

Radiomic Machine Learning for Characterization of Prostate Lesions with MRI: Comparison to ADC Values

David Bonekamp, MD* • Simon Kohl, MSc* • Manuel Wiesenfarth, PhD • Patrick Schelb • Jan Philipp Radtke, MD • Michael Götz, PhD • Philipp Kickingereder, MD • Kaneschka Yaqubi • Bertram Hittahler • Nils Gähler, MSc • Tristan Anselm Kuder, PhD • Fenja Deister, BSc • Martin Freitag, MD • Markus Hobenfellner, MD • Boris A. Hadaschik, MD • Heinz-Peter Schlemmer, MD, PhD • Klaus H. Maier-Hein, PhD

From the Department of Radiology (D.B., P.S., J.P.R., P.K., K.Y., M.F., H.P.S.), Division of Medical Image Computing (S.K., M.G., N.G., K.H.M.H.), Division of Statistics (M.W.), and Department of Medical Physics (T.A.K., F.D.), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg 69120, Germany; German Cancer Consortium (DKTK), Heidelberg, Germany (D.B., H.P.S., K.H.M.H.); and Departments of Urology (J.P.R., B.H., M.H., B.A.H.) and Neuroradiology (P.K.), University of Heidelberg Medical Center, Heidelberg, Germany. Received December 28, 2017; revision requested February 19, 2018; revision received May 28; accepted May 30. Address correspondence to D.B. (e-mail: d.bonekamp@dkfz-heidelberg.de).



Study supported by STIFTUNG KREBSFORSCHUNG Europa.

P.K. is a fellow of the Medical Faculty Heidelberg Postdoc-Program.

*D.B. and S.K. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

See also the editorial by Choyke in this issue.

Radiology 2018; 289:128–137 • <https://doi.org/10.1148/radiol.2018173064> • Content codes:  

Purpose: To compare biparametric contrast-free radiomic machine learning (RML), mean apparent diffusion coefficient (ADC), and radiologist assessment for characterization of prostate lesions detected during prospective MRI interpretation.

Materials and Methods: This single-institution study included 316 men (mean age \pm standard deviation, 64.0 years \pm 7.8) with an indication for MRI—transrectal US fusion biopsy between May 2015 and September 2016 (training cohort, 183 patients; test cohort, 133 patients). Lesions identified by prospective clinical readings were manually segmented for mean ADC and radiomics analysis. Global and zone-specific random forest RML and mean ADC models for classification of clinically significant prostate cancer (Gleason grade group \geq 2) were developed on the training set and the fixed models tested on an independent test set. Clinical readings, mean ADC, and radiomics were compared by using the McNemar test and receiver operating characteristic (ROC) analysis.

Results: In the test set, radiologist interpretation had a per-lesion sensitivity of 88% (53 of 60) and specificity of 50% (79 of 159). Quantitative measurement of the mean ADC (cut-off 732 mm²/sec) significantly reduced false-positive (FP) lesions from 80 to 60 (specificity 62% [99 of 159]) and false-negative (FN) lesions from seven to six (sensitivity 90% [54 of 60]) ($P = .048$). Radiologist interpretation had a per-patient sensitivity of 89% (40 of 45) and specificity of 43% (38 of 88). Quantitative measurement of the mean ADC reduced the number of patients with FP lesions from 50 to 43 (specificity 51% [45 of 88]) and the number of patients with FN lesions from five to three (sensitivity 93% [42 of 45]) ($P = .496$). Comparison of the area under the ROC curve (AUC) for the mean ADC ($AUC_{\text{global}} = 0.84$; $AUC_{\text{zone-specific}} \leq 0.87$) vs the RML ($AUC_{\text{global}} = 0.88$, $P = .176$; $AUC_{\text{zone-specific}} \leq 0.89$, $P \geq .493$) showed no significantly different performance.

Conclusion: Quantitative measurement of the mean apparent diffusion coefficient (ADC) improved differentiation of benign versus malignant prostate lesions, compared with clinical assessment. Radiomic machine learning had comparable but not better performance than mean ADC assessment.

© RSNA, 2018

Online supplemental material is available for this article.

Interpretation of multiparametric MRI according to the Prostate Imaging Reporting and Data System (PI-RADS) and its recent update to version 2.0 (1–3), in combination with transrectal US fusion biopsy, has shown promise in detecting clinically significant cancer at radical prostatectomy, with sensitivities reaching 97% (4,5). However, identifying patients with high sensitivity also comes at the cost of possible overdiagnosis, and further improvement in the differentiation of nonaggressive and aggressive prostate cancer is necessary (6–9). While PI-RADS uses qualitative assessment of diffusion-weighted imaging (DWI), recently the ability of quantitative apparent diffusion coefficient (ADC) measurements to improve interreader concordance

has been pointed out (10,11). DWI, and specifically the ADC, can be considered the current best monoparametric component of prostate MRI assessment, resulting from its ability to probe the microenvironment of neoplastic tissues efficiently and detect alterations in compartmental volumes (epithelium, stroma, and lumen space) and cellularity (12).

Radiomics uses advanced image-processing techniques to extract a large number of quantitative parameters from imaging data, and its potential to improve diagnostic accuracy is increasingly being studied (13–16). Initial studies have reported promising performance of radiomics with and without the use of machine learning in the prediction of

Abbreviations

ADC = apparent diffusion coefficient, AUC = area under the ROC curve, DWI = diffusion-weighted imaging, FN = false-negative, FP = false-positive, PI-RADS = Prostate Imaging Reporting and Data System, PZ = peripheral zone, RML = radiomic machine learning, ROC = receiver operating characteristic, TZ = transition zone, VOI = volume of interest

Summary

Quantitative measurement of the mean apparent diffusion coefficient (ADC) was more accurate than prospective clinical assessment in classifying a lesion as clinically significant prostate cancer rather than a benign lesion or clinically insignificant prostate cancer. No added benefit of radiomic machine learning was present compared with mean ADC values alone.

Implications for Patient Care

- Quantitative apparent diffusion coefficient (ADC) values improved characterization of prostate lesions, compared with qualitative clinical assessment.
- Quantitative ADC alone successfully characterized a lesion as clinically significant prostate cancer in both the peripheral and transition zone; no further benefit was demonstrated with complex radiomics and machine learning methods.

the prostate cancer Gleason score (14–16). However, accuracy varies depending on the machine learning approach used (16). Various textural features have been associated with prostate cancer aggressiveness and the pathologic index lesion: Homogeneity gray-level co-occurrence matrix texture features from T2-weighted images and ADC maps have been suggested to be superior to ADC parameters (14), and radiomics has been reported advantageous compared with and in combination with retrospective PI-RADS assessment (15). Further examination of multiparametric quantitative models, preferentially with the use of an independent test set in larger cohorts and with direct comparison to established monoparameters and clinical assessment, is required to better understand their predictive value.

We hypothesized that the characterization of MRI-detected lesions can be improved by radiomics. The purpose of this study was to compare radiomics and mean ADC for characterization of prostate lesions by using a sensitivity threshold equivalent to clinical reporting.

