Detection of Prostate Cancer Using Biparametric Prostate MRI, Radiomics, and Kallikreins: A Retrospective Multicenter Study of Men With a Clinical Suspicion of Prostate Cancer

Ileana Montoya Perez, MSc,[1,2,3*] Harri Merisaari, PhD,[1,2,3] Ivan Jambor, MD, PhD,[1,3,4] Otto Ettala, MD, PhD,[5] Pekka Taimen, MD, PhD,[6] Juha Knaapila, MD,[5] Henna Kekki, MSc,[7] Ferdhos L. Khan, MSc,[7] Elise Syrjälä, MSc,[2] Aida Steiner, MD, PhD,[1,3] Kari T. Syvänen, MD, PhD,[5] Janne Verho, MD,[1,3] Marjo Seppänen, MD,[8] Antti Rannikko, MD, PhD,[9] Jarno Riikonen, MD, PhD,[10] Tuomas Mirtti, MD, PhD,[11] Tarja Lamminen, MSc,[5] Jani Saunavaara, PhD,[12] Ugo Falagario, MD,[13] Alberto Martini, MD,[14] Tapio Pahikkala, PhD,[2] Kim Pettersson, PhD,[7] Peter J. Boström, MD, PhD,[5] and Hannu J. Aronen, MD, PhD[1,3]

Background: Accurate detection of clinically significant prostate cancer (csPCa), Gleason Grade Group ≥ 2, remains a challenge. Prostate MRI radiomics and blood kallikreins have been proposed as tools to improve the performance of biparametric MRI (bpMRI).
Purpose: To develop and validate radiomics and kallikrein models for the detection of csPCa.
Study Type: Retrospective.
Population: A total of 543 men with a clinical suspicion of csPCa, 411 (76%, 411/543) had kallikreins available and 360 (88%, 360/411) did not take 5-alpha-reductase inhibitors. Two data splits into training, validation (split 1: single center, $n = 72$; split 2: random 50% of pooled datasets from all four centers), and testing (split 1: 4 centers, $n = 288$; split 2: remaining 50%) were evaluated.
Field strength/Sequence: A 3 T/1.5 T, TSE T2-weighted imaging, 3x SE DWI.
Assessment: In total, 20,363 radiomic features calculated from manually delineated whole gland (WG) and bpMRI suspicion lesion masks were evaluated in addition to clinical parameters, prostate-specific antigen, four kallikreins, MRI-based qualitative (PI-RADSv2.1/IMPROD bpMRI Likert) scores.

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.27811

Received Apr 27, 2021, Accepted for publication Jun 18, 2021.

*Address reprint requests to: I.M.P., 4th Floor, Vesilinnantie 5, 20500 Turku, Finland. E-mail: iimope@utu.fi

From the [1]Department of Diagnostic Radiology, University of Turku, Turku, Finland; [2]Department of Computing, University of Turku, Turku, Finland; [3]Medical Imaging Centre of Southwest Finland, Turku University Hospital, Turku, Finland; [4]Department of Radiology and Biomedical Imaging, Yale University School of Medicine, New Haven, Connecticut, USA; [5]Department of Urology, University of Turku, Turku University Hospital, Turku, Finland; [6]Institute of Biomedicine, Department of Pathology, University of Turku, Turku University Hospital, Turku, Finland; [7]Department of Biotechnology, University of Turku, Turku, Finland; [8]Department of Surgery, Satakunta Central Hospital, Pori, Finland; [9]Department of Urology, Helsinki University, Helsinki University Hospital, Helsinki, Finland; [10]Department of Urology, Tampere University Hospital, University of Tampere, Tampere, Finland; [11]Department of Pathology, University of Helsinki, Helsinki, Finland; [12]Department of Medical Physics, Turku University Hospital, Turku, Finland; [13]Department of Urology, University of Foggia, Foggia, Italy; and [14]Department of Oncology/Unit of Urology, Urological Research Institute, IRCCS Ospedale San Raffaele, Milan, Italy

Additional supporting information may be found in the online version of this article

Statistical Tests: For the detection of csPCa, area under receiver operating curve (AUC) was calculated using the DeLong's method. A multivariate analysis was conducted to determine the predictive power of combining variables. The values of *P*-value < 0.05 were considered significant.
Results: The highest prediction performance was achieved by IMPROD bpMRI Likert and PI-RADSv2.1 score with AUC = 0.85 and 0.85 in split 1, 0.85 and 0.83 in split 2, respectively. bpMRI WG and/or kallikreins demonstrated AUCs ranging from 0.62 to 0.73 in split 1 and from 0.68 to 0.76 in split 2. AUC of bpMRI lesion-derived radiomics model was not statistically different to IMPROD bpMRI Likert score (split 1: AUC = 0.83, *P*-value = 0.306; split 2: AUC = 0.83, *P*-value = 0.488).
Data Conclusion: The use of radiomics and kallikreins failed to outperform PI-RADSv2.1/IMPROD bpMRI Likert and their combination did not lead to further performance gains.
Level of Evidence: 1
Technical Efficacy: Stage 2

J. MAGN. RESON. IMAGING 2021.

The use of prostate MRI in men with a clinical suspicion of prostate cancer (PCa) is currently recommended by major professional societies such as European Urologist Association and American Urologic Association. The use of prostate MRI and MRI-targeted biopsy can result in a decreased number of men undergoing biopsy procedures while maintaining or improving the detection of clinically significant PCa (csPCa), commonly defined as Gleason Grade Group (GGG) ≥ 2 or > 2, meaning Gleason score ≥ 3 + 4 or > 3 + 4, respectively.

Qualitative and quantitative parameters derived from prostate MRI are not typically used in routine clinical practice beyond the prostate lesion size.[1] Multiple research groups have been working on developing different machine learning methods for prostate MRI aiming to improve the diagnostic performance relative to qualitative report provided by radiologists in a supervised or unsupervised fashion.[2] The use of prostate MRI-derived variables combined with clinical and laboratory findings has attracted substantial interest. A large number of models have been proposed to improve risk stratification of PCa, which use PI-RADS,[3] Likert,[1] IMPROD bpMRI Likert[4–6] score, and/or radiomics analysis and deep learning of prostate MRI images.[7]

In previous retrospective studies, a statistical model based on four kallikrein panel (total-prostate-specific antigen [PSA], free-PSA, intact-PSA, and kallikrein-related peptidase 2 [hK2]) have shown to be useful in predicting biopsy outcome and reducing unnecessary biopsies.[8–10] Furthermore, retrospective studies have shown that the 4Kscore, a test based on the four kallikreins, together with the PI-RADS score lead to improved risk stratification of PCa compared to PI-RADS score alone.[11,12] However, the role of prostate MRI radiomics either from the whole prostate gland (WG) and/or from MRI-based suspicious lesions together with the kallikreins has not been evaluated.

In this study, we aimed to develop and validate radiomics and kallikrein models for the detection of csPCa (GGG ≥ 2) using multi-institutional datasets and compare those with routinely used clinical parameters, PSA, and qualitative MRI parameters (IMPROD bpMRI Likert, PI-RADSv2.1).

Materials and Methods

Study Design and Study Population

All trials involved in this study were approved by the local ethics committee. All enrolled men had given written informed consent before enrolment into the study. Between April 01, 2011 and March 31, 2017, 543 men with a clinical suspicion of PCa underwent prostate MRI followed by biopsy as a part of a single-center trial (cohort A and cohort B) or multicenter trial (cohort C) (Fig. 1). All prostate MRI examinations were performed based on elevated PSA (PSA > 2.5 ng/mL) and/or abnormal digital rectal examination (DRE), and men with a history of PCa were not eligible for enrolment. Criteria described by the standards of reporting for MRI-targeted biopsy studies (START) and the reporting of diagnostic accuracy (STARD) consortium were followed in reporting the results of these trials.[13,14] All anonymized datasets, including bpMRI data and reports, scanned prostatectomy images and biopsy reports, follow-up information, in cohort B and C are available at the following address: http://petiv.utu.fi/improd, http://petiv.utu.fi/multiimprod/.

