An integrated nomogram combining deep learning, Prostate Imaging–Reporting and Data System (PI-RADS) scoring, and clinical variables for identification of clinically significant prostate cancer on biparametric MRI: a retrospective multicentre study

Amogh Hiremath, Rakesh Shiradkar, Pingfu Fu, Amr Mahran, Ardeshir R Rastinehad, Ashutosh Tewari, Sree Harsha Tirumani, Andrei Purysko, Lee Ponsky, Anant Madabhushi

Summary

Background Biparametric MRI (comprising T2-weighted MRI and apparent diffusion coefficient maps) is increasingly being used to characterise prostate cancer. Although previous studies have combined Prostate Imaging–Reporting & Data System (PI-RADS)-based MRI findings with routinely available clinical variables and with deep learning-based imaging predictors, respectively, for prostate cancer risk stratification, none have combined all three. We aimed to construct an integrated nomogram (referred to as ClaD) combining deep learning-based imaging predictions, PI-RADS scoring, and clinical variables to identify clinically significant prostate cancer on biparametric MRI.

Methods In this retrospective multicentre study, we included patients with prostate cancer, with histopathology or biopsy reports and a screening or diagnostic MRI scan in the axial view, from four cohorts in the USA (from University Hospitals Cleveland Medical Center, Icahn School of Medicine at Mount Sinai, Cleveland Clinic, and Long Island Jewish Medical Center) and from the PROSTATEx Challenge dataset in the Netherlands. We constructed an integrated nomogram combining deep learning, PI-RADS score, and clinical variables (prostate-specific antigen, prostate volume, and lesion volume) using multivariable logistic regression to identify clinically significant prostate cancer on biparametric MRI. We used data from the first three cohorts to train the nomogram and data from the remaining two cohorts for independent validation. We compared the performance of our ClaD integrated nomogram with that of integrated nomograms combining clinical variables with either the deep learning-based imaging predictor (referred to as DIN) or PI-RADS score (referred to as PIN) using area under the receiver operating characteristic curves (AUCs). We also compared the ability of the nomograms to predict biochemical recurrence on a subset of patients who had undergone radical prostatectomy. We report cross-validation AUCs as means for the training set and used AUCs with 95% CIs to assess the performance on the test set. The difference in AUCs between the models were tested for statistical significance using DeLong's test. We used log-rank tests and Kaplan-Meier curves to analyse survival.

Findings We investigated 592 patients (823 lesions) with prostate cancer who underwent 3T multiparametric MRI at five hospitals in the USA between Jan 8, 2009, and June 3, 2017. The training data set consisted of 368 patients from three sites (the PROSTATEx Challenge cohort [n=204], University Hospitals Cleveland Medical Center [n=126], and Icahn School of Medicine at Mount Sinai [n=38]), and the independent validation data set consisted of 224 patients from two sites (Cleveland Clinic [n=151] and Long Island Jewish Medical Center [n=73]). The ClaD clinical nomogram yielded an AUC of 0·81 (95% CI 0·76–0·85) for identification of clinically significant prostate cancer in the validation data set, significantly improving performance over the DIN (0·74 [95% CI 0·69–0·80], p=0·0005) and PIN (0·76 [0·71–0·81], p<0·0001) nomograms. In the subset of patients who had undergone radical prostatectomy (n=81), the ClaD clinical nomogram resulted in a significant separation in Kaplan-Meier survival curves between patients with and without biochemical recurrence (HR 5·92 [2·34–15·00], p=0·044), whereas the DIN (1·22 [0·54–2·79], p=0·65) and PIN nomograms did not (1·30 [0·62–2·71], p=0·51).

Interpretation Risk stratification of patients with prostate cancer using the integrated ClaD nomogram could help to identify patients with prostate cancer who are at low risk, very low risk, and favourable intermediate risk, who might be candidates for active surveillance, and could also help to identify patients with lethal prostate cancer who might benefit from adjuvant therapy.

Funding National Cancer Institute of the US National Institutes of Health, National Institute for Biomedical Imaging and Bioengineering, National Center for Research Resources, US Department of Veterans Affairs Biomedical Laboratory Research and Development Service, US Department of Defense, US National Institute of Diabetes and

Lancet Digit Health 2021; 3: e445–54

Department of Biomedical Engineering (A Hiremath MS, R Shiradkar PhD, Prof A Madabhushi PhD) **and Department of Population and Quantitative Health Sciences** (Prof P Fu PhD), **Case Western Reserve University, Cleveland, OH, USA; Urology Institute** (A Mahran MD, Prof L Ponsky MD) **and Department of Radiology** (S H Tirumani MD), **University Hospitals Cleveland Medical Center, Cleveland, OH, USA; Department of Urology, Lenox Hill Hospital, Northwell Health, New York, NY, USA** (A R Rastinehad DO); **Department of Radiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA** (Prof A Tewari MD); **Department of Radiology and Nuclear Medicine, Cleveland Clinic, Cleveland, OH, USA** (A Purysko MD); **Louis Stokes Cleveland Veterans Administration Medical Center, Cleveland, OH, USA** (Prof A Madabhushi)

Correspondence to: Prof Anant Madabhushi, Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH 44106, USA
axm788@case.edu

Digestive and Kidney Diseases, The Ohio Third Frontier Technology Validation Fund, Case Western Reserve University, Dana Foundation, and Clinical and Translational Science Collaborative.

Introduction

Biparametric MRI, which comprises T2-weighted MRI and apparent diffusion coefficient maps derived from diffusion-weighted imaging, is increasingly being used for the detection and characterisation of prostate cancer.[1] The Prostate Imaging–Reporting and Data System (PI-RADS) has standardised the diagnosis of prostate cancer using MRI and is effective for characterisation of prostate cancer.[2] However, MRI is still restricted by benign confounding appearances and substantial intra-reader and inter-reader variability.[3] Therefore, Gleason grade grouping (GGG), which assigns a prostate cancer lesion to one of five categories (1–5) based on invasive biopsies, remains the standard of care in determining the aggressiveness of prostate cancer.

