Interreader variability in prostate MRI reporting using Prostate Imaging Reporting and Data System version 2.1

Brembilla, G.; Dell'Oglio, P.; Stabile, A.; Damascelli, A.; Brunetti, L.; Ravelli, S.; ... ; Cobelli, F. de

European Radiology (2020) 30:3383–3392
https://doi.org/10.1007/s00330-019-06654-2

UROGENITAL

Interreader variability in prostate MRI reporting using Prostate Imaging Reporting and Data System version 2.1

Giorgio Brembilla[1] · Paolo Dell'Oglio[2] · Armando Stabile[2] · Anna Damascelli[1] · Lisa Brunetti[1] · Silvia Ravelli[1] · Giulia Cristel[1] · Elena Schiani[1] · Elena Venturini[1] · Daniele Grippaldi[1] · Vincenzo Mendola[3] · Paola Maria Vittoria Rancoita[4] · Antonio Esposito[1] · Alberto Briganti[2] · Francesco Montorsi[2] · Alessandro Del Maschio[1] · Francesco De Cobelli[1]

Received: 11 September 2019 / Revised: 16 November 2019 / Accepted: 19 December 2019 / Published online: 12 February 2020

Abstract

Objectives To evaluate the agreement among readers with different expertise in detecting suspicious lesions at prostate multiparametric MRI using Prostate Imaging Reporting and Data System (PI-RADS) version 2.1.

Methods We evaluated 200 consecutive biopsy-naïve or previously negative biopsy men who underwent MRI for clinically suspected prostate cancer (PCa) between May and September 2017. Of them, 132 patients underwent prostate biopsy. Seven radiologists (four dedicated uro-radiologists and three non-dedicated abdominal radiologists) reviewed and scored all MRI examinations according to PI-RADS v2.1. Agreement on index lesion detection was evaluated with Conger's *k* coefficient, agreement coefficient 1 (AC1), percentage of agreement (PA), and indexes of specific positive and negative agreement. Clinical and radiological features that may influence variability were evaluated.

Results Agreement in index lesion detection among all readers was substantial (AC1 0.738; 95% CI 0.695–0.782); dedicated radiologists showed higher agreement compared with non-dedicated readers. Clinical and radiological parameters that positively influenced agreement were PSA density ≥ 0.15 ng/mL/cc, pre-MRI high risk for PCa, positivity threshold of PI-RADS score 4 + 5, PZ lesions, homogeneous signal intensity of the PZ, and subjectively easy interpretation of MRI. Positive specific agreement was significantly higher among dedicated readers, up to 93.4% (95% CI 90.7–95.4) in patients harboring csPCa. Agreement on absence of lesions was excellent for both dedicated and non-dedicated readers (respectively 85.1% [95% CI 78.4–92.3] and 82.0% [95% CI 77.2–90.1]).

Conclusions Agreement on index lesion detection among radiologists of various experiences is substantial to excellent using PI-RADS v2.1. Concordance on absence of lesions is excellent across readers' experience.

Key Points

- *Agreement on index lesion detection among radiologists of various experiences is substantial to excellent using PI-RADS v2.1.*
- *Concordance between experienced readers is higher than between less-experienced readers.*
- *Concordance on absence of lesions is excellent across readers' experience.*

Keywords Magnetic resonance imaging · Prostate cancer · Inter-observer variability

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-019-06654-2) contains supplementary material, which is available to authorized users.

✉ Giorgio Brembilla
brembilla.giorgio@hsr.it

[1] Department of Radiology, Centre for Experimental Imaging, IRCCS San Raffaele Scientific Institute, Milan, Italy

[2] Department of Urology and Division of Experimental Oncology, URI, Urological Research Institute, IRCCS San Raffaele Scientific Institute, Milan, Italy

[3] Vita-Salute San Raffaele University, Milan, Italy

[4] University Centre for Statistics in the Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University, Milan, Italy

Abbreviations

csPCa	Clinically significant prostate cancer
mpMRI	Multiparametric magnetic resonance imaging
PA	Percentage of agreement
PCa	Prostate cancer
P_{neg}	Proportion of negative agreement
P_{pos}	Proportion of positive agreement
PSAD	PSA density
PZ	Peripheral zone
SI	Signal intensity
TRUS-Bx	Transrectal ultrasound guided biopsy
TZ	Transitional zone

Introduction

In men with clinically suspected prostate cancer (PCa), multiparametric magnetic resonance imaging (mpMRI) reduces the number of unnecessary biopsies and improves the detection of clinically significant PCa (csPCa), while decreasing the detection of clinically insignificant PCa (non-csPCa) [1, 2]. Recently, international urological guidelines recommended to perform MRI before prostate biopsy both in biopsy-naïve patients and in patients with prior negative biopsy [3, 4]. Given the widespread adoption of prostate MRI in clinical practice, Prostate Imaging Reporting and Data System (PI-RADS) has been introduced to standardize interpretation and reporting of MRI examinations [5, 6]. Specifically, PI-RADS v2.1 has addressed limitations in interreader agreement of the previous versions [7]. Prior reports revealed only moderate reproducibility of PI-RADS v2 [8], with poor to moderate agreement in lesion detection [9–14]. However, these studies had several limitations in methodology (e.g., interpretation of screen captures) and in patient selection (only biopsy or radical prostatectomy cohorts) that may prevent the generalizability of their findings to a real-life clinical setting. Moreover, the statistical approach commonly used (*k* coefficient) is known to be exposed to severe paradoxes in determined circumstances [15, 16], potentially underestimating the true extent of the agreement [17].

