Title

Multiparametric MRI of the prostate: diagnostic performance and interreader agreement of two scoring systems

Journal

British Journal of Radiology, 89(1062)

ISSN

0007-1285

Authors

Lin, Wei-Ching
Muglia, Valdair F
Silva, Gyl EB
et al.

Publication Date

2016-06-01

DOI

10.1259/bjr.20151056

Peer reviewed

BJR

Received: 16 December 2015 | Revised: 24 February 2016 | Accepted: 21 March 2016

http://dx.doi.org/10.1259/bjr.20151056

Cite this article as:
Lin W-C, Muglia VF, Silva GEB, Chodraui Filho S, Reis RB, Westphalen AC. Multiparametric MRI of the prostate: diagnostic performance and interreader agreement of two scoring systems. *Br J Radiol* 2016; **89**: 20151056.

FULL PAPER

Multiparametric MRI of the prostate: diagnostic performance and interreader agreement of two scoring systems

[1,2]WEI-CHING LIN, MD, [3]VALDAIR F MUGLIA, MD, PhD, [4]GYL E B SILVA, MD, PhD, [3]SALOMÃO CHODRAUI FILHO, MD, [5]RODOLFO B REIS, MD, PhD and [6]ANTONIO C WESTPHALEN MD, PhD

[1]Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA
[2]Department of Radiology, School of Medicine, China Medical University, Tai Chung City, Central Taiwan, Taiwan
[3]Division of Radiology, Department of Internal Medicine, Ribeirão Preto School of Medicine, University of São Paulo, Ribeirão Preto, São Paulo, Brazil
[4]Department of Pathology, Ribeirão Preto School of Medicine, University of São Paulo, Ribeirão Preto, São Paulo, Brazil
[5]Division of Urology, Department of Surgery and Anatomy, Ribeirão Preto School of Medicine, University of São Paulo, Ribeirão Preto, São Paulo, Brazil
[6]Departments of Radiology and Biomedical Imaging, and Urology, University of California, San Francisco, CA, USA

Address correspondence to: Dr Antonio Carlos Westphalen
E-mail: *antonio.westphalen@ucsf.edu*

Objective: To compare the diagnostic accuracies and interreader agreements of the Prostate Imaging Reporting and Data System (PI-RADS) v. 2 and University of California San Francisco (UCSF) multiparametric prostate MRI scale for diagnosing clinically significant prostate cancer.
Methods: This institutional review board-approved retrospective study included 49 males who had 1.5 T endorectal MRI and prostatectomy. Two radiologists scored suspicious lesions on MRI using PI-RADS v. 2 and the UCSF scale. Percent agreement, 2 × 2 tables and the area under the receiver operating characteristic curves (Az) were used to assess and compare the individual and overall scores of these scales. Interreader agreements were estimated with kappa statistics.
Results: Reader 1 (R1) detected 78 lesions, and Reader 2 (R2) detected 80 lesions. Both identified 52 of 65 significant cancers. The Az for PI-RADS v. 2 and UCSF scale for R1 were 0.68 and 0.69 [T_2 weighted imaging (T2WI)], 0.75 and 0.68 [diffusion-weighted imaging (DWI)] and 0.64 and 0.72 (overall score), respectively, and were 0.72 and 0.75 (T2WI), 0.73 and 0.67 (DWI) and 0.66 and 0.75 (overall score) for R2. The dynamic contrast-enhanced percent agreements between scales were 100% (R1) and 95% (R2). PI-RADS v. 2 DWI of R1 performed better than UCSF DWI (Az = 0.75 *vs* Az = 0.68; $p = 0.05$); no other differences were found. The interreader agreements were higher for PI-RADS v. 2 (T2WI: 0.56 *vs* 0.42; DWI: 0.60 *vs* 0.46; overall: 0.61 *vs* 0.42). The UCSF approach to derive the overall PI-RADS v. 2 scores increased the Az for the identification of significant cancer (R1 to 0.76, $p < 0.05$; R2 to 0.71, $p = 0.35$).
Conclusion: Although PI-RADS v. 2 DWI score may have a higher discriminatory performance than the UCSF scale counterpart to diagnose clinically significant cancer, the utilization of the UCSF scale weighing system for the integration of PI-RADS v. 2 individual parameter scores improved the accuracy its overall score.
Advances in knowledge: PI-RADS v. 2 is moderately accurate for the identification of clinically significant prostate cancer, but the utilization of alternative approaches to derive the overall PI-RADS v. 2 score, including the one used by the UCSF system, may improve its diagnostic accuracy.

INTRODUCTION

Multiparametric MRI plays an incremental role in the detection, characterization and management of prostate cancer, including assistance in guiding biopsies,[1] treatment planning and patient selection for active surveillance,[2,3] guidance of focal therapies[4,5] and assessment of post-treatment effects.[6–8] However, the interpretation of multiparametric MRI remains difficult and with substantial interreader variability.[9]

In 2012, the European Society of Urogenital Radiology (ESUR) proposed the Prostate Imaging Reporting and Data System (PI-RADS)[10] in an attempt to standardize scanning protocols and diagnostic criteria for multiparametric MRI. Various studies have evaluated this scoring system and suggested it improves imaging interpretation and diagnostic accuracy;[11–13] yet, changes were recommended to further enhance it.[14]

Because of the limitations of this initial version of PI-RADS, some institutions have used other systems.[15] A modified version of the ESUR PI-RADS is utilized at the University of California, San Francisco (UCSF). The main differences between the two scoring systems are the

diagnostic criteria for diffusion-weighted imaging (DWI), dynamic contrast-enhanced (DCE) MRI and introduction of a guide to integrate individual scores and assign a five-point overall score (Table 1).

More recently, a new version of PI-RADS was announced (PI-RADS v. 2).[16] This second version, proposed by the American College of Radiology, ESUR and AdMeTech Foundation, corrects the shortcomings of the previous one. More specifically, it gives more precise definitions to each score and sequences and clearly explains how to derive a five-point overall score.[16] These changes are likely to reduce variability in imaging interpretations, enhance communication with referring clinicians and facilitate appropriate management of prostate cancer.