According to the 2017 European Association of Urology prostate cancer guidelines,[4] patients with intermediate-risk and high-risk prostate cancer with clinically significant disease (defined as GGG ≥2 on pathology) are recommended a definitive treatment, such as radical prostatectomy, whereas those with very low-risk or low-risk disease (clinically insignificant prostate cancer, defined as GGG 1) and some with intermediate favourable risk (GGG 2) are recommended to follow an active surveillance strategy wherein patients are closely monitored without being provided any definitive treatment. Therefore, low-risk patients with clinically insignificant prostate cancer lesions should be accurately identified to avoid overdiagnosis and overtreatment. Avoiding under-treatment and identifying patients with lethal prostate cancer that will recur and metastasise despite definitive therapy is equally important. If identified, these patients can benefit further from adjuvant therapy. However, invasive biopsies still remain a de-facto standard in assessment and grading of prostate cancer via GGG.

Prostate cancer nomograms[5,6] are widely used as prognostic tools to understand the nature of prostate cancer, assess risk of disease, and predict probable outcomes of treatment. Although some studies showed that combining prostate MRI with clinical nomograms did not improve performance over prostate MRI alone[7] or over clinical nomograms alone,[8] Rayn and colleagues[9] found that using MRI in combination with clinical nomograms provides significant added benefit in predicting adverse pathologies. Although pre-biopsy prostate cancer risk calculators, such as the Prostate Cancer Prevention Trial[10] and Prostate Biopsy Collaborative

Research in context

Evidence before this study

We searched PubMed for research articles published between database inception and May 27, 2021, using the search terms "prostate cancer risk stratification", or "clinically significant prostate cancer" and "MRI", together with at least one of the following: "machine learning", "artificial intelligence", "deep learning", or "nomogram". This search returned 190 titles and abstracts, which we subsequently reviewed. Although several previous studies have proposed pre-biopsy risk calculators for prostate cancer aggressiveness, most do not include MRI findings. Prostate cancer nomograms have been widely used as prognostic tools for prostate cancer risk assessment, and some of the previous studies have shown the advantage of combining Prostate Imaging–Reporting and Data System (PI-RADS)-based MRI findings with routinely available clinical variables, such as prostate-specific antigen, prostate volume, lesion volume, and digital rectal examination. However, to our knowledge, none of these studies have combined automated interpretation of MRI using artificial intelligent tools (such as deep learning and machine learning with PI-RADS-based MRI interpretation) with routine clinical variables for prostate cancer risk stratification. Additionally, we did not find any studies that had validated the approaches on large multi-site cohorts.

Added value of this study

This study presents a novel integrated nomogram, termed CIaD, constructed by integrating deep learning predictions on biparametric MRI with PI-RADS score and routinely used clinical variables (prostate-specific antigen, prostate volume, and lesion volume). The CIaD nomogram was able to identify clinically significant prostate cancer across a large multi-institutional cohort of 592 patients with prostate cancer. CIaD used deep learning-derived image patterns from both the intratumoural and the peritumoural regions. The CIaD nomogram could not only identify clinically significant prostate cancer regions, but could also predict biochemical recurrence-free survival in a subset of 81 patients who had undergone radical prostatectomy.

Implications of all the available evidence

Accurate risk stratification of patients with prostate cancer by the integrated nomogram could help to identify patients with prostate cancer who are at low risk, very low risk, and favourable intermediate risk of prostate cancer, who might be candidates for active surveillance, and could also help to identify patients with lethal prostate cancer who might benefit from adjuvant therapy.

Group (PBCG)[11] risk calculators, exist, they do not include imaging parameters.

Radiomic features derived from prostate biparametric MRI, which evaluate various quantitative measures in an image (such as shape, volume, surface, and texture), might improve stratification of patients into different risk categories compared with routine imaging.[12] However, this approach requires considerable effort in carefully engineering the features and using feature selection strategies before training a machine learning classifier.

Deep learning approaches with the use of convolutional neural networks are increasingly being investigated in research studies for the characterisation of prostate cancer.[13,14] Deep learning has previously been used to automatically grade prostate cancer or distinguish between clinically significant and clinically insignificant prostate cancer; however, most approaches have used data from a single site.[13,14] Although deep learning-based approaches have been compared with PI-RADS assessment scores,[13] to our knowledge, deep learning has not yet been integrated with PI-RADS assessment scores and other clinical factors to construct a clinical nomogram.

Studies suggest that the peritumoural region (the region immediately surrounding the visible tumour) on imaging includes important information related to the type and characteristics of the disease.[15,16] Although patch-based deep learning approaches using convolutional neural networks are increasingly being used for disease detection and characterisation,[17,18] these studies involve patches extracted around a region of interest delineated within the lesion without making use of important information in the peritumoural region. To our knowledge, no previous deep learning studies have used peritumoural regions to improve prostate cancer characterisation.

We aimed to construct an integrated nomogram (referred to as ClaD) combining deep learning-based imaging predictions, PI-RADS scoring, and clinical variables using multivariable logistic regression to identify clinically significant prostate cancer on biparametric MRI. Since previous research[19] has suggested the importance of the peritumoural region for characterisation of prostate cancer on biparametric MRI, we explored multiple input configurations of the deep learning network, with patches extracted at different scales and each subsequent scale incorporating additional information from the peritumoural region.

Methods

Study design and participants

We retrospectively included patients with prostate cancer who had histopathology or biopsy reports and a screening or diagnostic MRI scan in the axial view from the publicly available PROSTATEx Challenge dataset (Radboud University Medical Centre, Nijmegen, the Netherlands; Cancer Imaging Archive) and four US cohorts from University Hospitals Cleveland Medical Center, Cleveland, OH; Icahn School of Medicine at Mount Sinai, New York, NY; Cleveland Clinic, Cleveland, OH; and Long Island Jewish Medical Center, NY). Descriptions of the inclusion criteria and cohorts are provided in the appendix (pp 2–3). Patients with MRI scans with incomplete sequences or severe motion or other susceptibility artifacts were excluded.

See Online for appendix

We constructed our ClaD clinical nomogram combining deep learning-based imaging predictions, PI-RADS score, and clinical variables (prostate-specific antigen [PSA], prostate volume, and lesion volume) using multivariable logistic regression to identify clinically significant prostate cancer on biparametric MRI. Our clinical nomogram was trained on data from three different sites (D_{train}; from the PROSTATEx Challenge cohort [D1], University Hospitals Cleveland Medical Center [D2], and Icahn School of Medicine at Mount Sinai [D3]) and validated on data from two independent sites (D_{test}; from Cleveland Clinic [D4] and Long Island Jewish Medical Center [D5]). We also used our ClaD nomogram to predict biochemical recurrence in a subset of patients ($\check{D}_{test}$) from D4 who had undergone radical prostatectomy.

