.A.H. were compensated by Google LLC for their consultation and annotations as expert uropathologists. All other authors declare no competing interests.

Additional information

Extended data Extended data are available for this paper at <https://doi.org/10.1038/s41591-021-01620-2>.

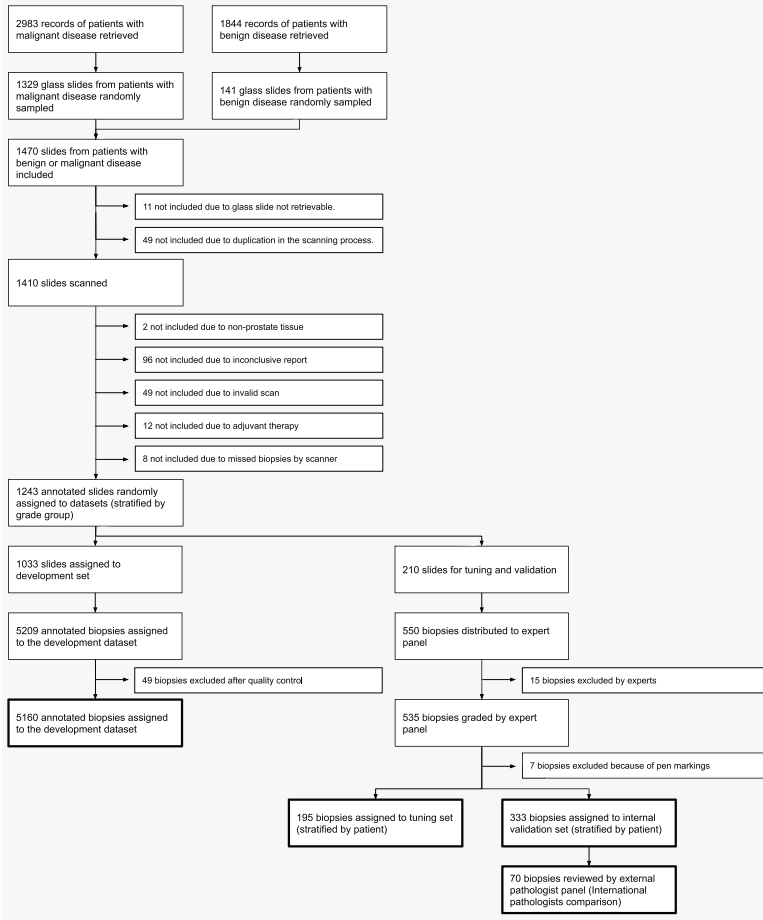
Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-021-01620-2>.

Correspondence and requests for materials should be addressed to Wouter Bulten, Kimmo Kartasalo or Po-Hsuan Cameron Chen.

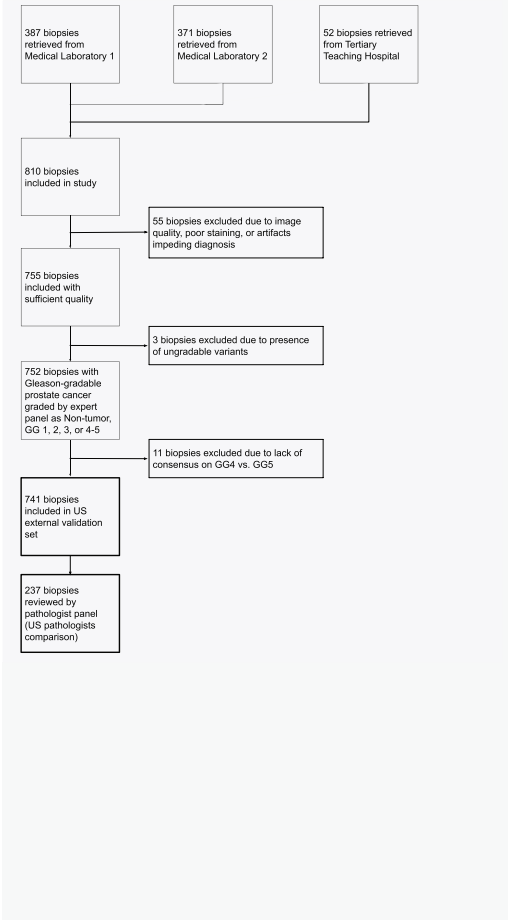
Peer review information *Nature Medicine* thanks Moritz Gerstung, Jonathan Epstein, Priti Lal and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Javier Carmona was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

