Radiology: Imaging Cancer

ORIGINAL RESEARCH

Improving Clinically Significant Prostate Cancer Detection with a Multimodal Machine Learning Approach: A Large-Scale Multicenter Study

Ana Carolina Rodrigues, PhD[*1,2] • José Guilherme de Almeida, PhD*[*1] • Nuno Rodrigues, PhD[1,3] • Raquel Moreno, MD[1] • Ana Sofia Castro Verde, MSc[1] • Ana Mascarenhas Gaivão, MD[4] • Carlos Bilreiro, MD, PhD[4] • Inês Santiago, MD, PhD[4] • Joana Ip, MD[4] • Sara Belião, MD[4] • Sara Silva, PhD[3] • Inês Domingues, PhD[5,6] • Manolis Tsiknakis, PhD[7] • Konstantinos Marias, PhD[7] • Daniele Regge, MD, PhD[8] • Nikolaos Papanikolaou, PhD[1] • for the ProCAncer-I Consortium*

*A.C.R. and J.G.d.A. contributed equally to this work.

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

Radiology: Imaging Cancer 2025; 7(5):e240507 • https://doi.org/10.1148/rycan.240507 • Content codes: GU OI AI

Purpose: To develop and prospectively validate a clinical and radiologic model to predict clinically significant prostate cancer (csPCa) using biparametric MRI (bpMRI).

Materials and Methods: Retrospective data (acquired before March 31, 2022) from 12 medical centers were collected. Radiomic features were extracted from the whole prostate gland using segmentations generated by an automatic deep learning algorithm. A model incorporating bpMRI radiomics, age, prostate-specific antigens, the Prostate Imaging Reporting and Data System (PI-RADS), and the prostate zone lesion location was trained. A retrospective validation set and prospective data (acquired after March 31, 2022) were used to compare PI-RADS scoring (area under the receiver operating characteristic curve [AUC] and specificity at PI-RADS >3). Sensitivity analyses for sequence (T2-weighted, apparent diffusion coefficient, diffusion-weighted imaging) and scanner vendor (GE, Philips, Siemens) were performed, in addition to fairness analyses for relevant categories.

Results: The retrospective dataset for model development included 7157 male patients (mean age, 64.78 years; 3342 [46.7%] with csPCa), and the prospective dataset for model validation included 1629 patients (mean age, 66.19 years; 592 [36.3%] with csPCa). The multimodal model outperformed PI-RADS in the retrospective (AUC, 0.88 vs 0.80, $P = .005$; specificity of 71% vs 58%, $P = .002$) and prospective validation sets (AUC, 0.91 vs 0.85, $P < .001$; specificity of 77% vs 66%, $P < .001$), leading to 22.7% fewer biopsies compared with PI-RADS. Sensitivity analyses showed the importance of multiple sequences and vendors in achieving model generalization, as using specific sequences or vendors alone led to worse performance. Fairness analysis showed generalizability across different categories but highlighted increased sensitivity with higher PI-RADS and reduced performance in one medical center.

Conclusion: A multimodal model provided a temporally generalizable predictor of csPCa that outperformed PI-RADS.

Supplemental material is available for this article.

Prostate cancer (PCa) has the second highest incidence among cancers in men worldwide and ranks fifth in terms of mortality (1,2), posing a substantial burden on health care systems worldwide. An essential imaging modality in the clinical management of prostate cancer is biparametric MRI (bpMRI). It is recommended as a primary diagnostic approach for individuals with a suspected clinical diagnosis of PCa (3,4). The current standard of care recommends using the Prostate Imaging Reporting and Data System (PI-RADS) score to analyze prostate bpMRI (5), but this has a relatively high rate of false positives (6), which can lead to unnecessary biopsy procedures, overtreatment, and increased risks and discomfort for patients.

Ongoing advancements in the fields of artificial intelligence and radiomics (7–10)—a computational method used for extracting and conducting statistical analysis of imaging features that are not readily visible to the human eye—are continuously shaping the field of clinical predictive models. Using complex mathematical algorithms, radiomic features (such as shape and texture-related information) can be extracted from MR images (11). This field shifts the focus from the visual assessment of medical images to a more objective quantitative evaluation, which can complement the present standard of care.

The use of radiomics across the spectrum of PCa is frequently reported in the literature (12), especially in relation to the detection of clinically significant PCa (csPCa). Traditionally, PCa diagnosis is based on the histologic analysis of a systematic biopsy specimen, where a Gleason score (GS) is assigned by a pathologist. More recently, the International Society of Urological Pathology (ISUP) pointed out the significant differences in patient outcome in GS of 3+4 and 4+3 (13), deciding to restructure the GS into the ISUP grades: ISUP = 1 corresponding to GS of 3+3 or lower; ISUP = 2 corresponding to GS of 3+4; ISUP = 3 corresponding to GS of 4+3; ISUP = 4 corresponding to GS of 8; ISUP = 5 corresponding to GS of 9 or higher. Overall, progression and relapse-free survival has been reported as significantly associated with ISUP grades (14–16). Some studies have reported that ISUP grade 2 can be considered low risk (17), but others have shown that significant differences in progression-free survival exist between ISUP grades 1 and 2 (14). This threshold—ISUP = 1 versus ISUP = 2, 3, 4, or 5—can

Abbreviations

AUC = area under the receiver operating characteristic curve, ADC = apparent diffusion coefficient, bpMRI = biparametric MRI, csPCa = clinically significant prostate cancer, DWI = diffusion-weighted imaging, ERC = endorectal coil, FOV = field of view, GS = Gleason score, ISUP = International Society of Urologic Pathology, mpMRI = multiparametric MRI, PI-RADS = Prostate Imaging Reporting and Data System, PSA = prostate-specific antigen, SHAP = Shapley additive explanation

Summary

A multimodal machine learning model combining radiomic, radiologic, and clinical variables reduced unnecessary prostate biopsies while maintaining high sensitivity for clinically significant cancer detection.

Key Points

- In this study of 8786 men, a multimodal machine learning model outperformed standard Prostate Imaging Reporting and Data System scoring in predicting clinically significant prostate cancer and reduced unnecessary biopsies by 20.2% (806 of 1031 vs 670 of 1031; $P < .001$) in prospective validation.
- Sensitivity analyses demonstrated that models trained on single-vendor data or that used only MRI sequences showed clinically significant and reduced performance (mean AUC decrease of 0.18; $P < .001$) compared with multivendor, multimodal approaches.
- While the model showed consistent performance across age groups, scanner vendors, and field strengths, performance was reduced in settings with wide field of view diffusion sequences (sensitivity decreased by 15%; $P < .001$).

Keywords

Algorithm Development, Machine Learning, Model Validation, Model Training, Genital/Reproductive, Neoplasms-Primary, Oncology, Comparative Studies, Technology Assessment

be considered a better determinant of clinical significance as it has the least amount of undetected clinically significant disease (18) and is motivated by other works in csPCa detection using MRI (19,20).

In this work, we developed a multimodal machine learning model incorporating clinical, demographic, radiologic semantic, and agnostic radiomic features using multicentric data to predict csPCa. We evaluated the impact of MRI scanner vendor, presence or absence of endorectal coil (ERC), and inclusion of clinical variables on model performance. To demonstrate temporal generalizability, our multimodal model was validated with contemporary data (from both internal and external centers) against PI-RADS scoring. Fairness analyses were performed to understand the cause and potential cases of failure in our model.

Materials and Methods

Ethics committee approval, as well as informed consent waivers (for retrospective data) and patient consent (for prospective data), were obtained independently at each clinical site.

