Original Article

Diagnostic Performance and Interobserver Consistency of the Prostate Imaging Reporting and Data System Version 2: A Study on Six Prostate Radiologists with Different Experiences from Half a Year to 17 Years

Zan Ke[1], Liang Wang[1], Xiang-De Min[1], Zhao-Yan Feng[1], Zhen Kang[1], Pei-Pei Zhang[1], Ba-Sen Li[1], Hui-Juan You[1], Sheng-Chao Hou[2]

[1]Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China
[2]Department of Library, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China

Abstract

Background: One of the main aims of the updated Prostate Imaging Reporting and Data System Version 2 (PI-RADS v2) is to diminish variation in the interpretation and reporting of prostate imaging, especially among readers with varied experience levels. This study aimed to retrospectively analyze diagnostic consistency and accuracy for prostate disease among six radiologists with different experience levels from a single center and to evaluate the diagnostic performance of PI-RADS v2 scores in the detection of clinically significant prostate cancer (PCa).
Methods: From December 2014 to March 2016, 84 PCa patients and 99 benign prostatic shyperplasia patients who underwent 3.0T multiparametric magnetic resonance imaging before biopsy were included in our study. All patients received evaluation according to the PI-RADS v2 scale (1–5 scores) from six blinded readers (with 6 months and 2, 3, 4, 5, or 17 years of experience, respectively, the last reader was a reviewer/contributor for the PI-RADS v2). The correlation among the readers' scores and the Gleason score (GS) was determined with the Kendall test. Intra-/inter-observer agreement was evaluated using κ statistics, while receiver operating characteristic curve and area under the curve analyses were performed to evaluate the diagnostic performance of the scores.
Results: Based on the PI-RADS v2, the median κ score and standard error among all possible pairs of readers were 0.506 and 0.043, respectively; the average correlation between the six readers' scores and the GS was positive, exhibiting weak-to-moderate strength ($r = 0.391$, $P = 0.006$). The AUC values of the six radiologists were 0.883, 0.924, 0.927, 0.932, 0.929, and 0.947, respectively.
Conclusion: The inter-reader agreement for the PI-RADS v2 among the six readers with different experience is weak to moderate. Different experience levels affect the interpretation of MRI images.

Key words: Benign Prostatic Hyperplasia; Diagnosis; Magnetic Resonance Imaging; Prostate Cancer; Prostate Imaging Reporting and Data System Version 2

Introduction

For the past decade, benign prostatic hyperplasia (BPH) and prostate cancer (PCa) have remained the most common diseases of the male prostate. In 2017, 161,360 new PCa cases and 26,730 PCa deaths are projected to occur in the United States according to the American Cancer Society.[1] PCa is the third leading cause of cancer-related death among males,[1] followed by lung/bronchus and colon/rectum-related neoplastic diseases, and prostate disease remains a significant challenge not only for urologists and oncologists, but also for radiologists.

Advances in computer software and hardware have led to multiparametric magnetic resonance imaging (mp-MRI),

Access this article online
Quick Response Code:
Website: www.cmj.org
DOI: 10.4103/0366-6999.235872

Address for correspondence: Dr. Liang Wang,
Department of Radiology, Tongji Hospital, Tongji Medical College,
Huazhong University of Science and Technology, Hubei,
Wuhan 430030, China
E-Mail: wang6@tjh.tjmu.edu.cn

Received: 28-02-2018 **Edited by:** Ning-Ning Wang
How to cite this article: Ke Z, Wang L, Min XD, Feng ZY, Kang Z, Zhang PP, Li BS, You HJ, Hou SC. Diagnostic Performance and Interobserver Consistency of the Prostate Imaging Reporting and Data System Version 2: A Study on Six Prostate Radiologists with Different Experiences from Half a Year to 17 Years. Chin Med J 2018;131:1666-73.

combining anatomical T2-weighted imaging (T2WI) and functional MRI sequences, such as diffusion-weighted imaging (DWI), apparent-diffusion coefficient (ADC) maps, or dynamic contrast-enhanced (DCE) imaging,[2,3] which has become the preferred imaging method for the prostate and periprostatic structures. This approach provides more accurate localization and high-quality images for the detection of prostate diseases, especially for PCa.[4-7] Due to differences in magnetic resonance scanners, acquisition parameter settings, and subjective evaluation criteria, the interpretation of mp-MRI findings by radiologists differs from different clinicians. Therefore, how to unify diagnostic systems and bridge the gap between different radiologists is increasingly recognized as an important clinical problem.

To address this issue, the European Society of Urogenital Radiology (ESUR) launched the first version of a global prostate standardization guide called the Prostate Imaging Reporting and Data System (PI-RADS; herein referred to as the PI-RADS v1) in 2012.[8] The PI-RADS v1 was widely distributed, but some limitations in its clinical application caused significant controversy regarding inter-reader reproducibility and the feasibility of the guidelines.[9-13] First, the PIRADS v1 does not include a rating scheme, and no weights for individual parameters were defined. Second, the PI-RADS v1 does not combine all imaging sequences into a comprehensive assessment.[14] Third, the value of DCE in evaluating the transition zone (TZ) is overestimated by the PI-RADS v1[15] and no value was assigned or recommended for DCE. In addition, DWI in the peripheral zone (PZ) has previously been reported to exhibit superior performance.[12] Considering these issues, in 2014, the updated PI-RADS version 2 (herein referred to as the PI-RADS v2) was released by the International Collaboration of the American College of Radiology, ESUR, and AdMetech Foundation, based on the best available evidence and expert consensus opinion worldwide.[16] Compared with the PI-RADS v1, the PI-RADS v2 has a simplified scoring system and uses only a 5-point scale for comprehensive evaluation of all imaging sequences. In addition, the PI-RADS v2 uses a more differentiated weighting system based on the concept of dominant techniques and does not recommend the magnetic resonance spectroscopic (MRS) imaging for PI-RADS assessment but rather DWI as the dominant sequence in PZ and T2WI as the dominant sequence in TZ. Third, the PI-RADS v2 recommends optimal technical parameters for T2WI, DWI, and DCE sequences and introduces a new size threshold of 15 mm for T2WI, DWI, and ADC to differentiate between PI-RADS scores of 4 and 5. Moreover, the aims of the PI-RADS v2 are to improve the detection of clinically significant cancer and increase the accuracy of risk assessment for patients with suspected PCa, to enhance diagnostic confidence in benign diseases, to establish the most simplified MRI capture process globally and diminish variation in the acquisition and interpretation of prostate images, and to promote communication between clinicians and radiologists.[16,17]

Based on these advantages and aims of the PI-RADS v2, investigation of the utility of the PI-RADS among readers is crucial, not only between two readers for a small number of cases, but also among more readers with varying experience levels for a high number of cases. Therefore, the purpose of this study was to retrospectively analyze consistency and accuracy in diagnosing prostate disease among six radiologists with different experience levels and to evaluate the diagnostic performance of the PI-RADS v2 in detecting clinically significant PCa.