To overcome these issues, we evaluated interreader agreement of prostate mpMRI using PI-RADS v2.1 in a cohort of patients referred for prostate MRI at our institution for clinically suspected PCa, reproducing the typical clinical workflow. In this setting, we investigated the reproducibility of multiple readers with different expertise in lesion detection, and we evaluated which clinical and radiological features may influence variability. We hypothesized that, using proper statistical analyses in a non-selected cohort of patients, observed agreement may be higher than previously reported.

Materials and methods

Study population

This retrospective study was approved by our Institutional Review Board and written informed consent was obtained from all patients. Our prospectically acquired local database was used to identify 219 consecutive biopsy-naïve or previously negative biopsy men who underwent prostate mpMRI at our institution (San Raffaele Hospital, Milan) for a clinically suspected PCa between May and September 2017. All MRI examinations were formerly interpreted and scored using PI-RADS v2 by one of six dedicated radiologists at our institute. Pre-MRI clinical information (PSA values, digital rectal examination, familiar history, previous local treatments) were collected for all patients. We subsequently excluded patients who had previous transurethral resection of the prostate ($n = 4$), incomplete mpMRI protocol ($n = 1$), low image quality or severe image artifacts ($n = 3$), and missing clinical information ($n = 11$). The final study cohort consisted of 200 men. Of those, 70.5% ($n = 141$) and 29.5% ($n = 59$) were biopsy-naïve and prior negative biopsies, respectively.

MRI acquisition

MRI images were acquired on a 1.5-T scanner (Achieva and Achieva dStream, Philips Medical Systems) with surface and endorectal coil (Prostate eCoilTM, Medrad®); acquisition protocols were in line with PI-RADS v2 standards [6]. Gastrointestinal peristalsis was suppressed by intramuscular administration of 20 mg of scopolamine-butylbromide (Buscopan, Boehringer) immediately before MR scanning. The imaging protocol consisted of multiplanar turbo spin-echo T2-weighted images; echo-planar DWI with *b* values of 50, 800, and 1600 s/mm^2 (ADC maps were automatically elaborated on a pixel-by-pixel basis using *b* values of 50 and 800 s/mm^2); 3-D fast field-echo dynamic contrast–enhanced (DCE) MRI; and delayed axial turbo spin-echo T1-weighted images with fat suppression. For DCE-MRI, an IV bolus of 0.1 mmol/kg of gadobutrol (Gadovist, Bayer Schering Pharma) at a flow rate of 4 mL/s was injected. For patients who had previously undergone prostatic biopsies, mpMRI scans were performed at least after 4 weeks from biopsies, and pre-contrast T1-weighted images were performed to rule out post-biopsy hemorrhagic artifacts.

MRI interpretation

For interreader agreement analyses, all examinations were retrospectively reviewed, interpreted, and scored according to PI-RADS v2.1 [7] by seven radiologists from a single tertiary care referral center. Four were dedicated radiologists with specific clinical and research interests in prostate MR imaging

and 4 to 8 years of experience in prostate MRI (referred to as "dedicated" readers). Three were abdominal radiologists who underwent a specific training but who were not specifically dedicated to prostate MRI in clinical routine (approximately < 10 prostate MRI examination per month), and had < 2 years of experience in the field (referred to as "non-dedicated" readers).

Study design

To replicate as much as possible the typical prostate MRI interpretation workflow, the 7 radiologists had full access to anonymized MRI examinations on a PACS workstation (Fig. 1). For each patient, readers were provided with all pre-MRI clinical information (age, PSA values, DRE, family history, and pre-MRI biopsy status), while they were blinded to original MRI report and post-MRI information (e.g., biopsy results). After image interpretation, findings were reported on a standardized form (Appendix) that included (1) presence/absence of equivocal or suspicious lesions (PI-RADS score ≥ 3); (2) PI-RADS v2.1 score for each lesion; (3) lesion localization and diameters; (4) a score for peripheral zone (PZ) signal intensity (SI) homogeneity, as proposed by Hötker et al [18] (a scale from 1 to 5 indicating the grade of homogeneity in the PZ, where 1 means markedly inhomogeneous SI and 5 indicates a highly homogeneous SI); (5) a subjective score on interpretative difficulty of the MRI images (scale from 1 to 3, where 1: easy; 2: intermediate; 3: difficult). Readers were asked to provide screenshots of each PI-RADS score ≥ 3 lesion on T2W images, which were used to determine lesion-specific agreement. After all readers completed the reviewing process, a consensus revision was made for all cases to determine the radiological standard of reference.

Biopsy and histopathology

The decision to perform or to avoid biopsy was made at the time of MRI examination and was based on the original report. All patients with at least one PI-RADS score ≥ 3 lesion at original MRI report ($n = 110$, 55%) underwent targeted biopsy with fusion or cognitive approach, as previously described [19]. Each patient was also concomitantly submitted to a standard 12-core random systematic biopsy (TRUS-Bx) [20]. Patients with negative MRI (maximum PI-RADS score < 3) underwent TRUS-Bx if deemed necessary by the treating physician ($n = 22$, 11%). The remaining patients with negative MRI did not undergo prostate biopsy ($n = 68$, 34%), neither immediately after MRI nor during routine follow-up. All prostate biopsy specimens were analyzed by dedicated uro-pathologists. Clinically significant PCa (csPCa) was considered as presence of any Gleason ≥ 3 + 4 (ISUP grade ≥ 2) at biopsy.

Biopsy results were then used to perform analyses on subgroup of patients harboring (or not) csPCa. It has to be noted that, since targeted biopsies were performed based on the original MRI report, their results could not correspond to the lesions identified after review using PI-RADS v2.1. Thus, all analyses based on biopsy should be considered on a per-patient level rather than a per-lesion level (i.e., agreement in men harboring or not csPCa).

