Genitourinary Imaging • Original Research

Multicenter Multireader Evaluation of an Artificial Intelligence–Based Attention Mapping System for the Detection of Prostate Cancer With Multiparametric MRI

OBJECTIVE. The purpose of this study was to evaluate in a multicenter dataset the performance of an artificial intelligence (AI) detection system with attention mapping compared with multiparametric MRI (mpMRI) interpretation in the detection of prostate cancer.

MATERIALS AND METHODS. MRI examinations from five institutions were included in this study and were evaluated by nine readers. In the first round, readers evaluated mpMRI studies using the Prostate Imaging Reporting and Data System version 2. After 4 weeks, images were again presented to readers along with the AI-based detection system output. Readers accepted or rejected lesions within four AI-generated attention map boxes. Additional lesions outside of boxes were excluded from detection and categorization. The performances of readers using the mpMRI-only and AI-assisted approaches were compared.

RESULTS. The study population included 152 case patients and 84 control patients with 274 pathologically proven cancer lesions. The lesion-based AUC was 74.9% for MRI and 77.5% for AI with no significant difference (p = 0.095). The sensitivity for overall detection of cancer lesions was higher for AI than for mpMRI but did not reach statistical significance (57.4% vs 53.6%, p = 0.073). However, for transition zone lesions, sensitivity was higher for AI than for MRI (61.8% vs 50.8%, p = 0.001). Reading time was longer for AI than for MRI (4.66 vs 4.03 minutes, p < 0.001). There was moderate interreader agreement for AI and MRI with no significant difference (58.7% vs 58.5%, p = 0.966).

CONCLUSION. Overall sensitivity was only minimally improved by use of the AI system. Significant improvement was achieved, however, in the detection of transition zone lesions with use of the AI system at the cost of a mean of 40 seconds of additional reading time.

Prostate cancer is the most common noncutaneous cancer type among men [1]. Unlike most other cancers, prostate cancer is difficult to detect with conventional imaging techniques such as ultrasound and CT. Therefore, until recently, imaging has not been accepted as standard-of-care practice for prostate cancer detection. Over the last 2 decades, major advances in prostate MRI have led to considerable improvements in prostate cancer detection. Although initially the use of MRI was limited [2], with the development of higher magnetic field strengths, higher quality of imaging, and the combined use of anatomic and functional MRI sequences, multiparametric MRI (mpMRI) has emerged as an important method of detecting prostate cancer [3]. Reports in the current literature, however, indicate that 5–30% of prostate cancers are missed at mpMRI [4–6]. The causes of such misses may be related to the complex nature of the prostate tissues and the limited spatial resolution of MRI.

One proposed solution is to use artificial intelligence (AI)-based detection systems. With the help of machine learning, classification algorithms can be trained to predict results and outcomes, provided that enough training data are available. In 2017, we at the National Cancer Institute [7] proposed an AI system based on intensity and texture analysis and a random forest classification algorithm. This system was validated in a large multireader multicenter study in 2018 [8]. Results of that study revealed an increase in detection of transition zone lesions among moderately experienced readers only. Overall, however, the AI system was equivalent to conventional MRI interpretation [8]. In that study, color-coded prediction maps were used to draw attention to AI-detected lesions. Feedback from the study suggested that prediction maps compromised the interaction between the radiologists and the AI system with resultant decreased accuracy for some readers. To address this issue a new AI detection system with more expert annotated...
Keywords: artificial intelligence, laparoscopic, MRI, multiparametric, prostate cancer, radical prostatectomy, robot-assisted

doi.org/10.2214/AJR.19.22573

Received November 5, 2019; accepted after revision February 4, 2020.

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.

Supported in whole or in part by federal funds from the National Cancer Institute, National Institutes of Health (contract HHSN26120080001E).

