Multi-Multimodality Integrated Stack-Ensemble Learning for the Prediction of Gleason Grade and Prognostic Outcome in Prostate Cancer: A Proof-of-Concept Study

Jie Bao
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 Soochow University
Haibin Shi
The First Affiliated Hospital of Nanjing Medical University
Chunhong Hu
The First Affiliated Hospital of Soochow University
Yu-Dong Zhang (njmu_zyd@163.com)
The first affiliated hospital of Nanjing Medicial University

Research Article

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

Posted Date: May 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-512084/v1

Abstract

Purpose

To develop a generalizable model, namely PRISK, for the prediction of Gleason grade and prognostic outcome in prostate cancer (PCa) with multiple clinical factors and multiparametric (mp) MRI using stack-ensemble learning.

Methods

PRISK is developed to primarily assess PCa Gleason grade between benign (pG0), 3 + 3 (pG1), 3 + 4 (pG2), 4 + 3 (pG3) and ≥ 4 + 4 (pG4) and secondly predict the biochemical recurrence (BCR) after radical prostatectomy (RP). PRISK was developed with a stacked-ensemble learning of large-scale clinical identifications and mpMRI data in 671 training datasets, and was validated in 232 internal and 539 external datasets.

Results

The stacked-ensemble learning of mpMRI delivered a Radiomics-score and 5 transfer learning signatures from 5 deep transfer learning embedders. The PRISK, build with 10 clinical and imaging embedded predictors, achieved an area under the roc curve of 0.783, 0.798 and 0.762 in training, internal and external validation data for classifying Gleason grade, respectively. Specially, combined use of prostate-specific antigen (PSA), PI-RADS and Radiomics-score had excellent negative predictive value (94.1%) for clinical insignificant disease (pG0-1) and high positive predictive value (79.8%) for high-risk PCa (pG4). PSA ≥ 20 ng/ml (odds ratio [OR], 1.58; 95% confidence intervals [CIs], 1.20–2.08; p = 0.001) and PRISK ≥ G3 (OR, 1.45; 95% CI, 1.12–1.88; p = 0.005) were independent predictors of BCR, with a C-index of 0.76 (95% CI, 0.73–0.79) for predicting BCR by Cox analysis.

Conclusions

We concluded that the PRISK can offer a noninvasive alternative to stratify PCa Gleason grade. This enables a step towards PCa risk stratification.

Introduction

Patients diagnosed with prostate cancer (PCa) often reveal significant heterogeneity in clinical outcomes [1–3]. Gleason score is currently the best prognostic factor of PCa that are clinically used to determine PCa aggressiveness and therapeutic schedule. 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 approach for accurate risk stratification of biopsy-naïve patients would have a significant impact on clinical decision making, therapeutic schedule and prediction of outcomes for patients and spare them from painful biopsies and their accompanying risk of complications.

Although multiparametric magnetic resonance imaging (mpMRI) is a state-of-the-art tool for PCa detection and stage characterization [8, 9], there is no clear consensus on the specific imaging marker that is effective to stratify the PCa pathological grade. T2-weighted imaging (T2WI) and quantified diffusion-weighted imaging (DWI) have been identified

valuable for differentiating PCa Gleason grade. PCa lesions with higher Gleason grade showed significant difference in gray-level imaging features on high-resolution T2WI and apparent diffusion coefficient (ADC) from DWI compared to the lower grade diseases [10, 11]. However, studies differed in these specific imaging markers used to distinguish between the tumors. Recently, radiomics and deep learning has revolutionized the field of medical image. High throughput images with machine learning or deep learning have predominantly been used for PCa detection and classification, with achieved improvements in diagnostic and predictive accuracy [12–14]. Such data-derived analysis on prostate MRI would be an important advance, while remains a great challenge. Generally, training a deep network requires a large number of images, increasing dimensionality relative to limited cohort sizes and the inherently complex networks of internal correlations between the measured cancer types and images present unique computational challenges [15]. In addition, in real world clinical settings, there are numerous data representations from clinical, serum, demographic and multimodal imaging findings. Such analyses should leverage the integration of those multimodality data but face up to challenges.

Deep generative approach, a newly developed hybrid analysis of deep transfer learning and machine learning, can translate complex and high-dimensional images into relevant computational feature representations. Like the visual cortex can adapt to the analysis of many scenes and images, a deep network pre-trained on a sufficiently diverse images might infer useful features from a broad range of new image sets. A deep transfer learning network stores the prior knowledge obtained from one problem in a trained model and applies it to another problem, thus does not require training on a closely related set of images. Secondly, it allows to image embedding by extracting deep imaging feature representations with pre-trained deep models, and their relations with task-specific sets can be inferred by novel stacked-ensemble learning scheme. Stacked-ensemble learning makes 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].

