Effect of Prostate MRI Interpretation Experience on PPV Using PI-RADS Version 2: A 6-Year Assessment Among Eight Fellowship-Trained Radiologists

Bassel R. Salka, BSE[1], Prasad R. Shankar, MD[2], Jonathan P. Troost, PhD[3], Shokoufeh Khalatbari, MS[3], Matthew S. Davenport, MD[2,4]

Genitourinary Imaging · Original Research

Keywords

experience, PPV, prostate, prostate MRI, quality

Submitted: Jan 19, 2022
Revision requested: Jan 28, 2022
Revision received: Feb 25, 2022
Accepted: Mar 15, 2022
First published online: Mar 23, 2022

M. S. Davenport receives royalties from Wolters Kluwer and uptodate.com. The remaining authors declare that there are no additional disclosures relevant to the subject matter of this article.

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as representing the views of the NIH.

Supported in part by NIH grant UL1TR002240 for statistical analysis.

BACKGROUND. Understanding the effect of specific experience in prostate MRI interpretation on diagnostic performance would help inform the minimum interpretation volume to establish proficiency.

OBJECTIVE. The purpose of this article is to assess for an association between increasing experience in prostate MRI interpretation and change in radiologist-level PPVs for PI-RADS version 2 (v2) categories 3, 4, and 5.

METHODS. This retrospective study included prostate MRI examinations performed between July 1, 2015, and August 13, 2021, that were assigned a PI-RADS v2 category of 3, 4, or 5 and with an MRI-ultrasound fusion biopsy available as the reference standard. All examinations were among the first 100–200 prostate MRI examinations interpreted using PI-RADS v2 by fellowship-trained abdominal radiologists. Radiologists received feedback through a quality assurance program. Radiologists' experience levels were classified using progressive subsets of 50 interpreted examinations. Change with increasing experience in distribution of individual radiologists' whole-gland PPVs for Gleason sum score 7 or greater prostate cancer, stratified by PI-RADS category, was assessed by hierarchic linear mixed models.

RESULTS. The study included 1300 prostate MRI examinations in 1037 patients (mean age, 66 ± 7 [SD] years), interpreted by eight radiologists (median, 13 years of postfellowship experience; range, 5–22 years). Aggregate PPVs were 20% (68/340) for PI-RADS category 3, 49% (318/652) for category 4, and 71% (220/308) for category 5. Interquartile ranges (IQRs) of PPVs overlapped for category 4 (51%; IQR, 42–60%) and category 5 (70%; IQR, 54–75%) for radiologists' first 50 examinations. IQRs of PPVs did not overlap between categories of greater experience; for example, at the 101–150 examination level, PPV for category 3 was 24% (IQR, 20–29%), category 4 was 55% (IQR, 54–63%), and category 5 was 81% (IQR, 77–82%). Hierarchic modeling showed no change in radiologists' absolute PPV with increasing experience (category 3, $p = .27$; category 4, $p = .71$; category 5, $p = .38$).

CONCLUSION. Absolute PPVs at specific PI-RADS categories did not change during radiologists' first 200 included examinations. However, resolution of initial overlap in IQRs indicates improved precision of PPVs after the first 50 examinations.

CLINICAL IMPACT. If implementing a minimum training threshold for fellowship-trained abdominal radiologists, 50 prostate MRI examinations may be sufficient in the context of a quality assurance program with feedback.

PI-RADS version 2 (v2) was developed in 2015 to standardize prostate MRI performance and interpretation [1]. This standardization enables automated and structured analysis of outcomes that in turn enables personalized radiologist performance improvement [2]. From the inception of PI-RADS, studies have shown that radiologists with a wide range of experience can obtain consistent PPVs, supporting the system's value in quality assurance benchmarking [3, 4].

Both aggregate and individual radiologist performances have been evaluated using PI-RADS v2 [3, 5, 6]. However, less is known about change over time in individual radiologist performance. It is generally anticipated that as an individual radiologist's interpretation volume increases accompanied by histopathologic feedback, their accuracy and risk

ARRS is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to provide continuing medical education activities for physicians. The ARRS designates this journal-based CME activity for a maximum of 1.00 AMA PRA Category 1 Credits™ and 1.00 American Board of Radiology©, MOC Part II, Self-Assessment CME (SA-CME). Physicians should claim only the credit commensurate with the extent of their participation in the activity. To access the article for credit, follow the prompts associated with the online version of this article.

doi.org/10.2214/AJR.22.27421
AJR 2022; 219:453–461
ISSN-L 0361–803X/22/2193–453

[1]Department of Radiology, University of Michigan School of Medicine, Ann Arbor, MI.
[2]Department of Radiology, Michigan Medicine, Ann Arbor, MI.
[3]Michigan Institute for Clinical and Health Research (MICHR), Michigan Medicine, Ann Arbor, MI.
[4]Department of Urology, Michigan Medicine, 1500 E Medical Center Dr, Ann Arbor, MI 48109. Address correspondence to M. S. Davenport (matdaven@med.umich.edu).

stratification using PI-RADS v2 will improve. For example, if a radiologist recognizes that their PPV for higher-risk prostate cancer (Gleason sum score ≥ 7) is below nationally accepted thresholds when assigning the highest-risk designation (PI-RADS category 5), they might review those cases for which they assigned that category and then modify their future behavior (e.g., assign PI-RADS category 5 less often). Through this process, radiologists' PPV is expected to iteratively approach established benchmarks as they gain more experience (i.e., as they interpret an increased volume of cases with ongoing feedback). If applied across radiologists in a practice, this feedback mechanism should also result in increased precision (i.e., less overlap) for the PPV of PI-RADS categories across the practice's radiologists as the group collectively gains experience.