Materials and Methods

This retrospective analysis was performed in a cohort of men undergoing MRI–transrectal US fusion biopsy. The institutional and governmental ethics committee approved the study and waived informed consent (institutional ethics approval number S-156/2018). All patients had a clinical indication for prostate biopsy (details given in Appendix E1 [online]) based on prostate-specific antigen elevation and clinical examination or participation in our active surveillance program. MRI data of 316 consecutive patients (median age, 64 years; interquartile range [IQR], 58–71 years) examined with a single 3-T MRI system in 2015–2016 were included in the analysis. One hundred eighty-three patients examined from May 2015 until January 2016 were included for model training and parameter optimization (median age, 64.5 years; IQR, 59–71 years), while patients examined

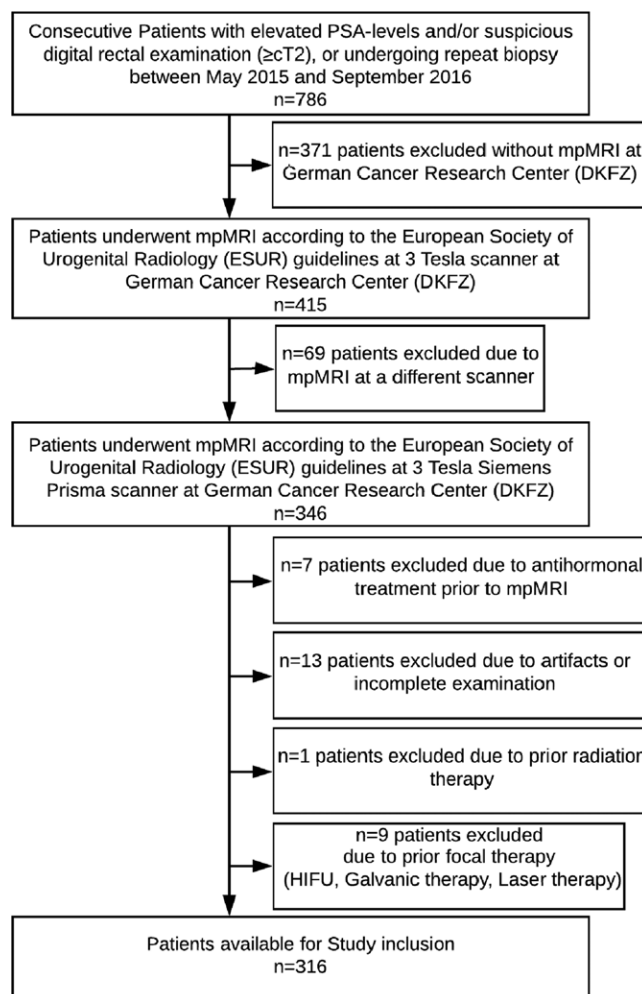


Figure 1: Diagram for inclusion of patients into the study. PSA = prostate-specific antigen, mpMRI = multiparametric MRI, HIFU = high-intensity focused ultrasound.

between January 2016 and September 2016 comprise the independent test cohort ($n = 133$; median age, 63 years; IQR, 58–71 years). Inclusion criteria were (a) imaging performed on our main institutional 3-T MRI system and (b) extended systematic and targeted MRI–transrectal US fusion biopsy performed after MRI. Exclusion criteria were (a) history of treatment for prostate cancer (antihormonal therapy, radiation therapy, focal therapy, prostatectomy); (b) biopsy within the past 6 months prior to the MRI examination; and (c) incomplete sequences or severe artifacts on MRI images (Fig 1). Baseline epidemiologic and clinical characteristics, including tumor location, pathologic findings, and clinical assessment, are shown in Table 1.

MRI Examination

MR images at 3 T were acquired prior to biopsy according to the European Society of Urogenital Radiology, or ESUR, guidelines (Magnetom Prisma, Siemens Healthcare, Erlangen, Germany), by using the standard multichannel body coil and integrated spine phased-array coil. T2-weighted, DWI, and dynamic contrast material–enhanced MR images were acquired according to the institutional prostate MRI protocol (Table E1 [online]). Interpretation of multiparametric MRI images was

Table 1: Demographic and Clinical Characteristics of Included Patients

Variable	Training Cohort (<i>n</i> = 183)	Test Cohort (<i>n</i> = 133)
Median age (y)*	64.5 (59–71)	63 (58–71)
Median PSA (ng/mL)*	6.6 (4.9–9.5)	7.5 (5.4–11)
Median PSA density*	0.16 (0.10–0.26)	0.16 (0.11–0.23)
No. of patients without MRI-detected lesions	26	12
No. of patients with MRI-detected lesions†	157 (100)	121 (100)
1 lesion	86 (55)	47 (39)
2 lesions	58 (37)	53 (44)
3 lesions	11 (7)	18 (15)
4 lesions	2 (1)	3 (2)
No. of patients with specified maximum Gleason score†		
No prostate cancer	76 (42)	50 (38)
6 (3+3)	35 (19)	34 (25)
7a (3+4)	49 (27)	31 (23)
7b (4+3)	8 (4)	7 (5)
8 (4+4)	4 (2)	8 (6)
9a (4+5)	7 (4)	2 (2)
9b (5+4)	4 (2)	1 (1)
No. of patients with specified MRI index lesion†		
No lesion	26 (14)	12 (9)
PI-RADS 2	11 (6)	1 (1)
PI-RADS 3	42 (23)	30 (23)
PI-RADS 4	60 (33)	54 (40)
PI-RADS 5	44 (24)	36 (27)
No. of MRI-detected lesions negative for sPC†	163 (67)	159 (73)
No. of MRI-detected lesions positive for sPC†	80 (33)	60 (27)
Peripheral zone	54 (22)	37 (17)
Transition zone	26 (11)	23 (10)
No. of lesions with specified MRI assessment†		
Total	243 (100)	219 (100)
PI-RADS 2	21 (9)	4 (2)
PI-RADS 3	80 (33)	82 (37)
PI-RADS 4	91 (37)	88 (40)
PI-RADS 5	51 (21)	45 (21)
No. of MRI-detected lesions with specified zone distribution†		
Peripheral zone	144 (59)	131 (60)
Transition zone	75 (31)	79 (36)
Anterior fibromuscular stroma	17 (7)	9 (4)
Central zone	7 (3)	0 (0)

Note.—PSA = prostate-specific antigen, PI-RADS = Prostate Imaging Reporting and Data System, sPC = clinically significant prostate cancer.

* Data in parentheses are the interquartile range.

† Data in parentheses are percentages.

performed by board-certified radiologists during clinical routine; eight radiologists, seven of them with at least 3 years of experience in prostate MR image interpretation, read 311 (98%) studies, and one younger colleague who joined the team after completing a departmental training period interpreted five (2%) studies. All examinations were reviewed again in an interdisciplinary conference prior to biopsy for quality assurance and all radiologists participated in regular retrospective review of MRI reports and biopsy results. Clinical reports included PI-RADS assessment for each detected lesion (version 2 for 305 lesions,

version 1 for 157 lesions recorded before adaptation of version 2 in the department) and a pictogram indicating lesion location. For subsequent quantitative analysis, ADC images, diffusion-weighted images with *b* value of 1500 sec/mm², and T2-weighted images were extracted and upsampled to 0.25-mm in-plane resolution and 3-mm section thickness by using the medical imaging toolkit (MITK, www.mitk.org) (17).

MRI Lesion Segmentation

Three-dimensional volumes of interest (VOIs) of clinical lesions were segmented by one investigator (P.S., with 6 months of experience in prostate MRI), using series and section information and pictograms given in the clinical reports in consensus with and under supervision of a board-certified radiologist with 8 years of experience in prostate MRI (D.B.) using MITK, and performed separately on T2-weighted images and ADC images. VOIs were drawn on consecutive axial sections by using the polygon tool, encompassing the whole lesion while avoiding areas of partial volume effects at the border and in regions of diffuse tumor infiltration. A total of 462 lesions were segmented. In addition, background peripheral zone (PZ) was segmented for reference, excluding any lesion and reducing diffuse signal changes to a minimum while encompassing at least 50 voxels on at least three adjacent sections.