Therefore, the purpose of this study was to develop and validate a generalizable risk assessment model, designated PRISK, by incorporating clinical identifications with high-dimensional mpMRI data for the prediction of PCa Gleason grade. The PRISK model was build based on a multi-class classification with weighted classifier selection and stacked-ensemble learning. The study included 1,442 biopsy-naïve patients from two tertiary care medical centers, comprising 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 Institutional Review Board (protocol 2019-SR-396) of the first affiliated hospital with Nanjing Medical University (NUH) and need for written informed consent was waived. All procedures performed in studies involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments.

The primary patients comprised of an evaluation of two collaborative centers' database (Center 1, NUH; and Center 2, the first affiliated hospital with Soochow University [SUH]) for medical records and were histologically proven between January 2014 and December 2019. The inclusion criteria were followed: (1) patients with biopsy or RP proved PCa; (2) standard prostate 3.0 T MRI performed within 4 weeks before the biopsy or RP; (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 systematic biopsy by five urologists who had a prior experience of more than 1,000 TRUS-guided prostate needle biopsies. Patients with PI-RADS 1–2 underwent TRUS-guided systematic 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 ≥ 4 + 4 as pG0, pG1, pG2, pG3 and pG4 group, respectively. We secondly grouped pG0-1 into clinically insignificant (CIS) disease, pG2 and pG3 into intermediate-risk PCa and pG4 into high-risk PCa.

The PRISK model was primarily designed for a multi-class classification of pG0, pG1, pG2, pG3 and pG4 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 after RP and then patients were consistently followed-up at intervals of 3 to 6 months based on PSA. The time of a biochemical recurrence (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 3.0 T MRI scanners with pelvic phased-array coils (MAGNETOM Skyra; Siemens Healthcare, Munich, Germany) at the two institutes. The mpMRI consisted of T2WI in three planes, DWI with high b value of 1500 s/mm^2 and 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 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, 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 histopathological 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. Finally, histopathologic-imaging matched specimens were identified in total 1006 patients. In patients without RP (n = 436), the reference standard for Gleason grade was based on the biopsy findings using MRI/TRUS-fusion targeted biopsy followed by 11-gauge core systematic 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 challenge to achieve a per-lesion imaging correlation with whole-mount prostatectomy specimens in 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 2.1.2 [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 radiomics (Rad) features (**Supporting data, S-text-1**) were computed from the target volume on T2WI, b-value of 1500 s/mm^2 DWI and ADC images that provide rich descriptions on the intratumor heterogeneity.

We further investigated the interaction between tumor and tumor-related regions using deep transfer learning analysis of mpMRI. Tumor-related region is a 5-mm extended region around the target lesion, which was determined using an erosion approach descripted in our previous study [23]. The transfer learning imaging features were measured through five pre-trained deep neural networks including DeepLoc, Inception v3, SqueenzeNet, VGG16 and VGG19, models of which were constructed delicately from a sufficiently large number of diverse images as described in [24]. Before we used the pre-trained models, the images were resampled as an inner-resolution of 0.5 × 0.5 mm^2 to remove the scale bias. Center slice of each PCa was extracted from each sequence and cropped into a patch with a size of 200 × 200. Each patch was normalized by Z-score to avoid the intensity scale. In the purest form of our transfer learning imaging feature analysis, it did not require training the model on a closely related set of images. The transfer learning embedder directly calculates a feature vector on the penultimate layer and returns an enhanced image descriptor, i.e., transfer learning imaging signature (TLIS), serving as another set of imaging feature representations in parallel to Rad signature for the next-step stacked-ensemble learning.

Reducing the feature space dimension aims to select informative characteristics, reduce the risk of bias and potential overfitting. The Pearson product-moment correlation coefficient (PCC) and false-discovery-rate (FDR) U-test was estimated between each pair of features, and random features were removed if the PCC was larger than 0.85 and FDR-test *p*-value was larger than 0.05. Then the retained features from Radiomics and 5 deep transfer learning embedders were assessed using mean decrease Gini index (MDGI) with a Random Forests (RFs) analysis. MDGI represents the importance of individual features for correctly classifying a residue into linker and non-linker factors. For the prevention of overfitting, only the top-ranked (top 0.1% by MDGI) factors in each imaging profile were selected as the linker factors. Next, the linker features selected from imaging embedders such as Radiomics, DeepLoc, SqueezeNet, Inception v3, VGG16 and VGG19 were integrated into six imaging signatures using multi-class classification with weighted classifier selection and stacked-ensemble learning. The first layer of our stacked ensemble learning framework has 11 base learners including AdaBoost, k-Nearest Neighbors (k-NNs), Support Vector Machine (SVM), Neural Networks (NNs), Naïve Bayes (NB), Logistic Regression (LR), RFs, Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGB), Extreme Gradient Boosting with Random Forest (XGB-RFs) and CatBoost, 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. 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 stacked model on 5-fold cross-validation datasets. The details of hyper-parameter set configurations of 11 base learners are summarized in supplementary data (**Table S2**). The outputs calculated from the stacked ensemble predictors indicated the relative risk that the patient has pG0, pG1, pG2, pG3 or pG4 diseases. Finally, six imaging embedded signatures, i.e., Rad, TLIS-DeepLoc, TLIS-SqueezeNet, TLIS-Inception v3, TLIS-VGG16 and TLIS-VGG19, were built for characterizing PCa Gleason grade with mpMRI.

