Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.264095

ABDOMINAL IMAGING

ORIGINAL ARTICLE

Development and validation of a multimodal artificial intelligence-based model for predicting post-prostatectomy treatment outcomes from baseline biparametric prostate magnetic resonance imaging

Benjamin D. Simon[1,2]
Esra Akcicek[3]
Stephanie A. Harmon[1]
Lei Clifton[4,5]
Anshul Thakur[2]
Sandeep Gurram[6]
David Clifton[2]
Bradford J. Wood[7,8]
Ali Devrim Karaosmanoglu[3]
Peter L. Choyke[1]
Deniz Akata[3]
Peter A. Pinto[6]
Baris Turkbey[1]

[1]Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, Maryland, United States of America

[2]University of Oxford, Institute of Biomedical Engineering, Oxford, United Kingdom

[3]Hacettepe University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

[4]Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

[5]Big Data Institute, University of Oxford, Oxford, United Kingdom

[6]Urology Oncology Branch, National Cancer Institute, National Institutes of Health, Maryland, United States of America

[7]Center for Interventional Oncology, National Cancer Institute, National Institutes of Health, Maryland, United States of America

[8]Department of Radiology, Clinical Center, National Institutes of Health, Bethesda, Maryland, United States of America

Handling editor: Nurullah Dağ

Corresponding author: Baris Turkbey

E-mail: turkbeyi@mail.nih.gov

Received 28 April 2026; revision requested 13 May 2026; accepted 06 June 2026.

Epub: 25.06.2026

DOI: 10.4274/dir.2026.264095

PURPOSE

Prostate cancer (PCa) is the second most common cancer and cause of cancer deaths among American men. Existing risk prediction methods have limited accuracy and reproducibility, resulting in difficulty in predicting treatment outcomes. We demonstrate the development and external validation of an automated multimodal artificial intelligence (AI) algorithm using biparametric magnetic resonance imaging (bpMRI) and clinical covariates for predicting biochemical recurrence (BCR) after radical prostatectomy (RP) in patients with PCa.

METHODS

The development cohort included 80% of patients from center 1 (n = 240) who underwent prostate MRI prior to RP between January 2008 and December 2018, with a minimum of 2 years of follow-up after RP. The test cohort included the remaining 20% of center 1 patients (n = 71) and an external validation cohort from center 2 (n = 168). Center 2 patients included those who underwent prostate MRI and RP between January 2015 and January 2024, with a minimum of 2 years of follow-up. Clinical comparisons were made using the Cancer of the Prostate Risk Assessment Postsurgical (center 1) and International Society of Urological Pathology Gleason Grade Group (ISUP GGG) scoring systems from post-RP pathology (center 2). The models developed were as follows: clinical (M0), automated clinical (M1), radiomics (M2), and a multimodal model (M3). Clinical variables (M0) included prostate-specific antigen (PSA), age, primary Gleason, and ISUP GGG. Automated clinical variables (M1 and M3) included PSA and age. Radiomic features (M2 and M3) were extracted from bpMRI using a lesion detection AI model. Accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) were calculated, and log-rank tests compared BCR-free survival to assess the models' ability to discriminate relative to clinical standards. Intermediate-risk groups were also assessed.

RESULTS

The multimodal model (M3) had the highest AUC across test sets (combined: 0.71; center 1: 0.70; center 2: 0.75). This was the only model that significantly differentiated BCR-free survival outcomes in intermediate-risk groups across both centers ($P < 0.05$).

CONCLUSION

This automated multimodal model leveraging radiomics and clinical covariates can predict BCR after RP, approaching clinical gold standards, and may enhance imaging-based prognostication following further validation.

CLINICAL SIGNIFICANCE

Given that this model demonstrated the potential to outperform pre-surgical and post-surgical clinical gold standards in an external cohort's intermediate-risk patient subgroup (for whom it is more challenging to predict disease trajectory), this model may contribute to enhanced personalized care in PCa after further validation.

KEYWORDS

Biochemical recurrence, MRI, multi-modal AI, prostate cancer, prostatectomy

You may cite this article as: Simon BD, Akcicek E, Harmon SA, et al. Development and validation of a multimodal artificial intelligence-based model for predicting post-prostatectomy treatment outcomes from baseline biparametric prostate magnetic resonance imaging. *Diagn Interv Radiol*. 25 June 2026 DOI: 10.4274/dir.2026.264095 [Epub Ahead of Print].

Prostate cancer (PCa) is the second leading cause of cancer deaths among American men. In 2025, there are projected to be 313,780 new diagnoses and 35,770 deaths from cancer in America and 712,000 global deaths per year in 2040.[1,2] Although PCa has a reputation for being treatable due to the high rates of low-grade disease and availability of blood-test screening through prostate-specific antigen (PSA) testing, it is a heterogeneous disease with multiple treatment options and risk of metastases and mortality.[3] High-grade and distant-stage disease are on the rise despite advances in treatment technologies. There are several common tools to stratify patients' risk of metastasis to avoid over- and under-treatment. Widely used scoring systems, such as the National Comprehensive Cancer Network (NCCN) risk groups, Gleason grading, Cancer of the Prostate Risk Assessment Postsurgical (CAPRA-S), and Prostate Imaging Reporting and Data System (PI-RADS), suffer from biopsy sampling error and inter-reader disagreement. Although these metrics can differentiate outcomes on a large-scale level, for an individual patient, it is unclear how sensitive or specific these scores may be, as they mostly rely on subjective "expert"-dependent categorical evaluations. Despite their shortcomings, these systems still inform treatment decisions, presenting an opportunity to improve outcomes by exploring a precision medicine approach. There is a need to introduce technology that is consistent across centers, and accurately prognostic on an individual patient level, to determine optimal treatments. With magnetic resonance imaging (MRI) being relatively new to PCa guidelines, this signal-rich imaging technology provides an exciting opportunity for modern artificial intelligence (AI) and computer vision to play a role.

Main points

- This study developed and validated an automated multimodal artificial intelligence (AI) model using biparametric magnetic resonance imaging (MRI) and clinical features such as age and prostate-specific antigen to predict biochemical recurrence after radical prostatectomy (RP).
- The fully automated multimodal model outperformed unimodal models using imaging and clinical data alone, achieving the best predictive performance in the internal test-set and external validation cohort, with area under the receiver operating characteristic curve values of 0.70 and 0.75, respectively.
- This automated model was also the only approach that was consistently able to separate intermediate-risk patients into prognostic groups with significantly different recurrence-free survival outcomes in both centers.
- These results suggest that automated MRI-based AI may improve risk assessment prior to RP, providing a potential avenue for personalized treatment planning in prostate cancer, although larger multicenter validation is still needed.