1Department of Urology and Pediatric Urology, University Medical Center, Mainz, Germany.
2Urologic Oncology Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD.
3Molecular Imaging Program, National Cancer Institute, National Institutes of Health, 10 Center Dr, MSC 1182, Bldg 10, Rm B3B85, Bethesda, MD 20892-1088. Address correspondence to B. Turkbey (turkbeyi@mail.nih.gov).
4Clinical Research Directorate, Leidos Biomedical Research, Inc., Frederick, MD.
5Division of Cancer Treatment and Diagnosis: Biometric Research Program, National Cancer Institute, National Institutes of Health, Rockville, MD.
6Department of Urology, Acibadem University, Istanbul, Turkey.
7Department of Radiology, University of Udine, Udine, Italy.
8Diagnostic Imaging Department, Albert Einstein Hospital, Sao Paolo, Brazil.
9Department of Urology, Koç University, School of Medicine, Istanbul, Turkey.
10Department of Radiology, Ankara City Hospital, Ankara, Turkey.
11Department of Radiology, Acibadem University, Istanbul, Turkey.
12Department of Radiology, Cleveland Clinic, Cleveland, OH.
13Department of Radiology, University of Alabama at Birmingham, Birmingham, AL.
14Department of Radiology, University of Alabama at Birmingham, Birmingham, AL.
15O’Neal Comprehensive Cancer Center at UAB, University of Alabama at Birmingham, Birmingham, AL.
16Department of Pathology, Cleveland Clinic, Cleveland, OH.
17Department of Pathology, University of Alabama at Birmingham, Birmingham, AL.
18Present address: Department of Pathology, Vanderbilt University, Nashville, TN.
19Pathology Department, Albert Einstein Hospital, Sao Paolo, Brazil.
20Department of Pathology, National Cancer Institute, National Institutes of Health, Bethesda, MD.
21Department of Pathology, Ankara Yildirim Beyazit University, School of Medicine, Ankara, Turkey.
22Department of Pathology, Acibadem University, Istanbul, Turkey.
23Department of Pathology, University of Udine, Udine, Italy.
24Department of Pathology, University of Cambridge, Cambridge, UK.
25Department of Radiology, University of Cambridge, Cambridge, UK.
26Department of Radiology, Federal Fluminense University, Rio de Janeiro, Brazil.
27DASA Company, Rio de Janeiro, Brazil.
28Department of Radiology, University of Health Sciences Dr. Behget Uz Child Disease and Pediatric Surgery Training and Research Hospital, İzmir, Turkey.
29Department of Radiology, Walter Reed Medical Center, Bethesda, MD.
30Department of Radiology, Singapore General Hospital, Singapore.
31Department of Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Bethesda, MD.
32Weill Cornell Imaging, Cornell University, New York, NY.
33Department of Radiology, Medical Imaging Centre, Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
34Center for Interventional Oncology, National Cancer Institute and Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Bethesda, MD.
35National Institutes of Health Clinical Center, Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, Bethesda, MD.

Mehralivand et al.

Patient Population

This HIPAA-compliant evaluation of multi-institutional data was approved by the National Cancer Institute ethics committee. Inclusion of anonymized data from the other institutions was approved in concordance with National Institutes of Health Office of Human Subjects Resources protocol 11617. Local ethics approvals to share data were obtained as needed. Patients from five institutions were included in this study. Those in the case group underwent mpMRI and had subsequent prostate biopsy results positive for adenocarcinoma and then underwent radical prostatectomy. Final histopathologic results for the prostatectomy specimens with lesion mapping were available for all case population patients. All participants in the control group underwent mpMRI with no visible lesions detected. Prostate cancer was ruled out by means of 12-core transrectal systematic biopsy. Three patients with missing final radical prostatectomy histopathologic lesion maps were excluded. The final study population included 152 case and 84 control subjects. The distribution of case and control subjects among institutions is summarized in Table 1.