Figure 2 depicts example segmentations of PZ, prostate boundary, and a PI-RADS 5 lesion in the anterior transition zone (TZ) and anterior fibromuscular stroma performed in a representative patient with volume renderings.

Image Postprocessing and Analysis

T2-weighted images and images with *b* value of 1500 sec/mm² were normalized by dividing voxel intensities with the mean value of background PZ tissue. ADC, being a quantitative measurement, was not normalized. ADC VOIs were used

for analysis of b value of 1500 sec/mm² images due to the natural coregistration of ADC maps to the source b -value images. Radiomic feature calculations were performed by using the pyradiomics package (<https://github.com/Radiomics/pyradiomics>) (18) according to analysis steps depicted in Figure 3. Within each VOI, (a) 19 first-order features, (b) 16 volume and shape features, and (c) 59 texture features were calculated, leading to 282 features per VOI. Because these features were calculated on the available ADC maps and on T2-weighted images and images with b value of 1500 sec/mm², a total of 846 radiomics features were available for each lesion. The feature set was chosen to represent optimal feature diversity for discovery, while maintaining statistical balance of sample size and data dimensionality. Mean ADC was extracted from the radiomic set for separate analysis. Radiomic features were reduced by univariate feature selection with the remaining features input into random forests (RFs), an approach well-established in radiomics (19). Details on radiomic features and machine learning are given in Appendix E1, sections S-2 and S-3 (online). RF models were computed using scikit-learn (<http://scikit-learn.org/stable/index.html>).

Statistical Analysis

Clinically significant prostate cancer, defined as Gleason grade group 2 or higher (20,21) at histopathologic examination, was used as the standard of reference. Independent variables were radiomic machine learning (RML), mean ADC, and PI-RADS; their relationship to the dependent variable, clinically significant prostate cancer, was evaluated. MRI was considered positive if PI-RADS assessment was greater than or equal to four. Receiver operating characteristic (ROC) curves were generated for mean ADC and RML models, using bootstrap and standard ROC curves in the training and test cohort, respectively, and compared using the Delong test (22). Cut-off values for RML and mean ADC models were selected to match the sensitivity of clinical assessment in the training cohort in order to construct working points of the models that maintain clinically achieved detection rates for significant prostate cancer. The sensitivity and specificity of the models were compared based on the reduction of false-positive (FP) lesions or patients with FP lesions compared with the clinical reference and by expressing this reduction as a ratio with the number of observations and using the McNemar test (23). Correction for multiple comparisons was performed by using the Holm method (24). Thirty-eight patients had no MRI-detected lesions and did not contribute to the lesion-based analysis, which focused entirely on the task of lesion classification. Patient-based

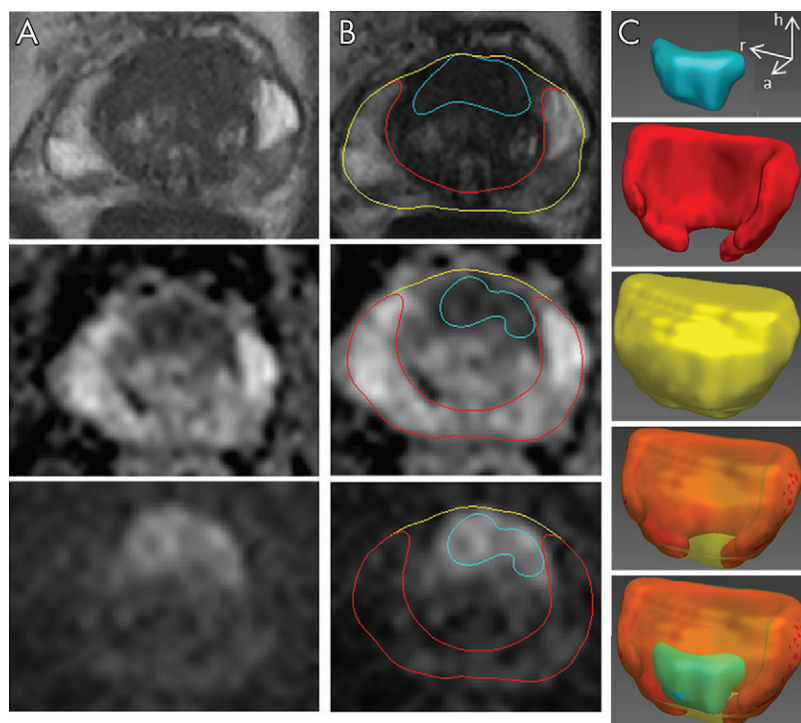


Figure 2: Example segmentations in a 56-year-old patient with initial prostate-specific antigen level of 7 ng/mL. Highly suspicious lesion in the anterior transition zone and anterior stroma (Prostate Imaging Reporting and Data System category 5). A. Top: T2-weighted image demonstrates erased charcoal sign and ill-defined margins; Middle: apparent diffusion coefficient image shows moderate diffusion restriction; Bottom: image with $b = 1500$ sec/mm² shows moderate diffusion restriction. B. Segmentations of prostate (yellow), peripheral zone (red), and suspicious lesion (cyan) overlaid on corresponding images. C. Three-dimensional renderings from top to bottom of lesion, peripheral zone, prostate, prostate combined with peripheral zone, and all volumes of interest combined. Targeted biopsy revealed prostate cancer with a Gleason score of 7 (3+4) in 95% of the four targeted cores. h = head, a = anterior, r = right.

analysis metrics were calculated for the entire cohort to assess the overall combined radiologist lesion detection and model-based lesion classification performance. A global radiomics model was trained on all lesions independent of lesion location (ie, TZ or PZ). For zone-specific performance assessment, previously proposed as advantageous (25), separate radiomics models were trained on TZ and PZ lesions independently and their predictions combined to obtain a performance assessment in the entire cohort. Statistical analysis was implemented in Python (using scikit-learn 0.18.2 and pandas 0.19.2).

Results

Type and Distribution of Clinically Significant Prostate Cancer and MRI-detected Lesions

Of the 183 patients in the training cohort, 26 (14%) had no MRI-detected lesions, while 11 (6%) had PI-RADS 2, 42 (23%) had PI-RADS 3, 60 (33%) had PI-RADS 4, and 44 (24%) had PI-RADS 5 index lesions (Table 1). A total of 243 lesions were recorded, including 21 (9%) PI-RADS 2, 80 (33%) PI-RADS 3, 91 (37%) PI-RADS 4, and 51