Statistical analysis

Medians and interquartile ranges, as well as frequencies and proportions, were reported for continuous and categorical variables, respectively. Interreader agreement for multiple readers was evaluated using Conger's generalized kappa coefficient [21], contextually reporting raw percent agreement (PA).

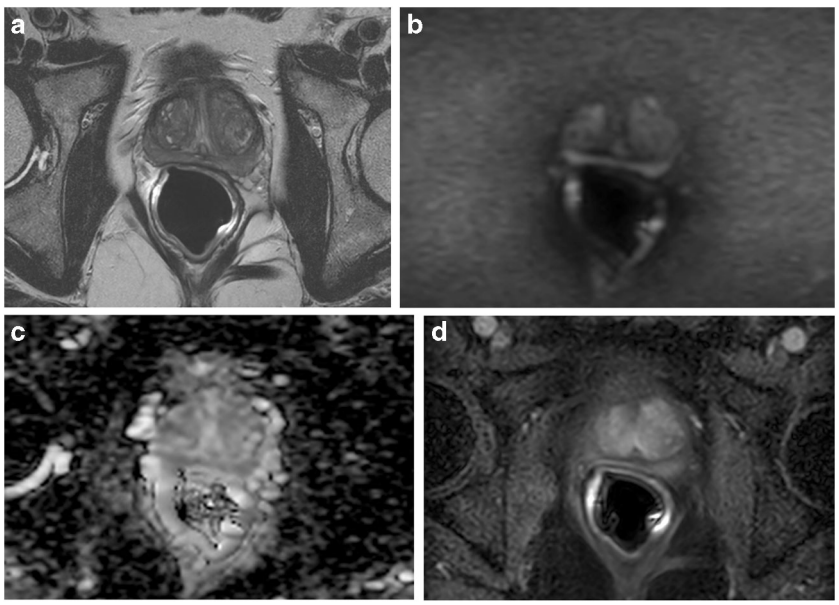

Fig. 1 mpMRI from a 56-year-old biopsy-naïve patient. PSA was 7.07 ng/mL, prostate volume 34 cc (PSA density 0.21 ng/mL/cc), negative digital rectal examination, and family history. Four readers reported the diffuse bilateral PZ signal changes as equivocal (PI-RADS score = 3), while three readers reported diffuse changes as likely benign (PI-RADS score = 2). Post-MRI biopsies were performed and histopathologic examination was negative for presence of prostate cancer. **a** T2-weighted images. **b** DWI (b1600). **c** ADC map. **d** Early post-injection DCE images

Similarly to the other kappa coefficients, Conger's kappa is exposed to the same well-known paradoxes [15, 16]; in particular, k values can be unexpectedly low even in presence of high agreement, depending on marginal frequencies. To overcome this issue, we additionally calculated agreement with alternative methods. First, agreement coefficient 1 (AC1) was computed [22]: based on the assumption that chance agreement is likely to affect only a portion of the observations, and not relying on assumed independence between observations of the readers, AC1 is less prone to paradoxes than k coefficients. Second, we measured interreader agreement for presence or absence of lesions using indexes of specific positive and negative agreement (P_{pos} and P_{neg}) [16]. The mathematical calculation of P_{pos} is identical to the Index of Specific Agreement (ISA) proposed by Shih et al [17] and represents the proportion of specific agreement relative to positive scores. Similarly, we calculated reader agreement on negative scores (negative MRI) by means of proportion of negative agreement (P_{neg}). Computing P_{pos} and P_{neg} allows to assess eventual differences in the agreement among readers on positive or negative cases. Even though these indexes are not chance-corrected (as k and AC1 coefficients), they are particularly suitable in this setting since the probability that more readers detect the same lesion in the same location by chance is negligible [17].

Levels of agreement were defined using the conventional classification of Landis and Koch [23]: slight (0–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and excellent (0.81–1). Even though it was originally proposed for k statistics, to simplify the comparison between different coefficients, we extended this categorization also to PA, AC1, and indexes of specific agreement.

Statistical tests were performed using AgreeStat 2015.6 and RStudio graphical interface v.0.98 for R software environment v.3.0.2 (R Foundation).

Subgroup analysis

To evaluate specific features that may influence variability, agreement analyses in index lesion detection (using a positivity threshold of PI-RADS score ≥ 3) were performed in subgroups of patients defined upon clinical, bioptic, or radiological parameters.

PSA values were considered relative to prostate volume and expressed as PSA density (PSAD); PSAD threshold was set at 0.15 ng/mL/cc [24]. Pre-MRI clinical risk was assessed with ESPRC risk calculator 6 [25]; patients were then divided in quartiles according to calculated risk, and divided in low risk (≤ 25th percentile), intermediate risk (25–50th percentile), and high risk (≥ 75th percentile).

With regard to post-MRI biopsy results, we divided patients in three groups: patients harboring csPCa (ISUP group ≥ 2), patients harboring non-csPCa (ISUP group 1), and patients with negative biopsies (no PCa group). We included patients without post-MRI biopsy in no PCa group given the low risk of harboring PCa [2], and considering acceptably this approximation as diagnostic performance analyses were not a purpose of our study. Biopsy results were also used to assess the concordance in true-positive (TP) vs. false-positive (FP) findings at MRI. Then, in men who underwent biopsy, we compared the prevalence of csPCa in concordant-positive cases vs. discordant cases, using all possible pairwise combination of dedicated and non-dedicated readers.