For higher-risk PCa, treatment often comes down to a decision between surgery or radiation therapy with or without androgen deprivation therapy. With roughly equivalent long-term outcomes, this decision often depends on how patients and providers choose to weigh side effects.[4] Sometimes, factors such as extraprostatic extension (EPE) or age will shift the decision away from surgery due to a higher likelihood of cancerous margins or risk of complications.[5] These factors, too, are subjective and may vary across physicians and hospitals.[6]

One solution to the inter-reader disagreement across gold standard risk stratification techniques is AI, which has demonstrated the potential to assist various medical decision-making processes across various modalities objectively.[7] Although there is some exploration on the use of AI in PCa, this mostly focuses on detection, and it is difficult to find multimodal high-quality work with sufficient external validation.[3,8] There is a need for automated multimodal pipelines to predict outcomes for patients undergoing radical prostatectomy (RP) with external validation for PCa. Here, we aim to develop and evaluate a multimodal AI-based algorithm using biparametric MRI (bpMRI) and clinical covariates for predicting biochemical recurrence (BCR) after RP in patients with PCa.

Methods

Development cohort—center 1—patient population

This retrospective, Institutional Review Board (IRB)-approved, and Health Insurance Portability and Accountability Act-compliant study (ClinicalTrials.gov identifiers NCT02594202 and NCT03354416) included patients at the National Institutes of Health (National Cancer Institute) who underwent prostate MRI prior to RP between January 2008 and December 2018 (n = 683). The criteria for exclusion are described in detail in Figure 1a and resulted in a final dataset of 311 patients. Biopsy methods and reader details are available in the Supplementary Methods. Clinical covariates (age, PSA, Gleason scores, and ISUP GGG) were recorded at the time of MRI or biopsy, and follow-up for BCR after RP was assessed using follow-up PSA values (using consistent PSA ^{3}0.2 ng/mL as the definition of BCR). The RP pathology evaluation was used to calculate a post-surgical CAPRA-S score for predicting BCR after surgery.[9] Data were split through pseudo-randomization into n = 240 training and n = 71 test images, consistent with the split used in the authors' previously developed open-source algorithm for lesion detection to avoid data leakage and maintain test-set independence.[10] Center 1 image acquisition details are shown in the Supplementary Table 1.

External validation cohort—center 2—patient population

This retrospective IRB-approved study included patients who underwent prostate MRI prior to RP and received a diagnosis of PCa at center 2 between January 2015 and January 2024 (n = 682) at Hacettepe University. The criteria for exclusion are described in detail in Figure 1b and resulted in a final dataset of 168 patients. Biopsy methods and reader details for this cohort are available in the Supplementary Methods. Clinical covariates (age, PSA, Gleason scores, and ISUP GGG) were recorded at the time of MRI or biopsy, and follow-up for BCR after RP was based on follow-up PSA values (using two consecutive PSA values of ^{3}0.2 ng/mL as the definition of BCR). Center 2 image acquisition details are shown in the Supplementary Table 1.

Lesion segmentation artificial intelligence model and radiomics extraction

The lesion segmentation AI model was built on a previously developed and publicly available lesion detection and segmentation model (GitHub).[10] This AI model leveraged T2-weighted (T2W) imaging, apparent diffusion coefficient maps, and high-b-value diffusion-weighted imaging to produce a lesion segmentation. Further work has leveraged this model for various purposes, such as EPE detection,[11] external validation,[12] and more.[13] AI-derived lesion masks were then used to measure 112 T2W MRI radiomic features using PyRadiomics version 3.10.0 (AIM, Harvard, Boston, MA, USA) in python. The T2W MR images were normalized and resampled to isotropic 1-mm^3 voxel spacing with an intensity discretization of 20. Features included both texture and volumetric features.

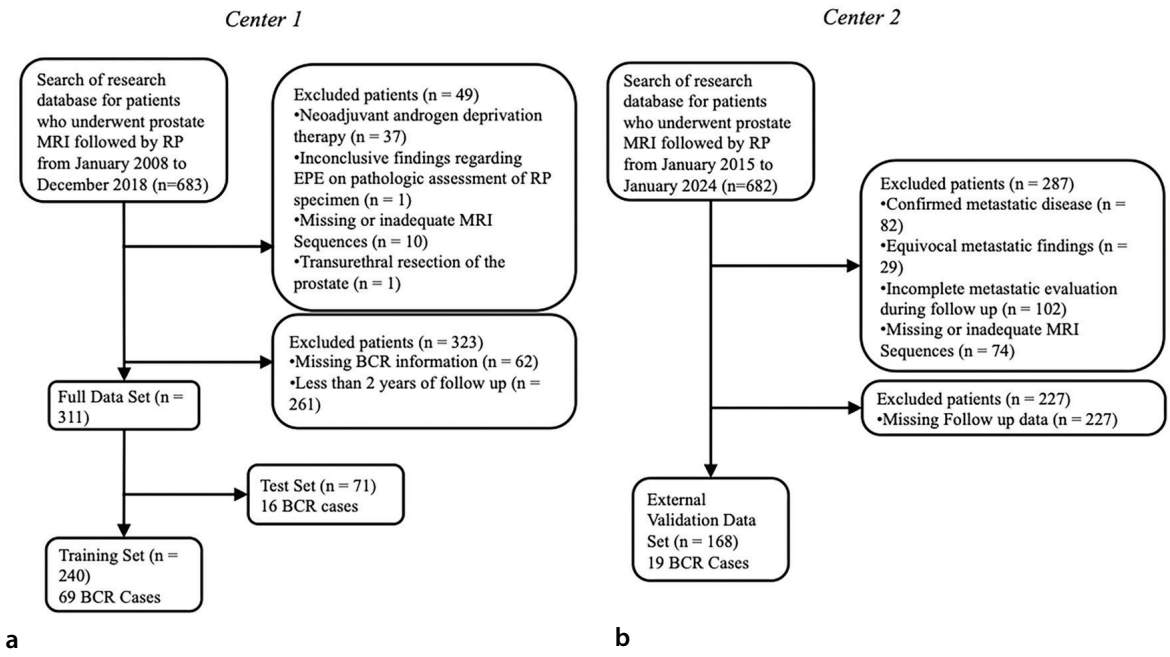

Figure 1. Patient population flowchart. **(a)** Describes retrospective query for individuals with prostate cancer (PCa) at center 1 who underwent radical prostatectomy (RP) during the specified interval and exclusions made based on other treatments, image quality, and follow-up. **(b)** Describes a retrospective query for individuals with PCa who underwent RP during a specified interval and had appropriate evaluation for metastatic disease, imaging, and follow-up data. MRI, magnetic resonance imaging; BCR, biochemical recurrence; EPE, extraprostatic extension.

