Integration of Clinical Identifications With Deep Transferrable Imaging Feature Representations Can Help Predict Prostate Cancer Aggressiveness and Outcome

Jie Bao

The first affiliated hospital of Soochow university

ying Hou

the first affiliated hospital of nanjing medical university

Rui Zhi

the first affiliated hospital of nanjing medical university

Ximing Wang

The First Affiliated Hospital of Nanjing Medical University

Hai-bin Shi

The First Affiliated Hospital of Nanjing Medical University

Chunhong Hu

The First Affiliated Hospital of Soochow University

yudong zhang (✉ njmu_zyd@163.com)

The first affiliated hospital of Nanjing Medicial University

Original Article

Keywords: Machine Learning, Deep Learning, Multiparametric Magnetic Resonance Imaging, Prostate Cancer, Gleason Score

Posted Date: February 11th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-180726/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Objective: To develop a generalizable machine learning platform, designated PI-Risk, which incorporates clinicians' prior identifications with deep transferrable imaging feature representations into predictive models for PCa Gleason grade.

Patients and Methods: A retrospective study included 1442 biopsy-naïve patients from two tertiary care medical centers between January 2014 and December 2019. We investigated an interpretable risk assessment model (PI-Risk) to predict risk stratification of PCa using . The performance of PI-Risk model was independently tested on 232 internal test datasets, and on 539 external validation datasets. Model performance was typically evaluated against a "ground truth" with imaging-histopathologic annotations using receiver operating characteristic (ROC). Detection rates such as true positive, true negative, false positive and false negative rate were reported using a confusion matrix analysis. The Cox model's performance was evaluated based on Harrell's concordance index (C-index), calibration curves and Kaplan–Meier survival analysis.

Results: The PI-Risk integrated with 10 risk factors is formed to accurate risk stratification. In multinomial regression analyzing model, predicted IntraT-Rad G0 and PI-RADS score 2 were two independent predictors of G0 stage. Predicted IntraT-Rad G1 (OR, 2.42; 95% CIs, 2.13–2.86, $p < 0.001$) and PSA 4-10 ng/ml were two independent predictors of G1 stage. PSA 10-20 ng/ml, Predicted PeriT-DLR-SqueezeNet G1 and Predicted IntraT-Rad G3 were three independent predictors of G2 stage. PI-RADS score 5 were the independent predictor of G3 stage. PI-RADS score 5 and PSA > 100 ng/ml were two independent predictors of G4 stage. Combined use of PSA 4-10 ng/ml, PI-RADS 1-2 and stacked IntraT-Rad G0-1 resulted in excellent NPV (94.1%) for CIS diseases; and combined use of PSA >100 ng/ml and PI-RADS 5 resulted in high PPV (79.8%) for high-risk PCa. In follow-up, patients stratified by PI-Risk showed significantly different biochemical recurrence rate after surgery.

Conclusions: We concluded that the PI-Risk can offer a noninvasive alternative tool to stratify PCa aggressiveness. This enables a step towards PCa risk stratification.

Introduction

Men diagnosed with same prostate cancer (PCa), however, often reveal significant heterogeneity in clinical outcomes. In general, low-risk PCa is mostly indolent that would never progress or cause harm to the patient if left untreated [1, 2]. Similarly, intermediate-risk PCa has lower biochemical recurrence (BCR) rates and significantly better survival rates than high-risk PCa [3]. Gleason score is currently the best prognostic factor of PCa that are clinically used to determine PCa aggressiveness and treatment planning. However, owing to random sampling, biopsies might misestimate the Gleason score compared to that do radical prostatectomy (RP) [4, 5]. Additionally, the widespread use of prostate-specific antigen (PSA) screening and the introduction of reduced PSA thresholds for biopsy have contributed to a significant increase in unnecessary biopsies in men who do not have PCa [6, 7]. Therefore, to develop a noninvasive tool for accurate risk stratification of biopsy-naïve patients would have a significant impact on clinical decision making, treatment planning and prediction of outcomes for patients and spare them from painful biopsies and their accompanying risk of complications.

Multiparametric magnetic resonance imaging (mpMRI), a clinically used tool for PCa detection, can also provide additional information for PCa aggressiveness and prognostications from its high-resolution characterization of tumorous heterogeneity and cellularity [8, 9]. Previous studies have indicated that PCa with progressive pathological characteristics was associated with significantly different gray-level imaging patterns on T2-weighted imaging (T2WI) and diffusion-weighted imaging (DWI) [10, 11]. Several studies also revealed that machine learning or deep learning on

high-dimensional image features can provide improved diagnostic and predictive accuracy for localized PCa [12-14]. However, the increased dimensionality relative to limited cohort sizes in clinical settings, as well as the inherently complex networks of internal correlations between the measured tumor types and images present unique computational challenges [15]. Additionally, there are numerous multimodal feature representations such as clinical and demographic indicators, and large-scale imaging identifications generated in varying clinical settings. A clinical tool should leverage the integration and interactions of multimodalities for risk stratification. However, there are challenges associated for such analyses. First, effective approaches for integrating these multimodality data are lacking, especially in the context of gray-level imaging features. Second, although high-throughput deep networks have matured to a point that enables detailed discoveries of diseases in task-specific settings, the limited cohort size and high dimensionality of data increase the possibility of false-positive discoveries and overfitting. Deep generative approach, like deep transferable learning, can translate complex and high-dimensional images into relevant computational feature representations. Third, as every algorithm has its strength and weakness, there is no single algorithm that works best for every problem. Stacked-ensemble learning can make use of meta-algorithms to learn predictions of various algorithms and builds a stacked-ensemble model with them, which increases prediction accuracy and reduces the false positive rate of the base model predictions [16].