Table 1. The University of California, San Francisco, scoring system for multiparametric MRI of the prostate

Score	Criteria
T2WI for the peripheral zone	
1	Homogeneous high SI
2	Streaky, triangular, geographic areas of low SI
3	Intermediate appearances not in categories 1/2 or 4/5
4	Discrete homogeneous low SI, confined to the prostate
5	Same as 4 but with extracapsular extension, or ≥1.5 cm in greatest dimension contact with the surface
T2WI for the transition zone	
1	Heterogeneous SI with well-defined margins: "organized chaos"
2	More homogeneous low SI focus with distinct margins
3	Intermediate appearances not in categories 1/2 or 4/5
4	More homogeneous low SI areas with burred borders: "erased charcoal sign"
5	Same as 4 but with other component invasion; ≥1.5 cm in greatest dimension
DWI	
1	No reduction on ADC compared with normal glandular tissue. No increase in SI on high *b*-value DWI
2	High SI on high *b*-value DWI but no reduction on ADC
3	Low or iso SI on high *b*-value DWI and low SI on ADC
4	High SI on high *b*-value DWI with low SI on ADC but the ADC value $>850 \times 10^{-6}$ mm^2 s^{-1}
5	High SI on high *b*-value DWI with low SI on ADC and the ADC value $\leq 850 \times 10^{-6}$ mm^2 s^{-1}
DCE	
Positive	Focal, asymmetric lesion with fast washin
Negative	Diffuse lesion with any kind of enhancement pattern or progressive enhancement of focal lesion
Overall	DWI score + T2WI score/2; round up the average score if DCE is positive

ADC, apparent diffusion coefficient; DCE, dynamic contrast material–enhanced imaging; DWI, diffusion-weighted imaging; SI, signal intensity; T2WI, T_2 weighted imaging.

Compared with the UCSF scale, the PI-RADS v. 2 is more subjective and possibly more generalizable, as its overall score is mostly dependent on one sequence, DWI for peripheral zone (PZ) tumours or T_2 weighted for transition zone (TZ) ones. However, the accuracy and interreader agreement of PI-RADS v. 2 has not yet been definitely established, although one recent study suggests it is moderately reproducible for detection of clinically relevant disease.[17] Although PI-RADS v. 2 might indeed be a very good way of interpreting multiparametric MRI of the prostate, other systems may demonstrate better performance. If so, these should not be entirely discarded as they could be better for serving a subset of patients. Furthermore, elements of these schemes could be taken into consideration for future PI-RADS updates. Accordingly, the aim of this study was to compare the diagnostic accuracies and interreader agreements of the PI-RADS v. 2 and UCSF multiparametric prostate MRI scales.

METHODS AND MATERIALS

Patients

The institutional review board of the Ribeirão Preto School of Medicine, Ribeirão Preto, São Paulo, Brazil, approved this retrospective study with a waiver of the written informed consent. Between May 2011 and June 2014, 192 males with biopsy-proven prostate cancer underwent multiparametric MRI for staging purposes. All scans were performed between 6 and 9 weeks after transrectal ultrasound-guided biopsy to avoid post-biopsy haemorrhage. 54 of these patients underwent radical prostatectomy after multiparametric MRI and were eligible for this study. The time interval between MRI and prostatectomy was less than 6 months. Only patients without interval prostate cancer treatment between MRI and prostatectomy were included in this study. No other inclusion criteria were applied. Five patients were excluded; one patient received radiation therapy before MRI and four MRI scans were incomplete and therefore not interpretable. A total of 49 patients were included in our study. The median patient age was 63 years (range, 46–73 years). The median serum prostate-specific antigen level at diagnosis was 13.27 ng ml^{-1} (range, 1.75–41.40 ng ml^{-1}). The median Gleason score at biopsy was 7 (range, 5–8). Most males had clinically localized disease: T1c = 7, T2a = 15, T2b = 12, T2c = 6, T3a = 9.

MRI acquisition

MRI was performed on a 1.5-T MRI scanner (Achieva®; Philips Healthcare, Best, Netherlands). A five-channel phased-array surface coil combined with a balloon-covered expandable endorectal coil (Medrad; Bayer Healthcare, Warrendale, PA) was used. The scanning protocol included T_2 weighted imaging (T2WI), DWI and DCE MRI. DWI was acquired with *b*-values of 0 and 1000 s^{-1} mm^{-2} and with an inline reconstruction of apparent diffusion coefficient (ADC) map. DCE MRI of the prostate was performed following administration of 0.1 mmol of gadopentetate dimeglumine (Magnevist®; Bayer HealthCare Pharmaceuticals,

Montville, NJ) per kilogram of body weight followed by a 20 ml saline flush at a rate of 3 ml s^{-1}. The temporal resolution ranged between 8 and 22 s, with 80% of patients having scans obtained with a temporal resolution of ≤10 s. Details of the imaging acquisition protocol are given in Appendix A.

Image interpretation

Two readers from two other institutions, Reader 1 (R1: ACW) and Reader 2 (R2: W-CL) with 12 and 5 years' of experience in prostate MRI, interpreted images independently on a picture archiving and communication system workstation (Infinitt Healthcare, Phillipsburg, NJ). Readers knew that patients had biopsy-proven prostate cancer and underwent radical prostatectomy but were unaware of any other clinical data or histopathological results. Readers reviewed all sequences on a single session. No post-biopsy haemorrhage was seen on T_1 weighted MR images. Readers identified the most suspicious prostate lesions (maximum three lesions per patient), recording each lesion's bidimensional size in millimeters and its location on a 15-region diagram, as proposed by Dickinson et al.[18] Readers also assigned scores to all detected lesions by using the PI-RADS v. 2[16] and UCSF scales (Table 1).

As shown in Table 1, the UCSF T2WI and DCE criteria are similar to that of the PI-RADS v. 2, but those of DWI are modified. If a focal lesion shows decreased signal intensity on the ADC map, the given score will be ≥3. If the low signal intensity focal lesion found on ADC map displays high signal intensity on high *b*-value DWI, it will be categorized as Score 4 or 5; those lesions with a mean ADC value $<850\times10^{-6}$ mm^2 s^{-1} are assigned Score 5. The option for this threshold is based on previous data that determined the mean ADC values of tumours with Gleason Pattern 4 and showed an inverse relationship between ADC values and Gleason scores.[19–24] Although not part of the DWI scoring criteria of PI-RADS v. 2, the use of an ADC threshold of $750–900\times10^{-6}$ mm^2 s^{-1} is described in its publication as a possible adjunct feature that correlates with clinically significant cancers,[16] and ADC value cut-off utilized by the UCSF scale is within this range. The mean ADC value of a lesion was measured on the ADC map, utilizing a region of interest that was drawn to occupy approximately 75% of its diameter.

Another difference between the two systems is how the overall scores are determined. The overall UCSF suspicion score is given by the formula: (DWI + T2WI)/2. The overall score is rounded up if DCE MRI is positive or down when it is negative. In PI-RADS v. 2, DWI is considered as dominant in assessing PZ tumours, whereas T2WI is the primary determinant in evaluating TZ lesions. The overall score is equal to the score of the dominant sequence, except for a lesion that has Score 3 on the dominant sequence. When a PZ lesion has a score of 3 on DWI, a positive finding on DCE MRI will increase the overall Score 4, whereas a negative finding on DCE MRI will keep it as 3. When a TZ lesion receives a Score 3 on T2WI, a score of 5 on DWI will increase the overall score to 4, whereas a DWI score of ≤4 will keep the overall score at 3.

