Published in final edited form as:
J Magn Reson Imaging. 2019 June ; 49(6): 1694–1703. doi:10.1002/jmri.26555.

Intra and Inter-Reader Reproducibility of PI-RADSv2: A multi-reader study

Clayton P. Smith, BA[#1,2], **Stephanie A. Harmon, PhD**[#3], **Tristan Barrett, MD**[4], **Leonardo K. Bittencourt, MD**[5,6], **Yan Mee Law, MD**[7], **Haytham Shebel, MD**[8], **Julie Y. An, BS**[9], **Marcin Czarniecki, MD**[1], **Sherif Mehralivand, MD**[1,10,11], **Mehmet Coskun, MD**[12], **Bradford J. Wood, MD**[13], **Peter A. Pinto, MD**[10], **Joanna H. Shih, PhD**[14], **Peter L. Choyke, MD**[1], and **Baris Turkbey, MD**[1]

[1] Molecular Imaging Program, National Cancer Institute, NIH, Bethesda, MD, U.S.A.

[2] Georgetown University School of Medicine, Washington, D.C., U.S.A.

[3] Clinical Research Directorate/Clinical Monitoring Research Program, Leidos Biomedical Research, Inc., NCI Campus at Frederick, Frederick, MD, U.S.A.

[4] Department of Radiology, Addenbrooke's Hospital and the University of Cambridge, Cambridge, CB2 0QQ, UK.

[5] Department of Radiology, Fluminese Federal University, Rio de Janeiro, Brazil.

[6] CDPI Clinics, DASA, Rio de Janeiro, Brazil.

[7] Department of Diagnostic Radiology, Singapore General Hospital, Singapore.

[8] Department of Radiology, Urology and Nephrology Center, Mansoura University, Mansoura City, Egypt.

[9] Northeast Ohio Medical University, Rootstown, OH, U.S.A.

[10] Urologic Oncology Branch, National Cancer Institute, NIH, Bethesda, MD, U.S.A.

[11] Department of Urology and Pediatric Urology, University Medical Center, Mainz, Germany.

[12] Department of Radiology, Dr. Behcet Uz Child Disease and Pediatric Surgery Training and Research Hospital, University of Health Sciences, İzmir, Turkey

[13] Department of Interventional Oncology, National Cancer Institute, NIH, Bethesda, MD, U.S.A.

[14] Biometric Research Program, National Cancer Institute, NIH, Rockville, MD, U.S.A.

[#] These authors contributed equally to this work.

Abstract

Corresponding Author: Baris Turkbey M.D, Address: 10 Center Drive, room B3B85 Bethesda MD, 201892 USA, turkbeyi@mail.nih.gov, Phone: 240-760-6112.

Conflict of Interest:
Author BJW is supported by the Intramural Research Program of the NIH and the NIH Center for Interventional Oncology and NIH Grant # Z1A CL040015–08. NIH and Philips / InVivo Inc have a cooperative Research and Development Agreement. NIH and Philips / InVivo Inc have a patent license agreement and NIH and BJW, BT, PAP, PLC may receive royalties. The remaining authors have no disclosures.

Background: Prostate Imaging Reporting and Data System version 2 (PI-RADSv2) has been in use since 2015, while inter-reader reproducibility has been studied, there has been a paucity of studies investigating the intra-reader reproducibility of PI-RADSv2.

Purpose: To evaluate both intra and inter-reader reproducibility of PI-RADSv2 assessment of intra-prostatic lesions at multiparametric magnetic resonance imaging (mpMRI).

Study Type: Retrospective

Population/Subjects: 102 consecutive biopsy-naïve patients who underwent prostate MRI subsequent MR/transrectal ultrasonography (MR/TRUS) guided biopsy.

Field Strength/Sequences: Prostate mpMRI at 3T using endorectal with phased array surface coils (TW MRI, DW MRI with ADC maps and b2000 DW MRI, DCE MRI).

Assessment: Previously detected and biopsied lesions were scored by four readers from four different institutions using PI-RADSv2. Readers scored lesions during two read out rounds with a four-week washout period.

Statistical Tests: Kappa (κ) statistics, specific agreement (P_o) were calculated to quantify intra and inter-reader reproducibility of PI-RADSv2 scoring. Lesion measurement agreement was calculated using Intraclass Correlation Coefficient (ICC).

Results: Overall intra-reader reproducibility was moderate to substantial ($\kappa = 0.43 – 0.67$, $P_o = 0.60 – 0.77$), while overall inter-reader reproducibility was poor to moderate ($\kappa = 0.24$, $P_o = 46$). Readers with more experience showed greater inter-reader reproducibility than readers with intermediate experience in the whole prostate ($p = 0.026$), peripheral zone ($p = 0.002$). Sequence specific inter-reader agreement for all readers was similar to overall PI-RADSv2 score, with κ=0.24, 0.24, and 0.23 and P_o =0.47, 0.44, and 0.54 in T2W, DWI, and DCE, respectively. Overall intra-reader and inter-reader ICC for lesion measurement was 0.82 and 0.71, respectively.

Data Conclusion: PI-RADSv2 provides moderate intra-reader reproducibility, poor inter-reader reproducibility, and moderate inter-reader lesion measurement reproducibility. These findings suggest a need for more standardized reader training in prostate MRI.

Keywords

Prostate cancer; MRI; PI-RADS; reader reproducibility

INTRODUCTION

The role of magnetic resonance imaging (MRI) of prostate cancer has expanded within the last decade, and now includes tumor detection and characterization, risk stratification, local staging, image guidance for biopsy and focal therapy, and treatment planning for radiation therapy (1–4). Most recently, multiparametric magnetic resonance imaging (mpMRI) of the prostate has been proven to be effective for previously undiagnosed screen-positive prostate cancer patients (5, 6). As prostate mpMRI has become increasingly utilized, the need for a standardized system for reporting findings has become even more apparent. To address this issue, the European Society of Urogenital Radiology (ESUR) introduced the first structured reporting system: the Prostate Imaging Reporting and Data System (PI-RADS), which was modeled on the BI-RADS system for breast cancer (7, 8). After its initial implementation, PI-RADS was found to have an overly complicated scoring system with wide ranging inter-reader reproducibility (2, 9–12). Following the recommendations of an international working group, PI-RADS version 2 (PI-RADSv2) emerged aiming to simplify the scoring system and standardize the nomenclature (3).

While PI-RADSv2 has simplified scoring by changing to a 5-point scale, there remain many concerns about its inter-reader reproducibility, with studies showing poor to moderate agreement (13–15). Results are often dependent on reader experience level and lesion location within the prostate, with a higher inter-reader reproducibility among lesions in the posterior peripheral zone (PZ) than in the transition zone (TZ) (13). While inter-reader reproducibility has been studied, there has been a paucity of studies investigating the intra-reader reproducibility of PI-RADSv2.

The purpose of this study was to investigate the intra and inter-reader reproducibility of PI-RADSv2 and provide further insight into the current landscape of the PI-RADSv2 reader agreement now, several years, into its implementation.