Study End Points

The primary end point of this retrospective analysis was the diagnostic accuracy of different individual features and models for the detection of csPCa, defined as GGG ≥ 2.[15] The models were based on clinical, laboratory, and bpMRI-derived variables aiming to predict csPCa in men who underwent prostate bpMRI before biopsy due to clinical suspicion of PCa. The "ground truth" for predicting csPCa was based on biopsy or prostatectomy findings (men who underwent prostatectomy following biopsy procedure—cohort A: 18, cohort B: 64, cohort C: 96).

MRI Protocol and MRI Reporting

All prostate MRI examinations were performed either at 3 T or 1.5 T magnetic field using the same T2-weighted imaging (T2W) and diffusion weighted imaging (DWI) acquisitions throughout the duration (2011–2017) of all three trials, no changes occurred in the acquisition protocols during duration of the trials. Prostate MRI examination was performed using body array coils (no endorectal coil) at 3 T MRI scanners in Turku, Finland (Verio, Siemens), Helsinki, Finland (Skyra, Siemens) and Tampere, Finland (Skyra, Siemens) while 1.5 T (Aera, Siemens) MRI scanner was used in Pori, Finland. The same MRI acquisition protocol was used throughout the duration of each trial and no changes in the MRI acquisition protocol were made. Imaging consisted of Turbo Spin Echo T2W acquisitions in axial and sagittal planes. Three

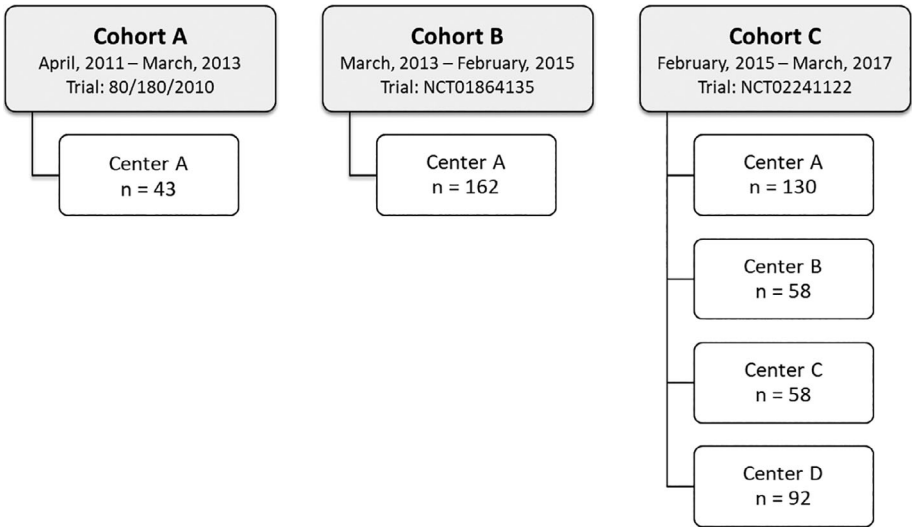

FIGURE 1: Patients included in cohorts by center.

separate Spin Echo DWI acquisitions were utilized: 1) b values 0, 100, 200, 300, 500 s/mm^2[16]; 2) b values 0, 1500 s/mm^2; and 3) b values 0, 2000 s/mm^2. The overall imaging time using 3 T scanners was 13–17 minutes including shimming and calibration while the corresponding time on 1.5 T was about 3 minutes longer. Basic MRI acquisition parameters are presented in the Supporting Material Table S1. DWI datasets were postprocessed using vendor-specific software with monoexponential fit to generate apparent diffusion coefficient (ADC) maps. Radiomics calculated from monoexponential fit of DWI done using 5 b-values in the range of 0–500 s/mm^2. Detailed MRI protocol and importable MRI protocols are publicly available at http://mrc.utu.fi/protocols/prostate and http://petiv.utu.fi/multiimprod/.

All image datasets were reported by a local radiologist (1–2 years of prostate MRI experience at the beginning of the trials in 2011) and re-reported or confirmed centrally by one designated central reader (IJ, 3 years of prostate MRI experience at the beginning of trial A in 2011) to guarantee reporting integrity prior to performing prostate biopsy. Studies in trials A, B, and C were prospectively reported using a dedicated IMPROD bpMRI Likert scoring system developed before initiation of the trials (see details at http://petiv.utu.fi/multiimprod/). The central reader was blind to all clinical data such as PSA, age, and patient past medical history. Following completion of the each trial, all bpMRI datasets were reported using PI-RADSv2.1 scoring system by the same central reader.[3] Since dynamic contrast-enhanced MRI was not performed, the peripheral zone lesions were scored solely by DWI.[17,18]

Inter-reader variability in reporting IMPROD bpMRI Likert and PI-RADsv2.1 in this dataset has been reported previous using a random selection of 81 patients.[4]

Biopsy Procedure and Histopathological Analysis

Cognitive targeting without MRI-TransRectal UltraSonography (TRUS) fusion was performed in cohort A and B. In cohort C, one of the centers (17%, 58/338) used MRI-TRUS fusion (UroNav Fusion Biopsy, Invivo Corporation) while others used cognitive targeting. In the case of a suspicious lesion on MRI (IMPROD bpMRI Likert score 3–5), systematic + targeted biopsy was performed, while men with no MRI-based suspicious lesions (IMPROD bpMRI Likert score 1–2), underwent systematic biopsy (12 cores). All prostate biopsies were performed by experienced urologists ($n = 7$) transrectally without enema and with periprostatic block.

All biopsy and prostatectomy specimens were reported locally at each center by a dedicated pathologist, each with at least 5 years of experience in genitourinary pathology at the beginning of the trial A, using the 2014 International Society of Urological Pathology Modified Gleason Grading System.[19]

Prostate Lesion and Whole Gland Segmentation

The central reader delineated the prostate capsule and bpMRI suspicious lesions (PI-RADSv2.1/IMPROD bpMRI Likert score > 2) on axial T2W imaging and individually on ADC maps (DWI done using 5 b-values in the range of 0–500 s/mm^2) using Carimas (version 2.9, Turku PET center, Turku, Finland) software. Suspicious lesions on bpMRI were delineated manually without knowledge of clinical or laboratory parameters such as PSA level. The lesion extent was determined by the largest signal abnormality of the lesion relative to the normal appearing surrounding tissue seen on T2W imaging and/or the separate three DWI acquisitions.

Data Analyses and Modeling

The study postprocessing pipeline is presented in Fig. 2. In the initial phase of the study, radiomic features of the manually delineated whole prostate gland (WG) and lesion masks were extracted from ADC maps and T2W images. These features included statistical descriptors (Moments), corners edge detector (EdgesCorner2D3D), Fourier transform filter (FFT2D), three-dimensional laws (Laws3D), features describing shape (Shapes), and texture features (Pyradiomics). The size and shape of the lesion were used to capture size properties of the lesion. In addition, the applied features aim to capture voxel level patterns and inhomogeneities in the lesion, corresponding to inhomogeneities in T2W and DWI signal in tumorous tissue in relation to healthy prostate tissue. Further, WG radiomics were included to evaluate usefulness of information outside delineated lesions, and for the purpose of methods which do not require manual delineation of bpMRI suspicious lesions. Details about the features utilized have been described