In order to evaluate synergistic effects of multimodal integrations on the prediction of PCa Gleason grade, the 6 newly developed imaging signatures were combined 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 (PRISK) was developed using a multinomial LR with elastic net penalty. The PRISK model is based on proportionally converting each regression coefficient in multivariate LR 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 PRISK model was independently tested on 232 internal test datasets, and on 539 external validation datasets. The entire flowchart of the study design is showed in Fig. 1.

Predictors of prognostic outcome

Additionally, we prospectively evaluated a Cox proportional hazard regression model using 5 clinical and imaging factors including age, PSA, PI-RADS score and a PRISK score to assess the incremental aspect of our PRISK calculator on predicting BCR of PCa after RP in 462 PCa patients who underwent RP treatment.

Statistical Analysis

Quantitative variables were expressed as mean ± standard deviation (mean ± SD) or median and range or median and range, as appropriate. Model performance was typically evaluated against a "ground truth" with imaging-histopathological 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.

Imaging signatures

Using top 0.1% MDGI as criteria for feature selection, 25 of 2553 top-ranked features from radiomics, 16 of 1,553 features from DeepLoc, 61 of 6,144 features from Inception v3, 30 of 3,000 features from SqueeezeNet, 112 of 12,288 features from VGG16 and 112 of 12,288 features from VGG19, respectively, were selected for the construction of new imaging signatures (supplementary **Table S3**). The top-ranked feature candidates by RFs with MDGI are illustrated in supplementary **Fig. S2**. Their integrated effect on Gleason grade classification is plotted with a t-distributed stochastic neighbor embedding (t-SNE) analysis in supplementary **Fig. S3**. The distribution of the leading feature values in each imaging embedder between group of pG0, pG1, pG2, pG3 and pG4 are visualized in a violin plot (**Fig. S4**). Results of weighted classifier selection and the final stacked learner are summarized in **Table 2**.

Development, performance, and validation of PRISK

The PRISK, comprising of 4 clinical factors and 6 new imaging signatures, was formed as interpretable multi-class classification model using multinomial LR analysis (Fig. 2). Among all those clinical and imaging predictors, Rad-G0 (odds ratio [OR], 3.22; 95% confidence intervals [CIs], 2.76–3.45, $p < 0.001$) and PI-RADS score 2 (OR, 3.01; 95% CIs, 2.63–3.36, $p = 0.001$) were two independent predictors of pG0. Rad-G1 (OR, 2.42; 95% CIs, 2.17–2.87, $p < 0.001$) and PSA 4–10 ng/ml (OR, 1.18; 95% CIs, 1.01–1.39, $p = 0.036$) were two independent predictors of pG1. Rad-G3 (OR, 1.47; 95% CIs, 1.16–1.74, $p = 0.004$), PSA 10–20 ng/ml (OR, 1.44; 95% CIs, 1.21–1.73, $p = 0.004$) were two independent predictors of pG2. PI-RADS score 4 (OR, 1.84; 95% CIs, 1.73–2.34, $p < 0.001$) were the independent predictor of pG3. PI-RADS score 5 (OR, 2.91; 95% CIs, 2.61–3.21, $p < 0.001$) and PSA > 100 ng/ml (OR, 1.61; 95% CIs, 1.26–1.93, $p = 0.004$) were two independent predictors of pG4. Overall in PRISK model, PI-RADS score, Rad score and PSA level were three top-ranked variables for Gleason grade. The discrimination ability of PRISK for PCa Gleason grade was summarized with confusion matrix, AUC, accuracy, F1, precision and Recall in training, testing and external validation, respectively (**Table 3**).

As part of this study, we considered predictive aspects of abridgedly combined use of independent factors at PRISK 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. 3, 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 Rad G0-1, PI-RADS 1–2 and PSA 4–10 ng/ml 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 PRISK

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 PRISK score), PSA ≥ 20 ng/ml (OR, 1.58; 95% CI, 1.20–2.08; $p = 0.001$) and predicted PRISK score ≥ 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 PRISK is shown in Fig. 4.

Discussion

This proof-of-concept study contributes important methodology accompanied with model interpretability to address a critical clinical question for PCa risk stratification. Our results from a large cohort of 1,442 biopsy-naïve patients in two tertiary care medical centers show promises of PRISK 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 PRISK for disease progression risk stratification.