With regard to MRI parameters, the radiological gold standard was used to define the presence or absence of index lesion in PZ and TZ and to determine the homogeneity score of the PZ (SI classified as "homogeneous" for homogeneity scores ≥ 3). MRI interpretation was classified as "difficult" when at least one dedicated reader gave a subjective interpretative difficulty score of 3/3, while it was considered "easy" when none of the readers gave a subjective interpretative difficulty score of 3/3. Multifocality was defined as the presence of more than one PI-RADS score ≥ 3 lesion at MRI.

Results

Clinical, radiological, and pathological characteristics of the study cohort are summarized in Table 1.

Overall agreement

Table 2 shows the agreement in assessing lesions at mpMRI with PI-RADSv2.1 based on kappa coefficient, AC1, and PA.

Overall, dedicated readers showed higher concordance than non-dedicated readers. Using a cutoff of PI-RADS score ≥ 3 for index lesion detection, agreement was moderate among all readers ($k = 0.591$, 95% CI 0.529–0.653) and non-dedicated readers ($k = 0.562$, 95% CI 0.481–0.643), while it was substantial for dedicated readers ($k = 0.621$, 95% CI 0.548–0.694). Using a cutoff of PI-RADS > 3, agreement was substantial for all readers, with highest scores among dedicated readers ($k = 0.779$, 95% CI 0.711–0.846).

Considering AC1 values, agreement was substantial in all groups of readers using a cutoff of PI-RADS score ≥ 3, while it was excellent with a cutoff of PI-RADS 4 + 5, with highest scores among dedicated readers (AC1 = 0. 876, 95% CI 0.836–0.916).

Subgroup analysis

Analyses made on subgroups of patients according to clinical, post-MRI biopsy, and radiological parameters (Table 2) showed that k values were unreliable when the observations were unbalanced toward a specific category (i.e., high prevalence of positive or negative MRIs), being disproportionately

Table 1 Baseline characteristics of the study cohort ($n = 200$)

Parameters	
No. of patients	200
Age (years)	65 (58–70)
PSA (ng/mL)	6.0 (4.1–8.4)
Prostate volume (mL)	58.9 (42.4–79.2)
PSA density, PSAD (ng/mL/cc)	0.10 (0.07–0.16)
Pre-MRI biopsy status	
Biopsy-naïve	67% (135/200)
Prior negative biopsy	33% (65/200)
Clinical stage	
cT1	88% (176/200)
cT2,3	12% (24/200)
Positive familiar history	9% (17/200)
Original MRI report	
Negative MRI	45% (90/200)
Positive MRI	55% (110/200)
Post-MRI biopsy (positive MRI)	110
No PCa	47% (52/110)
Any PCa	53% (58/110)
csPCa	38% (42/110)
Post-MRI biopsy (negative MRI)	22
No PCa	86% (19/22)
Any PCa	14% (3/22)
csPCa	5% (1/22)
No biopsy (negative MRI)	68

Values are reported as frequencies, medians (interquartile range in parentheses), or percentages (proportions in parentheses). Positive MRI: at least one PI-RADS ≥ 3 lesion; negative MRI: absence of PI-RADS ≥ 3 lesion; csPCa: ISUP group ≥ 2

low compared with the raw percent of agreement (PA). Conversely, AC1 provided stable results paralleling more faithfully PA values. Also in subgroup analyses, dedicated readers showed overall higher concordance than non-dedicated readers.

When accounting for clinical and radiological parameters, we observed higher agreement among all groups of readers in patients with PSA density (PSAD) ≥ 0.15 ng/mL/cc, pre-MRI high, and low risk of PCa, PZ lesions, homogeneous SI of the peripheral zone, and easy interpretation of MRI, compared with patients with PSAD < 0.15 ng/mL/cc, intermediate pre-MRI risk, TZ lesions, inhomogeneous SI of the peripheral zone, and subjectively difficult interpretation of MRI, respectively (Table 2). Agreement was not significantly different in biopsy-naïve or previously negative biopsy patients.

When accounting for post-MRI biopsy results, agreement was excellent in patients harboring csPCa (AC1 = 0.859, 95% CI 0.782–0.936) and substantial in patients without PCa (AC1 = 0.744, 95% CI 0.693–0.795). Conversely, concordance was significantly lower in patients with non-csPCa at biopsy (AC1 = 0.522, 95% CI 0.344–0.701). Agreement in true-positive findings of MRI was excellent (AC1 = 0.858, 95% CI 0.780–0.936), while it was low in false-positive findings (AC1 = 0.528, 95% CI 0.436–0.619). Among men who underwent biopsy, mean prevalence of csPCa in concordant positive cases across dedicated and non-dedicated readers was 52.3% (range 48.1–56.5%), while in discordant cases was 10.7% (4.5–17.9%). Accordingly, mean false-positive rate of MRI was 47.7% (43.5–51.9%) and 89.3% (82.1–95.5%) in concordant positive cases and discordant cases, respectively.

Concordance on the presence of more than one lesion at MRI was substantial (AC1 = 0.721; 95% CI 0.603–0.838).

Indexes of specific agreement

Table 3 shows indexes of specific positive and negative agreement between dedicated and non-dedicated readers.

When accounting for percentages of positive specific agreement, concordance was significantly higher for dedicated than for non-dedicated radiologists. Agreement on index lesion was substantial for a cutoff of PI-RADS score ≥ 3 and excellent for dedicated readers for a cutoff of PI-RADS score 4 + 5. Positive agreement was as high as 93.4% (95% CI 90.7–95.4) between dedicated readers in patients with csPCa. Agreement of non-dedicated readers with a positivity threshold of PI-RADS score 4 + 5 approached that of dedicated readers with a positivity threshold PI-RADS score ≥ 3, even if it was still significantly lower.