MATERIALS AND METHODS

Patient Population

This retrospective study was compliant with the Health Insurance Portability and Accountability Act and was approved by the local ethics committee. Written informed consent was obtained from all patients for future use of imaging and pathology data. One hundred and two consecutive biopsy-naïve patients (median age 64 [44–84]; median PSA 5.95 ng/ml [1.17–113.6]) who underwent mpMR imaging between May 2015 and May 2017 and subsequently underwent MRI/transrectal ultrasound (TRUS) fusion-guided biopsy were included in the study. Patients were accrued through an IRB-approved institutional protocol in which all patients undergo MRI/TRUS fusion-guided biopsy of all mpMRI-defined lesions (i.e. any lesion PI-RADS 2–5), as well as systematic 12-core biopsy. Patients were excluded if they had a previous prostate biopsy, non-diagnostic mpMRI or if they did not undergo fusion biopsy (Supplementary Figure 1).

MR Imaging Protocol

All imaging studies were performed with a combination of an endorectal coil (BPX-15; Medrad, Pittsburgh, PA) tuned to 127.8 MHz and a sixteen-channel cardiac coil with a parallel imaging (sensitivity-encoding [SENSE]; Philips Medical Systems, Best, the Netherlands) technique with a 3T MR scanner (Achieva; Philips Medical Systems). The endorectal coil was filled with 45 mL of Galden Perfluorinated Fluid (Solvay Specialty Polymers, Milan, Italy). MR imaging parameters included T1-weighted imaging, triplanar (coronal, sagittal, and axial) T2-weighted (T2W) turbo-spin-echo, apparent diffusion coefficient maps, high *b*-value diffusion weighted image (DWI) (b2000 sec/mm^2) and dynamic contrast enhanced (DCE) MRI sequences. Axial DCE images were obtained before, during, and after a single dose of gadopentetate dimeglumine (Magnevist; Berlex, Wayne, NJ) or gadoterate meglumine (Dotarem, Guerbet, Bloomington, IN) administered at a dose of 0.1 mmol/kg of body weight through a peripheral vein at a rate of 3 mL/sec by using a mechanical injector (Spectris MR Injection System; Medrad, Pittsburgh, PA). MR imaging pulse sequence parameters are defined in Table 1.

Reference Standard

All lesions prospectively detected on mpMRI by an in-house prostate radiologist (10 years of experience with approximately 6,500 examinations reported) using PI-RADSv2 guidelines were targeted for MRI/TRUS fusion-guided biopsy. Biopsies were performed using an office-based fusion platform (UroNav; InVivo Corp, Gainesville, FL) by either a urologist or interventional radiologist (both of whom had performed >1500 MRI/TRUS fusion guided biopsies at the time of the study) (16). All mpMRI-defined target lesions were sampled in the axial and sagittal plane, which resulted in two cores per target (17). Standard clinical evaluation of biopsy specimens was done by an experienced pathologist (>25 years of experience in interpretation of genitourinary histopathology). The pathological outcome (Gleason score) for each lesion was recorded from clinical pathology reports.

Repeat PI-RADSv2 Scoring Procedure

Four readers from different institutions with substantial levels of experience reading mpMRI of the prostate were included in this study (reader 1, H.S., with 6 years of experience [approximately 1,500 examinations]; reader 2, Y.M.L. with 5 years of experience [approximately 1,500 examinations]; reader 3, T.B., with 7 years of experience [approximately 2,500 examinations]; reader 4, L.K.B. with 10 years of experience [approximately 8,500 examinations]). Readers 1 and 2 were considered moderate experienced readers, readers 3 and 4 were considered expert readers. Readers were provided de-identified mpMR imaging (T2W [axial, sagittal, coronal acquisitions], DWI [ADC, b2000], and DCE), screenshots of lesion locations on T2W imaging (Figure 1) for all lesions sampled by MRI/TRUS guided biopsy sessions, and general lesion location (level and zone). For each target lesion, readers followed PI-RADSv2 guidelines for scoring each pulse sequence (e.g. T2W, DWI and DCE), longest axial dimension, and an overall PI-RADSv2 score (3). Readers were also asked to report the pulse sequence they used to make the measurement. Cases were sent to readers at outside and separate institutions to be viewed using RadiAnt DICOM Viewer (Meixant, Poznan, Poland). Following the first round of review, the order of paired mpMRI imaging and targets were re-randomized and distributed to readers following a four-week washout period. A research fellow independent of all readers coordinated all preparation/randomization, distribution and collection of DICOM and scoring data.

Statistical Analysis

Scoring agreement was examined both within (intra-reader) and across (inter-reader) readers for all reported PI-RADSv2 scoring (Overall, T2W, DWI, and DCE) and lesion size. Scoring agreement was calculated using two methods, kappa statistic and specific agreement. Specific agreement describes the overall proportion of observed agreement (P_o) across all scoring categories, defined as the sum of category-wise agreements relative to the total number of category-wise opportunities for agreement (18). Cohen's kappa was applied in the case of pairwise reader comparison (intra-reader or paired inter-reader stratified by experience) and Fleiss' kappa was applied in the case of four-reader comparison (overall inter-reader agreement).

Lesion size agreement was assessed using Intraclass Correlation Coefficient (ICC), estimated from a mixed effects model of log-transformed size measurements consisting of nested random effects for patient, lesions, and readers and fixed effect for reader experience. Significance of fixed effects was determined by the likelihood ratio test. As the difference in un-transformed lesion size measurements between round 1 and round 2 for all readers were more stable than the log-transformed counterpart, limits of agreement (LOA) for lesion size measurements were estimated from a mixed effects model of the difference in un-transformed lesion size measurements. Coverage probability (CP) was calculated as the percent of lesions contained within varying limits of size measurement agreement, assessed for each reader as the absolute difference in lesion size between round 1 and round 2 measurements (19).

Standard errors and 95% confidence intervals (CI) were estimated from 2000 bootstrap samples by random sampling on the patient-level to account for intra- and inter-lesion correlations. The Wald test was used to test the difference in inter-reader agreement for moderate vs. expert readers. All p-values correspond to two-sided tests, with a p-value <0.05 considered to represent a significant difference.

RESULTS

Patient and Lesion Characteristics

Table 2 shows baseline characteristics of our study population. Of note, 121/205 lesions were benign and 55/205 were clinically significant cancers (GS ≥ 3+4 = 7). The majority of lesions were located in the PZ (N=155), with fewer lesions in the TZ (N=49) and one lesion in the central zone (CZ).

Lesion Scoring Distribution

Considering the distribution of PI-RADSv2 scores in both rounds, category 5 showed highest frequency of agreement and category 3 showed lowest prevalence of agreement (Figure 3). For lesions scored category 5 by any reader, there was consensus by majority (3+) readers in 34% (12/35) in round 1 and 39% (12/31) in round 2, with 10/12 lesions validated to be cancer-positive in each round. A similar pattern of majority consensus was observed in category 4 lesions read by any reader, 32% (47/145) in round 1 and 34% (46/137) in round 2, with 30/47 and 24/46 validated to be cancer-positive, respectively In category 2 lesions, a consensus by majority was reached in 23% (34/148) and 30% (44/148) for the first and second round, with 7/34 and 13/44 cancer-positive, respectively. Agreement by majority of readers was lowest in PI-RADSv2 category 3 at 9% (9/102) and 8% (8/102) for rounds 1 and 2, respectively.