Standard of reference

A single genitourinary pathologist (GEBS, 11 years' of experience), blinded to clinical information, reviewed standard step-section slides from radical prostatectomy and recorded the size, location and Gleason score of all cancer foci with volume >0.5 cm^3 on a standardized map of the prostate. Clinically significant prostate cancer was characterized based on the definition described in the PI-RADS v. 2 publication—a tumour with volume >0.5 cm^3 with primary or secondary Gleason Pattern 4 or 5.[16] Histopathological tumour volumes were estimated using the formula for tumour volume of $(4/3)\pi(D/2)^3$, where D is the average of the maximum and minimum axial diameters of the tumour obtained from the slide demonstrating the maximum tumour area.[25] Using anatomic landmarks and tumour laterality, size and location, one investigator (VFM), not involved in the review of MR images, compared the histopathological tumour maps with the standardized maps generated by the MRI readers to determine which visualized lesions were true-positive and false-positive findings.

Post hoc analysis

After our initial analyses, which were planned prior to data collection, we noted that in spite of a better performance of the PI-RADS v. 2 DWI score for Reader 1, the diagnostic accuracy of the UCSF overall score was higher than that of the PI-RADS v. 2 overall score. Based on this finding, we hypothesized that deriving the PI-RADS v. 2 overall score utilizing the UCSF method, *i.e.* averaging the T_2 and DWI scores and rounding the mean up or down based on the results of DCE, could improve the diagnostic accuracy of the overall score of PI-RADS v. 2.

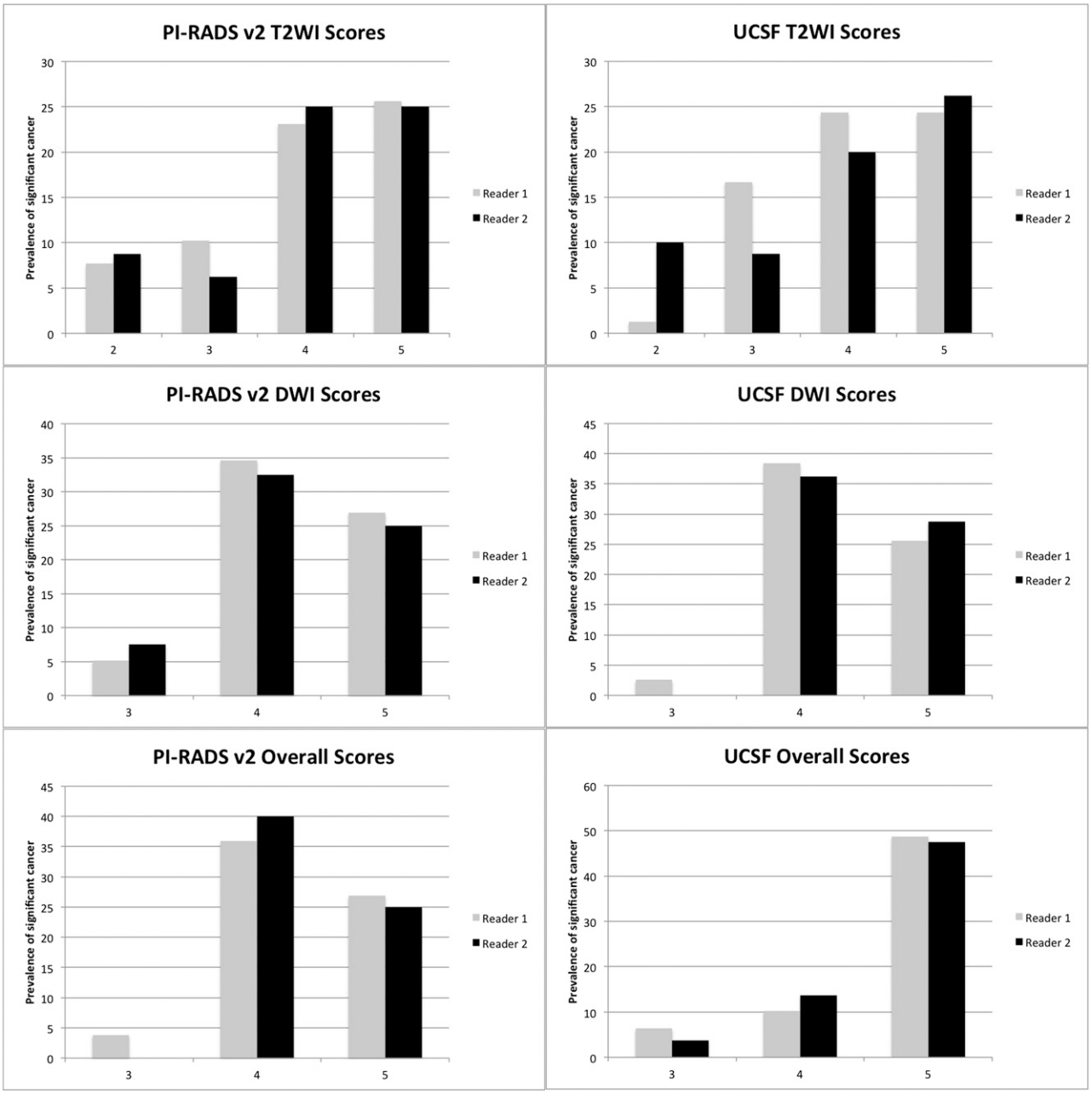
Statistical analysis

Analyses were performed for each reader and were lesion based. We calculated the prevalence of clinically significant cancers among the presumed tumours for the T2WI, DWI and the overall scores of each scoring system. The sensitivity, specificity, accuracy, positive-predictive value and negative-predictive value were calculated using a cut-off value of Score 3, 4 or 5 for T2WI, DWI and overall scores and using positive finding on DCE. Receiver operating characteristic curve analysis was used to assess the diagnostic performance of the individual sequences and overall scores of both scoring systems; and the equality of the areas under the receiver operating characteristic curves (Az) analyzed using the method proposed by DeLong et al.[26] The agreement of DCE findings between scoring systems was also calculated.

The interreader agreement was calculated using custom-weighted kappa statistics that gave 5° different weighting from 0 to 1 for each disagreement for T2WI, DWI and an overall score of each scoring system (Appendix B). We made this option because, generally, a score of 3, 4 or 5 will lead to a biopsy, but scores of 1 and 2 do not. In addition, the Scores 4 and 5 imply different outcomes because a Score 5 lesion may be associated with extraprostatic extension. We used bootstrapping to construct bias-corrected 95% confidence intervals (CIs). The interreader agreement was defined excellent ($\kappa>0.81$), good ($\kappa=0.61–0.80$), moderate ($\kappa=0.41–0.60$), fair ($\kappa=0.21–0.40$) and poor ($\kappa\leq0.20$).[27] The percent agreement between the readers was calculated for DCE MRI of each scoring system.

Statistical analyses were performed using IBM SPSS® Statistics 20 for Windows (IBM Corp., New York, NY; formerly SPSS Inc., Chicago, IL) and STATA®/IC 13.1 for Mac (StataCorp, College Station, TX). Statistical significance was defined as a p-value <0.05.