Image embedding by 5 pre-trained deep transfer learning 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. Zupan et al [24] initially explored such democratized image analytic tool box by integrating transfer embedding, which, to a nicety, got out of the way of ‘training will be overfitting’. In our experiments, the newly-build imaging biomarkers combined use of PSA level and PI-RADS score did make the impact on the prediction of PCa Gleason grade. Even without incorporating clinical indicators, the new imaging biomarker can determine the pathological Gleason grade with a maximum accuracy of 90%. The auto stack-ensemble learning is another core to our method. Different from typical ML approaches that focuses on the task of combined algorithm selection and hyperparameter optimization, our stack-

ensemble approach performed advanced data processing, deep learning and model ensembling. In particular, duo to the ability to employ multi-layer stack ensembling that combines the aggregated predictions of the base models as its features, it did 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 level and PI-RADS score on stratifying PCa aggressiveness [25–28]. Importantly, our findings enhanced the clinical indications of these factors especially for pG1 and pG2 lesions. Our PRISK allows for only dedicative features 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 and radiological identifications for improving risk stratification in biopsy-naïve patients. Importantly, abridgedly combined use of these factors, like combined used of radiomics with PI-RADS for CIS diseases, and combined use of PSA with PI-RADS for high-risk PCa, have valuable clinical implications for treating planning.

PSA recurrence is currently the strongest clinical end point of PCa, driving almost all initial disease management decisions after primary treatment [25]. It had correctly demonstrated that patient with high-risk PCa had significantly worse BCR-free survival compared with low-risk to intermediate-risk groups [29, 30]. Our preliminary results showcase the PRISK stratification even revealed a potential role in predicting the prognostic of disease progression risk preoperatively. We found that predicted PRISK score $\geq$ G3, PSA level $\geq$ 20 ng/ml were significantly associated with the worse BCR-free survival, implying the prognostic relevance of our PRISK assessment on short and long-term management of patients. Combined use of PRISK 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 PRISK 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. 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 [31, 32]. Third, currently, the deep transfer learning only used the center slice instead of the 3D full tumor volume, so the effect comparison between RADs and TLRs 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 TLR features are still informative.

Conclusions

In summary, we proposed an interpretable tool for PCa aggressiveness assessment. We provided a robust auto stacked-ensemble learning for integrating multimodality data in relation to PCa Gleason grade. The interpretability of PRISK is particularly imperative towards building trustable classification tools for clinical applications. Our study on two cohorts showed that PRISK 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
PI-RADS
Prostate Imaging Reporting and Data System

Declarations

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

Consent for publication: Not applicable

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 supvised the project; J.B., X.W., R.Z., Y.Z. 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.

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 Key Laboratory of health information technology; contract grant number: SZS201818 (to C.H.H)
3. Contract grant sponsor: National Key R&D Program of China; contract grant number: 2017YFC0114300 (to C.H.H)

References

1. Verhoef EI, Kweldam CF, Kümmerlin IP, 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;72:760–5.
2. Cornford P, Bellmunt J, Bolla M, et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer. Eur Urol. 2017;71:630–42.
3. 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.
4. Porten SP, Whitson JM, Cowan JE, et al. Changes in prostate cancer grade on serial biopsy in men undergoing active surveillance. J Clin Oncol. 2011;29:2795–800.
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;61:1019–24.
6. Schroder FH, Hugosson J, Roobol MJ, et al. Screening and prostate-cancer mortality in a randomized European study. N Engl J Med. 2009;360:1320–8.
7. Zhu X, Albertsen PC, Andriole GL, et al. Risk-based prostate cancer screening. Eur Urol. 2012;61:652–61.
8. Ullrich T, Arsov C, Quentin M, et al. Multiparametric magnetic resonance imaging can exclude prostate cancer progression in patients on active surveillance: a retrospective cohort study. Eur Radiol 2020.
9. Gandaglia G, Ploussard G, Valerio 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;77:733–41.
10. Zhang Y, Chen W, Yue X, et al. Development of a Novel, Multi-Parametric, MRI-Based Radiomic Nomogram for Differentiating Between Clinically Significant and Insignificant Prostate Cancer. Front Oncol. 2020;10:888.
11. Zhao K, Wang C, Hu J, et al. Prostate cancer identification: quantitative analysis of T2-weighted MR images based on a back propagation artificial neural network model. Sci China Life Sci. 2015;58:666–73.
12. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology. 2020;21:233–41.
13. Bonekamp D, Kohl S, Wiesenfarth M, et al. Radiomic Machine Learning for Characterization of Prostate Lesions with MRI: Comparison to ADC Values. Radiology. 2018;289:128–37.
14. Fehr D, Veeraraghavan H, Wibmer A, et al. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. Proc Natl Acad Sci U S A. 2015;112:E6265–73.
15. Lee JG, Jun S, Cho YW, et al. Deep Learning in Medical Imaging: General Overview. Korean J Radiol. 2017;18:570–84.