This study was compliant with Health Insurance Portability and Accountability Act and was approved by the institutional review board at Case Western Reserve University and University Hospitals Cleveland, which allowed for computational analysis of retrospectively acquired, de-identified imaging data for research purposes. The datasets from all five institutions were acquired under specific data use agreements.

Deep learning

We compared a deep learning-based imaging predictor with two different base architectures, AlexNet[20] and DenseNet,[21] to identify clinically significant prostate cancer lesions on biparametric MRI (figure 1A). The networks comprised three distinct input channels: T2-weighted MRI, apparent diffusion coefficient maps, and the corresponding binary lesion segmentation masks. The networks were trained at the lesion level. We used the highest GGG and the maximum predicted value obtained by the network among lesions to evaluate performance at the patient level.

We used multiple input configurations with different scales (S0–S4; defined in the appendix, pp 4–5) of patches (figure 1B) extracted from T2-weighted MRI and apparent diffusion coefficient maps, with each subsequent scale adding information from the peritumoural region. Briefly, the patches with respect to the first scale (S0) were extracted by drawing a bounding box around the segmented lesion and the patches with subsequent scales (S1–S4) were extracted by extending the bounding box from periphery of the delineated lesion up to 3 mm, 6 mm, 9 mm, and 12 mm, respectively (appendix pp 4–5). For each of the scales, 3-fold cross-validation was done on the training data set (D_{train}). The scale resulting in the best cross-validation

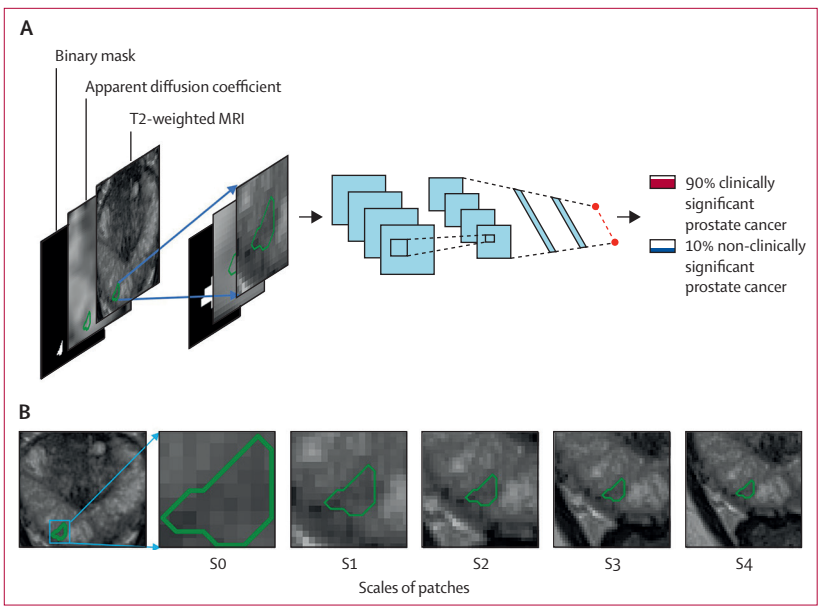


Figure 1: **Architecture and input configuration of the deep learning-based imaging predictor**
Architectural diagram and input configuration of the deep learning-based imaging predictor to identify clinically significant prostate cancer lesions on biparametric MRI. (A) Architectural diagram of the imaging predictor depicting the three different input channels of the network (T2-weighted MRI, apparent diffusion coefficient maps, and binary lesion segmentation masks). (B) Input configurations to the deep learning-based imaging predictor.

area under the receiver operating characteristic (ROC) curve (AUC) was chosen to evaluate performance on the validation data set (D_{test}). We used a binary entropy loss function when training the deep learning network, and an early stopping criterion was used to stop the network training with respect to the leave-one-out cross-validation loss. The network training was optimised using an Adam optimiser, with a weight decay of 10^{-5} and a learning rate of 10^{-6}. Further details on network implementation are given in the appendix (p 5).

To evaluate the stability of the deep learning-based imaging predictor, we used a Quantitative Imaging Network-PROSTATE-Repeatability dataset[22] from The Cancer Imaging Archive, consisting of baseline and repeat prostate multiparametric MRI scans of 15 individuals taken 2 weeks apart (inclusion criteria and data processing details are provided in the appendix, p 5). We calculated repeatability between the output network predictions on baseline and repeat scans in terms of the intraclass correlation coefficient.

To interpret the results of the deep learning-based imaging predictor, we used gradient-weighted class activation mapping[23] to identify specific regions in the image patches that contributed the most to successful predictions.

Clinical nomograms

Since clinical information and PI-RADS assessment scores were not available for one of the cohorts (PROSTATEx Challenge), we used data for the other two cohorts in the training data set to construct the nomograms.

We estimated lesion volume and prostate volume by multiplying the number of voxels from their corresponding manual annotations with their corresponding voxel spacing. We did a univariate analysis and multivariable analysis on clinical parameters (PSA, prostate volume, and lesion volume), PI-RADS score, and output probabilities from the deep learning-based imaging predictor. We then trained a logistic regression classifier on the independently predictive parameters from the multivariable analysis to obtain the clinical nomogram. Descriptions of data pre-processing and the data augmentation steps followed to construct the clinical nomogram are available in the appendix (pp 2–4).

We compared our clinical nomogram with a logistic regression classifier trained on independently predictive deep learning-based and clinical variables from the multivariable analysis (DIN), and with a logistic regression classifier trained on independently predictive PI-RADS and clinical variables from the multivariable (PIN). We used R software (rms package, with nomogram and lrm functions) to build the nomogram models and used a penalty factor of 20 for regularisation and to avoid over-fitting of the model.

We did four post-hoc analyses: (1) biopsy and ClaD undergrading (data from D4); (2) distinguishing between GGG 1 and GGG 2 lesions (data from D4 and D5); (3) comparison of nomograms with PBCG (data from D5); and (4) biochemical recurrence analysis (data from D4).