In this study, we introduce a generalizable auto machine learning (ML) platform, designated PI-Risk, which incorporates clinical identifications with high-dimensional images into predictive models for PCa Gleason grade. The imaging phenotypes of observations are quantitatively characterized by radiomic descriptors on intratumorous region and by deep transferable learning feature representations on region around the tumor. We built end-to-end sparse model to integrate multimodal data representations for multi-task classification. Our study included 1,442 biopsy-naïve patients from two tertiary care medical centers, consisting of 671 datasets for model development and 232 patients for internal test and 539 patients for external validation.

Patients And Methods

Study cohort

This study was retrospective and approved by the local Institutional Review and need for written informed consent was waived. All included of consecutive patients who underwent prostate mpMRI at two tertiary care medical centers were reviewed. All procedures performed in studies involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments.

The primary patients comprised an evaluation of the two institutions database for medical records and were histologically proven between January 2014 and December 2019. The inclusion criteria were followed: (1) patients with biopsy or prostatectomy proved PCa; (2) standard prostate 3.0 T MRI performed within 4 weeks before the biopsy or prostatectomy; (3) with standard histologic tissue slices of dissected prostatectomy specimens. Patients were excluded if (1) absence of biopsy, surgical intervention or medical records within 8 weeks after MRI examination (n = 9554); (2) noncompliance with imaging quality or imaging exam from outside institutions (n = 141); (3) previous surgery, radiotherapy or drug therapies for PCa (interventions for benign prostatic hyperplasia or bladder outflow obstruction were deemed acceptable) (n = 436). Finally, 903 patients from center 1 and 539 patients from center 2 were eligible for clinical evaluation.

As a standard part of patient management in our two medical centers, the lesion with a Prostate Imaging Reporting and Data System (PI-RADS) [17] scored ≥ 3 underwent fusion or cognitive targeted MR-guided biopsy in conjunction with systemic biopsy by five urologists who had a prior experience of at least 1000 TRUS-guided prostate needle biopsies. Patients with PI-RADS 1-2 underwent TRUS-guided systemic biopsy. two high-experience uropathologists

reviewed the available histopathological slides according to the 2014 WHO/ISUP recommendation [18]. From histopathology, we primarily defined biopsy-benign, Gleason score 3+3, 3+4, 4+3 and $\geq 4+4$ as G0, G1, G2, G3 and G4 group, respectively. We secondly grouped G0-1 into clinically insignificant (CIS) disease, G2-3 into intermediate-risk PCa and G4 into high-risk PCa.

The PI-Risk model was primarily designed for a multi-task classification of G0, G1, G2, G3 and G4 diseases. We randomly split the data of center 1 into training ($n = 671$) and test ($n = 232$) group, respectively, for model development and internal test. We also used the data from center 2 with 539 patients for external validation. A flow diagram of patient selection with inclusion and exclusion criteria is showed in supplementary **Fig. S1**.

Follow-up

The first postoperative visit was 6 weeks later after RP and then patients were consistently followed-up at intervals of 3 to 6 months based on PSA. The time of a BCR was recorded. Patients were censored in case of emigration, or on 30th Jul 2020, whichever came first. The definition of BCR was referred to criteria previously reported [19, 20].

Prostate mpMRI

All imaging exams were performed on two 3.0T MRI scanners with pelvic phased-array coils (MAGNETOM Skyra; Siemens Healthcare, Munich, Germany) at the two institutes. The mpMRI consisted of T2WI in three panes, DWI with high b value of 1500 s/mm^2 and apparent diffusion coefficient (ADC) map in axial plane (supplementary data, **Table S1**).

Lesion Segmentation

Entire volume of interest (VOI) of lesion was segmented using an in-house software (Oncology Imaging Analysis v2; Shanghai Key Laboratory of MRI, ECNU, Shanghai, China) based on histopathologic-imaging matching by two dedicated radiologists (reader 1 and reader 2 with 3-yr and 5-yr experience of prostate imaging). The contours of VOIs were then rechecked in consensus with a board-certified radiologist (reader 3, with 15-yr experience of prostate imaging). In patients with RP ($n = 1006$), postsurgical ex vivo prostates were processed using a previously described protocol [21]. Key steps included sectioning, digitization, and annotation of cancer regions by highly experienced urological pathologists. The histopathologic specimens were then assembled into pseudo-whole-mount sections and coregistered to the MRI using a previously described registration method [21]. In this way, regions of annotated PCa were mapped onto the images to produce the ground truth maps. In total, histopathologic-imaging matched specimens were identified. In patients without RP ($n = 436$), all subjects underwent MRI/TRUS-fusion targeted biopsy followed by 11-gauge core systemic needle biopsy. A central challenge in image labeling is the presence of ambiguous regions, where the true tumor boundary cannot be deduced from the image, and thus multiple equally plausible interpretations exist. To fill this gap, the VOI of each lesion was drawn twice by each of two independent radiologists. Regional identification overlapping in two instances was identified as the authorized VOI of the targeted lesion. Because it is inaccessible to achieve an imaging correlation with whole-mount prostatectomy specimens in our retrospective data, the unit of assessment in this study was per-patient. When patients had multiple lesions, only the index lesion with the largest lesion size and/or Gleason score was assessed.

Development, performance, and validation of predictive models

Volumetric radiomics features were analyzed from the target lesions using an open-source Python package Pyradiomics [22]. Image normalization was performed using a method that remapped the histogram to fit within $\mu \pm 3\sigma$ (μ : mean gray-level within the VOI and σ : gray-level standard deviation). A total of 2,553 radiomic features such as intensity, shape, texture, and wavelets, were computed from the target volume on T2WI, b-value of 1500 s/mm² DWI and ADC images that provide rich descriptions on the heterogeneity of entire-volumetric intratumor regions (IntraT-Rad).

