Title

Performance of Deep Learning and Genitourinary Radiologists in Detection of Prostate Cancer Using 3-T Multiparametric Magnetic Resonance Imaging

Authors

Cao, Ruiming
Zhong, Xinran
Afshari, Sohrab
et al.

Publication Date

2021-03-12

DOI

10.1002/jmri.27595

Peer reviewed

Performance of Deep Learning and Genitourinary Radiologists in Detection of Prostate Cancer using 3 T Multi-parametric MRI

Ruiming Cao MS[1, *], Xinran Zhong PhD[2], Sohrab Afshari MD[3], Ely Felker MD[3], Voraparee Suvannarerg MD[3, 4], Teeravut Tubtawee MD[3, 5], Sitaram Vangala PhD[6], Fabien Scalzo PhD[7], Steven Raman MD[3], Kyunghyun Sung PhD[3]

[1] Department of Bioengineering, UC Berkeley
[2] Department of Radiation Oncology, UT Southwestern
[3] Department of Radiology, UCLA
[4] Department of Radiology, Faculty of Medicine, Siriraj Hospital, Mahidol University
[5] Department of Radiology, Faculty of Medicine, Prince of Songkla University
[6] Department of Medicine Statistics Core, UCLA
[7] Department of Neurology, UCLA
[*] Corresponding author: rcao@berkeley.edu

Grant support: This work was supported by National Institutes of Health R01-CA248506 and funds from the Integrated Diagnostics Program, Department of Radiological Sciences and Pathology, David Geffen School of Medicine at UCLA.

Abstract

Background: Several deep learning-based techniques have been developed for prostate cancer (PCa) detection using multi-parametric MRI (mpMRI), but few of them have been rigorously evaluated relative to radiologists' performance or whole-mount histopathology (WMHP).

Purpose: To compare the performance of a previously proposed deep learning algorithm, FocalNet, and expert radiologists in the detection of PCa on mpMRI with WMHP as the reference.

Study type: Retrospective, single-center study.

Subjects: 553 patients (development cohort: 427 patients; evaluation cohort: 126 patients) who underwent 3 T mpMRI prior to radical prostatectomy from October 2010 to February 2018.

Field Strength/Sequence: 3 T, T2-weighted imaging and diffusion-weighted imaging.

Assessment: FocalNet was trained on the development cohort to predict PCa locations by detection points, with a confidence value for each point, on the evaluation cohort. Four fellowship-trained genitourinary (GU) radiologists independently evaluated the evaluation cohort to detect suspicious PCa foci, annotate detection point locations, and assign a five-point suspicion score (1:least suspicious, 5:most suspicious) for each annotated detection point. The PCa detection performance of FocalNet and radiologists were evaluated by the lesion detection sensitivity versus the number of false-positive detections at different thresholds on suspicion scores. Clinically significant lesions: Gleason Group≥2 or pathological size≥10 mm. Index lesions: the highest Gleason Group and the largest pathological size (secondary).

Statistical tests: Bootstrap hypothesis test for the detection sensitivity between radiologists and FocalNet.

Results: For the overall differential detection sensitivity, FocalNet was 5.1% and 4.7% below the radiologists for clinically significant and index lesions, respectively; however, the differences were not statistically significant (P=0.413 and P=0.282, respectively).

Data Conclusion: FocalNet achieved slightly lower but not statistically significant PCa detection performance compared to GU radiologists. Compared with radiologists, FocalNet demonstrated similar detection performance for a highly sensitive setting (suspicion score≥1) or a highly specific setting (suspicion score=5) while lower performance in between.

Keywords: Deep learning; prostate cancer; automatic cancer detection; multi-parametric MRI.

Introduction

Multi-parametric MRI (mpMRI) acquired at 3 T has been shown to be highly sensitive for the detection of high-grade and index prostate cancer (PCa) (1). The main prostate mpMRI sequences for PCa detection include T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI), and dynamic contrast enhanced imaging (DCE). Prostate mpMRI is commonly interpreted by a combination of anatomical (T2WI) and functional (DWI and DCE) information, relying on a zonal-based subjective scoring of qualitative characteristics, as specified by the current iteration of Prostate Imaging Reporting and Data System version 2 (PI-RADS v2) (2). This subjective system has inter- and intra-reader variability (3–5) and has a finite range of sensitivity due to the qualitative scoring.

Artificial intelligence for automatic prostate cancer detection from mpMRI has shown promise to overcome such limitations of subjective and qualitative interpretation (6–14). An automatic detection system locates regions of suspicion and assesses their levels of suspiciousness from the input volumetric imaging of mpMRI. In particular, deep learning methods can recognize target patterns after training (15) and have achieved promising prostate cancer detection performance (9). However, the performance of these deep learning algorithms with respect to diagnostic accuracy in comparison to genitourinary (GU) radiologists has not been well studied.

FocalNet is a previously proposed deep learning system that detect PCa from volumetric 3 T mpMRI (16). Thus, the aim of our study was to evaluate and compare the performance of FocalNet with GU radiologists experienced in the interpretation of 3 T mpMRI for PCa detection.

Materials and Methods

This study was approved by the local institutional review board with a waiver for written informed consent and compliant with the 1996 United States Health Insurance Portability and Accountability Act.

Patients

The evaluation cohort for the study cohort included 126 patients from 184 patients who underwent robotic assisted laparoscopic radical prostatectomy (RALP) from July 2016 to February 2018 at a single academic medical center. The excluded patients had MRI from outside provider (n=37), hormonal, chemo or radiation treatment prior to prostatectomy (n=7), 1.5 T MRI (n=4), incomplete MRI imaging sequence (n=3), no WMHP (n=2), no MRI (n=2), or other reasons (n=3, Figure 2). The development cohort included 427 patients from October 2010 to June 2016, using the same exclusion criteria as the evaluation cohort, for the training and validation of FocalNet. Of the total 553 included patients in either cohort, 417 had been previously included in a pilot study describing the technical development of FocalNet (16). The development cohort was split into five folds for cross-validation, and the model parameters and optimization settings were tuned using the 5-fold cross-validation. After the model and parameter settings were determined, we tested the FocalNet algorithm with the evaluation cohort (Figure 1). The mean age for the 427 patients in the development cohort was 61.1±7.1 (mean±std) years (IQR: 56-66), and the mean age for the 126 patients in the evaluation cohort was 62.4±6.4 years (IQR: 58-68). The median duration between pre-operative MRI and radical prostatectomy was 63 days (IQR: 23-101) and 86 days (IQR: 45-130) for development and evaluation cohorts, respectively.