The purpose of this investigation was to assess for a potential association between increasing experience in prostate MRI interpretation and change in radiologist-level PPVs for PI-RADS v2 categories 3, 4, and 5. This information will help inform the minimum number of positive examinations that must be interpreted with subsequent feedback according to a reference standard to establish proficiency in prostate MRI interpretation.

HIGHLIGHTS

Key Finding

- *Increasing experience did not affect absolute PPVs for specific PI-RADS v2 categories in a cohort of eight fellowship-trained abdominal radiologists who interpreted up to 200 prostate MRI examinations without prior PI-RADS v2 experience; however, the precision of radiologists' PPVs improved after the first 50 examinations.*

Importance

- *Practices adopting PI-RADS v2 might wish to consider radiologists' first 50 prostate MRI examinations, interpreted with histopathologic feedback, as part of a structured training experience.*

Methods

This was a single-center HIPAA-compliant retrospective cohort analysis that was exempt from institutional review board approval given the review board's determination that the analysis had nonregulated status because of its use of quality assurance data.

Patient Selection

Institutional databases (described later in the Methods) were searched for consecutive multiparametric prostate MRI examinations performed at Michigan Medicine between July 1, 2015 (date of initiation of PI-RADS v2 utilization and templated reporting at the study institution), to August 13, 2021 (date of initiation of this investigation). This search identified 8761 prostate MRI examinations interpreted by 31 fellowship-trained abdominal radiologists. Prostate MRI examinations were eligible for inclusion regardless of the patient's biopsy history (e.g., biopsy-naive, prior negative biopsy, or prior positive biopsy). The search did not capture examinations performed at outside institutions. Of the identified prostate MRI examinations, 4346 were excluded because they had been assigned a PI-RADS category 1 or 2, and 1877 were excluded because of a lack of an MRI-ultrasound (US) fusion biopsy of the reported lesion(s) performed within 2 years of the MRI to serve as a reference standard. After these exclusions, 2538 prostate MRI examinations interpreted by 23 radiologists remained potentially eligible for inclusion. Then, 456 prostate MRI examination examinations were excluded because they were interpreted by a radiologist who had interpreted fewer than 100 examinations; 610 examinations were excluded because of the inclusion of only the first 100, first 150, or first 200 examinations interpreted by each radiologist (as described later in the Methods); and 172 examinations were excluded because they were interpreted by a radiologist with faculty-level experience in prostate MRI interpretation using PI-RADS v2 before the study period and for whom performance data from before the study period were unavailable. These exclusions resulted in a final study sample of 1300 prostate MRI examinations assigned a PI-RADS category of 3, 4, or 5 with a reference standard available and that were among the first 100, first 150, or first 200 such examinations interpreted by

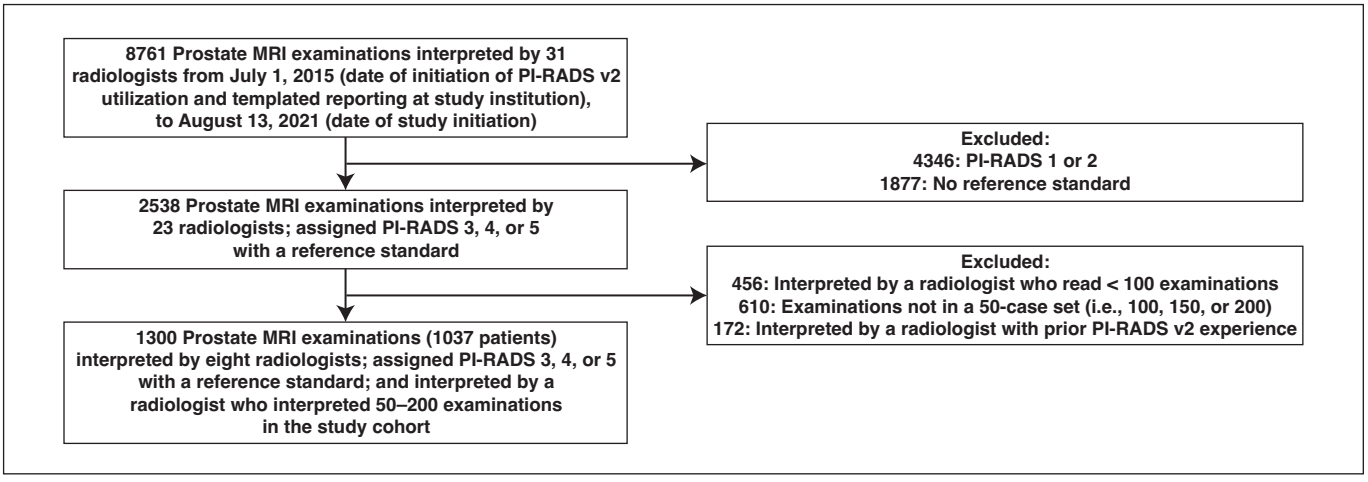

Fig. 1—Flow diagram shows study sample. PI-RADS v2 = PI-RADS version 2.

one of eight radiologists within the sample. The examinations were performed in a total of 1037 patients (mean age at the time of the first MRI of 66 ± 7 [SD] years). The study sample flow diagram is provided in Figure 1. A total of 704 patients in the current study were also included in an earlier investigation [3] that evaluated dispersion of radiologists' PPVs when using PI-RADS v2 but that did not evaluate the impact of experience and feedback on changes in performance.