Methods

Ethical approval

This retrospective, single-center study was approved by the Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology Institutional Review Board, and written informed consent was provided by all patients before examination.

Study design

Between December 2014 and March 2016, patients with clinically suspected PCa due to elevated prostate-specific antigen (PSA) levels and/or abnormal signal nodules were recruited for this study. Initially, we reviewed 317 patients who underwent 3.0T prostate MRI. However, 134 patients were excluded for the following reasons: (1) the patients had no histopathologically confirmed results (n = 30); (2) the patients underwent prior treatment, including surgical therapies, irradiation, cryosurgery, or hormonotherapy (n = 15); (3) previous biopsies were performed within 6 weeks before the MR examination (n = 2); (4) DCE imaging was not performed in the patient due to renal dysfunction and/or unwillingness to undergo the procedure (n = 84); and (5) the quality of the MRI images was poor due to movement artifacts, catheter artifacts, or the presence of hip implants (n = 3). Finally, the remaining 183 patients were included and a flowchart of the patient selection process is provided in Figure 1.

Magnetic resonance imaging protocol

All examinations were performed with a 3.0T system (MAGNETOM Skyra, Siemens Healthcare, Erlangen, Germany), using an anterior 18-element body coil combined with a posterior spine coil array. The scan sequences included T2WI, DWI, and DCE, which were performed using the parameters shown in Table 1. In DWI, the b values consisted of 0, 50, 200, 400, 600, 800, 1000, and 1500 s/mm^2. ADC maps were automatically reconstructed for qualitative and quantitative assessments of DWI. Axial DCE images were obtained before, during, and after rapid injection of gadolinium chelate (35 phases and 8 s for each phase) using a power injector (Medtron, Saarbruecken, Germany), followed by a 20 ml saline flush injected at a rate of 2.5 ml/s. All axial images were copied at the same location.

Magnetic resonance imaging interpretation and PI-RADS scoring

For each patient, mp-MRI images of the prostate were shown to six independent readers with varying levels of experience in the diagnosis of prostate diseases (reader 1, Zhen Kang, with 6 months of experience [approximately 100 examinations]; reader 2, Pei-Pei Zhang, with 2 years of experience [approximately 400 examinations]; reader 3, Zan Ke, with 3 years of experience [approximately 600 examinations]; reader 4, Xiang-De Min, with 4 years of experience [approximately 800 examinations]; reader 5, Zhao-Yan Feng, with 5 years of experience [approximately 1000 examinations]; and reader 6, Liang Wang, with 17 years of experience who was a reviewer/contributor for the PI-RADS v2 [approximately 10,000 examinations]). These six readers were blinded to all identifying information of the patients and their clinicopathologic outcomes. During scoring, the T2WI, DWI, and DCE images of each patient were shown to the readers at the same location on one single screen by an assistant fellow who assigned a scoring region but was not involved in the scoring process, and then the readers independently provided a single score (on a scale from 1 to 5 scores) based on the PI-RADS v2 and their own experience and comprehensive judgment after browsing all sequences. After 2 weeks, reader 3 repeated the scoring process to test intrareader reproducibility.

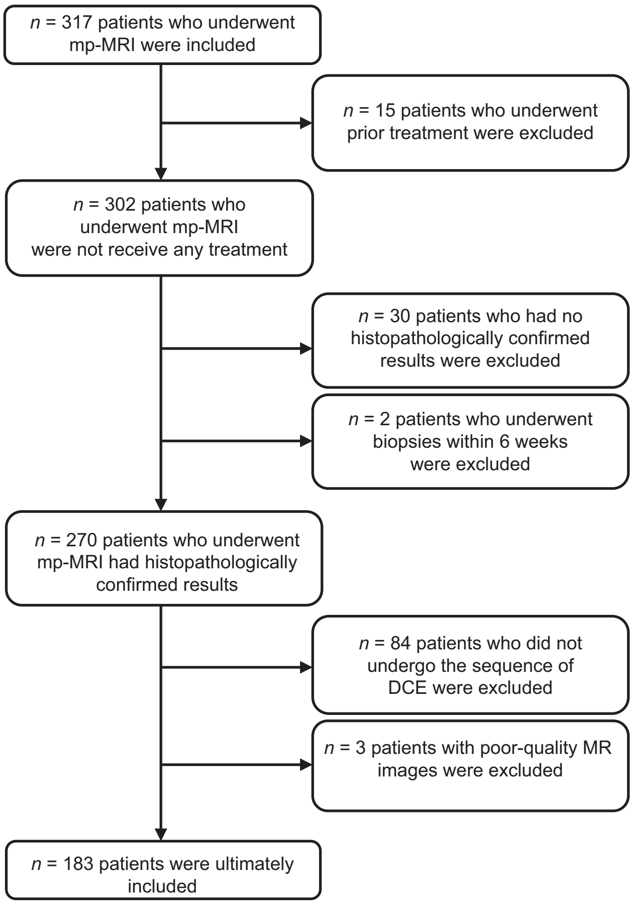
n = 317 patients who underwent mp-MRI were included
n = 15 patients who underwent prior treatment were excluded
n = 302 patients who underwent mp-MRI were not receive any treatment
n = 30 patients who had no histopathologically confirmed results were excluded
n = 2 patients who underwent biopsies within 6 weeks were excluded
n = 270 patients who underwent mp-MRI had histopathologically confirmed results
n = 84 patients who did not undergo the sequence of DCE were excluded
n = 3 patients with poor-quality MR images were excluded
n = 183 patients were ultimately included

Figure 1: Flowchart for the selection of patients in the present study. mp-MRI: Multiparametric magnetic resonance imaging; DCE: Dynamic contrast enhanced.