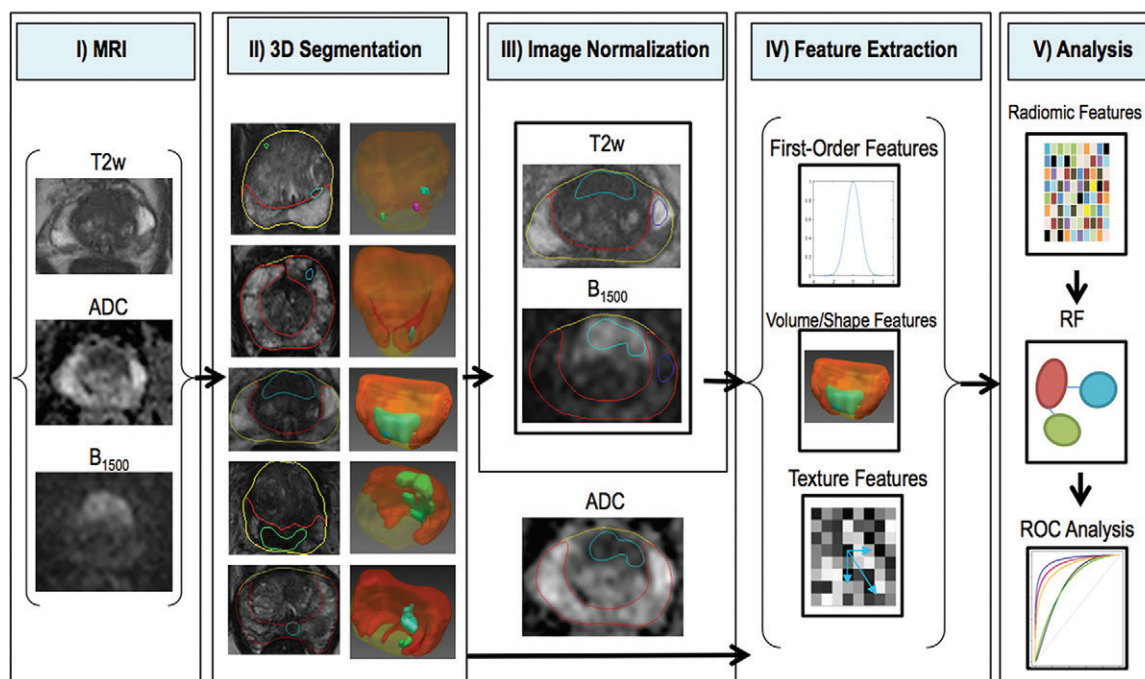


Figure 3: Radiomics workflow. From left to right. Processing follows this sequence: I) T2-weighted (T2w) and diffusion-weighted images (apparent diffusion coefficient [ADC], $b = 1500 \text{ sec/mm}^2$ [B_{1500}]) are extracted from the multiparametric MRI protocol; II) three-dimensional segmentations of lesions, peripheral zone, and prostate are shown in five representative patients overlaid on representative axial T2-weighted images and using volume rendering; III) segmentation of normal-appearing peripheral zone (blue) is used to normalize T2-weighted image and image with $b = 1500 \text{ sec/mm}^2$; IV) radiomic features are extracted, including first-order, volume and shape, and texture features (see text); V) the radiomic feature matrix is combined with clinical information and entered into machine learning analysis (random forest [RF]). Performance is assessed by using receiver operating characteristics (ROC) analysis.

(21%) PI-RADS 5 lesions. Targeted biopsies were positive for clinically significant prostate cancer in 80 (33%) lesions. Of the 133 patients in the test cohort, 12 (9%) had no MRI-detected lesion, while one (1%) had PI-RADS 2, 30 (23%) had PI-RADS 3, 54 (40%) had PI-RADS 4, and 36 (27%) had PI-RADS 5 index lesions. A total of 219 lesions were recorded, including four (2%) PI-RADS 2, 82 (37%) PI-RADS 3, 88 (40%) PI-RADS 4, and 45 (21%) PI-RADS 5 lesions. Targeted biopsies were positive for clinically significant prostate cancer in 60 (27%) lesions.

Radiomics, Mean ADC, and Radiologist Interpretation in Training Cohort

Lesion-based performance.—This cross-validated analysis included 243 MRI-detected lesions (33% of which were positive for clinically significant prostate cancer) found in 157 of 183 patients in the training cohort (Table 1). The classifier area under the ROC curve (AUC) was almost identical for the mean ADC (0.79) and the RML (0.78) (Fig 4a). Model cut-off values for mean ADC were $732 \text{ mm}^2/\text{sec}$ for the global, $760 \text{ mm}^2/\text{sec}$ for PZ, and $714 \text{ mm}^2/\text{sec}$ for TZ models. Model cut-off values for RML were 0.28 for the global, 0.33 for PZ, and 0.19 for TZ models. At the fixed radiologist sensitivity of 79% (63 of 80), the specificity of mean ADC (67% [110 of 163]) and RML (63% [103 of

163]) was improved, compared with 52% (84 of 163) for lesion classification by radiologists (Table 2). In comparison to clinical interpretation by radiologists, measurement of the mean ADC reduced FP lesions by 26 (10.7%) and RML reduced FP lesions by 19 (7.8%), both leaving false-negative (FN) lesions unchanged (Table 2). The top 10 parameters of the RML classifier are given in Table 3.

Patient-based performance.—The cross-validated specificity of the mean ADC (67% [80 of 120]) and RML classifier (62% [74 of 120]) was higher compared to that of radiologist interpretation (57% [68 of 120]). Compared to radiologist interpretation, measurement of the mean ADC reduced the number of patients with FP lesions by 12 and did not reduce the number of patients with FN lesions. RML showed a lower reduction in patients with FP lesions (by six) and an increase by two in the number of patients with FN lesions (Table 4).

Radiomics, Mean ADC, and Radiologist Interpretation in an Independent Test Cohort

Lesion-based performance.—This analysis included 219 MRI-detected lesions (27% of which were positive for prostate cancer) found in 121 of 133 patients of the independent test cohort (Table 1). The classifier AUC was not significantly