Study Sample

Our training dataset consisted of retrospective bpMRI examinations comprising T2-weighted, apparent diffusion coefficient (ADC), and diffusion-weighted imaging (DWI) sequences from the ProstateNet image archive (21) created under the scope of the ProCAncer-I project, as illustrated in Figure 1. The data came from 12 different clinical centers, nine countries, and three MRI scanner manufacturers (Table S1). The inclusion criteria were *(a)* bpMRI including high-spatial resolution T2-weighted imaging and a high *b* value ($b > 1000$ sec/mm^2) DWI (acquisition parameters in Table S2); *(b)* histology results (either biopsy or prostatectomy) or a minimum 1-year clinical follow-up for individuals with no disease evidence at baseline bpMRI; *(c)* 18 years or older at the time of diagnosis; and *(d)* collection date before March 31, 2022. Prospectively collected data used during validation followed identical histology inclusion criteria, with the exception of those who had a collection date after March 31, 2022. Cases were considered positive if they had a biopsy with ISUP higher than 1, and negative if they had a biopsy with no lesion or ISUP = 1 or no biopsy and a minimum 1-year follow-up with no evidence of disease (either normal prostate-specific antigen [PSA] or negative MRI).

Whole-Gland Segmentation

Automatic segmentation of the whole prostate gland was performed on T2-weighted sequences, based on an artificial intelligence model developed in prior work (22) trained on ProstateNet data (21) with a custom protocol (Appendix S1). Given that Rodrigues et al (22) used 638 studies from ProstateNet to develop prostate segmentation models and that we used ProstateNet in this study, the training data from Rodrigues et al (22) (542 studies) were used exclusively during training, ensuring there was no data leakage. T2-weighted sequences were coregistered to DWI, and the calculated transformation matrix was applied to the segmentation mask generated on T2-weighted images using nearest-neighbor interpolation (Fig S1). For wide field-of-view (FOV) examinations where the coregistration algorithm failed to converge, a center crop was applied on the x- and y-axis to 240 mm. The transformed mask was then used for the radiomics extraction of DWI and ADC. The full parameters can be found in Appendix S1. A radiologist (R.M., with 8.5 years of experience) rated the segmentation quality of 125 randomly selected studies considering both T2-weighted and DWI (Appendix S1).

Feature Extraction

Bias field correction was performed on T2-weighted sequences using the N4ITK algorithm (23,24). Since the studies are anisotropic, only in-plane features were considered. Additionally, x and y spacings were resampled to the 95th quantile value for both T2-weighted images and high *b* value DWI and ADC (0.6875 and 2.0, respectively). The bin width was selected for each image transform to produce an average of 80 bins, as recommended in the literature (25). Radiomic features were extracted using the PyRadiomics package (version 3.0) (26); the PyRadiomics configuration is available in Table S3. A total of 1223 features pertaining to whole prostate gland shape, texture, and first-order features were calculated per sequence. The clinical variables included were age (continuous), PI-RADS (1–5 for increasing lesion severity in multiparametric MRI [mpMRI] as evaluated by clinical radiologists at each institution), PSA (continuous), and index lesion anatomic location (a set of four indicator variables for central zone, peripheral zone, transition zone, and anterior stroma). Missing PSA values were imputed using k-nearest neighbors imputation.

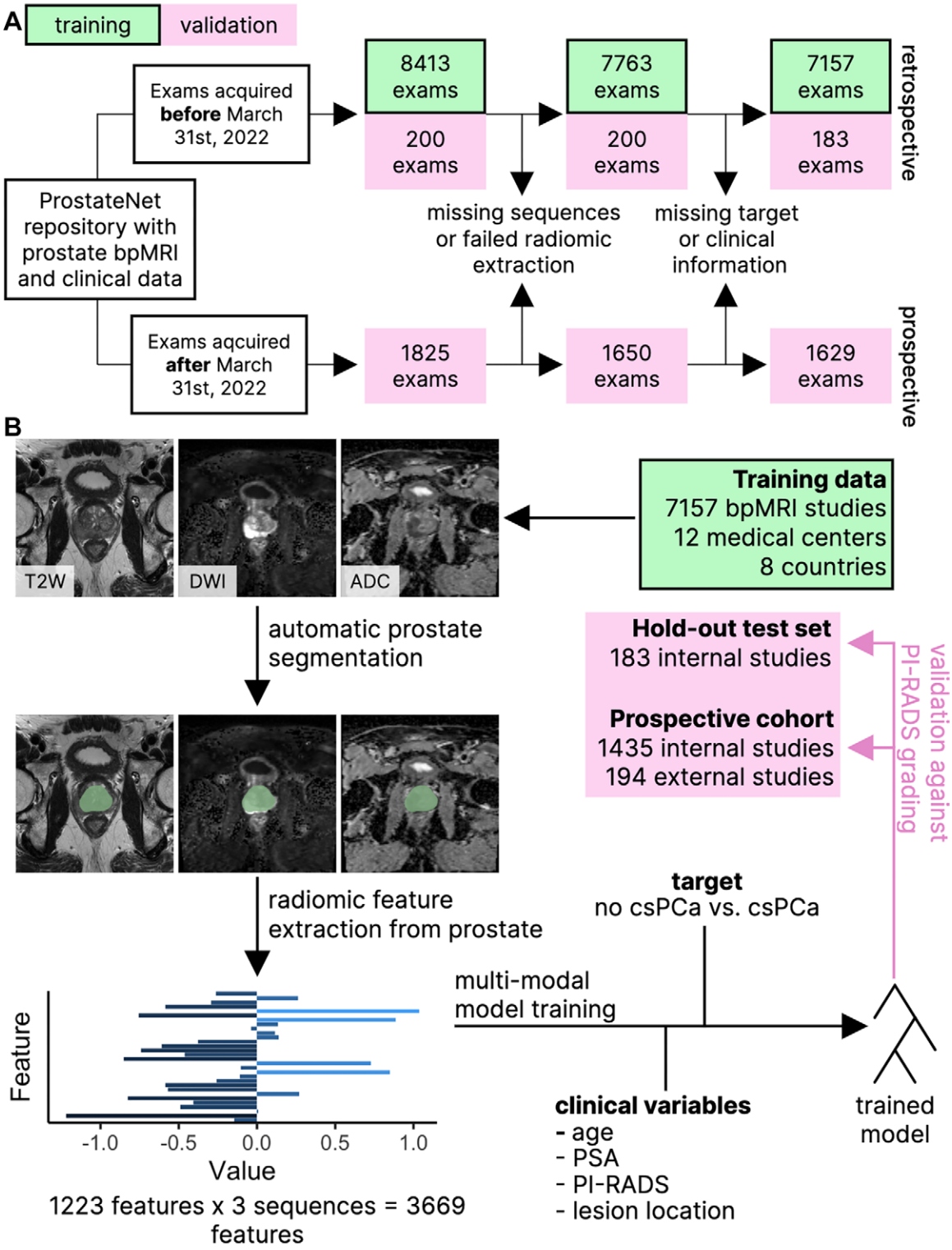

Figure 1: **(A)** Flowchart of the inclusion and exclusion criteria of retrospective and prospective biparametric MRI (bpMRI) examinations in ProstateNet. Exclusion criteria were missing sequence in the bpMRI (either T2W [T2-weighted images], DWI [diffusion-weighted imaging], or ADC [apparent diffusion coefficient maps]), file corruption, coregistration failure, and missing clinical or target information. **(B)** Schematic representation of our modeling and validation protocol. After studies are retrieved from ProstateNet, prostate segmentation masks are automatically generated from T2-weighted imaging and coregistered to DWI and ADC. Radiomic features are extracted from all series. These, together with clinical variables (age, prostate serum antigen [PSA], Prostate Imaging Reporting and Data System [PI-RADS], lesion location), were used to train a radiomics model predicting either clinically significant prostate cancer (csPCa) or no csPCa. Validation was performed against PI-RADS scoring using two datasets—a retrospective validation set and a prospective cohort including 194 external studies (also available in ProstateNet).