Agreement on absence of lesions (negative MRI) was excellent both for dedicated and non-dedicated readers (respectively 85.1%, 95% CI 78.4–92.3; 82.0%, 95% CI 77.2–90.1), and it was as high as 93.6 (95% CI 90.5–96.5) for dedicated readers with a cutoff of PI-RADS score 4 + 5. In patients without PCa, negative specific agreement was excellent both for dedicated and non-dedicated readers (respectively 87.8%, 95% CI 81.7–91.8; 86.0%, 95% CI 79.8–90.6).

Discussion

In our study, we assessed the reproducibility of mpMRI reporting using PI-RADS v2.1 among multiple readers on a large cohort of patients who underwent mpMRI for a suspected PCa, reproducing a typical clinical workflow. We found overall good concordance among readers for index lesion detection, with excellent agreement in the subgroup of men harboring csPCa. As expected, concordance between experienced readers was generally higher than that between less-experienced readers.

Of note, agreement on absence of lesions was excellent across reader experience. To the best of our knowledge, this represents the first available study estimating the agreement on absence of lesions at MRI. This information is of

Table 2 Agreement on index lesion detection of Prostate Imaging Reporting and Data System version 2.1

Feature	All	Dedicated	Non-dedicated
All patients			
PI-RADS ≥ 3			
k	0.591 (0.529–0.653)	0.621 (0.548–0.694)	0.562 (0.481–0.643)
AC1	0.738 (0.695–0.782)	0.762 (0.713–0.811)	0.712 (0.652–0.771)
PA (%)	78.4 (74.9–81.9)	80.3 (76.3–84.3)	76.3 (71.5–81.1)
PI-RADS 4 + 5			
k	0.699 (0.634–0.764)	0.779 (0.711–0.846)	0.671 (0.589–0.753)
AC1	0.841 (0.806–0.878)	0.876 (0.836–0.916)	0.821 (0.772–0.870)
PA (%)	86.5 (83.5–89.6)	90.3 (87.3–93.4)	84.8 (80.8–88.9)
Clinical parameters			
Prev. Neg. Bx			
k	0.562 (0.435–0.690)	0.601 (0.430–0.771)	0.534 (0.388–0.679)
AC1	0.763 (0.687–0.839)	0.779 (0.675–0.884)	0.726 (0.628–0.824)
PA (%)	79.9 (73.7–86.1)	81.4 (72.8–89.9)	78.8 (71.7–85.9)
Biopsy-naïve			
k	0.587 (0.513–0.661)	0.639 (0.552–0.727)	0.533 (0.438–0.629)
AC1	0.730 (0.677–0.783)	0.769 (0.709–0.829)	0.685 (0.612–0.758)
PA (%)	77.8 (73.5–82.1)	80.9 (76.1–85.8)	74.2 (68.4–80.1)
PSAD > 0.15			
k	0.671 (0.545–0.797)	0.734 (0.605–0.864)	0.595 (0.433–0.758)
AC1	0.797 (0.715–0.880)	0.822 (0.732–0.913)	0.743 (0.630–0.857)
PA (%)	83.2 (76.4–90.0)	86.7 (80.0–93.3)	78.8 (69.6–88.0)
PSAD < 0.15			
k	0.544 (0.473–0.616)	0.585 (0.490–0.680)	0.536 (0.440–0.631)
AC1	0.716 (0.664–0.767)	0.740 (0.674–0.806)	0.699 (0.627–0.771)
PA (%)	76.5 (72.3–80.6)	78.5 (73.1–83.8)	75.2 (69.5–81.0)
Low risk			
k	0.532 (0.493–0.662)	0.546 (0.377–0.714)	0.551 (0.364–0.739)
AC1	0.738 (0.649–0.826)	0.753 (0.649–0.857)	0.738 (0.602–0.873)
PA (%)	77.9 (70.9–84.9)	79.1 (70.6–87.6)	78.0 (67.2–88.9)
Intermediate risk			
k	0.510 (0.422–0.599)	0.536 (0.431–0.641)	0.483 (0–367-0.559)
AC1	0.681 (0.618–0.745)	0.703 (0.631–0.775)	0.655 (0.569–0.742)
PA (%)	73.8 (68.7–78.9)	75.5 (69.6–81.4)	71.7 (64.9–78.7)
High risk			
k	0.754 (0.640–0.869)	0.807 (0.682–0.932)	0.683 (0.536–0.831)
AC1	0.844 (0.766–0.921)	0.879 (0.797–0.961)	0.794 (0.689–0.898)
PA (%)	87.1 (80.8–93.4)	90.0 (83.2–96.8)	83.0 (74.6–91.5)
MRI parameters			
PZ			
k	0.248 (0.141–0.354)	0.237 (0.093–0.380)	0.289 (0.137–0.441)
AC1	0.694 (0.615–0.774)	0.694 (0.592–0.796)	0.689 (0.589–0.789)
PA (%)	73.1 (66.8–79.4)	74.4 (67.0–82.0)	72.8 (64.8–80.9)
TZ			
k	0.210 (0.058–0.362)	0.175 (0.033–0.383)	0.252 (0.052–0.557)
AC1	0.633 (0.461–0.805)	0.728 (0.549–0.906)	0.535 (0.294–0.776)
PA (%)	68.2 (54.8–81.5)	75.4 (60.7–90.2)	61.4 (42.6–80.2)

Table 2 (continued)