MR Imaging

We included one pre-operative 3 T mpMRI scan for each patient in this work. All images were acquired with one of four 3 T MR scanners (Trio, Verio, Skyra, Prisma; Siemens Healthineers, Erlangen, Germany). T2WI, DWI, and DCE MRI in the transverse plane were acquired using a standardized protocol following ESUR PI-RADS guidelines (2). The pelvic phased-array coil was used for all patients, and endorectal coil was used for a partial cohort (Table 1). The same protocol was applied in the whole cohort regardless of the coil. T2WI was acquired using axial turbo spin-echo (TSE) imaging with 3,800-5,040 ms repetition time (TR) and 101 ms echo time (TE). T2WI had the in-plane resolution of 0.625 x 0.625 mm^2 with a matrix size of 320 x 310 for a field of view (FOV) of 20 x 20 cm^2. The slice thickness was 3 mm with no gap between slices. DWI was acquired using echo-planar imaging (EPI) with 4800 ms TR and 80 ms TE using four different b-values (0/100/400/800 s/mm^2). DWI had the in-plane resolution of 1.6 x 1.6 mm^2 with a matrix size of 160 x 94 for a FOV of 26 x 21 cm^2. The slice thickness of DWI was 3.6 mm with no gap between slices. Apparent diffusion coefficient (ADC) maps were calculated by linear least-square fitting with all b-values.

Whole-mount histopathology

RALP was performed on average 79.7 (interquartile range (IQR): 27 - 107) days after the 3 T mpMRI. The resected prostate specimens were sectioned in the axial plane from the inked basal margin to the apex in approximately 5 mm intervals and mounted on large slides. The sliced WMHP specimen were examined by GU pathologists (with 12, 6, and 3 years of experience in clinical prostate histopathology interpretation) independent of radiological findings, and each PCa lesion was contoured and received a Gleason Score (GS) as a part of clinical routine. At a separate matching meeting with at least one GU radiologist and one GU pathologist, individual lesions detected on WMHP were matched on a slice-by-slice basis with MRI as a part of the standard of

care. After the matching meeting, GU radiology research fellows led by GU radiologists (S.R., E.F.) retrospectively identified and contoured MRI-visible lesions on T2WI. We defined prospectively missed lesions (i.e., false negatives) as MRI-visible lesions. The remaining MRI non-visible lesions, retrospectively unidentifiable in MRI with WMHP, were not included nor contoured for this study as we were unable to locate them or to confirm their presence at the time of MRI. The characteristics of MRI non-visible lesions are in Supplement Table S1.

Image Processing

The ADC map was resized and aligned to the T2WI using the field-of-view coordinates. An 80 mm x 80 mm window, centered at the prostate, was cropped in the transverse plane. The intensity of the T2WI was clipped and normalized linearly with a lower threshold of zero and an upper threshold determined by the bladder's intensity to account for significant intensity variations between mpMRI scans with and without an endorectal coil.

Deep learning: training and lesion detection

FocalNet took a slice of 3 T mpMRI as the input and predicted the cancer probability map for that slice. The T2WI and ADC of that slice were passed into FocalNet as two image channels, and the predicted probability maps for individual slices were stacked together to create a volumetric prediction probability map. For training, FocalNet was optimized by the stochastic gradient descent for the cancer probability map with binary lesion masks from the groundtruth lesion contours. For evaluation, detection points to localize PCa lesions were identified by searching for local maxima from the volumetric prediction probability map, and each detection point is

associated with a confidence value which is the predicted probability value at the detection point (16).

We made two modifications to the original FocalNet model (16) to accommodate full volumetric imaging. First, for the prediction of a given slice, that slice and its two adjacent slices were used as input into FocalNet. When the given middle slice is the first/last slice, the middle slice was duplicated as the missing adjacent slice. We expect this to help FocalNet utilize inter-slice information to promote or reject cancer prediction and to suppress false positive detections for non-prostate regions (e.g., bladder just outside of prostate gland base). Second, the model was trained using focal loss (17) only to alleviate the GPU memory consumption due to the additional slices as input.

Radiologists: lesion detection

Four board-certificated GU radiologists (19, 12, 10, and 5 years of experience in clinical prostate MRI interpretation) independently evaluated every case of the 126 cases in the evaluation cohort using Osirix MD 10.0 (Pixmeo SARL, Geneva, Switzerland). Each radiologist was instructed to place detection points to localize PCa lesions on T2WI after reviewing the T2WI and ADC map only, blinded to all clinical information or pathology reports. The radiologists scored each detection point using a 5-point Likert score from suspicion score 1 (least suspicious to be a PCa lesion) to suspicion score 5 (most suspicious to be a PCa lesion) based on their own experience.

In a session before the reading, the radiologists were instructed the general working principles of FocalNet and the detailed evaluation setup; e.g., multiple detection points on a single lesion or detection points on indolent prostate cancer (defined as a lesion with Gleason Group (GG) 1 and

pathological size less than 10mm) would not be penalized, nor would they improve the detection sensitivity.

Evaluation of Lesion Detection

We evaluated the overall detection accuracy of both FocalNet and GU radiologists using the free-response receiver operating characteristics analysis (FROC) (11). The FROC analysis measured detection sensitivity versus the number of false positives by thresholding on the confident value (FocalNet) or suspicion score (radiologist). For each radiologist and each threshold score, we found the matching point on the FROC curve of FocalNet based on the number of false positives per patient, and we computed the difference in detection sensitivity between a radiologist and FocalNet. The overall differential detection sensitivity was obtained by averaging the difference detection sensitivity between each radiologist and FocalNet at five suspicion score thresholds.

A detection point in or within a 5 mm margin of an annotated lesion contour was considered as a true positive and detection points outside of the 5 mm margin of any lesion contours were otherwise considered as false positives (16,18). The margin was assessed in 3 dimensions. More than one true-positive detection point on a single lesion did not affect the detection sensitivity. False-positive detection points were counted individually. Lesions with Gleason Group equal to or greater than 2 (GS 3+4) or pathological size larger than or equal to 10 mm were defined as clinically significant PCa (csPCa) (19), and the csPCa lesion with the highest Gleason Group and the largest pathological size (in the case of multiple lesions with the same Gleason Group) was considered as the index lesion for a patient (20,21). For the evaluation of a specific group (e.g., csPCa and index lesion), only lesions in the group were counted for detection sensitivity, and the detection of lesions outside of the specific group would contribute to neither detect sensitivity nor

the number of false-positive detection. The detection performance of FocalNet for lesions grouped by size and Gleason Group was also investigated. For size, all csPCa lesions were placed into three groups based on their pathological sizes (diameter): less than or equal to 10mm, 10–20mm, and greater than 20mm. Similarly, for Gleason Group, all csPCa lesions were placed into 3 groups: Gleason Group 1, Gleason Group 2, and Gleason Group 3 or above.