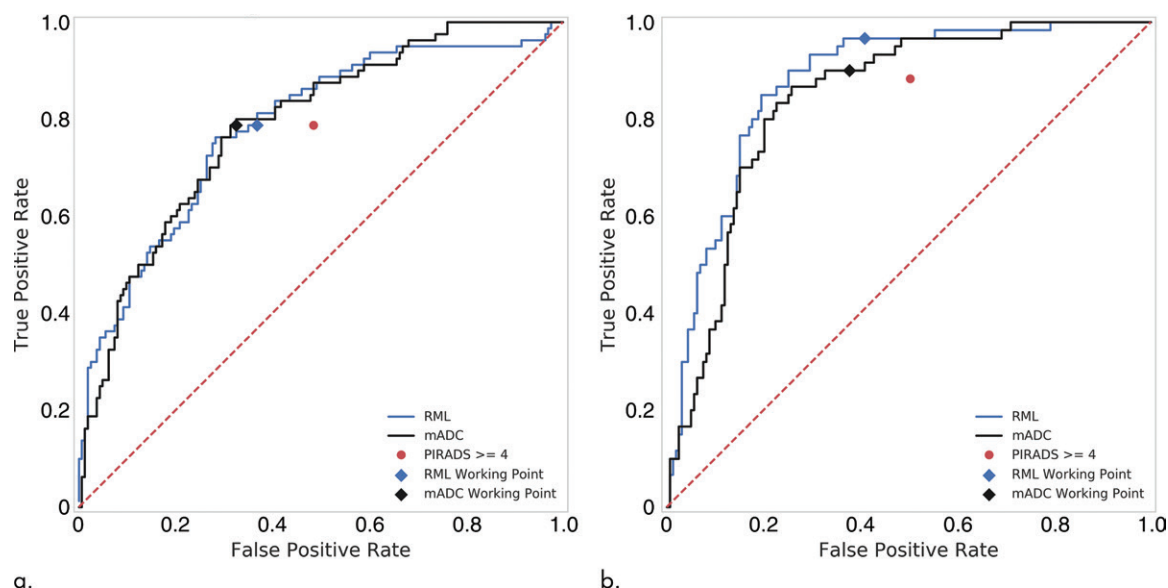


Figure 4: Receiver operating characteristics (ROC) curves, working points of radiomic machine learning (RML), and mean apparent diffusion coefficient (mADC) compared with radiologist performance. Radiologist performance is indicated by a red circle for Prostate Imaging Reporting and Data System (PI-RADS) category of four or greater. The lesion-based radiologist performance shown excludes patients without MRI-detected lesions to focus on the discriminative performance in the decision task (for patient-based performance in the entire cohort, see text). Working points are calibrated to produce matched sensitivity to clinical assessment in the training cohort. ROC curve (solid line) and working points (diamond) for RML are shown in blue, and for mean ADC are shown in black. **(a)** Training cohort: RML and mean ADC model bootstrapped ROC curves are nearly superimposed with clinical assessment showing lower performance. By calibration to clinical sensitivity, both models achieve higher specificity. **(b)** Test cohort: Both RML and mean ADC models generalize without indication for overfitting. Standard ROC curves lie above clinical performance, with RML showing a larger area under the ROC curve. The working point of mean ADC shows a smaller increase of sensitivity; however, a larger increase of specificity compared with the working points of RML. For quantification of differences, see text and Tables 2 and 3.

different for the mean ADC (0.84) versus RML (0.88) ($P = .176$) (Fig 4b). Radiologist interpretation of multiparametric MRI provided a per-lesion sensitivity of 88% (53 of 60) and specificity of 50% (79 of 159). In comparison, measurement of the mean ADC reduced the number of FP lesions from 80 to 60 (specificity, 62% [99 of 159]) and the number of FN lesions from seven to six (sensitivity, 90% [54 of 60]) ($P = .048$). RML reduced the number of FP lesions from 80 to 66 (specificity, 58% [93 of 159]) and the number of FN lesions from seven to two (sensitivity, 97% [58 of 60]) ($P = .176$) (Table 2). Patient examples are shown in Figures 5 and 6.

Patient-based performance.—Radiologist interpretation provided a per-patient sensitivity of 89% (40 of 45) and specificity of 43% (38 of 88). In comparison, measurement of the mean ADC reduced the number of patients with FP lesions from 50 to 43 (specificity, 51% [45 of 88]) and the number of patients with FN lesions from five to three (sensitivity, 93% [42 of 45]) ($P = .496$). The use of RML reduced the number of patients with FP lesions from 50 to 43 (specificity, 51% [45 of 88]) and the number of patients with FN lesions from five to two (sensitivity, 96% [43 of 45]) ($P = .496$) (Table 4).

Radiomics, Mean ADC, and Radiologist Interpretation Using Zone-specific Models

Zone-specific RML performance is shown in Table 2 on a per-lesion basis and in Table 4 on a per-patient basis. In both cases,

the performance was lower than that for the global model. For example, on a per-lesion basis in the test set, sensitivity was 92% (55 of 60) and specificity was 53% (84 of 159), compared with 97% (58 of 60) and 58% (93 of 159), respectively, for the global model. The AUC for RML was 0.84 compared with 0.83 for mean ADC in the PZ ($P = .822$), and 0.89 compared with 0.87, respectively, in the TZ ($P = .493$).

Discussion

Prostate MRI may have the potential to spare patients the discomfort and potential morbidity of biopsies (26), but limitations in the detection of clinically significant prostate cancer with prostate MRI are well known (27). In particular, diagnostic accuracy varies based on the individual radiologist (28–30). Radiomics and quantitative assessment may have potential for decision support (13–15,31). Our results provide an assessment of their ability to aid standardization of MRI interpretation. In the present study, quantitative measurement of the mean ADC, when compared with clinical assessment, is able to significantly reduce the misclassification of MRI-detected lesions. This is in agreement with recent reports of the advantage of quantitative ADC measurement over qualitative clinical assessment in the TZ (10), and also with a recent study in equivocal PI-RADS 3 lesions (11). We found improved lesion classification in both the PZ and TZ, confirming the high value of mean ADC for whole gland assessment. Further, we also investigated the performance of

Table 2: Diagnostic Performance of Radiologist Interpretation, Apparent Diffusion Coefficient, and Radiomic Machine Learning on a Per-Lesion Basis

Cohort and Method	Sensitivity (%) [*]	Specificity (%) [*]	No. of FP Lesions [†]	No. of FN Lesions [†]	Reduction Misclassification [‡]	P Value [§]
Global Model						
Training						
Radiologist	79 (63/80)	52 (84/163)	79/163 (reference)	17/80 (reference)		
mADC	79 (63/80)	67 (110/163)	53/163 (26)	17/80 (0)	26/243 (10.7)	.008 [#]
RML	79 (63/80)	63 (103/163)	60/163 (19)	17/80 (0)	19/243 (7.8)	.035 [#]
Test						
Radiologist	88 (53/60)	50 (79/159)	80/159 (reference)	7/60 (reference)		
mADC	90 (54/60)	62 (99/159)	60/159 (20)	6/60 (1)	21/219 (9.6)	.048
RML	97 (58/60)	58 (93/159)	66/159 (14)	2/60 (5)	20/219 (9.1)	.176
Combined Zone-specific Models						
Training						
Radiologist	79 (63/80)	52 (84/163)	79/163 (reference)	17/80 (reference)		
mADC	79 (63/80)	67 (110/163)	53/163 (26)	17/80 (0)	26/243 (10.7)	.010 [#]
RML	79 (63/80)	58 (94/163)	69/163 (10)	17/80 (0)	10/243 (4.1)	.289 [#]
Test						
Radiologist	88 (53/60)	50 (79/159)	80/159 (reference)	7/60 (reference)		
mADC	92 (55/60)	62 (98/159)	61/159 (19)	5/60 (2)	21/219 (9.6)	.038
RML	92 (55/60)	53 (84/159)	75/159 (5)	5/60 (2)	7/219 (3.2)	.596

Note.—FP = false-positive, FN = false-negative, mADC = mean apparent diffusion coefficient, RML = radiomic machine learning.

^{*} Data in parentheses are raw data.

[†] Data in parentheses show the reduction in the number of lesions compared to the reference value (radiologist method).

[‡] Reduction in FP plus FN lesions compared to the reference, divided by the number of FP plus FN lesions. The number in parentheses is the ratio of FP plus FN reduction, divided by all lesions and expressed as a percentage.

[§] McNemar test for differences in specificity to radiologist performance.

^{||} Statistical significance.

[#] P values derived after bootstrap analysis have to be regarded with caution.

Table 3: Top 10 Parameters with Highest Variable Importance according to (Unscaled) Mean Decrease in Accuracy (Ranking according to Mean Decrease in Node Impurity across Ensemble Members and Splits) for the Global RML Model

Parameter	Gini Index
ADC1500_original_firstorder_Maximum	41.72
ADC1500_original_firstorder_RootMeanSquared	39.41
ADC1500_original_firstorder_Median	36.84
ADC1500_original_firstorder_Mean	33.24
ADC1500_original_firstorder_90Percentile	30.63
ADC1500_original_firstorder_10Percentile	29.54
T2_original_shape_Maximum2DDiameterColumn	14.26
ADC1500_original_firstorder_Minimum	13.73
T2_original_shape_MinorAxis	13.61
T2_original_shape_SurfaceArea	12.37

biparametric contrast-free radiomics with machine learning to assess if there is added value that can be gained from such an approach over measurement of the mean ADC alone. The radiomic models did not perform better than mean ADC, as assessed by ROC. While the performance of mean ADC and radiomics with machine learning was comparable, both methods reduced misclassification more in the per-lesion

than per-patient analysis. This is consistent with a recent report that prospective PI-RADS assessment has excellent performance on a per-patient basis but weaknesses in the identification and characterization of individual lesions (27).