Multimodal Model Development

Throughout our work, a single study acquired before any potential treatment intervention was included for each patient, so there were no risks of data leakage. From the totality of the retrospective data, two distinct datasets were constructed—one for development and another for validation (retrospective validation set). To do so,

a random set of 200 studies, stratified according to ISUP grade distribution, was selected for the retrospective validation set, while the remaining studies were used for model development.

On the training set, numerical variables were standardized, constant features were excluded, and highly correlated features (Spearman coefficient > 0.8) were removed (if two features had Spearman coefficient > 0.8, we removed the feature with the highest average correlation across all features). For radiomics-only models (described later for sensitivity analysis), a light gradient boosting machine (27) was trained; whereas for models incorporating both radiomics and clinical features, a CatBoost (28) algorithm was used. Hyperparameter tuning was performed for each algorithm with a random search using nested cross-validation (Appendix S1).

Sensitivity Analysis

While the primary model developed in our work comprises radiomics, semantic, and clinical features, other models were developed to understand *(a)* the impact of clinical variables on our model, *(b)* the impact of having large amounts of diverse data in terms of ERC use and scanner vendor, and *(c)* the impact of using only specific sequences during training and inference. To do this, different data subsets were tested for their predictive performance by stratifying the data into six different subsets, considering scanner manufacturer and ERC usage (GE, Philips, Siemens, GE with no ERC, all vendors, and all vendors with no ERC). This resulted in 24 training combinations (six scanner/ERC × four volumes). The analysis of these results was performed using a linear model where the area under the receiver operating characteristic curve (AUC) is the dependent variable, and the sequence, feature set, and manufacturer are the independent variables. The discriminated training set sizes are shown in Table S4. Given that the clinical centers providing data for prospective validation used different PI-RADS versions (Table S5), we compared model performance stratified according to PI-RADS version use (centers not using PI-RADS version 2 vs all centers) and dynamic contrast enhancement use (all centers vs only centers using dynamic contrast enhancement to derive PI-RADS) for both our model and the baseline performance (PI-RADS).

Statistical Analysis

We compared age and PSA among training, retrospective testing, and prospective testing using Wilcoxon tests, and ISUP, lesion location, PI-RADS, normal follow-up PSA, country, and biopsy type using a χ^2 test. To ground our assessment in the current best radiologic practice, we used PI-RADS >3 as the classification threshold for a possible indication of csPCa (which can be confirmed through a biopsy), which represents the standard of care. We selected three distinct assessments: the AUC, the specificity at PI-RADS >3 sensitivity (proxy for unnecessary biopsy sparing), and the sensitivity at PI-RADS >3 specificity (proxy for false-negative reduction). AUC, specificity, and sensitivity comparisons were performed using the DeLong test and bootstrapping (29), respectively, with the pROC package for the R programming language (R Foundation for Statistical Computing). These give a general performance metric for our model and quantify the unnecessary biopsy-sparing potential of our method, respectively. A Shapley additive explanation (SHAP) analysis (Python package SHAP, version 0.42.1) (30) was used to identify the extent to which features or groups of features are important for prediction. We made use of the additivity of SHAP values to calculate feature group SHAP values for each instance. We also performed a calibration curve analysis to understand whether our model is well-calibrated using the CalibrationCurves package for R (31) and a decision curve analysis using the dcurves package for R (32,33) to better understand the added benefit at detecting csPCa and avoiding biopsies. We assessed both the added net benefit and biopsies avoided at a threshold corresponding to PI-RADS >3 sensitivity. This allowed us to quantify how our model could improve on PI-RADS without excluding patients from potentially life-saving treatment. We used a statistical significance threshold of .05.

For validation, two distinct datasets were used—the aforementioned retrospective validation set randomly constructed from the retrospective data distribution obtained before March 31, 2022, and a prospective validation set collected from consecutive patients from March 31, 2022, to August 2024. The extraction of radiomic features was performed following the same method and parameters described for the retrospective data. Prospective cases with missing PI-RADS scoring were assigned PI-RADS = 1 if no lesions were detectable in the MRI examination (MRI negative) and excluded otherwise. We additionally used the prospective validation set to establish a learning curve—how performance evolves as the size of the training dataset increases by sampling percentages between 1% and 100% (1%, 2%, 5%, 10%, 25%, 50%, 75%, 100%).

Exploratory fairness analyses—a performance analysis on subgroups focusing on different categories that can create disparities—were performed for the prospective data. We considered country, scanner vendor, age, ERC use, diffusion FOV, and PI-RADS on the model developed during this work; we excluded categories with fewer than 50 cases. Given that no Greek centers were used during training and were only used in the prospective cohort (194 of 1825 [10.6%]), we considered the performance on Greek studies to be an external validation proxy.

The METRICS statement is available in Table S6; please note that our approach achieves an "excellent" score of 94.6% (34).

Results

Study Sample Characteristics

The total training dataset included 8596 patients, 7778 of whom had a complete bpMRI with radiomic features (cases were excluded due to file corruption or coregistration failure). After exclusion of cases in which clinical features or ground truths (ISUP) were unavailable, a total of 7157 training samples remained. Of the 200 individuals from the retrospective validation cohort (not included in the previously described 8596 patients), 17 were excluded due to missing PI-RADS. Finally, of 1825 individuals in the prospective cohort, 196 were excluded due to missing data or missing follow-up information (Fig 1A). We observed a wide variety of scanner manufacturers and models, as well as a variety of receive coils and magnetic field strength (Table S7). The analysis of predicted segmentations performed by a radiologist showed that these were correct in most cases (no low-quality T2-weighted segmentations and only 11 of 125 [9%] low-quality DWI segmentations; Appendix S1, Tables S8 and S9).

Table 1: Description of Clinical Variables across All Cohorts

Variable	Training	Retrospective Validation	*P* Value (Training vs Retrospective)	Prospective Validation	*P* Value (Training vs Prospective)
Age (y)			.34		<.001
Mean (SD)	64.78 (7.78)	64.13 (8.03)		66.19 (7.98)	
PSA (ng/mL)			.16		.34
Mean (SD)	11.30 (40.01)	10.94 (15.34)		11.18 (52.03)	
ISUP (Gleason grade group)			.081		<.001
0 (not assessed)	3364 (40.0)	58 (29.0)		761 (41.7)	
1	1383 (16.4)	34 (17.0)		385 (21.1)	
2	1982 (23.6)	56 (28.0)		370 (20.3)	
3	858 (10.2)	26 (13.0)		156 (8.5)	
4	368 (4.4)	11 (5.5)		77 (4.2)	
5	458 (5.4)	15 (7.5)		76 (4.2)	
Lesion location			.50		<.001
Peripheral zone	4332 (51.5)	116 (58.0)		1048 (57.4)	
Transitional zone	1124 (13.4)	29 (14.5)		259 (14.2)	
Anterior stroma	295 (3.5)	4 (2.0)		86 (4.7)	
Central zone	256 (3.0)	5 (2.5)		61 (3.3)	
PI-RADS			.052		<.001
1	3216 (38.2)	58 (29.0)		557 (30.5)	
2	96 (1.1)	1 (0.5)		90 (4.9)	
3	517 (6.1)	11 (5.5)		138 (7.6)	
4	2449 (29.1)	68 (34.0)		599 (32.8)	
5	2135 (25.4)	62 (31.0)		441 (24.2)	
Normal follow-up PSA (for negative cases only)			.38		<.001
No	1947 (23.1)	36 (18.0)		541 (29.6)	
Yes	1404 (16.7)	21 (10.5)		242 (13.3)	
Country			<.0001		<.001
France	294 (3.5)	2 (1.0)		90 (4.9)	
Greece	NA	NA		194 (10.6)	
Italy	1088 (12.9)	30 (15.0)		260 (14.2)	
Lithuania	619 (7.4)	18 (9.0)		214 (11.7)	
the Netherlands	2196 (26.1)	73 (36.5)		79 (4.3)	
Portugal	928 (11.0)	20 (10.0)		157 (8.6)	
Spain	781 (9.3)	6 (3.0)		356 (19.5)	
Turkey	1929 (22.9)	29 (14.5)		440 (24.1)	
UK	578 (6.9)	22 (11.0)		35 (1.9)	
Biopsy type (if available)			.17		<.001
In-bore	194 (2.3)	5 (2.5)		27 (1.5)	
Systematic	1156 (13.7)	29 (14.5)		442 (24.2)	
Fusion	1411 (16.8)	55 (27.5)		101 (5.5)	
Systematic plus fusion	1242 (14.8)	34 (17.0)		179 (9.8)	
Complete	7157	183		1629	
Total	8413	200		1825	