The IntraT-Rad features focus on the inner regions of PCa. We further investigated a tumor-related region around the target lesion using novel deep transferable learning feature representations (PeriT-DLR). PeriT-DLR features were directly measured on MRI data using an image embedding toolbox (<https://github.com/biolab>) through five pre-trained deep neural networks, i.e. DeepLoc, Inception v3, SqueezeNet, VGG-16 and VGG-19 as embedders [23]. In order to obtain the representative imaging features of the target lesion, we used hand-cropped VOI as an attention to gate each embedder for analyzing PeriT-DLRs (i.e., regions around the PCa) in the center slice of an MRI scan. Each embedder calculates a feature vector for each image and returns an enhanced image descriptor. For image embedding, we used the penultimate layer of embedders to produce new image profiles, serving as another set of imaging feature representations (PeriT-DLRs) in parallel to IntraT-Rad for PCa. The detailed parameters of each embedder are summarized in a supplementary data (**Table S2**).

Reducing the feature space dimension aims to select informative characteristics, reduce the risk of bias and potential overfitting. To obtain the quantitative imaging hallmarks, we first assessed multi-scale imaging profiles including 2553 vectors from Pyradiomics, 6144 vectors from Inception v3, 3000 vectors from SqueezeNet, 12288 vectors from VGG16, 12288 vectors from VGG19 and 1536 vectors from DeepLoc, respectively, using the mean decrease Gini index (MDGI) calculated by a Random Forest algorithm. The MDGI represents the importance of individual features for correctly classifying a residue into linker and non-linker regions. The MDGI was calculated by classifying 200 randomly selected linker features and 200 non-linker features, and the mean MDGI was calculated as the averaged MDGI over 100 trials. The mean MDGI z-score of each feature was calculated as: $\frac{MDGI_i - \mu}{\sigma}$, where $MDGI_i$ is the individual MDGI of the feature dedicated; and μ and σ is the mean and standard deviation of all MDGIs, respectively. Vector elements with MDGI z-score larger than 2.0 were selected as optimum feature candidates. Next, MDGI-selected features from each embedder were analyzed using an auto stacked-ensemble ML based on an open-source auto ML platform (<https://github.com/aws-labs/autogluon>). The first layer of our auto ML framework has 5 base learners such as a k-nearest neighbors (kNN), AdaBoost, Random Forests, Logistic Regression (LR) and a Support Vector Machine (SVM), whose outputs are concatenated and then fed into the next layer, which itself consists of multiple stacker models. These stackers then act as base models to an additional layer. It merely employs random search for hyperparameter tuning, model selection, ensembling, feature engineering, data preprocessing, data splitting, etc., thus offers us an enticing alternative to deploy high-performance stack-ensemble models. We performed a random search over the parameter configuration, and chose the optimal parameters with the best score based on the evaluation of log-loss of ML model on 5-fold cross-validation datasets. The outputs calculated from ML predictor indicated the relative risk that the patient had G0, G1, G2, G3 or G4 disease. By this way, six quantitative imaging hallmarks, i.e., PeriT-DLR-DeepLoc, PeriT-DLR-SqueezeNet, PeriT-DLR-Inception v3, PeriT-DLR-VGG16, PeriT-DLR-VGG19 and IntraT-Rad, were developed from mpMRI, respectively, for decoding the heterogeneity of PCa.

In order to evaluate synergistic effect of multimodal features for the prediction of Gleason grade, the obtained 6 new imaging signatures were integrated with 4 clinical variables such as patient age (≤ 60 yrs, > 60 yrs), PSA level (4-10 ng/ml, 10-20 ng/ml, 20-100 ng/ml and > 100 ng/ml), location of observation (peripheral zone [PZ], transition zone [TZ]) and a PI-RADS score from radiologists' reports. An interpretable risk assessment model (PI-Risk) was finally developed using a multinomial LR with elastic net penalty. The PI-Risk model is based on proportionally converting each

regression coefficient in multivariate logistic regression to a 0- to 100-point scale. The effect of the variable with the highest β coefficient (absolute value) is assigned 100 points. The points are added across independent variables to derive total points, which are converted to predicted probabilities (Pi). The performance of PI-Risk model was independently tested on 232 internal test datasets, and on 539 external validation datasets. The entire flowchart of auto ML analysis for the PI-Risk model development is showed in **Fig. 1**.

Predictors of clinical outcome

Additionally, we prospectively evaluated a Cox model in using 5 clinic-imaging risk factors including dedicated age, PSA, PI-RADS score and predicted PI-Risk to assess the incremental aspect of our imaging signatures for predicting biochemical recurrence (BCR) of PCa after RP in 462 PCa patients who underwent RP treatment.

Statistical Analysis

By using biopsy and/or prostatectomy specimens as reference standard, the extents of lesions were divided into G0, G1, G2, G3 and G4 group. Quantitative variables were expressed as mean $\pm$ standard deviation (mean $\pm$ SD) or median and range or median and range, as appropriate. Model performance was typically evaluated against a “ground truth” with imaging-histopathologic annotations using receiver operating characteristic (ROC). Detection rates such as true positive, true negative, false positive and false negative rate were reported using a confusion matrix analysis. The Cox model's performance was evaluated based on Harrell's concordance index (C-index), calibration curves and Kaplan–Meier survival analysis. All the statistics were two-sided, and a p-value less than 0.05 was considered statistically significant. All statistical analyses were performed using MedCalc software (V.15.2; 2011 MedCalc Software bvba, Mariakerke, Belgium) and R software package (V.4.0.2; <https://www.r-project.org>).

Results

Baseline characteristics

Out of all patients included, PCa was diagnosed in explanted tissue of 557/671 patients (83.0%) in training group, 189/232 (81.5%) in internal test group and 360/539 (66.8%) in external validation group. The demographic/clinical factors included the age, PSA level, lesion location, measured diameter and PI-RADS score. Histopathological factors included Gleason score, number of positive cores, perineural invasion positive core and pathological T and N stage at biopsy and/or RP specimens. Detailed baseline characters of the patients are summarized in Table 1.