Our findings revealed strengths of ADC measurements when compared with both qualitative clinical assessment and radiomic models. In the radiomic model, all highly ranked features were closely related to the ADC values themselves rather than to textural or morphologic information. Our results refute findings of prior studies (16,32) of an added value of radiomic features and lesion morphology derived from T2-weighted images, which may be explained by differences in study populations and patient selection criteria. The size of our training and the presence and size of the test cohort exceeded that of other radiomics studies for prostate cancer (15,25,33,34), providing an incremental gain in methodology and available data for signature discovery.

In our study, machine learning and radiomics do not surpass the predictive performance of monoparametric ADC assessment. This will potentially change in the future with the development of next-generation machine learning techniques on larger-scale cohorts in multicentric setups as machine learning approaches depend on a large amount of training and test data. Overall radiologist and model performance increased in the test cohort, likely reflecting the

Table 4: Diagnostic Performance of Radiologist Interpretation, Apparent Diffusion Coefficient, and Radiomic Machine Learning on a Per-Patient Basis

Cohort and Method	Sensitivity (%) [*]	Specificity (%) [*]	No. Patients with FP Lesion(s) [†]	No. Patients with FN Lesion(s) [†]	Reduction Misclassification [‡]	P Value [§]
Global Model						
Training						
Radiologist	86 (54/63)	57 (68/120)	52/120 (reference)	9/63 (reference)		
mADC	86 (54/63)	67 (80/120)	40/120 (12)	9/63 (0)	12/183 (6.6)	.180
RML	83 (52/63)	62 (74/120)	46/120 (6)	11/63 (-2)	-4/183 (-2.2)	.429
Test						
Radiologist	89 (40/45)	43 (38/88)	50/88 (reference)	5/45 (reference)		
mADC	93 (42/45)	51 (45/88)	43/88 (7)	3/45 (2)	9/133 (6.8)	.496
RML	96 (43/45)	51 (45/88)	43/88 (7)	2/45 (3)	10/133 (7.5)	.496
Combined Zone-specific Models						
Training						
Radiologist	86 (54/63)	57 (68/120)	52/120 (reference)	9/63 (reference)		
mADC	86 (54/63)	68 (81/120)	39/120 (13)	9/63 (0)	13/183 (7.1)	.134
RML	84 (53/63)	54 (65/120)	55/120 (-3)	10/63 (-1)	-4/183 (-2.2)	.749
Test						
Radiologist	89 (40/45)	43 (38/88)	50/88 (reference)	5/45 (reference)		
mADC	93 (42/45)	51 (45/88)	43/88 (7)	3/45 (2)	9/133 (6.8)	.530
RML	93 (42/45)	38 (33/88)	55/88 (-5)	3/45 (2)	-3/133 (-2.3)	.530

Note.—FP = false-positive, FN = false-negative, mADC = mean apparent diffusion coefficient, RML = radiomic machine learning.

^{*} Data in parentheses are raw data.

[†] Data in parentheses show the reduction in the number of patients compared to the reference value (radiologist method).

[‡] Reduction in FP plus FN lesions compared to the reference, divided by the number of FP plus FN lesions. The number in parentheses is the ratio of FP plus FN reduction, divided by all lesions and expressed as a percentage.

[§] McNemar test for differences in specificity to radiologist performance.

^{||} P values derived after bootstrap analysis have to be regarded with caution.

radiologists' learning curve since the introduction of the PI-RADS version 2 system and, upon review of the clinical images, also attributable to a larger number of patients with small solitary PZ lesions in the training cohort. In this study, all measurements were derived from a single MRI unit with the same protocol. Consecutive patients were included, providing the advantage to test the method with a nonpreselected data set representing the typical difficulties of prostate imaging, such as MRI-detected lesions not caused by clinically significant prostate cancer, which have been excluded in some published approaches (35). We focus on radiologist-informed use of quantitative analysis in lesion differentiation. This approach is different from studies using histopathologically identified lesions as their basis (16) and allows us to directly assess the added benefit of the developed models in the clinical setting. The use of lesions detected during clinical routine has the advantage of targeted core biopsies being available for all lesions and avoiding bias inherent in retrospective assessment.

In our study, radiomics was based on a biparametric MRI protocol, while clinical interpretation used additional evaluation of dynamic contrast-enhanced (DCE) MRI data. The importance of DCE MRI is currently debated (2,3). Biparametric protocols have recently been reported to match the performance of multiparametric MRI including DCE

(36,37). Our single-center/single-MRI unit study design likely benefited our mean ADC findings. Translation of these findings to other sites requires standardization of ADC measurements between MRI units and institutions.

Clinical performance of MRI interpretations with multiple readers in this study may be lower than that of a single expert and may differ from teams at other institutions. However, the approach we used should most directly model the clinical task of prospective decision making in which the model would be used after full validation (27). All lesions were biopsied such that all lesions presented to the learning algorithms are valid learning examples. The ability of the machine learning system to learn from the provided data likely also benefits from increased heterogeneity in MRI lesion detection, thus not becoming specific to a single reader. The lesion-based statistical analysis examined pooled performance and did not consider group effects within patients; however, the biology of cancer is assumed to provide a larger effect than between-patient differences in MRI parameters.

Limitations of our study include the use of radiomics for lesion characterization but not for lesion detection, thus not examining if mean ADC or RML are better than radiologists at cancer detection. Lesions were drawn manually, a time-consuming task that has to be regarded as prohibitive when very large databases are evaluated in the future, requiring the