Owing to the random sampling nature of 12-core systematic biopsies and the confounding factors involved in magnetic resonance-guided and ultrasound fusion-guided biopsies, prostate cancer lesions can be undergraded. Since the integrated nomogram ClaD was trained with ground-truth from 12-core systematic biopsies and magnetic resonance-guided and ultrasound fusion-guided biopsies, we analysed the under-grading of ClaD with respect to the ground-truth from surgical specimens, and compared the under-grading with 12-core systematic biopsies and magnetic resonance-guided and ultrasound fusion-guided biopsies. We analysed under-grading in a subset of patients from Cleveland Clinic (D4) because they had data on ground-truth from all three sources (12-core systematic biopsies, magnetic resonance-guided and ultrasound fusion-guided biopsies, and surgical specimens).

To have clinical benefit, the nomograms must perform well in the diagnostic grey area, which means distinguishing between patients with GGG 1 and GGG 2 lesions. Therefore, we chose a subset of patients from the D_{test} validation data set (both cohorts) to evaluate the performance of the deep learning-based imaging predictor, and the DIN, PIN, and integrated ClaD models in distinguishing between patients with GGG 1 and GGG 2 lesions.

The PBCG risk calculator,[11] which has replaced the earlier Prostate Cancer Prevention Trial risk calculator,[10] is one of the most widely used prediction tools to

estimate the risk of high grade prostate cancer (GGG >1). We compared the performance of our integrated nomogram, ClaD, with the PBCG risk calculator in the validation set (D5 only). We chose the D5 cohort because only D5 had all the information available to evaluate the PBCG risk calculator against DIN, PIN, and ClaD.

We did a post-hoc analysis ($\check{D}_{test}$) of a subset of patients (Cleveland Clinic, D4) in the D_{test} validation data set who had undergone radical prostatectomy (inclusion criteria listed in the appendix, p 6) to investigate biochemical recurrence-free survival. We dichotomised (positive or negative for biochemical recurrence) the predicted class probability obtained from the models for the deep learning-based imaging predictor, and the DIN, PIN, and ClaD nomograms using the chosen cutoff point on the ROC curve. We then used the dichotomised labels for survival analysis using Kaplan-Meier curves.

Statistical analysis

For the identification of clinically significant lesions on biparametric MRI, we trained the models by labelling clinically significant lesions as 1 and clinically insignificant lesions as 0. We used AUC and other performance metrics (accuracy, sensitivity, and specificity) to compare performance between the models (the deep learning-based imaging predictor, and the DIN, PIN, and ClaD nomograms). We tested the differences in AUC between the models for statistical significance using DeLong's test.[24] We report the training cross-validation AUCs as means with SD; performance in the validation set is reported as AUCs with 95% CIs. 95% CIs were calculated by bootstrapping the ROC curve more than 2000 times. We chose optimal cutoff points by maximising the accuracy in the training ROC curves.

For test-retest analysis of the deep learning-based imaging predictor, we analysed the repeatability of the output predictions using biparametric MRI scans of patients taken 2 weeks apart. Intraclass correlation coefficient (3,1)[25] scores are reported as a measure of repeatability.

We used univariate and multivariable analyses to construct the nomograms. We computed AUCs, log odds ratios (ORs), and p values of individual variables for univariate analysis; log ORs and p values are reported for multivariable analysis. We used decision curve analyses to illustrate the overall net-benefit of using one model versus another.

We used Kaplan-Meier survival curves for time-to-event analysis for biochemical recurrence. We used the dichotomised values (negative or positive for biochemical recurrence) of predictions from the models (the deep learning-based imaging predictor, and the DIN, PIN, and ClaD nomograms), obtained from the chosen cutoff point on the ROC curve, to construct the Kaplan-Meier curves. We used log-rank tests to determine statistically significant differences ($p<0\cdot05$) in biochemical recurrence-free survival.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

We investigated 592 patients (823 lesions) with prostate cancer who underwent 3T multiparametric MRI at one hospital in the Netherlands and four hospitals in the USA between Jan 8, 2009, and June 3, 2017. The training data set consisted of 368 patients from three sites (from the PROSTATEx Challenge cohort [D1; n=204], University Hospitals Cleveland Medical Center [D2; n=126], and Icahn School of Medicine at Mount Sinai [D3; n=38]), and the validation data set consisted of 224 patients from

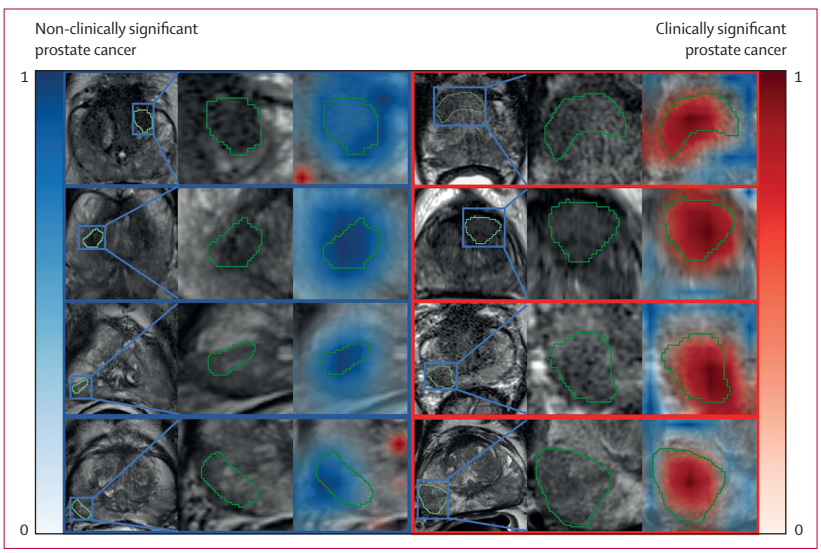
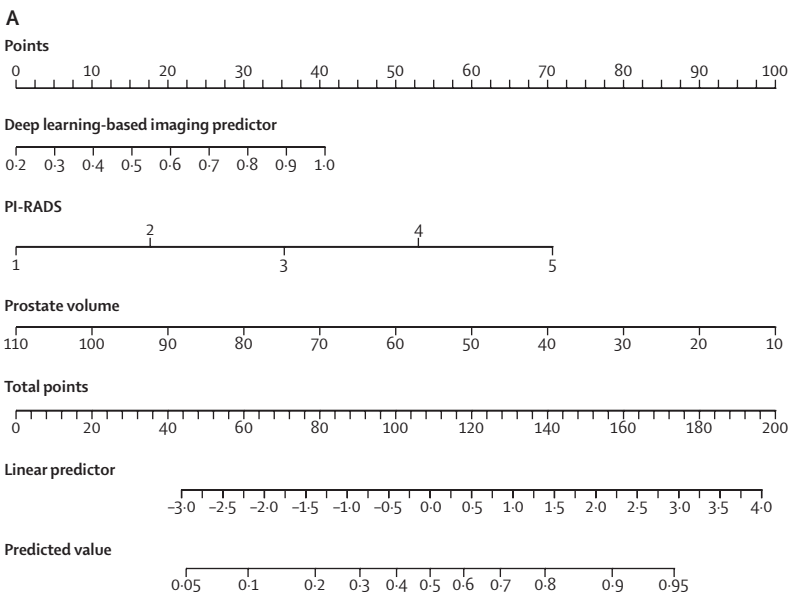