Intra-Reader Reproducibility

Intra-reader kappa (κ) and specific agreement (P_o) reproducibility estimates for PI-RADSv2 overall score assignment are listed for all readers and all zones in Table 3. Overall, intra-reader reproducibility by kappa was moderate to high in all readers (κ=0.43–0.67), with average specific agreement 70% across all readers (P_o range 0.60–0.74) reflecting the overall proportion of lesions with concordant scoring. Expert readers demonstrated higher intra-reader agreement (κ=0.62–0.66; P_o=0.76–0.79) compared to moderate readers (κ=0.39–0.49; P_o=0.60–0.69) in PZ lesions. This trend was not observed in TZ lesions (Table 3).

Intra-reader reproducibility of PI-RADSv2 pulse sequence scoring (T2W, DWI, DCE) in all lesions is shown in Figure 4. Comparing to PI-RADSv2 overall scores, intra-reader reproducibility was lower in sequence specific scoring for a majority of readers. Expert readers demonstrated higher intra-reader reproducibility (κ=0.53–0.65; P_o=0.67–0.73) compared to moderate readers (κ=0.33–0.44; P_o=0.56–62) in DWI (Supplementary Table 1). No differences in intra-reader reproducibility across reader experience were observed for T2W and DCE scoring. Zone-specific assessment of intra-reader reproducibility for all sequences is reported in Supplementary Table 1.

Inter-Reader Reproducibility

Inter-reader agreement for all readers was poor (κ=0.24) with specific agreement occurring in only 46% as summarized in Table 4. Pairwise agreement between moderate and expert readers demonstrated superior scoring agreement between expert readers, with better agreement for PZ lesions (κ=0.23 and 0.43 in moderates and experts, respectively, p=0.002) and specific agreement (P_o=0.51 and 0.62 in moderates and experts, respectively, p=0.009). No significant differences in pairwise inter-reader agreement stratified by reader experience was observed in TZ lesions (Table 4).

Figure 4 shows inter-reader agreement by pulse sequence (T2W, DWI, and DCE) for all readers with pairwise stratification by reader experience. Sequence specific inter-reader agreement for all readers was similar to the results for overall score (κ=0.24, 0.24, and 0.23 and P_o=0.47, 0.44, and 0.54 in T2W, DWI, and DCE, respectively; Supplementary Table 2). Expert readers demonstrated higher agreement in DWI scoring using both similarity metrics (Figure 5). In T2W, the observed proportion of agreement was 45% in moderate readers vs 58% in expert readers (p=0.0002), while kappa showed a trend favoring expert readers (κ=0.25 vs 0.35, p=0.077). No differences in DCE scoring were observed among all readers or by pairwise agreement stratified by reader experience. Zone-specific sequence-based inter-reader reproducibility for all readers, expert readers, and moderate readers is reported in Supplementary Table 2.

Lesion Measurement Agreement

Overall intra-reader ICC was 0.82, with similar performance across both prostatic zones (Table 5). Agreement in lesion size for all readers was more moderate, with overall inter-reader ICC 0.71 influenced by a relatively poorer agreement in TZ (ICC=0.58). Reader experience was found to significantly contribute to lesion size variability in the mixed effect model (X^2=70.5, p<0.0001). Variance estimates derived from mixed effects model demonstrate lesion size differences up to +/−4.8 mm on two measurements are within 95% statistical limits (LOA) of expected variability across all readers for all lesions. The proportion of lesions falling (CP) within1 mm incremental limits of lesion size difference is shown in Figure 5. Despite the average of all readers achieving 95% CP at +/−5 mm limit, reader-specific 95% CP is achieved at variable limits (range 3–6 mm). Reader 2 (expert) showed the highest reproducibility in lesion size measurements, nearing 95% CP at 2 mm.

DISCUSSION

Our study suggests that PI-RADSv2 has overall moderate to good intra-reader reproducibility and poor to moderate inter-reader reproducibility and lesion measurement agreement. Our overall intra-reader reproducibility findings show moderate to good agreement, with expert readers having overall higher intra-reader reproducibility than moderate experience level readers. Reproducibility was lower in T2W and DWI sequences than overall scoring, except for reader 1. Overall, our study shows similar scoring reproducibility compared to BI-RADS literature, which reports moderate to good intra-reader reproducibility ranging from κ=0.53–0.77 (8, 20–22). As indicated in other reader agreement studies, alternative statistical metrics can be utilized to inform raw agreement in addition to the traditional kappa statistic due to the dependence of kappa on the presence and distribution of scoring observations (9, 23, 24). In this study, we report the overall proportion of specific agreement for intra-reader reproducibility in PI-RADSv2 overall scoring to be 60–77% across all readers, with a higher proportion of lesion-based scoring agreement in expert readers compared to moderate readers. The intra-reader reproducibility has implications for within-institution lesion interpretation and reporting. Aside from carrying out in-house investigations of reader reproducibility to specify areas of scoring disagreement, a solution to improve intra and inter-reader reproducibility is implementation of a PI-RADSv2 scoring template. This could decrease the subjectivity involved with interpreting mpMRI signal patterns and the PI-RADSv2 lexicon itself. Assessment for in-house reproducibility at individual Prostate MRI centers for purposes of quality control may further improve diagnostic precision. The moderate intra-reader reproducibility observed in our study has implications on the interpretation of agreeability and performance in multi-reader studies.

Inter-reader reproducibility was poor to moderate in our study, which aligns with the current PI-RADSv2 inter-reader literature. Sonn et al. (15) had 9 single center radiologists prospectively evaluate prostate mpMRIs and found “considerable variability in PIRADS score assignment.” Both Muller et al. (13) and Rosenkrantz et al. (14) found only moderate reproducibility among beginner to moderately experienced readers while expert readers had only slightly better performance. In our study, the greatest consensus among readers was reached for category 4 and 5 lesions, and poor consensus was reached for category 3 lesions. This supports current PI-RADS reader agreement literature, which shows greatest agreement among higher PI-RADSv2 categories (14, 25). We additionally report differences in reader agreement across prostate zones. Experts show highest agreement in the PZ using the DWI sequence. Although still only moderate in reproducibility, it is reassuring that the best agreement occurs for the dominant sequence used in the zone where most cancerous lesions exist (26). The TZ also contains the majority of benign prostatic hyperplasia, which can appear similar to cancer, complicating interpretation of this zone. This could contribute to the considerable number of lesions observed to have equal proportion of PI-RADSv2 category 2 and 4 assignments across all readers in both rounds.

Lesion size agreement was moderate across all readers, leading to overall limits of agreement of +/−4.8mm to capture approximately 95% of measurements. While lesion size is only incorporated within PI-RADSv2 for distinguishing upgrade to Category 5 (>1.5cm), there are several clinical scenarios that utilize lesion size measurements. The poor reproducibility observed in this study, resulting in wide confidence intervals, is important to consider in longitudinal assessment of prostate lesions on mpMRI. Recent work in the use of mpMRI for patients on active surveillance has suggested 2–3mm increases can represent meaningful progression (27–29). Only one of four readers achieved accuracy within this range, with others ranging from 4–6 mm to achieve 95% coverage making accurate diagnosis of progression by changes in lesion size challenging. The results of this study should further be evaluated in an active surveillance population and in the context of true disease progression. Lesion size agreement also has important implications on focal therapy planning. An expert consensus suggested using 5 mm treatment planning margins for focal ablation of prostate cancer, while Le Nobin et al. (30) showed that a 9 mm treatment margin achieved complete histological tumor ablation in 100% of patients (31). These treatment margins compensate for both disagreement in lesion size measurement reported in our study and possible tumor size underestimation on mpMRI (32).