Outcome prediction models

AI-derived lesion segmentation-based texture and volumetric features, along with clinical covariates (age and PSA), were used as inputs to XGBoost models (Figure 2). Hyperparameter tuning was completed using 5-fold cross-validation using the area under the receiver operating characteristic curve (AUC). Training did not consider the time to BCR. Four models (M0–M3) were developed:

- M0 is a comprehensive clinical model without MRI information; M0 inputs include only age, PSA, primary Gleason, and ISUP GGG at biopsy. This model serves as a control to evaluate M2 and M3.
- M1 is a purely automated clinical model without MRI information; M1 inputs include only age and PSA. This model also serves as a control to evaluate M2 and M3.
- M2 is a purely imaging-based model with only the 112 radiomic features associated with the AI-derived lesion mask on T2W MRI.

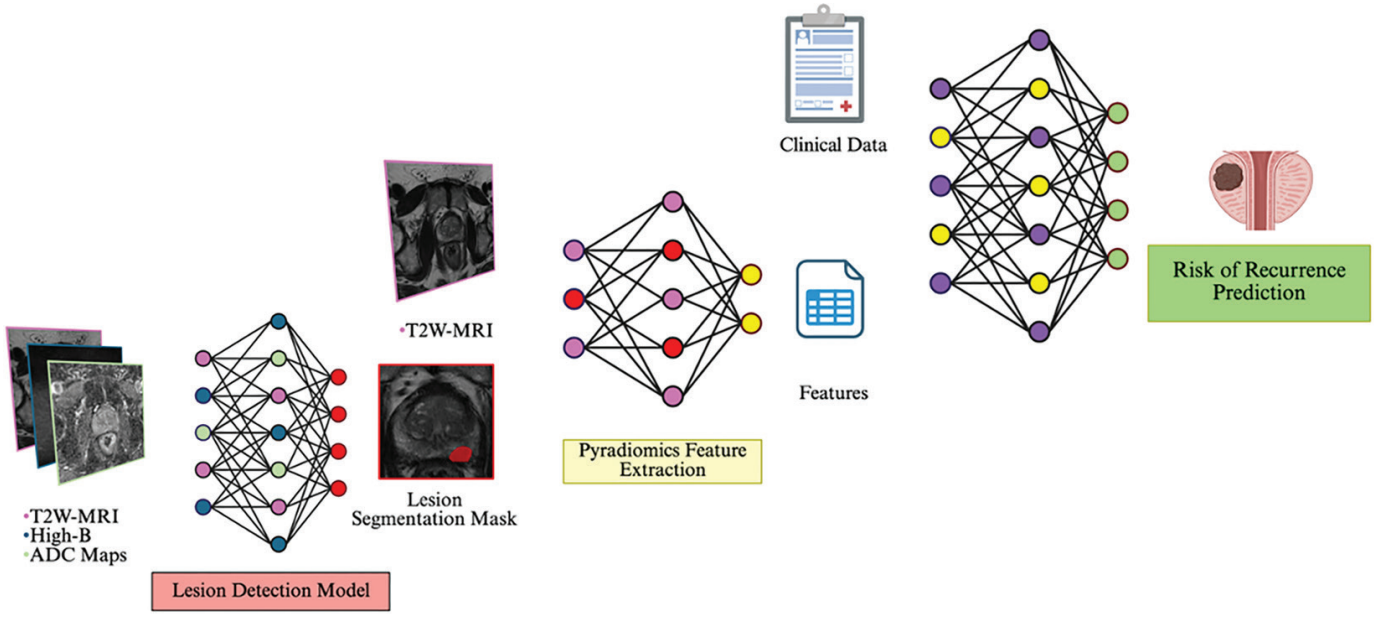

Figure 2. Model Pipeline Schematic. Pipeline schematic illustrating three imaging sequences used in the lesion segmentation model, followed by feature extraction of T2-weighted magnetic resonance imaging using PyRadiomics and finally combining with clinical data to predict the biochemical recurrence. T2W, T2-weighted; MRI, magnetic resonance imaging; ADC, apparent diffusion coefficient.

- M3 is the automated multimodal model combining these imaging features with age and PSA.

These models and associated code are freely available at https://github.com/NIH-MIP/MultimodalRadiomicsSurgeryModel.

Statistical analysis

For each model and cohort, the AUC was calculated with pairwise DeLong's tests for comparison. Youden's J Statistic in the training set was used as a cut-off for the model's final prediction, and accuracy, sensitivity, specificity, positive predictive value, and negative predictive value are reported with bootstrapped (n = 1,000) 95% confidence intervals. Kaplan–Meier curves were generated for each cohort and combined to evaluate prediction in the context of time to BCR. Log-rank tests were performed to evaluate these curves. The CAPRA-S score was evaluated and compared for the development cohort test as a clinical gold standard baseline, and only the RP specimen ISUP grade was used for the external validation cohort due to a lack of margin status and regional biopsy information. Furthermore, a comparison with the pre-operative biopsy ISUP GGG is included. The equivalent statistical analyses were conducted for intermediate- and low-risk patients, for whom it is most difficult to identify which patients will have BCR.

Results

Patient population—center 1

Of the 311 patients from center 1, 85 had BCR. This constituted 29% of the training set (69/240) and 23% of the test set (16/71).

The postsurgical CAPRA-S score was calculated for patients, with most receiving an intermediate score of 3–5 [66% (157/240) in the training set and 72% (51/71) in the test set]. Of those with BCR, a higher percentage had high-risk CAPRA-S scores of ≥ 6 [54% (37/69) in the training set and 38% (6/16) in the test set].

Patient population—center 2

Of the 168 patients in center 2, there were 19 patients with BCR documented. This constituted 11.3% of the external validation set (19/168).

The postsurgical CAPRA-S score could not be calculated for patients at this center; therefore, the ISUP GGG at post-surgical biopsy specimens was used as a clinical baseline. The majority of patients had an intermediate ISUP GGG score of 2–3 (58% or 97/168 in the external validation set) compared with 50/168 ISUP 1 and 21/168 ISUP 4–5. Those with high-risk ISUP had a larger proportion of BCR events (6/21) than those with ISUP 2–3 (13/97) and low-risk ISUP 1 (2/50). Complete demographic information for each center is presented in Table 1.