Reader Profiles

To prevent bias, the case and control MRI data were obtained from five different institutions but were interpreted by nine readers from independent institutions. Readers ranged in experience in interpreting prostate mpMRI and were stratified into three levels of experience: low, moderate, and high. Determination was based on years of experience and number of prostate MRI studies read per year according to the following criteria: a low level of experience was less than 1 year or fewer than 100 examinations per year; moderate, 1–3 years or 100–300 examinations per year; high, more than 3 years or more than 300 examinations per year.
Artificial Intelligence in MRI of Prostate Cancer

Study Design and Statistical Powering

Because it was unrealistic for every reader to read every case, each patient imaging examination was randomly assigned to three different readers to ensure balanced and unbiased distribution. Our primary hypothesis was that AI-assisted mpMRI can achieve higher lesion-based sensitivity than mpMRI without the benefit of AI. To test this hypothesis, nine readers read assigned cases according to the balanced incomplete block design, in which each randomly selected case stratified by patient disease status was assigned to each triple-wise combination of readers [9]. Among the 236 patients (152 with cancer, 84 without cancer), each reader evaluated a mean of 78 patients (range, 75–81). The primary endpoint was the difference in mean lesion-level reader-specific sensitivity between AI-assisted detection and mpMRI alone. In the power analysis, sensitivity was set at 57% and the SD of the primary endpoint was estimated from previous studies [8, 10]. This study had 93% power to detect an 8% difference in sensitivity using the Z test at the two-sided 5% significance level.

MRI Acquisition Technique

All imaging examinations were performed at 3 T without an endorectal coil with equipment from a variety of vendors. Axial T2-weighted imaging, DWI with at least two b values, and dynamic contrast-enhanced imaging sequences were performed. Because DWI with a b value of 1500 mm/s² was necessary for the AI processing, in cases in which this acquisition was not available, the high b value was calculated by use of a monoexponential decay model. Image acquisition protocols were compliant with Prostate Imaging Reporting and Data System version 2 (PI-RADSv2) technical recommendations.

Patient Data Deidentification

In compliance with U.S. Office of Human Subjects Resources guidelines, all medical images were fully anonymized by provider centers before submission to our center (National Cancer Institute). For this purpose, standard scripts were used that remove all DICOM tags except for technical image acquisition–related information. When the data were received, a second deidentification was performed to ensure patient confidentiality at the highest standard.

Artificial Intelligence System

The AI system was based on a custom multi-task random forest similar to the Hough forest and the regression forest [11]. Each random tree was trained on 3-T MR images of 161 patients from five different institutions; an alternating information gain function was used that was either defined to optimize for classification accuracy or minimize the 𝐢ₙ residual of predicted bounding box extents. The training population was patients and institutions different from the test population in this study. This learning system entailed a combination of patch-based intensity and Haralick texture features and operated only on pre-segmented transition and peripheral zones.

Automatic segmentations were performed on T2-weighted images and manually corrected by an expert radiologist. Each zone had its own specialized AI model. The result was a collection of 10 random trees per zone that each evaluated an image patch from T2-weighted images, apparent diffusion coefficient maps, and b1500 images and predicted both the probability of clinically significant cancer and the bounding box width and height of the lesion. The predictions from all trees were averaged, and box candidates were postprocessed with nonmaximum suppression to choose the final predicted boxes (up to four were kept). The result was a probability map for clinically significant cancer (Gleason score > 3 + 3) and a collection of attention map boxes for suspicious lesions, the latter of which were used in this study. The attention map boxes were picked by use of a threshold corresponding to a 67% tumor detection rate at 2.71 false-positive results per patient.

The genitourinary pathologist at each provider center was blinded to the mpMRI results. For each case patient’s specimen, cancer lesions were mapped, and a corresponding Gleason score according to the International Society of Urological Pathology 2014 consensus guidelines was assigned [14].
Statistical Analysis

Patient-based sensitivity and specificity were calculated at each PI-RADSv2 threshold, and AUC was estimated for each reader; the maximum PI-RADSv2 category assigned by a given reader represented each patient’s outcome. The comparison between AI and mpMRI was made at PI-RADSv2 ≥ 1 or greater representing all detected lesions and PI-RADSv2 ≥ 3 or greater representing suspicious lesions. For lesion-based analysis, reader sensitivity and free-response ROC analysis was performed [15]. Reader statistics were averaged across all readers and by experience level. Reader agreement on lesion detection in the same location was assessed by the index of specific agreement [16]. Statistical inference was obtained by the bootstrap resampling procedure with 2000 bootstrap samples drawn at the patient level. The 95% confidence limits were the 2.5% and 97.5% percentiles of the bootstrap resampling distribution. All test statistics were based on two-sided Wald test and bootstrap standard error. Values of \( p < 0.05 \) were considered statistically significant.