Table 1. The baseline characteristics of training cohort, internal and external cohort.									
Variable	Training cohort (center 1, n = 671)			Internal cohort (center 1, n = 232)			External cohort (center 2, n = 539)		
	PCa	Benign	<i>P</i>	PCa	Benign	<i>P</i>	PCa	Benign	<i>P</i>
No. of subjects	557	114		189	43		360	179	
Age (y), median (range)	70 (65, 75)	65 (61, 72)	<0.001	70 (65,74)	67 (61, 74)	0.08	71 (66, 76)	66 (60, 71)	<0.001
< 60 y, n (%)	47/557 (8.4%)	25/114 (21.9%)		17/189 (9.0%)	9/43 (20.9%)		23/360 (6.4%)	43/179 (24.0%)	
≥ 60 y, n (%)	508/557 (91.6%)	89/114 (78.1%)		172/189 (91.0%)	34/43 (79.1%)		337/360 (93.6%)	136/179 (76.0%)	
PSA level, median (range)	17.9 (6.6, 36.9)	8.1 (5.9,11.8)	0.001	14.5 (9.6, 31)	9.53 (6.1,17.3)	0.018	15.7 (8.5, 34.4)	8.93 (6.2, 13.8)	<0.001
4-10 ng/ml, n (%)	168/557 (30.2%)	66/114 (57.9%)		58/189 (30.7%)	19/43 (44.2%)		96/360 (26.7%)	101/179 (56.4%)	
10-20 ng/ml, n (%)	174/557 (31.2%)	35/114 (30.7%)		65/189 (34.4%)	16/43 (37.2%)		95/360 (26.4%)	55/179 (30.7%)	
20-100 ng/ml, n (%)	206/557 (37.0%)	13/114 (11.4%)		57/189 (30.2%)	7/43 (16.3%)		126/360 (35.0%)	22/179 (12.3%)	
> 100 ng/ml, n (%)	9/557 (1.6%)	0/114 (0)		9/189 (4.8%)	1/43 (2.3%)		43/360 (11.9%)	1/179 (0.6%)	
D-max (cm), median (range)	1.7 (1, 2.5)	1.1 (0.8, 1.4)	<0.001	1.4 (0.9, 2.2)	1.0 (0.7, 1.6)	0.03	2 (1.5, 3)	1.6 (1.2, 2)	<0.001
Prostate Zone, n	557	114	<0.001	189	43	<0.001	360	179	
PZ, n (%)	362/557 (65.0%)	41/114 (36.0%)		137/189 (72.5%)	17/43 (39.5%)		243/360 (67.5%)	76/179 (42.5%)	
TZ/ AFMS/CZ, n (%)	195/557 (35.0%)	73/114 (64.0%)		52/189 (27.5%)	26/43 (60.5%)		117/360 (32.5%)	103/179 (57.5%)	
MRI index lesion per patient, n (%)	557	114	<0.001	189	43	<0.001	360	179	0.044
PI-RADS 1-2	28/557	80/114		10/189	22/43		35/360	95/179	

	(5%)	(70.2%)		(5.3%)	(51.2%)		(9.7%)	(53.1%)	
PI-RADS 3	88/557	28/114		38/189	18/43		28/360	44/179	
	(15.8%)	(24.6%)		(20.1%)	(41.9%)		(7.8%)	(24.6%)	
PI-RADS 4	189/557	4/114		61/189	2/43		73/360	19/179	
	(33.9%)	(3.5%)		(32.3%)	(4.7%)		(20.3%)	(10.6%)	
PI-RADS 5	252/557	2/114		80/189	1/43		224/360	21/179	
	(45.2%)	(1.8%)		(42.3%)	(2.3%)		(62.2%)	(11.7%)	
Biopsy GG, n (%)	557	114	<0.001	189	43	<0.001	360	179	0.044
Negative		114			43		3/360	179	
							(0.8%)		
GG 1 (3+3)	150/557			41/189			36/360		
	(26.9%)			(21.7%)			(10%)		
GG 2 (3+4)	122/557			37/189			81/360		
	(21.9%)			(19.6%)			(22.5%)		
GG 3 (4+3)	142/557			49/189			84/360		
	(25.5%)			(25.9%)			(23.3%)		
GG 4 (8)	126/557			55/189			68/360		
	(22.6%)			(29.1%)			(18.9%)		
GG 5 (5+5)	17/557			7/189			88/360		
	(3.1%)			(3.7%)			(24.4%)		
Surgical GG, n (%)	557			189			260		
GG 1 (3+3)	81/557			26/189			18/260		
	(14.5%)			(13.8%)			(6.9%)		
GG 2 (3+4)	146/557			58/189			60/260		
	(26.2%)			(30.7%)			(23.1%)		
GG 3 (4+3)	178/557			51/189			79/260		
	(32.0%)			(27.0%)			(30.4%)		
GG 4 (8)	113/557			40/189			31/260		
	(20.3%)			(21.2%)			(11.9%)		
GG 5 (5+5)	39/557			14/189			72/260		
	(7.0%)			(7.4%)			(27.7%)		
ECE, n (%)	557			189			260		
+	141			50/189			113/260		

	(25.3%)	(26.5%)	(43.5%)
-	416 (74.7%)	139/189 (73.5%)	147/260 (56.5%)
SVI, n (%)	557	189	260
+	90 (16.2%)	36/189 (19.0%)	45/260 (17.3%)
-	467 (83.8%)	153/189 (81.0%)	215/260 (82.7%)
LNI, n (%)	355	118	150
+	49/355 (13.8%)	21/118 (17.8%)	10/150 (6.7%)
-	306/355 (86.2%)	97/118 (82.2%)	140/150 (93.3%)
PCa = prostate cancer; PZ = peripheral zone; TZ = transition zone; CZ = center zone; AFMS = anterior fibromuscular stroma; GG = Gleason grade; ECE = extracapsular extension; SVI = seminal vesicle infiltration; LNI = lymph node invasion			

Imaging feature analysis

Imaging features analyzed from mpMRI data were selected and integrated into 6 quantitative signature scores with the image embedders abovementioned. Using the criteria of MDGI z-score > 2.0, the step-wise random forest analyzes selected 34/1,553 (2.2%) features from DeepLoc, 287/6,144 (4.7%) features from Inception v3, 114/3,000 (3.8%) features from SqueezeNet, 406/12,288 (3.3%) features from VGG16, 378/12,288 (3.1%) features from VGG19 embedder and 102/2,553 (4.0%) features from Pyradiomics. Results achieved for feature ranking and feature selection are summarized in supplementary **Fig. S2**. Performance of selected features achieved for the base models is shown in supplementary **Fig. S3**. As the results, the individual importance of each dedicative model is determined (supplementary **Fig. S4**).