Figure 1. Prevalence of clinically significant prostate cancer by T_2 weighted imaging (T2WI), diffusion-weighted imaging (DWI) and overall Prostate Imaging Reporting and Data System (PI-RADS) v. 2 and University of California San Francisco (UCSF) scores for both readers.



RESULTS

Pathological findings

A total of 74 cancers with volume >0.5 cm^3 were identified. The number of the cancers of each Gleason score was as follows: 3 + 3 (n = 9/13.8%), 3 + 4 (n = 21/32.3%), 4 + 3 (n = 30/46.2%), 4 + 4 (n = 6/9.2%), 5 + 3 (n = 2/3.1%) and 5 + 4 (n = 6/9.2%). Accordingly, 65 clinically significant cancers were identified. Of these, 63 were involved only or predominantly in the PZ and 2 were involved only or predominantly in the TZ. The average diameter of these clinically significant cancers was 18.6 mm (range, 10–42 mm) on histopathology.

Tumour detection

R1 detected 78 suspicious foci on MRI, and R2 detected 80 suspicious foci on MRI. Both readers identified 52 of 65 (80.0%) clinically significant cancers. Neither reader identified suspicious MRI findings in one patient who had a single tumour that was consistent with a clinically significant cancer on histology. For both readers T2WI, DWI and overall scores of both scales showed a tendency to a higher prevalence of cancer at higher scores (Figure 1).

Diagnostic accuracy of the PI-RADS v. 2 and UCSF scale

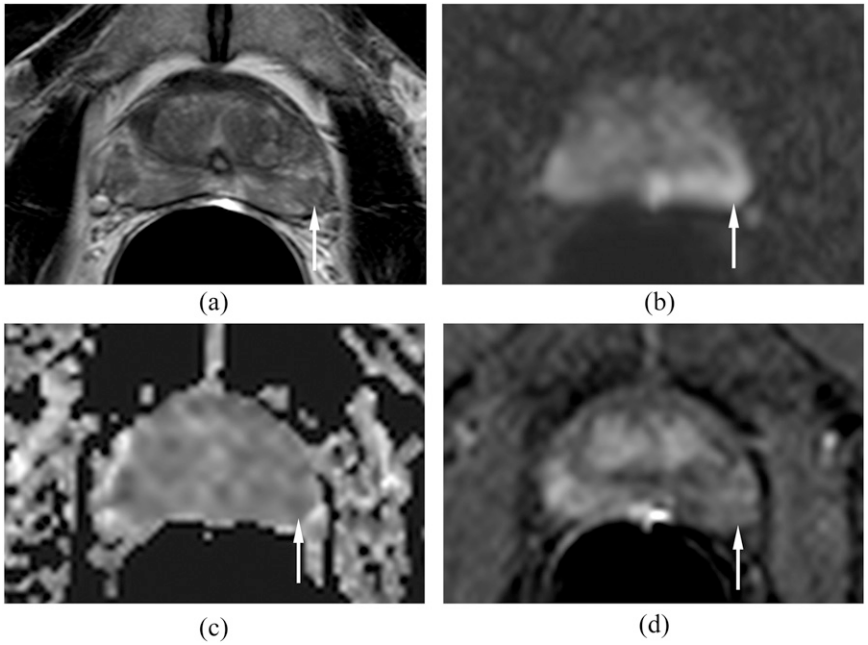
Table 2 presents the sensitivity, specificity, accuracy, positive-predictive value and negative-predictive value for the diagnosis of clinically significant cancer at cut-off values of 3, 4 and 5 for T2WI, DWI and overall score, and positive finding for DCE of both scales for both readers. For both scales, the highest specificity (>85.7%) is seen with individual T2WI and DWI scores, using

Table 2. Diagnostic accuracies of the University of California, San Francisco (UCSF) scale and Prostate Imaging Reporting and Data System (PI-RADS) v. 2 for identification of clinically significant prostate cancer