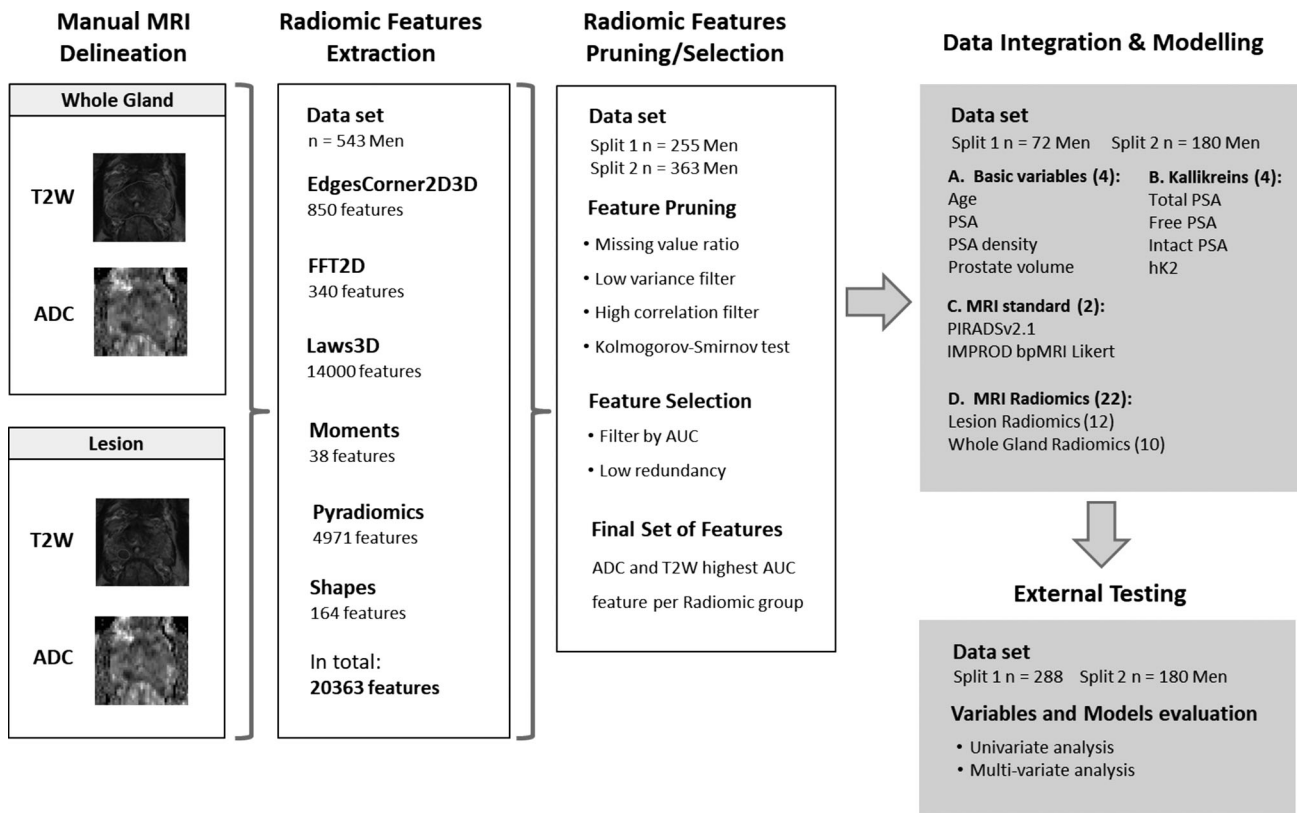

FIGURE 2: The study postprocessing pipeline.

previously,[20,21] and the extraction algorithms are available at https://github.com/haanme/ProstateFeatures.

In the next phase, a pruning and feature selection strategy was applied to obtain a set of radiomic features that would accurately predict csPCa. The strategy consisted of first removing features with missing values. Then applying a low variance filter to drop features with zero or near-zero variance. Next Pearson's r was used as a correlation filter to avoid highly redundant features, and Kolmogorov–Smirnov to test robustness of features between cohorts. In the feature selection process, the five features with the highest area under the received operating curve (ROC) curve (AUC) from each radiomic group were selected. Pearson's r correlation filter was again used to finally choose the top AUC feature per radiomic group taking into account the correlation with other selected features. The same feature pruning and selection strategy were applied to WG and lesion radiomic features separately.

In the data integration and modeling phase, the following four variable groups were considered individually and combined: basic variables (age, PSA, PSA-density, prostate volume), Kallikreins (total-PSA, free-PSA, intact-PSA, hK2), MRI qualitative features (IMPROD bpMRI Likert, PI-RADSv2.1), and top selected MRI radiomic features (10 WG and 12 lesion features). PSA-density was defined as PSA divided by prostate volume (volume of WG masks). Lastly, external evaluation of individual variables and multivariate models was performed on an independent test set unseen during the development and validation phase.

To evaluate and compare the kallikreins performance against other features, in the modeling phase and final external testing, we only considered cases with kallikrein data available. However, in the radiomic feature selection phase, cases without kallikreins were included to improve sample size. Subsequently, to avoid bias brought by the effect that 5-alpha-reductase inhibitor (5-ARI) medication has on PSA and PSA-based kallikreins,[22–24] we identified cases that were not taking 5-ARI medication within 6 months of bpMRI examination and only included those in the modeling, validation and external testing phases. Inclusion of cases in each of the study phases is shown in Fig. 3.

In addition, we considered the effect of multicenter data on modelling, validation, and external testing by performing our analyses in two different data splitting approaches. In the first approach (Data Split 1), models were trained using data from a single-center (i.e., cohort B, $n = 72$) and externally evaluated on multicenter data (i.e., cohort C, $n = 288$). In the second approach (Data Split 2), multicenter data were pooled (i.e., cohort B and C, $n = 360$) and randomly split into 50% training and 50% testing (Fig. 3). In both approaches, the data used for external testing were never used in the other study phases (i.e., radiomic feature pruning/selection, integration, and modeling). Datasets for radiomic feature selection and model training are available at http://mrc.utu.fi/data.

Statistical Methods

To evaluate the ability of each variable/feature in detecting csPCa in men with a clinical suspicion of PCa, a univariate analysis based on AUC with 95% confidence intervals (CI) was performed using the DeLong's method.[25,26] A multivariate analysis, using regularized least-squares (RLS)[27] with regularization parameter 1, was conducted to determine the predictive power of combining variables. Each RLS model was externally validated using an independent test set.