Pathologic evaluation

After the MRI examination, all patients underwent a 12-core transrectal ultrasound (TRUS)-guided prostate biopsy (within 6 weeks; median: 1 week) to obtain tissue samples for histopathological examination. To match biopsy sextants and MR images, the prostate was divided into 12 regions, and each specimen was individually labeled according to its location and histologically analyzed. The targeted biopsy was performed using an ultrasound system (Hawk 2102, BK Medical, Denmark) equipped with a 5.1-MHz endocavitary probe and a spring-loaded biopsy gun with an 18G core biopsy needle; a single urologist with 20 years of skilled experience performed these biopsies. The samples were assessed by an experienced genitourinary pathologist with more than 10 years of experience, who was blinded to the MRI results. The cases were obtained from standard pathologic reports, and each sample was histologically analyzed as cancerous or noncancerous and then given a respective Gleason score (GS) if the sample was classified as PCa. Finally, we selected the GS matching the scoring region in the MRI images as the final GS.

Statistical analysis

SPSS 19.0 (SPSS, Chicago, IL, USA) and MedCalc version 11.4.2.0 (MedCalc statistical software, Mariakerke, Belgium) were used for the data analysis, and all data were expressed as the mean ± standard deviation (SD). The normality and equality of variances of the parameter value

Table 1: mp-MR imaging sequence parameters at 3.0T

Parameter	T2WI	T1WI	DWI	DCE
Repetition time (ms)	6874.00	807.00	4500.00	5.08
Echo time (ms)	104.00	13.00	85.00	1.77
Section thickness (mm)	3.00	5.00	3.00	3.50
Intersection gap (mm)	0	0	0	0.70
Field of view (mm^2)	180 × 180	300 × 356	214 × 171	260 × 260
Matrix	384 × 384	320 × 240	90 × 72	192 × 154
Parallel imaging factor	2	NA	2	2
Flip angle (°)	160	160	90	15
Time of acquisition (s)	196	186	248	284

T2WI: T2-weighted imaging; T1WI: T1-weighted imaging; DWI: Diffusion-weighted imaging; DCE: Dynamic contrast enhanced; NA: Not applicable; mp-MRI: Multiparametric magnetic resonance imaging.

distributions were tested by the Kolmogorov-Smirnov test and Levene's F-test. Differences in reader grouping variables were evaluated by the Kruskal-Wallis H-test and a comparison between all possible pairs of readers was performed using the Nemenyi test. Intra- and inter-reader agreement was evaluated using κ statistics,[18] and κ coefficients were assessed as follows:[19] 0.01–020: slight agreement; 0.21–0.40: fair agreement; 0.41–0.60: moderate agreement; 0.61–0.80: substantial agreement; and 0.81–0.99: almost perfect agreement. The correlation among the readers' scores of PCa and the GS was determined with the Kendall τ correlation coefficient (presented as "r"), which is a nonparametric statistical method used for variables that do not meet normality. The r ranged from −1 to 1, with 1 corresponding to a 100% positive correlation, −1 corresponding to a 100% negative correlation, and 0 corresponding to independence.[18] The Wald test was used to obtain the P value of the final Kendall τ estimate. A receiver operating characteristic curve (ROC) analysis was performed, and the area under the curve (AUC) was obtained to evaluate diagnostic performance. The AUC values from the six readers were compared using the Z-test. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated by dichotomizing the PI-RADS criteria according to cutoff values of 3 and 4, which were used as the threshold to distinguish benign cases from cancer and low-risk cancer (defined as GS $\leq 3 + 4 = 7$) from clinically significant cancer (defined as GS $\geq 4 + 3 = 7$)[16] in PCa patients. A $P < 0.05$ was used to identify a statistically significant difference.

Results

Patient characteristics

One-hundred and eighty-three patients were included in this retrospective, single-center study, including 84 patients who were diagnosed with PCa and 99 patients whose 1188 biopsy specimens were all diagnosed as benign hyperplasia tissue, representing the BPH group in the study. The mean age of our study population was 65.4 ± 8.5 years (range: 46–88 years). The mean PSA level was 134.48 ± 230.97 ng/ml (range: 1.51–1000.00 ng/ml, excluding one patient without a PSA value) in the PCa group and 14.29 ± 19.17 ng/ml (range: 0.26–115.04 ng/ml, excluding six patients without a PSA value) in the BPH group. The biopsy results confirmed clinically significant PCa (GS $\geq 4 + 3 = 7$) in 58 (69%) patients and low-risk PCa (GS $\leq 3 + 4 = 7$) in 26 (31%) patients. Patient characteristics are shown in Table 2.

Interobserver agreement

The κ statistics of all possible pairs of readers were calculated, and pairwise κ statistics and standard errors are shown in Table 3. In general, the inter-reader agreement was weak to moderate, while the intrareader agreement was good. The median κ statistic and standard error among all possible pairs of readers for the PI-RADS v2 were 0.506 and 0.043, respectively. Figures 2 and 3 show representative lesions for BPH and PCa with inter-reader variability, respectively.

Differences in grouping variables

The data did not conform to the criteria for normality or homogeneity of variance. The Kruskal-Wallis H-test

Table 2: Characteristics of patients enrolled in this study

Characteristics	Total	PCa	BPH
Number of patients	183	84	99
Age (years), mean (range)	65.4 (46.0–88.0)	66.1 (50.0–88.0)	64.9 (46.0–85.0)
PSA (ng/ml), mean (range)	70.97 (0.26–1000.00)	134.48 (1.51–1000.00)	14.29 (0.26–115.04)
Prostate volume (ml), mean (range)	54.61 (12.31–271.47)	49.02 (12.31–271.47)	59.34 (13.31–232.55)
Clinically significant PCa, *n* (%)	58 (31.7)	58 (69.0)	NA
Low-risk PCa, *n* (%)	26 (14.2)	26 (31.0)	NA
GS, *n*			
GS of 2 + 3	NA	1	NA
GS of 3 + 3	NA	7	NA
GS of 3 + 4	NA	18	NA
GS of 4 + 3	NA	14	NA
GS of 4 + 4	NA	25	NA
GS of 4 + 5	NA	7	NA
GS of 5 + 4	NA	10	NA
GS of 5 + 5	NA	2	NA
Clinical stage, *n*			
cT2a	NA	10	NA
cT2b	NA	15	NA
cT2c	NA	1	NA
cT3a	NA	7	NA
cT3b	NA	32	NA
cT4	NA	19	NA

PSA: Prostate-specific antigen; PCa: Prostate cancer; BPH: Benign prostatic hyperplasia; NA: Not applicable; GS: Gleason score.

showed significant differences in the overall variables, and the Nemenyi test indicated that some of the possible pairs of readers presented significant differences [Supplementary Table 1]. For the 183 patients, including 84 PCa patients and 99 BPH patients, significant differences among the six readers were identified (F = 39.42, $P < 0.001$; F = 32.09, $P < 0.001$; and F = 97.45, $P < 0.001$, respectively).

