PI–RADS v2: Current standing and future outlook

Clayton P. Smith, Barış Türkbey

Cite this article as: Smith CP, Türkbey B. PI–RADS v2: Current standing and future outlook. Turk J Urol 2018; 44(3): 189-94.

ABSTRACT

The Prostate Imaging-Reporting and Data System (PI–RADS) was created in 2012 to establish standardization in prostate multiparametric magnetic resonance imaging (mpMRI) acquisition, interpretation, and reporting. In hopes of improving upon some of the PI–RADS v1 shortcomings, the PI–RADS Steering Committee released PI–RADS v2 in 2015. This paper reviews the accuracy, interobserver agreement, and clinical outcomes of PI–RADS v2 and comments on the limitations of the current literature. Overall, PI–RADS v2 shows improved sensitivity and similar specificity compared to PI–RADS v1. However, concerns exist regarding interobserver agreement and the heterogeneity of the study methodology.

Keywords: Biopsy; magnetic resonance imaging; prostate cancer.

Prostate cancer is one of the most common cancers worldwide, ranking among the top five cancers for both incidence and mortality. \[1\] There have been many changes in the way health care providers screen, diagnose and treat the disease over the past few decades. Multiparametric magnetic resonance imaging (mpMRI) of the prostate has become more commonly utilized for numerous purposes including tumor detection and characterization, risk stratification, and image guidance for biopsy.\[2\] Despite the increased utilization of prostate mpMRI, no universal system existed for radiologists to use when reading, interpreting, and reporting these scans prior to the conception of the Prostate Imaging-Reporting and Data System (PI–RADS). This made it difficult to ascertain diagnostic accuracy of mpMRI in the detection of the prostate cancer.\[3\] Addressing this unmet need in 2012, the European Society of Urogenital Radiology (ESUR) created PI–RADS.\[4\] PI–RADS version 1 (PI–RADS v1) included instructions for the interpretation of T2-weighted (T2W), diffusion-weighted imaging (DWI), dynamic contrast-enhanced (DCE), and magnetic resonance spectroscopy (MRS) and instructed readers to score each lesion using a 5-point scale. However, the guidelines did not explain how to establish an overall score for each lesion, which caused confusion and inability to standardize scoring.\[5\] Multiple PI–RADS v1 validation studies ensued with reasonable accuracy and moderate to good inter-reader agreement values.\[6-8\]

In a 14 study meta-analysis, Hamoen et al.\[9\] reported a pooled sensitivity and specificity for PI–RADS v1 as 0.78 (95% CI, 0.70-0.84) and 0.79 (95% CI, 0.68-0.86), respectively. Limitations of PI–RADS v1 included use of a summed final score contributed equally by each pulse sequence, DCE MRI evaluated with curve type analysis and incorporation of MRS, which is used very rarely.\[10-12\] Realizing the shortcomings and strengths of PI–RADS v1 and also in consideration of additional clinical experience, a joint steering committee consisting of the American College of Radiology (ACR), ESUR, and the non-profit organization AdMeTech Foundation came together to revise the original system to create PI–RADS v2.

PI–RADS v2

Since the conception of PI–RADS v2, many studies have evaluated its accuracy, interobserver agreement, and clinical outcomes.\[13-31\]
Herein, we will summarize clinical results, interobserver agreement level and limitations of PI-RADS v2.