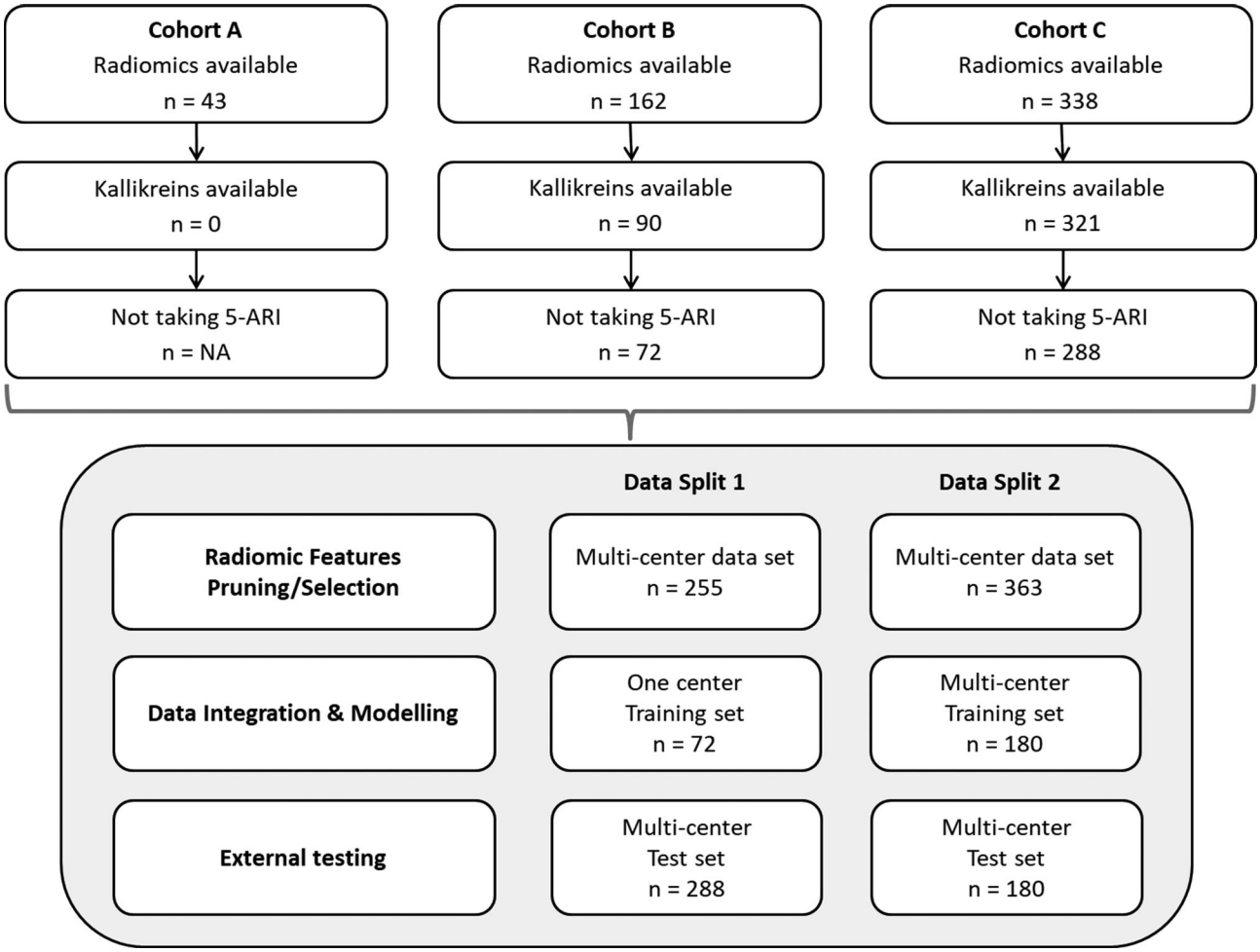

FIGURE 3: Training, validation, and testing data splitting approaches.

DeLong's test was used to compare RLS model test set AUC against qualitative IMPROD bpMRI Likert score AUC. Additionally, ROC curves obtained using 10-fold cross-validation[28] on cohort C ($n = 288$) were plotted to present the diagnostic ability of the models when trained and tested on multicenter data. Radiomic features were scaled between zero and one using min–max normalization. Analyses in modeling and external evaluation phases were performed on patient level, where only the dominant lesion was included. The dominant lesion was defined as the lesion with the highest PI-RADSv2.1/IMPROD bpMRI Likert score followed by lesion size.

RLS models were implemented in Python v. 3.6 using publicly available RLScore software version 0.8.1[29] (https://github.com/aatapa/RLScore). Stratified 10-fold cross-validation was implemented using scikit-learn v. 0.20.0.[30] Statistical analyses were conducted using R v. 3.4.3 software (R Foundation for Statistical Computing, Vienna, Austria). For multiple comparisons, Bonferroni-adjustment was carried out. A total of 14 models were compared to IMPROD bpMRI Likert using DeLong's test and Bonferroni-adjusted alpha level of .004 (0.05/14). Results with P-value < 0.05 were considered significant.

Results

Patient characteristics of each clinical trial are presented in Table 1.

Feature Selection and Pruning

The number of remaining radiomic features after pruning varied between data splits. In Data Split 1, the number of features derived from bpMRI suspicious lesions was reduced from 12,525 to 1315, while in Data Split 2, it was reduced to 1797. Despite the difference of 482 features between splits, the features' AUC ranges were close (0.5–0.82) and (0.5–0.84) in Data Split 1 and 2, respectively. WG radiomic features were reduced from 7838 to 522 in Data Split 1 and to 692 in Data Split 2, respectively, the remaining features' AUC ranges were (0.5–0.76). From the radiomic groups (Table 2), Laws3D WG features in both ADC and T2W were pruned out completely.

The final selected features, which were included in further analyses, are presented in Tables 3 and 4. A total of 22 radiomic features, 12 from lesion and 10 from WG were selected in both Data Split 1 and Data Split 2. The selected features differed between splits, except for the lesion features for ADC Moments and T2W Moments, and the WG features for ADC FFT2D, T2W EdgesCorner2D3D and T2W Moments. The AUC with 95% CI for each selected feature and Pearson correlation in both splits are presented in the Supporting Materials (Table S2–S5 and Figure S1–S4).

TABLE 1. Patients' Characteristics by Study Cohort

	Cohort A	Cohort B	Cohort C
	n = 43	*n* = 162	*n* = 338
Trial duration	April 01, 2011 to March 31, 2013	March 01, 2013 to February 27, 2015	February 01, 2015 to March 31, 2017
Age, years; median (IQR)	67 (63–70)	65 (61–69)	65 (59–69)
PSA, mg/L; median (IQR)	7.7 (6.2–9.1)	7.5 (5.7–9.6)	6.9 (5.1–9.0)
PSA density, %; median (IQR)	0.16 (0.11–0.24)	0.20 (0.13–0.29)	0.16 (0.11–0.24)
Prostate volume; median (IQR)	46.0 (33.5–58.0)	37.5 (28.0–49.0)	39.0 (30.0–53.8)
5-ARI; *n* (%)	NA	27 (17)	34 (10)
IMPROD bpMRI Likert score; *n* (%)			
1	8 (19)	31 (19)	32 (10)
2	9 (21)	8 (5)	44 (13)
3	10 (23)	23 (14)	63 (19)
4	3 (7)	21 (13)	62 (18)
5	13 (30)	79 (49)	137 (41)
PI-RADSv2.1 score; *n* (%)			
1	6 (14)	29 (18)	32 (10)
2	11 (26)	8 (5)	44 (13)
3	9 (21)	21 (13)	63 (19)
4	6 (14)	50 (31)	86 (25)
5	11 (26)	54 (33)	113 (33)
Gleason Grade Group, *n* (%)			
Benign	10 (23)	57 (35)	131 (39)
1 (Gleason score 3 + 3)	11 (26)	16 (10)	54 (16)
2 (Gleason score 3 + 4)	11 (26)	41 (25)	56 (17)
3 (Gleason score 4 + 3)	3 (7)	20 (12)	43 (13)
4 (Gleason score 4 + 4, 3 + 5, 5 + 3)	6 (14)	25 (15)	32 (9)
5 (Gleason score 4 + 5, 5 + 4)	2 (5)	3 (2)	22 (6)

IQR = min-max values interquartile range; PSA = prostate-specific antigen; 5-ARI = five-alfa-reductase inhibitors; bpMRI = biparametric MRI; PI-RADSv2.1 = prostate imaging reporting and data system version 2.1.

Univariate Analysis

Individual feature diagnostic performance using the external validation data in both data splits is shown in Fig. 4. The highest performance was achieved by MRI qualitative score IMPROD bpMRI Likert with AUC (95%CI) of 0.85 (0.81–0.89) in Data Split 1 and 0.85 (0.80–0.90) in Data Split 2. PI-RADSv2.1 had the same AUC as IMPROD bpMRI Likert in Data Split 1, and in Data Split 2 it had AUC of 0.83 (0.77–0.89). Lesion radiomics were the second group of features that presented high diagnostic performance with feature AUCs ranging from 0.58 to 0.84 in Data Split 1 and from 0.75 to 0.85 in Data Split 2. From basic and kallikreins feature groups, PSA-density and total-PSA had the highest AUC within their group, with AUCs of 0.71 and 0.67, respectively. Selected WG radiomic features AUCs ranged from 0.52 to 0.74 in Data Split 1 and from 0.52 to 0.71 in Data Split 2. In this set of features, the selected ADC

TABLE 2. Lesion and Whole Gland Number of Remaining Features per Radiomic Group After Pruning for Two Data Splits