16. Li F, Chen J, Ge Z, et al. Computational prediction and interpretation of both general and specific types of promoters in Escherichia coli by exploiting a stacked ensemble-learning framework. Brief Bioinform 2020.
17. Barrett T, Rajesh A, Rosenkrantz AB, et al. PI-RADS version 2.1: one small step for prostate MRI. Clinical radiology 2019.
18. Epstein JI, Egevad L, Amin MB, et al. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. 2015; 40: 244–252.
19. Cookson MS, Aus G, Burnett AL, 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. J Urol. 2007;177:540–5.
20. Brockman JA, Alanee S, Vickers AJ, et al. Nomogram Predicting Prostate Cancer-specific Mortality for Men with Biochemical Recurrence After Radical Prostatectomy. Eur Urol. 2015;67:1160–7.
21. Zhang YD, Wang Q, Wu CJ, et al. The histogram analysis of diffusion-weighted intravoxel incoherent motion (IVIM) imaging for differentiating the gleason grade of prostate cancer. Eur Radiol. 2015;25:994–1004.
22. Van Griethuysen JJ, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer research. 2017;77:e104–7.
23. Xu X, Zhang HL, Liu QP, et al. Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma. J Hepatol. 2019;70:1133–44.
24. Godec P, Pancur M, Ilenic N, et al. Democratized image analytics by visual programming through integration of deep models and small-scale machine learning. Nat Commun. 2019;10:4551.
25. Epstein JI, Zelefsky MJ, Sjoberg DD, et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. Eur Urol. 2016;69:428–35.
26. 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;4:89–92.
27. Guazzoni G, Nava L, Lazzeri M, 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;60:214–22.
28. Alberts AR, Roobol MJ, Verbeek JFM, 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;75:310–8.
29. Hamada R, Nakashima J, Ohori M, et al. Preoperative predictive factors and further risk stratification of biochemical recurrence in clinically localized high-risk prostate cancer. Int J Clin Oncol. 2016;21:595–600.
30. Van den Broeck T, van den Bergh RCN, Arfi N, et al. Prognostic Value of Biochemical Recurrence Following Treatment with Curative Intent for Prostate Cancer: A Systematic Review. Eur Urol. 2019;75:967–87.
31. Siddiqui MM, Rais-Bahrami S, Truong H, et al. Magnetic resonance imaging/ultrasound-fusion biopsy significantly upgrades prostate cancer versus systematic 12-core transrectal ultrasound biopsy. Eur Urol. 2013;64:713–9.
32. Siddiqui MM, Rais-Bahrami S, Truong H, et al. Magnetic resonance imaging/ultrasound–fusion biopsy significantly upgrades prostate cancer versus systematic 12-core transrectal ultrasound biopsy. 2013; 64: 713–719.

Tables

Table 1. The baseline characteristics of training cohort, internal and external cohort.

Variable	Training cohort (center 1, n = 672)			Internal cohort (center 1, n = 231)			External cohort (center 2, n = 539)		
	PCa	Benign	*P*	PCa	Benign	*P*	PCa	Benign	*P*
No. of subjects	557	115		188	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%)	26/115 (22.6%)		16/188 (8.5%)	9/43 (20.9%)		23/360 (6.4%)	43/179 (24.0%)	
≥ 60 y, n (%)	508/557 (91.6%)	89/115 (77.4%)		172/188 (91.5%)	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%)	66/115 (57.4%)		57/188 (30.3%)	19/43 (44.2%)		96/360 (26.7%)	101/179 (56.4%)	
10-20 ng/ml, n (%)	174/557 (31.2%)	36/115 (31.3%)		65/188 (34.6%)	16/43 (37.2%)		95/360 (26.4%)	55/179 (30.7%)	
20-100 ng/ml, n (%)	206/557 (37.0%)	13/115 (11.3%)		57/188 (30.3%)	7/43 (16.3%)		126/360 (35.0%)	22/179 (12.3%)	
> 100 ng/ml, n (%)	9/557 (1.6%)	0/115 (0)		9/188 (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	115	<0.001	188	43	<0.001	360	179	
PZ, n (%)	362/557 (65.0%)	42/115 (36.5%)		135/188 (71.8%)	17/43 (39.5%)		243/360 (67.5%)	76/179 (42.5%)	
TZ/ AFMS/CZ, n (%)	195/557 (35.0%)	73/115 (63.5%)		53/188 (28.2%)	26/43 (60.5%)		117/360 (32.5%)	103/179 (57.5%)	
MRI index lesion per patient, n (%)	557	115	<0.001	188	43	<0.001	360	179	0.044
PI-RADS 1-2	28/557 (5%)	80/115 (69.6%)		10/188 (5.3%)	22/43 (51.2%)		35/360 (9.7%)	95/179 (53.1%)	