The imperfect agreement of PI-RADSv2 readings is likely due to many factors. First, there is a high degree of signal pattern complexity within the prostate on mpMRI. Despite working with PI-RADSv2 since 2015, readers of all experience levels continue to have difficulty agreeing with each other on category assignment, especially, category 3 (13). Readers have been shown to have poor agreement when attempting to localize identified lesions on a PI-RADSv2 sector map (33). In addition, readers have also been shown to have difficulty in applying the PI-RADSv2 lexicon, as expert readers only show moderate agreement when using it (14). It is also possible that the current PI-RADSv2 lexicon does not yet encompass the diverse manifestations of prostate cancer on mpMRI. Finally, readers may have distinctive styles in their application of PI-RADS. For example, one reader may attempt to strictly utilize the lexicon to score lesions while another reader may score lesions based on how the score will affect clinical outcomes. We believe that these factors, and possibly others, could be influencing reader decisions, reducing agreement by creating variability in intra-reader assessment which influences inter-reader agreement and clinical mpMRI utilization.

Our study has several limitations. This study was designed to allow for robust characterization of inter- and intra-reader agreement in PI-RADSv2 interpretation and scoring in a selection of pre-defined lesions. Due to the retrospective nature and use of biopsy validation, readers were only permitted to score lesions presented to them. Of the four readers two were designated as expert experienced and two as moderately experienced based on overall experience reading prostate MRI (duration and number of studies), though the moderately experienced readers in this case still had many years experience and thus these results cannot be extrapolated to accurately reflect inter- or intra-reader agreement of novice or inexperienced radiologists. Within this biopsy-validated cohort, most lesions were benign (121/205). Performance metrics of cancer detection at each score level were not evaluated due to the inherent bias from pre-selected lesions by an independent radiologist. The majority benign lesion population may be an effect of the protocol our patients are enrolled in, where all reported lesions undergo MRI/TRUS fusion-guided biopsy. However, we believe that our population reflects real life referrals considering more wide use of prostate mpMRI for biopsy guidance rather than staging in surgery populations. A major strength of this study is a relatively large cohort of biopsy-naïve, consecutive patient who were evaluated by multiple experts at varying institutions experienced in interpreting prostate MRI.

In conclusion, our study found PI-RADSv2 to have moderate to substantial intra-reader reproducibility but poor to moderate inter-reader reproducibility among moderate to expert level readers. Readers disagree most strongly on PI-RADS category 3 lesions and least on PI-RADS category 4 and 5 lesions. Lesion measurement agreement is moderate at best and requires further investigation. Thus, a major issue in the use of PI-RADS v2 is its high inter-reader variability and inconsistency in measurement, and research should be directed at reducing these differences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding:

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This research was also made possible through the NIH Medical Research Scholars Program, a public-private partnership supported jointly by the NIH and generous contributions to the Foundation for the NIH from the Doris Duke Charitable Foundation, the American Association for Dental Research, the Colgate-Palmolive Company, Genentech, Elsevier, and other private donors.

REFERENCES

1. Murphy G, Haider M, Ghai S, Sreeharsha B. The expanding role of MRI in prostate cancer. AJR Am J Roentgenol 2013;201(6):1229–38. [PubMed: 24261361]
2. Röthke M, Blondin D, Schlemmer HP, Franiel T. [PI-RADS classification: structured reporting for MRI of the prostate]. Rofo 2013;185(3):253–61. [PubMed: 23404430]
3. 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. European urology 2016;69(1):16–40. [PubMed: 26427566]
4. Carroll PR, Parsons JK, Andriole G, Bahnson RR, Castle EP, Catalona WJ, et al. NCCN Guidelines Insights: Prostate Cancer Early Detection, Version 2.2016. J Natl Compr Canc Netw 2016;14(5): 509–19. [PubMed: 27160230]
5. Kasivisvanathan V, Rannikko AS, Borghi M, Panebianco V, Mynderse LA, Vaarala MH, et al. MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. N Engl J Med 2018.
6. Mehralivand S, Shih JH, Rais-Bahrami S, Oto A, Bednarova S, Nix JW, et al. A Magnetic Resonance Imaging-Based Prediction Model for Prostate Biopsy Risk Stratification. JAMA Oncol 2018.
7. Barentsz JO, Richenberg J, Clements R, Choyke P, Verma S, Villeirs G, et al. ESUR prostate MR guidelines 2012. Eur Radiol 2012;22(4):746–57. [PubMed: 22322308]
8. D'Orsi CJ. ACR BI-RADS atlas: breast imaging reporting and data system: American College of Radiology; 2013.

9. Rosenkrantz AB, Lim RP, Haghighi M, Somberg MB, Babb JS, Taneja SS. 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(4):W612–8. [PubMed: 24059400]
10. 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(11):3185–90. [PubMed: 23756958]
11. Quentin M, Arsov C, Röhlen S, Klasen J, Antoch G, Albers P, et al. Inter-reader agreement of multi-parametric MR imaging for the detection of prostate cancer: evaluation of a scoring system. Rofo 2012;184(10):925–9. [PubMed: 22744328]
12. 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(4):477–94. [PubMed: 21195536]
13. 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(3):741–50. [PubMed: 26098458]
14. 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(3):793–804. [PubMed: 27035179]
15. Sonn GA, Fan RE, Ghanouni P, Wang NN, Brooks JD, Loening AM, et al. Prostate Magnetic Resonance Imaging Interpretation Varies Substantially Across Radiologists. Eur Urol Focus 2017.
16. Pinto PA, Chung PH, Rastinehad AR, Baccala AA, Kruecker J, Benjamin CJ, et al. Magnetic resonance imaging/ultrasound fusion guided prostate biopsy improves cancer detection following transrectal ultrasound biopsy and correlates with multiparametric magnetic resonance imaging. J Urol 2011;186(4):1281–5. [PubMed: 21849184]
17. Hong CW, Rais-Bahrami S, Walton-Diaz A, Shakir N, Su D, George AK, et al. Comparison of magnetic resonance imaging and ultrasound (MRI-US) fusion-guided prostate biopsies obtained from axial and sagittal approaches. BJU Int 2015;115(5):772–9. [PubMed: 25045781]
18. Uebersax JS. A design-independent method for measuring the reliability of psychiatric diagnosis. Journal of Psychiatric Research 1982;17(4):335–42. [PubMed: 7187777]
19. Barnhart HX, Yow E, Crowley AL, Daubert MA, Rabineau D, Bigelow R, et al. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. Stat Methods Med Res 2016;25(6):2939–58. [PubMed: 24831133]
20. Redondo A, Comas M, Macià F, Ferrer F, Murta-Nascimento C, Maristany MT, et al. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. Br J Radiol 2012;85(1019):1465–70. [PubMed: 22993385]
21. Irshad A, Leddy R, Ackerman S, Cluver A, Pavic D, Abid A, et al. Effects of Changes in BI-RADS Density Assessment Guidelines (Fourth Versus Fifth Edition) on Breast Density Assessment: Intra- and Interreader Agreements and Density Distribution. AJR Am J Roentgenol 2016;207(6): 1366–71. [PubMed: 27656766]
22. Ekpo EU, Ujong UP, Mello-Thoms C, McEntee MF. Assessment of Interradiologist Agreement Regarding Mammographic Breast Density Classification Using the Fifth Edition of the BI-RADS Atlas. AJR Am J Roentgenol 2016;206(5):1119–23. [PubMed: 26999655]
23. Greer MD, Brown AM, Shih JH, Summers RM, Marko J, Law YM, et al. Accuracy and agreement of PIRADSv2 for prostate cancer mpMRI: A multireader study. J Magn Reson Imaging 2017;45(2):579–85. [PubMed: 27391860]
24. Rosenkrantz AB, Oto A, Turkbey B, Westphalen AC. Prostate Imaging Reporting and Data System (PI-RADS), Version 2: A Critical Look. American Journal of Roentgenology 2016;206(6):1179–83. [PubMed: 26913638]
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(1):108–16. [PubMed: 26836049]