		Lesion Radiomics				Whole Gland Radiomics			
		Data Split 1 NL = 219, csPCa = 138		Data Split 2 NL = 292, csPCa = 177		Data Split 1 N = 255, csPCa = 131		Data Split 2 N = 363, csPCa = 175	
	Radiomic Group	No. Features	AUC range	No. Features	AUC range	No. Features	AUC range	No. Features	AUC range
ADC	EdgesCorners2D3D	94	0.50–0.78	105	0.51–0.80	4	0.58–0.70	21	0.50–0.68
	FFT2D	35	0.50–0.74	35	0.50–0.74	29	0.50–0.68	31	0.50–0.63
	Laws3D	206	0.50–0.63	227	0.50–0.62	0	NA	0	NA
	Moments	9	0.58–0.78	9	0.56–0.78	7	0.50–0.69	8	0.51–0.67
	Pyradiomics	501	0.50–0.82	519	0.50–0.84	346	0.50–0.76	380	0.50–0.76
	Shapes	27	0.50–0.79	31	0.51–0.79	12	0.51–0.67	15	0.50–0.66
T2W	EdgesCorners2D3D	166	0.50–0.75	180	0.50–0.76	4	0.56–0.67	20	0.50–0.66
	FFT2D	36	0.50–0.71	42	0.50–0.72	12	0.50–0.57	23	0.50–0.59
	Laws3D	5	0.55–0.59	198	0.50–0.68	0	NA	0	NA
	Moments	8	0.52–0.75	8	0.55–0.76	7	0.51–0.68	7	0.50–0.67
	Pyradiomics	205	0.50–0.75	420	0.50–0.76	97	0.50–0.74	178	0.50–0.75
	Shapes	23	0.50–0.75	23	0.51–0.76	4	0.53–0.66	9	0.50–0.66
	Total	1315	0.50–0.82	1797	0.50–0.84	522	0.50–0.76	692	0.50–0.76

NL = number of lesions; csPCa = clinically significant prostate cancer; AUC = area under the ROC curve; NA = not applicable; ADC = apparent diffusion coefficient; T2W = T2-weighted imaging; EdgesCorners2D3D = corners edges detector; FFT2D = Fourier transform filter; Laws3D = three-dimensional laws; Shapes = features describing shape; Pyradiomics = texture features.

and T2W Pyradiomics had generally higher AUC than the corresponding top feature from other radiomic groups.

Multivariate Analysis

AUCs from RLS models that included lesion radiomic features were not found to be significantly different from IMPROD bpMRI Likert AUC in either of the Data Splits 1 and 2 (Table 5). In contrast, AUCs of RLS models that did not include lesion radiomic features were significantly lower than IMPROD bpMRI Likert AUC.

In 10-fold cross-validation, mean ROC curve with one standard deviation (SD) for IMPROD bpMRI Likert, PI-RADSv2.1 and RLS models using individual or combined features from Data Split 1 and test set (n = 288) showed that IMPROD bpMRI Likert (AUC = 0.85, SD = 0.09), PI-RADSv2.1 (AUC = 0.85, SD = 0.07) and lesion radiomics (AUC = 0.84, SD = 0.07) models had a higher mean AUC estimate than RLS models based on basic (AUC = 0.74, SD = 0.14), kallikreins (AUC = 0.73, SD = 0.10), or the selected WG radiomics (AUC = 0.74, SD = 0.10) (Fig. 5). RLS models combining kallikreins with WG radiomics (AUC = 0.79, SD = 0.08) demonstrated higher mean AUC estimate and stability than using models from individual basic, kallikreins, and WG radiomic features.

Discussion

In this retrospective multicenter study, we have developed and validated models for predicting csPCa in men with a clinical suspicion of PCa using basic clinical variables, four kallikreins and qualitative and quantitative features of bpMRI. Our analyses showed that basic variables, four kallikrein markers, and top selected WG radiomic features, alone or combined, were not superior to PI-RADSv2.1/IMPROD bpMRI Likert score assigned by an experienced radiologist for predicting csPCa. However, the prediction performance of the lesion radiomics model was similar (although slightly lower) to PI-RADSv2.1/IMPROD bpMRI Likert score; this is in line with other prostate MRI machine learning studies,[2,7,31,32] indicating the necessity of lesion delineation for the successful prediction of csPCa.

TABLE 3. Selected Lesion Features per Radiomic Group in Data Split 1 and Data Split 2

	Lesion Radiomic Group	Data Split 1		Data Split 2	
		Feature	Physical Interpretation	Feature	Physical Interpretation
ADC	Corner edge detector (EdgesCorners2D3D)	Harris–Stephens filter b = 4 ks = 3 k = 0.05 Corner density Lesion/WG ratio overall	Amount of nonhomogeneous lesion locations in contrast to WG	Shi–Tomasi filter b = 1000, ks = 0.100, k = 2.00 Number of corners Lesion/WG ratio	Amount of nonhomogeneous Lesion locations in contrast to WG
	Fourier transform filter (FFT2D)	f = 1.0 FWHM = 2.00 mm Lesion Interquartile range of Low-Pass filtered data	Variation of values when noise is removed	f = 1.0 FWHM = 1.00 mm Lesion Interquartile range of Low-Pass filtered data	Variation of values when noise is removed
	Three-dimensional laws (Laws3D)	3D-Laws component W5W5W5 f = 1.0 25% Percentile of filtered Lesion region	Lesion intensity of low values in 3D pattern	3D-Law component L5L5W5 f = 0.5 Kurtosis of filtered Lesion region	Lesion inhomogeneity of values in 3D pattern
	Statistical descriptors (Moments)	Range of Lesion Intensity	Lesion intensity inhomogeneity	Range of Lesion Intensity	Lesion intensity inhomogeneity
	Pyradiomics	1 mm log_sigma 2.0 mm 3D glszm ZoneEntropy	Lesion intensity inhomogeneity	3 mm original glszm Gray Level Non-Uniformity Normalized	Lesion intensity inhomogeneity
	Shapes	Lesion/WG relative surface area smoothed	Relative lesion shape curvature and size	Lesion/WG relative surface area of mesh faces	Relative lesion shape curvature and size
T2W	Corner edge detector (EdgesCorners2D3D)	Harris–Stephens filter b = 4 ks = 7 k = 0.50 Corner density Lesion/WG ratio	Amount of nonhomogenous lesion locations in contrast to WG	Harris–Stephens filter b = 4 ks = 7 k = 0.50 Corner density Lesion/WG ratio overall	Amount of nonhomogenous lesion locations in contrast to WG
	Fourier transform filter (FFT2D)	F = 1.0 FWHM = 4.00 mm Kurtosis of Low-Pass filtered Lesion region	Kurtosis of values when noise is removed	F = 1.0 FWHM = 3.00 mm Kurtosis of Low-Pass filtered Lesion region	Kurtosis of values when noise is removed
	Three-dimensional laws (Laws3D)	3D-Laws component W5W5W5 f = 1.0 Lesion/WG relative intensity	Relative intensity of low values in 3D pattern	3D-Laws component W5W5W5 f = 1.0 Lesion SD	Variation of intensity of low values in 3D pattern
	Statistical descriptors (Moments)	Lesion volume (mL)	Lesion size	Lesion volume (mL)	Lesion size
	Pyradiomics	1 mm log sigma 2.0 mm 3D glszm ZoneEntropy	Lesion intensity inhomogeneity	1 mm wavelet filter LLL component glcmJointEntropy	Lesion low-frequency inhomogeneity
	Shapes	Lesion/WG relative surface area of mesh faces	Relative lesion shape curvature and size	CSM_mean_curvaturesm Lesion mean surface curvature	Lesion shape curvature

ADC = apparent diffusion coefficient; T2W = T2-weighted imaging; WG = prostate whole gland; SD = standard deviation.