Hyperparameter tuning

Hyperparameter tuning was completed for all models based on cross-validation AUC (Table 2).

Model results

For each model, the performance was evaluated using the AUC overall and for each cohort, separately treating BCR as a binary outcome. The AUC in the combined set of test patients in centers 1 and 2 was 0.54 for M0, 0.61 for M1, 0.65 for M2, and 0.71 for M3. Using the optimal threshold as determined by Youden's J Statistic in the training set, M3 achieved the highest sensitivity of 94%, and M0 had the highest specificity of 70% and accuracy of 66%. Complete binary results are shown in Table 3, with AUCs presented in

Table 1. Patient characteristics table

		Min	25th percentile	Median	75th percentile	Max
Center 1	PSA	0.21	4.58	6.8	11.5	113.6
	Age (years)	42	56	61	65	76
	ISUP GGG	0	1	2	4	5
	Gleason primary	3	3	3	4	9
	BCR free survival time	2.08	2.7	3.4	4.7	9.0
Center 2	PSA	0.01	5.19	7.08	10.96	334.00
	Age (years)	45	59	64	69	77
	ISUP GGG	1	1	2	2.25	5
	Gleason primary	3	3	3	3	5
	BCR free survival time (years)	2.1	2.4	3.8	5.0	6.1
	BCR	No BCR				
Center 1	85	226				
Center 2	19	149				
Intermediate risk patients	BCR	No BCR				
Center 1 (CAPRA-S 3–5)	41	168				
Center 2 (ISUP 2–3)	11	86				

Basic clinical distribution statistics for each center and BCR event counts. PSA, prostate-specific antigen; ISUP GGG, International Society of Urological Pathology Gleason Grade Group; CAPRA-S, Cancer of the Prostate Risk Assessment Postsurgical; BCR, biochemical recurrence; Min, minimum; Max, maximum.

Table 2. Hyperparameter tuning selections

	Hyperparameter distribution	M0	M1	M2	M3
Estimators	100, 150, 200, 250, 300	250	200	100	200
Learning rate	0.01, 0.05, 0.1, 0.2, 0.3	0.05	0.01	0.3	0.01
Maximum depth	3, 4, 5, 6, 7, 8, 9, 10	5	5	5	5
Subsample	0.6, 0.8, 1.0	0.80	0.60	0.60	0.60
Minimum child weight	1, 3, 5	3	5	1	5
Gamma	0, 0.1, 0.2, 0.3	0	0.3	0.2	0.3
Lambda	0, 0.1, 0.5, 1, 2, 5	1	5	1	5
Alpha	0, 0.1, 0.5, 1, 2, 5	0.1	5	5	5
Subsample of columns ratio (per tree)	0.6, 0.8, 1.0	0.80	0.60	0.60	0.60

Distribution of hyperparameters evaluated using 5-fold cross-validation of the training set and selected parameters for each model.

Figure 3. DeLong's tests pairwise for each model in combined and separate center cohorts can be found in Table 4.

In the center 1 cohort, the AUC was 0.53 for M0, 0.61 for M1, 0.58 for M2, and 0.70 for M3. Using the optimal threshold, M3 achieved the highest sensitivity of 88%, and M0 had the highest specificity of 85% and accuracy of 72%. In the center 2 cohort, the AUC was 0.60 for M0, 0.61 for M1, 0.76 for M2, and 0.75 for M3. Using the optimal threshold, M3 achieved the highest sensitivity of 100%, and M0 had the highest specificity of 65% and accuracy of 64%. The feature importances for each model are presented in Table 5.

Outcome analysis

Although each model was trained on binary outcomes, time to BCR is an important factor. Accordingly, Kaplan–Meier curves for BCR-free survival were created for each model across multiple subgroups, including the combined cohort, individual cohorts, and intermediate-risk patients. A log-rank test was performed between the two prediction groups (BCR vs. BCR-free) using the optimal cut-off. Models M1 and M3 stratified patients into risk groups with significantly different BCR-free survival outcomes in the combined cohorts; M0, M1, and M3 demonstrated significant stratification in center 1, whereas in center 2, M1, M2, and M3 significantly stratified patients ($P < 0.05$). These survival curves are shown in Figure 4 (a–d for the combined test set, e–h for center 1, and i–l for center 2). Center 1 and center 2 intermediate-risk results are presented in Figure 5a–d and Figure 5e–h, respectively, with clinical comparisons in Figure 5i for center 1 (CAPRA-S) and 5j for center 2 (post-surgical ISUP GGG).

Clinical comparison

Given that scoring systems such as post-surgical ISUP, CAPRA-S, and NCCN prognostic groupings perform relatively well in low- or high-risk individuals, it is important to investigate how these models perform in the intermediate-risk subgroups.[14,15] Across all center 1 patients, CAPRA-S stratified patients into risk groups with significantly different BCR-free survival outcomes ($P < 0.05$). For the intermediate-risk PCa center 1 test-set patients with CAPRA-S scores of 3–5, neither CAPRA-S nor models M0, M1, or M2 could stratify patients ($P > 0.05$). However, M3 successfully stratified patients in the same cohort ($P < 0.05$).

For all center 2 external validation patients, ISUP GGG was able to significantly stratify patients. However, looking at intermediate-risk patients only with ISUP GGG groups 2–3, neither ISUP nor models M0, M1, or M2 could stratify patients successfully ($P > 0.05$). In the same cohort, M3 successfully stratified patients ($P < 0.05$). The multimodal model M3 was the only model that consistently stratified patients into risk groups with significantly different outcomes across each cohort and subgroup. Complete clinical comparison results are shown in Supplementary Figure 1. Because post-surgical scores could not be combined across centers, combined pre-surgical ISUP GGG results are presented in Supplementary Figure 2. Pre-surgical ISUP GGG was able to stratify patients into meaningful risk groups ($P < 0.05$) in the entire test-set population but was unable to do so in the intermediate-risk subgroup ($P > 0.05$).