Figure 2: **Identification of regions contributing to predictions of clinically significant prostate cancer by the deep learning-based imaging predictor**
Model interpretability results of guided gradient-weighted class activation mapping to identify regions that most contributed to predictions of clinically insignificant or benign prostate cancer lesions (shown in blue) and clinically significant prostate cancer lesions (shown in red) by the deep learning-based imaging predictor on scale S1 (ie, patches extracted with the bounding box extended 3 mm from the tumour periphery). The colour gradient corresponds to the contribution of a pixel towards a clinically insignificant or clinically significant prostate cancer region, with darker colours showing a greater contribution.

	AUC (95% CI)	Log odds ratio	p value
Deep learning-based imaging predictor (0–1)	0·76 (0·67–0·86)	3·98	<0·0001
PI-RADS (1–5)	0·72 (0·61–0·82)	0·52	0·001
PSA, ng/mL	0·62 (0·51–0·74)	0·015	0·01
Prostate volume, mm³	0·76 (0·65–0·87)	–0·06	<0·0001
Lesion volume, mm³	0·75 (0·66–0·84)	3·8	0·003

AUC=area under the receiver operating characteristic curve. PI-RADS=Prostate Imaging–Reporting and Data System. PSA=prostate-specific antigen.

Table 1: **Univariate analysis**



B

Variable	Log odds ratio	p value
Deep learning-based imaging predictor (0–1)	1·44	0·033
PI-RADS (1–5)	0·42	0·016
PSA (ng/mL)	0·0051	0·76
Prostate volume (mm³)	−0·026	0·0038
Lesion volume (mm³)	0·070	0·45

Figure 3: **Integrated ClaD nomogram, combining deep learning, PI-RADS, and clinical variables**
(A) An integrated nomogram combining deep learning, PI-RADS, and clinical variables constructed by integrating output probability score of the deep learning-based imaging predictor, PI-RADS score, and clinical variables to distinguish between clinically significant and clinically insignificant prostate cancer. (B) Multivariable logistic regression analysis of the deep learning-based imaging predictor, PI-RADS scoring, and clinical variables. PI-RADS=Prostate Imaging-Reporting and Data System. PSA=prostate-specific antigen.

	Deep learning-based imaging predictor	DIN	PIN	ClaD
Threshold	0·361	0·549	0·639	0·571
AUC (95% CI)	0·76 (0·71–0·81)	0·74 (0·69–0·80)	0·76 (0·71–0·81)	0·81 (0·76–0·85)
p value	0·004	0·0005	<0·0001	..
Accuracy	69·36%	76·10%	75·67%	77·92%
Sensitivity	69·94%	86·70%	81·50%	83·23%
Specificity	67·34%	38·77%	55·10%	59·18%

AUC=area under the receiver operating characteristic curve. ClaD=integrated clinical nomogram combining deep learning, PI-RADS, and clinical variables. DIN=integrated nomogram combining deep learning-based imaging prediction and clinical variables. PIN=integrated nomogram combining PI-RADS scoring and clinical variables. PI-RADS=Prostate Imaging–Reporting and Data System.

Table 2: **Performance metrics**

two sites (Cleveland Clinic [D4; n=151] and Long Island Jewish Medical Center [D5; n=73]). The study profile is shown in the appendix (p 7, with data on patient demographics on p 14). All patients had 3-Tesla multiparametric MRI scans, using either endorectal coil (n=108) or pelvic phased-array coil (n=484; data on MRI acquisition criteria are provided in the appendix p 15). Pathological assessment of GGG using either radical prostatectomy specimens or biopsy reports identified 398 clinically significant prostate cancer lesions (GGG ≥2) and 425 lesions with either clinically insignificant prostate cancer (GGG 1) or no prostate cancer on biopsy results (detailed data in appendix p 14).

For the deep learning-based imaging predictor, the performances of AlexNet[20] and DenseNet[21] were similar; however, we chose AlexNet as our base architecture because it had slightly better performance than DenseNet on the training data set (appendix p 16). Among different input configurations to the network, patches with scale S1 best identified clinically significant prostate cancer lesions with a cross validation AUC of 0·75 (SD 0·01) using the training data set (D_{train}; appendix p 16). Patient-level analysis using scale S1 resulted in an AUC of 0·72 (SD 0·03) for identification of patients with clinically significant prostate cancer lesions. The optimal deep learning-based imaging predictor resulted in an AUC of 0·76 (95% CI 0·71–0·81) for identification of patients with clinically significant prostate cancer lesions in the D_{test} validation data set.

The gradient-weighted class activation maps showed that, together with the tumoural region, the peritumoural region (S1; 3 mm from the tumour periphery) also contributes to the predictions (figure 2). Although, other input configurations with peritumoural (S2–4) were included in the analysis, S1 resulted in the best performance, indicating the importance of peritumoural regions between 0 mm and 3 mm. Moderate repeatability, with an intraclass correlation coefficient of 0·71 (n=10; appendix p 5), was seen between deep learning-based imaging predictions on baseline and repeat scans from the of Quantitative Imaging Network-PROSTATE-Repeatability dataset.