Development, performance, and validation of PI-Risk model

The PI-Risk integrated with 10 risk factors is formed as interpretable nomograms (Fig. 2). In multinomial regression analyzing model, predicted IntraT-Rad G0 (odds ratio [OR], 3.01; 95% confidence intervals [CIs], 2.74–3.34, $p < 0.001$) and PI-RADS score 2 (OR, 2.88; 95% CIs, 2.47–3.17, $p = 0.002$) were two independent predictors of G0 stage. Predicted IntraT-Rad G1 (OR, 2.42; 95% CIs, 2.13–2.86, $p < 0.001$) and PSA 4–10 ng/ml (OR, 1.29; 95% CIs, 1.06–1.43, $p = 0.037$) were two independent predictors of G1 stage. PSA 10–20 ng/ml (OR, 1.52; 95% CIs, 1.32–1.67, $p = 0.005$), Predicted PeriT-DLR-SqueezeNet G1 (OR, 1.31; 95% CIs, 1.14–1.52, $p = 0.009$) and Predicted IntraT-Rad G3 (OR, 1.29; 95% CIs, 1.07–1.44, $p = 0.011$) were three independent predictors of G2 stage. PI-RADS score 5 (OR, 1.84; 95% CIs, 1.74–2.31, $p < 0.001$) were the independent predictor of G3 stage. PI-RADS score 5 (OR, 2.84; 95% CIs, 2.62–3.17, $p < 0.001$) and PSA > 100 ng/ml (OR, 1.69; 95% CIs, 1.36–1.83, $p = 0.007$) were two independent predictors of G4 stage. The discrimination ability of classification model for PCa Gleason grade was summarized with confusion matrix, AUC, accuracy, F1, precision and Recall in training, testing and external validation, respectively (Fig. 3).

As part of this study, we considered predictive aspects of abridgedly combined use of independent factors at PI-Risk model for CIS disease, intermediate-risk PCa and high-risk PCa. As patients with CIS disease are often recommended for active surveillance, patients with intermediate-risk PCa are the candidates for RP treatment, while high-risk PCa are often implicated with adverse clinical outcomes. As shown in Fig. 4, the negative predictive value (NPV) and positive predictive value (PPV) of individual and combined factors for CIS, intermediate-risk PCa and high-risk PCa were plotted. Combined use of PSA 4–10 ng/ml, PI-RADS 1–2 and stacked IntraT-Rad G0-1 resulted in excellent NPV (94.1%) for CIS diseases; and combined use of PSA > 100 ng/ml and PI-RADS 5 resulted in high PPV (79.8%) for high-risk PCa.

Prognostic evaluation of PI-Risk

As of Jul 2020, we collected a cohort of 462 PCa patients who had completed 3-yr BCR follow-up after surgery. The median BCR-free survival of the patients was 40.7 (range, 37.7–40.7) months. And the multivariate Cox analysis model shows that, among the 5 pretreatment risk factors (age, PSA, PCa location, PI-RADS score and PI-Risk score), PSA ≥ 20 ng/ml (OR, 1.58; 95% CI, 1.20–2.08; $p = 0.001$) and PI-Risk $\geq G3$ (OR, 1.45; 95% CI, 1.12–1.88; $p = 0.005$) were the independent predictors of BCR. The resulting Cox model produces a C-index of 0.76 (95% CI, 0.73–0.79) for predicting 3-yr BCR. The Kaplan-Meier survival curve analysis of BCR according to the PSA and PI-Risk is shown in Fig. 5.

Discussion

Gleason score is the determining factor of treating planning of PCa and postoperative survival prediction [24]. In this study, we proposed a collaborative framework that enables integration of clinicians' prior knowledges and deep transferable image feature representations into an interpretable PI-Risk tool to improve the predictions of Gleason score. This integrated approach to data analysis can be generalized under the 'task-free image embedding with privileged deep networks' paradigm described by Zupan et al [23]. This study contributes important methodology accompanied with model interpretability to address a critical clinical question for PCa risk stratification. Our results on a large cohort of 1,442 biopsy-naïve patients from two tertiary care medical centers show promises of PI-Risk model and potential utilities of this strategy for studying similar clinical questions. Additionally, results on the follow-up of BCR show favorable prognostic aspect of PI-Risk for disease progression risk stratification.

Deep image embedding by 5 pretrained models is a core to our study. Feature learning with problem-specific algorithms is implicit, however, training a deep network usually requires large number of images, which limits its utility. Deep image embedding does not need training on a closely related set of images. A pretrained deep model on a sufficiently large number of diverse images may infer useful features from a broad range of new image sets. This idea was proposed by Zupan et al [23], who explored a democratized image analytic tool box by integrating deep learning embedding. In our experiments, we used the dedicative embedders to build 6 new imaging hallmarks, providing an additional information for stratifying patients into groups with G0 to G4 risks. Even without incorporating clinical indicators, the new hallmarks can determine G0, G1, G2, G3 and G4 disease with accuracy of 0.806-0.918, 0.718-0.789, 0.591-0.680, 0.604-0.655 and 0.775-0.792, respectively. An auto ML platform using stack-ensemble is another core to our method. Different from prior approaches that focused on the task of combined algorithm selection and hyperparameter optimization, our approach performs advanced data processing, deep learning and model ensembling. It allows to automatically recognize the data type in each column for robust data preprocessing, including special handling of high-dimensional imaging datasets; In particular, duo to the ability to employ multi-layer stack ensembling that combines the aggregated predictions of the base models as its features, the stacker model can improve upon shortcomings of the individual base predictions and exploit interactions between base models that offer enhanced predictive power.