Table 3. Model M0–M3 results at optimal threshold

		M0	M1	M2	M3
Combined	Acc	66% (60%–72%)	62% (56%–68%)	55% (48%–60%)	50% (44%–56%)
	Sens	40% (24%–56%)	60% (43%–75%)	60% (44%–77%)	94% (86%–100%)
	Spec	70% (64%–77%)	62% (56%–68%)	54% (47%–60%)	43% (36%–50%)
	AUC	0.54 (0.43–0.65)	0.61 (0.50–0.71)	0.65 (0.54–0.74)	0.71 (0.62–0.79)
Center 1	Acc	72% (61%–82%)	63% (52%–73%)	62% (51%–73%)	62% (51%–73%)
	Sens	25% (6%–47%)	56% (32%–81%)	38% (15%–62%)	88% (69%–100%)
	Spec	85% (76%–93%)	65% (53%–78%)	69% (57%–81%)	55% (42%–67%)
	AUC	0.53 (0.38–0.70)	0.61 (0.44–0.78)	0.58 (0.42–0.75)	0.70 (0.53–0.84)
Center 2 (external)	Acc	64% (57%–71%)	61% (54%–69%)	52% (45%–58%)	45% (38%–52%)
	Sens	53% (30%–76%)	63% (39%–85%)	79% (59%–95%)	100% (100%–100%)
	Spec	65% (58%–73%)	61% (54%–68%)	48% (41%–56%)	39% (32%–46%)
	AUC	0.60 (0.45–0.73)	0.61 (0.46–0.74)	0.76 (0.63–0.87)	0.75 (0.64–0.84)

Binarized result statistics at the optimal threshold determined by the training set by center and combined. Acc, accuracy; Sens, sensitivity; Spec, specificity; AUC, area under the curve.

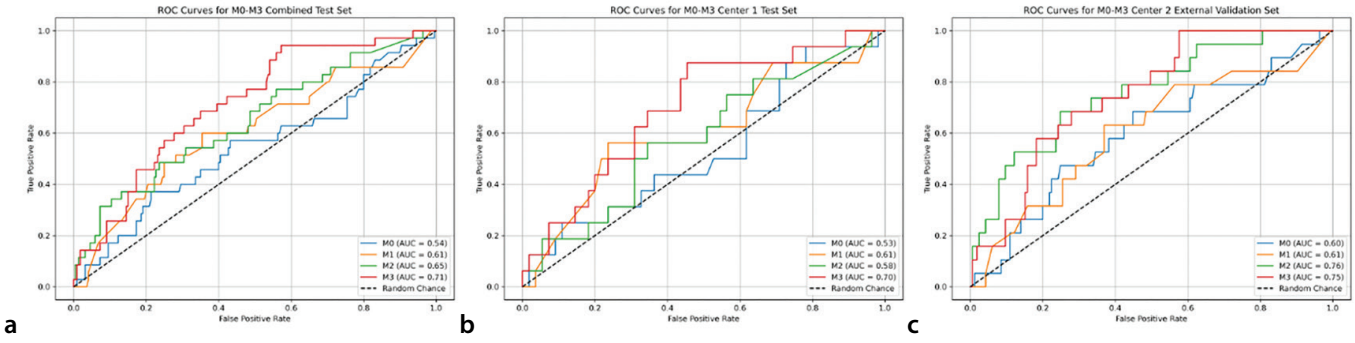

Figure 3. Area under the receiver operating characteristic (ROC) curve (AUC) results by model and cohort. **(a)** The AUC for each model M0–M3 in the combined test set. **(b)** The AUC for each model in the center 1 test set. **(c)** The AUC for each model in the center 2 external validation set.

Table 4. Pairwise AUC DeLong's tests

	Model a	Model_b	AUC a	AUC b	*P* value
Combined	M0	M1	0.54	0.61	0.148
Combined	M0	M2	0.54	0.65	0.141
Combined	**M0**	**M3**	0.54	0.71	**0.000218**
Combined	M1	M2	0.61	0.65	0.628
Combined	**M1**	**M3**	0.61	0.71	**0.0172**
Combined	M2	M3	0.65	0.71	0.215
Center 1	M0	M1	0.53	0.61	0.323
Center 1	M0	M2	0.53	0.58	0.679
Center 1	**M0**	**M3**	0.53	0.70	**0.0337**
Center 1	M1	M2	0.61	0.58	0.824
Center 1	M1	M3	0.61	0.70	0.251
Center 1	M2	M3	0.58	0.70	0.154
Center 2	M0	M1	0.60	0.61	0.836
Center 2	M0	M2	0.60	0.76	0.0850
Center 2	**M0**	**M3**	0.60	0.75	**0.00471**
Center 2	M1	M2	0.61	0.76	0.141
Center 2	**M1**	**M3**	0.61	0.75	**0.00765**
Center 2	M2	M3	0.76	0.75	0.918

DeLong's test *P* values for each center and model pairwise comparison with significant *P* values ($P < 0.05$) bolded.

Table 5. Feature importance

M0		M1		M2		M3	
Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
PSA	30.5%	PSA	89.6	GLSZM—zone variance	8.4%	PSA	3.7%
Primary Gleason	28.9%	Age	10.4	Shape sphericity	6%	Shape—least axis length	2.9%
ISUP GGG	22.2%			First order range	5%	Shape—maximum 2D diameter column	2.6%
Age	18.4%			GLSZM—small area emphasis	3.3%	Shape—mesh volume	2.4%
				GLRLM—run length non-uniformity	3.1%	GLSZM—gray level variance	2.4%

Top five feature importance weights for models M0–M3 and corresponding weights. PSA, prostate-specific antigen; ISUP GGG, International Society of Urological Pathology Gleason Grade Group; GLSZM, Gray Level Size Zone Matrix; GLRLM, Gray Level Run Length Matrix; 2D, two-dimensional.