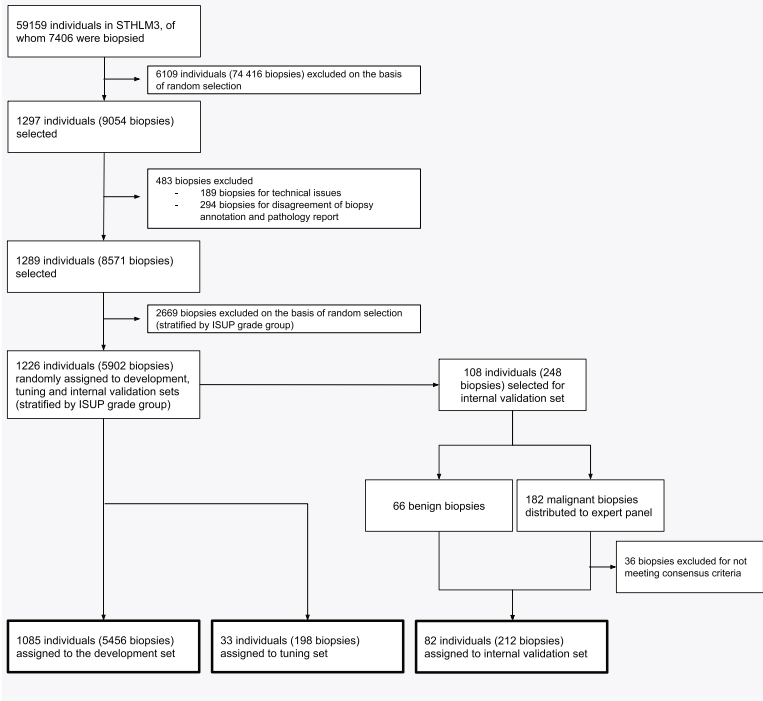
(A) Radboud University Medical Center



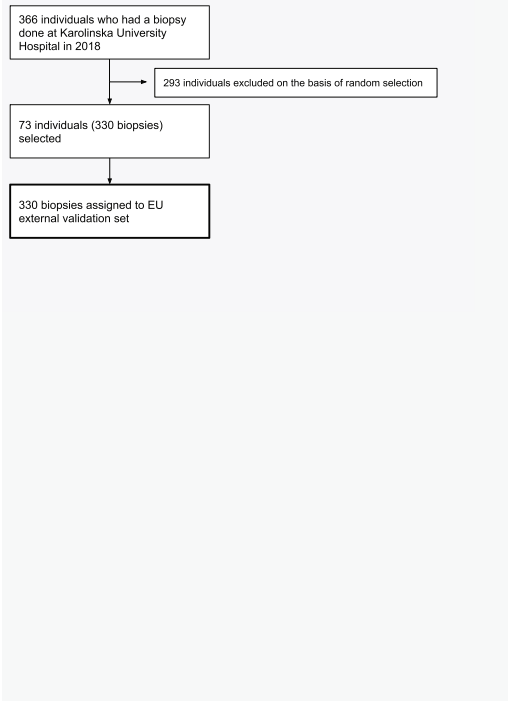
(C) United States



(B) Karolinska Institutet

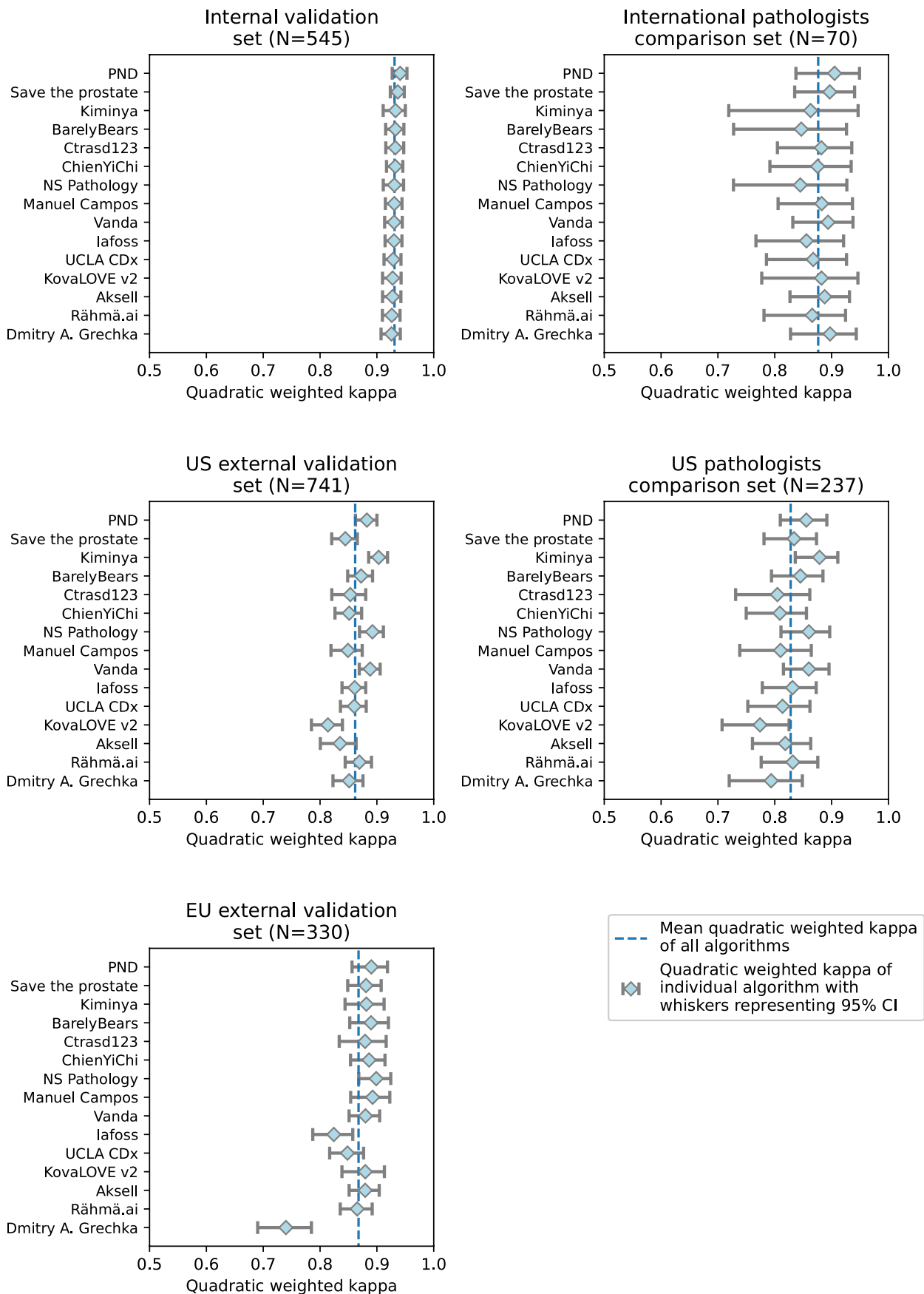


(D) Karolinska University Hospital

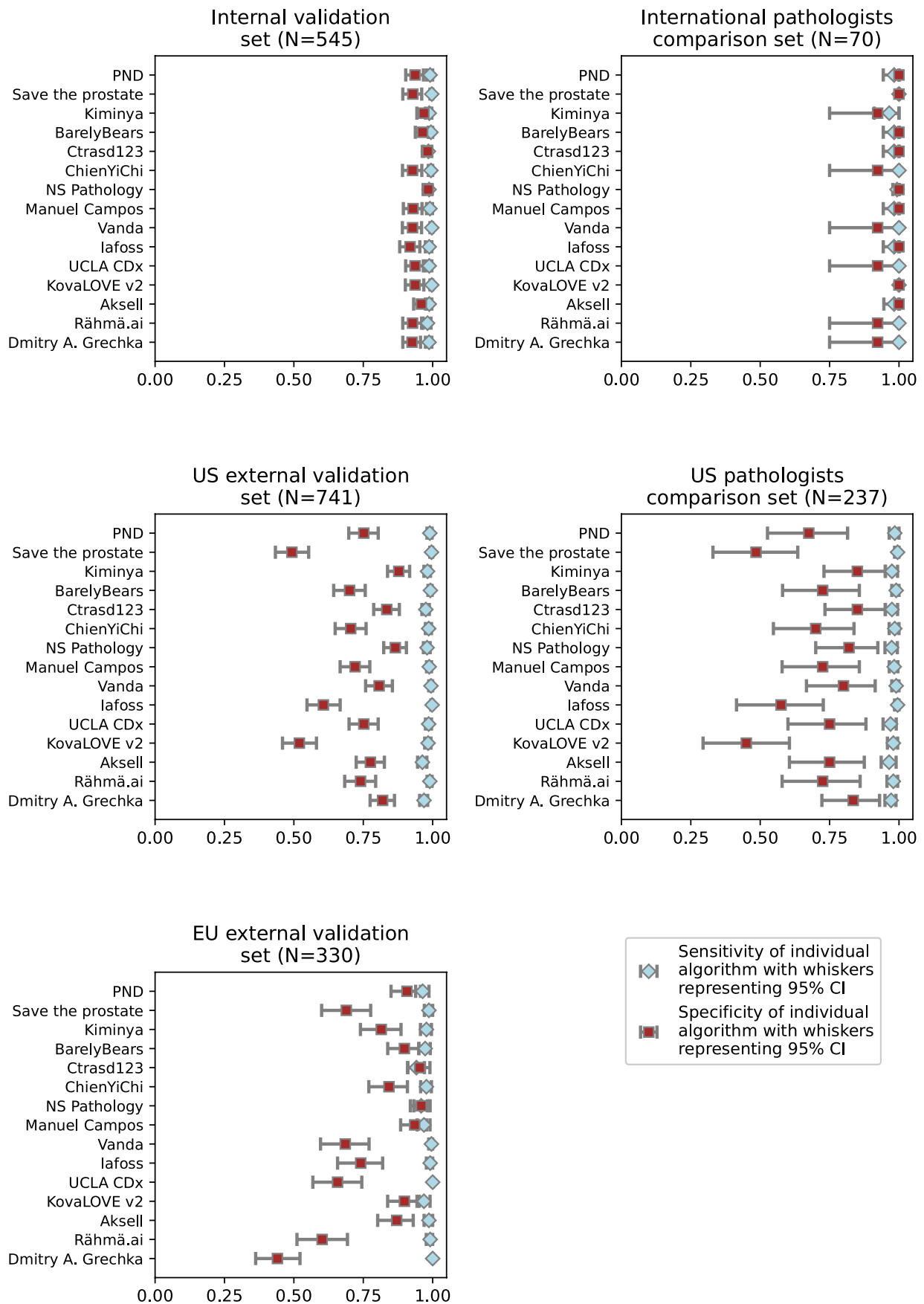


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Flow charts of inclusion and exclusion for the various datasets. (a) Data