Results

Study Population and Lesion Characteristics

The final study population consisted of 152 case and 84 control subjects. Except in 38 case patients, the final histopathologic result was grade group 2 or higher. There were 274 pathologically proven cancer lesions with 188 of the 274 in the peripheral zone, 77 in the transition zone, and nine spanning both zones. Among all 274 lesions, 38 were assigned grade group 1, 130 group 2, 45 group 4, and 15 group 5 at final histopathologic analysis. The final study population consisted of 152 case and 84 control subjects. Except in 38 case patients, the final histopathologic result was grade group 2 or higher. There were 274 pathologically proven cancer lesions with 188 of the 274 in the peripheral zone, 77 in the transition zone, and nine spanning both zones. Among all 274 lesions, 38 were assigned grade group 1, 130 group 2, 45 group 4, and 15 group 5 at final histopathologic analysis.

Multiparametric MRI and Artificial Intelligence Performance at Patient Level

The overall AUCs were 81.6% for MRI and 78% for AI (\( p = 0.053 \)). Readers with a low experience level had AUCs of 80.9% for MRI and 73.3% for AI (\( p = 0.018 \)); moderate experience, 80% for MRI and 76.9% for AI (\( p = 0.28 \)); and a high level of experience, 83.8% for MRI and 83.6% for AI (\( p = 0.95 \)).

Sensitivity and specificity plotted against PI-RADSv2 thresholds are shown in Figure 2. For the detection of all ground truth lesions (threshold, PI-RADSv2 ≥ 1), no significant difference in sensitivity was observed between MRI and AI (89.6% vs 87.9%, \( p = 0.364 \)). However, in the subgroup of experienced readers, sensitivity of AI was significantly greater than that of MRI (95.5% vs 89.0%, \( p = 0.013 \)). For lesions considered suspicious with MRI (threshold, PI-RADSv2 ≥ 3) no significant differences in sensitivity were observed in the whole group (81.7% vs 83.5%, \( p = 0.453 \)). This finding held true for all subgroups of reader experience.

For the detection of all ground truth lesions (threshold, PI-RADSv2 ≥ 1), specificity was significantly lower for AI (30.0% vs 51.5%, \( p < 0.001 \)) in the whole group. This finding held true for all subgroups of reader experience. For lesions considered suspicious on MRI studies (threshold, PI-RADSv2 ≥ 3), specificity was significantly lower for AI in the whole group (51.4% vs 60.7%, \( p = 0.01 \)). Although it was observed in all subgroups of reader experience, this result did not reach statistical significance in the groups of readers with moderate and high experience levels.

Multiparametric MRI and Artificial Intelligence Performance at Lesion Level

The free-response ROC AUCs were 74.9% for MRI and 77.5% for AI (\( p = 0.095 \)). Readers with a low experience level had AUCs of 76.6% for MRI and 78.4% for AI (\( p = 0.095 \)); moderate experience, 78.1% for MRI and 78.6% for AI (\( p = 0.747 \)); and a high experience level, 79.3% for MRI and 81.1% for AI (\( p = 0.003 \)).