Univariate analysis showed that all five variables (deep learning-based imaging prediction, PI-RADS score, PSA, prostate volume, and lesion volume) were predictive (p≤0·01) of clinically significant prostate cancer with AUCs in the range 0·62–0·76 (table 1). Multivariable analysis using DIN showed that only the deep learning-based imaging prediction and prostate volume were independently predictive of clinically significant prostate cancer (appendix p 8). Similarly, multivariable analysis using PIN suggested that only PI-RADS and prostate volume were independently predictive of clinically significant prostate cancer (appendix p 9).

On the D_{test} validation data set, the AUCs were 0·74 (95% CI 0·69–0·80) for DIN and 0·76 (0·71–0·81) for PIN. DeLong's test showed that predictions of DIN and PIN did not differ significantly (p=0·35). Multivariable analysis of the deep learning-based imaging predictor,

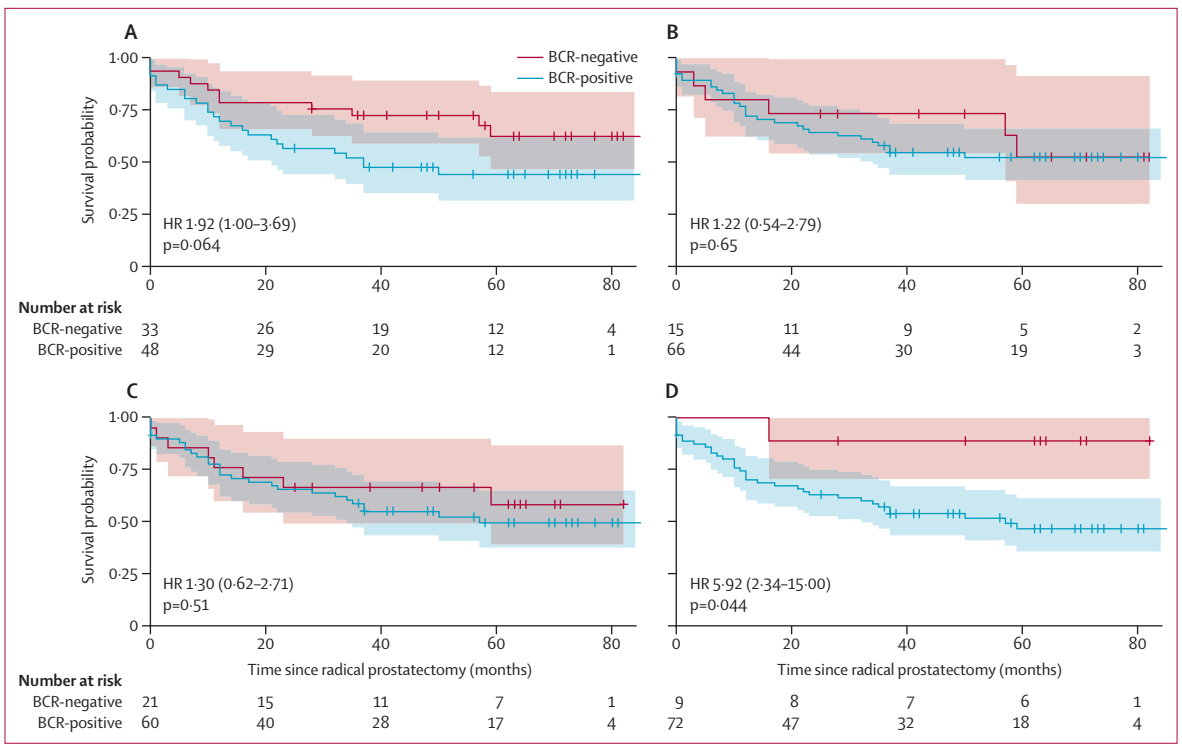


Figure 4: **Kaplan-Meier analysis of BCR-free survival**
Predictions obtained from the deep learning-based imaging predictor (A), integrated nomogram combining deep learning and clinical variables (B), integrated nomogram combining PI-RADS scoring and clinical variables (C), and integrated nomogram combining deep learning, PI-RADS scoring, and clinical variables (D). The predicted class probability obtained from models were dichotomised (BCR-negative or BCR-positive) using the chosen operating point on the receiver operating characteristic curve. Dichotomised labels were then used for the survival analysis. BCR=biochemical recurrence. HR=hazard ratio. PI-RADS=Prostate Imaging-Reporting and Data System.

PI-RADS, and clinical variables showed that the deep learning-based imaging predictor (p=0·03), PI-RADS (p=0·02), and prostate volume (p=0·004) were independently predictive of other clinical variables (lesion volume and PSA; figure 3). ClaD improved performance over DIN (p=0·0005) and PIN (p<0·0001), yielding an AUC of 0·81 (95% CI 0·76–0·85) on the D_{test} validation data set. This shows that including PI-RADS in the nomogram led to a statistically significant increase of 7% in AUC compared with DIN, which does not include PI-RADS.

The ClaD clinical nomogram had greater accuracy than the deep learning-based imaging predictor, DIN, and PIN (table 2). The ClaD clinical nomogram predicted that a score of greater than 0·571 (optimal cutoff point on ROC curve) suggested the presence of clinically significant prostate cancer, whereas scores less than or equal to 0·571 indicated clinically insignificant prostate cancer and benign lesions. Choosing a cutoff of 0·571 on ClaD could help to avoid 59·18% of unnecessary biopsies among low-risk patients with clinically insignificant or benign lesions, at the cost of missing the clinically significant disease in 16·77% of intermediate-risk or high-risk patients with prostate cancer.

In patients with reference ground-truth from radical prostatectomy specimens (the Cleveland Clinic cohort, D4), the ClaD clinical nomogram had an AUC of 0·83 (95% CI 0·75–0·91). In patients with ground-truth obtained from magnetic resonance-guided or ultrasound fusion-guided biopsies (the Long Island Jewish Medical Center cohort), the ClaD nomogram had an AUC of 0·88 (0·82–0·94). The decision curve analysis on the D_{test} validation data set showed an added net-benefit of using the integrated ClaD model over the deep learning-based imaging predictor and PIN (appendix p 10). The standardised net-benefit observed for a high-risk score is greater than 0·3 (appendix p 10).