PI-RADS 3	88/557 (15.8%)	29/115 (25.2%)		37/188 (19.7%)	18/43 (41.9%)		28/360 (7.8%)	44/179 (24.6%)	
PI-RADS 4	189/557 (33.9%)	4/115 (3.4%)		61/188 (32.4%)	2/43 (4.7%)		73/360 (20.3%)	19/179 (10.6%)	
PI-RADS 5	252/557 (45.2%)	2/115 (1.7%)		80/188 (42.6%)	1/43 (2.3%)		224/360 (62.2%)	21/179 (11.7%)	
Biopsy GG, n (%)	557	115	<0.001	188	43	<0.001	360	179	0.044
Negative		115			43		3/360 (0.8%)	179	
GG 1 (3+3)	150/557 (26.9%)			40/188 (21.3%)			36/360 (10%)		
GG 2 (3+4)	122/557 (21.9%)			37/188 (19.7%)			81/360 (22.5%)		
GG 3 (4+3)	142/557 (25.5%)			49/188 (26.1%)			84/360 (23.3%)		
GG 4 (4+4)	126/557 (22.6%)			55/188 (29.3%)			68/360 (18.9%)		
GG 5 (5+5)	17/557 (3.1%)			7/188 (3.7%)			88/360 (24.4%)		
Surgical GG, n (%)	557			188			260		
GG 1 (3+3)	81/557 (14.5%)			25/188 (13.3%)			18/260 (6.9%)		
GG 2 (3+4)	146/557 (26.2%)			58/188 (30.8%)			60/260 (23.1%)		
GG 3 (4+3)	178/557 (32.0%)			51/188 (27.1%)			79/260 (30.4%)		
GG 4 (4+4)	113/557 (20.3%)			40/188 (21.3%)			31/260 (11.9%)		
GG 5 (5+5)	39/557 (7.0%)			14/188 (7.4%)			72/260 (27.7%)		
ECE, n (%)	557			188			260		
+	141 (25.3%)			50/188 (26.6%)			113/260 (43.5%)		

-	416 (74.7%)	138/188 (73.4%)	147/260 (56.5%)
SVI, n (%)	557	188	260
+	90 (16.2%)	36/188 (19.1%)	45/260 (17.3%)
-	467 (83.8%)	152/188 (80.9%)	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

Due to technical limitations, table 2,3 is only available as a download in the Supplemental Files section.

Figures

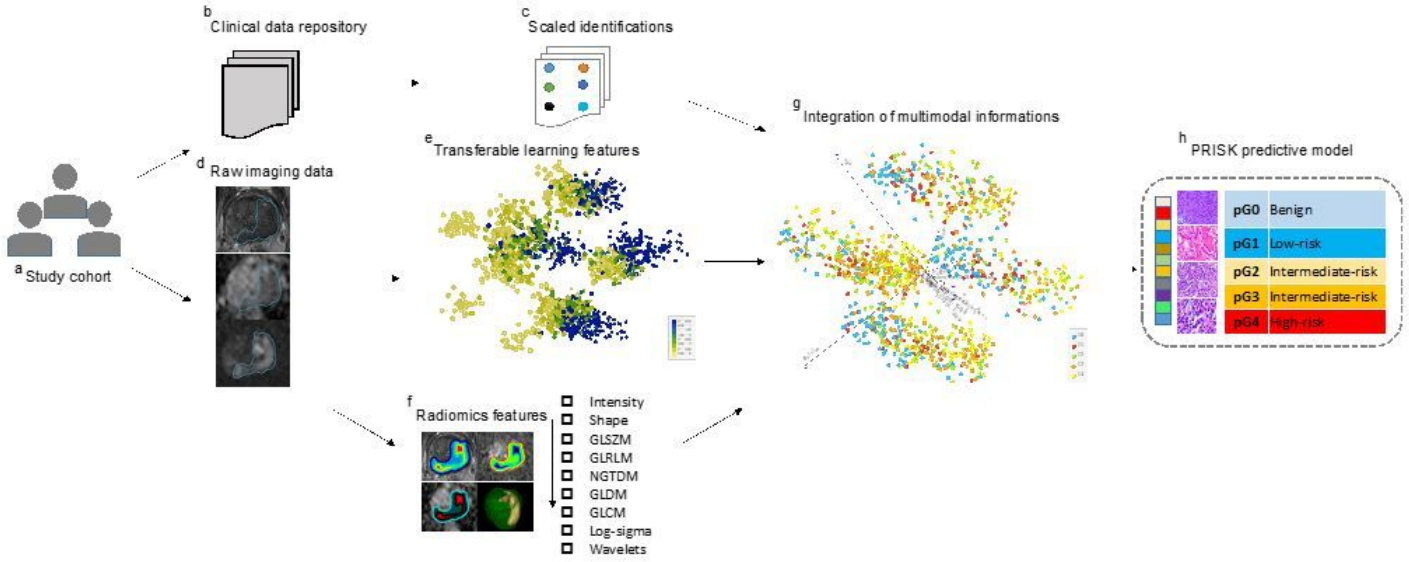

Figure 1

The integrated PRISK 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.