For medical data, the potential phenotype information conveyed in images is more complex than simple variables, and it is also delicate and thus needs to be analyzed more carefully. Results of previous studies have revealed the critical role of clinical identifications, such as PSA and PI-RADS score, on stratifying patients of PCa [24-27]. Importantly, our findings can enhance the clinical indications of these dedicative clinical factors. The PI-Risk model, based on the data-driven approach, allows for the dedicative features only to be incorporated when an improvement in the model is observed. Functionally, this reduces the regularization of the features consistent with clinicians' prior identifications, resulting in the development of sparse models that prioritize features in line with previous studies. This not only increases the predictive performance, but also facilitates clinical interpretation and translation of the results. By this strategy, we did determine the significant imaging predictors, which were combined with prior clinical identifications for improving risk stratification in biopsy-naïve patients. Importantly, abridgedly combined use of these factors, like incorporating IntraT-Rad with PI-RADS for CIS diseases and incorporating PSA with PI-RADS for high-risk PCa, have valuable clinical implications for treating planning. As a pre-treatment identification of CIS can help spare unnecessary invasive biopsies from these patients. And a pre-treatment identification of high-risk PCa can help select the candidates for neoadjuvant therapy before RP treatment, thus reduce the risk of post-operative recurrence and long-term mortality.

PSA recurrence is currently the strongest clinical end point of PCa, driving almost all initial disease management decisions after primary treatment [24]. It had correctly demonstrated that patient with high-risk PCa had significantly worse BCR-free survival compared with low- to intermediate-risk groups [28, 29]. Results from some studies indicated that imaging findings such as PI-RADS and radiomics features had prognostic value on BCR after prostatectomy [30, 20, 31]. Our preliminary results also showcase the PI-Risk stratification even revealed a potential role in predicting the prognostic of disease progression risk preoperatively. We found that PI-Risk > G3, PSA level ≥ 20 ng/ml were significantly associated with the worse BCR-free survival, implying the prognostic relevance of our PI-Risk assessment on short and long-term management of patients. Combined use of PI-Risk and PSA can result in a C-index of 0.76 for predicting 3-yr BCR in our primary cohort. After all, this was a preliminary result in a small population, the prognostic aspects of PI-Risk on external independent cohorts warrant further validations.

There are several limitations of our research. Firstly, although the data of this study originated from two medical centers with internal and external validation, the cohort size was still limited for our data-driven approach which is expected larger data sets. A larger studied population will be needed to optimize the performance of the model. Secondly, part of our external data used MRI-guided biopsy as the reference standard, even targeted prostate biopsy was identified as a reliable method for PCa detection, the accuracy of which might be impacted by technical variations in the features or operation of equipment [32, 33]. Third, currently, the deep transfer learning only used the center slice instead of the 3D full tumor volume, so the effect comparison between IntraT-Rad features and PeriT-DLR features may not be comprehensive. The center slices have been shown very close performance to using the 3D volume in many cancer imaging-based studies, thus our results on the PeriT-DLR features are still informative. In our next-step research, we will implement 3D-based deep learning approach by leveraging more powerful computational resources.

Conclusions

In summary, we proposed an interpretable tool for PCa aggressiveness assessment. We provided a robust auto-ML framework for integrating multimodality data in relation to PCa aggressiveness. The interpretability of PI-Risk is particularly imperative towards building trustable auto-ML tools for clinical applications. Our study on two cohorts showed that PI-Risk may serve as a great alternative to enhance biopsy-naïve patients' stratification and prognostication. Further evaluation of our methods on a multi-center setting is needed and a goal of our future work.

Abbreviations

PCa: prostate cancer

ROC: receiver operating characteristic

BCR: biochemical recurrence

RP: radical prostatectomy

PSA: prostate-specific antigen

mpMRI: Multiparametric magnetic resonance imaging

T2WI: T2-weighted imaging

DWI: diffusion-weighted imaging

ML: machine learning

PI-RADS: Prostate Imaging Reporting and Data System

CIS : clinically insignificant

VOI: volume of interest

kNN: k-nearest neighbors

LR: Logistic Regression

SVM: Support Vector Machine

PZ: peripheral zone

TZ : transition zone

ROC: receiver operating characteristic

Declarations

Ethics Committee approval was granted by the local institutional ethics review board, and the requirement of written informed consent was waived.

Conflict of Interest statement: The authors declare that they have no Conflict of Interests.

Funding

1. Contract grant sponsor: Key research and development program of Jiangsu Province; contract grant number: BE2017756 (to Y.D.Z.)
2. Contract grant sponsor: Suzhou Science and Technology Bureau-Science and Technology Demonstration Project; contract grant number: SS201808 (to X.M.W)
3. Contract grant sponsor: National Key R&D Program of China; contract grant number: 2017YFC0114300 (H.C.H.)

Conflict of Interest statement

The authors declare that they have no Conflict of Interests.

Availability of data and material:

The imaging studies and clinical data used for algorithm development are not publicly available, because they contain private patient health information. Interested users may request access to these data, where institutional approvals along with signed data use agreements and/or material transfer agreements may be needed/negotiated. Derived result data supporting the findings of this study are available upon reasonable requests.

Acknowledgements:

The authors thank all those who helped us during the writing of this research. We also thank the department of Ultrasound, Urology and Pathology of the two hospitals for their valuable help and feedback.

Author contributions:

Y.Z. and C.H. conceived, designed and supervised the project; J.B., X.W., R.Z., Y.Z. H.S. and Y.H. collected and pre-processed all data and performed the research; J.B., Y.H. and Y.Z. performed imaging data annotation and clinical data review; Y.Z. proposed the model; Y.Z. and J.B. drafted the paper; all authors reviewed, edited and approved the final version of article.