Lesion-level sensitivity and positive predictive value for all PI-RADSv2 thresholds are shown in Table 2 for MRI and AI. Sensitivity plotted against PI-RADSv2 thresholds

<table>
<thead>
<tr>
<th>TABLE 2: Lesion-Level Diagnostic Performance Statistics on Artificial Intelligence (AI) and MRI for PI-RADS Category Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI-RADS Threshold for Performance Metric Evaluation</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>PI-RADS ≥ 1</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
</tr>
<tr>
<td>PPV (%)</td>
</tr>
<tr>
<td>PI-RADS ≥ 2</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
</tr>
<tr>
<td>PPV (%)</td>
</tr>
<tr>
<td>PI-RADS ≥ 3</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
</tr>
<tr>
<td>PPV (%)</td>
</tr>
<tr>
<td>PI-RADS ≥ 4</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
</tr>
<tr>
<td>PPV (%)</td>
</tr>
<tr>
<td>PI-RADS ≥ 5</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
</tr>
<tr>
<td>PPV (%)</td>
</tr>
</tbody>
</table>

Note—Values in parentheses are 95% CI. PI-RADS = Prostate Imaging Reporting and Data System, PPV = positive predictive value.

\(^a p < 0.01\).
for the whole prostate, peripheral zone, and transition zone are shown in Figure 3. For the detection of all ground truth lesions (threshold, PI-RADSv2 ≥ 1), sensitivity was higher for AI than for MRI but did not reach a statistical significance (57.4% vs 53.6%, p = 0.073). However, in the subgroup of highly experienced readers, AI had significantly greater sensitivity than MRI did (62.7% vs 54.1%, p = 0.002). There were no statistically significant differences between AI and MRI in the subgroups of readers with low and intermediate experience levels. For lesions considered suspicious on MRI studies (threshold, PI-RADSv2 ≥ 3), there was no significant difference in sensitivity between AI and MRI (53% vs 49.4%, p = 0.238). This was also true among all subgroups of reader experience. The mean numbers of region proposals not corresponding to reportable findings according to PI-RADSv2 guidelines were 2.53 (range, 0–4) for control patients and 1.76 (range, 0–4) for case patients.

**Interreader Agreement**

There was moderate interreader agreement in the whole group for AI and MRI with no statistically significant difference (58.7% vs 58.5%, p = 0.966). This was also true of readers with low (55.2% vs 48.5%, p = 0.403) and moderate (55.7% vs 55.5%, p = 0.993) experience levels. Among highly experienced readers, interreader agreement was substantial for AI and MRI with no statistically significant difference (0.644 vs 0.645, p = 0.959).

**Image Interpretation Times**

The overall reading time was significantly longer for AI than for MRI (4.66 vs 4.03 minutes, p < 0.001). This was particularly pronounced among moderately (5.41 vs 4.68 minutes, p = 0.001) and highly (4.22 vs 3.33 minutes, p < 0.001) experienced readers, but there was no statistically significant difference among readers with a low level of experience (4.33 vs 4.1 minutes, p = 0.289).

**Discussion**

In this study an AI detection system entailing region-based attention mapping showed little or no improvement over conventional interpretation of prostate MRI across multiple readers of various experience levels. The notable exception was overall improvement in the detection of transition zone lesions, which are more difficult to diagnose. This result agrees with those of other AI studies, which to date have not shown dramatic improvements in performance over conventional interpretation. Interestingly, the AI system did not improve interreader variability or improve the performance of readers with a low level of experience, which are features commonly touted for AI systems.

This study revealed some interesting findings in subset analysis. For instance, among highly experienced readers, sensitivity was significantly higher for AI than for MRI for all lesions; this difference was not observed