Clinical results
Although PI-RADS v2 is a relatively new system incorporated into radiology and urology practice in early 2015, there have been several research articles, majority of which are retrospective in nature. Based on this growing literature, in a meta-analysis of 21 studies (3,857 patients) concerning the diagnostic performance of PI-RADS v2, Woo et al.\[^{13}\] reported a pooled sensitivity and specificity of 0.89 (95% CI 0.86-0.92) and 0.73 (95% CI 0.60-0.83), respectively. In 6 of 21 studies, head-to-head comparisons performed between PI-RADS v1 and 2, showed a higher pooled sensitivity for PI-RADS v2 0.95 (95% CI 0.85-0.98) than PI-RADS v1 0.88 (95% CI 0.80-0.93) (p=0.04). However, pooled specificity was similar between the two systems (0.73 [95% CI 0.47-0.89] vs. 0.75 [95% CI 0.36-0.94]; p=0.90).\[^{20,21,27-35}\] Woo et al.\[^{13}\] did not perform a meta-regression analysis or subgroup analysis of the quality of the reader (experience level), which may have contributed greatly to the results of the studies included in the meta-analysis. Another meta-analysis by Zhang et al.\[^{14}\] focused on the detection and accuracy of PI-RADS v2, and consisted of 13 studies (2,049 patients). The analysis revealed pooled sensitivity and specificity of 0.85 (95% CI 0.78-0.91) and 0.71 (95% CI 0.60-0.80), respectively. Both meta-analyses revealed high overall sensitivity and moderate specificity. Nine of the 13 studies in the Zhang et al.\[^{14}\] analysis were also included in the Woo et al.\[^{13}\] analysis. One of the main limitations of the literature that was mentioned by both meta-analyses was the large heterogeneity in methodology in both cancer threshold definition and tissue confirmation. Radical prostatectomy (the gold standard) was the reference standard in 5/21 and 4/13 studies performed by Woo et al.\[^{13}\] and Zhang et al.\[^{14}\], respectively. The remaining studies used systematic biopsy, targeted biopsy or both for tissue confirmation. Both meta-analyses also commented on the heterogeneous mix of per-lesion and per-patient analysis among studies. Finally, all but three studies included in both analyses were retrospective in nature. One of these prospective studies was conducted by Rastinehad et al.\[^{27}\]. In a multi-institutional, multi-reader (3 experts) study including 312 patients using a cut-off of PI-RADS category 3, they found an overall sensitivity and specificity of 0.96 (95% CI 0.90-0.97) and 0.32 (95% CI 0.25-0.40), respectively. The overall quality of the PI-RADS v2 literature is limited by the lack of prospective studies.

PI-RADS v2 has been somewhat disappointing in its ability to detect clinically significant (CS) cancer.\[^{23-25,28,36}\] Most retrospective and prospective studies that have evaluated the cancer detection rate (CDR) of PI-RADS v2 have suggested a correlation between higher PI-RADS v2 categories and higher CDRs for all prostate cancer and CS cancer.\[^{18,25,29,37,38}\] However, some studies have shown relatively poor correlation. Mehralivand et al.\[^{25}\] showed highest CDR in PI-RADS 5 lesions (86.9%) and only 39.1% for category 4 lesions. This low CDR for category 4 is contradictory to its definition of “clinically significant cancer likely to be present”.\[^{19}\] An editorial comment was written at the end of the Mehrelivand et al.\[^{25}\] study remarking that the inclusion criteria and screening patterns may have contributed to the low sensitivity found for categories 3 and 4 for significant cancer, mentioning the high sensitivity and specificity found for categories 4 and 5 in the study by Kasel-Seibert et al.\[^{20}\].

Few studies have compared the accuracy and detection rate of PI-RADS v1 to PI-RADS v2. Auer et al.\[^{32}\] had two radiologists read the same 50 mpMRIs using different PI-RADS scoring algorithms. PI-RADS v1 had a higher AUC 0.96 (95% CI 0.94-0.98) vs. AUC 0.90 (95% CI 0.86-0.94) for PI-RADS v2. Lack of demonstration of reader agreement and difference in reader performance may be a factor in the AUC variance, however. Despite the small number of lesions included (n=39 malignant lesions), Kasel-Seibert et al.\[^{20}\] showed improved AUC values with PI-RADS v2 for experienced and inexperienced readers (0.79 vs. 0.83 and 0.70 vs. 0.83, respectively). All of these studies reveal a considerable heterogeneity in the PI-RADS v2 literature like differences in study design, accrual of subjects, experience of the assessors, and validation criteria.

Reader agreement
The simplified PI-RADS v2 lesion assessment algorithm was created to further standardize lesion scoring to produce more reliable and repeatable prostate mpMRI scoring.\[^{39}\] Several multi-reader studies have investigated the inter-reader agreement of PI-RADS v2.\[^{17-23}\] Overall, inter-reader agreement varies from poor to good depending on the study. This wide variation in reader agreement is most likely due to inconsistent methodology among studies.\[^{36}\] Single-center studies with highly experienced readers generally report higher inter-reader agreement than multicenter studies with readers having variable experience levels.\[^{17,19,21}\] The results from multicenter studies with readers of all experience levels are likely more generalizable and reflective of the current true PI-RADS v2 reader agreement landscape than those with single-center, single-reader experience level.