TABLE 4. Selected Whole Gland Feature per Radiomic Group in Data Split 1and Data Split 2

		Data Split 1		Data Split 2	
	Whole Gland (WG) Radiomic Group	**Feature**	**Physical Interpretation**	**Feature**	**Physical Interpretation**
ADC	Corner edge detector (EdgesCorners2D3D)	Harris–Stephens filter b = 2 ks = 1 k = 0.01 Corner locations density secondary component	Amount of nonhomogenous locations in WG	Harris–Stephens b = 2 ks = 1 k = 0.01 Mean Corner locations density	Amount of nonhomogenous locations in WG
	Fourier transform filter (FFT2D)	F = 1.0 FWHM = 1.00 mm Standard Deviation of Low-Pass filtered WG region	Variation of values when noise is removed	F = 1.0 FWHM = 1.00 mm 75% Standard Deviation of Low-Pass filtered WG region	Variation of values when noise is removed
	Statistical descriptors (Moments)	WG Standard Deviation	General variation inside WG	WG volume (mL)	Prostate size
	Pyradiomics	1 mm wavelet HHH gldm Small Dependence Emphasis	WG intensity high-frequency inhomogeneities	1 mm wavelet HHH gldm Large dependenceEmphasis	WG intensity high-frequency inhomogeneities
	Shapes	WG median distance to Center of Gravity	WG size correlate	WG CSM_mean_curvature	WG shape
T2W	Corner edge detector (EdgesCorners2D3D)	Harris–Stephens filter b = 2 ks = 1 k = 0.01 Corner locations density secondary component	Amount of nonhomogenous locations in WG	Harris–Stephens b = 2 ks = 1 k = 0.01 Corner locations density secondary component	Amount of nonhomogenous locations in WG
	Fourier transform filter (FFT2D)	F = 1.0 FWHM = 1.00 mm 75% Percentile of Low-Pass filtered WG region	Intensity of high values when noise is removed	F = 1.0 FWHM = 5.00 mm 75% Kurtosity of Low-Pass filtered WG region	Intensity of high values when noise is removed
	Statistical descriptors (Moments)	WG intensity 25% Percentile	WG low intensity	WG intensity 25% percentile	WG low intensity
	Pyradiomics	1 mm original shape Sphericity	WG shape	1 mm original shape Surface Volume Ratio	WG shape curvature
	Shapes	WG CSM Surface mean curvature	WG shape	WG median distance to Center of Gravity	WG size correlate

ADC = apparent diffusion coefficient; T2W = T2-weighted imaging.

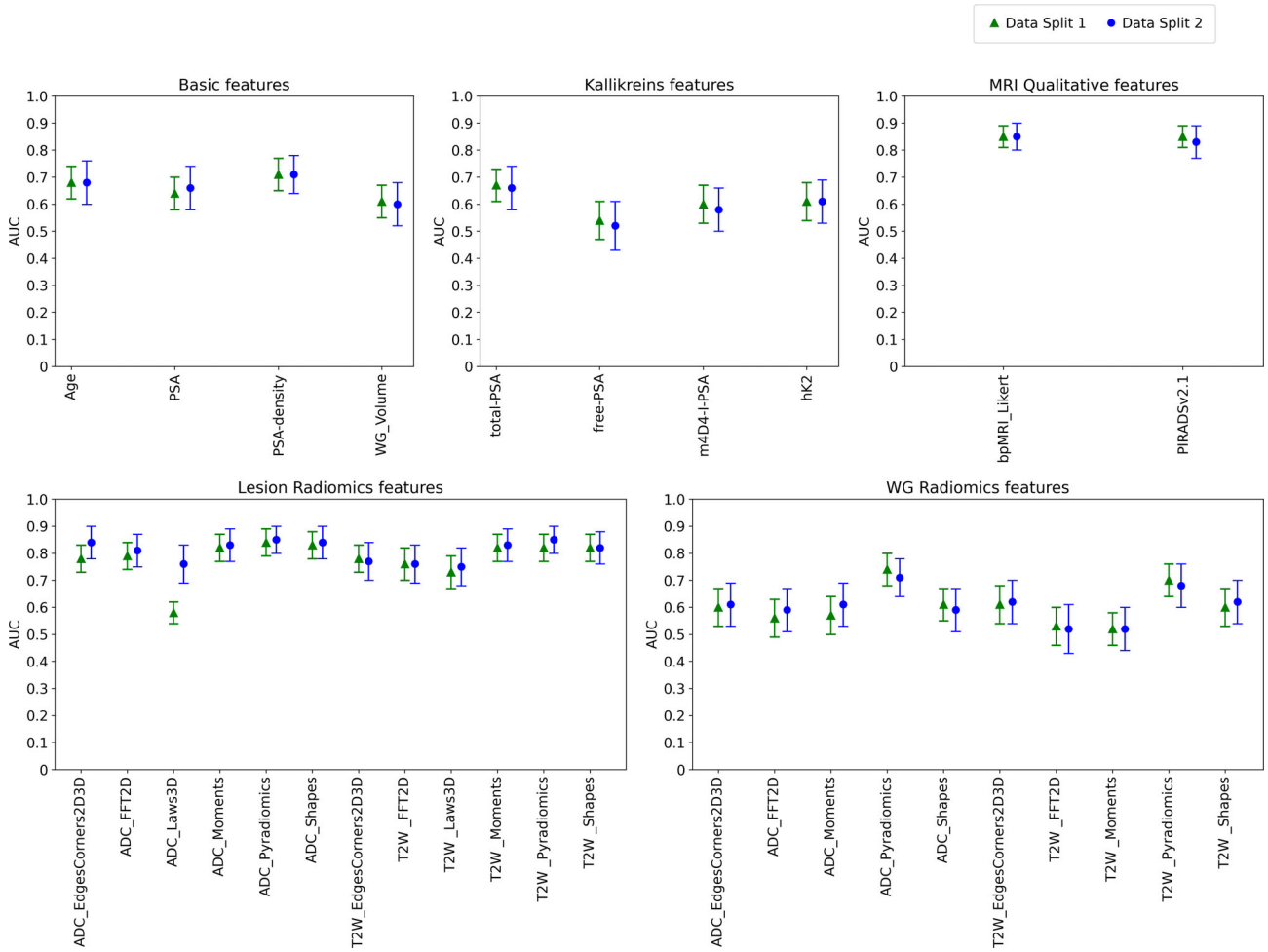

FIGURE 4: Diagnostic performance of features per group, using the external validation data set for Data Split 1 and Data Split 2.

In contrast to prior studies, which evaluated a relatively small number of radiomic features, we used a very large number of radiomics (a total of 20,363) to ensure that all commonly used radiomic features[33] were included along with different combinations of parameters. We hypothesize that the choice of large set of features may help in distinguishing disease pathophysiology better and hence likely improve prediction than a small set of features, which may miss certain tissue characteristics.[34–38] Furthermore, we employed open-source software and provide access to datasets as well as details on our feature pruning/selection process to ensure comparability and transparency. However, direct comparison of our results with prior studies is limited since none of prior studies applying radiomics and machine learning methods for prostate MRI provide access to datasets and postprocessing code. In contrast, free public access to datasets used in the study is provided enabling external validation of our results.

The combination of radiomic features with other relevant variables, such as clinical variables, for improving csPCa detection has been investigated.[38,39] In our study, in addition to the evaluation of clinical variables in combination with WG and lesion radiomics, we included four kallikrein markers that have shown to be valuable in predicting csPCa.[8–10] Although, in this study, the kallikrein model performed poorly compared to PI-RADSv2.1/IMPROD bpMRI Likert score alone, a 10-fold cross-validation on our test set of multicenter datasets showed the potential and stability that the four kallikreins combined with the WG radiomic features in predicting csPCa. This result indicates adequate performance without the need for an experienced radiologist to assign either a PI-RADSv2.1/IMPROD bpMRI Likert score or delineate a bpMRI suspicious lesion. However, WG delineations would still be required. We argue that WG delineations are easier to be obtained by less experience readers rather than assigning either a PI-RADSv2.1/IMPROD bpMRI Likert score and/or performing voxel level annotations of bpMRI suspicious lesions.