Table 3: Pair-wise inter-reader κ statistic of the PI-RADS v2 (n = 183)

Reader pairs	κ score*	Standard error
1 and 2	0.475	0.043
1 and 3a	0.558	0.044
1 and 4	0.478	0.043
1 and 5	0.369	0.038
1 and 6	0.455	0.045
2 and 3a	0.553	0.045
2 and 4	0.536	0.044
2 and 5	0.441	0.041
2 and 6	0.620	0.043
3a and 4	0.569	0.045
3a and 5	0.385	0.040
3a and 6	0.520	0.045
4 and 5	0.642	0.041
4 and 6	0.559	0.044
5 and 6	0.430	0.039
Mean	0.506	0.043
3a and 3b[†]	0.788	0.036

*κ score of the overall PI-RADS score; [†]3a-The first score of reader 3; 3b-The second score of reader 3 (the second score was performed two weeks after the first score). PI-RADS v2: Prostate Imaging Reporting and Data System Version 2.

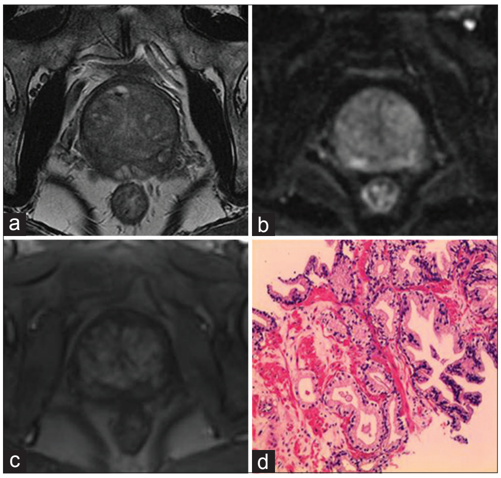

Figure 2: Images from a 55-year-old man who was diagnosed with PCa (GS = 3 + 3 = 6), with a PSA level of 22.43 mg/ml. The readers evaluated the prostate based on (a) T2-WI, (b) DWI (b = 1500 s/mm^2), and (c) an axial early DCE image, and the results were confirmed by (d) a pathology image (hematoxylin and eosin staining, ×200). Five of the six readers did not note the right peripheral zone lesion; only reader 6 noticed it. The DCE image showed that the lesion presented slight early enhancement, but the other five readers considered the prostate as a whole to be negative for DCE. Finally, the overall PI-RADS v2 scores assigned by the six readers were 2, 2, 2, 2, 2, and 4, respectively. T2WI: T2-weighted imaging; DWI: Diffusion-weighted imaging; DCE: Dynamic contrast enhanced; PSA: Prostate-specific antigen; PCa: Prostate cancer; GS: Gleason score; PI-RADS v2: Prostate Imaging Reporting and Data System Version 2.

Correlations of the Prostate Imaging Reporting and Data System with pathologic results

The double-variable data in our study did not meet the requirements for normality, and the correlations between the PI-RADS v2 scores of the readers and the GSs of the pathologic results are shown in Table 4. On the basis of the PI-RADS v2, the average correlation between the six readers' scores and the GS was positive (r = 0.319; P = 0.006), exhibiting significance and weak-to-moderate strength. The scores of reader 3 were most significant in relation to the GS (r = 0.464; P = 0.000), while the correlation between the scores of readers 2, 4, and 6 and the GS was weak.

Receiver operating characteristic curves and diagnostic performance

Supplementary Table 2 shows that in all cases, readers 2 and 6 showed the highest accuracy (90.2%), reader 4 showed the highest sensitivity (96.4%), and reader 1 showed the highest specificity (91.9%). For PCa detection, reader 6 showed the lowest accuracy (70.2%), reader 3 showed the highest accuracy (79.8%), and reader 5 showed the highest sensitivity (96.6%). Figure 4 shows the ROC curves of the six readers with different experience levels. The comparison of the AUC values is shown in Supplementary Table 3. In addition to readers 1 and 6 (Z = 2.341; P = 0.019), no significant differences were found in the overall AUC values

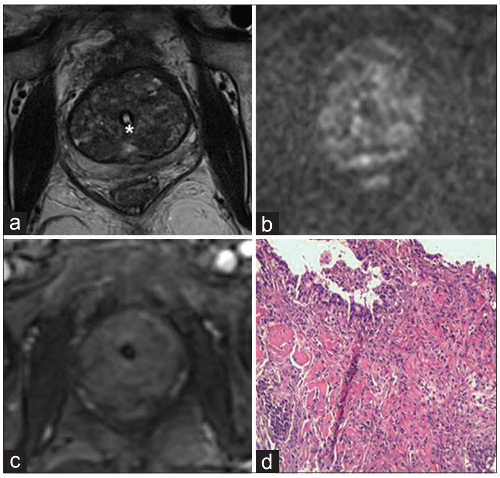

Figure 3: Images from a 73-year-old man who was diagnosed with BPH with a PSA level of 15.948 mg/ml. The readers evaluated the prostate based on (a) T2WI (the asterisk represents a urethral catheter), (b) DWI (b = 1500 s/mm^2), and (c) an axial early DCE image, and the results were confirmed by (d) a pathology image (hematoxylin and eosin staining, ×200). All the readers considered the prostate as a whole to be negative for DCE, with only slight diffusion restriction on DWI. Finally, the overall PI-RADS v2 scores assigned by the six readers were 2, 3, 2, 2, 3, and 3, respectively. T2WI: T2-weighted imaging; DWI: Diffusion-weighted imaging; DCE: Dynamic contrast enhanced; PSA: Prostate-specific antigen; BPH: Benign prostatic hyperplasia; PI-RADS v2: Prostate Imaging Reporting and Data System Version 2.

among the readers. However, readers 2 and 3, readers 3 and 4, readers 3 and 6, readers 1 and 6, and readers 5 and 6 showed significant differences in AUC values for the PCa group.