Muller et al.\[^{18}\] included 94 biopsy-naïve patients who underwent mpMRI of the prostate and subsequent transrectal ultrasonography-MRI (TRUS-MRI) fusion-guided biopsy. Five readers of varying experience levels scored lesions using PI-RADS v2. They revealed kappa interobserver agreement for overall suspicion score, T2W in the peripheral zone (PZ),
T2W in the transitional zone (TZ), DWI, and DCE MRI of 0.46, 0.47, 0.37, 0.40 and 0.46, respectively. In another study looking at specific agreement between readers using the pathology sector map provided by the PI-RADS v2 guidelines, six readers of varying experience levels read 30 consecutive prostate mpMRIs and drew tumor locations on their respective pathology sector maps.[40] Exact agreement (defined as agreement in each sector involved with tumor) was found to only be 21.2%. They also found poor agreement (39%) for indicating which sector was the primary sector involved in an index lesion. Another study had five radiologists (n=2 prostate dedicated, n=3 general body) read 34 consecutive prostate mpMRIs and found greater specific reader agreement between experts (0.70) than between moderate experience level readers (0.53).[23] Rosenkrantz et al.[17] looked specifically at the reader agreement of PI-RADS v2 scores and score features (focality, encapsulation, intensity, early enhancement, invasive behavior, etc.) among six highly experienced uroradiologists from six different institutions. Interestingly, even experts showed only moderate reproducibility (kappa=0.55). After undergoing a training session between reading sessions, readers’ reproducibility did not improve. Kappa in the TZ was 0.51 and higher in PI-RADS ≥4 than 3 lesions (0.55 versus 0.46, respectively). While the new PI-RADS scoring system is simpler, reader agreement for both PI-RADS v1 and 2 systems are comparable.[20,34,35] Despite the lack of major improvement in reader agreement for PI-RADS v2, there is improved standardization in research and clinical practice for lesion scoring with the new system thanks to clarifications in scoring by the PI-RADS Steering Committee.

**PI-RADS v2 limitations and future research**
The PI-RADS Steering Committee made changes to PI-RADS v1 with the goal of promoting standardization and diminishing the “variation in the acquisition, interpretation, and reporting of prostate mpMRI examinations”.[39] The sensitivity is significantly better for PI-RADS v2 than PI-RADS v1. However, there is still large heterogeneity in the PI-RADS literature methodology that needs to be addressed so that studies can provide the most ‘true’ results, which can then lead the way to better revise the current PI-RADS.

There are three fundamental areas of heterogeneity in the current literature: image acquisition, reader experience, and tissue confirmation. Esses et al.[41] assessed the variability in imaging facilities’ adherence to the minimum technical standards established by PI-RADS v2. The study reviewed 107 prostate mpMRIs from 107 unique imaging facilities after the release of PI-RADS v2 and showed that adherence was variable. Adherence was particularly poor for T2W imaging frequency resolution (16.8%) and DCE temporal resolution (31.5%). Adherence was not improved for examinations performed in 2016 than 2015.
any parameter (p>0.05). This study sheds light on the unlevel playing field that currently exists for prostate mpMRI readers and their patients. It suggests the need for greater community education and the possibility that some standards may be too stringent (Figures 1 and 2).

Variations in reader experience level may contribute to the variability in the PI-RADS v2 diagnostic literature. However, the studies that include variable ‘quality’ readers are still informative as they better reflect the PI-RADS landscape in non-academic centers. It is unknown whether reader experience affects reader agreement as some studies have shown experts have better mutual agreeability while others do not.\[17,23\]

PI-RADS studies evaluating accuracy and reader agreement use varying tissue confirmation procedures. Some use systematic/targeted biopsies and others use whole mount prostates after prostatectomy. Both have their limitations and advantages, however, comparing studies that use different practices may be misleading. Using biopsy as tissue confirmation may lead to a high false negative rate, as biopsies do not detect all cancer. Using prostatectomy specimens for tissue confirmation can select for a more aggressive lesion population, increasing the pretest likelihood of cancer. Finally, further research into the clinical benefit of DCE is warranted as the literature shows mixed results regarding its clinical benefit.\[30,43\]

Conclusion

It is obvious that PI-RADS v2, although limited in some respects, has been embraced in both radiology and urology and the scientific evidence about its clinical utility is growing with several recent papers. There are still limitations and ambiguities of PI-RADS v2 which need to be addressed by the imaging community along with their clinical collaborators.

References