TABLE 1: PPV of Prostate MRI Examinations Assigned a PI-RADS Version 2 Category of 3, 4, or 5, Stratified by Radiologist and Experience Level

Radiologist and PI-RADS Category	No. of MRI Examinations Read				All
	1–50	51–100	101–150	151–200	
Radiologist A					
Category 3	15 (2/13)	0 (0/11)	10 (1/10)	0 (0/10)	7 (3/44)
Category 4	67 (12/18)	52 (11/21)	54 (15/28)	43 (13/30)	53 (51/97)
Category 5	63 (12/19)	83 (15/18)	67 (8/12)	70 (7/10)	71 (42/59)
Radiologist B					
Category 3	28 (7/25)	21 (3/14)	24 (8/33)	15 (3/20)	23 (21/92)
Category 4	75 (12/16)	30 (6/20)	55 (6/11)	63 (12/19)	55 (36/66)
Category 5	89 (8/9)	56 (9/16)	83 (5/6)	64 (7/11)	69 (29/42)
Radiologist C					
Category 3	19 (4/21)	15 (3/20)	29 (2/7)	33 (2/6)	20 (11/54)
Category 4	46 (12/26)	32 (7/22)	63 (19/30)	56 (20/36)	51 (58/114)
Category 5	33 (1/3)	63 (5/8)	77 (10/13)	88 (7/8)	72 (23/32)
Radiologist D					
Category 3	21 (3/14)	22 (2/9)	36 (5/14)	40 (4/10)	30 (14/47)
Category 4	52 (12/23)	39 (12/31)	65 (13/20)	42 (10/24)	48 (47/98)
Category 5	69 (9/13)	70 (7/10)	81 (13/16)	81 (13/16)	76 (42/55)
Radiologist E					
Category 3	25 (3/12)	18 (2/11)	20 (1/5)	0 (0/6)	18 (6/34)
Category 4	54 (15/28)	56 (15/27)	38 (13/34)	53 (16/30)	50 (59/119)
Category 5	70 (7/10)	75 (9/12)	82 (9/11)	86 (12/14)	79 (37/47)
Radiologist F					
Category 3	25 (4/16)	18 (2/11)	—	—	22 (6/27)
Category 4	50 (10/20)	38 (10/26)	—	—	43 (20/46)
Category 5	71 (10/14)	77 (10/13)	—	—	74 (20/27)
Radiologist G					
Category 3	27 (4/15)	7 (1/15)	—	—	17 (5/30)
Category 4	38 (10/26)	48 (11/23)	—	—	43 (21/49)
Category 5	78 (7/9)	58 (7/12)	—	—	67 (14/21)
Radiologist H					
Category 3	0 (0/6)	33 (2/6)	—	—	17 (2/12)
Category 4	34 (12/35)	50 (14/28)	—	—	41 (26/63)
Category 5	44 (4/9)	56 (9/16)	—	—	52 (13/25)
All radiologists					
Category 3	22 (27/122)	15 (15/97)	25 (17/69)	17 (9/52)	20 (68/340)
Category 4	49 (95/192)	43 (86/198)	54 (66/123)	51 (71/139)	49 (318/652)
Category 5	67 (58/86)	68 (71/105)	78 (45/58)	78 (46/59)	71 (220/308)

Note—Values represent percentage with numerator and denominator in parentheses. Dash (—) indicates data not available.