26. McNeal JE, Redwine EA, Freiha FS, Stamey TA. Zonal distribution of prostatic adenocarcinoma. Correlation with histologic pattern and direction of spread. Am J Surg Pathol 1988;12(12):897–906. [PubMed: 3202246]

27. Bryk DJ, Llukani E, Huang WC, Lepor H. Natural History of Pathologically Benign Cancer Suspicious Regions on Multiparametric Magnetic Resonance Imaging Following Targeted Biopsy. J Urol 2015;194(5):1234–40. [PubMed: 26003206]

28. Rosenkrantz AB, Rice SL, Wehrli NE, Deng FM, Taneja SS. Association between changes in suspicious prostate lesions on serial MRI examinations and follow-up biopsy results. Clin Imaging 2015;39(2):264–9. [PubMed: 25457528]

29. Felker ER, Wu J, Natarajan S, Margolis DJ, Raman SS, Huang J, et al. Serial Magnetic Resonance Imaging in Active Surveillance of Prostate Cancer: Incremental Value. J Urol 2016;195(5):1421–7. [PubMed: 26674305]

30. Le Nobin J, Rosenkrantz AB, Villers A, Orczyk C, Deng FM, Melamed J, et al. Image Guided Focal Therapy for Magnetic Resonance Imaging Visible Prostate Cancer: Defining a 3-Dimensional Treatment Margin Based on Magnetic Resonance Imaging Histology Co-Registration Analysis. J Urol 2015;194(2):364–70. [PubMed: 25711199]

31. Donaldson IA, Alonzi R, Barratt D, Barret E, Berge V, Bott S, et al. Focal therapy: patients, interventions, and outcomes--a report from a consensus meeting. Eur Urol 2015;67(4):771–7. [PubMed: 25281389]

32. Borofsky S, George AK, Gaur S, Bernardo M, Greer MD, Mertan FV, et al. What Are We Missing? False-Negative Cancers at Multiparametric MR Imaging of the Prostate. Radiology 2018;286(1):186–95. [PubMed: 29053402]

33. Greer MD, Shih JH, Barrett T, Bednarova S, Kabakus I, Law YM, et al. All over the map: An interobserver agreement study of tumor location based on the PI-RADSv2 sector map. J Magn Reson Imaging 2018.

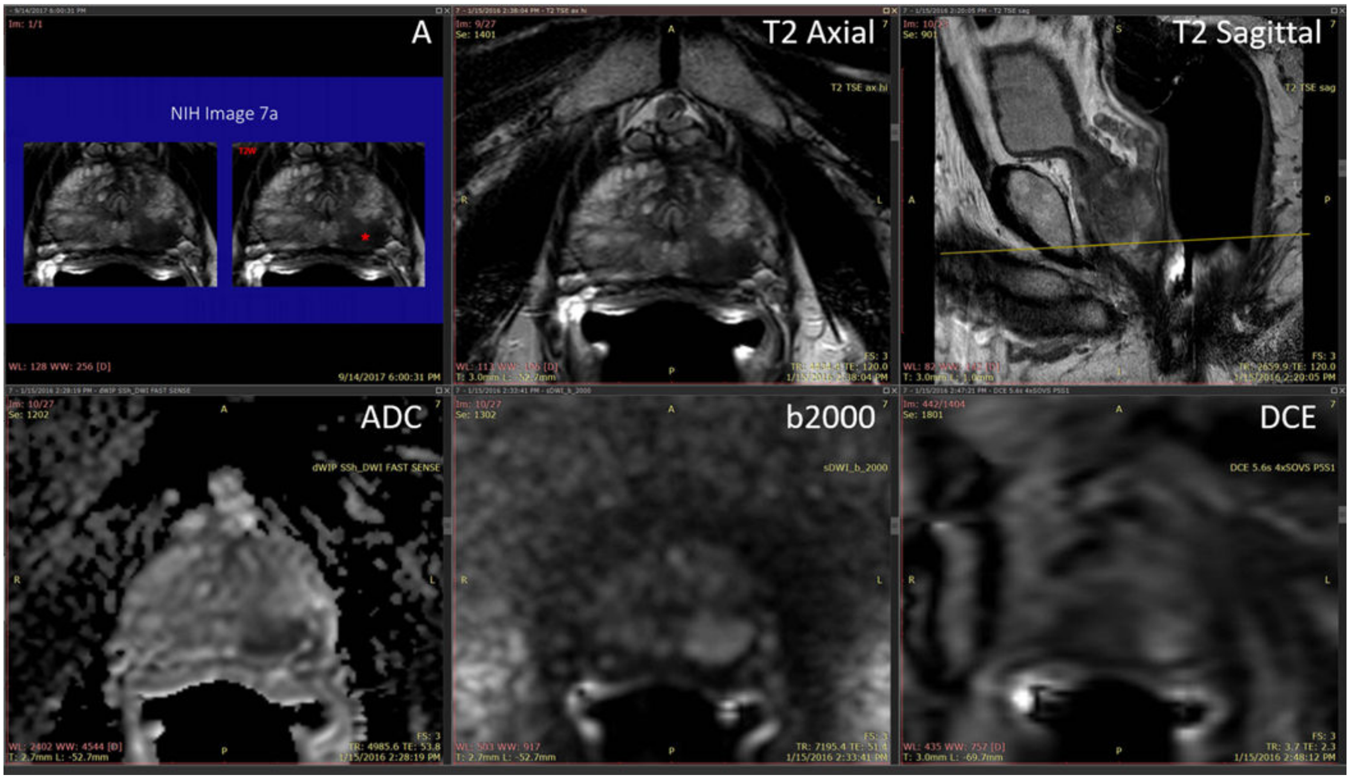

Figure 1. Example of lesion identification presented to all readers.