**TABLE 3: Diagnosis Performance Statistics on Artificial Intelligence (AI) and MRI for PI-RADS Category Thresholds:**

<table>
<thead>
<tr>
<th>PI-RADS Threshold for Performance Metric Evaluation</th>
<th>Overall</th>
<th>Reader Experience Level</th>
<th>Overall</th>
<th>Reader Experience Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AI</td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>PI-RADS ≥ 1</td>
<td>Sensitivity (%)</td>
<td>56.8(51.3–63.1)</td>
<td>54.5(46.5–63.0)</td>
<td>56.6(48.7–63.0)</td>
</tr>
<tr>
<td></td>
<td>PPV (%)</td>
<td>52.1(47.9–56.5)</td>
<td>54.1(47.5–61.4)</td>
<td>47.7(41.0–54.2)</td>
</tr>
<tr>
<td>PI-RADS ≥ 2</td>
<td>Sensitivity (%)</td>
<td>56.8(51.3–63.1)</td>
<td>54.5(46.5–63.0)</td>
<td>55.6(48.7–63.0)</td>
</tr>
<tr>
<td></td>
<td>PPV (%)</td>
<td>52.5(48.2–56.9)</td>
<td>54.1(47.5–61.4)</td>
<td>48.3(41.6–54.8)</td>
</tr>
<tr>
<td>PI-RADS ≥ 3</td>
<td>Sensitivity (%)</td>
<td>50.1(44.2–57.1)</td>
<td>47.6(39.7–56.5)</td>
<td>50.4(43.2–58.4)</td>
</tr>
<tr>
<td></td>
<td>PPV (%)</td>
<td>60.4(55.9–65.3)</td>
<td>60.7(53.5–68.6)</td>
<td>54.1(46.7–61.3)</td>
</tr>
<tr>
<td>PI-RADS ≥ 4</td>
<td>Sensitivity (%)</td>
<td>44.0(38.4–50.7)</td>
<td>41.9(34.8–50.0)</td>
<td>44.1(38.9–52.2)</td>
</tr>
<tr>
<td></td>
<td>PPV (%)</td>
<td>67.5(62.8–72.5)</td>
<td>64.2(55.9–72.9)</td>
<td>59.5(51.5–67.5)</td>
</tr>
<tr>
<td>PI-RADS ≥ 5</td>
<td>Sensitivity (%)</td>
<td>17.8(13.4–22.8)</td>
<td>20.0(13.9–26.5)</td>
<td>18.0(12.0–25.0)</td>
</tr>
<tr>
<td></td>
<td>PPV (%)</td>
<td>88.6(82.4–94.1)</td>
<td>90.8(82.7–97.9)</td>
<td>83.1(72.4–93.3)</td>
</tr>
</tbody>
</table>

Note—Values in parentheses are 95% CI. PI-RADS = Prostate Imaging Reporting and Data System, PPV = positive predictive value.
TABLE 4: Diagnosis Performance Statistics on Artificial Intelligence and MRI for PI-RADS Category Thresholds: Transition Zone Lesions

<table>
<thead>
<tr>
<th>PI-RADS Threshold for Performance Metric Evaluation</th>
<th>AI</th>
<th>Reader Experience Level</th>
<th>MRI</th>
<th>Reader Experience Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>61.8 (53.9–70.1) 1</td>
<td>48.1 (34.9–61.9)</td>
<td>67.2 (54.2–80.4)</td>
<td>70.0 (60.8–80.1)</td>
</tr>
<tr>
<td>Moderate</td>
<td>34.4 (29.6–39.9)</td>
<td>39.9 (29.8–50.8)</td>
<td>28.5 (21.6–37.3)</td>
<td>34.7 (28.4–41.7)</td>
</tr>
<tr>
<td>High</td>
<td>61.8 (53.9–70.1) 1</td>
<td>48.1 (34.9–61.9)</td>
<td>67.2 (54.2–80.4)</td>
<td>70.0 (60.8–80.1)</td>
</tr>
<tr>
<td>PPV (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>46.2 (40.7–52.9)</td>
<td>41.3 (30.9–52.9)</td>
<td>42.1 (34.3–50.7)</td>
<td>55.2 (45.5–65.4)</td>
</tr>
<tr>
<td>Moderate</td>
<td>46.2 (40.7–52.9)</td>
<td>41.3 (30.9–52.9)</td>
<td>42.1 (34.3–50.7)</td>
<td>55.2 (45.5–65.4)</td>
</tr>
<tr>
<td>High</td>
<td>46.2 (40.7–52.9)</td>
<td>41.3 (30.9–52.9)</td>
<td>42.1 (34.3–50.7)</td>
<td>55.2 (45.5–65.4)</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>41.8 (33.8–51.1)</td>
<td>35.3 (25.5–46.4)</td>
<td>51.0 (38.4–65.5)</td>
<td>38.9 (28.2–50.8)</td>
</tr>
<tr>
<td>Moderate</td>
<td>64.9 (57.3–73.1)</td>
<td>51.7 (38.2–66.2)</td>
<td>66.5 (55.9–77.4)</td>
<td>76.5 (64.7–88.4)</td>
</tr>
<tr>
<td>High</td>
<td>64.9 (57.3–73.1)</td>
<td>51.7 (38.2–66.2)</td>
<td>66.5 (55.9–77.4)</td>
<td>76.5 (64.7–88.4)</td>
</tr>
<tr>
<td>PPV (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>23.9 (16.8–32.4)</td>
<td>26.4 (16.5–37.0)</td>
<td>24.2 (14.9–35.3)</td>
<td>21.1 (11.3–33.1)</td>
</tr>
<tr>
<td>Moderate</td>
<td>86.6 (79.3–94.4)</td>
<td>86.7 (76.2–100)</td>
<td>82.2 (63.9–97.0)</td>
<td>91.1 (81.5–100)</td>
</tr>
</tbody>
</table>