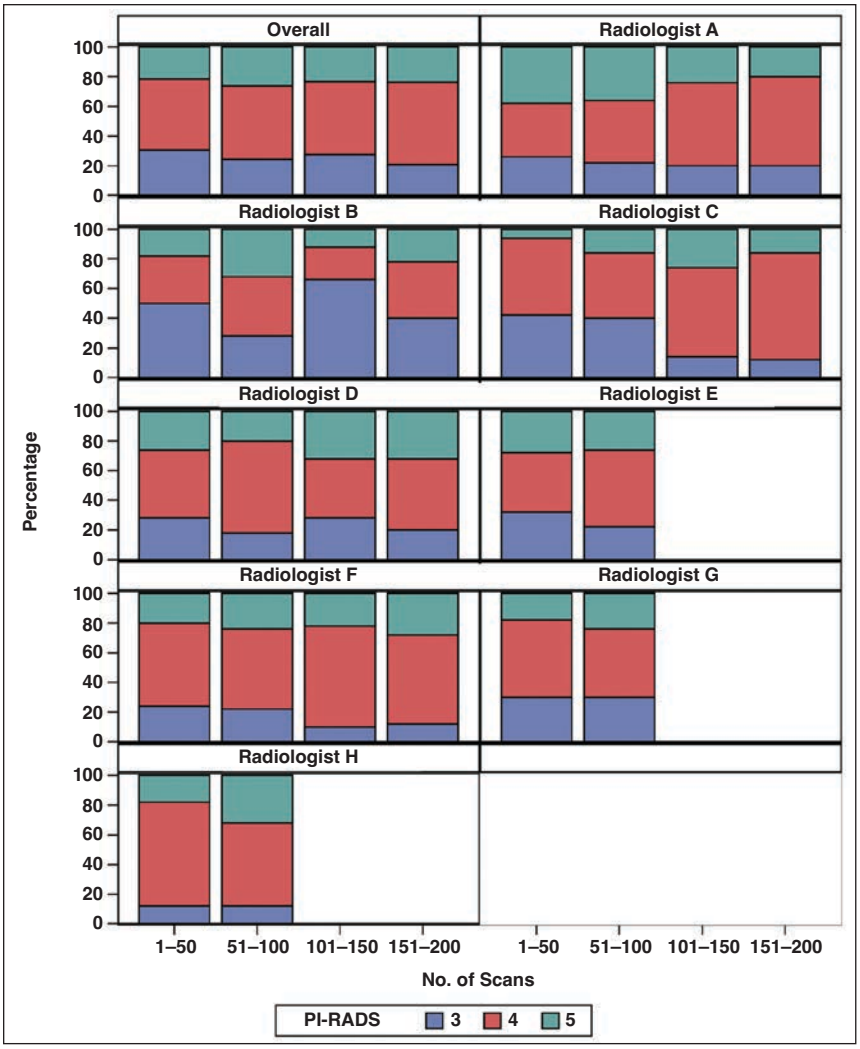

Fig. 2—Graphs show percentage of prostate MRI examinations assigned PI-RADS category of 3, 4, or 5, stratified by radiologist and experience level.

Study Setting

PI-RADS v2 was used for MRI interpretation at the institution for the entire study period. Before the start of the study period, the institution had not used PI-RADS version 1 (v1) or v2, and none of the radiologists included in the study had experience with PI-RADS v1 or PI-RADS v2 before the study period. In addition, the study institution did not adopt use of PI-RADS version 2.1 after its release in March 2019, pending publication of supporting evidence. Before the study period, the study institution performed fewer than 50 prostate MRI examinations annually, and prostate MRI was performed primarily for assessment of T category of known prostate cancer. However, during the study period, prostate MRI was performed primarily for detection of clinically important prostate cancer, and prostate MRI utilization increased throughout the study period to a volume of approximately 1500 examinations in the final study year. After the institution's adoption of PI-RADS v2 and development of quality assurance databases, radiologists received ongoing feedback during the study period. This feedback was variable in nature and consisted of peer learning; multidisciplinary conferences among radiologists, urologists, and pathologists with case-level feedback; one-on-one and group-level instruction; and dissemination of data from quality assurance databases.

Data Sources and Study Variables

The databases used for the initial patient search and subsequent data extraction included a manual institutional quality assurance database (used from July 1, 2015, to October 7, 2015) and, after an update within the radiology information system, an automated dashboard (Tableau Software, v2019.3) (used from October 8, 2015, to August 13, 2021) [3]. These databases were developed to improve radiologist performance by tracking diagnostic performance using PI-RADS v2 at aggregate and individual radiologist levels and by providing radiologists with case-level data regarding false-negative and false-positive findings. Variables extracted from the databases for each examination included date of MRI interpretation, maximum PI-RADS v2 category, interpreting radiologist, sequential count among included examinations interpreted by the radiologist, and maximum Gleason sum score from subsequent prostate biopsy.

MRI Examinations

In all patients, the multiparametric prostate MRI acquisition and clinical interpretation were performed in accordance with the institution's routine clinical practice, as previously reported

[3]. Examinations were performed according to PI-RADS v2 criteria on a 3-T scanner (Ingenia, Philips Healthcare, or Vida, Siemens Healthcare) using a variable-channel surface coil. The examinations were interpreted with PI-RADS v2 criteria using a standardized reporting template.

Reference Standard

Prostate biopsies were performed by one of approximately 10 faculty urologists and included two to four core samples of each target obtained using MRI-US fusion targeting and a 12-core systematic biopsy. The biopsy results were evaluated at the whole-gland level. Therefore, biopsies that showed higher-risk prostate cancer (i.e., Gleason sum score ≥ 7) on any core (targeted or systematic) were considered positive for purposes of analysis.

Sample Size, Primary Outcome Measurements, and Data Analysis

The sample comprised all examinations meeting study inclusion and exclusion criteria, without a priori study sample size determination. A minimum of 100 MRI examinations per radiologist was chosen to decrease bias from small sample sizes and to permit assessment of the effect on PPV of interpretation experience with feedback. The primary outcome was the change in individual radiologists' distributions (medians and 1st and 3rd quartiles) of whole-gland PPVs. PPV was calculated as the ratio between the number of true-positive examinations for higher-risk prostate cancer (i.e., biopsy showing Gleason sum score ≥ 7) and the total number of examinations (i.e., true-positive and false-positive examinations), expressed as a percentage. PPVs were calculated for the entire study sample (all radiologists) and for individual included radiologists. Aggregate PPV was also calculated for each year of the study period to assess for evidence of longitudinal bias. The proportion of examinations assigned to each PI-RADS category was calculated by individual radiologists and overall.