Scoring system	Score	Reader 1					Reader 2				
		Sensitivity	Specificity	Accuracy	PPV	NPV	Sensitivity	Specificity	Accuracy	PPV	NPV
UCSF											
T2WI	3	98.08 (51/52)	15.38 (4/26)	70.51 (55/78)	69.86 (51/73)	80.00 (4/5)	84.62 (44/52)	46.43 (13/28)	71.25 (57/80)	74.58 (44/59)	61.90 (13/21)
	4	73.08 (38/52)	53.85 (14/26)	66.67 (52/78)	76.00 (38/50)	50.00 (14/28)	71.15 (37/52)	67.86 (19/28)	70.00 (56/80)	80.43 (37/46)	55.88 (19/34)
	5	36.54 (19/52)	88.46 (23/26)	53.85 (42/78)	86.36 (19/22)	41.07 (23/56)	40.38 (21/52)	92.86 (26/28)	58.75 (47/80)	91.30 (21/23)	45.61 (26/57)
DWI	3	100.00 (52/52)	0.00 (0/26)	66.67 (52/78)	66.67 (52/78)	–	100.00 (52/52)	0.00 (0/28)	65.00 (52/80)	65.00 (52/80)	–
	4	96.15 (50/52)	19.23 (5/26)	70.51 (55/78)	70.42 (50/71)	71.43 (5/7)	100.00 (52/52)	7.14 (2/28)	67.50 (54/80)	66.67 (52/78)	100.00 (2/2)
	5	38.46 (50/52)	88.46 (23/26)	55.13 (43/78)	86.96 (20/23)	41.82 (23/55)	44.23 (23/52)	85.71(24/28)	58.75 (47/80)	85.19 (23/27)	45.28 (24/53)
DCE	Positive	86.54 (45/52)	34.62 (9/26)	69.23 (54/78)	72.58 (45/62)	56.25 (9/16)	92.31 (48/52)	28.57 (8/28)	70.00 (56/80)	70.59 (48/68)	66.67 (8/12)
Overall	3	98.08 (51/52)	3.85 (1/26)	66.67 (52/78)	67.11 (51/76)	50.00 (1/2)	100.00 (52/52)	0.00 (0/28)	65.00 (52/80)	65.00 (52/80)	–
	4	88.46 (46/52)	23.08 (6/26)	66.67 (52/78)	69.70 (46/66)	50.00 (6/12)	94.23 (49/52)	25.00 (7/28)	70.00 (56/80)	70.00 (49/70)	70.00 (7/10)
	5	73.08 (38/52)	73.08 (19/26)	73.08 (57/78)	84.44 (38/45)	57.58 (19/33)	73.08 (38/52)	75.00 (21/28)	73.75 (59/80)	84.44 (38/45)	60.00 (21/35)
PI-RADS v. 2											
T2WI	3	88.46 (46/52)	30.77 (8/26)	69.23 (54/78)	71.88 (46/64)	57.14 (8/14)	86.5 (45/52)	32.14 (9/28)	67.50 (54/80)	70.31 (45/64)	56.25 (9/16)
	4	73.08 (38/52)	50 (13/26)	65.38 (51/78)	74.51 (38/51)	48.15 (13/27)	76.92 (40/52)	57.14 (16/28)	70.00 (56/80)	76.92 (40/52)	57.14 (16/28)
	5	38.46 (20/52)	88.46 (23/26)	55.13 (43/78)	86.96 (20/23)	41.82 (23/55)	38.46 (20/52)	92.86 (26/28)	57.50 (46/80)	90.91 (20/22)	44.83 (26/58)
DWI	3	100.00 (52/52)	0.00 (0/26)	66.67 (52/78)	66.67 (52/78)	–	100.00 (52/52)	0.00 (0/28)	65.00 (52/80)	65.00 (52/80)	–
	4	92.31 (48/52)	42.31 (11/26)	75.64 (59/78)	76.19 (48/63)	73.33 (11/15)	88.46 (46/52)	42.86 (12/28)	72.50 (58/80)	74.19 (46/62)	66.67 (12/18)
	5	40.38 (21/52)	92.31 (24/26)	57.69 (45/78)	91.30 (21/23)	43.64 (24/55)	38.46 (20/52)	92.86 (26/28)	57.50 (46/80)	90.91 (20/22)	44.83 (26/58)
DCE	Positive	86.54 (45/52)	34.62 (9/26)	69.23 (54/78)	72.58 (45/62)	56.25 (9/16)	96.15 (50/52)	21.43 (6/28)	70.00 (56/80)	69.44 (50/72)	75.00 (6/8)
Overall	3	100.00 (52/52)	0.00 (0/26)	66.67 (52/78)	66.67 (52/78)	–	100.00 (52/52)	0.00 (0/28)	65.00 (52/80)	65.00 (52/80)	–
	4	94.23 (49/52)	7.69 (2/26)	65.38 (51/78)	67.12 (49/73)	40.00 (2/5)	100.00 (52/52)	7.14 (2/28)	67.50 (54/80)	66.67 (52/78)	100.00 (2/2)
	5	40.38 (21/52)	88.46 (23/26)	56.41 (44/78)	87.50 (21/24)	42.59 (23/54)	38.46 (20/52)	89.29 (25/28)	56.25 (45/80)	86.96 (20/23)	43.86 (25/57)
Modified overall PI-RADS[a]											
	3	98.08 (51/52)	3.85 (1/26)	66.67 (52/78)	67.11 (51/76)	50.00 (1/2)	100.00 (52/52)	7.14 (2/28)	67.5 (54/80)	66.67 (52/78)	100.00 (2/2)
	4	86.54 (45/52)	34.62 (9/26)	69.23 (54/78)	72.58 (45/62)	56.25 (9/16)	96.15 (50/52)	28.57 (8/28)	72.5 (58/80)	71.43 (50/70)	80.00 (8/10)
	5	73.08 (38/52)	80.77 (21/26)	75.64 (59/78)	88.37 (38/43)	60.00 (21/35)	67.31 (35/52)	67.86 (19/28)	67.5 (54/80)	79.55 (35/44)	52.78 (19/36)

DCE, dynamic contrast material–enhanced imaging; DWI, diffusion-weighted imaging; NPV, negative-predictive value; PPV, positive-predictive value; T2WI, T_2 weighted imaging.
[a]Using data of individual scores of PI-RADS v. 2 scale but generating an overall score based on the UCSF scale approach.

Figure 2. Prostate cancer Gleason score 3 + 4 in a 70-year-old male. Axial T_2 weighted imaging (T2WI) (a) demonstrates an 11-mm homogeneous focus of moderately low signal intensity (arrow) in the right mid-gland peripheral zone. It is mildly hyperintense on high *b*-value DWI (arrow in b) and shows focal mildly low signal intensity on the apparent diffusion coefficient (ADC) map (arrow in c) with an ADC value of $973 \times 10^{-6}\,mm^2\,s^{-1}$. No suspicious enhancement is seen on dynamic contrast-enhanced (DCE) MRI (arrow in d). Both readers provided scores of 4, 4, negative and 4 for T2WI, diffusion-weighted imaging (DWI), DCE MRI and overall University of California, San Francisco, scores, respectively; but 4, 3, negative and 3 for T2WI, DWI, DCE MRI and overall Prostate Imaging Reporting and Data System v. 2 scores.



a cut-off of 5 (3–4 = negative and 5 = positive) but associated with very low sensitivities (<44.2%). The overall UCSF scale and PI-RADS v. 2 accuracies for this same cut-off value were 73.1% (R1) and 73.8% (R2), and 56.4% (R1) and 56.3% (R2), respectively. Conversely, the highest sensitivity is seen using a cut-off value of 3, 98.1% (R1) and 84.6% (R2) for T2WI and 100.0% (both readers) for DWI. The overall UCSF and PI-RADS v. 2 accuracies then were 66.7% (R1, both scales) and 65.0% (R2, both scales). Figure 2 depicts a case in which there was a discrepancy between the overall scores derived by each approach.

Comparison of the diagnostic performance of the PI-RADS v. 2 and UCSF scale

The Az of the UCSF scale and PI-RADS v. 2 for the identification of clinically significant cancer with T2WI, DWI and overall scores ranged from 0.64 to 0.75, as shown in Table 3. The percent agreements of DCE for UCSF scale and PI-RADS v. 2 were 100% for R1 and 95% for R2. Except for the PI-RADS v. 2 DWI of R1 that had a tendency to perform better than the UCSF DWI score (Az = 0.75, Az = 0.68; $p = 0.05$), the comparison of Az of all other individual and overall scores, for both readers, showed no significant differences.

Interreader agreement

Table 4 shows that the interreader agreements for PI-RADS v. 2 were generally higher than those of the UCSF scale; moderate agreements were found for all UCSF scores but moderate to good for PI-RADS v. 2. The percent agreements of DCE for both readers were 81.8 % (95% CI = 71.4–89.7%) for UCSF scale and 84.4 % (95% CI = 74.4–91.7%) for PI-RADS v. 2.

Table 3. Receiver operating characteristic curve analyses and comparisons of the diagnostic performance of the University of California, San Francisco (UCSF) scale and Prostate Imaging Reporting and Data System (PI-RADS) v. 2 for identification of clinically significant prostate cancer

Criteria	Reader 1			Reader 2		
	UCSF scale	PI-RADS v. 2	*p*-value	UCSF scale	PI-RADS v. 2	*p*-value
T2WI	0.69 (0.57–0.81)	0.68 (0.56–0.80)	0.56	0.75 (0.64–0.86)	0.72 (0.62–0.83)	0.36
DWI	0.68 (0.57–0.78)	0.75 (0.65–0.86)	0.05	0.67 (0.57–0.77)	0.73 (0.63–0.84)	0.26
Overall	0.72 (0.61–0.83)	0.64 (0.54–0.74)	0.11	0.75 (0.65–0.86)	0.66 (0.57–0.75)	0.11

DWI, diffusion-weighted imaging; T2WI, T_2 weighted imaging.