Statistical analysis

We used the non-parametric bootstrap to obtain inferences on this mean difference in sensitivity adjusting for number of false positives per patient. 10,000 bootstrap samples were obtained by resampling both patients and radiologists. We hypothesized that the radiologists' detection sensitivity was statistically different from the detection sensitivity of FocalNet. Two-tailed t-test (p=0.05) was performed on the differential detection sensitivity by inverting the bootstrap confidence interval.

Results

Lesion Characteristics

A total of 883 and 224 lesions were defined in WMHP in the development and evaluation cohorts, respectively, of which 709 and 185, respectively, were identified as MRI-visible on radiology-pathology review. Of the development cohort, at least one MRI-visible lesion was identified in 97.2% (415 of 427) patients, and 98.4% (124 of 126) of patients in the evaluation cohort had MRI-visible lesions. Overall 275 and 53 lesions in the development and evaluation cohorts, respectively,

that were prospectively missed on clinical MRI reports were identified on radiology-pathology review and included in this study. The detailed statistics for patients and their cancer characteristics are summarized in Table 1.

Lesion Detection

The FROC analysis results of our deep learning system, FocalNet, for the detection of prostate cancer index and clinically significant lesions in evaluation cohort are shown in Figures 4 and 5, respectively. The detection sensitivity of FocalNet was 50%, 80%, and 90% for index lesions at the cost of 0.24, 2.08, and 4.98 false-positive detections per patient on average, respectively. The detection sensitivity for clinically significant lesions was 50%, 80%, and 90% at 0.43, 3.39, and 11.7 false-positive detections per patient, respectively.

All readers evaluated the 3T mpMRI sequences after the development and evaluation of FocalNet were concluded. Each reader placed on average 9.0±2.2 detection points per patient. The detection performance of the radiologists and FocalNet for index prostate cancer lesions is reported in Figure 4 and Table 3. For high specificity setting (detection points with suspicion score = 5), the detection sensitivity of the radiologists on average was 0.4%±12.2% (range: -11.5% – 20.5%) lower than the detection sensitivity of FocalNet at the same false-positive detections per patient. For detection points with suspicion score ≥ 4, the radiologists' detection sensitivity was 8.4%±8.9% (range: -4.1% – 19.7%) higher than FocalNet. For high sensitivity setting (detection points with suspicion score ≥ 1), the radiologists' detection sensitivity was 0.4%±6.6% (range: -0.4% – 6.5%) lower than FocalNet. The overall differential detection sensitivity of radiologists for index lesions was 4.7% (95% CI: -4.9%–14.3%) over FocalNet, which was not statistically significant (p=0.413).

The detection performance of the radiologists and FocalNet for clinically significant prostate cancer lesions is reported in Figure 5 and Table 4. For high specificity setting (detection points with suspicion score = 5), the radiologists' detection sensitivity was 0.9%±10.3% (range: -15.7% – 10.6%) higher than FocalNet. For detection points with suspicion score ≥ 3, the radiologists' detection sensitivity was 8.2%±5.5% (range: -0.7% – 13.6%) higher than FocalNet. For high sensitivity setting (detection points with suspicion score ≥ 1), the sensitivity of the radiologists was 0.9%±4.6% (range: -2.5% – 8.8%) higher than FocalNet. The overall differential detection sensitivity of radiologists for clinically significant lesions was 5.1% (95% CI: -3.4%–13.2%) over FocalNet, which was not statistically significant (p=0.282). As shown in the last two columns of Table 4, on average 9.0% (range 5.0% – 23.6%) of total clinically significant lesions were missed by radiologists while detected by FocalNet at the same number of false-positive detections per patient, even though the radiologists' overall detection sensitivity was slightly higher than FocalNet. At high sensitivity settings (suspicion score ≥ 1 or 2), on average 8.5% (range 5.0% – 10.6%) of clinically significant lesions can be additionally detected by FocalNet on top of the detected lesions by a radiologist.

An example case is shown in Supplement Figure S3 to further illustrate that FocalNet's potential of providing predictions complementary to radiologists' output. The unifocal, transition zone prostate cancer lesion with GG1 and pathological size 20mm was detected by FocalNet with a 0.892 confidence value and no false-positive detection. In contrast, in radiologists' reading, four radiologists marked this lesion by a suspicion score of 2, 1, 1, 3, respectively, and radiologists made 6, 9, 9, 1 false-positive detection, respectively, in order to identify this true lesion.

Detection for Lesions Grouped by Size, Gleason Group

The detection sensitivity for lesions less than or equal to 10mm, 10–20mm, and greater than 20mm was 26.7%, 60.3%, and 73.2%, respectively, at 1 false-positive detection per patient and 40.0%, 68.3%, and 84.2%, respectively, at 2 false-positive detections per patient (Supplemental Figure S1). Additionally, the detection sensitivity for lesions with Gleason Group 1, Gleason Group 2, and Gleason Group 3 or above was 55.6%, 62.3%, and 74.1%, respectively, at 1 false-positive detection per patient and 59.3%, 73.9%, and 81.5%, respectively, at 2 false-positive detections per patient (Supplemental Figure S2).

Discussion

In this study, FocalNet had a comparable detection performance for a high specificity setting, which provided evidence that the detection of highly suspicious PCa regions could be automated using the deep learning systems. While the benefits for the annotation and scoring of less suspicious regions are currently being established, the deep learning systems could be competent to detect and stratify less suspicious regions in an efficient and systematic way.

We analyzed the sets of detected lesions and counted clinically significant lesions that were detected by either FocalNet or radiologists but missed by the other. FocalNet identified a subset of targets that might be difficult for expert radiologists to identify, which suggested that FocalNet may allow improved detection of clinically significant lesions when incorporated into radiologists' workflow.

Our results showed that the performance of FocalNet, after training, was comparable to that of experienced GU radiologists, especially in either a highly sensitive setting (suspicion score ≥ 1) or a highly specific setting (suspicion score = 5). This finding potentially suggests that deep learning can be integrated into the clinical workflow, as an assistant tool, to identify all potential

PCa lesions (high sensitivity setting) or highly suspicious PCa lesions only (high specificity setting).

Schelb et al. (7) developed a deep learning system to classify PCa with Gleason Group ≥ 2 from the biopsy sample and achieved comparable performance with clinically assigned PI-RADS scores of either 3 or 4. Similarly, Zhong et al. (22) reported similar classification performance of clinically significant PCa as PI-RADS scores either 3 or 4 with the WMHP reference.