PPVs were calculated stratified by both PI-RADS v2 category and by radiologists' experience in use of PI-RADS v2 for interpretation, stratified into levels for each radiologist according to the number of included MRI examinations interpreted after the study start date: 1st through 50th, 51st through 100th, 101st through 151st, and 151st through 200th. For example, six PPVs were calculated for a radiologist who interpreted 100 examinations: three PPVs for each PI-RADS v2 category (3, 4, and 5) in the 1–50 experience level and three PPVs for each PI-RADS v2 category in the 51–100 experience level. The number of radiologists who interpreted more than 200 MRI examinations was insufficient to permit analysis for an experience level beyond 200 examinations. The number of examinations (i.e., 50 examinations) within the experience levels was determined a priori to ensure stable point estimates. Only complete sets of 50 examinations were analyzed. For example, if a radiologist interpreted 175 MRI examinations, only the first 150 MRI examinations were analyzed (stratified into experience levels as 1–50, 51–100, and 101–150), and the remaining 25 examinations interpreted by that radiologist were excluded. The distribution of PPVs for each radiologist, stratified by PI-RADS v2 category and experience level, was depicted visually using box-and-whisker plots. Overlap in the interquartile ranges (IQRs) of PPVs between PI-RADS categories at a given experience level was interpreted as evidence of a lack of precision in radiologists' PPV. Hierarchic linear mixed models were used to test for experience-related changes in PPV. The linear mixed

TABLE 2: Frequencies of PI-RADS Category Assignments and of PPV by PI-RADS Categories, Stratified by Experience Level

Measure	No. of MRI Examinations Read			
	1–50	51–100	101–150	151–200
No. of radiologists	8	8	5	5
PI-RADS category 3				
Frequency	31 (122/400)	24 (97/400)	28 (69/250)	21 (52/250)
Aggregate PPV	22 (27/122)	15 (15/97)	25 (17/69)	17 (9/52)
Rater-specific PPV	23 [17–26]	18 [11–22]	24 [20–29]	15 [0–33]
Absolute change in rater PPV	—	7 [5–18]	10 [3–13]	9 [5–10]
PI-RADS category 4				
Frequency	48 (192/400)	50 (198/400)	49 (123/250)	56 (139/250)
Aggregate PPV	49 (95/192)	43 (86/198)	54 (66/123)	51 (71/139)
Rater-specific PPV	51 [42–60]	43 [35–51]	55 [54–63]	53 [43–56]
Absolute change in rater PPV	—	14 [10–15]	25 [17–26]	10 [9–15]
PI-RADS category 5				
Frequency	22 (86/400)	26 (105/400)	23 (58/250)	24 (59/250)
Aggregate PPV	67 (58/86)	68 (71/105)	78 (45/58)	78 (46/59)
Rater-specific PPV	70 [54–75]	66 [57–76]	81 [77–82]	81 [70–86]
Absolute change in rater PPV	—	16 [5–25]	14 [11–17]	4 [3–11]

Note—Unless otherwise indicated, values represent percentage with numerator and denominator in parentheses (when applicable) and interquartile range in brackets (when applicable). Dash (—) indicates no comparison can be performed because there are no preceding cases.

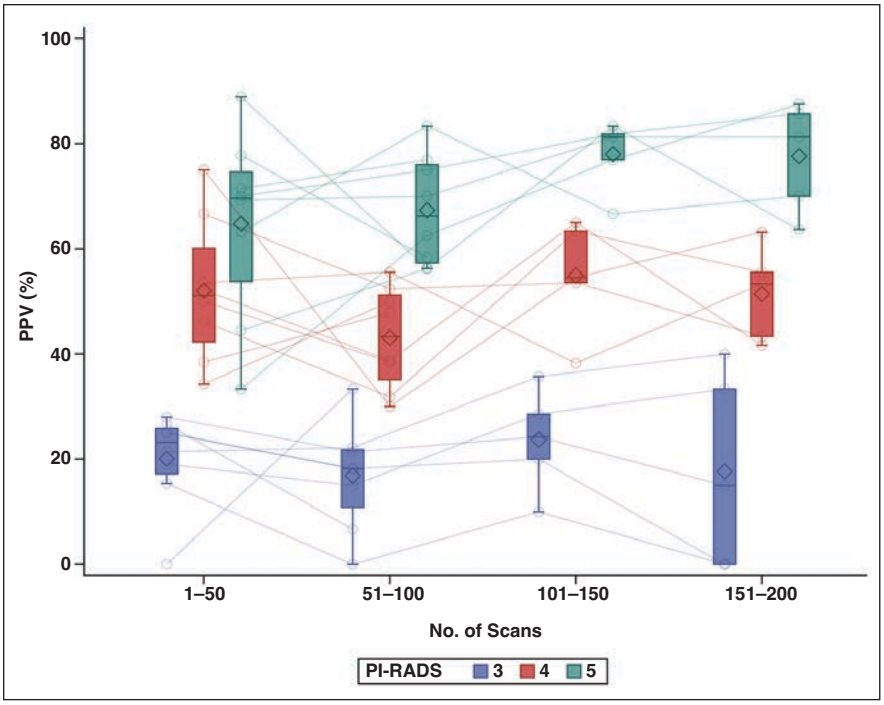

Fig. 3—Box-and-whisker plots of radiologist-level PPVs stratified by PI-RADS category and experience level. Centerlines within boxes indicate medians, diamonds within boxes indicate means, edges of boxes indicate interquartile ranges, and whiskers indicate interdecile ranges. Linear connectors between boxes indicate changes between experience levels in individual radiologists' PPVs for given PI-RADS categories.

effects models evaluated for absolute changes in radiologist-level PPV for each PI-RADS v2 category between consecutive experience levels (e.g., 51–100 vs 1–50), with random effects terms to account for clustering of results for individual radiologists. Logistic regression analysis was also used to assess for a change over time in the frequency of Gleason sum score of 7 or greater prostate cancer. Statistical analysis was performed using SAS 9.4 TS1M7 (SAS).