Table 4. Interreader agreement for University of California, San Francisco (UCSF) scale and Prostate Imaging Reporting and Data System (PI-RADS) v. 2 using weighted kappa statistics

Criteria	UCSF scale	PI-RADS v. 2
T2WI	0.42 (0.29–0.56)	0.56 (0.41–0.70)
DWI	0.46 (0.30–0.63)	0.60 (0.41–0.72)
Overall	0.42 (0.27–0.62)	0.61 (0.44–0.76)

DWI, diffusion-weighted imaging; T2WI, T_2 weighted imaging.
Numbers in parentheses are 95% confidence intervals.

Post hoc analysis

Our analysis showed that utilizing the UCSF approach of deriving the overall score to individual PI-RADS v. 2 scores leads to an increase in the Az of overall score of PI-RADS v. 2 for the identification of clinically significant cancer. For R1, the Az changed from 0.64 (95% CI = 0.54 to 0.74) to 0.76 (95% CI = 0.65 to 0.87); for R2, it increased from 0.66 (95% CI = 0.57 to 0.75) to 0.71 (95% CI = 0.60 to 0.82). The difference, however, was statistically significant only for R1 ($p = 0.005$ for R1 and 0.35 for R2). As expected, Table 2 shows that there was also an increase in the diagnostic accuracy of the overall score for both readers, in particular when a cut-off value of 5 is used. For R1, it increased from 56.4% to 75.6%, and for R2, it changed from 56.3% to 67.5%. Table 5 summarizes the net changes in the overall scores of PI-RADS derived using the official and UCSF approaches.

DISCUSSION

Our results showed that either method is only moderately reproducible and moderately accurate for the detection of clinically significant tumours and that the accuracy of the overall PI-RADS v. 2 scores may be improved by the use of a different weighing system for the integration of its individual parameter scores.

No significant differences were found when we compared the Az of T2WI of the PI-RADS v. 2 and UCSF scales, a finding that can be explained by the similarity of the criteria used by these two approaches. In general, the interreader agreements of both scales were similar to those described in previous studies that evaluated the ESUR PI-RADS scale,[11,12] and agree with the results found by Muller et al,[17] who evaluated PI-RADS v. 2. Yet, the interreader agreement for T2WI of PI-RADS v. 2 was higher than that of the UCSF scale. Although one cannot be certain about the reasons for this finding, PI-RADS v. 2 provides an additional descriptor for Score 3 and defines the degree of low signal intensity to each score, features that may have contributed to this result.

The Az of PI-RADS v. 2 DWI was higher than that of the UCSF scale, but this difference only reached statistical significance for R1. The two scales differ slightly in the way signal intensity changes are assessed on DWI to establish a score of 3. Also, Scores 4 and 5 of the UCSF scale are distinguished using a mean ADC value threshold of $850 \times 10^{-6}\,mm^2\,s^{-1}$. In general, one would expect a better diagnostic performance and interreader agreement when using an optimal and objective criterion. Yet, the UCSF DWI score had only moderate accuracy and its interreader agreement was worse than that of PI-RADS v. 2 DWI. This might be because the chosen ADC value threshold was not optimal for this patient population, as ADC values are known to vary across institutions. Reproducibility of measured ADC values can be affected by using different MRI scanners, imaging sequences, parameters setting and potentially limiting its application as a diagnostic criterion.[28–31] Another possible contributing factor is tumour heterogeneity, which might have been captured by each reader differently when regions of interest were drawn. Methods such as standardized ADC value measurements, diffusional kurtosis imaging and volumetric measurements of ADC values could minimize the effect of tumour heterogeneity on ADC measurements, but further studies are required to determine their applicability. Last, PI-RADS v. 2 uses a lesion size threshold of 1.5 cm and/or presence of extraprostatic extension or invasive behaviour to discriminate between scores DWI 4 and 5, and this could have also contributed to the outperformance of PI-RADS v. 2 DWI.

Both UCSF and PI-RADS v. 2 scales describe DCE results in a "positive" and "negative" binary fashion and use it to adjust overall scores. The binary results of DCE lead to good

Table 5. Net changes in the overall scores of Prostate Imaging Reporting and Data System v. 2 derived using the official American College of Radiology (ACR) and the University of California, San Francisco (UCSF) methods

	ACR									
	Reader 1					Reader 2				
	Score	3	4	5	Total	Score	3	4	5	Total
UCSF	2	2	0	0	2	2	2	0	0	2
	3	3	11	0	14	3	0	8	0	8
	4	0	19	0	19	4	0	25	1	26
	5	0	19	24	43	5	0	22	22	44
	Total	5	49	24	78	Total	2	55	23	80

interreader agreement; however, the accuracy of DCE is only 69–70% for both scales and both readers. These results are similar to those of a previous report of ESUR PI-RADS[32] and are in line with other publications that evaluated the accuracy of DCE for prostate cancer detection.[33–35]

Our study found no clear difference in the accuracy of overall scores, for both readers, using a cut-off value of 3 or 4. However, the accuracy of the UCSF scale seems to be better than that of PI-RADS v. 2 when using a cut-off value of 5. Noticeably, the Az of the overall score was higher for the UCSF scale, although not reaching statistical significance. Because we found this potential difference between the Az for the overall scores, in spite of better discrimination of the PI-RADS v. 2 DWI, which is its dominant sequence for PZ tumours, we hypothesized that the weighing method proposed by PI-RADS v. 2 is suboptimal and that deriving the overall score utilizing the UCSF approach could improve its accuracy. The UCSF system gives equal weight to T2WI and DWI, irrespective of lesion location. Also, DCE is not generally integrated into the overall score of PI-RADS v. 2, as its use is mostly limited to tumours located in the PZ and that receive a score of 3 based on DWI. In the UCSF system, however, it is utilized for suspicious lesions seen in the PZ and TZ and taken into account in all cases. The results of our *post hoc* analysis indeed showed that the overall accuracy of PI-RADS v. 2 increased, suggesting that it might be possible to refine the integration of PI-RADS v. 2 individual parameter scores into the overall score utilizing a different weighing system.

This study has limitations. First, this was a retrospective study with all inherent limitations of the design. Second, we only included patients undergoing radical prostatectomy. These males tend to have localized disease, but not very aggressive tumours, as often seen in patients treated with radiation therapy. Similarly, this sample does not represent the population who elect active surveillance, and our results may not be generalizable to all males with prostate cancer. Third, readers assessed MR images and assigned both PI-RADS v. 2 and UCSF scores in a single session. It is conceivable that scores for one scheme could have influenced scores assigned using the other approach. If this happened, we underestimated the difference between the accuracies of the PI-RADS v. 2 and UCSF systems. This would be particularly true for DWI, as the T2WI and DCE MRI criteria are very similar. Readers knew all patients had prostate cancer treated with prostatectomy, and it is therefore possible that readers searched for findings more thoroughly than would have been carried out prospectively. If so, our study may have overestimated the diagnostic accuracy of both scoring systems. Yet, this would not affect the relative differences between the two scales, as both would be subject to the same bias. Last, only 2 of 65 clinically significant tumours were located in the TZ. Therefore, our results are mostly applicable to disease found in the PZ.