originating from Radboud University Medical Center (development, tuning and internal validation sets, and international pathologist comparison), (b) Data originating from Karolinska Institutet (development, tuning and internal validation sets), (c) Data originating from the United States (US external validation set, and US pathologist comparison), (d) Data originating from Karolinska University Hospital (EU external validation set).

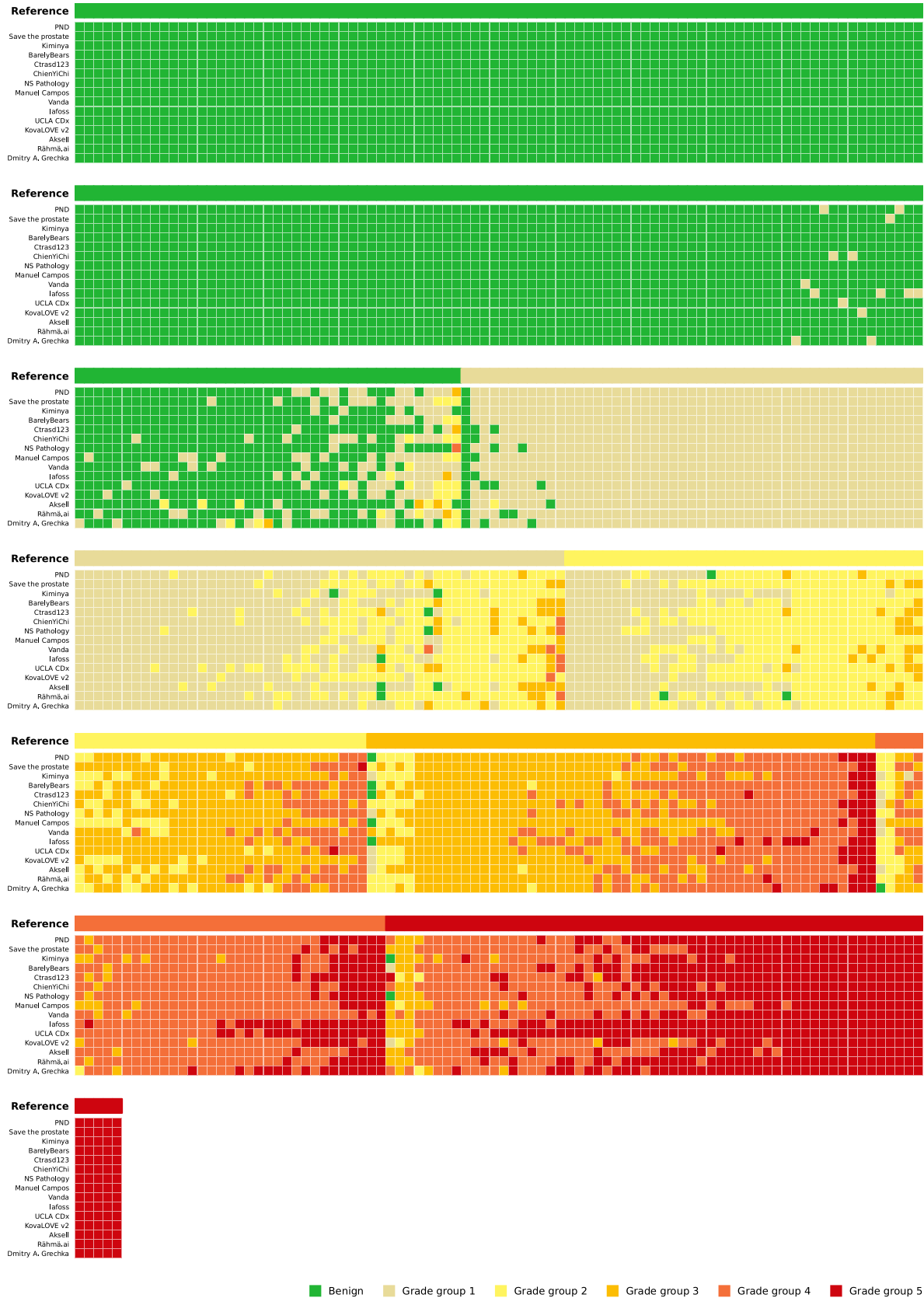


Extended Data Fig. 2 | Individual algorithms' agreement with the reference standard for the validation sets. Concordance with ISUP GG of the reference standard (Cohen's quadratically weighted kappa with 95% CI over cases) is shown for each algorithm on each validation set. The dashed line indicates the mean of all teams on the validation set in question.