Note.—Unless otherwise indicated, data are numbers of cases, with percentages in parentheses. Statistical testing (*P*) was performed using Wilcoxon tests for numerical variables (age, prostate-specific antigen [PSA]), and the χ^2 test was performed for categorical variables (International Society of Urological Pathology [ISUP], lesion location, Prostate Imaging Reporting and Data System [PI-RADS], normal follow-up, PSA, country, biopsy type). "Complete" represents the cases that were complete and available for model training testing. Complete = biparametric MRI and clinical variables, NA = not applicable.

The mean ages for the retrospective training and prospective validation cohorts were 64.78 and 66.19 years, respectively (Table 1). Differences in clinical variables were observed between the training and retrospective validation cohorts for ISUP distribution, lesion location, and country, and between the training and prospective validation in terms of age, ISUP, lesion location, whether individuals with negative MRI had normal PSA at follow-up, biopsy type, and country (Table 1). There was no evidence of a difference in terms of PSA.

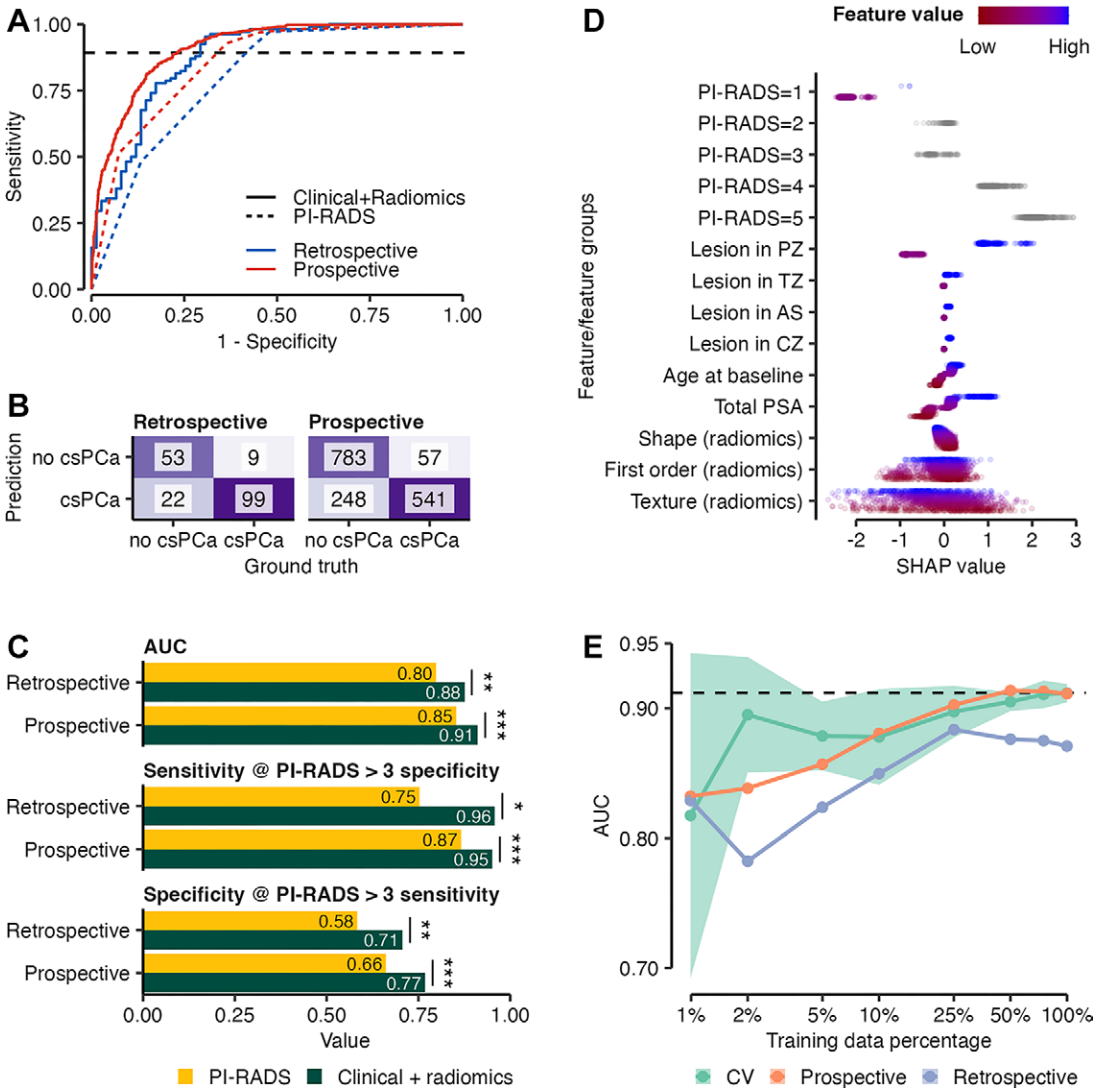

Figure 2: **(A)** Receiver operating characteristics curve of our model compared with Prostate Imaging Reporting and Data System (PI-RADS) on both the retrospective validation set (blue) and prospective validation set (red). The dashed line corresponds to the sensitivity of PI-RADS >3. **(B)** Confusion matrices for the retrospective and prospective cohorts. **(C)** Comparison of PI-RADS and the clinical plus radiomics models in terms of area under the receiver operating characteristic curve (AUC), sensitivity of PI-RADS >3 specificity, and specificity of PI-RADS >3 sensitivity stratified according to retrospective or prospective cohorts. **(D)** Shapley additive explanation (SHAP) values for contribution of features for model predictions on the prospective cohort. Each dot in the graph represents the SHAP value of a feature (clinical features) or of a group of features (radiomic features) for one observation. Positive and negative SHAP values correspond to contributions to classification as clinically significant prostate cancer (csPCa) or non-csPCa, respectively. **(E)** Learning curve for cross-validation (CV), retrospective, and prospective model performance. The shaded green area represents the SD of the CV performance estimate. *P < .05. **P < .01. ***P < .001. AS = anterior stroma, CZ = central zone, PZ = peripheral zone, TZ = transition zone.

Artificial Intelligence Reduces Unnecessary Biopsies

Our multimodal machine learning model (Fig 1B) shows remarkable performance, improving on the standard-of-care threshold of PI-RADS >3 (Fig 2A–2C). Compared with PI-RADS scoring, our model offers improvements—for the retrospective validation set, performance improved in both the retrospective validation set (AUC, 0.88 [95% CI: 0.82, 0.93] for our model and AUC, 0.80 [95% CI: 0.73, 0.86] for PI-RADS, P = .005) and the prospective validation set (AUC, 0.91 [95% CI: 0.90, 0.93] for our model and AUC, 0.85 [95% CI: 0.84, 0.87] for PI-RADS, P < .001).