Unlike prior studies using labels from biopsy-based histopathology as reference standards (11,13,23), our study included all retrospectively identifiable lesions in MRI with WMHP as a reference standard. This minimized biases for the assessment of PCa detection performance of our deep learning system and radiologist readers based on the whole prostate pathology. Our definition of MRI-visible lesions included lesions that were missed in prospective MRI reading before RALP but retrospectively identified in MRI based on WMHP. While the matching meeting conducted by radiologists and pathologists ensured the accuracy of the MRI-histopathology correlation, a small number of WMHP lesion findings were not identifiable in MRI. These MRI non-visible lesions were not included in our study, as they could not be located or contoured in MRI.

Limitations

Both the development and evaluation cohorts consisted of patients from a single medical center, and all the MRI scans were acquired using consistent imaging parameters by MRI scanners from a single manufacturer. Our deep learning system may need some adjustments to accommodate potential imaging heterogeneity caused by different MRI scanners and settings (12). Additionally, all data was retrospectively assessed. A prospective study, where the system is fixed and trained prior to data collection, is necessary to validate the system in the future study. This study population may not fully represent prostate cancer screening that contains a large percentage

of cancer-negative patients, as all patients in our study population later underwent radical prostatectomy due to prostate cancer positivity.

Our study focused on a comparison of prostate cancer detection performance between in-house developed deep learning-based system, FocalNet, and expert GU-radiologists, in order to have a better understanding of the capability of deep learning-based detection systems. We considered the deep learning system as an independent tool to perform the detection on its own, which helped determine the system's strength and weaknesses when compared with radiologists. However, the potential of deep learning detection systems as assistive tools to radiologists in the clinical workflow (24) was not fully explored in this study and left for future investigation.

The radiologists were only provided with T2WI and ADC maps in axial orientation, which were the input for the deep learning system. While a previous study suggested that using T2WI and ADC maps together can achieve similar detection as using all the series (25), radiologists usually have more imaging series (e.g., DCE MRI, high b-value DWI, and imaging in different orientations) to look at for their clinical interpretation. Although we provided the same images to the radiologists as FocalNet for a fair comparison, the radiologists' performance may be underestimated compared to the actual clinical setting where they would have access to additional image types.

Finally, all four radiologists in this study were from academic medical centers and specialized in prostate MRI. The 5-point Likert suspicion scores assigned by our readers were based on their experience as the PI-RADS v2 scoring was not applicable for 1) DWI, DCE imaging was not provided and 2) PI-RADS score 1 or 2 was rarely used (PI-RADS score 1 indicated non-suspicious area). For this reason, their outputs may not be fully representative of or reproducible by less experienced radiologists. Also, the relatively small number of recruited readers caused a

variation of performance, and future study may expand thhe reader recruitment to better understand the inter-reader variability.

Conclusion

On the evaluation cohort, FocalNet achieved slightly lower but not statistically significant PCa detection performance compared to GU radiologists. Compared with radiologists, FocalNet demonstrated similar detection performance for a highly sensitive setting (suspicion score ≥ 1) or a highly specific setting (suspicion score = 5) while lower performance in between. This suggested that the detection of highly suspicious PCa could be automated using the deep learning systems.

References

1. Turkbey B, Choyke PL. Multiparametric MRI and prostate cancer diagnosis and risk stratification. Current Opinion in Urology. 2012.
2. Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, et al. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. Eur Urol. 2016;
3. Ruprecht O, Weisser P, Bodelle B, Ackermann H, Vogl TJ. MRI of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy. Eur J Radiol. 2012;81(3):456–60.
4. Kasel-Seibert M, Lehmann T, Aschenbach R, Guettler F V., Abubrig M, Grimm MO, et al. Assessment of PI-RADS v2 for the Detection of Prostate Cancer. Eur J Radiol. 2016;

5. Greer MD, Shih JH, Lay N, Barrett T, Bittencourt L, Borofsky S, et al. Interreader variability of prostate imaging reporting and data system version 2 in detecting and assessing prostate cancer lesions at prostate MRI. Am J Roentgenol. 2019;212(6):1197–205.
6. Wang S, Burtt K, Turkbey B, Choyke P, Summers RM. Computer aided-diagnosis of prostate cancer on multiparametric MRI: A technical review of current research. BioMed Research International. 2014.
7. Schelb P, Kohl S, Radtke JP, Wiesenfath M, Kickingereder P, Bickelhaupt S, et al. Classification of cancer on prostate MRI: Deep learning vs. clinical PI-RADS assessment. Radiology. 2019;
8. Bonekamp D, Kohl S, Wiesenfarth M, Schelb P, Radtke JP, Götz M, et al. Radiomic machine learning for characterization of prostate lesions with MRI: Comparison to ADC values. Radiology. 2018;
9. Tsehay YK, Lay NS, Roth HR, Wang X, Kwak JT, Turkbey BI, et al. Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images. In: Medical Imaging 2017: Computer-Aided Diagnosis. 2017.
10. Sumathipala Y, Lay N, Turkbey B. Prostate cancer detection from multi-institution multiparametric MRIs using deep convolutional neural networks. J Med Imaging. 2018;
11. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. Computer-aided detection of prostate cancer in MRI. IEEE Trans Med Imaging. 2014;
12. Gibson E, Hu Y, Ghavami N, Ahmed HU, Moore C, Emberton M, et al. Inter-site Variability in Prostate Segmentation Accuracy Using Deep Learning. In: International

Conference on Medical Image Computing and Computer-Assisted Intervention. 2018. p. 506–14.

13. Armato SG, Huisman H, Drukker K, Hadjiiski L, Kirby JS, Petrick N, et al. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. J Med Imaging. 2018;

14. Wang Z, Liu C, Cheng D, Wang L, Yang X, Cheng KT. Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network. IEEE Trans Med Imaging. 2018;

15. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. 2012.

16. Cao R, Bajgiran AM, Mirak SA, Shakeri S, Zhong X, Enzmann D, et al. Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet. IEEE Trans Med Imaging. 2019;

17. Lin T, Girshick R, Doll P. Focal Loss for Dense Object Detection Online Appendix. In: International Conference on Computer Vision. 2017. p. 2999–3007.

18. Priester A, Natarajan S, Khoshnoodi P, Margolis DJ, Raman SS, Reiter RE, et al. Magnetic Resonance Imaging Underestimation of Prostate Cancer Geometry: Use of Patient Specific Molds to Correlate Images with Whole Mount Pathology. J Urol. 2017;

19. Le JD, Tan N, Shkolyar E, Lu DY, Kwan L, Marks LS, et al. Multifocality and prostate cancer detection by multiparametric magnetic resonance imaging: Correlation with whole-mount histopathology. Eur Urol. 2015;

20. Tan N, Margolis DJ, Lu DY, King KG, Huang J, Reiter RE, et al. Characteristics of detected and missed prostate cancer foci on 3-T multiparametric MRI using an endorectal

coil correlated with whole-mount thin-section histopathology. Am J Roentgenol. 2015;

21. Ploussard G, Epstein JI, Montironi R, Carroll PR, Wirth M, Grimm MO, et al. The contemporary concept of significant versus insignificant prostate cancer. European Urology. 2011.

