Inter-Reader Variability Using PI-RADS v2 Versus PI-RADS v2.1: Most New Disagreement Stems from Scores 1 and 2

Ziel Analyse der möglichen Differenzen in der Interrater-Reliabilität zwischen PI-RADS Version 2 (v2) und Version 2.1

ABSTRACT

Purpose To analyze possible differences in the inter-reader variability between PI-RADS version 2 (v2) and version 2.1 (v2.1) for the classification of prostate lesions using multiparametric MRI (mpMRI) of the prostate.

Methods In this retrospective and randomized study, 239 annotated and histopathologically correlated prostate lesions (104 positive and 135 negative for prostate cancer) were rated twice by three experienced uroradiologists using PI-RADS v2 and v2.1 with an interval of at least two months between readings. Results were tabulated across readers and reading timepoints and inter-reader variability was determined using Fleiss’ kappa (κ). Thereafter, an additional analysis of the data was performed in which PI-RADS scores 1 and 2 were combined, as they have the same clinical consequences.

Results PI-PI-RADS v2.1 showed better inter-reader agreement in the peripheral zone (PZ), but poorer inter-reader agreement in the transition zone (TZ) (PZ: \( \kappa = 0.63 \) vs. \( \kappa = 0.58 \); TZ: \( \kappa = 0.47 \) vs. \( \kappa = 0.57 \)). When PI-RADS scores 1 and 2 were combined, the use of PI-RADS v2.1 resulted in almost perfect inter-reader agreement in the PZ and substantial agreement in the TZ (PZ: \( \kappa = 0.81 \); TZ: \( \kappa = 0.80 \)).

Conclusion PI-RADS v2.1 improves inter-reader agreement in the PZ. New differences in inter-reader agreement were mainly the result of the assignment of PI-RADS v2.1 scores 1 and 2 to lesions in the TZ. Combining scores 1 and 2 improved inter-reader agreement both in the TZ and in the PZ, indicating that refined definitions may be warranted for these PI-RADS scores.

Key Points:
- PI-RADSv2.1 improves inter-reader agreement in the PZ but not in the TZ.
- New differences derived from PI-RADSv2.1 scores 1 and 2 in the TZ.
- Combined PI-RADSv2.1 scores of 1 and 2 yielded better inter-reader agreement.
- PI-RADSv2.1 appears to provide more precise description of lesions in the PZ.
- Improved inter-reader agreement in the PZ stresses the importance of appropriate lexicon description.
Introduction

Multiparametric magnetic resonance imaging (mpMRI) of the prostate for suspected prostate cancer aims at detecting clinically significant prostate cancer (csPCa) and has gained support from international guidelines in order to reduce overdiagnosis and overtreatment [1–3]. The PROMIS and PRECISION trials have shown that mpMRI not only reduces unnecessary biopsies and overdiagnosis of clinically non-significant prostate cancer (cnPCa) but also improves the detection of csPCa [4, 5]. Additionally, mpMRI is superior in detecting anteriorly located prostate cancers, which are frequently missed by systematic biopsies [6, 7]. Consequently, MRI/ultrasound-fusion biopsies have shown better prediction of the final Gleason score, which is crucial for initial treatment recommendation [8, 9].

In order to standardize prostate MRI examination and reporting, the European Society of Urogenital Radiology (ESUR) introduced Prostate Imaging Reporting and Data System (PI-RADS) version 1 (v1), the first structured reporting scheme for evaluating prostate MRI in pre-therapy patients, in 2012 [10]. Limitations of PI-RADS v1 included poor inter-reader agreement, as well as an unsharp definition of positive and negative examinations regarding suspected prostate cancer [11, 12]. In collaboration with the AdMeTech Foundation and American College of Radiology (ACR), the ESUR updated the reporting scheme to PI-RADS version 2 (v2) in 2015, successfully improving diagnostic accuracy and acceptance among urologists and radiologists [13–15].