Assuming a sensitivity identical to that of PI-RADS >3 in ProstateNet (89%), we observed improved specificity in the retrospective validation for our model (0.71; specificity of 58% for PI-RADS >3, P = .002) and the prospective validation sets for our model (77%; specificity of 66% for PI-RADS >3, P < .001). Concretely, at identical sensitivity, our model would have reduced false positives (unnecessary biopsies) by 22.7% (54 vs 44 true negatives for our model vs PI-RADS >3, respectively) and 20.1% (806 vs 670 true negatives for our model vs PI-RADS >3, respectively, of 1031 total negatives) in the retrospective and prospective testing cohorts, respectively.

On the other hand, at a specificity identical to that of PI-RADS >3, we observed improved sensitivity both on the retrospective validation for our model and for PI-RADS >3 (96% vs 75%, respectively, P = .003) and in the prospective test set (95% vs 87%, respectively, P < .001). In other words, an additional 5.1% (104 vs 99 true positives) and 2.7% (573 vs 558 true positives) of csPCa cases would have been detected in the retrospective and prospective validation cohorts, respectively, by our model when compared with using PI-RADS >3 alone.

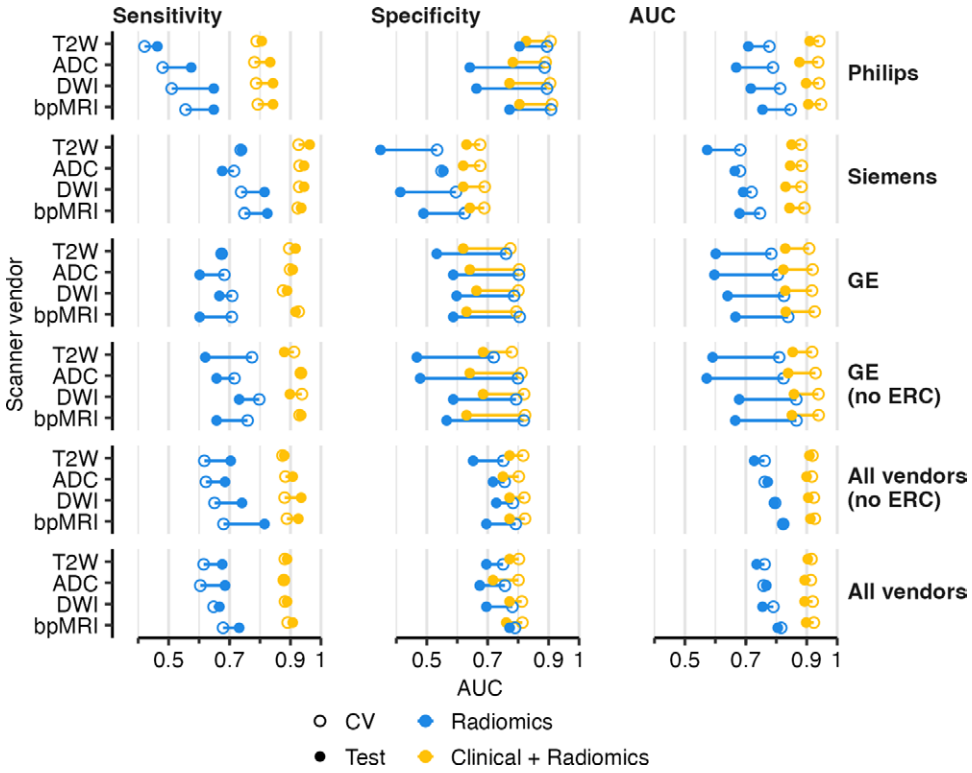

Figure 3: Cross-validation (CV) and retrospective validation set area under the receiver operating characteristic curve (AUC) sensitivity, specificity, and the AUC for models trained to predict clinical significance (ISUP 2–5), stratified according to sequence type radiomics used during training, vendor, and whether clinical features were used to train the model. ADC = apparent diffusion coefficient, bpMRI = biparametric MRI (combined T2W, ADC and DWI), DWI = diffusion-weighted imaging, ERC = endorectal coil, ISUP = International Society of Urological Pathology, T2W = T2-weighted.

While this is sufficient motivation for the use of multimodal features, we undertook a SHAP analysis of feature importance considering clinical and semantic features separately and by aggregating radiomic features by feature type (Fig 2D). This analysis shows that, while PI-RADS and lesion in the peripheral zone offered larger contributions, both first-order and texture features contribute to prediction (given their large absolute SHAP values). Conversely, shape and lesion location features (with the exception of peripheral zone lesion location) contribute little. In addition, according to the SHAP analysis, the model made reasonable assumptions by increasing the probability of clinically significant disease when the PI-RADS score increases (Fig 2D).

Our learning curve analysis (Fig 2E) shows performance saturating at approximately 50% of training data (3578 cases), with the error associated with the cross-validation estimate decreasing as training data size increases. Using calibration analysis, we also observed that our model overestimates the risk with a Brier score of 0.13 (Fig S2A). A decision curve analysis showed that, when compared with the threshold corresponding to PI-RADS >3 sensitivity, there was an added net benefit of 0.15 (Fig S2B) and an added net biopsies avoided of 0.16 (Fig S2C).

Multiple Vendors and Sequences Are Crucial for Generalizability

Multicentric, diverse datasets such as ProstateNet facilitate the retraining of models with specific scanner vendors and ERC use. As shown in Figure 3, more significant performance drops between cross-validation and testing were observed when models were trained on specific scanner vendors, indicating that generalization to testing data with multiple vendors requires the training data to have a similar composition in terms of scanner vendors. Interestingly, when training with data from all vendors, including or excluding ERC has little impact on the performance of the retrospective validation set (Table 2).

Additionally, we tested how sequence-specific radiomics contribute to prediction. On their own, there was a general trend for radiomic features in terms of used sequence—performance was best when all sequences (bpMRI) or DWI alone is used (Table 2). We also note that radiomic models do not perform particularly well on their own, with multimodal models performing higher (on average >0.177 AUC). Finally, we observed no difference between assessing performance of all centers versus centers using exclusively PI-RADS version 2.1: AUC of 0.85 versus AUC of 0.86 (P = .79) for the PI-RADS model, and AUC of 0.91 versus AUC of 0.92 (P = .98) for our model. Furthermore, comparing models using all centers systematically using dynamic contrast enhancement does not lead to performance differences either: AUC of 0.85 versus AUC of 0.85 (P = .68); and AUC of 0.91 versus AUC of 0.91 (P = .71) for PI-RADS and our model, respectively.

Fairness Analysis Shows Consistent Performance

Performance consistency across different centers, scanner vendors, ages, ERC use, and PI-RADS is important, as this can help us understand when to apply the model to different subcohorts. We observed stable performance across various age ranges, ERC use, and scanner vendor categories, highlighting broad generalizability to these categories (Fig 4). However, variations in diffusion FOV

Table 2: Features Contributing to AUC in Detection of Clinically Significant Prostate Cancer

Feature	Estimate	SEE	*t* Value	*P* Value
Intercept	0.764	0.014	54.1	<.001*
Sequence (vs bpMRI)				
T2-weighted	−0.037	0.013	−2.9	.006*
ADC	−0.035	0.013	−2.7	.009*
DWI	−0.012	0.013	−1.0	.34
Feature set (vs radiomics only)				
Clinical plus radiomics	0.177	0.009	19.8	<.001*
Vendors (vs all vendors)				
Philips	−0.027	0.015	−1.7	.093
Siemens	−0.084	0.015	−5.4	<.001*
GE	−0.104	0.015	−6.7	<.001*
GE (no ERC)	−0.09	0.015	−6.0	<.001*
All vendors (no ERC)	0.011	0.015	0.7	.48

Note.—Linear model coefficient estimates (Estimate), standard error of the estimate (SEE), and associated *t* value and *P* value. The linear model was parameterized as the retrospective validation set area under the receiver operating characteristic curve (AUC) as a function of training sequence, feature set, vendors, and an intercept term. ADC = apparent diffusion coefficient, bpMRI = biparametric MRI (combined T2-weighted, ADC, and DWI), DWI = diffusion weighted imaging, ERC = endorectal coil.
* Indicates statistical significance for a two-sided *t* test.