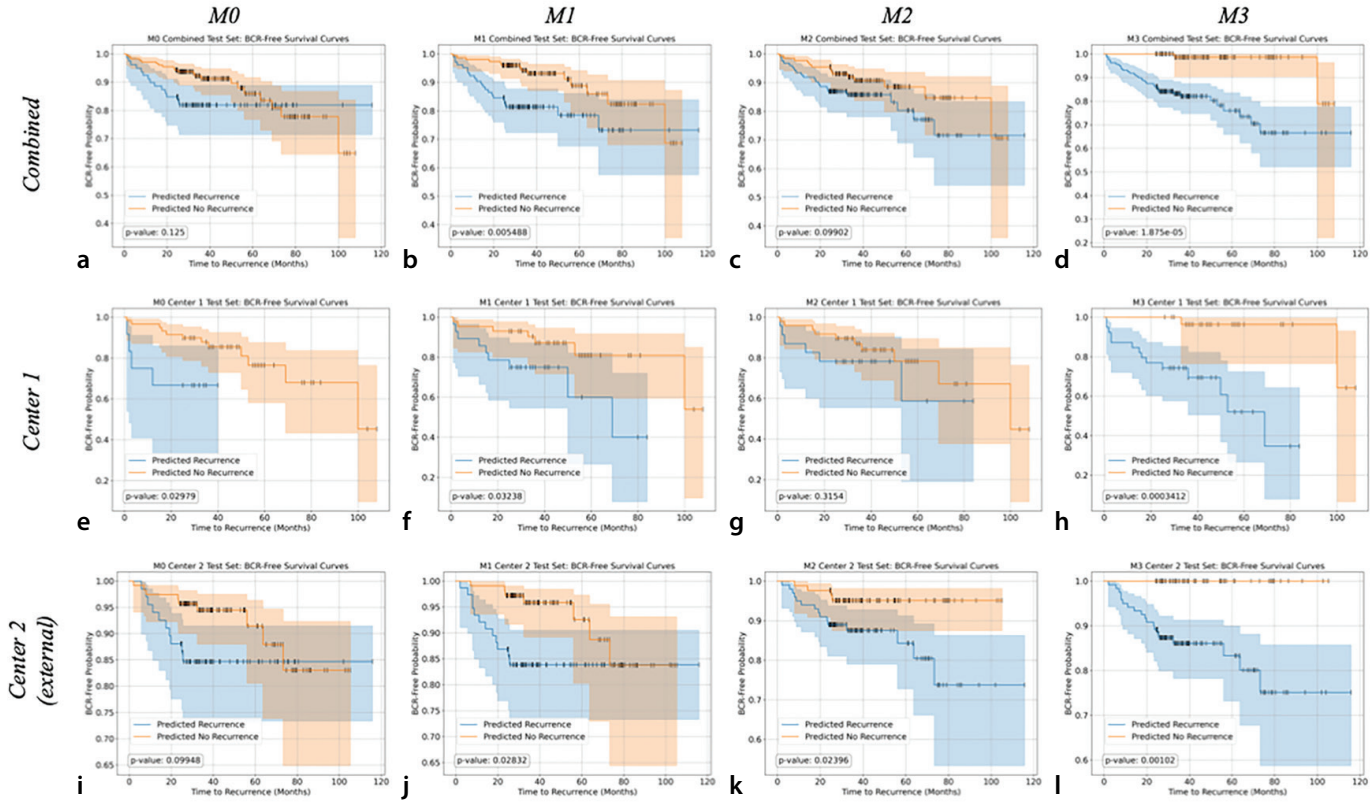

Figure 4. Recurrence-free survival curves by model and center with log-rank tests. Kaplan–Meier survival curves with error bars for artificial intelligence prognostic predictions from each model in the combined center test set. **(a)** M0: complete clinical model **(b)** M1: automated clinical model. **(c)** M2: imaging model. **(d)** M3: multimodal model. **(e–h)** Center 1 test set by model (note limited sample size). **(i–l)** Center 2 external validation set by model.

Discussion

We present the development and external validation of an automated multimodal AI-based model to predict outcomes following RP from baseline bpMRI, age, and PSA. Across both the internal and external validation sets, M3—the fully automated multimodal AI model—was the only model that stratified patients into meaningful risk groups for predicting BCR-free survival across each center and within intermediate-risk sub-groups. With intermediate-risk patients being the most difficult group for outcome prediction based on current gold standards, a model that can reproducibly and accurately predict patient outcomes within this group, independent of current metrics, is an important contribution and could potentially further delineate patient risk and inform treatment strategies. Although reader-based scoring systems, such as NCCN groups, Gleason grades, and PI-RADS, are incredibly important, they may yield inconsistent predictions due to reader variability. An automated approach, such as the one presented here, could be instrumental after further multicenter external validation and ethical testing to inform physicians and patients of individualized surgical risk, independent of reader analyses.

The field of AI in prostate imaging has garnered considerable attention in the past decade; however, existing research often faces barriers to being realistically implemented in the clinic. Leading research tends to be non-automated through the use of reader contours or biopsy grades. Single-center studies lacking external validation are common but do not provide evidence of reproducibility. With bpMRI being a relatively new component of the PCa prognostication process, its evaluation may be inconsistent and relatively under-explored, with guidelines incorporating it only in the past decade or so. Our radiomics extraction pipeline is fully automated, outperforming CAPRA-S in our internal test cohort from center 1 and ISUP GGG grades in the external center 2 test set, providing evidence that our results may be generalizable and that these features have clinical value. Our results provide evidence that AI may further refine prognostic abilities independently of existing clinical gold standards. This could be especially helpful for intermediate-risk patients.

Several studies have explored the clinical relevance of quantitative MRI-based assessments, although few venture to take a multimodal approach integrating clinical covariates.[16] Those that do so generally focus on single-center studies and do not provide evidence that these results generalize to external centers[13,17] or incorporate non-automated features, such as contours or biopsy reads, which are subject to reader variability.[18,19] Some also focus on comparisons between low/intermediate vs. high-risk groupings rather than focusing on challenging comparisons, such as within the intermediate-risk group.[16] More common in the PCa AI space is a focus on histopathology or even integration of pathology and MRI.[17,20] However, digitized pathology is not necessarily feasible in many clinical settings, and many studies focus on post-operative pathology, which is not relevant for stratifying pre-treatment risk.

In comparing our results with the existing literature, we observe that few models are directly comparable given our focus on multimodal (radiomics and clinical covariates) and automated pre-operative MRI in PCa. Our AUC of 0.75 on an external validation set ap-