Note—Values in parentheses are 95% CI. AI = artificial intelligence, PI-RADS = Prostate Imaging Reporting and Data System, PPV = positive predictive value.

1p < 0.01.

for lesions deemed suspicious for prostate cancer with MRI. In other words, the AI system increased the sensitivity of highly experienced radiologists in detecting invisible (PI-RADS 1) and low-category (PI-RADS 2) lesions, whereas it was not as contributory for clearly visible lesions (PI-RADS ≥ 3).

One possible explanation for the more pronounced effect on highly experienced readers could be the different output format of our AI system. Although our previous and most other AI systems use color-coded cancer probability maps, we chose an attention-based mapping box to decrease distraction and subjectivity caused by background noise and false-positive lesions on AI maps and focus the reader’s attention on the areas of highest likelihood of cancer. It is possible that highly experienced readers benefitted most from this approach because they were more confident in detecting cancer-suspicous lesions and could spend additional time on the regions highlighted by the AI derived boxes, whereas less experienced readers might be less confident and not weight the AI data as strongly.

The AI system performed differently in different zones of the prostate. The peripheral and transition zones have very different radiologic and histopathologic properties. As a consequence, the PI-RADS consensus guidelines recommend different criteria for the assignment of risk categories in the peripheral and transition zones [12]. The transition zone usually has a heterogeneous signal-intensity pattern on T2-weighted images, especially in patients with benign prostatic hyperplasia. This complicates the detection of prostate cancer lesions, which therefore can be easily overlooked. As a result, the sensitivity of mpMRI in general is lower for transition zone than for peripheral zone lesions [17]. In the subgroup analysis of peripheral and transition zone lesions in our study, the AI system had significantly higher sensitivity for all MRI-detected transition zone lesions than did mpMRI. This may represent an important contribution of this detection system.

Current AI systems are not fully automated detection systems but rather adjunct tools to aid radiologists reading prostate mpMRI studies. Therefore, the AI information along with prostate mpMRI findings can be considered an additional parameter potentially increasing complexity for no benefit. Our study did not show significant improvement in interreader agreement among readers. Reading time was slightly longer (mean, 40 seconds) done because PI-RADS allows evaluation of up to four lesions and there was no method for varying the number of boxes for each case. This likely resulted in more false-positive readings, resulting in significantly lower specificity on the patient level compared with prior AI detection systems.

Second, the training of the algorithm was based on a fairly small patient population.
Artificial Intelligence in MRI of Prostate Cancer

This might have negatively affected the performance of the model because larger and more diverse patient populations improve generalizability of classification algorithms.

Third, our classification algorithm was based on a random forest classifier. With advances in computational resources, big data, and more sophisticated deep neural network algorithms, deep learning is gaining popularity in medicine. In particular, convolutional neural networks appear to be superior to classic machine learning techniques and other deep neural network architectures in performance and generalization in imaging-related tasks [19–21]. As a result, we are currently working on procuring and annotating larger imaging datasets and developing deep neural network architectures for designing a stronger prostate cancer AI detection system.