Limitations

First, our population consists exclusively of Caucasian men, presenting with rising PSA and/or lower urinary tract

TABLE 5. Multivariate Analysis in Two Data Splitting Approaches for Prediction of Prostate Cancer Gleason Grade Group ≥ 2

		Data Split 1		Data Split 2	
Feature Group	**Num. of Features**	**Test set $n = 288$, csPCa = 133 AUC (CI 95%)**	***P*-value**	**Test set $n = 180$, csPCa = 91 AUC (CI 95%)**	***P*-value**
IMPROD bpMRI Likert score	1	0.85 (0.81–0.89)	REF.	0.85 (0.80–0.90)	REF.
PI-RADSv2.1 score	1	0.85 (0.81–0.89)	1.0	0.83 (0.77–0.89)	0.209
Basic	4	0.73 (0.67–0.79)	<0.001**	0.73 (0.66–0.80)	0.003**
Kallikreins	4	0.62 (0.55–0.69)	<0.001**	0.76 (0.69–0.83)	0.017*
Lesion Radiomics	12	0.83 (0.78–0.88)	0.306	0.83 (0.78–0.90)	0.488
WG Radiomics	10	0.65 (0.59–0.71)	<0.001**	0.68 (0.60–0.76)	<0.001**
Basic and Kallikreins	8	0.70 (0.64–0.76)	<0.001**	0.74 (0.67–0.81)	0.005*
Basic and Lesion Radiomics	16	0.84 (0.79–0.89)	0.604	0.84 (0.78–0.90)	0.720
Basic and WG Radiomics	14	0.72 (0.66–0.78)	<0.001**	0.75 (0.68–0.82)	0.009*
Kallikreins and Lesion Radiomics	16	0.81 (0.76–0.86)	0.079	0.84 (0.78–0.90)	0.723
Kallikreins and WG Radiomics	14	0.68 (0.62–0.74)	<0.001**	0.73 (0.66–0.80)	0.002**
Lesion and WG Radiomics	22	0.82 (0.77–0.87)	0.162	0.83 (0.77–0.89)	0.484
Basic, Kallikreins and Lesion Radiomics	20	0.82 (0.77–0.87)	0.180	0.84 (0.78–0.90)	0.713
Basic, Kallikreins and WG Radiomics	18	0.72 (0.66–0.78)	<0.001**	0.76 (0.69–0.83)	0.017*
Basic, Kallikreins, Lesion and WG Radiomics	30	0.80 (0.75–0.85)	0.035	0.84 (0.78–0.90)	0.714

REF = IMPROD bpMRI Likert score as reference for DeLong's test; bpMRI: biparametric MRI; PI-RADSv2.1: prostate imaging reporting and data system version 2.1; Basic: basic clinical variables (Age, PSA, PSA density, and prostate volume); Kallikreins: four kallikrein markers (Free PSA, Total PSA, intact PSA, hK2); Lesion radiomics: 12 top selected lesion radiomic features; WG radiomics: 10 top selected whole gland radiomic features.
*Significant level *P*-value < 0.05.
**Bonferroni-adjusted significant level *P*-value < 0.004.

symptoms and/or family history of PCa, thus limiting extension of the findings to a wider population. Both 1.5 T (one center) and 3 T (three enters) MRI scanners were used in the prospective trials, which provided datasets for this retrospective analysis. Thus, variation in the performance due to difference between scanners and field strengths is unknown. All enrolled men underwent biopsy, and prostatectomy findings were only available for some patients; thus, true csPCa prevalence in men who did not undergo prostatectomy was unknown. Histopathology findings were reported locally by adhering to international standards, however without a central reviewing framework, systematic differences between institutions could not be assessed. In the study, we have not explored inter-reader variability in whole dataset. However, in a prior study[4] using a portion of the current dataset, we have shown a moderate agreement in assigning IMPROD bpMRI

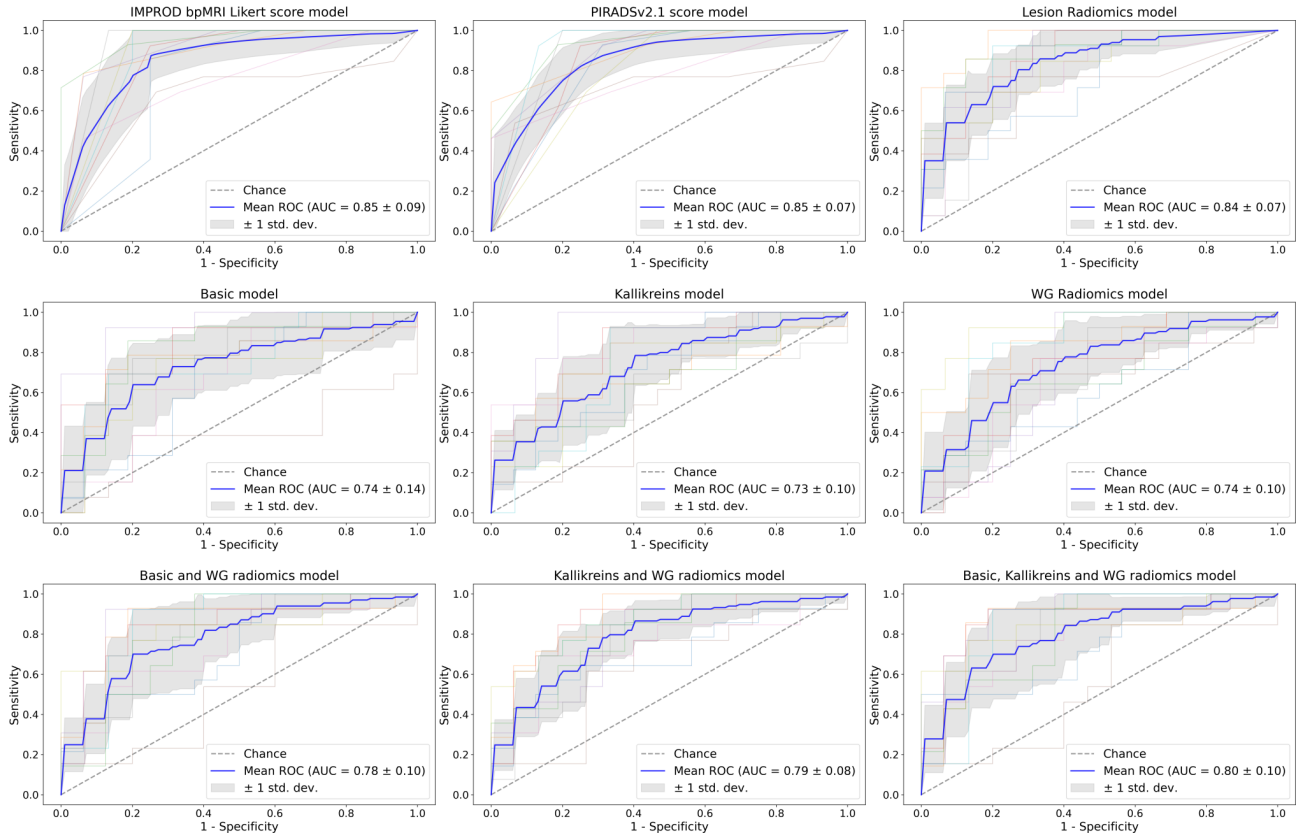

FIGURE 5: Regularized least-squares 10-fold cross-validation ROC curves for IMPROD bpMRI Likert score, PI-RADSv2.1 score, lesion radiomics, basic variables, kallikreins and whole gland radiomics models using the data from the 288 men in test set of data Split 1.

Likert (k = 0.59 or k = 0.68 after dichotomization) and PI-RADSv2.1 (k = 0.54 or k = 0.60 after dichotomization) score between the central reader and an inexperienced reader suggesting that an acceptable performance can be achieved by an inexperienced reader. Voxel level annotations performed by one central reader for datasets of both clinical trials used in this retrospective analysis were used, thus, not inter-readers variation on the voxel level was evaluated. Future studies are needed to evaluate inter-reader variability in voxel level masks/delineations for both whole gland as well as bpMRI suspicious lesions.

Conclusions

We found that the models based on basic variables (age, PSA, PSA density, and prostate volume), four kallikreins and selected WG radiomic features, alone or combined, had inferior performance in csPCa detection than the qualitative score (PI-RADSv2.1 or IMPROD bpMRI Likert) reported by an experienced radiologist. In contrast, a model based on selected lesion radiomic features had comparable performance to PI-RADSv2.1/IMPROD bpMRI Likert score, while combination with the other variables/features did not improve performance in an external validation.