22. Zhong X, Cao R, Shakeri S, Scalzo F, Lee Y, Enzmann DR, et al. Deep transfer learning-based prostate cancer classification using 3 Tesla multi-parametric MRI. Abdom Radiol. 2019;

23. Tsehay Y, Lay N, Wang X, Kwak JT, Turkbey B, Choyke P, et al. Biopsy-guided learning with deep convolutional neural networks for Prostate Cancer detection on multiparametric MRI. In: Proceedings - International Symposium on Biomedical Imaging. 2017.

24. Sim Y, Chung MJ, Kotter E, Yune S, Kim M, Do S, et al. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology. 2020;

25. Delongchamps NB, Rouanne M, Flam T, Beuvon F, Liberatore M, Zerbib M, et al. Multiparametric magnetic resonance imaging for the detection and localization of prostate cancer: combination of T2-weighted, dynamic contrast-enhanced and diffusion-weighted imaging. BJU Int. 2011;107(9):1411–8.

Table 1: Study cohort patient characteristics.

Variable	Development Cohort (n=427)	Evaluation Cohort (n=126)
Median age (years)	61 (IQR: 56-66)	62.5 (IQR: 58-68)
Median PSA (ng/mL)	6 (IQR: 4.6-8.3)	6.2 (IQR: 4.9-9.5)
Median PSA density	0.16 (IQR: 0.11-0.24)	0.17 (IQR: 0.11-0.25)
Scanner		
Skyra	255	118
Prisma	18	7
Trio	130	1
Verio	24	0
With endorectal coil	234	23
Without endorectal coil	193	103
No. of pathological lesions	883	224
No. of MRI-visible pathological lesions	709	185
No. of MRI-visible pathological lesions that were missed in prospective MRI reading before RALP	257	57

No. of patients with pathological lesions (including not MRI-visible)		
1 lesion	157	57
2 lesions	141	46
3 lesions	83	18
>= 4 lesions	46	5
No. of patients with MRI-visible pathological lesions		
No MRI-visible lesions	12	2
1 lesion	217	74
2 lesions	130	40
3 lesions	46	9
>= 4 lesions	22	1
No. of patients with highest Gleason Group (Gleason Score) MRI-visible lesion		
Gleason Group 1 (GS 3+3)	57	12
Gleason Group 2 (GS 3+4)	218	55
Gleason Group 3 (GS 4+3)	95	34
Gleason Group 4 and 5 (GS >=8)	45	23

Note: data are reported as number of 427 (development) and 126 (evaluation) patients. IQR: Interquartile range.

Table 2: MRI-visible prostate cancer lesion characteristics.

Variable	Development Cohort (n=709)	Evaluation Cohort (n=185)
Gleason Group (Gleason Score)		
1 (3+3)	281	52
2 (3+4)	281	69
3 (4+3)	102	37
4 and 5 (>=8)	45	27
Median pathological size (mm)	1.5 (IQR: 0.8-2.2)	1.8 (IQR: 1.18-2.6)
Lesion size		
<=10 mm	224	40
10-20 mm	269	63
>20 mm	209	82

Note: data are reported as number of 427 (development) and 126 (evaluation) patients. IQR: Interquartile range.

Table 3: Index lesion detection performance results.

Suspicion Score	Radiologist	FocalNet Sensitivity (%)	Radiologist Sensitivity (%)	FPPP	Difference (%)	Average Difference (%)
5	1	47.5 (58/122)	54.1 (66/122)	0.15	+6.6	-0.4±12.2
	2	34.4 (42/122)	45.9 (56/122)	0.06	+11.5	
	3	48.4 (59/122)	27.9 (34/122)	0.18	-20.5	
	4	24.6 (30/122)	25.4 (31/122)	0.01	+0.4	
≥ 4	1	59.0 (72/122)	72.1 (88/122)	0.50	+13.1	+8.4±8.9
	2	57.4 (70/122)	77.1 (94/122)	0.44	+19.7	
	3	54.1 (66/122)	50.0 (61/122)	0.33	-4.1	
	4	48.4 (59/122)	53.3 (65/122)	0.19	+4.9	
≥ 3	1	70.5 (86/122)	84.4 (103/122)	1.24	+13.9	+9.4±6.9
	2	65.0 (79/122)	81.2 (99/122)	0.90	+16.2	
	3	60.7 (74/122)	59.0 (72/122)	0.56	-1.7	
	4	62.3 (76/122)	71.3 (87/122)	0.63	+9.0	
≥ 2	1	81.2 (99/122)	87.7 (107/122)	2.35	+6.5	+5.1±7.3
	2	81.2 (99/122)	86.1 (105/122)	2.43	+4.9	
	3	89.3 (109/122)	83.6 (102/122)	4.56	-5.7	
	4	73.8 (90/122)	88.5 (108/122)	1.43	+14.7	

≥ 1	1	92.6 (113/122)	90.2 (110/122)	7.23	-2.5	-0.4±6.6
	2	92.6 (113/122)	89.3 (109/122)	7.15	-3.3	
	3	92.6 (113/122)	86.1 (105/122)	7.15	-6.5	
	4	78.7 (96/122)	89.3 (109/122)	1.90	+0.4	

Sensitivity (Sen.) in percent, number of false positive per patient (FPPP). Sensitivity difference = radiologists' sensitivity – FocalNet's sensitivity.

Table 4: Clinically significant lesion detection performance results.