An update, PI-RADS version 2.1 (v2.1), was published in 2019, aiming to further improve inter-reader and intra-reader agreement [16]. It introduces three major changes to the scoring system: first, typical benign prostate hyperplasia (BPH) nodules (completely encapsulated and round) in the transition zone (TZ) are now classified as PI-RADS score 1. Second, atypical BPH nodules in the TZ are assigned PI-RADS score 2 with possible upgrading to PI-RADS score 3 based on the results of diffusion-weighted imaging (DWI) and corresponding apparent diffusion coefficient (ADC) maps. Only atypical BPH nodules (without or with incomplete encapsulation) or a homogeneous mildly hypointense area between nodules are assigned a score of 2. Third, in the peripheral zone (PZ), a score of 2 is now defined as a wedge-shaped or linear DWI-hyperintense or ADC-hypointense lesion, whereas a score of 3 is defined as a focal DWI hyperintense and/or focal ADC-hypointense lesion.

Although the PI-RADS score does not lead to definite clinical recommendations and other factors including prostate-specific antigen (PSA) level and clinical history have to be considered, PI-RADS scores 4 and 5 indicate a high likelihood of csPCa and favor prostate biopsy. In contrast, PI-RADS scores 1 and 2 suggest a low likelihood of csPCa and prostate biopsy is usually not performed [17, 18].

Given the significance of the PI-RADS score for clinical decision making, further improvement of inter-reader agreement of the updated PI-RADS v2.1 is desired. Therefore, this randomized and controlled study with a retrospective dataset aims to compare inter-reader agreement between PI-RADS v2 and PI-RADS v2.1. In addition, we tested if combining PI-RADS score 1 and score 2 may have an impact on inter-reader agreement.
Materials and Methods

Study design
In this single-center study, three highly experienced uroradiologists interpreted annotated prostate lesions on mpMRI from a retrospective dataset in a randomized fashion, first according to PI-RADS v2 and second according to PI-RADS v2.1. The institutional review board approved this study.

Reference standard
All patients underwent 10-core systematic and 2-8 targeted MRI/TRUS fusion-guided biopsies with histopathological analysis, which served as a reference standard. Prostate cancer was graded using the Gleason classification system [19]. Gleason scores of $\geq 3 + 4$ were considered to indicate csPCa [20, 21]. Reported benign findings included chronic and acute prostatitis, BPH, and normal prostate tissue. The biopsies were obtained using one of two biopsy devices (Hitachi Medical Systems, HI VISION Preirus, or Toshiba, Aplio 500) and 18-gauge needles. All biopsies were executed by a team of experienced urologists and uroradiologists at our tertiary university center.

Lesion annotation and reviewing software
For each prostate mpMRI dataset, 2–4 lesions representative for the relevant histopathological findings were defined and annotated in consensus by two radiologists (N.L.B. and T.P.) based on a review of histopathological and imaging data. These annotations included normal appearing prostate tissue, benign changes, and cancerous lesions. Unclear or ambiguous areas (e.g., if the targeted biopsy in a specific area was non-cancerous but an adjacent systematic biopsy yielded prostate cancer) were not used as possible review targets.

A dedicated proprietary reviewing software was used to ensure a randomized and controlled rating process as well as blinding of readers to clinical data and patient information. Annotated lesions were automatically presented in random order. The readers had access to the complete examination and were allowed to scroll through the dataset without limitation.

PI-RADS v2 and v2.1 reading
In the first round the three readers were instructed to independently rate the annotated lesion according to PI-RADS version 2. After an interval of at least 8 weeks, the readers were instructed to independently re-evaluate the same lesions according to PI-RADS version 2.1. During the review process, the annotated lesions were automatically displayed in all MRI sequences. The reading results were recorded using a digital questionnaire within the reviewing software.

Each mpMRI dataset was read by three senior board-certified uroradiologists from a high-volume university center. All three readers (A.J.D.B., M.H., and P.A.) are attending-level uroradiologists with several years (> 7 years) of experience in prostate MRI and hold the highest available certificate for prostate mpMRI from the German Society of Radiology (DRG Q2 certificate). Each of them has read and supervised at least 1000 prostate mpMRI datasets within the last 5 years.