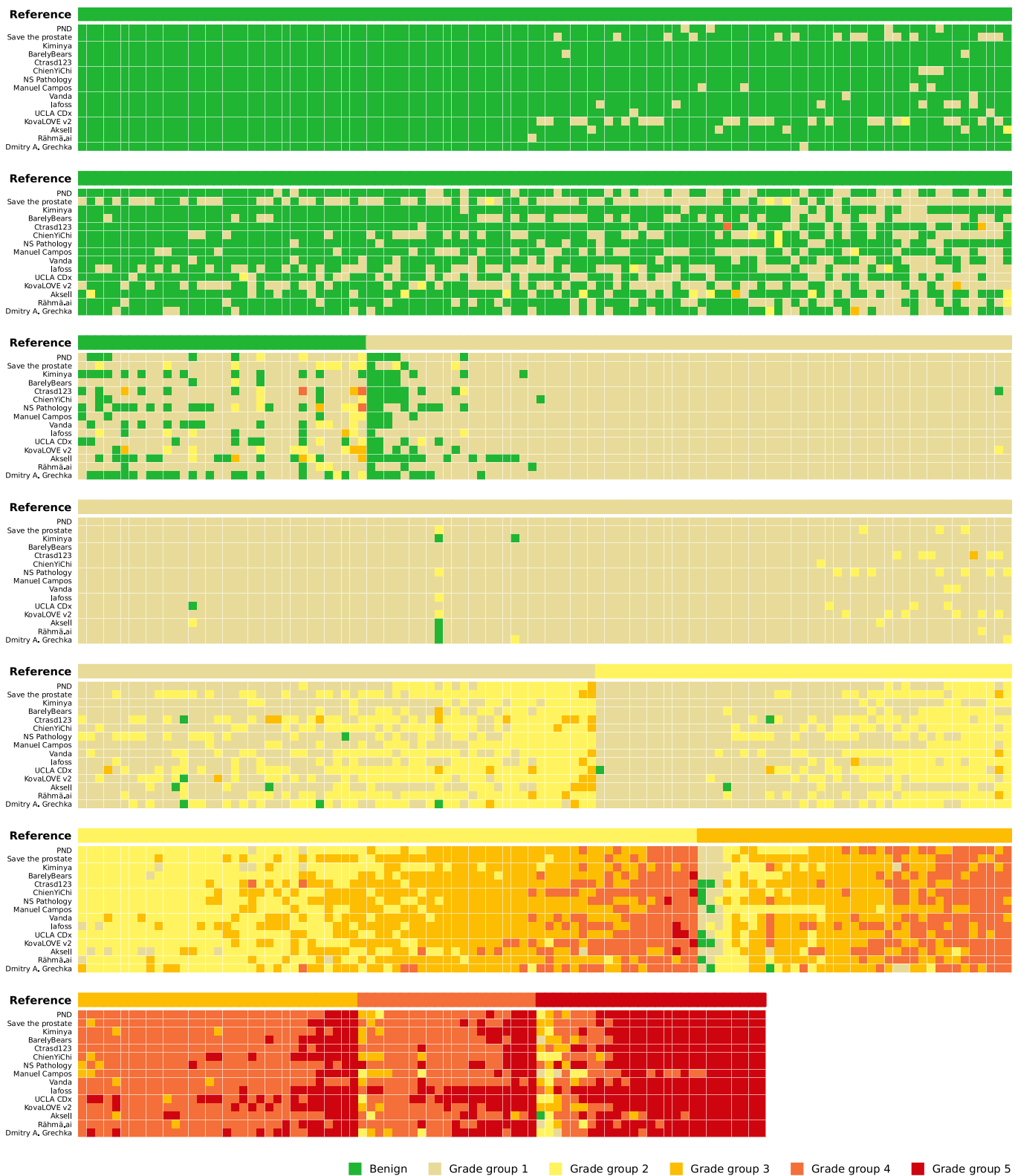
Extended Data Fig. 3 | Individual algorithms' Sensitivity and specificity for the validation sets. Performance in detecting biopsies containing cancer (sensitivity and specificity with 95% CI over cases) is shown for each algorithm on each validation set.