Suspicion Score	Radiologist	FocalNet Sensitivity (%)	Radiologist Sensitivity (%)	FPPP	Difference (%)	Average Difference (%)	Detected by FocalNet but not by Radiologist (%)	Detected by Radiologist but not by FocalNet (%)
5	1	40.6 (65/160)	48.7 (78/160)	0.15	+8.1	+0.9±10.3	10.6	18.8
	2	30.0 (48/160)	40.6 (65/160)	0.06	+10.6		6.9	17.5
	3	41.3 (66/160)	25.6 (41/160)	0.18	-15.7		21.9	6.3
	4	22.5 (36/160)	23.1 (37/160)	0.01	+0.6		5.6	6.3
≥ 4	1	52.5 (84/160)	65.0 (104/160)	0.50	+12.5	+8.5±7.9	8.1	20.6
	2	51.3 (82/160)	69.4 (111/160)	0.44	+18.1		5.6	23.7
	3	46.9 (75/160)	43.7 (70/160)	0.33	-3.2		13.8	10.6
	4	41.3 (66/160)	48.1 (77/160)	0.19	+6.8		8.8	15.6
≥ 3	1	65.6 (105/160)	77.5 (124/160)	1.24	+11.9	+8.2±5.5	5.6	17.5
	2	60.8 (97/160)	74.4 (119/160)	0.90	+13.6		5.6	19.3
	3	54.4 (87/160)	53.7 (86/160)	0.56	-0.7		13.8	13.1
	4	56.3 (90/160)	64.4 (103/160)	0.63	+8.1		6.9	15.0
≥ 2	1	75.6 (121/160)	80.6 (129/160)	2.35	+5.0	+5.0±4.6	8.8	13.8
	2	75.6 (121/160)	81.2 (130/160)	2.43	+5.6		8.1	13.8
	3	82.5 (132/160)	80.6 (129/160)	4.56	-1.9		10.6	8.8
	4	68.8 (110/160)	80.0 (128/160)	1.43	+11.2		6.3	17.5
≥ 1	1	87.5 (140/160)	85.6 (137/160)	7.23	-1.9	+0.9±4.6	9.4	7.5
	2	87.5 (140/160)	86.9 (139/160)	7.15	-0.6		8.8	8.1
	3	87.5 (140/160)	85.0 (136/160)	7.15	-2.5		10.6	8.1
	4	73.1 (117/160)	81.9 (131/160)	1.90	+8.8		5.0	13.8

Sensitivity (Sen.) in percent, number of false positive per patient (FPPP). Sensitivity difference = radiologists' sensitivity – FocalNet's sensitivity.

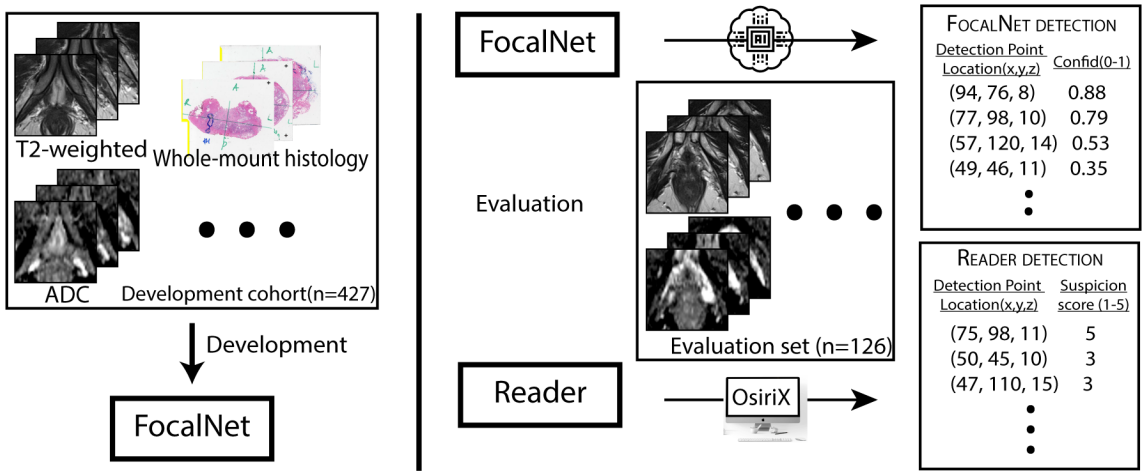


Figure 1: Illustration of the processes for FocalNet development (left) and the evaluation (right) of cancer detection performance by FocalNet and experienced genitourinary radiologists under the same setting. ADC: apparent diffusion coefficient. Confid: confidence value.

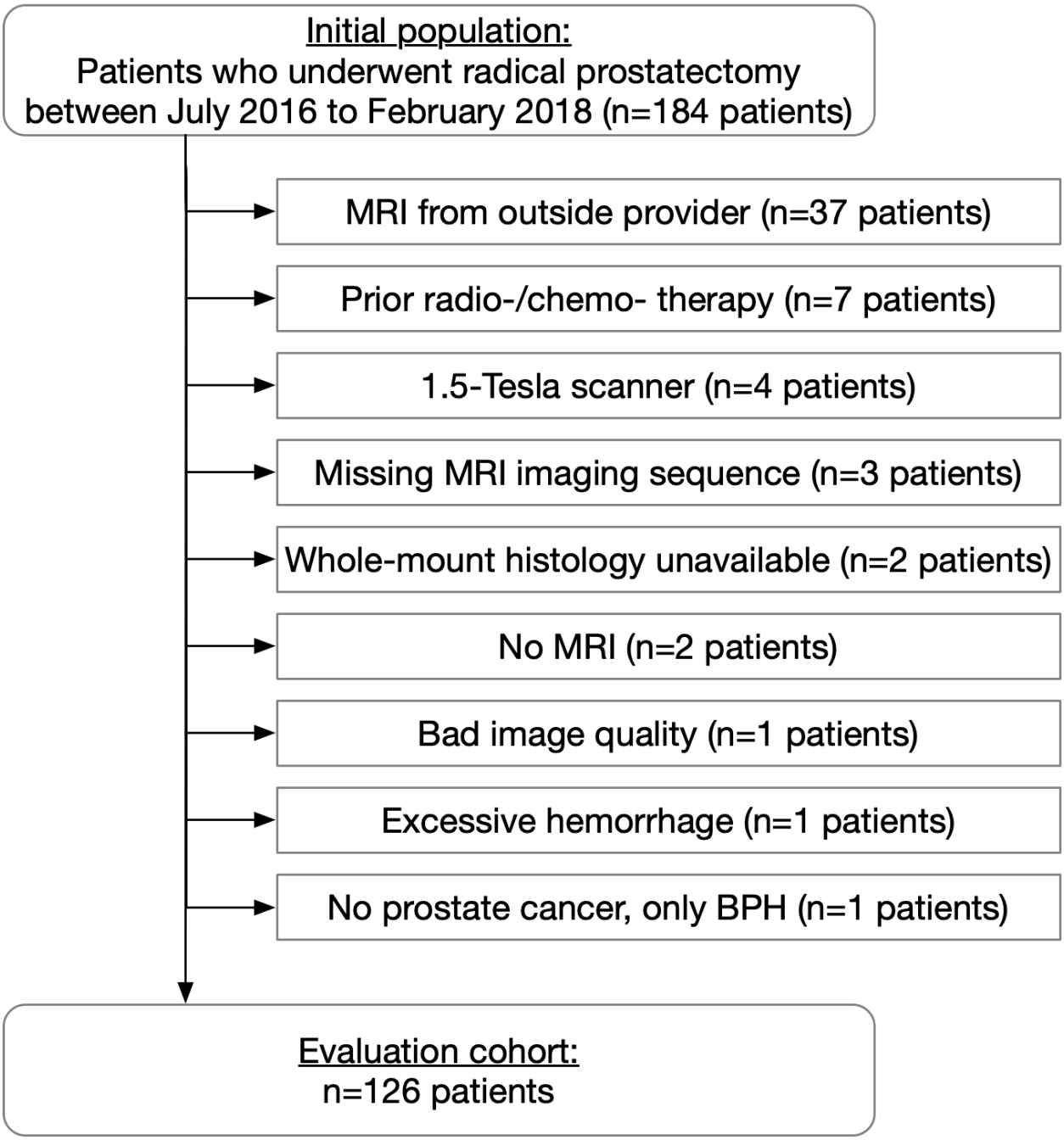


Figure 2: Flowchart for study inclusion for the evaluation cohort.