Feature	All	Dedicated	Non-dedicated
Homogeneous SI			
k	0.766 (0.675–0.858)	0.804 (0.705–0.904)	0.677 (0.549–0.806)
AC1	0.839 (0.773–0.905)	0.807 (0.708–0.906)	0.767 (0.663–0.871)
PA (%)	88.0 (83.2–92.9)	90.3 (85.3–95.2)	82.9 (75.4–90.3)
Inhomogeneous SI			
k	0.496 (0.420–0.573)	0.523 (0.430–0.617)	0.499 (0.396–0.602)
AC1	0.671 (0.615–0.728)	0.693 (0.626–0.760)	0.666 (0.588–0.744)
PA (%)	73.0 (68.5–77.6)	74.7 (69.4–80.1)	72.7 (66.4–78.9)
Difficult			
k	0.185 (0.072–0.297)	0.191 (0.032–0.350)	0.138 (0.008–0.297)
AC1	0.460 (0.355–0.565)	0.436 (0.293–0.578)	0.387 (0.227–0.548)
PA (%)	55.8 (47.8–63.7)	58.1 (48.5–67.8)	50.4 (38.4–62.4)
Easy			
k	0.685 (0.619–0.751)	0.718 (0.641–0.795)	0.665 (0.578–0.753)
AC1	0.811 (0.770–0.853)	0.833 (0.785–0.881)	0.796 (0.738–0.854)
PA (%)	84.3 (80.9–87.7)	86.1 (82.1–90.0)	83.0 (78.3–87.8)
Multifocality			
k	0.508 (0.356–0-660)	0.475 (0.301–0.649)	0.432 (0.241–0.623)
AC1	0.721 (0.603–0.838)	0.718 (0.583–0.852)	0.644 (0.482–0.806)
PA (%)	82.1 (76.1–88.1)	81.6 (74.5–88.7)	77.8 (69.4–86.2)
Post-MRI biopsy			
ISUP group ≥ 2			
k	0.370 (0.108–0.632)	0.394 (0.058–0.730)	0.443 (0.174–0.712)
AC1	0.859 (0.782–0.936)	0.888 (0.809–0.968)	0.843 (0.740–0.945)
PA (%)	86.9 (80.1–93.7)	89.8 (82.9–96.6)	85.6(76.8–94.5)
ISUP group 1			
k	0.184 (0.071–0.439)	0.207 (0.061–0.475)	0.115 (0.024–0.425)
AC1	0.522 (0.344–0.701)	0.501 (0.271–0.731)	0.458 (0.149–0.768)
PA (%)	59.9 (46.1–73.8)	61.8 (46.1–77.4)	54.9 (31.5–78.3)
No PCa			
k	0.478 (0.391–0.565)	0.501 (0.395–0.608)	0.456 (0.345–0.566)
AC1	0.744 (0.693–0.795)	0.764 (0.705–0.823)	0.693 (0.615–0.772)
PA (%)	78.0 (73.9–82.1)	79.6 (74.7–84.5)	76.0 (70.4–81.7)
True positive			
k	0.229 (0.079–0.378)	0.209 (0.021–0.439)	0.352 (0.085–0.619)
AC1	0.858 (0.780–0.936)	0.891 (0.814–0.967)	0.841 (0.737–0.944)
PA (%)	86.6 (79.7–93.5)	89.5 (82.5–96.5)	85.3 (76.2–94.3)
False positive			
k	0.144 (0.067–0.221)	0.198 (0.035–0.361)	0.115 (0.019–0.213)
AC1	0.528 (0.436–0.619)	0.533 (0.400–0.666)	0.529 (0.419–0.638)
PA (%)	60.1 (53.2–67.0)	60.8 (50.5–71.1)	59.9 (51.6–68.3)

Values in parenthesis are 95% confidence intervals

PA, percentage of agreement; *Prev. Neg. Bx*, previous negative biopsy; *PSAD*, PSA density (expressed in ng/mL/cc); *PZ*, peripheral zone; *TZ*, transitional zone; *SI*, signal intensity; *PCa*, prostate cancer

paramount importance, given that the main strength of prostate MRI relies on its sensitivity and negative predictive value [2], and the most significant effect of its implementation has been to avoid biopsy in a substantial proportion of men [3].

Table 3 Indexes of specific agreement of index lesion detection

	Dedicated	Non-dedicated
P_{pos} (%)		
PI-RADS ≥ 3	75.7 (75.1–77.2)	63.6 (62.7–64.6)
PI-RADS 4 + 5	82.8 (81.2–84.3)	70.0 (68.6–71.4)
csPCa	93.4 (90.7–95.4)	86.0 (84.2–89.1)
P_{neg} (%)		
PI-RADS ≥ 3	85.1 (78.4–92.3)	82.0 (77.2–90.1)
PI-RADS 4 + 5	93.6 (90.5–96.5)	90.9 (87.3–94.5)
No PCa	87.8 (81.7–91.8)	86.0 (79.8–90.6)

Values in parenthesis are 95% confidence intervals

P_{pos}, proportion of positive agreement; P_{neg}, proportion of negative agreement

While agreement based on *k* values was comparable to previous multireader studies [8–14], we confirmed how this statistical index may actually underestimate the true extent of the agreement, and how it could be unreliable in situations of unbalanced marginal frequencies compared with other coefficients (i.e., AC1 coefficient) [15, 16, 22]. These evidences raise questions about its suitability in prostate MRI image analysis, given that the probability of chance agreement in this setting is negligible [17]. Correspondingly, our results are in line with other studies that evaluated indexes of specific agreement [26, 27], and confirm that concordance in mpMRI reporting using PI-RADS scoring system may be actually higher than previously reported.