Shown is a demarcated left apical peripheral zone lesion on T2W axial pulse sequence (A) with associated T2W axial and sagittal imaging, ADC map, calculated b2000 DWI, and DCE MR images. Readers were provided the full mpMRI image sets for evaluation and scoring was only completed on provided, pre-identified lesions.

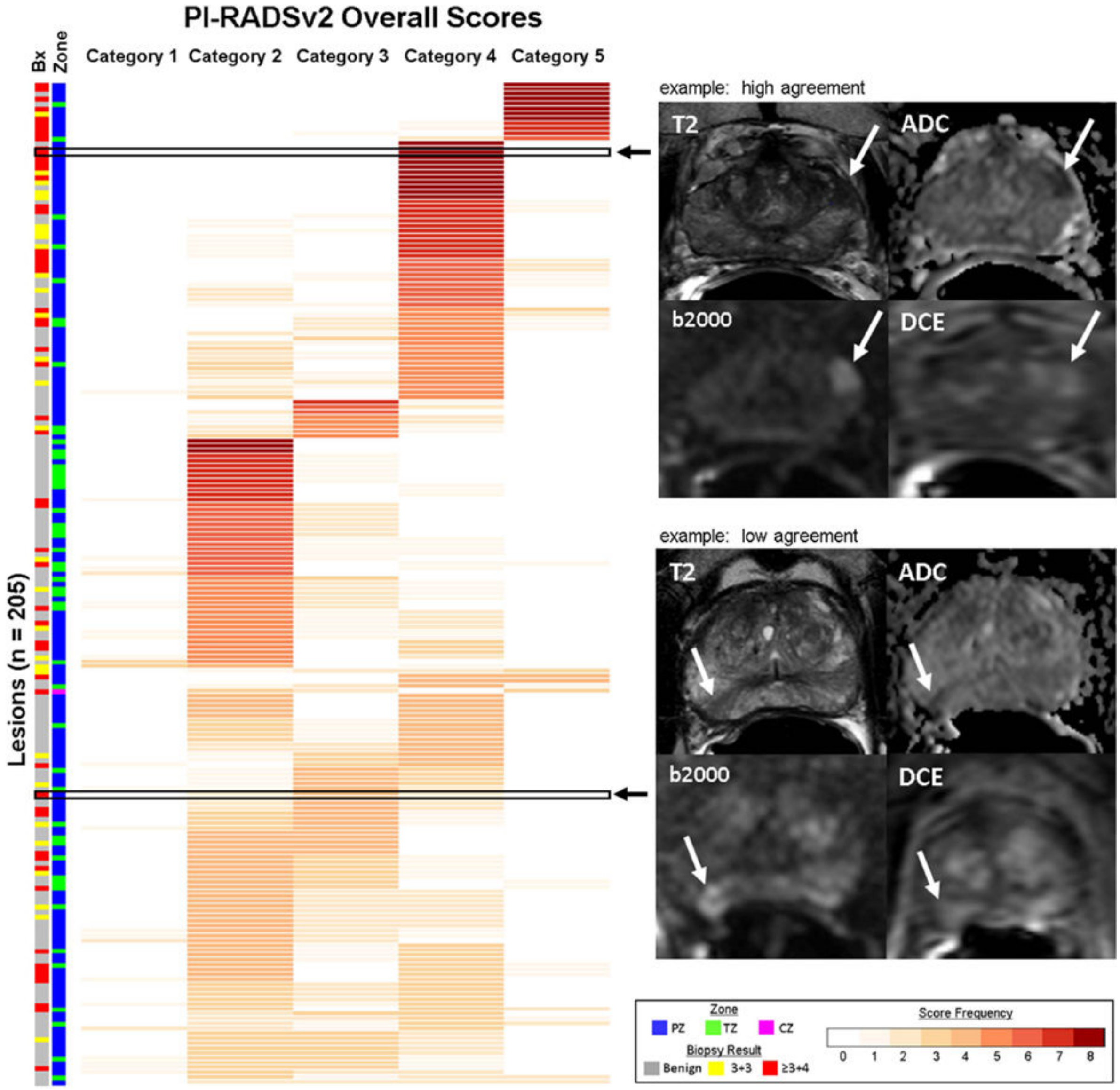

Figure 2. Overall PI-RADSv2 scoring distribution chart.

Each lesion (n=205) represented as a row with the frequency of scores reported for PI-RADSv2 categories. Lesions with consistent scoring, indicating high agreement, are represented by darker shades of red (max count = 4 readers x 2 rounds = 8). Lesions with poor agreement in scoring demonstrate lower frequencies across multiple PI-RADSv2 scoring categories. Two representative examples demonstrating high and low scoring agreement are shown with T2-weighted (T2), ADC, high *b*-value DWI (b2000) and DCE pulse sequences. **Top – high agreement:** A left base anterior peripheral zone lesion (Gleason Score 3+4 = 7) with reproducible scoring, in which all 4 readers assigned PI-RADSv2 Category 4 in both rounds. **Bottom – low agreement:** A right apical-mid peripheral zone lesion (Gleason Score 3+4 = 7) showing poor consensus, with two PI-RADSv2 Category 2 scores, four PI-RADSv2 Category 3 scores, and two PI-RADSv2 Category 4 scores. Two readers reported the same score in both rounds (Reader 1 PI-

RADSv2 score 3, Reader 4 PI-RADSv2 score 2), with other readers demonstrating inconsistent scoring between round 1 and round 2. Prostate zonal location of each lesion and targeted biopsy results are indicated along the left side of the figure. PZ = peripheral zone, TZ = transition zone, and CZ = central zone.