Discussion

In the current study, we invited six radiologists with varying experience levels to read prostate MRI images and assign scores using the PI-RADS v2. Our study revealed a moderate level of interobserver agreement among these readers, indicating that different experience levels may affect the interpretation of images, even under the guidance of the PI-RADS v2. A similar level of interobserver agreement was reported by Muller *et al*.,[18] who showed that the interobserver reproducibility for the overall suspicion score of the PI-RADS v2 was moderate (κ statistic score: 0.46; standard error: 0.03) as scored by five independent readers with varying experience levels (12 years, 7 years, 1 year for two readers, and 6 months). However, the readers in their study showed a narrow range of experience, while our study involved six readers with a broad range of experience (2, 3, 4, 5, and 17 years, and 6 months). In another study, Rosenkrantz *et al*.[20] found that the interobserver agreement was 0.593 for the PZ and 0.509 for the TZ based on a PI-RADS v2 score of 4 or greater; their analysis included six experienced radiologists from six separate institutions, consisted of two sessions, and included an intersession training period with discussion. However, no substantial difference in interobserver agreement was observed between the two sessions, and a training session was neither required nor provided an added benefit.

Table 4: Correlation coefficient of Kendall test and *P* values between six readers' PI-RADS v2 scores and GS on PCa patients (*n* = 84)

Reader	*r*	*P*
1	0.377	0.000
2	0.284	0.004
3a	0.464	0.000
4	0.253	0.011
5	0.306	0.002
6	0.231	0.020
Mean	0.319	0.006

PCa: Prostate cancer; GS: Gleason score; PI-RADS v2: Prostate Imaging Reporting and Data System Version 2.

Significant differences were identified among the scores of the six readers in our study. Therefore, radiologist experience is a crucial factor when evaluating MR images. However, differences were not noted between each pair of readers, and most of the differences were associated with reader 1 who had only 6 months of experience, suggesting that lack of experience has an impact on MRI interpretation, even though according to the PI-RADS v2 which is based on expert consensus opinion worldwide, lack of experience may correspond to a lack of understanding. Our results also indicated that the average correlation between the scores of the six readers for the 84 PCa patients and the GS was positive and moderate according to the PI-RADS v2. NiMhurchu *et al*.[21] showed that the correlation between a positive targeted biopsy and both the T2WI and overall PI-RADS scores was also significant ($P < 0.001$), while the correlation between a targeted biopsy and the DWI score was significant only for PZ tumors. However, this study was based on the PI-RADS v1, and whether the PI-RADS v2 would have led to the same result is difficult to determine.

In distinguishing between benign and malignant lesions, the most experienced reader (reader 6) in our study achieved the highest accuracy and AUC when the cutoff value was set at 3; however, this reader showed neither the highest nor the lowest percentage in terms of sensitivity, specificity, PPV, and NPV. Meanwhile, the least experienced reader (reader 1) achieved the lowest AUC, sensitivity, NPV, and accuracy and the highest specificity and PPV among the readers, which may be due to the different experience levels of the readers. In the study of Baldisserotto *et al*.,[22] a PI-RADS score of 3 was applied as an indicator of the absence of cancer, and the accuracy, sensitivity, specificity,

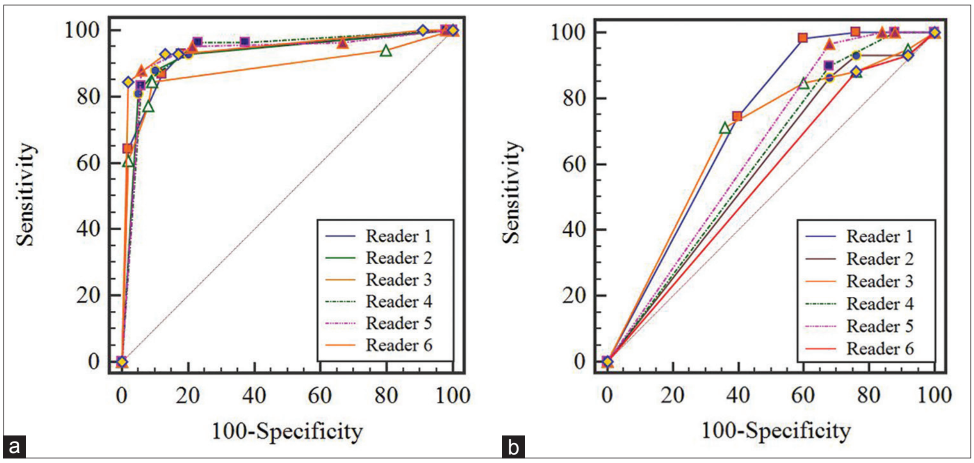

Figure 4: ROC analysis results of the six readers with different experience levels for the PCa patients. (a) ROC curves of the six readers with different experience levels for all the 183 patients. (b) ROC curves of the six readers with different experience levels for the 84 PCa patients. ROC: Receiver operating characteristic; PCa: Prostate cancer.