Besides readers' experience, other factors may affect the agreement. When positivity threshold of MRI was set to PI-RADS 4 + 5, variability has been significantly reduced; this is a clue of how PI-RADS score 3 lesions are a substantial source of interreader variability, especially in less-experienced readers. Interestingly, to reach a similar level of positive agreement between experienced and non-experienced readers, the positivity threshold of MRI for less-experienced readers should be heightened to PI-RADS scores 4 + 5 (Table 3). Furthermore, we observed more consistent scores between readers in patients at higher risk of PCa, for PZ lesions, in presence of homogeneous SI of the PZ and when MRI interpretation was subjectively judged easily (Table 2). In general terms, these findings may indicate that there is not one absolute value of interreader agreement: reproducibility is expected to be higher in those patients at higher risk of having an obvious lesion in the PZ, with less background PZ inhomogeneity, and thus in MRI examinations that are objectively easier to score. Accordingly, while agreement is higher for true-positive findings, the majority of discordant cases (i.e., harder to score cases) are more probably related to false-positive findings. Interestingly, within the subgroup of patients who underwent biopsy, we observed that the false positivity rate of MRI was as high as 88.8% in presence of discordant findings between two different readers.

We also found good agreement on multifocality assessment, defined as the presence of more than one suspicious lesion (PI-RADS score ≥ 3) at MRI. However, taking into account the possibility of false-positive agreement on non-index lesions, actual concordance on multifocal disease is expected to be lower, as previously reported [27]. This consideration is relevant, as good concordance on multifocal disease in the presence of low lesion-specific agreement may furtherly support the need to perform random biopsies in adjunction to target biopsies in presence of suspicious lesions at MRI [19, 28].

Compared with other studies on interreader agreement, our study design has the advantage to reproduce a typical clinical scenario both in terms of image interpretation process and in patients' cohort, providing greater generalizability of the findings. Specifically, our readers had full access both to MRI examinations on a PACS workstation and to pre-MRI clinical information; conversely, studies based on scoring predetermined lesions on screen captures or blinding readers to relevant clinical information do not reflect the routine radiological workflow [8–10, 13]. Second, all available studies so far included only patients who underwent biopsy or radical prostatectomy [8–10, 12–14, 26, 27], biasing the cohort toward high prevalence of positive MRI and clinically significant PCa. However, although useful for sub-analyses, histopathologic reference standard is not necessary for inter-observer studies per se. Accordingly, we included all consecutive men who underwent MRI for clinically suspected PCa, regardless of post-MRI biopsy results. This allowed us to have a cohort of patients that was more representative of the contemporary general population of men referred for prostate MRI for a clinical suspicion of PCa; at the same time, we were able to calculate agreement on negative MRI including patients that do not routinely undergo biopsy in clinical practice, according to latest guidelines [3].

Our study is not devoid of limitations. First, despite our cohort was as much as possible representative of the general population, the retrospective nature of this study may have introduced selection biases. Therefore, prospective studies are needed to further validate our results.

Moreover, all readers came from a single tertiary care referral center, where less-experienced radiologists undergo specific training in prostate MRI led by experienced colleagues; this may induce readers to approach cases with similar interpretation schemes, reducing variability a priori. Thus, the generalizability of our findings may be limited to centers with a similar training. In the real world, the actual extent of variability between experienced radiologists from academic centers and non-experienced radiologists from non-academic centers may be significantly higher [29]. Multicenter studies should be performed to address this limitation.

Finally, the lack of a direct comparison between PI-RADS v2 and v2.1 limited our possibility to assess the potential improvements in agreement that have been auspicated in the

latest update. However, since PI-RADS v2.1 added minor changes to the previous version (some of which apply to relatively rare conditions, e.g., central zone and anterior fibromuscular stromal tumors, atypical TZ nodules), to reliably detect significant differences between the two systems a larger number of cases is required. Future multicenter studies are needed to overcome this issue. Nonetheless, our study represents a valuable standpoint on agreement using PI-RADS scoring system, giving insights on its clinical implications and on the investigative methodology that could be used for future studies.

In conclusion, we observed overall good reproducibility of prostate MRI interpretation between appropriately trained radiologists of different expertise using PI-RADS v2.1. In particular, agreement is excellent between experienced readers in index lesion detection and across readers' experience in determining the absence of lesions at MRI.

Funding information This work has not received any funding.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Francesco De Cobelli, M.D.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry Paola Maria Vittoria Rancoita kindly provided statistical advice for this manuscript, and is one of the authors.

Informed consent Written informed consent was obtained from all subjects (patients) in this study.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap None of the study subjects or cohorts has been previously reported.