lead to reduced sensitivity. When considering country, we observed a drop in performance for Lithuania across both AUC and sensitivity ($P < .001$ for both) and a drop in AUC for Spain ($P = .009$). While the latter is likely to be associated with the relatively lower prevalence of negative cases in data coming from these sites, the former was likely attributable to centers from Lithuania (NCI) acquiring diffusion sequences (DWI and ADC) with a considerably wider FOV than that of T2-weighted, which leads to poor segmentation results, as illustrated by our analysis of segmentation quality (Appendix S1); this highlights the importance of careful inspection of outputs by radiologists. For PI-RADS, we noted that—as expected—there is an inverse relationship between specificity and sensitivity as PI-RADS increases, an outcome of using this feature as part of the classifier. Finally, we had access to lesion size for 150 cases in the prospective cohort. When comparing individuals whose lesion size was greater than 1.5 cm with those whose lesion size was 1.5 cm or smaller, we observed no evidence of a difference in sensitivity ($P = .32$), indicating that the effect of small lesions in csPCa classification was negligible for our model.

Discussion

In our work, the added value of bpMRI prostate radiomics to the standard of care is reported for the prediction of PCa clinical significance. Our study shows physicians could potentially reduce over 20% of unnecessary biopsies for csPCa diagnosis by using a relatively simple and easily trainable multimodal machine learning model that combines clinical and demographic variables, semantic features defined by visual inspection from a human radiologist, and agnostic radiomic features. Additionally, this model does not require any lesion annotation masks.

The study demonstrated consistently high predictive performance, which is in line with previous studies for csPCa prediction using large multicentric datasets. Saha et al showed that a deep learning detection model trained on a large dataset (>9000 cases) could simultaneously increase the specificity and sensitivity when fixing the sensitivity and specificity at specific PI-RADS operating points (19). Through this work, we showed that combining standard radiologic evaluation with radiomic features leads to comparable AUC (0.91 for Cai et al [20] and 0.88 and 0.91 for our work on retrospective and prospective cohorts, respectively). Additionally, we presented not only a more holistic prospective validation but also an external prospective validation, as well as sensitivity analyses into the effect of scanner vendor and ERC use and fairness analysis to better understand where our model can fail. Finally, our calibration curve analysis demonstrates that our model, while overestimating risk, does so at a relatively small level, and our decision curve analysis showed that both the added net benefit and the added biopsies avoided demonstrated the improvement of our model over PI-RADS.

Previous studies have also shown that combining information from standard radiologic practice (ie, PI-RADS, radiology reports) outperforms both automatic methods and radiologists (20,35). Here, we confirmed this added benefit by showing that the combination of PI-RADS, lesion location, clinical features, and radiomic features leads to state-of-the-art performance. Importantly, our model is designed to support radiologists rather than replace them. It builds on PI-RADS scoring and is more interpretable, thanks to SHAP analysis of individual feature contributions. In contrast to deep learning models, which often act as "black boxes" and may face resistance from radiologists due to their opacity, our approach provided explainability and aligned with existing clinical workflows, encouraging broader adoption by radiologists.

The noninvasive method presented in this study for clinical significance classification outperforms PI-RADS, the clinical standard for noninvasive clinical significance risk assessment. Additionally, there are relatively few biomarkers available for csPCa detection. PSA density has been described as a potentially useful biomarker (36); at a threshold of 0.185 (37), a sensitivity of 0.55 and a specificity of 0.70 can be expected; this is a significantly lower performance than the values presented here. Nonetheless, combining several different biomarkers may lead to improved sensitivity, as illustrated in our work.

Our radiomic pipeline considers the whole prostate gland, unlike other works focusing on individual lesions. This is mainly due to the high prevalence of multifocal PCa, which includes most PCa cases (38) and is characterized by high clonal heterogeneity (39). The literature on whole prostate gland radiomics is relatively recent, but results have shown that whole prostate gland radiomics has predictive power for csPCa classification (40,41). Other works using only MRI studies with no lesion location information confirm that these approaches are promising, albeit with limited results or no external or prospective validations (42,43). Our sensitivity analysis determined that radiomics alone is insufficient: previous works are aligned with our results (AUC between approximately 0.7 and 0.8 for radiomics-only models) (40,41,44). Because of the substantial size of our dataset, we

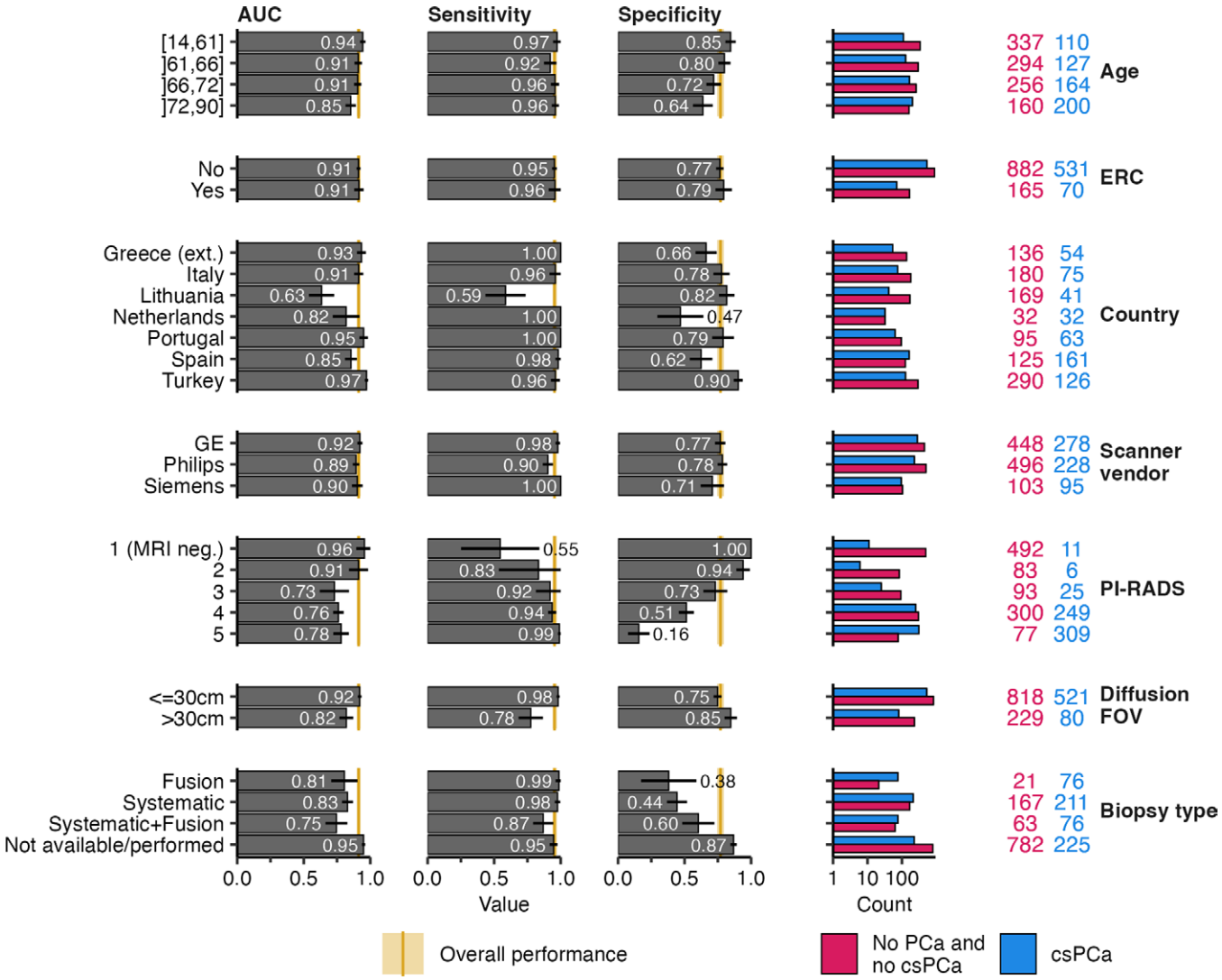

Figure 4: Performance metrics for our model are applied to the prospective cohort, stratified according to relevant subgroups (age, endorectal coil [ERC], country, scanner vendor, Prostate Imaging Reporting and Data System [PI-RADS]). The gold vertical line in the first three columns corresponds to the performance observed for the whole prospective dataset. Horizontal black lines in the first three columns correspond to the parametric 95% CIs (DeLong test for area under the receiver operating characteristic curve [AUC] and *z* scores for sensitivity and specificity). Categories with fewer than 50 instances were excluded. csPCa = clinically significant prostate cancer, FOV = field of view, PCa = prostate cancer.