Grades provided by algorithms for Internal validation set (N=545)



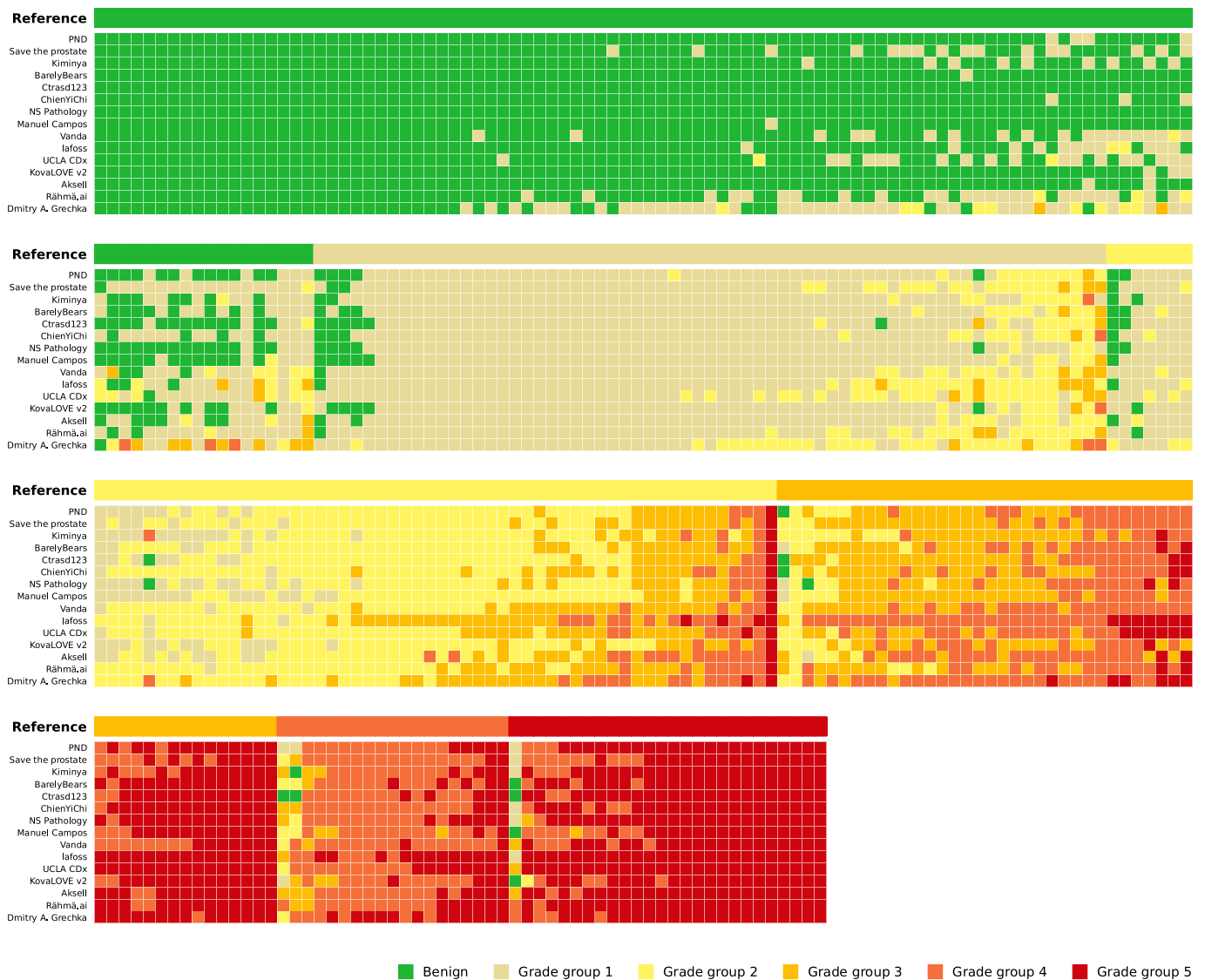
Extended Data Fig. 4 | Visualization of grade assignment by algorithms for the internal validation set. Cases are ordered by the reference ISUP grade group and average grade group of the AI cohort.

Grades provided by algorithms for US external validation (N=741)