In the first round, the three readers were instructed to independently rate the annotated lesions according to PI-RADS v2. After an interval of at least two months, the readers were instructed to re-evaluate the same annotated lesions according to PI-RADS v2.1. In both rounds, the annotated lesions were presented randomly to control for potential bias. The readers were handed reporting charts describing the rating system according to PI-RADS v2 and PI-RADS v2.1 for reference (Fig. 1 and Fig. S1).

Inter-reader agreement was assessed for the primary scoring results, followed by a secondary analysis to assess how inter-reader agreement is affected when PI-RADS score 1 and score 2 are combined.

Patient cohort, patient characteristics, and biopsy results
The study included patients with suspected prostate cancer due to an abnormal digital rectal examination or an elevated PSA level. The availability of a prostate mpMRI and a subsequent in-house targeted MRI/TRUS fusion biopsy combined with a 10-core systematic biopsy were further inclusion criteria.

Exclusion criteria were prior prostate-related therapies such as brachytherapy, radiation, or hormonal treatment. Additionally, six patients were excluded because of non-diagnostic image quality due to motion artifacts and hip replacement. A total of 102 patients with a mean age of 68.7 ± 8.1 years and a mean PSA level of 10.1 ± 9.3 ng/dl were included. Overall, 239 previously annotated lesions were rated in the study.

The cohort is a subset of patients from a consecutive MRI/TRUS biopsy cohort recruited at a tertiary university hospital who met the inclusion criteria. The subset included 107 patients, of whom 8 were on active surveillance and 33 had undergone biopsy before.

In the TZ, 68 (65 %) of the annotated lesions were normal prostate tissue, 2 (2 %) inflammation, 12 (12 %) corresponded to a Gleason score $3 + 3$, 7 (7 %) to a Gleason score $3 + 4$, 4 (4 %) to a Gleason score $4 + 3$, 8 (8 %) to a Gleason score $4 + 4$ and 1 each (1 %) to a Gleason score $5 + 3$, $5 + 4$ and $5 + 5$. In the PZ, 59 (44 %) of the annotated lesions were normal prostate tissue, 6 (4 %) inflammation, 25 (19 %) corresponded to a Gleason score $3 + 3$, 12 (9 %) to a Gleason score $3 + 4$, 2 (1 %) to a Gleason score $3 + 5$, 7 (5 %) to a Gleason score $4 + 3$, 14 (10 %) to a Gleason score $4 + 4$, 6 (4 %) to a Gleason score $4 + 5$, and 4 (3 %) to a Gleason score $5 + 4$. The histopathological biopsy results of all 239 lesions are compiled in Table 1. The classifications according to the International Society of Urological Pathology (ISUP) are compiled in Table 2.

Image acquisition
A standardized mpMRI protocol in accordance with the ESUR guidelines was used for prostate imaging in all patients [22]. The protocol included axial T1-weighted imaging ($3.0 \times 0.6 \times 0.6$ mm, 32 cm FoV), axial and coronal T2-weighted imaging ($3.0 \times 0.47 \times 0.47$ mm, 18 cm FoV), DWI with generation of ADC maps ($3.0 \times 1.4 \times 1.4$ mm, 17 cm FoV, with b-values of 0, 50, 500, 1000, and calculated $b = 1400$ s/mm²), and dynamic contrast-en-
Hanced (DCE) imaging (3.0 × 1.4 × 1.4 mm, 18.6 cm FoV, 5 s temporal resolution) after intravenous injection of Gd-DO3A-butrol (Gadovist, Bayer Healthcare) at a rate of 3 ml/s. All mpMRI examinations were performed on one of two identical 3-Tesla MRI scanners (Skyra, Siemens Healthineers).

**Table 1** Histopathological results of 239 lesions used to assess PI-RADS in terms of inter-reader agreement in a) the transition zone and b) the peripheral zone. GS = Gleason score.