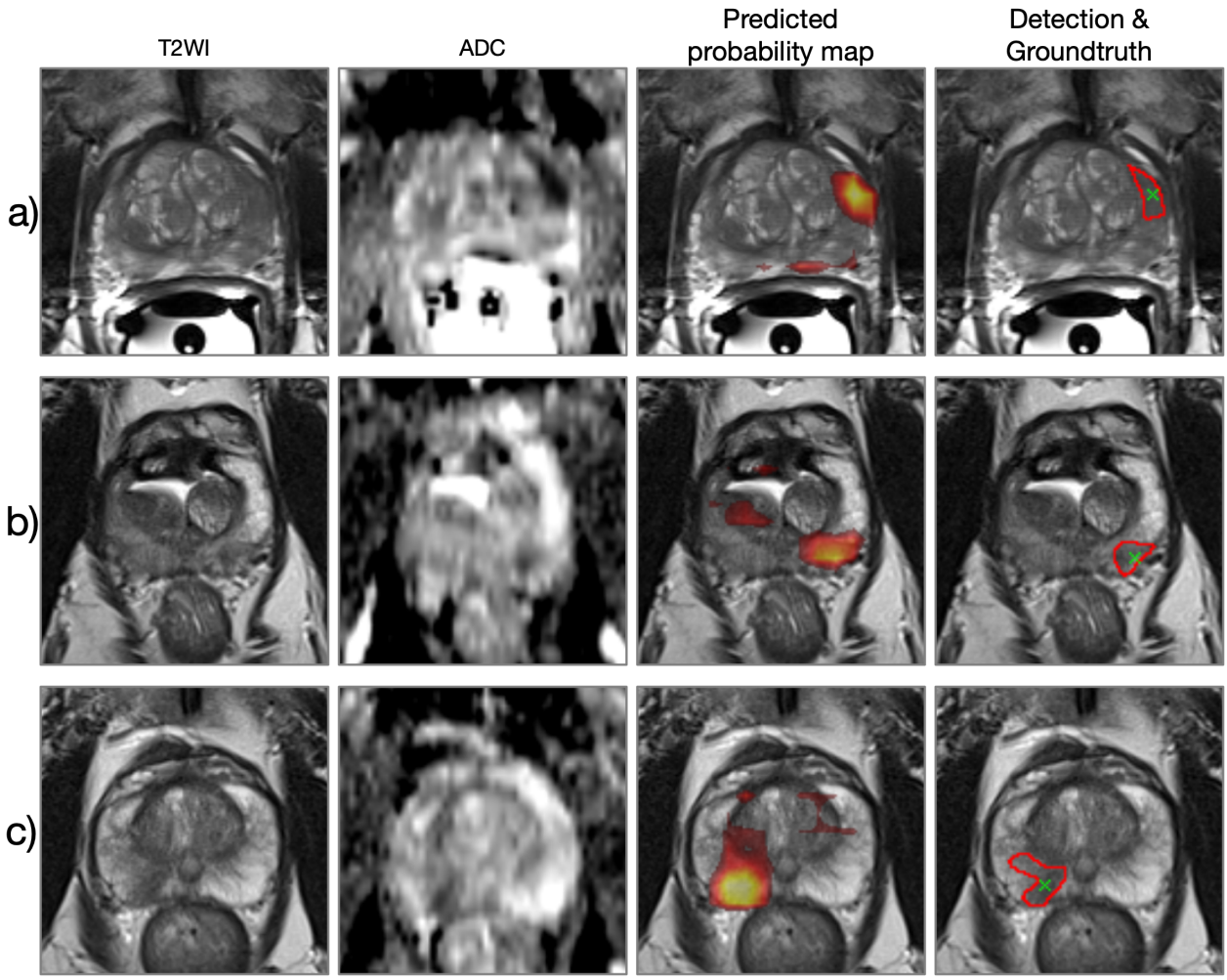


Figure 3: Examples of lesion detection. The left two columns show the input T2WI and ADC map, respectively. The right two columns show the FocalNet predicted lesion probability map and detection points (green crosses) with reference lesion annotation (red contours), respectively. a) patient at age 66, with a PCa lesion at left anterior peripheral zone with Gleason Group 5 (Gleason Score 4+5). b) patient at age 68, with a PCa lesion at left posterolateral peripheral zone with Gleason Group 2 (Gleason Score 3+4). c) patient at age 69, with a PCa lesion at right posterolateral peripheral zone with Gleason Group 3 (Gleason Score 4+3).