Methodology
- Retrospective
- Observational
- Performed at one institution

References

1. Kasivisvanathan V, Rannikko AS, Borghi M et al (2018) MRI-targeted or standard biopsy for prostate-cancer diagnosis. N Engl J Med 378:1767–1777. https://doi.org/10.1056/NEJMoa1801993
2. Ahmed HU, El-Shater Bosaily A, Brown LC et al (2017) Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet. https://doi.org/10.1016/S0140-6736(16)32401-1
3. Mottet N, van den Bergh RCN, Briers E et al (2019) Guidelines on prostate cancer 2019. Eur Assoc Urol. Available via https://uroweb.org/guideline/prostate-cancer/
4. Drost FJH, Osses DF, Nieboer D et al (2019) Prostate MRI, with or without MRI-targeted biopsy, and systematic biopsy for detecting prostate cancer. Cochrane Database Syst Rev. https://doi.org/10.1002/14651858.CD012663.pub2
5. Barentsz JO, Richenberg J, Clements R et al (2012) ESUR prostate MR guidelines 2012. Eur Radiol. https://doi.org/10.1007/s00330-011-2377-y
6. Weinreb JC, Barentsz JO, Choyke PL et al (2016) PI-RADS Prostate Imaging - Reporting and Data System: 2015, version 2. Eur Urol. https://doi.org/10.1016/j.eururo.2015.08.052
7. Turkbey B, Rosenkrantz AB, Haider MA et al (2019) Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System version 2. Eur Urol. https://doi.org/10.1016/j.eururo.2019.02.033
8. Rosenkrantz AB, Ginocchio LA, Cornfeld D et al (2016) Interobserver reproducibility of the PI-RADS version 2 lexicon: a multicenter study of six experienced prostate radiologists. Radiology. https://doi.org/10.1148/radiol.2016152542
9. Schimmöller L, Quentin M, Arsov C et al (2013) Inter-reader agreement of the ESUR score for prostate MRI using in-bore MRI-guided biopsies as the reference standard. Eur Radiol. https://doi.org/10.1007/s00330-013-2922-y
10. Muller BG, Shih JH, Sankineni S et al (2015) Prostate cancer: interobserver agreement and accuracy with the revised Prostate Imaging Reporting and Data System at multiparametric MR imaging. Radiology. https://doi.org/10.1148/radiol.2015142818
11. Sonn GA, Fan RE, Ghanouni P et al (2018) Prostate magnetic resonance imaging interpretation varies substantially across radiologists. Eur Urol Focus. https://doi.org/10.1016/j.euf.2017.11.010
12. Pickersgill NA, Vetter JM, Andriole GL et al (2018) Accuracy and variability of prostate multiparametric magnetic resonance imaging interpretation using the prostate imaging reporting and data system: a blinded comparison of radiologists. Eur Urol Focus. https://doi.org/10.1016/j.euf.2018.10.008
13. Smith CP, Harmon SA, Barrett T et al (2019) Intra- and interreader reproducibility of PI-RADSv2: a multireader study. J Magn Reson Imaging. https://doi.org/10.1002/jmri.26555
14. Girometti R, Giannarini G, Greco F et al (2019) Interreader agreement of PI-RADS v. 2 in assessing prostate cancer with multiparametric MRI: a study using whole-mount histology as the standard of reference. J Magn Reson Imaging 49:546–555. https://doi.org/10.1002/jmri.26220
15. Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. the problems of two paradoxes. J Clin Epidemiol. https://doi.org/10.1016/0895-4356(90)90158-L
16. Cicchetti DV, Feinstein AR (1990) High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. https://doi.org/10.1016/0895-4356(90)90159-M
17. Shih JH, Greer MD, Turkbey B (2018) The problems with the kappa statistic as a metric of interobserver agreement on lesion detection using a third-reader approach when locations are not prespecified. Acad Radiol. https://doi.org/10.1016/j.acra.2018.01.030
18. Hötker AM, Dappa E, Mazaheri Y et al (2019) The influence of background signal intensity changes on cancer detection in prostate MRI. Am J Roentgenol 212:823–829. https://doi.org/10.2214/AJR.18.20295
19. Stabile A, Dell'Oglio P, De Cobelli F et al (2018) Association between Prostate Imaging Reporting and Data System (PI-RADS) score for the index lesion and multifocal, clinically significant prostate cancer. Eur Urol Oncol 1:29–36. https://doi.org/10.1016/j.euo.2018.01.002
20. Dell'Oglio P, Stabile A, Soligo M et al (2019) There is no way to avoid systematic prostate biopsies in addition to multiparametric magnetic resonance imaging targeted biopsies. Eur Urol Oncol 3:112–118. https://doi.org/10.1016/j.euo.2019.03.002

21. Conger AJ (1980) Integration and generalization of kappas for multiple raters. Psychol Bull. https://doi.org/10.1037/0033-2909.88.2.322
22. Gwet KL (2008) Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 61:29–48. https://doi.org/10.1348/000711006X126600
23. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics. https://doi.org/10.2307/2529310
24. Rais-Bahrami S, Siddiqui MM, Vourganti S et al (2015) Diagnostic value of biparametric magnetic resonance imaging (MRI) as an adjunct to prostate-specific antigen (PSA)-based detection of prostate cancer in men without prior biopsies. BJU Int 115:381–388. https://doi.org/10.1111/bju.12639
25. Roobol MJ, Verbeek JFM, van der Kwast T, Kümmerlin IP, Kweldam CF, van Leenders GJLH (2017) Improving the Rotterdam European randomized study of screening for prostate cancer risk calculator for initial prostate biopsy by incorporating the 2014 International Society of Urological Pathology Gleason Grading and Cribriform growth. Eur Urol. https://doi.org/10.1016/j.eururo.2017.01.033
26. Greer MD, Brown AM, Shih JH et al (2017) Accuracy and agreement of PIRADSv2 for prostate cancer mpMRI: a multireader study. J Magn Reson Imaging 45:579–585. https://doi.org/10.1002/jmri.25372
27. Greer MD, Shih JH, Lay N et al (2019) Interreader variability of prostate imaging reporting and data system version 2 in detecting and assessing prostate cancer lesions at prostate MRI. AJR Am J Roentgenol 212:1197–1205. https://doi.org/10.2214/AJR.18.20536
28. Rouvière O, Puech P, Renard-Penna R et al (2019) Use of prostate systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-naive patients (MRI-FIRST): a prospective, multicentre, paired diagnostic study. Lancet Oncol 20:100–109. https://doi.org/10.1016/S1470-2045(18)30569-2
29. Luzzago S, Petralia G, Musi G et al (2019) Multiparametric magnetic resonance imaging second opinion may reduce the number of unnecessary prostate biopsies: time to improve radiologists' training program? Clin Genitourin Cancer 17:88–96. https://doi.org/10.1016/j.clgc.2018.10.006