<table>
<thead>
<tr>
<th>a) Transition zone</th>
<th>b) Peripheral zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biopsy result</td>
<td>Number</td>
</tr>
<tr>
<td>Total</td>
<td>104</td>
</tr>
<tr>
<td>Normal tissue/BPH</td>
<td>68</td>
</tr>
<tr>
<td>Inflammation</td>
<td>2</td>
</tr>
<tr>
<td>GS 3 + 3</td>
<td>12</td>
</tr>
<tr>
<td>GS 4 + 3</td>
<td>7</td>
</tr>
<tr>
<td>GS 4 + 4</td>
<td>8</td>
</tr>
<tr>
<td>GS 5 + 3</td>
<td>1</td>
</tr>
<tr>
<td>GS 5 + 4</td>
<td>1</td>
</tr>
<tr>
<td>GS 5 + 5</td>
<td>1</td>
</tr>
</tbody>
</table>

**Statistical analysis and performance correlation**

**Agreement analysis:** Fleiss' kappa (κ) and Krippendorf's alpha (α) were used to assess inter-reader agreement. Both measures were analyzed together to determine inter-reader agreement between...
all three readers. Additionally, inter-reader agreement for combined PI-RADS scores 1 and 2 was calculated. Fleiss’ kappa results were interpreted as follows: ‘κ’ values of 0.41–0.60 indicate moderate agreement, ‘κ’ values of 0.61–0.80 indicate substantial agreement and ‘κ’ values > 0.81 indicate almost perfect agreement. Krippendorff’s alpha (α) was designed to compute inter-reader agreement between multiple ratings (readers 1, 2, 3). Values range between α = 1 for perfect reliability and α = 0 for absence of reliability.

Accuracy analysis: The study has been modeled as a factorial diagnostic trial involving the factors Reader (levels: readers 1, 2, 3) and Modality (PI-RADS versions 2.1 and 2.0). Diagnostic accuracies of the methods were assessed with the areas under the receiver operating characteristic curves (AUC) for each reader and modality combination. It should be noted that we assumed all lesions to be independent. A clustered-data AUC analysis revealed the same conclusions [23].

R statistical software was used for all data analysis (version 3.6.1; www.r-project.org).

Results

PI-RADS classification

In both reading rounds, a majority of 91% of annotated lesions were assigned PI-RADS scores of 1 or 2 for low likelihood of csPCa and PI-RADS scores 4 or 5 for high likelihood of csPCa based on mpMRI. Only 9% of annotated lesions were PI-RADS score 3 for indeterminate likelihood of csPCa based on mpMRI (64 lesions using PI-RADS v2 vs. 66 lesions using PI-RADS v2.1). An overview of PI-RADS scores assigned by all three readers is compiled in Table 3.

Distribution of PI-RADS v2 and v2.1 classification for different histopathological findings

Histopathological normal or benign findings were rated 8x score 1, 180x score 2, 9x score 3, and 7x score 5 according to PI-RADS v2, whereas when using PI-RADS v2.1 the same lesions were rated 101x score 1, 79x score 2, 11x score 3, and 5x score 4, and 8x score 5. Inflammation was rated 6x score 2 using PI-RADS v2 and 4x score 1 and 2x score 2 using PI-RADS v2.1. cnsPCa was rated 10x score 2, 9x score 3, 4x score 4, and 13x score 5 using PI-RADS v2, whereas with PI-RADS v2.1 the lesions were rated 1x score 1, 12x score 2, 7x score 3, 3x score 4, and 13x score 5. csPCa was rated 7x score 2, 4x score 3, 10x score 4, and 44x score 5 using PI-RADS v2, whereas with PI-RADS v2.1 the lesions were rated 1x score 1, 1x score 2, 15x score 3, 2x score 4, and 47x score 5. The results are shown in Table 4.