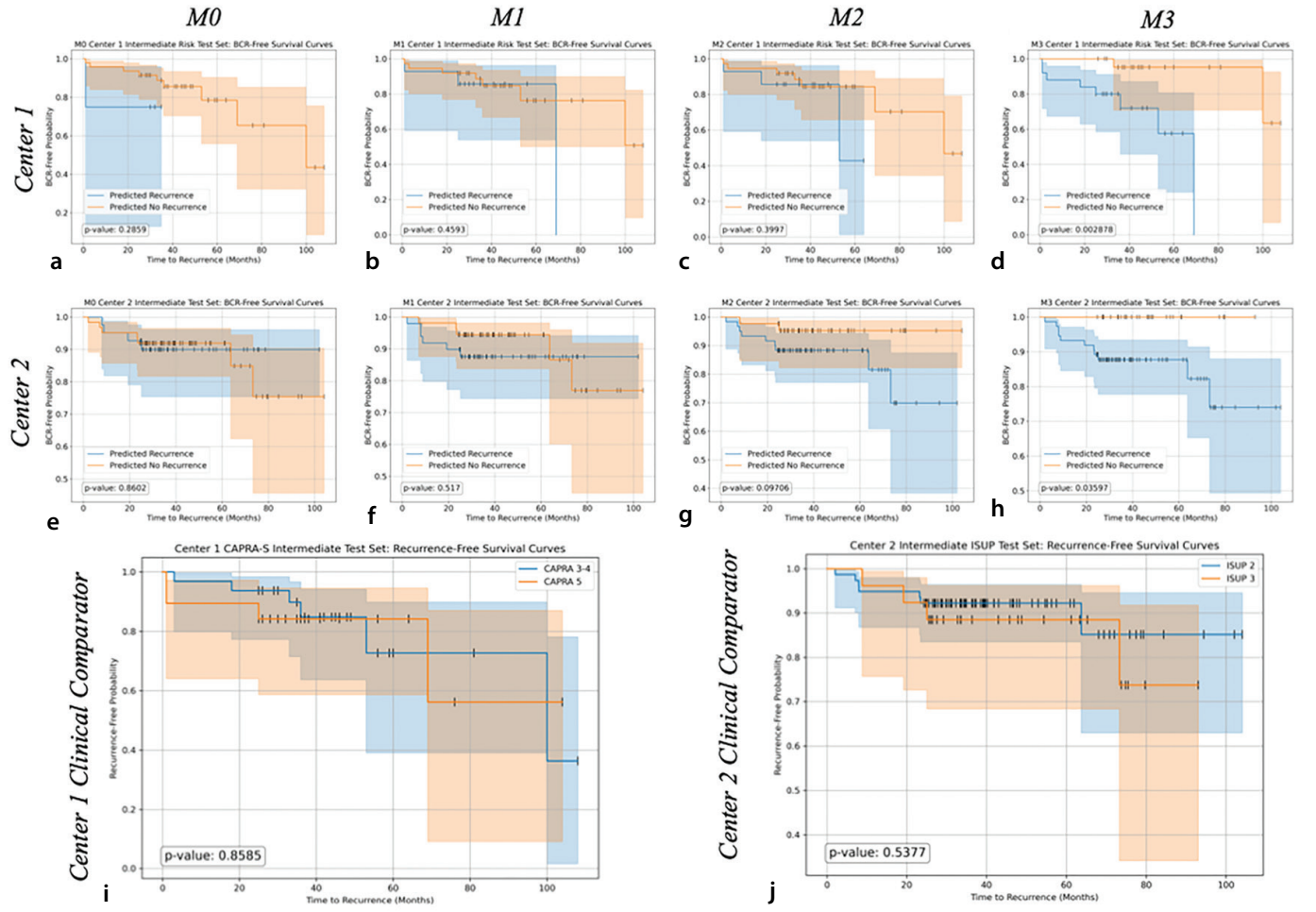

Figure 5. Intermediate-risk comparisons by center. Recurrence-free survival curves for intermediate-risk patients by model and center with clinical baseline and log-rank tests. **(a–d)** Kaplan–Meier curve for each model M0–M3 in center 1. **(i)** Kaplan–Meier curve for Cancer of the Prostate Risk Assessment Postsurgical scores in center 1. **(e–h)** Kaplan–Meier Curve for each model M0–M3 in center 2. **(j)** Kaplan–Meier curve for ISUP grade in center 2. CAPRA-S, Cancer of the Prostate Risk Assessment Postsurgical; ISUP, International Society of Urological Pathology.

pears consistent with the existing literature, and our sample size and external validation provide evidence of generalizability. One similar study focusing on imaging alone reported an AUC of 0.73 for predicting BCR.[21] However, their external validation set was considerably smaller, with n = 50 patients and seven BCR events. Another study featured an even smaller cohort of n = 18 test-set patients and reported an AUC of 0.73[22] Other studies reporting similar or higher AUCs focus solely on single centers or physician reads,[13,23] highlighting concerns about reproducibility and bias.[18,24] The fact that M3 can differentiate intermediate-risk outcomes with significance in an external center in a completely automated fashion that would not be subject to reader variability presents an exciting contribution to the PCa prognostic space. Nevertheless, it is important to acknowledge and consider differences in performance across centers for such a model and whether centers have similar patient populations and vendors. Although center 1 and center 2 perform similarly given the use of different vendors, it is possible that these patient populations are similar, and the model performance may degrade in further external validations. Multicenter analyses and heterogeneity assessment are imperative to ensure fair and comprehensive efficacy.

Although the main focus of this study is not explainability, the weighted feature importances provide some insight into the features the algorithm relies upon for decision making. For example, it is notable that PSA was the most heavily weighted feature in models M0, M1, and M3. Furthermore, radiomic features associated with tumor heterogeneity, spatial irregularity in tumor intensity patterns, and tumor size and shape were heavily weighted in models M2 and M3. A detailed discussion of features can be found in the Supplementary Discussion.

Although the results indicate that M3 may provide value in PCa prognostics beyond that of existing metrics, several aspects of the study limit its applications and demand further study before clinical implementation. First, we did not have information available to calculate CAPRA-S in center 2 nor information about adjuvant radiotherapy status. Although the external validation results are promising, they must be approached with caution given the difference in the clinical comparison standard. Future research should quantify equivalent gold standards across training and external validation cohorts. Second, BCR-free survival times present a limitation, as the model itself did not account for time to BCR, although our BCR-free survival analysis did. In addition, given that a substantial portion of our patients had < 3 years of follow-up, it is possible that some of these patients experienced BCR at a later date (25th percentile of 2.7 and 2.4 years in centers 1 and 2, respectively), thus imperfectly training the model and evaluating its results. Furthermore, although external validation provides some evidence of generalizability, given that this study relies on only

two centers, it is worth questioning the variability of clinical environments. A larger-scale multicenter or national study is needed to verify that these results are clinically reliable regardless of specific center demographics and imaging technology. Finally, the lower specificity of M3 compared with the other models presents a limitation that raises the question of what the clinical utility of such a model might be. Although this might be primarily a result of cut-off calibration given the improved AUCs, this model may tend to flag patients unnecessarily. This should be addressed before any clinical implementation.