Results

Distributions of PI-RADS v2 Categories and Frequencies of Higher-Risk Cancer

The 1300 MRI examinations were interpreted by eight radiologists with a median of 13 years of postfellowship experience (range, 5–22 years) (Table 1). Five radiologists interpreted 200 included examinations (mean of 11 years of postfellowship experience at study initiation), and three radiologists interpreted 100 included examinations (mean of 13 years of postfellowship experience at study initiation); no radiologist interpreted 150 included examinations. The radiologists assigned PI-RADS category 3 in 26% (340/1300), category 4 in 50% (652/1300), and category 5 in 24% (308/1300) of examinations. Qualitatively, the eight radiologists showed no systematic change in the distribution of assigned PI-RADS v2 categories with increasing experience level (Fig. 2). Subsequent prostate biopsy showed Gleason sum score of 7 or greater prostate cancer in 47% (606/1300) of examinations. This frequency did not change significantly over time (p = .50): 56% (97/174) in 2015; 41% (185/455) in 2016; 46% (59/128) in 2017; 46% (107/235) in 2018; 54% (115/212) in 2019; 46% (40/87) in 2020; and 33% (3/9) in 2021.

PPV Stratified by PI-RADS v2 Categories and Experience Levels

Aggregate PPVs (across all radiologists and all experience levels) for maximum PI-RADS v2 categories were 20% (68/340) for PI-RADS category 3, 49% (318/652) for PI-RADS category 4, and 71% (220/308) for PI-RADS category 5 (Table 1). Radiologist-level PPVs for PI-RADS v2 categories stratified by experience level are summarized by medians and IQRs in Table 2 and depicted visually by box-and-whisker plots in Figure 3. The PPVs showed overlapping IQRs for PI-RADS category 4 (51% [midpoint of 10/20 and 12/13]; IQR, 42–60%) and category 5 (70% [midpoint of 9/13 and 7/10]; IQR, 54–75%) for radiologists' first 50 examinations interpreted. However, the IQRs of the PPVs did not overlap between categories for greater experience levels. The greatest separation in PPVs between PI-RADS categories was observed for the 101–150 examination experience level (PI-RADS category 3, 24% [8/33]; IQR, 20–29%; PI-RADS category 4, 55% [6/11]; IQR, 54–63%; PI-RADS category 5, 81% [13/16]; IQR, 77–82%). Hierarchic linear modeling identified no statistically significant association between PPV and experience level (PI-RADS category 3, p = .27; PI-RADS category 4, p = .71; PI-RADS category 5, p = .38) (Fig. 4).

Discussion

The effect of prostate MRI-specific experience on radiologist performance is important to understand because it informs the necessary training required to establish proficiency. In our analysis of eight radiologists who assigned a PI-RADS v2 category of 3, 4, or 5 to a total of 1300 MRI examinations after a targeted biopsy, individual radiologists' absolute PPVs for specific PI-RADS categories did not change during their first 200 MRI examinations interpreted. This finding supports radiologists' use of PI-RADS v2 when beginning to interpret prostate MRI for tumor detection. However, the precision of PPVs appeared to improve with increasing experience. Specifically, overlap was observed for the IQRs of PPVs for PI-RADS categories 4 and 5 for the first 50 examinations, but it was resolved with additional experience.