PPV, and NPV of reader 1 (10 years of experience) were 77.8%, 73.5%, 85.0%, 89.3%, and 65.4%, respectively, and these values for reader 2 (4 years of experience) were 77.8%, 76.5%, 80.0%, 86.7%, and 66.7%, respectively. These values are lower than those of our study, which also demonstrates that the differences among readers may have been caused by varying experience levels. Previous studies without the PI-RADS criteria, such as that by Garcia-Reyes *et al.*,[23] have also shown that readers' experience influences the accuracy of mp-MRI regarding the diagnosis of PCa. Nevertheless, from the results of the distinction between low-risk cancer and clinically significant cancer, reader 1 achieved the highest PPV and specificity, while reader 6 showed the lowest specificity, PPV, and accuracy, which can be interpreted as the reader with more experience showing more conservative tendencies. A study in 2016 by Zhao *et al.*[24] revealed a significant correlation between a higher PI-RADS v2 score and the presence of clinically significant PCa ($P < 0.001$), and a PI-RADS score of 3 was identified as the best cutoff point with a sensitivity and specificity greater than 80%. Our results showed a similar average specificity and sensitivity using the same cutoff. In recent years, most studies have concluded that the PI-RADS v2 exhibits better diagnostic performance than the PI-RADS v1.[25,26] Another study by Wang *et al.*[27] that evaluated the PI-RADS v1 score with respect to the PCa detection rate in patients with PSA levels <20 ng/ml showed a good correlation between an increased PI-RADS score and an increased cancer detection rate, and the summed score of T2WI + DWI showed the highest accuracy for PCa detection. However, a few studies have produced different results, showing that although the PI-RADS v2 uses a simplified approach, this system can lead to a higher rate of false-negative results and lower diagnostic accuracy due to the risk of missing low PI-RADS-scored tumors. In a study by Auer *et al.*,[28] the authors included fifty PCa patients who underwent mp-MRI, and all the images were evaluated according to the PI-RADS v1 and PI-RADS v2 by two radiologists with a similar level of expertise. Their results showed that the PI-RADS v1 had a significantly larger discriminative ability for tumor detection regardless of whether the lesion was in the PZ or the TZ (PI-RADS v1 AUC: 0.96; PI-RADS v2 AUC: 0.90).

Several limitations existed in our study. The primary limitation is that the mean PSA level for the PCa population was slightly higher, and 70% of the PCa cases were locally advanced (stage T3/T4; 23% of the tumors were T4), which may have biased the study because larger, more aggressive tumors will be found by most radiologists; therefore, the agreement will be high and the diagnostic accuracy will be good. This phenomenon has also been observed in other studies.[29] However, PSA screening is not common in China, so our patients usually visit a doctor when they have obvious clinical symptoms, which often reflect an advanced disease stage. Therefore, we hope to improve this aspect in future research. Second, our readers provided only one final score for each case, and the results were not separately analyzed according to the PZ, TZ, T2WI, T1WI, or DWI. The aim of the PI-RADS v2 is to conduct a comprehensive evaluation of prostate lesions according to all major sequences rather than just one sequence. Therefore, providing a fast, accurate, and comprehensive judgment according to the PI-RADS v2 is important, which is why we conducted a comprehensive evaluation to adapt to these new conditions. Another potential limitation is that our reader and patient data all came from the same center, and although the readers were blinded to all identifying patient information, the readers may have been familiar with the cases in our database, which may have increased the inter-reader agreement or accuracy. Therefore, a larger dataset from a multicentric study is needed in the future. In addition, we selected a GS ≥4 + 3 as the definition of clinically significant PCa. However, no universally accepted consensus exists regarding the definition of clinically significant PCa. Finally, the reference standard that we used was TRUS-guided prostate biopsy, which may be less accurate than prostatectomy.[30] However, the primary goal of our study was to explore diagnostic performance and interobserver consistency among readers with different experience levels according to the latest PI-RADS version. Therefore, the impact of this limitation was very small, and we aim to enroll more patients with prostatectomy in the future to support the results of this study.

In conclusion, six prostate radiologists with different experience levels achieved weak-to-moderate inter-reader agreement using the PI-RADS v2 lexicon, and varying levels of experience have an impact on the interpretation of MR images. However, the PI-RADS v2 showed excellent diagnostic performance for different readers; therefore, our data suggested that as a living document, the PI-RADS will evolve and change in response to clinical needs and technical improvements in the future.

Supplementary information is linked to the online version of the paper on the Chinese Medical Journal website.

Financial support and sponsorship

This work was supported by grants from the National Natural Science Foundation of China (NSFC; No. 81671656 and No. 81171307).

Conflicts of interest

There are no conflicts of interest.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. CA Cancer J Clin 2017;67:7-30. doi: 10.3322/caac.21387.
2. Barrett T, Turkbey B, Choyke PL. PI-RADS version 2: What you need to know. Clin Radiol 2015;70:1165-76. doi: 10.1016/j.crad.2015.06.093.
3. Moosavi B, Flood TA, Al-Dandan O, Breau RH, Cagiannos I, Morash C, *et al.* Multiparametric MRI of the anterior prostate gland: Clinical-radiological-histopathological correlation. Clin Radiol 2016;71:405-17. doi: 10.1016/j.crad.2016.01.002.
4. Turkbey B, Choyke PL. Multiparametric MRI and prostate cancer diagnosis and risk stratification. Curr Opin Urol 2012;22:310-5. doi: 10.1097/MOU.0b013e32835481c2.
5. Ghai S, Haider MA. Multiparametric-MRI in diagnosis of prostate cancer. Indian J Urol 2015;31:194-201. doi:

10.4103/0970-1591.159606.
6. Mertan FV, Greer MD, Shih JH, George AK, Kongnyuy M, Muthigi A, *et al.* Prospective evaluation of the prostate imaging reporting and data system version 2 for prostate cancer detection. J Urol 2016;196:690-6. doi: 10.1016/j.juro.2016.04.057.
7. Lin WC, Westphalen AC, Silva GE, Chodraui Filho S, Reis RB, Muglia VF, *et al.* Comparison of PI-RADS 2, ADC histogram-derived parameters, and their combination for the diagnosis of peripheral zone prostate cancer. Abdom Radiol (NY) 2016;41:2209-17. doi: 10.1007/s00261-016-0826-4.
8. 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: 10.1007/s00330-011-2377-y.
9. Rosenkrantz AB, Lim RP, Haghighi M, Somberg MB, Babb JS, Taneja SS, *et al.* Comparison of interreader reproducibility of the prostate imaging reporting and data system and Likert scales for evaluation of multiparametric prostate MRI. AJR Am J Roentgenol 2013;201:W612-8. doi: 10.2214/AJR.12.10173.
10. 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: 10.1148/radiol.13122233.
11. 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: 10.1007/s00330-013-2922-y.
12. Junker D, Quentin M, Nagele U, Edlinger M, Richenberg J, Schaefer G, *et al.* Evaluation of the PI-RADS scoring system for mpMRI of the prostate: A whole-mount step-section analysis. World J Urol 2015;33:1023-30. doi: 10.1007/s00345-014-1370-x.
13. Grey AD, Chana MS, Popert R, Wolfe K, Liyanage SH, Acher PL, *et al.* Diagnostic accuracy of magnetic resonance imaging (MRI) prostate imaging reporting and data system (PI-RADS) scoring in a transperineal prostate biopsy setting. BJU Int 2015;115:728-35. doi: 10.1111/bju.12862.
14. Hamoen EHJ, de Rooij M, Witjes JA, Barentsz JO, Rovers MM. Use of the prostate imaging reporting and data system (PI-RADS) for prostate cancer detection with multiparametric magnetic resonance imaging: A Diagnostic meta-analysis. Eur Urol 2015;67:1112-21. doi: 10.1016/j.eururo.2014.10.033.
15. Rosenkrantz AB, Kim S, Campbell N, Gaing B, Deng FM, Taneja SS, *et al.* Transition zone prostate cancer: Revisiting the role of multiparametric MRI at 3 T. AJR Am J Roentgenol 2015;204:W266-72. doi: 10.2214/AJR.14.12955.
16. Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, *et al.* PI-RADS prostate imaging – Reporting and data system: 2015, version 2. Eur Urol 2016;69:16-40. doi: 10.1016/j.eururo.2015.08.052.
17. Turkbey B, Choyke PL. PIRADS 2.0: What is new? Diagn Interv Radiol 2015;21:382-4. doi: 10.5152/dir.2015.15099.
18. 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: 10.1148/radiol.2015142818.
19. Mendhiratta N, Meng X, Rosenkrantz AB, Wysock JS, Fenstermaker M, Huang R, *et al.* Prebiopsy MRI and MRI-ultrasound fusion-targeted prostate biopsy in men with previous negative biopsies: Impact on repeat biopsy strategies. Urology 2015;86:1192-8. doi: 10.1016/j.urology.2015.07.038.
20. Rosenkrantz AB, Ginocchio LA, Cornfeld D, Froemming AT, Gupta RT, Turkbey B, *et al.* Interobserver reproducibility of the PI-RADS version 2 lexicon: A Multicenter study of six experienced prostate radiologists. Radiology 2016;280:793-804. doi: 10.1148/radiol.2016152542.
21. NiMhurchu E, O'Kelly F, Murphy IG, Lavelle LP, Collins CD, Lennon G, *et al.* Predictive value of PI-RADS classification in MRI-directed transrectal ultrasound guided prostate biopsy. Clin Radiol 2016;71:375-80. doi: 10.1016/j.crad.2016.01.001.
22. Baldisserotto M, Neto EJ, Carvalhal G, de Toledo AF, de Almeida CM, Cairoli CE, *et al.* Validation of PI-RADS v. 2 for prostate cancer diagnosis with MRI at 3T using an external phased-array coil. J Magn Reson Imaging 2016;44:1354-9. doi: 10.1002/jmri.25284.
23. Garcia-Reyes K, Passoni NM, Palmeri ML, Kauffman CR, Choudhury KR, Polascik TJ, *et al.* Detection of prostate cancer with multiparametric MRI (mpMRI): Effect of dedicated reader education on accuracy and confidence of index and anterior cancer diagnosis. Abdom Imaging 2015;40:134-42. doi: 10.1007/s00261-014-0197-7.
24. Zhao C, Gao G, Fang D, Li F, Yang X, Wang H, *et al.* The efficiency of multiparametric magnetic resonance imaging (mpMRI) using PI-RADS version 2 in the diagnosis of clinically significant prostate cancer. Clin Imaging 2016;40:885-8. doi: 10.1016/j.clinimag.2016.04.010.
25. Park SY, Jung DC, Oh YT, Cho NH, Choi YD, Rha KH, *et al.* Prostate cancer: PI-RADS version 2 helps preoperatively predict clinically significant cancers. Radiology 2016;280:108-16. doi: 10.1148/radiol.16151133.
26. Kasel-Seibert M, Lehmann T, Aschenbach R, Guettler FV, Abubrig M, Grimm MO, *et al.* Assessment of PI-RADS v2 for the detection of prostate cancer. Eur J Radiol 2016;85:726-31. doi: 10.1016/j.ejrad.2016.01.011.
27. Wang X, Wang JY, Li CM, Zhang YQ, Wang JL, Wan B, *et al.* Evaluation of the prostate imaging reporting and data system for magnetic resonance imaging diagnosis of prostate cancer in patients with prostate-specific antigen <20 ng/ml. Chin Med J 2016;129:1432-8. doi: 10.4103/0366-6999.183419.
28. Auer T, Edlinger M, Bektic J, Nagele U, Herrmann T, Schäfer G, *et al.* Performance of PI-RADS version 1 versus version 2 regarding the relation with histopathological results. World J Urol 2017;35:687-93. doi: 10.1007/s00345-016-1920-5.
29. Feng ZY, Wang L, Min XD, Wang SG, Wang GP, Cai J, *et al.* Prostate cancer detection with multiparametric magnetic resonance imaging: Prostate imaging reporting and data system version 1 versus version 2. Chin Med J 2016;129:2451-9. doi: 10.4103/0366-6999.191771.
30. Siddiqui MM, Rais-Bahrami S, Truong H, Stamatakis L, Vourganti S, Nix J, *et al.* Magnetic resonance imaging/ultrasound-fusion biopsy significantly upgrades prostate cancer versus systematic 12-core transrectal ultrasound biopsy. Eur Urol 2013;64:713-9. doi: 10.1016/j.eururo.2013.05.059.