Ethics approval and Consent to participate

Ethics Committee approval was granted by the local institutional ethics review board, and the requirement of written informed consent was waived. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards

Consent for publication: Not applicable

References

1. Cornford P, Bellmunt J, Bolla M, Briers E, De Santis M, Gross T et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer. *European urology*. 2017;71(4):630-42. doi:10.1016/j.eururo.2016.08.002.
2. Mottet N, Bellmunt J, Bolla M, Briers E, Cumberbatch MG, De Santis M et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *European urology*. 2017;71(4):618-29. doi:10.1016/j.eururo.2016.08.003.
3. Verhoef EI, Kweldam CF, Kümmerlin IP, Nieboer D, Bangma CH, Incrocci L et al. Characteristics and outcome of prostate cancer patients with overall biopsy Gleason score 3 + 4 = 7 and highest Gleason score 3 + 4 = 7 or > 3 + 4 = 7. *Histopathology*. 2018;72(5):760-5. doi:10.1111/his.13427.
4. Porten SP, Whitson JM, Cowan JE, Cooperberg MR, Shinohara K, Perez N et al. Changes in prostate cancer grade on serial biopsy in men undergoing active surveillance. *J Clin Oncol*. 2011;29(20):2795-800. doi:10.1200/JCO.2010.33.0134.
5. Epstein JI, Feng Z, Trock BJ, Pierorazio PM. Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified Gleason grading system and factoring in

- tertiary grades. *Eur Urol.* 2012;61(5):1019-24. doi:10.1016/j.eururo.2012.01.050.
6. Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V et al. Screening and prostate-cancer mortality in a randomized European study. *The New England journal of medicine.* 2009;360(13):1320-8. doi:10.1056/NEJMoa0810084.
 7. Zhu X, Albertsen PC, Andriole GL, Roobol MJ, Schroder FH, Vickers AJ. Risk-based prostate cancer screening. *European urology.* 2012;61(4):652-61. doi:10.1016/j.eururo.2011.11.029.
 8. Ullrich T, Arsov C, Quentin M, Mones F, Westphalen AC, Mally D et al. Multiparametric magnetic resonance imaging can exclude prostate cancer progression in patients on active surveillance: a retrospective cohort study. *Eur Radiol.* 2020. doi:10.1007/s00330-020-06997-1.
 9. Gandaglia G, Ploussard G, Valerio M, Mattei A, Fiori C, Roumiguié M et al. The Key Combined Value of Multiparametric Magnetic Resonance Imaging, and Magnetic Resonance Imaging-targeted and Concomitant Systematic Biopsies for the Prediction of Adverse Pathological Features in Prostate Cancer Patients Undergoing Radical Prostatectomy. *Eur Urol.* 2020;77(6):733-41. doi:10.1016/j.eururo.2019.09.005.
 10. Zhang Y, Chen W, Yue X, Shen J, Gao C, Pang P et al. Development of a Novel, Multi-Parametric, MRI-Based Radiomic Nomogram for Differentiating Between Clinically Significant and Insignificant Prostate Cancer. *Frontiers in oncology.* 2020;10:888. doi:10.3389/fonc.2020.00888.
 11. Zhao K, Wang C, Hu J, Yang X, Wang H, Li F et al. Prostate cancer identification: quantitative analysis of T2-weighted MR images based on a back propagation artificial neural network model. *Science China Life sciences.* 2015;58(7):666-73. doi:10.1007/s11427-015-4876-6.
 12. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology.* 2020;21(2):233-41. doi:10.1016/s1470-2045(19)30739-9.
 13. Bonekamp D, Kohl S, Wiesenfarth M, Schelb P, Radtke JP, Gotz M et al. Radiomic Machine Learning for Characterization of Prostate Lesions with MRI: Comparison to ADC Values. *Radiology.* 2018;289(1):128-37. doi:10.1148/radiol.2018173064.
 14. Fehr D, Veeraraghavan H, Wibmer A, Gondo T, Matsumoto K, Vargas HA et al. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America.* 2015;112(46):E6265-73. doi:10.1073/pnas.1505935112.
 15. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB et al. Deep Learning in Medical Imaging: General Overview. *Korean J Radiol.* 2017;18(4):570-84. doi:10.3348/kjr.2017.18.4.570.
 16. Li F, Chen J, Ge Z, Wen Y, Yue Y, Hayashida M et al. Computational prediction and interpretation of both general and specific types of promoters in *Escherichia coli* by exploiting a stacked ensemble-learning framework. *Briefings in bioinformatics.* 2020. doi:10.1093/bib/bbaa049.
 17. Barrett T, Rajesh A, Rosenkrantz AB, Choyke PL, Turkbey B. PI-RADS version 2.1: one small step for prostate MRI. *Clinical radiology.* 2019.
 18. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PAJTAJoSP. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. 2015;40(2):244-52.
 19. Cookson MS, Aus G, Burnett AL, Canby-Hagino ED, D'Amico AV, Dmochowski RR et al. Variation in the definition of biochemical recurrence in patients treated for localized prostate cancer: the American Urological Association Prostate Guidelines for Localized Prostate Cancer Update Panel report and recommendations for a standard in the reporting of surgical outcomes. *The Journal of urology.* 2007;177(2):540-5. doi:10.1016/j.juro.2006.10.097.