In conclusion, although PI-RADS v. 2 DWI score may have a higher discriminatory performance than the UCSF scale counterpart to diagnose clinically significant cancer, the utilization of the UCSF scale weighing system for the integration of PI-RADS v. 2 individual parameter scores improved the accuracy its overall score.

REFERENCES

1. Sonn GA, Natarajan S, Margolis DJ, MacAiran M, Lieu P, Huang J, et al. Targeted biopsy in the detection of prostate cancer using an office based magnetic resonance ultrasound fusion device. *J Urol* 2013; **189**: 86–91. doi: http://dx.doi.org/10.1016/j.juro.2012.08.095
2. Margel D, Yap SA, Lawrentschuk N, Klotz L, Haider M, Hersey K, et al. Impact of multiparametric endorectal coil prostate magnetic resonance imaging on disease reclassification among active surveillance candidates: a prospective cohort study. *J Urol* 2012; **187**: 1247–52. doi: http://dx.doi.org/10.1016/j.juro.2011.11.112
3. Mullins JK, Bonekamp D, Landis P, Begum H, Partin AW, Epstein JI, et al. Multiparametric magnetic resonance imaging findings in men with low-risk prostate cancer followed using active surveillance. *BJU Int* 2013; **111**: 1037–45. doi: http://dx.doi.org/10.1111/j.1464-410X.2012.11641.x
4. Muller BG, van den Bos W, Brausi M, Cornud F, Gontero P, Kirkham A, et al. Role of multiparametric magnetic resonance imaging (MRI) in focal therapy for prostate cancer: a Delphi consensus project. *BJU Int* 2014; **114**: 698–707. doi: http://dx.doi.org/10.1111/bju.12548
5. Lee T, Mendhiratta N, Sperling D, Lepor H. Focal laser ablation for localized prostate cancer: principles, clinical trials, and our initial experience. *Rev Urol* 2014; **16**: 55–66.
6. Foltz WD, Wu A, Chung P, Catton C, Bayley A, Milosevic M, et al. Changes in apparent diffusion coefficient and T2 relaxation during radiotherapy for prostate cancer. *J Magn Reson Imaging* 2013; **37**: 909–16. doi: http://dx.doi.org/10.1002/jmri.23885
7. Liu L, Wu N, Ouyang H, Dai JR, Wang WH. Diffusion-weighted MRI in early assessment of tumour response to radiotherapy in high-risk prostate cancer. *Br J Radiol* 2014; **87**: 20140359. doi: http://dx.doi.org/10.1259/bjr.20140359
8. Barrett T, Gill AB, Kataoka MY, Priest AN, Joubert I, McLean MA, et al. DCE and DW MRI in monitoring response to androgen deprivation therapy in patients with prostate cancer: a feasibility study. *Magn Reson Med* 2012; **67**: 778–85. doi: http://dx.doi.org/10.1002/mrm.23062
9. Ruprecht O, Weisser P, Bodelle B, Ackermann H, Vogl TJ. MRI of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy. *Eur J Radiol* 2012; **81**: 456–60. doi: http://dx.doi.org/10.1016/j.ejrad.2010.12.076
10. Barentsz JO, Richenberg J, Clements R, Choyke P, Verma S, Villeirs G, et al. ESUR prostate MR guidelines 2012. *Eur Radiol* 2012; **22**: 746–57. doi: http://dx.doi.org/10.1007/s00330-011-2377-y
11. Rosenkrantz AB, Kim S, Lim RP, Hindman N, Deng FM, Babb JS, et al. Prostate cancer localization using multiparametric MR imaging: comparison of Prostate Imaging Reporting and Data System (PI-RADS) and Likert scales. *Radiology* 2013; **269**: 482–92. doi: http://dx.doi.org/10.1148/radiol.13122233
12. Schimmöller L, Quentin M, Arsov C, Lanzman RS, Hiester A, Rabenalt R, et al. Inter-reader agreement of the ESUR score for prostate MRI using in-bore MRI-guided biopsies as the reference standard. *Eur Radiol* 2013; **23**: 3185–90. doi: http://dx.doi.org/10.1007/s00330-013-2922-y