In conclusion, we present a fully automated deep learning-based multimodal model that can predict BCR after RP in patients with PCa using baseline bpMRI and routinely used clinical covariates and demonstrates promising generalizability in an external validation center.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

This work was supported by the intramural program at the National Institutes of Health, National Cancer Institute. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Supplementary Methods: https://d2v96fxpocvxx.cloudfront.net/cf9d60d6-523c-458a-a2e6-78728d3ffbb0/content-images/fa4ac6cb-c19d-4114-b712-942425d170d4.pdf

Supplementary Discussion: https://d2v96fxpocvxx.cloudfront.net/cf9d60d6-523c-458a-a2e6-78728d3ffbb0/content-images/84834bc0-f7a7-43d3-847c-68bd309afe16.pdf

Supplementary Table: https://d2v96fxpocvxx.cloudfront.net/a426c3a3-a110-40af-a6dd-1b2b563ce9ac/content-images/fdef2e70-13cf-45ae-a68c-b86128b3de85.pdf

Supplementary Figures: https://d2v96fxpocvxx.cloudfront.net/a426c3a3-a110-40af-a6dd-1b2b563ce9ac/content-images/bf9fcaf4-1c6b-4980-b529-15a256d86b62.pdf

References

1. Siegel RL, Kratzer TB, Giaquinto AN, Sung H, Jemal A. Cancer statistics, 2025. *CA Cancer J Clin*. 2025;75(1):10-45. [Crossref]
2. Schafer EJ, Laversanne M, Sung H, et al. Recent patterns and trends in global prostate cancer incidence and mortality: an update. *Eur Urol*. 2025;87(3):302-313. [Crossref]
3. Raychaudhuri R, Lin DW, Montgomery RB. Prostate cancer: a review. *JAMA*. 2025;333(16):1433-1446. [Crossref]
4. Hamdy FC, Donovan JL, Lane JA, et al. Fifteen-year outcomes after monitoring, surgery, or radiotherapy for prostate cancer. *N Engl J Med*. 2023;388(17):1547-1558. [Crossref]
5. Partin AW, Borland RN, Epstein JI, Brendler CB. Influence of wide excision of the neurovascular bundle(s) on prognosis in men with clinically localized prostate cancer with established capsular penetration. *J Urol*. 1993;150(1):142-146; discussion 146-148. [Crossref]
6. Zhu M, Gao J, Han F, et al. Diagnostic performance of prediction models for extraprostatic extension in prostate cancer: a systematic review and meta-analysis. *Insights Imaging*. 2023;14(1):140. [Crossref]
7. Simon BD, Ozyoruk KB, Gelikman DG, Harmon SA, Türkbey B. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. *Diagn Interv Radiol*. 2025;31(4):303-312. [Crossref]
8. Bhattacharya I, Khandwala YS, Vesal S, et al. A review of artificial intelligence in prostate cancer detection on imaging. *Ther Adv Urol*. 2022;14:17562872221128791. [Crossref]
9. Cooperberg MR, Hilton JF, Carroll PR. The CAPRA-S score: a straightforward tool for improved prediction of outcomes after radical prostatectomy. *Cancer*. 2011;117(22):5039-5046. [Crossref]
10. Mehralivand S, Yang D, Harmon SA, et al. A cascaded deep learning-based artificial intelligence algorithm for automated lesion detection and classification on biparametric prostate magnetic resonance imaging. *Acad Radiol*. 2022;29(8):1159-1168. [Crossref]
11. Simon BD, Merriman KM, Harmon SA, et al. Automated detection and grading of extraprostatic extension of prostate cancer at MRI via cascaded deep learning and random forest classification. *Acad Radiol*. 2024;31(10):4096-4106. [Crossref]
12. Belue MJ, Mukhtar V, Ram R, et al. External validation of an artificial intelligence algorithm using biparametric MRI and Its simulated integration with conventional PI-RADS for prostate cancer detection. *Acad Radiol*. 2025;32(7):3813-3823. [Crossref]
13. Simon BD, Harmon SA, Merriman KM, et al. A multimodal automated deep learning-based model for predicting biochemical recurrence of prostate cancer following prostatectomy from baseline MRI, presurgical clinical covariates. *Clin Imaging*. 2025;126:110579. [Crossref]
14. Cornford P, van den Bergh RCN, Briers E, et al. EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer-2024 update. Part I: screening, diagnosis, and local treatment with curative intent. *Eur Urol*. 2024;86(2):148-163. [Crossref]
15. Dess RT, Suresh K, Zelefsky MJ, et al. Development and validation of a clinical prognostic stage group system for nonmetastatic prostate cancer using disease-specific mortality results from the international staging collaboration for cancer of the prostate. *JAMA Oncol*. 2020;6(12):1912-1920. [Crossref]
16. Park SY, Oh YT, Jung DC, et al. Prediction of biochemical recurrence after radical prostatectomy with PI-RADS version 2 in prostate cancers: initial results. *Eur Radiol*. 2016;26(8):2502-2509. [Crossref]
17. Hu C, Qiao X, Huang R, Hu C, Bao J, Wang X. Development and validation of a multimodality model based on whole-slide imaging and biparametric MRI for predicting postoperative biochemical recurrence in prostate cancer. *Radiol Imaging Cancer*. 2024;6(3):e230143. [Crossref]
18. Li L, Shiradkar R, Leo P, et al. A novel imaging based nomogram for predicting post-surgical biochemical recurrence and adverse pathology of prostate cancer from pre-operative bi-parametric MRI. *EBioMedicine*. 2021;63:103163. [Crossref]
19. Manceau C, Beauval JB, Lesourd M, et al. MRI characteristics accurately predict biochemical recurrence after radical prostatectomy. J Clin Med. 2020;9(12):3841. [Crossref]
20. Gu WJ, Liu Z, Yang YJ, et al. A deep learning model, NAFNet, predicts adverse pathology and recurrence in prostate cancer using MRIs. *NPJ Precis Oncol*. 2023;7(1):134. [Crossref]
21. Shiradkar R, Ghose S, Jambor I, et al. Radiomic features from pretreatment biparametric MRI predict prostate cancer biochemical recurrence: Preliminary findings. *J Magn Reson Imaging*. 2018;48(6):1626-1636. [Crossref]
22. Zhong QZ, Long LH, Liu A, et al. Radiomics of multiparametric MRI to predict biochemical recurrence of localized prostate cancer after radiation therapy. *Front Oncol*. 2020;10:731. [Crossref]
23. Wu XH, Ke ZB, Chen ZJ, et al. Periprostatic fat magnetic resonance imaging based radiomics nomogram for predicting biochemical recurrence-free survival in patients with non-metastatic prostate cancer after radical prostatectomy. *BMC Cancer*. 2024;24:1459. [Crossref]
24. Wu SY, Wang Y, Fan P, et al. Bi-parametric MRI-based quantification radiomics model for the noninvasive prediction of histopathology and biochemical recurrence after prostate cancer surgery: a multicenter study. *Abdom Radiol (NY)*. 2025;50(9):4320-4330. [Crossref]