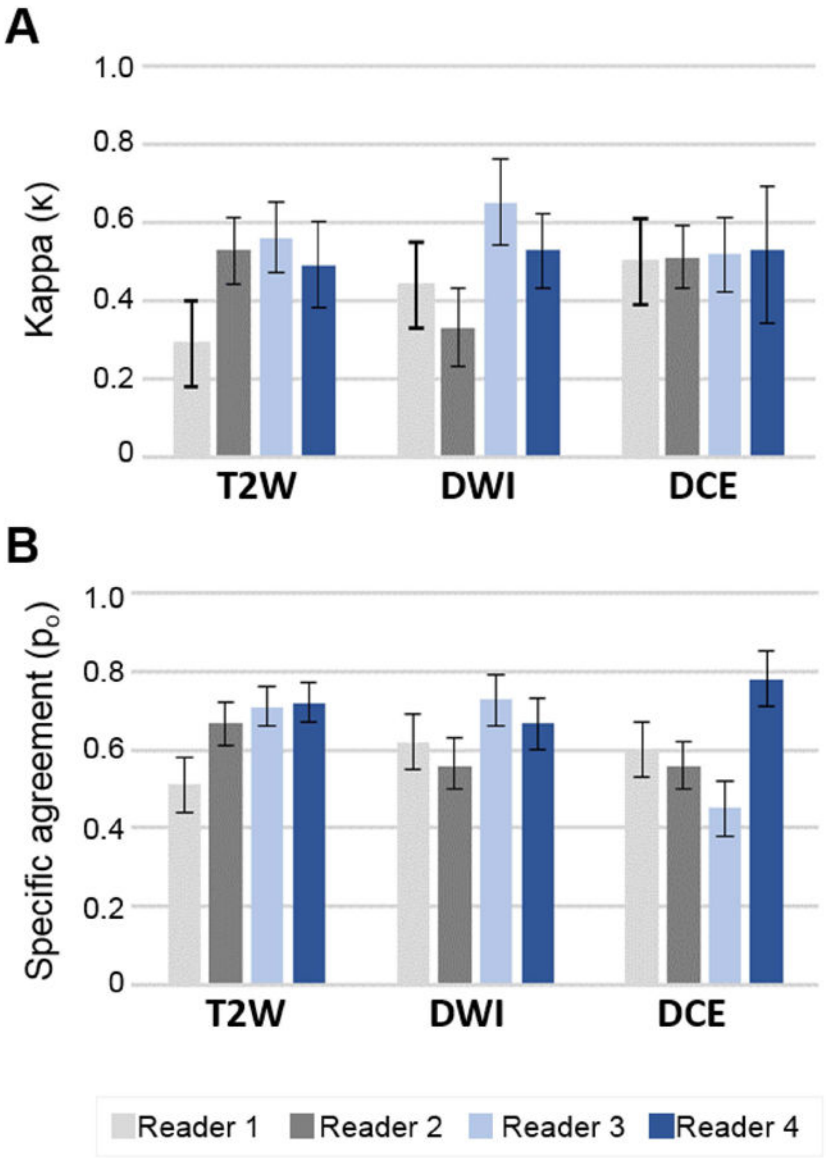

Figure 3.

Intra-reader repeatability using Kappa (A) and specific agreement (B) among 4 readers and 2 rounds of lesions scoring. Readers 1 and 2 were moderate experience level, and readers 3 and 4 were expert experience level.

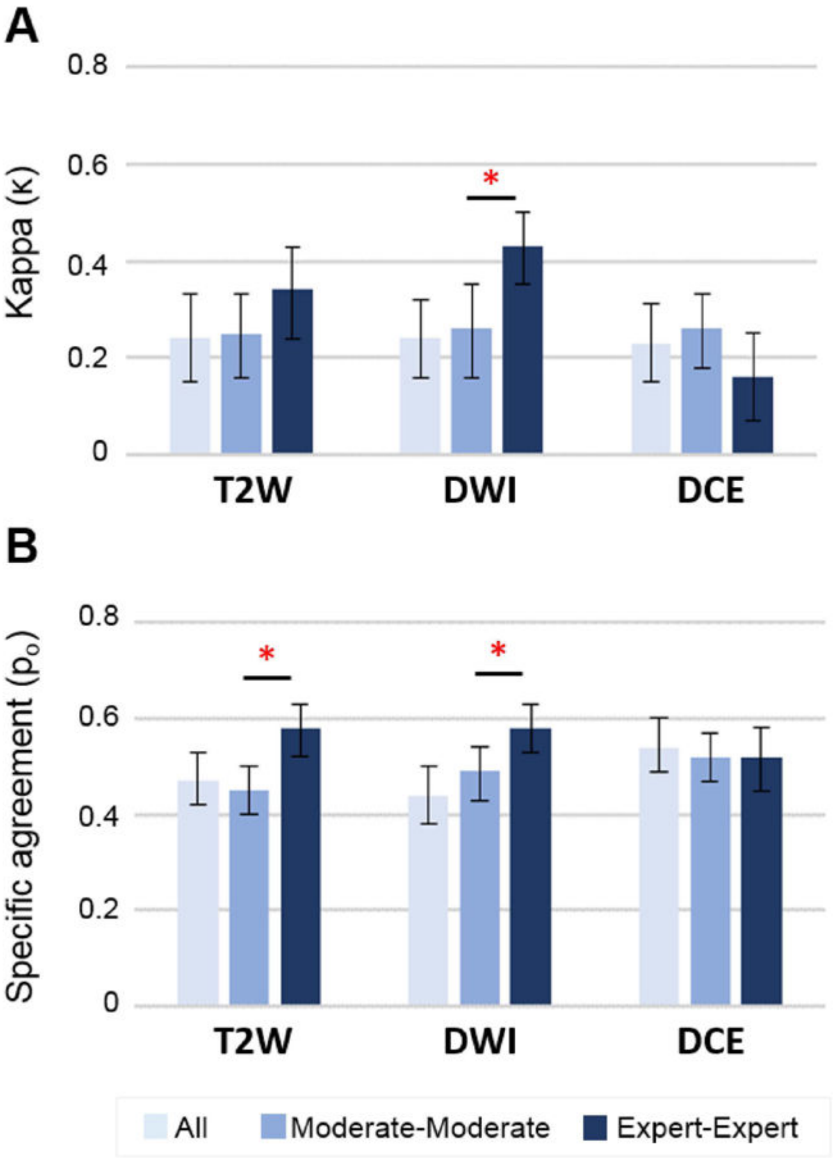

Figure 4.

Inter-reader reproducibility analysis of all, moderate and expert experience level readers using kappa (A) and specific agreement (B) statistics. *= significant value (P<.05) between moderate and expert level readers.