Inter-reader agreement

Inter-reader agreement for lesions in the PZ was substantial using PI-RADS v2.1 and moderate using PI-RADS v2 (α = 0.63 and κ = 0.63 vs. α = 0.58 and κ = 0.58). In the TZ the inter-reader agree-
ment was moderate for both PI-RADS v2.1 and PI-RADS v2 ($\alpha = 0.47$ and $\kappa = 0.47$ vs. $\alpha = 0.57$ and $\kappa = 0.57$).

Alluvial plots for the visualization of differences in scores assigned by the three readers using PI-RADS v2 versus PI-RADS v2.1 show that, for the reading using PI-RADS v2.1, most inter-reader differences were attributable to lesions assigned PI-RADS scores 1 and 2 (Fig. 2). We therefore performed an additional analysis combining these two scores, as both scores have identical clinical consequences. Examples of lesions showing improved inter-reader agreement in the PZ and decreased inter-reader agreement in TZ using PI-RADS v2.1 are shown in Fig. 3A-D.

Using PI-RADS v2.1, agreement differences of $\geq 2$ score points appeared in 17 out of 135 cases ($12\%$) in the PZ and in 14 out of 104 in the TZ ($13\%$). In the TZ there were 5 cases ($5\%$) in which csPCa was rated with a difference of $\geq 2$ score points: 4 lesions were rated PI-RADS v2.1 score 5 by two raters and score 3 by one rater, and another lesion was scored PI-RADS v2.1 score 5 by two raters and score 2 by one rater. In contrast, in the PZ no disagreement with a difference of $\geq 2$ score points was observed for csPCa. All differences of $\geq 2$ score points separated for benign/normal prostate tissue, inflammation, cnsPCa, and csPCa are shown in Table 5.

**Table 4** Distribution of PI-RADS v2 and v2.1 classification for different histopathological findings for a) the TZ and b) the PZ. One histological finding showed atypical small acinar proliferation in the PZ which is not shown in this table. TZ = transition zone, PZ = peripheral zone, csPCa = clinically significant prostate cancer, cnsPCa = clinically non-significant prostate cancer.

<table>
<thead>
<tr>
<th></th>
<th>Benign/normal</th>
<th>Inflammation</th>
<th>cnsPCa</th>
<th>csPCa</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) TZ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score 1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Score 2</td>
<td>180</td>
<td>79</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Score 3</td>
<td>9</td>
<td>11</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Score 4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Score 5</td>
<td>7</td>
<td>8</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>14</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Score 1</td>
<td>59</td>
<td>93</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Score 2</td>
<td>71</td>
<td>41</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Score 3</td>
<td>22</td>
<td>17</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Score 4</td>
<td>25</td>
<td>23</td>
<td>2</td>
<td>43</td>
</tr>
<tr>
<td>Score 5</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>b) PZ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score 1</td>
<td>8</td>
<td>101</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Score 2</td>
<td>180</td>
<td>79</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Score 3</td>
<td>9</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Score 4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Score 5</td>
<td>7</td>
<td>8</td>
<td>0</td>
<td>13</td>
</tr>
</tbody>
</table>

**Discussion**

In this study we compared inter-reader agreement between PI-RADS v2 and v2.1 in two reading rounds with randomized presentation of a set of predefined lesions in mpMRI datasets. When PI-RADS v2.1 was used and all scores were analyzed separately, inter-reader agreement improved from moderate to substantial in the TZ but worsened in the TZ compared with PI-RADS v2. Visualization of the results revealed that most differences in inter-reader agreement occurred in the assignment of the two lower PI-RADS scores of 1 and 2, which indicate low csPCa likelihood and usually have the same clinical consequences. In a further analysis combining the PI-RADS scores 1 and 2, PI-RADS v2.1 showed improved (almost perfect) inter-reader agreement in the TZ and comparable inter-reader agreement in the PZ.