Conclusion
The AI detection system had significantly higher sensitivity than mpMRI for the detection of transition zone lesions, especially those not visible to readers using the raw images alone. Overall, there was no significant gain from the AI detection system compared with MRI alone, and it did not improve the performance of readers with a low experience level or reduce interreader variability. AI did, however, improve the performance of radiologists in evaluating transition zone lesions. These results suggest a need for further work on deep learning convolutional neural networks in larger datasets to improve the performance of radiologists interpreting mpMRI of the prostate.

References
3. Hamoen EGH, de Rooij M, Witjes JA, Barentsz JO, Rovers MM. Use of the Prostate Imaging Reporting and Data System (PI-RADS) for prostate cancer detection with multiparametric magnetic resonance imaging: a diagnostic meta-analysis. Eur Urol 2015; 67:1112–1121
16. Shih JH, Greer MD, Turkbey B. The problems with the kappa statistic as a metric of interobserver agreement on lesion detection using a third reader approach when locations are not specified. Acad Radiol 2018; 25:1325–1332
Fig. 1—55-year-old man with prostate-specific antigen level of 4.68 ng/mL and Prostate Imaging Reporting and Data System category 5 lesion in left anterior transition zone correctly detected by artificial intelligence system. Final histopathologic result was Gleason 3 + 4 prostate cancer.
A, T2-weighted MR image.
B, Apparent diffusion coefficient map.
C, DW image (b = 2000 mm/s²).
D, Dynamic contrast-enhanced MR image.
E, T2-weighted MR image with attention box produced by means of artificial intelligence.
F, Photomicrograph of radical prostatectomy specimen. I = index lesion.

Fig. 2—Graphs show sensitivity and specificity of artificial intelligence (AI) and MRI for different Prostate Imaging Reporting and Data System (PI-RADS) category thresholds at patient level. Asterisk denotes $p < 0.05$; double asterisk, $p < 0.01$.
A, Sensitivity for all readers.
B, Sensitivity for readers with low level of experience.
C, Sensitivity for readers with moderate level of experience.

(Fig. 2 continues on next page)
Artificial Intelligence in MRI of Prostate Cancer

Fig. 2 (continued)—Graphs show sensitivity and specificity of artificial intelligence (AI) and MRI for different Prostate Imaging Reporting and Data System (PI-RADS) category thresholds at patient level. Asterisk denotes \( p < 0.05 \); double asterisk, \( p < 0.01 \).

D, Sensitivity for readers with high level of experience.
E, Specificity for all readers.
F, Specificity for readers with low level of experience.
G, Specificity for readers with moderate level of experience.
H, Specificity for readers with high level of experience.

Fig. 3—Graphs show sensitivity of artificial intelligence (AI) and MRI for different Prostate Imaging Reporting and Data System (PI-RADS) category thresholds at lesion level for whole prostate, peripheral zone, and transition zone. Asterisk denotes \( p < 0.05 \); double asterisk, \( p < 0.01 \).

A, Whole prostate, all readers.
B, Whole prostate, readers with low level of experience.
C, Whole prostate, readers with moderate level of experience.

(Fig. 3 continues on next page)
Fig. 3 (continued)—Graphs show sensitivity of artificial intelligence (AI) and MRI for different Prostate Imaging Reporting and Data System (PI-RADS) category thresholds at lesion level for whole prostate, peripheral zone, and transition zone. Asterisk denotes \( p < 0.05 \); double asterisk, \( p < 0.01 \).

D, Whole prostate, readers with high level of experience.
E, Peripheral zone, all readers.
F, Peripheral zone, readers with low level of experience.
G, Peripheral zone, readers with moderate level of experience.
H, Peripheral zone, readers with high level of experience.
I, Transition zone, all readers.
J, Transition zone, readers with low level of experience.
K, Transition zone, readers with moderate level of experience.
L, Transition zone, readers with high level of experience.