regard this as additional confirmatory evidence for the general performance of these models and a solid demonstration that clinical data can further improve clinical machine learning models. As shown by our sensitivity analysis, models trained on individual vendors fail to generalize when tested on data from other vendors, confirming previous findings in the literature pertaining to classification (44,45) and segmentation (22,46). Our fairness analysis, on the other hand, highlights the advantages of training models on multicentric data by showing generally consistent generalization across age, country, ERC use, and scanner vendor.

Our study had several limitations. First, due to the statistically significant dataset size, the segmentation masks were automatically generated for all volumes. Given that only a small sample (approximately 5%) was validated by a radiologist, this is a potential limitation. While keeping a small percentage of subpar segmentations (rare in our data as noted by the segmentation quality analysis) still led to remarkable results that outperform the current standard of care, we note that a practical clinical deployment of this tool should involve a segmentation quality confirmation analysis, as in previous works (47). Second, reduced performance in cases where the diffusion FOV is larger than that of T2-weighted imaging raises an important exception for the application of our model. Third, the inclusion of dynamic contrast enhancement, typical of mpMRI, could lead to better performing models. However, a large multicentric study has recently shown that bpMRI is noninferior to mpMRI (48), so the added value to computational models may be limited. Fourth, the lack of information on radiologist expertise makes these results harder to interpret, although PI-RADS scores and ground truths were obtained from patients treated at each center and represent the variability and quality observed in day-to-day practice. Fifth, a caveat common to these analyses is that not all positive cases are detected, as they depend on positive PSA and MRI examinations or other clinical determinants. Having a 1-year follow-up tentatively addresses this, but neither PSA nor MRI detect all csPCa (6,49), and other studies have used cohorts with longer follow-ups to define negative cases (47). Indeed, it is impractical to determine for such

a large cohort whether all centers followed similar criteria for biopsy recommendation in the retrospective data, the PI-RADS version they used, and whether their PI-RADS estimates made use of mpMRI or bpMRI. Here, we used PI-RADS >3 as a conservative cutoff point, but it is possible that some centers will recommend biopsies for lower PI-RADS cutoffs. However, modern PI-RADS leads to similar performance between bpMRI and mpMRI, as recently evidenced in an aforementioned large, multireader, multicentric study (48). Our sensitivity analyses showed no differences in conclusions when considering centers using PI-RADS version 2 or version 2.1, or bpMRI or mpMRI PI-RADS. Finally, a proper assessment of this system with simulated paired readings could lead to a better understanding of its real-world performance, with an adequate evaluation of potential automation bias (50).

In conclusion, our multicentric study used the ProstateNet repository, a large bpMRI archive, to train and prospectively validate a multimodal machine learning model to distinguish clinically significant cancer from nonsignificant or no cancer. The developed model could potentially help clinicians reduce the number of unnecessary biopsies, as the model showed extremely high generalization power and proved robust for most of the analyzed subcohorts. Future work should focus on incorporating additional sources of clinical information, such as race and ethnicity, familial history of PCa or other hereditary cancer syndromes, and PCa genetic risk (51).

Author affiliations:

[1] Champalimaud Research, Champalimaud Foundation, Computational Clinical Imaging, Av. Brasília, Doca de Pedrouços, Lisboa, Lisbon, PT 1400-038, Portugal

[2] Faculty of Medicine, University of Porto, Porto, Portugal

[3] LASIGE, Faculty of Sciences, University of Lisbon, Lisbon, Portugal

[4] Radiology Department, Champalimaud Clinical Center, Champalimaud Foundation, Lisbon, Portugal

[5] Instituto Politécnico de Coimbra, Instituto Superior de Engenharia, Coimbra, Portugal

[6] Centro de Investigação do Instituto Português de Oncologia do Porto (CI-IPOP): Grupo de Física Médica, Radiobiologia e Protecção Radiológica, Porto, Portugal

[7] FORTH, Institute of Computer Science, Computational BioMedicine Lab, Greece

[8] Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Turin, Italy

Received December 18, 2024; revision requested January 22, 2025; revision received April 1; accepted June 24.

Address correspondence to: N.P. (email: nickolas.papanikolaou@research.fchampalimaud.org).

Funding: Supported by European Union Horizon 2020 research and innovation program (grant no. 952159). Partially supported by Fundação para a Ciência e a Tecnologia, Portugal, through the LASIGE Research Unit (UIDB/00408/2020) and UIDP/00408/2020. N.R. supported by PhD grant no. 10.54499/2021.05322.BD.

Author contributions: Guarantors of integrity of entire study, **J.G.d.A., S.S., N.P.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **A.C.R., R.M., A.M.G., M.T., D.R.**; clinical studies, **A.C.R., J.G.d.A., R.M., A.S.C.V., A.M.G., C.B., I.S., J.I., S.B., N.P.**; experimental studies, **A.C.R., J.G.d.A., N.R., R.M., A.M.G., C.B.**; statistical analysis, **A.C.R., J.G.d.A., A.M.G., S.S., N.P.**; and manuscript editing, **A.C.R., J.G.d.A., N.R., R.M., A.M.G., S.S., M.T., K.M., D.R., N.P.**