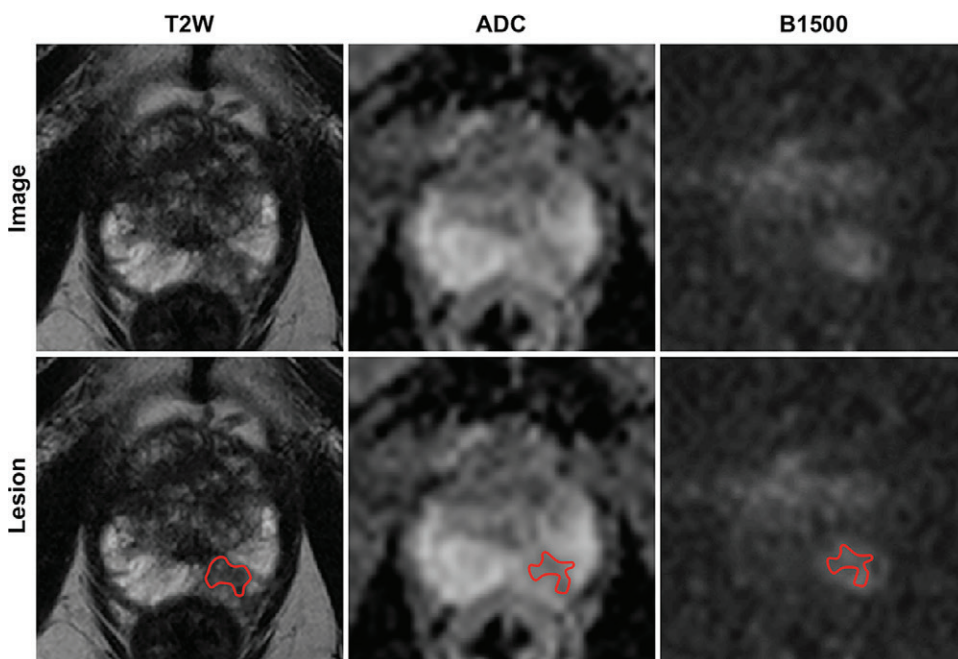


Figure 5: Images in a 74-year-old patient with prostate-specific antigen level of 6.2 ng/mL and negative digital rectal examination. T2-weighted (T2W) and apparent diffusion coefficient (ADC) images and images with b value of 1500 sec/mm² (B1500) are shown in columns, and images without and with the superimposed, outlined segmented lesion are shown in rows. A lesion in the left mid posteromedial peripheral zone is shown, which was read as Prostate Imaging Reporting and Data System category 4. This lesion was rated negative according to a mean ADC value of 895 mm²/sec (cut-off 732 mm²/sec), and it was also below the radiomic machine learning (RML) cut-off, with an RML score of 0.12 (cut-off 0.28). Targeted biopsy revealed no cancer at this location. International Society of Urological Pathology grade 1 prostate cancer was found in systematic cores and a targeted biopsy from the MRI index lesion in the left mid anterior transition zone (not shown).

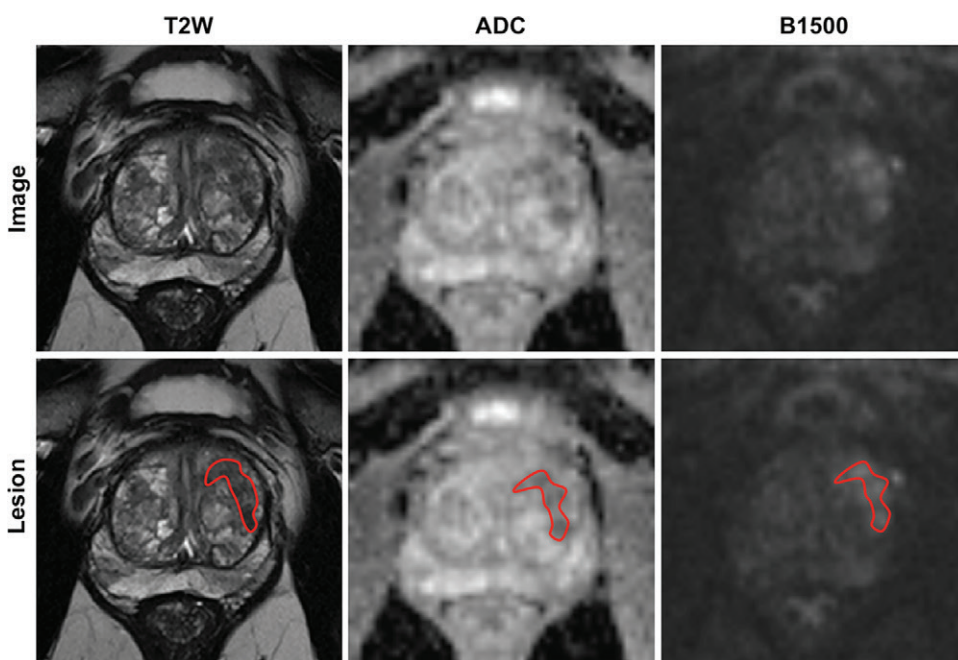


Figure 6: Images in a 65-year-old patient with prostate-specific antigen level of 19.5 ng/mL and negative digital rectal examination. T2-weighted (T2W) and apparent diffusion coefficient (ADC) images and images with b value of 1500 sec/mm² (B1500) are shown in columns, and images without and with the superimposed, outlined segmented lesion are shown in rows. A lesion in the left mid anterior transition zone is shown, which was read as Prostate Imaging Reporting and Data System category 4. This lesion was rated negative according to a mean ADC value of 756 mm²/sec, although it was above the radiomic machine learning (RML) threshold, with an RML score of 0.50. Targeted biopsy revealed no cancer at this location. There was no presence of any cancer in any of the systematic and targeted biopsy cores in this patient.

development of automated segmentation techniques. VOI definitions were performed by a single operator and require standardization for use by different operators. Furthermore, a histopathologic assessment based on saturation and MRI-targeted biopsy rather than radical prostatectomy specimens was used. However, our biopsy approach has been tested against radical prostatectomy as the reference standard and showed a sensitivity of 97% for significant prostate cancer at the final histopathologic examination (5).

To our knowledge, the data presented here comprise the largest data set reported to date in a radiomic analysis involving consecutive patients who are under suspicion for prostate cancer and undergoing a homogeneous protocol on a single MRI unit, followed by systematic saturation and MRI–transrectal US fusion biopsy. We assume that the variability of prostate cancer in the population is diverse enough to have provided us with sufficiently different manifestations of the disease to generate a valid test set, while the controlled technical variability appears beneficial when investigating radiomic signatures at the current cohort sizes. Importantly, inter-institutional and interscanner differences may lead to variations in radiomic parameters. The current data enable a direct comparison with other cohorts that can be formed in the future to answer important questions about such dependencies.

In conclusion, this study compared the use of mean ADC and radiomics with machine learning for the characterization of lesions that were prospectively detected during routine clinical interpretation.

Quantitative assessment of the mean ADC was more accurate than qualitative PI-RADS assessment in classifying a lesion as clinically significant prostate cancer. Radiomics provided additional data that ADC metrics (including mean ADC) were more valuable than other MRI features. In fact, at the current cohort size, no added benefit of the radiomic approach was found, and mean ADC is suggested as the best choice for quantitative prostate assessment.

Author contributions: Guarantors of integrity of entire study, D.B., S.K., H.P.S.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, D.B., M.G., N.G., F.D., M.H., B.A.H., H.P.S., K.H.M.H.; clinical studies, D.B., P.S., B.H., T.A.K., M.F., B.A.H., H.P.S.; experimental studies, K.Y., N.G., T.A.K., K.H.M.H.; statistical analysis, D.B., S.K., M.W., P.S., J.P.R., M.G., P.K., K.H.M.H.; and manuscript editing, D.B., S.K., P.S., J.P.R., M.G., P.K., T.A.K., M.F., M.H., B.A.H., H.P.S., K.H.M.H.

Disclosures of Conflicts of Interest: D.B. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received payment for lectures, including service on speakers' bureaus, from Profound Medical Inc. Other relationships: disclosed no relevant relationships. S.K. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received a paid research internship with Google DeepMind. Other relationships: disclosed no relevant relationships. M.W. disclosed no relevant relationships. P.S. disclosed no relevant relationships. J.P.R. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received payment for consultant work from Saegeling Medizintechnik, Siemens Healthineers and for development of educational presentations from Saegeling Medizintechnik. Other relationships: disclosed no relevant relationships. M.G. disclosed no relevant relationships. P.K. disclosed no relevant relationships. K.Y. Activities related to the present article: disclosed that his salary was paid by a grant from STIFTUNG KREBSFORSCHUNG Europa during the period when he contributed to this study. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. B.H. disclosed no relevant relationships. N.G. disclosed no relevant relationships. T.A.K. disclosed no relevant relationships. E.D. disclosed no relevant relationships. M.F. disclosed no relevant relationships. M.H. disclosed no relevant relationships. B.A.H. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received payment from Janssen R&D; BMS; Sanofi for board membership; payment from Janssen, Bayer, and Profound for lectures, including service on speakers' bureaus; royalties paid to his institution by Uromed; and payment from Pfizer and Astellas for travel/accommodations/meeting expenses. Other relationships: disclosed no relevant relationships. H.P.S. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received payment from Curagita for consultancy and payment from Bayer and Curagita for lectures, including service on speakers' bureaus. Other relationships: disclosed no relevant relationships. K.H.M. disclosed no relevant relationships.