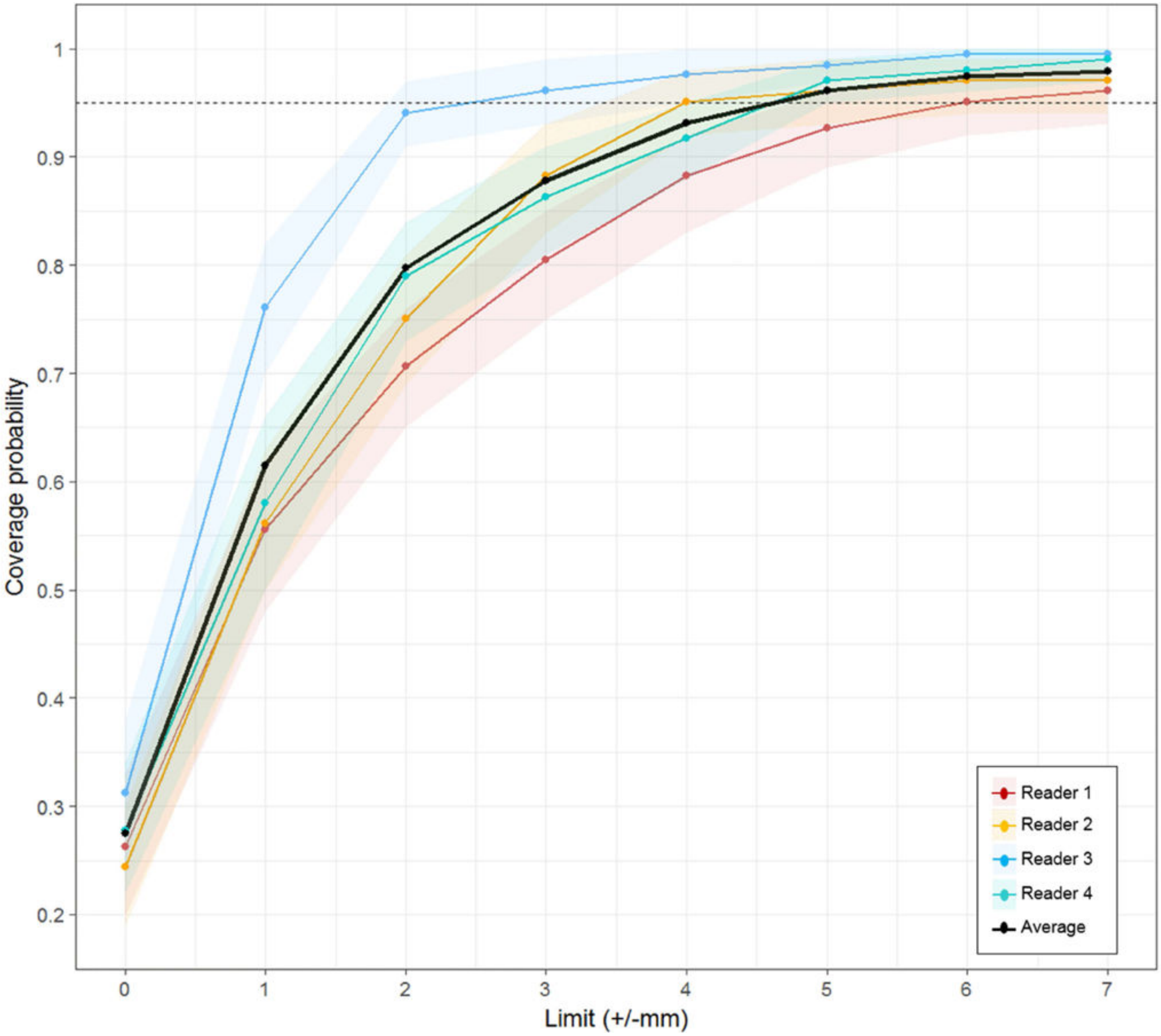

Figure 5.

Lesion coverage probability versus absolute change in lesion size measurement for all four readers in the whole prostate. For the average of four readers, 95% of all lesions had an absolute value ≤4.88 mm lesion measurement change. The shading that flanks each reader's line represents 95% CIs.

Table 1.

Multiparametric MR imaging sequence parameters at 3T.

Parameter	T2 Weighted	DWI[a]	High *b*-Value DWI[b]	DCE MR Imaging[c]
Field of view (mm)	140 × 140	140 × 140	140 × 140	262 × 262
Acquisition matrix	304 × 234	112 × 109	76 × 78	188 × 96
Repetition time (msec)	4434	4986	6987	3.7
Echo time (msec)	120	54	52	2.3
Flip angle (degrees)	90	90	90	8.5
Section thickness (mm), no gaps	3	3	3	3
Image reconstruction matrix (pixels)	512 × 512	256 × 256	256 × 256	256 × 256
Reconstruction voxel imaging resolution (mm/pixel)	0.27 × 0.27 × 3.00	0.55 × 0.55 × 2.73	0.55 × 0.55 × 2.73	1.02 × 1.02 × 3.00
Time for acquisition (min:sec)	2:48	4:54	3:50	5:16

[a] For ADC map calculation. Five evenly-spaced b values (0–750 sec/mm^2) were used.

[b] *b*= 2000 sec/mm^2

[c] DCE images obtained before, during, and after a single dose of gadopentetate dimeglumine 0.1 mmol/kg at 3 mL/sec. Each sequence obtained at 5.6 -sec intervals.

Table 2.

Summary of all demographic data of all study subjects.

Description	Summary	PZ	TZ	CZ
Patients (N)	102			
Age (yr.)	64 (44–84)			
PSA (ng/mL)	5.95 (1.17–113.6)			
Prostate Volume (mL)	51.5 (19–100)			
Lesions (n)	205	155	49	1
Benign	121	91	30	0
GS 3+3	29	23	6	0
GS 3+4	29	24	5	0
GS 4+3	10	7	3	0
GS 4+4	11	7	3	1
GS> 4+4	5	3	2	0

Note.- Median values with ranges reported in parentheses. PSA = prostate-specificantigen, PZ = peripheral zone, TZ = transition zone, GS =Gleason score.

Table 3.

PI-RADSv2 overall score intra-reader reproducibility by prostate zone

Prostate Zone	Kappa Value [κ]	Specific Agreement [P_o]
Whole Prostate		
Reader 1[M]	0.43 (0.32–0.54)	0.60 (0.53–0.67)
Reader 2[M]	0.54 (0.42–0.65)	0.70 (0.63–0.77)
Reader 3[E]	0.67 (0.59–0.75)	0.77 (0.71–0.83)
Reader 4[E]	0.55 (0.45–0.63)	0.74 (0.69–0.79)
Peripheral Zone		
Reader 1[M]	0.39 (0.27–0.52)	0.60 (0.53–0.67)
Reader 2[M]	0.49 (0.35–0.62)	0.69 (0.61–0.77)
Reader 3[E]	0.66 (0.57–0.75)	0.76 (0.70–0.82)
Reader 4[E]	0.62 (0.52–0.71)	0.79 (0.73–0.84)
Transition Zone		
Reader 1[M]	0.47 (0.27–0.65)	0.61 (0.47–0.75)
Reader 2[M]	0.57 (0.37–0.74)	0.71 (0.58–0.83)
Reader 3[E]	0.69 (0.49–0.86)	0.82 (0.69–0.92)
Reader 4[E]	0.32 (0.10–0.53)	0.59 (0.45–0.73)

Note.- Numbers in parentheses are 95% CIs.

[M] Moderate Experience Reader

[E] Expert Experience Reader

Table 4.

PI-RADSv2 overall score inter-reader reproducibility by zone

Prostate Zone	Kappa Value (κ)	P Value	Specific Agreement (P_o)	P Value
Whole Prostate				
All Readers	0.24 (0.18–0.30)		0.46 (0.43–0.50)	
Moderate Readers	0.27 (0.19–0.36)	0.026	0.51 (0.45–0.56)	0.007
Expert Readers	0.40 (0.32–0.48)		0.61 (0.55–0.66)	
Peripheral Zone				
All Readers	0.23 (0.16–0.30)		0.46 (0.42–0.51)	
Moderate Readers	0.23 (0.12–0.34)	0.002	0.51 (0.45–0.58)	0.009
Expert Readers	0.43 (0.34–0.53)		0.62 (0.56–0.68)	
Transition Zone				
All Readers	0.19 (0.10–0.28)		0.47 (0.40–0.53)	
Moderate Readers	0.27 (0.14–0.40)	0.929	0.49 (0.39–0.59)	0.347
Expert Readers	0.26 (0.06–0.43)		0.56 (0.47–0.66)	

Note.- P values show the statistical difference between moderate and expert readers.

Table 5.

Lesion measurement agreement

Prostate Zone	ICC
Intra-Reader	
WP	0.82 (0.78 – 0.86)
PZ	0.83 (0.77 – 0.87)
TZ	0.80 (0.72 – 0.86)
Inter-Reader	
WP	0.71 (0.66 – 0.76)
PZ	0.74 (0.67 – 0.80)
TZ	0.58 (0.45 – 0.68)

Note.- Numbers in parentheses represent 95%CIs.

ICC = Interclass correlation, WP = whole prostate,

PZ = peripheral zone, TZ = transition zone.