Disclosures of conflicts of interest: **A.C.R.** No relevant relationships. **J.G.d.A.** No relevant relationships. **N.R.** Grant from SciPROJ. **M.R.** No relevant relationships. **A.S.C.V.** No relevant relationships. **A.M.G.** No relevant relationships. **C.B.** No relevant relationships. **I.S.** No relevant relationships. **J.I.** No relevant relationships. **S.B.** No relevant relationships. **S.S.** No relevant relationships. **I.D.** No relevant relationships. **M.T.** No relevant relationships. **K.M.** No relevant relationships. **D.R.** Grants or contracts from AIRC 5x1000 Colon Cancer. **N.P.** Stock in MRIcons.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71(3):209–249.
2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023;73(1):17–48.
3. Mottet N, van den Bergh RCN, Briers E, et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer-2020 update. Part 1: screening, diagnosis, and local treatment with curative intent. Eur Urol 2021;79(2):243–262.
4. Albertsen PC. Prostate cancer screening and treatment: where have we come from and where are we going? BJU Int 2020;126(2):218–224.
5. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. Eur Urol 2019;76(3):340–351.
6. Cuocolo R, Verde F, Ponsiglione A, et al. Clinically significant prostate cancer detection with biparametric MRI: a systematic review and meta-analysis. AJR Am J Roentgenol 2021;216(3):608–621.
7. Scapicchio C, Gabelloni M, Barucci A, Cioni D, Saba L, Neri E. A deep look into radiomics. Radiol Med 2021;126(10):1296–1311.
8. Koh DM, Papanikolaou N, Bick U, et al. Artificial intelligence and machine learning in cancer imaging. Commun Med (Lond) 2022;2(1):133.
9. Papanikolaou N, Matos C, Koh DM. How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging 2020;20(1):33.
10. Horvat N, Papanikolaou N, Koh DM. Radiomics beyond the hype: a critical evaluation toward oncologic clinical use. Radiol Artif Intell 2024;6(4):e230437.
11. Wang J, Wu CJ, Bao ML, Zhang J, Wang XN, Zhang YD. Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer. Eur Radiol 2017;27(10):4082–4090.
12. Huynh LM, Hwang Y, Taylor O, Baine MJ. The use of MRI-derived radiomic models in prostate cancer risk stratification: a critical review of contemporary literature. Diagnostics (Basel) 2023;13(6):1128.
13. Stark JR, Perner S, Stampfer MJ, et al. Gleason score and lethal prostate cancer: does 3 + 4 = 4 + 3? J Clin Oncol 2009;27(21):3459–3464.
14. Milonas D, Venclovas Ž, Gudinaviciene I, et al. Impact of the 2014 International Society of Urological Pathology grading system on concept of high-risk prostate cancer: comparison of long-term oncological outcomes in patients undergoing radical prostatectomy. Front Oncol 2019;9:1272.
15. Offermann A, Hohensteiner S, Kuempers C, et al. Prognostic value of the new prostate cancer International Society of Urological Pathology grade groups. Front Med (Lausanne) 2017;4:157.
16. Spratt DE, Jackson WC, Abugharib A, et al. Independent validation of the prognostic capacity of the ISUP prostate cancer grade grouping system for radiation treated patients with long-term follow-up. Prostate Cancer Prostatic Dis 2016;19(3):292–297.
17. Williams IS, McVey A, Perera S, et al. Modern paradigms for prostate cancer detection and management. Med J Aust 2022;217(8):424–433.
18. Schoots IG, Osses DF, Drost FJH, et al. Reduction of MRI-targeted biopsies in men with low-risk prostate cancer on active surveillance by stratifying to PI-RADS and PSA-density, with different thresholds for significant disease. Transl Androl Urol 2018;7(1):132–144.
19. Saha A, Bosma JS, Twilt JJ, et al; PI-CAI consortium. Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. Lancet Oncol 2024;25(7):879–887.
20. Cai JC, Nakai H, Kuanar S, et al. Fully automated deep learning model to detect clinically significant prostate cancer at MRI. Radiology 2024;312(2):e232635.
21. ProstateNET. https://prostatenet.eu. Accessed August 5, 2024.
22. Rodrigues NM, Almeida JGd, Verde ASC, et al; ProCAncer-I Consortium. Analysis of domain shift in whole prostate gland, zonal and lesions segmentation and detection, using multicentric retrospective data. Comput Biol Med 2024;171:108216. [Published correction appears in Comput Biol Med 2024;173:108352.]
23. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310–1320.
24. Yaniv Z, Lowekamp BC, Johnson HJ, Beare R. SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. J Digit Imaging 2018;31(3):290–303. [Published correction appears in J Digit Imaging 2019;32(6):1118.]
25. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med 2011;52(3):369–378.

26. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21): e104–e107.
27. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. Red Hook, NY, USA: Curran Associates, 2017;3149–3157.
28. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, eds. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2018; 6639–6649.
29. Pepe M, Longton G, Janes H. Estimation and comparison of receiver operating characteristic curves. Stata J 2009;9(1):1.
30. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, eds. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2017; 4768–4777.
31. Bavo DC, Daan N, Ben VC, Ewout S, Yvonne V. CalibrationCurves: calibration performance. CRAN: Contributed Packages. The R Foundation; 2022. 10.32614/cran.package.calibrationcurves.
32. Sjoberg DD. Dcurves: decision curve analysis for model evaluation. CRAN: Contributed Packages. The R Foundation; 2021. 10.32614/cran.package.dcurves.
33. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019;3(1):18.
34. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al. METhodological radiomICs score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. Insights Imaging 2024;15(1):8.
35. Schrader A, Netzer N, Hielscher T, et al. Prostate cancer risk assessment and avoidance of prostate biopsies using fully automatic deep learning in prostate MRI: comparison to PI-RADS and integration with clinical data in nomograms. Eur Radiol 2024;34(12):7909–7920.
36. Omri N, Kamil M, Alexander K, et al. Association between PSA density and pathologically significant prostate cancer: the impact of prostate volume. Prostate 2020;80(16):1444–1449.
37. Sebastianelli A, Morselli S, Vitelli FD, et al. The role of prostate-specific antigen density in men with low-risk prostate cancer suitable for active surveillance: results of a prospective observational study. Prostate Int 2019;7(4):139–142.
38. Andreoiu M, Cheng L. Multifocal prostate cancer: biologic, prognostic, and therapeutic implications. Hum Pathol 2010;41(6):781–793.
39. Løvf M, Zhao S, Axcrona U, et al. Multifocal primary prostate cancer exhibits high degree of genomic heterogeneity. Eur Urol 2019;75(3):498–505.
40. Rodrigues A, Santinha J, Galvão B, Matos C, Couto FM, Papanikolaou N. Prediction of prostate cancer disease aggressiveness using bi-parametric MRI radiomics. Cancers (Basel) 2021;13(23):6065.
41. Filos D, Fotopoulos D, Rouni MA, Chouvarda I. Machine learning-based whole gland radiomics analysis for prostate cancer classification. 2024 IEEE International Symposium on Biomedical Imaging (ISBI). IEEE, Athens: 2024;1–5. 10.
42. Weißer C, Netzer N, Görtz M, et al. Weakly supervised MRI slice-level deep learning classification of prostate cancer approximates full voxel- and slice-level annotation: effect of increasing training set size. J Magn Reson Imaging 2024;59(4):1409–1422.
43. Redekop E, Sarma KV, Kinnaird A, et al. Attention-guided prostate lesion localization and grade group classification with multiple instance learning. In: Konukoglu E, Menze B, Venkataraman A, Baumgartner C, Dou Q, Albarqouni S, eds. Proceedings of the 5th International Conference on Medical Imaging with Deep Learning. Baltimore, MA: PMLR: 2022; 975–987.
44. Gresser E, Schachtner B, Stüber AT, et al. Performance variability of radiomics machine learning models for the detection of clinically significant prostate cancer in heterogeneous MRI datasets. Quant Imaging Med Surg 2022;12(11):4990–5003.
45. Kushol R, Parnianpour P, Wilman AH, Kalra S, Yang YH. Effects of MRI scanner manufacturers in classification tasks with deep learning models. Sci Rep 2023;13(1):16791.
46. Zavala-Romero O, Breto AL, Xu IR, et al. Segmentation of prostate and prostate zones using deep learning: A multi-MRI vendor analysis: a multi-MRI vendor analysis. Strahlenther Onkol 2020;196(10):932–942.
47. Spaanderman DJ, Hakkesteegt SN, Hanff DF, et al. Multi-center external validation of an automated method segmenting and differentiating atypical lipomatous tumors from lipomas using radiomics and deep-learning on MRI. EClinicalMedicine 2024;76:102802.
48. Twilt JJ, Saha A, Bosma JS, et al; PI-CAI Consortium. Evaluating biparametric versus multiparametric magnetic resonance imaging for diagnosing clinically significant prostate cancer: an international, paired, noninferiority, confirmatory observer study. Eur Urol 2025;87(2):240–250.
49. Merriel SWD, Pocock L, Gilbert E, et al. Systematic review and meta-analysis of the diagnostic accuracy of prostate-specific antigen (PSA) for the detection of prostate cancer in symptomatic patients. BMC Med 2022;20(1):54.
50. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121–127.
51. Ni Raghallaigh H, Eeles R. Genetic predisposition to prostate cancer: an update. Fam Cancer 2022;21(1):101–114.