In a subset of 36 patients from the Cleveland Clinic (D4) cohort who underwent radical prostatectomy, we compared the under-grading of the 12-core systematic biopsies, magnetic resonance-guided biopsies, or ultrasound fusion-guided biopsies, and ClaD predictions with respect to radical prostatectomy specimens. Our proposed nomogram failed to detect clinically significant prostate cancer lesions in seven (19%) of 36 patients, compared with 10 (28%) for the 12-core systematic biopsy and 13 (36%) for magnetic resonance-guided or ultrasound fusion-guided biopsy approaches, resulting in a sensitivity of 79·4% for ClaD (appendix p 11).

To test whether the nomograms could distinguish between patients with GGG 1 and GGG 2 lesions, we

analysed a subset of 158 patients (n=49 with GGG 1 and n=109 with GGG 2) from the D_{test} validation data set (both cohorts). The AUC of the models ranged from 0·72 to 0·75 (appendix p 12), with the ClaD clinical nomogram obtaining the highest AUC of 0·75 (95% CI 0·70–080; p<0·01). The ClaD clinical nomogram distinguished between patients with GGG 1 and GGG 2 with 72·2% accuracy, 77·1% sensitivity, and 61·2% specificity.

Among the validation set, only one cohort (Long Island Jewish Medical Center; n=73) had all the information available to evaluate the PBCG nomogram. In this cohort, the DIN, PIN, and ClaD nomograms outperformed the PBCG risk calculator (p<0·03; appendix p 13) with AUCs in the range of 0·83–0·88; the ClaD nomogram had the highest AUC (0·88; 95% CI 0·82–0·94).

In our post-hoc analysis of biochemical recurrence in 81 patients from Cleveland Clinic (D4) who had undergone radical prostatectomy, the ClaD clinical nomogram resulted in a significant separation in Kaplan-Meier survival curves between patients with and without biochemical recurrence using the dichotomised predictions (p=0·044). None of the other models distinguished between patients according to biochemical recurrence (p=0·064 for the deep learning-based imaging predictor, p=0·65 for DIN, and p=0·51 for PIN; figure 4).

Discussion

In this study, we constructed ClaD, an integrated clinical nomogram combining deep learning, PI-RADS score, and clinical variables to identify clinically significant prostate cancer lesions on biparametric MRI, and compared our nomogram with other integrated nomograms that combined either deep learning or PI-RADS score with clinical variables (DIN and PIN). The ClaD nomogram significantly outperformed DIN and PIN in identification of clinically significant prostate cancer, as indicated by DeLong's test. Given that the standardised net-benefit of ClaD for a high-risk score was greater than 0·3, clinicians and patients should engage in shared decision making regarding use of active treatment if the risk of significant cancer is greater than 0·3. We also showed that integrating imaging and clinical parameters into a nomogram can outperform risk calculators such as the PBCG risk calculator.

Patch-based deep learning approaches using convolutional neural networks have been used to detect and characterise prostate cancer;[17,18] however, these studies did not make use of important information in the peritumoural region, which has been shown to improve disease characterisation and classification performance.[15,16,26] Algohary and colleagues[19] showed the benefit of extracting radiomic representations from peritumoural regions along with intratumoural regions. They highlighted the differences in concentration of epithelial cells and lymphocytes between low-risk and high-risk regions by mapping the representative peritumoural regions on biparametric MRI to the whole-mount pathology slides. In our study, we explored multiple input configurations of the network, with patches extracted at different scales and each subsequent scale adding information from the peritumoural region. The patch scale (S1) that included the peritumoural region up to 3 mm from the periphery of the lesion resulted in optimum performance. The use of binary lesion segmentation masks as an auxiliary input to the network aids in setting an attention region,[27] while at the same time providing context of the peritumoural region.

Although a few studies have shown the association of radiomic features with biochemical recurrence-free survival,[28,29] to our knowledge, none of these studies have evaluated the association of deep learning predictions and representations with biochemical recurrence-free survival. Shiradkar and colleagues[28] showed that a machine learning classifier trained on radiomic signatures can predict biochemical recurrence. Similarly, in a study by our group, Li and colleagues[29] constructed a radiomics-based nomogram, including GGG, PSA, and radiomics-based imaging biomarkers, using multivariable analysis to predict biochemical recurrence. Here, our findings showed that a clinical nomogram that was trained to identify clinically significant prostate cancer lesions on biparametric MRI could, as an additional feature, also predict biochemical recurrence-free survival, which might help to identify patients who would benefit from adjuvant therapy.

Several studies have explored MRI-based prostate cancer diagnosis and disease characterisation;[13,30] however, most of these used data from a single site or evaluated their approach on only a cross-validation set. In this study, we cross-validated our approach using data from three different sites and independently validated it using data collected from two different sites. We found that our model generalised well across the external sites.

Our study had some limitations. First, the ground-truth GGG assessment of some patients was obtained from biopsy reports (12-core systematic, magnetic resonance-guided, or ultrasound fusion-guided biopsies), whereas others were obtained from radical prostatectomy specimens. Although the ClaD clinical nomogram was trained with reference ground-truth from 12-core and magnetic resonance-guided or ultrasound fusion-guided biopsies, the weight decay regularisation in the Adam optimiser[31] and early stopping criteria while training the deep learning network made sure that the network did not over-fit to the biopsy labels. The ClaD nomogram resulted in an AUC of 0·83 (95% CI 0·75–0·91) with respect to radical prostatectomy specimens. Second, since experienced pathologists from each of the institutions separately graded the biopsy and radical prostatectomy specimens, we acknowledge the possibility of inter-reader variability. A central review of Gleason grading for all patients will be done in future work. Third, we used only PSA, lesion volume, and prostate volume as clinical variables because other clinical

information, such as free-to-total PSA ratio and digital rectal examination, were not available across the datasets. In our work, manual delineations of the prostate gland were obtained, which is time consuming. Automatic lesion detection and segmentation would be desirable. Fourth, inter-reader variability in manual annotation of the prostate gland and lesion could lead to differences in the estimation of prostate and lesion volume. Although a single experienced radiologist from each institution delineated the lesions, using multiple readers could have yielded a consensus annotation that might have been less prone to the sensitivity of annotations made by individual readers. Fifth, we were not able to access data on race for all of the datasets; therefore, in terms of participant demographics, we could only report on age and clinical variables. Finally, our integrated ClaD nomogram was constructed by integrating deep learning predictions, PI-RADS, and three clinical variables. Including PI-RADS-based scoring in the nomogram might require manual reading or interpretation of the MRI, which could lead to inter-observer variability. However, we showed that including PI-RADS in the nomogram led to a statistically significant 7% increase in the AUC compared with DIN, which does not include PI-RADS.