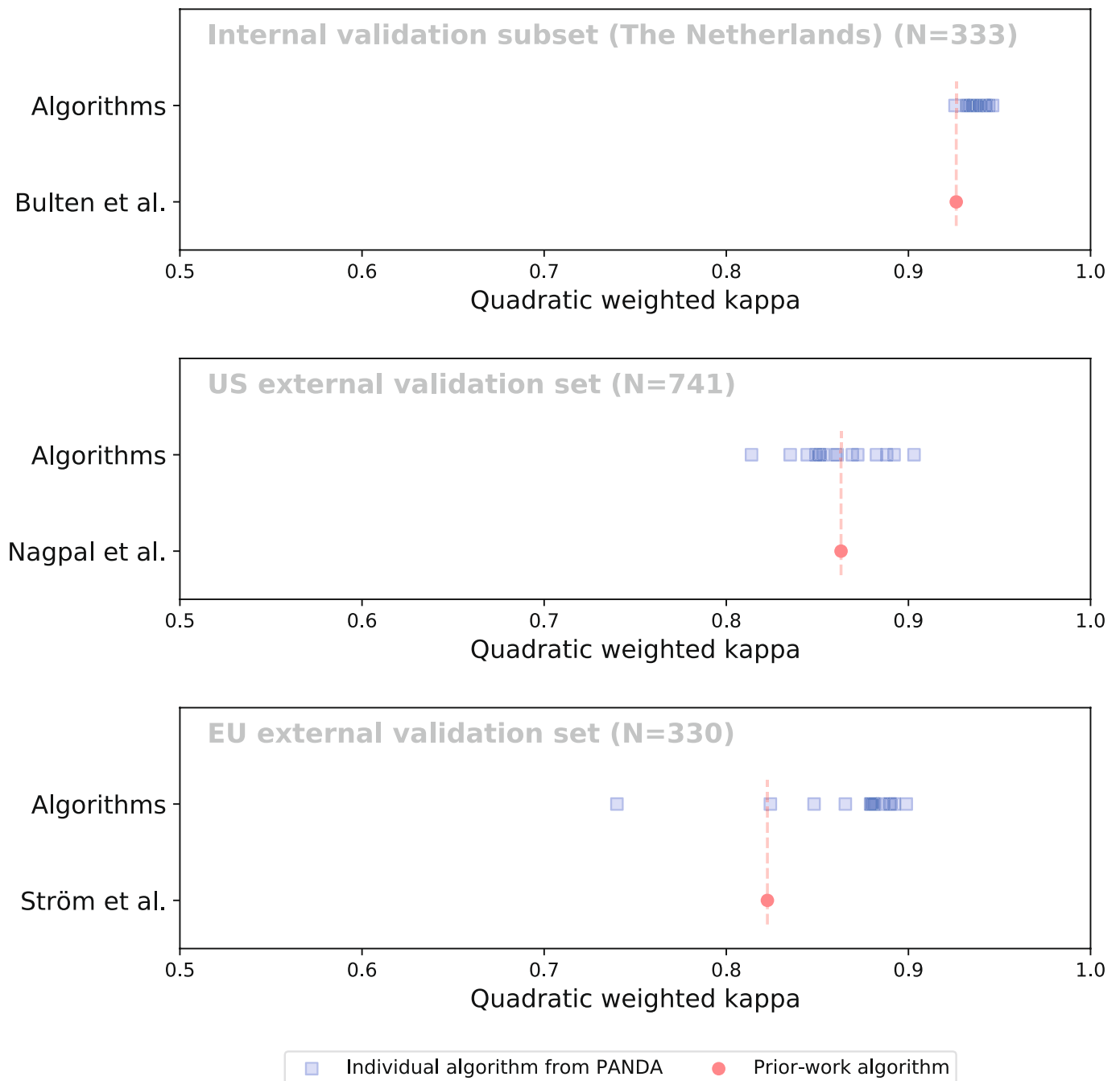


Extended Data Fig. 5 | Visualization of grade assignment by algorithms for the US external validation set. Cases are ordered by the reference ISUP grade group and average grade group of the AI cohort.

Grades provided by algorithms for EU external validation (N=330)



Extended Data Fig. 6 | Visualization of grade assignment by algorithms for the EU external validation set. Cases are ordered by the reference ISUP grade group and average grade group of the AI cohort.



Extended Data Fig. 7 | Comparison of challenge algorithms to prior work. The performance of the teams' algorithms was computed on validation (sub) sets of earlier work. For each validation set, we additionally show the performance of the original algorithm.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

After scanning the slides, data was converted and exported using the open source ASAP software (version 1.9, <https://github.com/computationalpathologygroup/ASAP>). Instructions on how to use the data are included in the following repository: <https://github.com/DIAGNijmegen/panda-challenge>

Data analysis

Analysis was performed using Python (version 3.8) in combination with the following software packages: scipy (1.5.4), pandas (1.1.4), mlxtend (0.18.0), numpy (1.19.4), scikit-learn (0.23.2), matplotlib (3.3.2), jupyterlab (2.2.9) and notebook (6.1.5).

Code for the analysis of algorithm performance is made publicly available through <https://github.com/DIAGNijmegen/panda-challenge/>

The Docker image that all the algorithms were based on is available online at <https://github.com/Kaggle/docker-python>. On the Puhti GPU cluster, the Docker images were automatically converted for use with Singularity (version 3.8.3). Details on the availability of specific models and the code of the contributed algorithms can be found in the supplementary algorithm descriptions.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The full development set, from here on named the PANDA Challenge dataset, of 10,616 digitized de-identified H&E stained prostate biopsies (383GB) will be made publicly available for further research. The data can be used under a Creative Commons BY-SA-NC 4.0 license. To adhere to the "Attribution" part of the license, we ask anyone who uses the data to cite the corresponding paper. The most up-to-date information regarding the dataset will be published on the challenge website at <https://panda.grand-challenge.org/>