Acknowledgments

This study was financially supported by grants from Instrumentarium Research Foundation, Sigrid Jusélius Foundation, Turku University Hospital, TYKS-SAPA research fund, Finnish Cancer Society, Finnish Cultural Foundation, and Orion Research Foundation. H.M. was supported by the Cultural Foundation of Finland, and Orion Pharma Research Fellowship. P.T. was supported by a Clinical Researcher Funding from the Academy of Finland.

References

1. Khoo CC, Eldred-Evans D, Jaenicke J, et al. Likert vs. PI-RADS v2: A comparison of two radiological scoring systems for detection of clinically significant prostate cancer. Eur Urol Suppl 2019;18:e1865-e1866.
2. Cuocolo R, Cipullo MB, Stanzione A, et al. Machine learning for the identification of clinically significant prostate cancer on MRI: A meta-analysis. Eur Radiol 2020;30:6877-6887.
3. Weinreb J, Barentsz J, Choyke P, et al. PI-RADS prostate imaging-reporting and data system:2015, version 2. Eur Urol 2016;69(1):16-40.
4. Perez IM, Jambor I, Kauko T, et al. Qualitative and quantitative reporting of a unique biparametric MRI: Towards biparametric MRI-based nomograms for prediction of prostate biopsy outcome in men with a clinical suspicion of prostate cancer (IMPROD and MULTI-IMPROD trials). J Magn Reson Imaging 2019;51:4-6.
5. Jambor I, Boström PJ, Taimen P, et al. Novel biparametric MRI and targeted biopsy improves risk stratification in men with a clinical

suspicion of prostate cancer (IMPROD trial). J Magn Reson Imaging 2017;46:1089-1095.

6. Jambor I, Verho J, Ettala O, et al. Validation of IMPROD biparametric MRI in men with clinically suspected prostate cancer: A prospective multi-institutional trial. PLoS Med 2019;16:e1002813.
7. Schelb P, Kohl S, Radtke JP, et al. Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. Radiology 2019;293:607-617.
8. Benchikh A, Savage C, Cronin A, et al. A panel of kallikrein markers can predict outcome of prostate biopsy following clinical work-up: An independent validation study from the European randomized study of prostate cancer screening, France. BMC Cancer 2010;10:1-7.
9. Bryant RJ, Sjoberg DD, Vickers AJ, et al. Predicting high-grade cancer at ten-Core prostate biopsy using four Kallikrein markers measured in blood in the ProtecT study. JNCI J Natl Cancer Inst 2015;107:95.
10. Braun K, Sjoberg DD, Vickers AJ, Lilja H, Bjartell AS. A four-kallikrein panel predicts high-grade cancer on biopsy: Independent validation in a community cohort. Eur Urol 2016;69:505-511.
11. Punnen S, Nahar B, Soodana-Prakash N, et al. Optimizing patient's selection for prostate biopsy: A single institution experience with multiparametric MRI and the 4Kscore test for the detection of aggressive prostate cancer. PLoS One 2018;13:e0201384.
12. Falagario UG, Martini A, Wajswol E, et al. Avoiding unnecessary magnetic resonance imaging (MRI) and biopsies: Negative and positive predictive value of MRI according to prostate-specific antigen density, 4Kscore and risk calculators prostate cancer multiparametric magnetic resonance imaging prost. Eur Urol Oncol 2020;3:700-704.
13. Moore CM, Kasivisvanathan V, Eggener S, et al. Standards of reporting for MRI-targeted biopsy studies (START) of the prostate: Recommendations from an international working group. Eur Urol 2013;64:544-552.
14. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration. BMJ Open 2016;6:e012799.
15. Filson CP, Natarajan S, Margolis DJA, et al. Prostate cancer detection with magnetic resonance-ultrasound fusion biopsy: The role of systematic and targeted biopsies. Cancer 2016;122:884-892.
16. Jambor I, Kähkönen E, Taimen P, et al. Prebiopsy multiparametric 3T prostate MRI in patients with elevated PSA, normal digital rectal examination, and no previous biopsy. J Magn Reson Imaging 2015;41:1394-1404.
17. Boesen L, Norgaard N, Logager V, et al. Assessment of the diagnostic accuracy of biparametric magnetic resonance imaging for prostate cancer in biopsy-naive men: The biparametric MRI for detection of prostate cancer (BIDOC) study. JAMA Netw Open 2018;1:1-28.
18. Boesen L, Nørgaard N, Løgager V, et al. Prebiopsy biparametric magnetic resonance imaging combined with prostate-specific antigen density in detecting and ruling out Gleason 7–10 prostate cancer in biopsy-naïve men. Eur Urol Oncol 2018;17:e2753.
19. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. 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. Am J Surg Pathol 2016;40:244-252.
20. Merisaari H, Taimen P, Shiradkar R, et al. Repeatability of radiomics and machine learning for DWI: Short-term repeatability study of 112 patients with prostate cancer. Magn Reson Med 2019;83:2293-2309.
21. Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77:e104-e107.
22. Sarkar RR, Parsons JK, Bryant AK, et al. Association of treatment with 5α-reductase inhibitors with time to diagnosis and mortality in prostate cancer. JAMA Intern Med 2019;179:812-819.
23. Kumar A, Nalawade V, Riviere P, et al. Association of treatment with 5α-reductase inhibitors and prostate cancer mortality among older adults. JAMA Netw Open 2019;2:e1913612.
24. Vertosick E, Vickers A, Goodman P, Lilja H. PD52-08 changes in blood levels of prostate kallikrein-related and microseminoprotein-beta marker levels following 5ARI therapy. J Urol 2020;203:e1092-e1093.
25. Hanley JA, Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. Acad Radiol 1997;4:49-58.
26. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988;44:837-845.
27. Rifkin R, Yeo G, Poggio T. Regularized least-squares classification. Nato Sci Ser Sub Ser III Comput Syst Sci 2003;190:131-154.
28. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Jt Conf Artif Intell. 1995;14:1137-1143.
29. Pahikkala T, Airola A. RLScore: Regularized least-squares learners. J Mach Learn Res 2016;17:1-5.
30. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825-2830.
31. Cao R, Mohammadian Bajgiran A, Afshari Mirak S, et al. Joint prostate cancer detection and Gleason score prediction in mp-MRI via FocalNet. IEEE Trans Med Imaging 2019;38:2496-2506.
32. Zhong X, Cao R, Shakeri S, et al. Deep transfer learning-based prostate cancer classification using 3 Tesla multi-parametric MRI. Abdom Radiol 2019;44:2030-2039.
33. Sun Y, Reynolds HM, Parameswaran B, et al. Multiparametric MRI and radiomics in prostate cancer: A review. Australas Phys Eng Sci Med 2019;42:3-25.
34. Bonekamp D, Kohl S, Wiesenfarth M, et al. Radiomic machine learning for characterization of prostate lesions with MRI: Comparison to ADC values. Radiology 2018;289:128-137.
35. Cuocolo R, Stanzione A, Ponsiglione A, et al. Clinically significant prostate cancer detection on MRI: A Radiomic shape features study. Eur J Radiol 2019;116(March):144-149.
36. Bernatz S, Ackermann J, Mandel P, et al. Comparison of machine learning algorithms to predict clinically significant prostate cancer of the peripheral zone with multiparametric MRI using clinical assessment categories and radiomic features. Eur Radiol 2020;30:6757-6769.
37. Germanese D, Colantonio S, Caudai C, et al. May radiomic data predict prostate cancer aggressiveness? Commun Comput Inf Sci. 2019; 1089:65-75.
38. Gong L, Xu M, Fang M, et al. Noninvasive prediction of high-grade prostate cancer via biparametric MRI radiomics. J Magn Reson Imaging 2020;52:1102-1109.
39. Woźnicki P, Westhoff N, Huber T, et al. Multiparametric MRI for prostate cancer characterization: Combined use of radiomics model with PI-RADS and clinical parameters. Cancers (Basel) 2020;12:1-14.