In summary, we constructed an integrated nomogram (ClaD) using a routinely available blood parameter (PSA), prostate volume, lesion volume, PI-RADS score, and deep learning predictions from biparametric MRI. Our multicentre results suggest that the ClaD nomogram could be used as a non-invasive computer-aided diagnostic tool to triage patients into very low, low, and intermediate favourable risk categories for prostate cancer, thus helping to avoid unnecessary biopsies and identifying patients who might be candidates for active surveillance, and also identify high-risk patients with clinically significant prostate cancer, who are likely to have disease recurrence after definitive therapy and might benefit from adjuvant therapy.

Contributors

AH oversaw the study and software. AMad, AH, and RS conceived and designed the study. AMah, ARR, SHT, and AP collected clinical data and were responsible for manual annotations. AH, PF, and AMad did the statistical analysis and interpretation. AMad, AH, and RS critically revised the manuscript for important intellectual content. AMad, AT, and LP supervised the study and managed resources. AMad and RS handled funding support. AMad, RS, and AH verified the data sets. All authors had full access to all the data in the study, approved the final manuscript, and had the final responsibility for the decision to submit for publication.

Declaration of interests

AH is a former employee of Philips Research. AMad reports previous scientific advisory board membership for Inspirata, AstraZeneca, Bristol Myers Squibb, and Merck; current advisory board membership for Aiforia; sponsored research agreements with Philips, AstraZeneca, and Bristol Myers Squibb; technology that has been licensed to Elucid Bioimaging; involvement in a US National Institutes of Health (NIH) U24 grant with PathCore (1U24CA199374-01); three different R01 grants with Inspirata (NIH 1R01CA202752-01A1, NIH 1 R01 CA216579-01A1, and NIH 1 R01 CA220581-01A1); and equity holder status in Elucid Bioimaging and Inspirata. AP reports a service contract with Profound Medical and Research support from The Radiological Society of North America Research and Education Foundation. RS, PF, AMah, ARR, AT, SHT, and LP declare no competing interests.

Data sharing

Requests to access the datasets from Cleveland Clinic, University Hospitals Cleveland Medical Center, Icahn School of Medicine at Mount Sinai, and Long Island Jewish Medical Center (used with permission for this study) should be made directly to the institutions via their data access request forms. The Cancer Imaging Archive data (PROSTATEx Challenge) are available online. We have included all the codes used for analyses following feature extraction, including those used for training the deep learning network and the integrated nomograms (ClaD, PIN, and DIN), on GitHub (https://github.com/amogh3892/Integrated-DL-PIRADS-and-clinical-nomogram-for-identifying-clinically-significant-prostate-cancer).

Acknowledgments

This work was supported by the National Cancer Institute of the US National Institutes of Health (1U24CA199374-01, R01CA202752-01A1, R01CA208236-01A1, R01 CA216579-01A1, R01 CA220581-01A1, 1U01 CA239055-01, 1U01CA248226-01 [AMad]); the National Institute for Biomedical Imaging and Bioengineering (1R43EB028736-01 [AMad]); the National Center for Research Resources (1 C06 RR12463-01 [AMad]); the US Department of Veterans Affairs Biomedical Laboratory Research and Development Service VA Merit Review Award (IBX004121A [AMad]); the US Department of Defense's Breast Cancer Research Program Breakthrough Level 1 Award (W81XWH-19-1-0668 [AMad]), Prostate Cancer Idea Development Award (W81XWH-15-1-0558 [AMad] and W81XWH-18-1-0524 [RS]), Lung Cancer Investigator-Initiated Translational Research Award (W81XWH-18-1-0440 [AMad]), and Peer Reviewed Cancer Research Program (W81XWH-16-1-0329 [AMad]); the National Institute of Diabetes and Digestive and Kidney Diseases (1K25 DK115904–01A1 [AMad]); the Ohio Third Frontier Technology Validation Fund (AMad); the Wallace H Coulter Foundation Program in the Department of Biomedical Engineering and The Clinical and Translational Science Award Program at Case Western Reserve University (AMad); the Dana Foundation David Mahoney Neuroimaging Program (AMad); and the Clinical and Translational Science Collaborative Cleveland Annual Pilot Award 2020 (UL1TR002548 [RS]). The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health, the US Department of Veterans Affairs, the US Department of Defense, or the US Government.

References

1 Boesen L, Nørgaard N, Løgager V, et al. Assessment of the diagnostic accuracy of biparametric magnetic resonance imaging for prostate cancer in biopsy-naive men: the Biparametric MRI for Detection Of prostate Cancer (BIDOC) study. *JAMA Netw Open* 2018; **1:** e180219.

2 Baruah SK, Das N, Baruah SJ, et al. Combining prostate-specific antigen parameters with Prostate Imaging Reporting and Data System score version 2.0 to improve its diagnostic accuracy. *World J Oncol* 2019; **10:** 218–25.

3 Sonn GA, Fan RE, Ghanouni P, et al. Prostate magnetic resonance imaging interpretation varies substantially across radiologists. *Eur Urol Focus* 2019; **5:** 592–99.

4 Mottet N, Bellmunt J, Bolla M, et al. EAU-ESTRO-SIOG guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent. *Eur Urol* 2017; **71:** 618–29.

5 Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *J Natl Cancer Inst* 1998; **90:** 766–71.

6 Memorial Sloan Kettering Cancer Center. Dynamic prostate cancer nomogram: coefficients. https://www.mskcc.org/nomograms/prostate/pre_op/coefficients (accessed July 8, 2020).

7 Zanelli E, Giannarini G, Cereser L, et al. Head-to-head comparison between multiparametric MRI, the partin tables, memorial sloan kettering cancer center nomogram, and CAPRA score in predicting extraprostatic cancer in patients undergoing radical prostatectomy. *J Magn Reson Imaging* 2019; **50:** 1604–13.

8 Weaver JK, Kim EH, Vetter JM, et al. Prostate magnetic resonance imaging provides limited incremental value over the Memorial Sloan Kettering Cancer Center preradical prostatectomy nomogram. *Urology* 2018; **113:** 119–28.