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal sample size calculation was performed. We collected as many samples as possible, considering samples that were readily available in digital format across the institutions involved in this study, and taking into account that the dataset size needed to remain feasible to download (at < 500 GB) for competition participants. For the validation sets, we combined the cohorts of Bulten et al. LO 2020, Ström et al. LO 2020, and Nagpal et al. JAMA Onc 2020, resulting in a sample size far surpassing earlier work and capturing a wide range of the morphological heterogeneity present in prostate needle core biopsies.
Data exclusions	<p>During the data collection for the development, tuning, and validation sets, a total of 111 biopsies were excluded due to risk of information leakage (e.g. pathologist pen markings visible on the tissue), poor staining, or image quality issues. While establishing the reference standard, 65 biopsies were excluded due to a lack of consensus among the pathologists providing the reference standard. More details on exclusion criteria are displayed in the supplementary appendix.</p> <p>No data were excluded from the analysis. All algorithms selected for the study among the competition participants were included in the analysis. All selected algorithms produced results for all included cases.</p>
Replication	<p>All training data is made publicly available, which allows for replication of the algorithms. For several algorithms, the code has also been made open source. Details on the availability of specific models and the code of the contributed algorithms can be found in the supplementary algorithm descriptions. Code for the analysis of algorithm performance is made publicly available through https://github.com/DIAGNijmegen/panda-challenge/</p> <p>The authors independently reproduced all algorithms contributed by the challenge participants. All details on this process are described in the main text and supplementary materials.</p>
Randomization	Samples were per data provider randomly allocated into development (N = 10,616), tuning (N = 393) and internal validation (N = 545) sets, stratified by Gleason score. All samples from a given patient were assigned to the same set. External validation sets were collected fully independently from the development, tuning and internal validation data and were thus not part of the randomization process.
Blinding	The challenge organizers had access to all data in the study. The algorithm developers had no access to the data used for the validation of the algorithms. Algorithms were independently applied to the validation sets by the challenge organizers, without the involvement of the original developers. Running the algorithms on the validation sets was done programmatically without manual intervention or any algorithmic modifications by the challenge organizers. The algorithms' output was fixed and stored in a repository before the statistical analysis was performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

All cases were retrospectively collected histological H&E stained tissue sections of prostate biopsy specimens, acquired from men who underwent a biopsy procedure due to suspicion of prostate cancer in one of the six institutions included in the study. The patient age distribution was as follows. EU internal validation set (Karolinska Institutet): <54 y (3.7%), 55-59 y (11.0%), 60-64 y (23.2%), 65-69 y (58.5%), >= 70 y (3.7%); EU external validation set (Karolinska University Hospital): <54 y (9.5%), 55-59 y (13.7%), 60-64 y (16.4%), 65-69 y (20.5%), >= 70 y (39.7%); US external validation set medical laboratory 1: <65 y (44.2%), >= 65 y (51.6%), not available (4.2%); US external validation set medical laboratory 2: <65 y (41.0%), >= 65 y (57.5%), not available (1.6%). Further details are provided in the supplementary appendix.

Recruitment

Cases were retrospectively included at random, sourced through three independent studies, across six sites. For the Radboud data, we retrieved all pathology reports dated between Jan 1, 2012, and Dec 31, 2017, for patients who underwent a prostate biopsy owing to a suspicion of prostate cancer. Patients were randomly sampled based on the highest reported Gleason score mentioned in each report. Additionally, a set of reports was sampled which only mentioned benign biopsies. The data from Karolinska comes from the Stockholm-3 diagnostic trial that was conducted between May 28, 2012 and Dec 30, 2014, (ISRCTN84445406). It was a prostate cancer screening-by-invitation trial of men aged 50–69 years living in Stockholm, Sweden. The purpose of the trial was to compare prostate specific antigen (PSA) to the Stockholm-3 model (S3M) for predicting the presence of cancer, and the criterion for referral to biopsy was either PSA above 3 ng/ml or a S3M probability of 10% or higher. A random sample from the biopsies included in the trial was taken, stratified on patient and the reported Gleason score to avoid including too many of the prevalent benign and low grade diseases. The US external validation set consisted of retrospective cases from three different sources. Briefly, cases were obtained from two medical laboratories and one tertiary teaching hospital. All tumor-containing cases available from the tertiary teaching hospital from 2005–2007 were included, and a fraction of the benign biopsies available were randomly sampled for inclusion. From the medical laboratories, all available ISUP grade group 4–5 cases were included in the study, and remaining benign and ISUP grade group 1–3 cases were randomly sampled for inclusion. The EU external validation set comprised biopsy cores assessed by L.E. at the Karolinska University Hospital during 2018. The set included all positive biopsy cores from all men diagnosed with an ISUP grade group 2, 3, 4, or 5 cancer as well as from a random selection of men diagnosed with ISUP grade group 1 cancer during that time period. In addition, the set included all cores from a random selection of men with only benign biopsies.

Ethics oversight

The study was approved by the institutional review board of Radboud University Medical Center (IRB 2016–2275), Stockholm regional ethics committee (permits 2012/572-31/1, 2012/438-31/3, and 2018/845-32), and Advarra (Columbia, MD; Pro00038251).

Note that full information on the approval of the study protocol must also be provided in the manuscript.