PI-RADS v2诊断效能对六名不同经验水平（半年至17年）的前列腺影像医师诊断一致性的评价研究

摘要

背景：最新版的前列腺影像报告和数据系统（PI-RADS v2）的主要目的之一是减少不同影像医师间对前列腺影像解读的差异性，尤其是针对具有不同经验水平的影像医师。本研究旨在回顾性分析6名具有不同经验水平的影像医师在诊断前列腺疾病中一致性和准确性，并评估使用PI-RADS v2检测临床上显著性前列腺癌的诊断效能。

方法：本研究共纳入183例（从2014年12月到2016年3月）在前列腺穿刺活检前均接受了3.0T多参数磁共振（Mp-MRI）检查的患者，其中包括84例前列腺癌（PCa）和99例良性前列腺增生（BPH）。6名具有不同经验水平的影像医师（分别为6个月、2、3、4、5及17年，最后一位曾参与PI-RADS v2撰写和讨论）基于PI-RADS v2对所有患者分别进行评分（1-5分）。采用Kendall相关系数来分析读者评分与Gleason评分（GS）之间的相关性；采用Kappa一致性检验来评估读者内及读者间的一致性；同时采用ROC曲线和曲线下面积（AUC）分析评估不同评分的诊断效能。

结果：在PI-RADS v2的基础上，6名读者间一致性的平均值及标准误分别为0.506和0.043；6名读者的评分与GS间为正相关，平均相关系数为r=0.319，*P*=0.006，相关程度为弱到中等。6名读者的AUC值分别为0.883，0.924，0.927，0.932，0.929和0.947。

结论：6名具有不同经验水平的影像医师间的平均一致性为弱到中等，因此不同的经验水平对MRI图像的解读具有一定影响。

Supplementary Table 1: Overall and pair-wise inter-reader differences according to PI-RADS v2 score

All reader pairs	*F*	*P*	PCa reader pairs	*F*	*P*	BPH reader pairs	*F*	*P*
N = 183	39.42	0.00	*n* = 84	32.09	0.00	*n* = 99	97.45	0.00
1 and 2	0.91	0.63	1 and 2	5.29	0.07	1 and 2	0.19	0.91
1 and 3a	4.15	0.13	1 and 3a	0.81	0.67	1 and 3a	9.47	0.01
1 and 4	5.35	0.07	1 and 4	8.26	0.02	1 and 4	7.19	0.03
1 and 5	14.16	0.00	1 and 5	12.22	0.00	1 and 5	35.72	0.00
1 and 6	0.35	0.84	1 and 6	8.41	0.01	1 and 6	0.67	0.72
2 and 3a	8.94	0.01	2 and 3a	10.25	0.01	2 and 3a	12.34	0.00
2 and 4	1.85	0.40	2 and 4	0.33	0.85	2 and 4	5.04	0.08
2 and 5	7.89	0.02	2 and 5	1.43	0.49	2 and 5	30.70	0.00
2 and 6	0.13	0.94	2 and 6	0.36	0.83	2 and 6	1.57	0.46
3a and 4	18.93	0.00	3a and 4	14.26	0.00	3a and 4	33.15	0.00
3a and 5	33.64	0.00	3a and 5	19.34	0.00	3a and 5	81.97	0.00
3a and 6	6.89	0.03	3a and 6	14.45	0.00	3a and 6	5.11	0.08
4 and 5	2.10	0.35	4 and 5	0.39	0.82	4 and 5	10.86	0.00
4 and 6	2.98	0.23	4 and 6	0.00	1.00	4 and 6	12.23	0.00
5 and 6	10.08	0.01	5 and 6	0.35	0.84	5 and 6	46.14	0.00

PCa: Prostate cancer; BPH: Benign prostatic hyperplasia; PI-RADS v2: Prostate Imaging Reporting and Data System Version 2.

Supplementary Table 2: Diagnostic performance of PI-RADS v2 scores from six readers

Readers	Threshold ≤3 (*n* = 183)*					
	AUC (95% *CI*)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)
1	0.883 (0.828–0.926)	77.4	91.9	89.0	82.7	85.3
2	0.924 (0.876–0.958)	88.1	89.9	88.1	89.9	90.2
3a	0.927 (0.879–0.960)	86.9	87.9	85.9	88.8	88.0
4	0.932 (0.885–0.963)	96.4	76.8	77.9	96.2	86.3
5	0.929 (0.882–0.962)	95.2	78.8	79.2	95.1	88.4
6	0.947 (0.903–0.974)	92.9	86.9	85.7	93.5	90.2
Readers	**Threshold ≥4 (*n* = 84)†**					
	AUC (95% *CI*)	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**	**NPV (%)**	**Accuracy (%)**
1	0.678 (0.567–0.776)	71.2	64.0	82.4	48.5	72.6
2	0.592 (0.479–0.698)	86.4	32.0	75.0	50.0	75.0
3a	0.721 (0.612–0.813)	74.6	60.0	81.5	50.0	79.8
4	0.615 (0.503–0.719)	89.8	32.0	75.7	57.1	73.8
5	0.646 (0.543–0.747)	96.6	32.0	77.0	80.0	73.8
6	0.557 (0.445–0.666)	88.1	24.0	73.2	46.2	70.2

*Cutoff value for differentiating between benign and malignant cases was set at 3, with values and Data System Version. †The cutoff value for differentiating between low-risk and clinically significant PCa was set at 4, with values ≥4 considered positive. PPV: Positive predictive value; NPV: Negative predictive value; AUC: Area under the curve; PI-RADS v2: Prostate Imaging Reporting and Data System Version 2; *CI*: Confidence interval.

Supplementary Table 3: AUC values of overall PI-RADS scores for cancer detection and PCa PI-RADS scores for clinically significant cancer detection

Reader pairs	Overall (*n* = 183)		PCa (*n* = 84)	
	Z	*P*	*Z*	*P*
1 and 2	1.680	0.929	1.671	0.095
1 and 3a	1.646	0.100	0.914	0.361
1 and 4	1.782	0.075	1.121	0.262
1 and 5	1.583	0.113	0.579	0.563
1 and 6	2.341	0.019	2.150	0.032
2 and 3a	0.147	0.883	2.630	0.009
2 and 4	0.376	0.707	0.633	0.527
2 and 5	0.214	0.831	1.450	0.142
2 and 6	1.323	0.186	1.189	0.234
3a and 4	0.271	0.787	2.048	0.041
3a and 5	0.121	0.904	1.477	0.140
3a and 6	0.940	0.347	3.024	0.003
4 and 5	0.175	0.861	0.942	0.346
4 and 6	0.789	0.430	1.682	0.093
5 and 6	0.844	0.399	2.442	0.015

AUC: Area under the curve; PCa: Prostate cancer; BPH: Benign prostatic hyperplasia; PI-RADS: Prostate Imaging Reporting and Data System.