20. Brockman JA, Alanee S, Vickers AJ, Scardino PT, Wood DP, Kibel AS et al. Nomogram Predicting Prostate Cancer-specific Mortality for Men with Biochemical Recurrence After Radical Prostatectomy. *European urology*. 2015;67(6):1160-7. doi:10.1016/j.eururo.2014.09.019.
21. Zhang YD, Wang Q, Wu CJ, Wang XN, Zhang J, Liu H et al. The histogram analysis of diffusion-weighted intravoxel incoherent motion (IVIM) imaging for differentiating the gleason grade of prostate cancer. *European radiology*. 2015;25(4):994-1004. doi:10.1007/s00330-014-3511-4.
22. Van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V et al. Computational radiomics system to decode the radiographic phenotype. *Cancer research*. 2017;77(21):e104-e7.
23. Godec P, Pancur M, Ilenic N, Copar A, Strazar M, Erjavec A et al. Democratized image analytics by visual programming through integration of deep models and small-scale machine learning. *Nat Commun*. 2019;10(1):4551. doi:10.1038/s41467-019-12397-x.
24. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. *Eur Urol*. 2016;69(3):428-35. doi:10.1016/j.eururo.2015.06.046.
25. Patel HD, Tosoian JJ, Carter HB, Epstein JI. Adverse Pathologic Findings for Men Electing Immediate Radical Prostatectomy: Defining a Favorable Intermediate-Risk Group. *JAMA Oncol*. 2018;4(1):89-92. doi:10.1001/jamaoncol.2017.1879.
26. Guazzoni G, Nava L, Lazzeri M, Scattoni V, Lughezzani G, Maccagnano C et al. Prostate-specific antigen (PSA) isoform p2PSA significantly improves the prediction of prostate cancer at initial extended prostate biopsies in patients with total PSA between 2.0 and 10 ng/ml: results of a prospective study in a clinical setting. *Eur Urol*. 2011;60(2):214-22. doi:10.1016/j.eururo.2011.03.052.
27. Alberts AR, Roobol MJ, Verbeek JFM, Schoots IG, Chiu PK, Osses DF et al. Prediction of High-grade Prostate Cancer Following Multiparametric Magnetic Resonance Imaging: Improving the Rotterdam European Randomized Study of Screening for Prostate Cancer Risk Calculators. *Eur Urol*. 2019;75(2):310-8. doi:10.1016/j.eururo.2018.07.031.
28. Hamada R, Nakashima J, Ohori M, Ohno Y, Komori O, Yoshioka K et al. Preoperative predictive factors and further risk stratification of biochemical recurrence in clinically localized high-risk prostate cancer. *International journal of clinical oncology*. 2016;21(3):595-600. doi:10.1007/s10147-015-0923-3.
29. Van den Broeck T, van den Bergh RCN, Arfi N, Gross T, Moris L, Briers E et al. Prognostic Value of Biochemical Recurrence Following Treatment with Curative Intent for Prostate Cancer: A Systematic Review. *Eur Urol*. 2019;75(6):967-87. doi:10.1016/j.eururo.2018.10.011.
30. Bourbonne V, Vallières M, Lucia F, Doucet L, Visvikis D, Tissot V et al. MRI-Derived Radiomics to Guide Post-operative Management for High-Risk Prostate Cancer. *Frontiers in oncology*. 2019;9:807. doi:10.3389/fonc.2019.00807.
31. Bourbonne V, Fournier G, Vallieres M, Lucia F, Doucet L, Tissot V et al. External Validation of an MRI-Derived Radiomics Model to Predict Biochemical Recurrence after Surgery for High-Risk Prostate Cancer. *Cancers (Basel)*. 2020;12(4). doi:10.3390/cancers12040814.
32. Siddiqui MM, Rais-Bahrami S, Truong H, Stamatakis L, Vourganti S, Nix J et al. Magnetic resonance imaging/ultrasound-fusion biopsy significantly upgrades prostate cancer versus systematic 12-core transrectal ultrasound biopsy. *Eur Urol*. 2013;64(5):713-9. doi:10.1016/j.eururo.2013.05.059.
33. Siddiqui MM, Rais-Bahrami S, Truong H, Stamatakis L, Vourganti S, Nix J et al. Magnetic resonance imaging/ultrasound-fusion biopsy significantly upgrades prostate cancer versus systematic 12-core transrectal ultrasound biopsy. 2013;64(5):713-9.

Figures

Figure 1

The integrated PI-Risk analysis pipeline. a, b, c, Clinical prior knowledge within the cohort of study, in response to each pathological status, is extracted by a panel of experts and encoded into a scaled clinical identification. d, e, f, Individual mpMRI data within the cohort of study are embedded with training-spared deep image embedders and Radiomics toolbox to extract various imaging features. This produces deep imaging feature measurements with auto learning algorithms, resulting in complex network representations of PCa pathological characteristics. g, h, This dataset is then fed into the elastic net algorithm for predictive modelling of the outcome of interest.

a

b

Interpretable nomogram (a) and estimated odds ratio (OR) with forest plots (b) for PI-Risk model. Only the significant ($p < 0.05$) factors at regression model are shown in the forest plots, shown is the mean, lower and upper of 95% confidence intervals of OR value.

Discrimination ability of PI-Risk classification model for PCa Gleason grade. a, b, c, Results of confusion matrix for G grade of PCa in training (a), test (b) and validation (c) data. d, Results of AUC, ACC, F1, Precision and Recall in the three data set group.

Figure 4

Clinical application of independent factors, i.e., PSA level, IntraT-Rad signature, PI-RADS score, as well as with the simplified combinations for determining clinically insignificant (CIS) disease, intermediate-risk PCa and high-risk PCa. For determining CIS (a), PSA 40-10 ng/ml, IntraT-Rad G0-1 and PI-RADS 1-2 were combined, which resulted in highest negative predictive value (yellow color marked). For determining high-risk PCa (b), PSA > 100 ng/ml and PI-RADS 5 were combined, resulting in highest positive predictive value (yellow color marked). This simple approach can produce additional clinical implications for better treatment decision making.

Figure 5

Kaplan-Meier survival curves of biochemical recurrence (BCR) according to the two independent prognostic factors in multivariate Cox model. a, BCR-free survival stratified by PI-Risk level. b, BCR-free survival stratified by PSA level.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- fig.s1.eps
- fig.s2.eps
- fig.s3.eps
- fig.s4.eps
- SupplementaryTables.docx