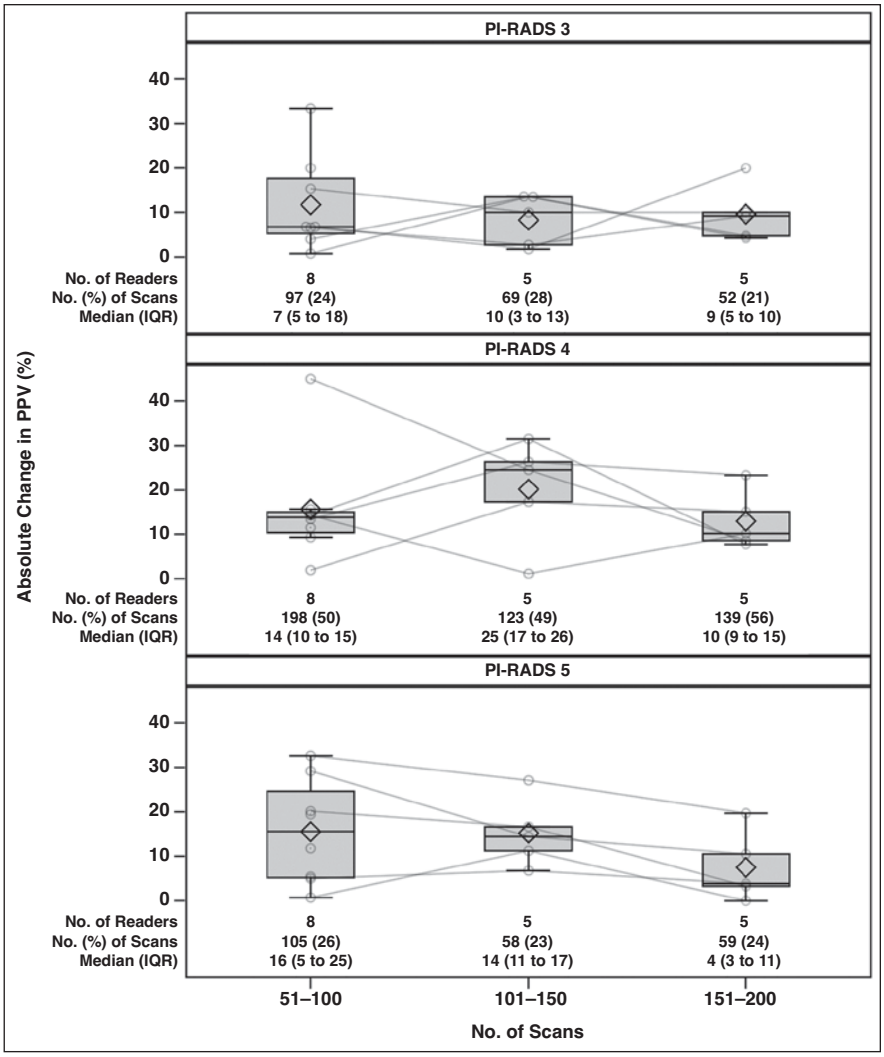

Fig. 4—Box-and-whisker plots of absolute changes in radiologist-level PPVs stratified by PI-RADS category and experience level, according to comparisons with immediately preceding experience level. Centerlines within boxes indicate medians, diamonds within boxes indicate means, edges of boxes indicate interquartile ranges (IQRs), and whiskers indicate interdecile ranges. Linear connectors between boxes indicate changes between experience levels in individual radiologists' PPVs for given PI-RADS categories.

At the start of the study period in 2015, none of the radiologists had prior experience with PI-RADS (v1 or v2). Therefore, our data represent a relatively pure sample from the perspective of training assessment. The aggregate PPVs by PI-RADS v2 category (category 3, 20%; category 4, 49%; category 5, 71%) are similar to those reported in earlier studies, indicating the inclusion of a representative sample [3–7]. For example, Westphalen et al. [7] analyzed lesion-level PI-RADS v2 categories at 26 centers for 3449 men and found PPVs of 15% (221/1462) for PI-RADS category 3, 38% (789/2071) for PI-RADS category 4, and 72% (635/887) for PI-RADS category 5. In addition, evidence was lacking in our results of a systematic group-level upgrading or downgrading in the distribution of assigned categories with increasing experience levels. Therefore, any such changes in distribution instead likely resulted from individual radiologist reporting adjustments with increasing experience. Moreover, the aggregate PPVs were stable over time, thus failing to provide evidence of longitudinal performance bias (e.g., from the COVID-19 pandemic or from a change in urologists performing the biopsies). Finally, as the radiologists and urologists gained experience in parallel, the improved precision over time in the observed PPVs may reflect a combination of gains in experience by both groups.

Others have evaluated the effect of increasing experience on performance in prostate MRI interpretation. For example, Rosenkrantz et al. [8] tested the performance of three second-year radiology residents in interpreting 60 prostate MRI examinations at two time points (before and after completion of an online educational training module). They observed improved sensitivity, NPV, and reader confidence after education. In comparison, we evaluated the performance among fellowship-trained radiologists in clinical interpretations rendered over the course of approximately 6 years. In a separate study, Rosenkrantz et al. [9] compared the diagnostic performance of six second-year radiology residents on 124 prostate MRI examinations, three with and three without case-by-case directed feedback. They observed a period of rapid performance improvement for the first 40 cases followed by diminishing gains with additional cases. Moreover, the early improvement occurred in both groups, suggesting that performance gains may largely reflect self-directed learning rather than external feedback. The rapid performance improvement for the trainees' first 40 cases is consistent with our findings of improved precision during

the first 50 cases for faculty radiologists. Finally, Curci et al. [4] likewise observed improved diagnostic performance over time for four faculty radiologists who interpreted a total of 249 prostate MRI examinations in the context of a 12-month long-distance longitudinal quality improvement program.

Our study has limitations. We evaluated only PPV as a measure of diagnostic performance. NPV could not be evaluated in our study sample given the inclusion of only positive examinations (i.e., those with a PI-RADS category of 3, 4, or 5). NPV is challenging to reliably evaluate because patients with negative MRI examinations may defer biopsy in the absence of elevated clinical risk and because of limitations of possible reference standards for negative results. Also, although the eight radiologists had no prior experience in the use of PI-RADS v1 or PI-RADS v2, they were from a single center with a small amount of earlier institutional experience in the use of prostate MRI for local staging. The performance may have differed among radiologists from other sites with different backgrounds. In addition, we did not stratify results by the clinical indication for prostate MRI. Another limitation is the 2-year window permitted between MRI and biopsy. This window was used to maximize the number of eligible examinations in consideration of prostate cancer's indolent nature. Finally, PPV was determined at the whole-gland level rather than the lesion level, reflecting a clinically relevant approach in patients who are not being evaluated for focal therapy.