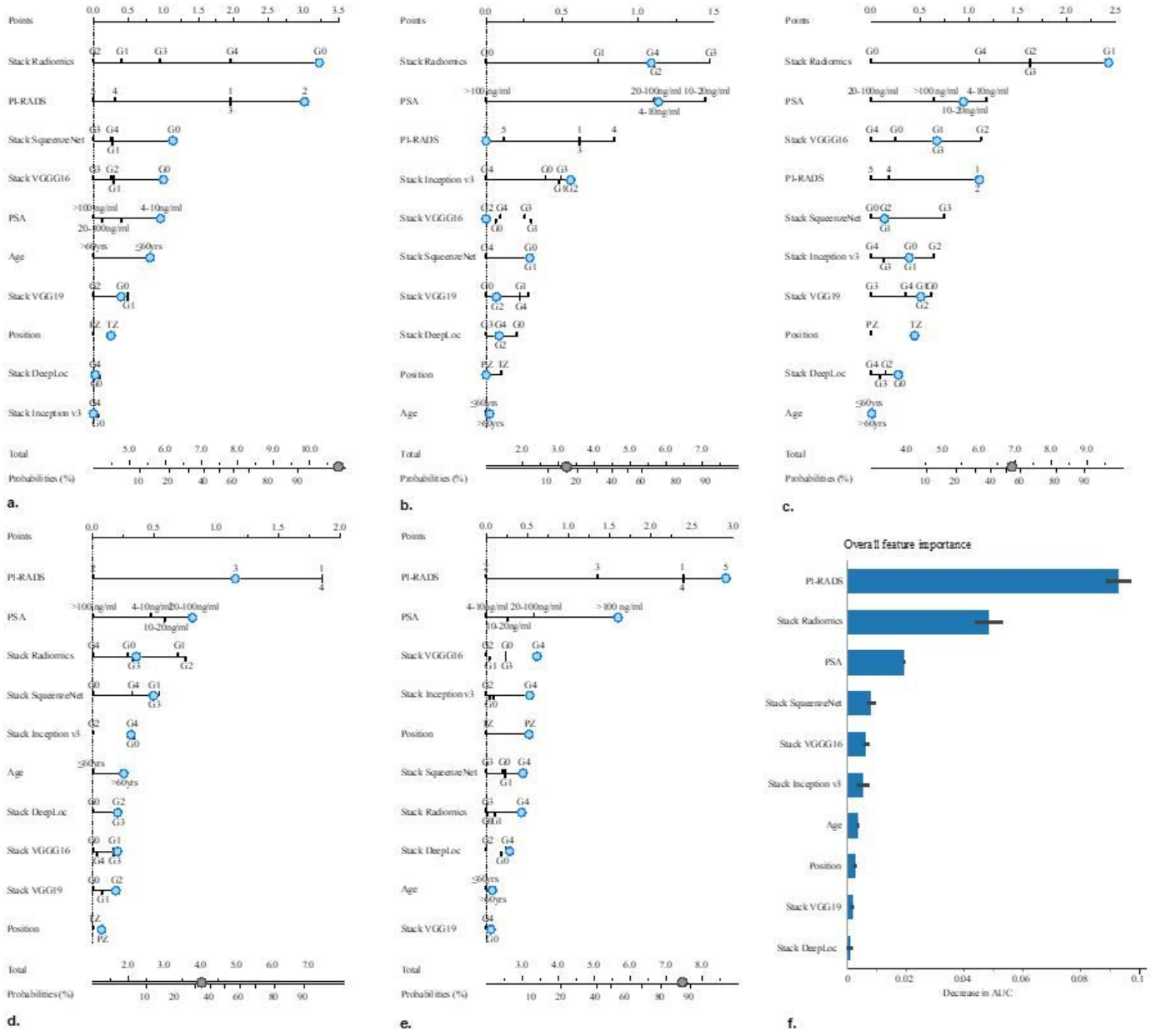

Figure 2

Interpretable nomograms for PRISK model and impact (scaled by odds ratio) of individual factor on the model outputs. Resulting nomograms for pG0 (a), pG1(b), pG2 (c), pG3 (d) and pG4 (e). Integrated impact of individual factor (scaled by decrease in AUC) on model output (f).