13. Schieda N, Quon JS, Lim C, El-Khodary M, Shabana W, Singh V, et al. Evaluation of the European Society of Urogenital Radiology (ESUR) PI-RADS scoring system for assessment of extra-prostatic extension in prostatic carcinoma. *Eur J Radiol* 2015; **84**: 1843–8. doi: http://dx.doi.org/10.1016/j.ejrad.2015.06.016
14. Westphalen AC, Rosenkrantz AB. Prostate imaging reporting and data system (PI-RADS): reflections on early experience with a standardized interpretation scheme for multiparametric prostate MRI. *AJR Am J Roentgenol* 2014; **202**: 121–3. doi: http://dx.doi.org/10.2214/AJR.13.10889
15. Turkbey B, Mani H, Aras O, Ho J, Hoang A, Rastinehad AR, et al. Prostate cancer: can multiparametric MR imaging help identify patients who are candidates for active surveillance? *Radiology* 2013; **268**: 144–52. doi: http://dx.doi.org/10.1148/radiol.13121325
16. ACR. MR Prostate Imaging Reporting and Data System version 2.0. In: ACR, ed. *Prostate Imaging Reporting and Data System (PI-RADS)*. 1st ed. **Volume 2015**. Washington, DC: DC American College of Radiology; 2015.
17. Muller BG, Shih JH, Sankineni S, Marko J, Rais-Bahrami S, George AK, et al. Prostate cancer: interobserver agreement and accuracy with the Revised Prostate Imaging Reporting and Data System at multiparametric MR imaging. *Radiology* 2015; **277**: 741–50. doi: http://dx.doi.org/10.1148/radiol.2015142818
18. Dickinson L, Ahmed HU, Allen C, Barentsz JO, Carey B, Futterer JJ, et al. Magnetic resonance imaging for the detection, localisation, and characterisation of prostate cancer: recommendations from a European consensus meeting. *Eur Urol* 2011; **59**: 477–94. doi: http://dx.doi.org/10.1016/j.eururo.2010.12.009
19. Nagarajan R, Margolis D, Raman S, Sheng K, King C, Reiter R, et al. Correlation of Gleason scores with diffusion-weighted imaging findings of prostate cancer. *Adv Urol* 2012; **2012**: 374805. doi: http://dx.doi.org/10.1155/2012/374805
20. Peng Y, Jiang Y, Yang C, Brown JB, Antic T, Sethi I, et al. Quantitative analysis of multiparametric prostate MR images: differentiation between prostate cancer and normal tissue and correlation with Gleason score—a computer-aided diagnosis development study. *Radiology* 2013; **267**: 787–96. doi: http://dx.doi.org/10.1148/radiol.13121454
21. Donati OF, Mazaheri Y, Afaq A, Vargas HA, Zheng J, Moskowitz CS, et al. Prostate cancer aggressiveness: assessment with whole-lesion histogram analysis of the apparent diffusion coefficient. *Radiology* 2014; **271**: 143–52. doi: http://dx.doi.org/10.1148/radiol.13130973
22. Lebovici A, Sfrangeu SA, Feier D, Caraiani C, Lucan C, Suciu M, et al. Evaluation of the normal-to-diseased apparent diffusion coefficient ratio as an indicator of prostate cancer aggressiveness. *BMC Med Imaging* 2014; **14**: 15. doi: http://dx.doi.org/10.1186/1471-2342-14-15
23. Turkbey B, Shah VP, Pang Y, Bernardo M, Xu S, Kruecker J, et al. Is apparent diffusion coefficient associated with clinical risk scores for prostate cancers that are visible on 3-T MR images? *Radiology* 2011; **258**: 488–95. doi: http://dx.doi.org/10.1148/radiol.10100667
24. Bains LJ, Studer UE, Froehlich JM, Giannarini G, Triantafyllou M, Fleischmann A, et al. Diffusion-weighted magnetic resonance imaging detects significant prostate cancer with high probability. *J Urol* 2014; **192**: 737–42. doi: http://dx.doi.org/10.1016/j.juro.2014.03.039
25. Taouli B, Goh JS, Lu Y, Qayyum A, Yeh BM, Merriman RB, et al. Growth rate of hepatocellular carcinoma: evaluation with serial computed tomography or magnetic resonance imaging. *J Comput Assist Tomogr* 2005; **29**: 425–9. doi: http://dx.doi.org/10.1097/01.rct.0000164036.85327.05
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**: 837–45. doi: http://dx.doi.org/10.2307/2531595
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–74. doi: http://dx.doi.org/10.2307/2529310
28. Rosenkrantz AB, Obele C, Rusinek H, Balar AV, Huang WC, Deng FM, et al. Whole-lesion diffusion metrics for assessment of bladder cancer aggressiveness. *Abdom Imaging* 2015; **40**: 327–32. doi: http://dx.doi.org/10.1007/s00261-014-0213-y
29. Braithwaite AC, Dale BM, Boll DT, Merkle EM. Short- and midterm reproducibility of apparent diffusion coefficient measurements at 3.0-T diffusion-weighted imaging of the abdomen. *Radiology* 2009; **250**: 459–65. doi: http://dx.doi.org/10.1148/radiol.2502080849
30. Kozlowski P, Chang SD, Goldenberg SL. Diffusion-weighted MRI in prostate cancer—comparison between single-shot fast spin echo and echo planar imaging sequences. *Magn Reson Imaging* 2008; **26**: 72–6. doi: http://dx.doi.org/10.1016/j.mri.2007.04.008
31. Lin WC, Chen JH. Pitfalls and limitations of diffusion-weighted magnetic resonance imaging in the diagnosis of urinary bladder cancer. *Transl Oncol* 2015; **8**: 217–30. doi: http://dx.doi.org/10.1016/j.tranon.2015.04.003
32. Hansford BG, Peng Y, Jiang Y, Vannier MW, Antic T, Thomas S, et al. Dynamic contrast-enhanced MR imaging curve-type analysis: is it helpful in the differentiation of prostate cancer from healthy peripheral zone? *Radiology* 2015; **275**: 448–57. doi: http://dx.doi.org/10.1148/radiol.14140847.
33. Baur AD, Maxeiner A, Franiel T, Kilic E, Huppertz A, Schwenke C, et al. Evaluation of the prostate imaging reporting and data system for the detection of prostate cancer by the results of targeted biopsy of the prostate. *Invest Radiol* 2014; **49**: 411–20. doi: http://dx.doi.org/10.1097/RLI.0000000000000030
34. Kuru TH, Roethke MC, Rieker P, Roth W, Fenchel M, Hohenfellner M, et al. Histology core-specific evaluation of the European Society of Urogenital Radiology (ESUR) standardised scoring system of multiparametric magnetic resonance imaging (mpMRI) of the prostate. *BJU Int* 2013; **112**: 1080–7. doi: http://dx.doi.org/10.1111/bju.12259
35. Hara N, Okuizumi M, Koike H, Kawaguchi M, Bilim V. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a useful modality for the precise detection and staging of early prostate cancer. *Prostate* 2005; **62**: 140–7. doi: http://dx.doi.org/10.1002/pros.20124

APPENDIX A

MR imaging parameters

Sequence	Technique	TR/TE (ms)	Flip angle (°)	Slice thickness (mm)	FOV (mm)	Matrix
T2WI						
Axial	TSE	3060/100	90	3	150 × 150	232 × 184
Coronal	TSE	2444/120	90	3	150 × 150	248 × 198
Sagittal	TSE	3770/120	90	3	260 × 260	360 × 275
DWI[a,b]						
Axial	SE EPI	1561/71	90	5	304 × 375	152 × 152
DCE[c,d]						
Axial	THRIVE	4/2	10	4	297 × 345	172 × 172
T1WI						
Axial	TSE	443/15	90	3	180 × 180	180 × 143

DCE, dynamic contrast material–enhanced imaging; DWI, diffusion-weighted imaging; FOV, field of view; SE EPI, spin-echo echo-planar imaging; T1WI, T_1 weighted imaging; T2WI, T_2 weighted imaging; TE, echo time; THRIVE, T_1 high-resolution isotropic volume excitation; TR, repetition time; TSE, turbo spin echo.
Number of excitations (NEXs) is two for all sequences, except for DCE (NEX is one).
[a]*b*-values were 0 and 1000 s mm^{-2}.
[b]In-plane dimension = 2.0 and 2.5 mm [Prostate Imaging Reporting and Data System (PI-RADS) v. 2 recommendation ≤2.5 mm].
[c]The section thickness was 4.0 mm, interpolated into 2.0 mm on DCE MR image.
[d]In-plane dimension = 1.7 and 2.0 mm (PI-RADS v. 2 recommendation ≤2.0 mm).

APPENDIX B

The weights for custom-weighted kappa

		Reader 1				
		Score 1	Score 2	Score 3	Score 4	Score 5
Reader 2	Score 1	1	1	0.25	0	0
	Score 2	1	1	0.25	0	0
	Score 3	0.25	0.25	1	0.75	0.5
	Score 4	0	0	0.75	1	0.75
	Score 5	0	0	0.5	0.75	1