Major changes introduced with the PI-RADS v2.1 update relate to the TZ, and it is therefore plausible to assume that the readers were mostly divided over rating a lesion as “typical” or “atypical” BPH nodules, even though it might be expected that further experience with PI-RADS v2.1 could cause inter-reader agreement to level off in the future. While the modifications regarding the
PZ are relatively minor, our results indicate that PI-RADS v2.1 appears to provide a more precise description of lesions in the PZ. This improvement stresses the importance of appropriate lexicon descriptors [24]. Fittingly, in our study there was no inter-reader disagreement with a difference of ≥2 score points for lesions harboring csPCa in the PZ, whereas in the TZ 5% of lesions harboring csPCa were rated with a difference of ≥2 score points using PI-RADS v2.1.

The assignment of a PI-RADS score of 1 or 2 indicates a low likelihood of csPCa, and, on the basis of these mpMRI findings, biopsy is not performed in most cases. A PI-RADS score of 4 or 5 indicates a high likelihood of csPCa, and targeted biopsy should be included in the further management recommendations [25]. In contrast, an indeterminate PI-RADS score of 3 might augment uncertainty regarding management recommendations for biopsy [26]. Consequently, a low rate of PI-RADS 3 assignments is favorable. In our study, only nine percent of all annotated lesions were assigned PI-RADS score 3 with use of both PI-RADS v2 and v2.1, indicating a high diagnostic certainty with both PI-RADS versions. Even though the number of csPCa lesions is low, one has to bear in mind that, in the clinical routine, there is a large proportion of PI-RADS scores 1, 2, and 3. Patients with these scores have a low to indeterminate likelihood of csPCa based on the mpMRI and therefore might not undergo prostate biopsy. Nevertheless, these ratings contribute to the overall inter-reader agreement using PI-RADS and need to be evaluated.

Several recently published studies have compared the inter-reader agreement of PI-RADS v2 and v2.1. First, Tamada et al.
report better inter-reader agreement for TZ lesions scored using PI-RADS v2.1 for the TZ, but the cohort was comparably small [27]. Brembilla et al. describe substantial inter-reader agreement for index lesions [28]. On the other hand, Bhayana et al. report an improved inter-reader agreement using PI-RADS v2.1 in the PZ but not in the TZ, which is consistent with our results [29]. None of these earlier investigators addressed possible underlying causes of the observed differences in inter-reader agreement.

We attribute the lower inter-reader agreement in the TZ to a notable discrepancy in the assignment of BPH nodules to PI-RADS scores 1 and 2, as confirmed by the improved inter-reader agreement in the TZ when PI-RADS v2.1 scores 1 and 2 were combined. Additionally, in accordance with other comparative studies we have demonstrated improved diagnostic performance of PI-RADS V2.1 in the TZ, whereas the diagnostic performance remained unchanged in the PZ [28, 30].
Whereas the first revision of PI-RADS introduced major changes by defining dominant MRI sequences for the PZ and the TZ, the next update v2.1 presented only fairly minor modifications. Indeed, PI-RADS v2 quickly gained recognition from urologists and radiologists, although initial inter-reader studies comparing PI-RADS v1 and v2 were controversial, too. Especially the new zone-specific definition of dominant MRI sequences and the downgrading of the DCE sequences initially raised skepticism [31, 32]. The PI-RADS classification is an ongoing “work-in-progress”, and further refinement of lesion descriptions, as seen in the PZ in PI-RADS v2.1, will help improve future versions of the PI-RADS scoring system.

Limitations

The study is limited by the use of a retrospective dataset. Moreover, owing to the greater level of longer experience with the earlier PI-RADS classification system, readers might have been biased and rated the annotated lesions more correctly according to PI-RADS v2 compared to PI-RADS v2.1. Our study design using predefined annotated lesions reduces bias but differs from clinical practice, which might decrease the overall generalizability of our results.

Conclusion

PI-RADS v2.1 improves inter-reader agreement in the PZ. Differences in inter-reader agreement were mainly the result of the assignment of PI-RADS v2.1 scores 1 and 2 to lesions in the TZ, which improved when combining scores 1 and 2. These findings suggest that more refined definitions of these scores may be needed.

Conflict of Interest

The authors declare that they have no conflict of interest.

References