References

- Barentsz JO, Richenberg J, Clements R, et al. ESUR prostate MR guidelines 2012. *Eur Radiol* 2012;22(4):746–757.
- Vargas HA, Hötter AM, Goldman DA, et al. Updated prostate imaging reporting and data system (PI-RADS v2) recommendations for the detection of clinically significant prostate cancer using multiparametric MRI: critical evaluation using whole-mount pathology as standard of reference. *Eur Radiol* 2016;26(6):1606–1612.
- Weinreb JC, Barentsz JO, Choyke PL, et al. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. *Eur Urol* 2016;69(1):16–40.
- Moldovan PC, Van den Broeck T, Sylvester R, et al. What is the negative predictive value of multiparametric magnetic resonance imaging in excluding prostate cancer at biopsy? A systematic review and meta-analysis from the European Association of Urology Prostate Cancer Guidelines Panel. *Eur Urol* 2017;72(2):250–266.
- Radtke JP, Schwab C, Wolf MB, et al. Multiparametric magnetic resonance imaging (MRI) and MRI-transrectal ultrasound fusion biopsy for index tumor detection: correlation with radical prostatectomy specimen. *Eur Urol* 2016;70(5):846–853.
- Donati OF, Mazaheri Y, Afari A, et al. Prostate cancer aggressiveness: assessment with whole-lesion histogram analysis of the apparent diffusion coefficient. *Radiology* 2014;271(1):143–152.
- Donati OF, Afari A, Vargas HA, et al. Prostate MRI: evaluating tumor volume and apparent diffusion coefficient as surrogate biomarkers for predicting tumor Gleason score. *Clin Cancer Res* 2014;20(14):3705–3711.
- Cooperberg MR, Carroll PR. Trends in management for patients with localized prostate cancer, 1990–2013. *JAMA* 2015;314(1):80–82.
- Woo S, Suh CH, Kim SY, Cho JY, Kim SH. Diagnostic performance of Prostate Imaging Reporting and Data System version 2 for detection of prostate cancer: a systematic review and diagnostic meta-analysis. *Eur Urol* 2017;72(2):177–188.
- Pierre T, Cornud F, Colléter L, et al. Diffusion-weighted imaging of the prostate: should we use quantitative metrics to better characterize focal lesions originating in the peripheral zone? *Eur Radiol* 2018;28(5):2236–2245.
- Hansen NL, Koo BC, Warren AY, Kastner C, Barrett T. Sub-differentiating equivocal PI-RADS-3 lesions in multiparametric magnetic resonance imaging of the prostate to improve cancer detection. *Eur J Radiol* 2017;95:307–313.
- Chatterjee A, Watson G, Myint E, Sved P, McEntee M, Bourne R. Changes in epithelium, stroma, and lumen space correlate more strongly with Gleason pattern and are stronger predictors of prostate ADC changes than cellularity metrics. *Radiology* 2015;277(3):751–762.
- 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(46):E6265–E6273.
- Vignati A, Mazzetti S, Giannini V, et al. Texture features on T2-weighted magnetic resonance imaging: new potential biomarkers for prostate cancer aggressiveness. *Phys Med Biol* 2015;60(7):2685–2701.
- Wang J, Wu CJ, Bao ML, Zhang J, Wang XN, Zhang YD. Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer. *Eur Radiol* 2017;27(10):4082–4090.
- Wibmer A, Hricak H, Gondo T, et al. Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. *Eur Radiol* 2015;25(10):2840–2850.
- Nolden M, Zelzer S, Seitel A, et al. The Medical Imaging Interaction Toolkit: challenges and advances: 10 years of open-source development. *Int J CARS* 2013;8(4):607–620.
- van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 2017;77(21):e104–e107.
- Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep* 2015;5(1):13087.
- Carter HB, Partin AW, Walsh PC, et al. Gleason score 6 adenocarcinoma: should it be labeled as cancer? *J Clin Oncol* 2012;30(35):4294–4296.
- Loeb S, Folkvaljon Y, Robinson D, Lissbrant IF, Egevad L, Stattin P. Evaluation of the 2015 Gleason grade groups in a nationwide population-based cohort. *Eur Urol* 2016;69(6):1135–1141.
- 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(3):837–845.
- McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947;12(2):153–157.
- Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979;6(2):65–70.
- Ginsburg SB, Algohary A, Pahwa S, et al. Radiomic features for prostate cancer detection on MRI differ between the transition and peripheral zones: preliminary findings from a multi-institutional study. *J Magn Reson Imaging* 2017;46(1):184–193.
- Ahmed HU, El-Shater Bosaily A, Brown LC, et al. Diagnostic accuracy of multiparametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet* 2017;389(10071):815–822.
- Borofsky S, George AK, Gaur S, et al. What are we missing? False-negative cancers at multiparametric MR imaging of the prostate. *Radiology* 2018;286(1):186–195.
- Greer MD, Brown AM, Shih JH, et al. Accuracy and agreement of PI-RADSv2 for prostate cancer mpMRI: a multireader study. *J Magn Reson Imaging* 2017;45(2):579–585.
- Rosenkrantz AB, Ayoola A, Hoffman D, et al. The learning curve in prostate MRI interpretation: self-directed learning versus continual reader feedback. *AJR Am J Roentgenol* 2017;208(3):W92–W100.
- Rosenkrantz AB, Babb JS, Taneja SS, Ream JM. Proposed adjustments to PI-RADS Version 2 decision rules: impact on prostate cancer detection. *Radiology* 2017;283(1):119–129.
- Sonn GA, Fan RE, Ghanouni P, et al. Prostate magnetic resonance imaging interpretation varies substantially across radiologists. *Eur Urol Focus* 2017;17:S2405–S2459.
- Nketiah G, Elschot M, Kim E, et al. T2-weighted MRI-derived textural features reflect prostate cancer aggressiveness: preliminary results. *Eur Radiol* 2017;27(7):3050–3059.
- Cameron A, Khalvati F, Haider MA, Wong A. MAPS: a quantitative radiomics approach for prostate cancer detection. *IEEE Trans Biomed Eng* 2016;63(6):1145–1156.
- Khalvati F, Wong A, Haider MA. Automated prostate cancer detection via comprehensive multi-parametric magnetic resonance imaging texture feature models. *BMC Med Imaging* 2015;15(1):27.
- Kwak JT, Sankineni S, Xu S, et al. Prostate cancer: a correlative study of multiparametric MR imaging and digital histopathology. *Radiology* 2017;285(1):147–156.
- Kuhl CK, Bruhn R, Krämer N, Nebelung S, Heidenreich A, Schrading S. Abbreviated biparametric prostate MR imaging in men with elevated prostate-specific antigen. *Radiology* 2017;285(2):493–505.
- Radtke JP, Boxler S, Kuru TH, et al. Improved detection of anterior fibromuscular stroma and transition zone prostate cancer using biparametric and multiparametric MRI with MRI-targeted biopsy and MRI-US fusion guidance. *Prostate Cancer Prostatic Dis* 2015;18(3):288–296.