In summary, individual radiologists' absolute PPVs at specific PI-RADS categories did not change during the radiologists' first 200 examinations interpreted (for examinations assigned a PI-RADS v2 category of 3–5 and that underwent subsequent targeted biopsy). However, overlap between IQRs of PPVs for different PI-RADS categories resolved after the initial 50 examinations, indicating that experience may improve the precision of PPV. Radiology practices contemplating a minimum training threshold for fellowship-trained abdominal radiologists might therefore consider 50 prostate MRI examinations to be sufficient in the context of a quality assurance program with feedback. Future studies could assess the effect of automated quality assurance histopathology feedback systems on diagnostic performance.

References

1. Weinreb JC, Barentsz JO, Choyke PL, et al. PI-RADS Prostate Imaging–Reporting and Data System: 2015, version 2. *Eur Urol* 2016; 69:16–40
2. Shankar PR, Davenport MS, Helvie MA. Prostate MRI and quality: lessons learned from breast imaging rad-path correlation. *Abdom Radiol (NY)* 2020; 45:4028–4030
3. Davenport MS, Downs E, George AK, et al. Prostate imaging and data reporting system version 2 as a radiology performance metric: an analysis of 18 abdominal radiologists. *J Am Coll Radiol* 2021; 18:1069–1076
4. Curci NE, Gartland P, Shankar PR, et al. Long-distance longitudinal prostate MRI quality assurance: from startup to 12 months. *Abdom Radiol (NY)* 2018; 43:2505–2512
5. Mazzone E, Stabile A, Pellegrino F, et al. Positive predictive value of Prostate Imaging Reporting and Data System version 2 for the detection of clinically significant prostate cancer: a systematic review and meta-analysis. *Eur Urol Oncol* 2021; 4:697–713
6. Barkovich EJ, Shankar PR, Westphalen AC. A systematic review of the existing Prostate Imaging Reporting and Data System version 2 (PI-RADSv2) literature and subset meta-analysis of PI-RADSv2 categories stratified by Gleason scores. *AJR* 2019; 212:847–854
7. Westphalen AC, McCulloch CE, Anaokar JM, et al. Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: experience of the Society of Abdominal Radiology Prostate Cancer Disease-focused Panel. *Radiology* 2020; 296:76–84
8. Rosenkrantz AB, Begovic J, Pires A, Won E, Taneja SS, Babb JS. Online interactive case-based instruction in prostate magnetic resonance imaging interpretation using Prostate Imaging and Reporting Data System version 2: effect for novice readers. *Curr Probl Diagn Radiol* 2019; 48:132–141
9. Rosenkrantz AB, Ayoola A, Hoffman D, et al. The learning curve in prostate MRI interpretation: self-directed learning versus continual reader feedback. *AJR* 2017; 208:[web]W92–W100

(Editorial Comment starts on next page)

Editorial Comment: What Is Prostate MRI Interpretation Experience?

PI-RADS provides an important framework to standardize interpretation of prostate MRI for clinically significant prostate cancer. However, the PPV of PI-RADS version 2 (v2) at category thresholds of 3 or 4 varies widely among institutions, which is problematic in the context of a standardized reporting system [1]. This article addresses one of many factors that could affect PI-RADS performance: radiologist experience.

In this study, PI-RADS v2 had been newly implemented at a large academic practice. The PPVs obtained for PI-RADS categories 3–5 were on par with the overall performance in an earlier multicenter study [1]. Although precision of the PPVs improved after the first 50 cases interpreted, the actual PPVs were not significantly different between the start and end of the study period (after 200 cases interpreted). The results effectively show consistent performance of PI-RADS throughout the period, including at the outset.

Radiologists' experience during the study period comprised case interpretation in the setting of an institutional culture rich with multiple opportunities for histopathologic feedback including peer learning, multidisciplinary conferences, formal instruction, and data from quality assurance databases. The unique institutional support around PI-RADS that the radiologists experienced may be difficult to replicate. It is unclear whether one, several, or all components of the institutionally mediated feedback helped in achieving the reported PPVs. On the other hand, it is possible that none of the components had an impact, as another study found that self-directed learning without histologic feedback can improve prostate MRI interpretation [2]. The academic radiologists in the current study performed well from the start, but for practices encountering PPVs below the median, what is the key to improvement? For novice readers of prostate MRI without the opportunity to gain experience in an academic environment with plentiful resources for feedback and improvement, what is the next best accessible experience?

Motoyo Yano, MD, PhD
Mayo Clinic Arizona
Scottsdale, AZ
yano.motoyo@mayo.edu

The author declares that there are no disclosures relevant to the subject matter of this article.

doi.org/10.2214/AJR.22.27781

References

1. Westphalen AC, McCulloch CE, Anaokar JM, et al. Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: experience of the Society of Abdominal Radiology Prostate Cancer Disease-focused Panel. *Radiology* 2020; 296:76–84
2. Rosenkrantz AB, Ayoola A, Hoffman D, et al. The learning curve in prostate MRI interpretation: self-directed learning versus continual reader feedback. *AJR* 2017; 208:[web]W92–W100