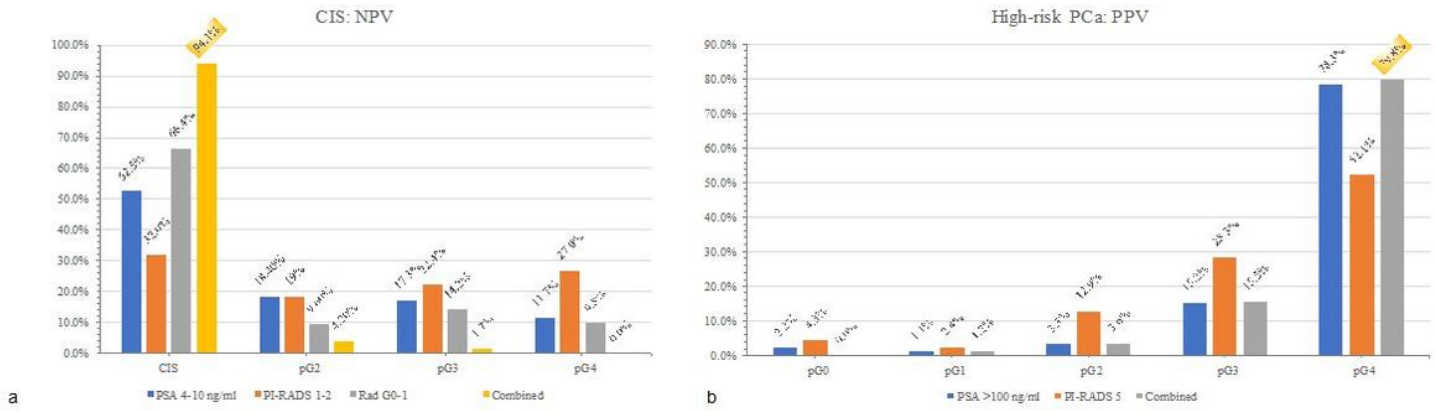

Figure 3

Clinical application of independent factors, i.e., PSA level, Rad score, 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, predicted 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.

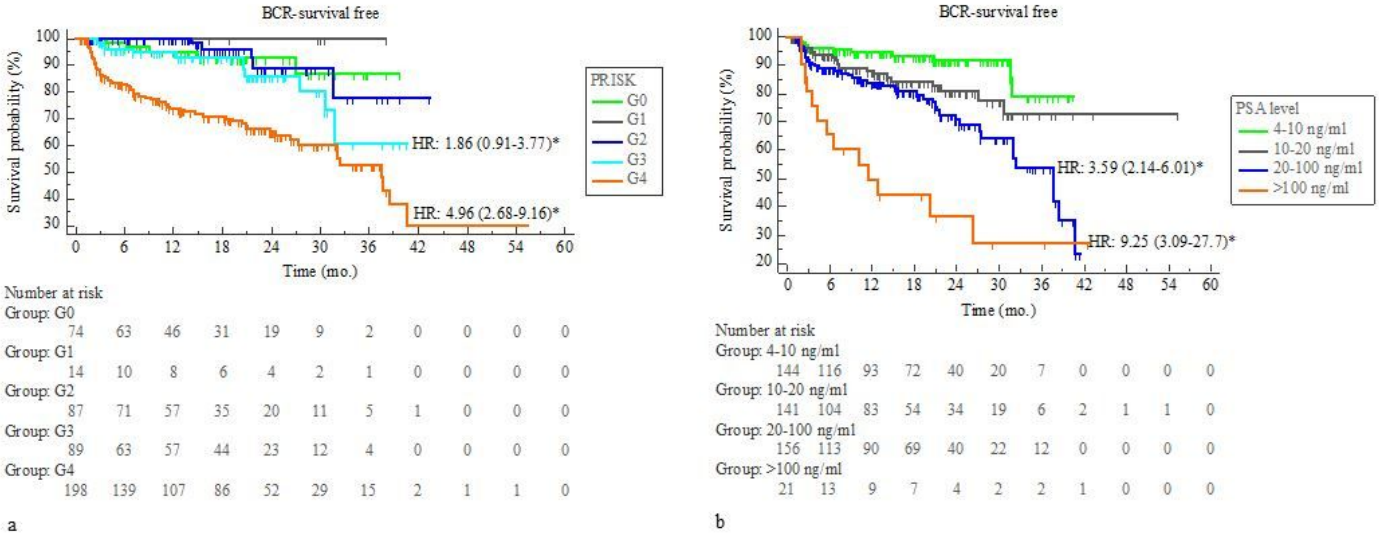

Figure 4

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 PRISK 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.

- SupplementaryTablesandfiguers.docx
- supplementarydata.docx
- Table2.jpg
- Table3.jpg