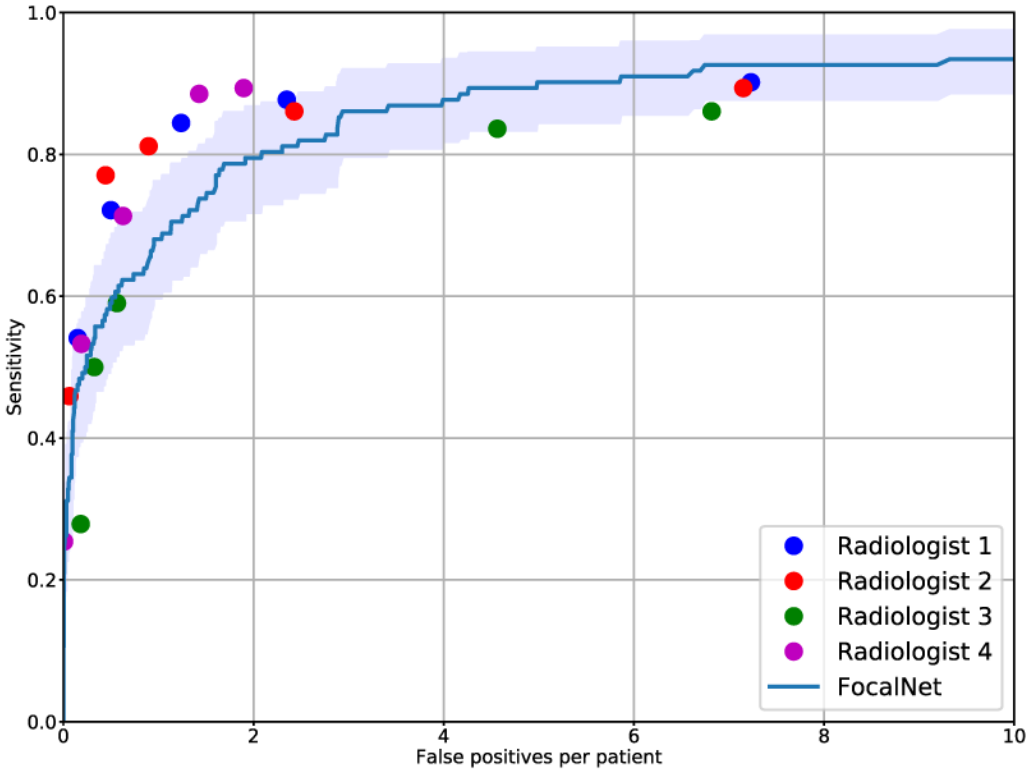


Figure 4: Free-response receiver operating characteristics (FROC) analysis for index lesion detection for 126 patients in the evaluation cohort with detection sensitivity plotted as a function of the number of false-positive detections for each patient on average. The shaded area surrounding the FocalNet curve (blue) shows the 95% confidence interval for detection sensitivity by bootstrapping the patient population. Dots indicate each radiologist performance at suspicion score thresholds.

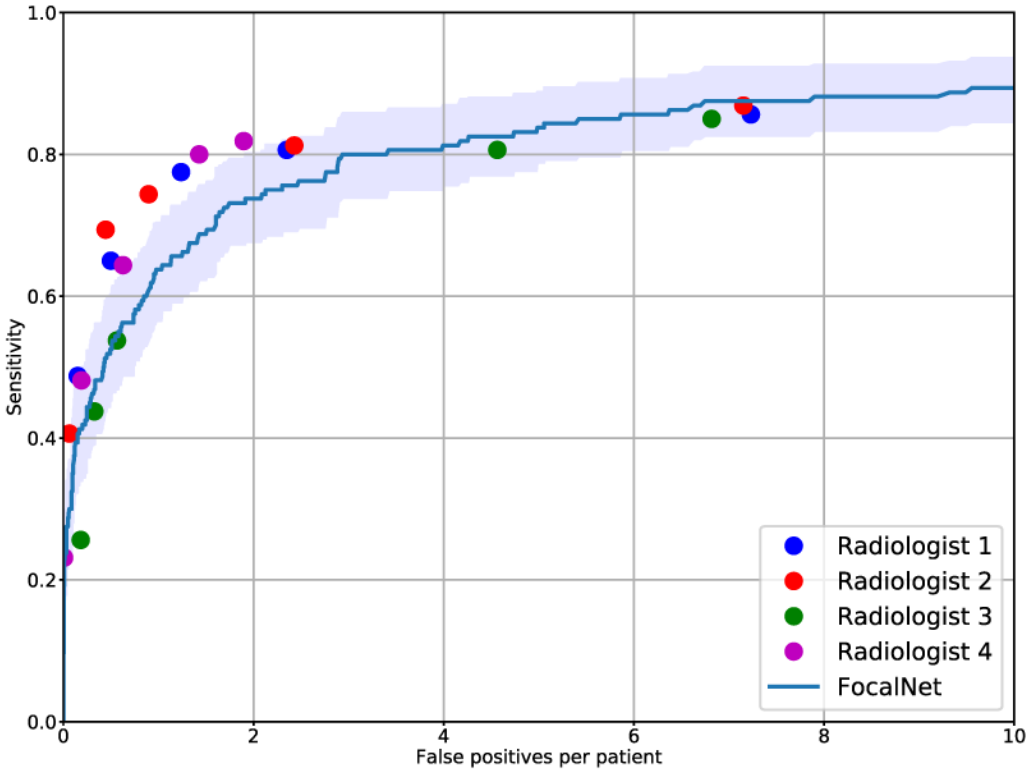
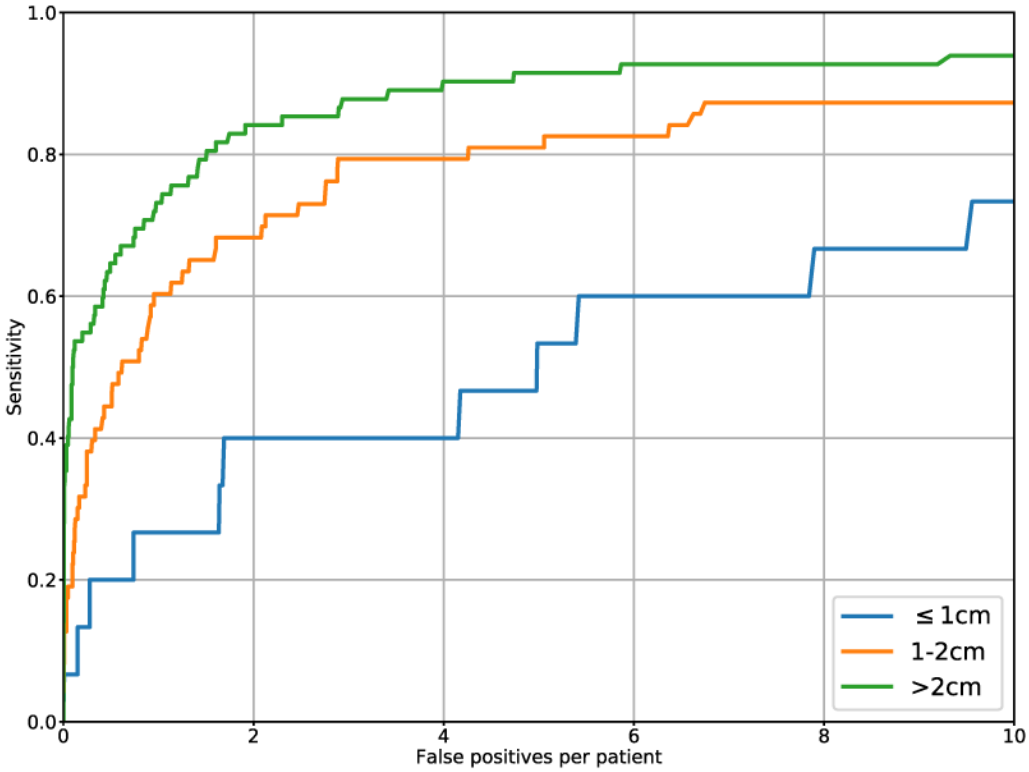


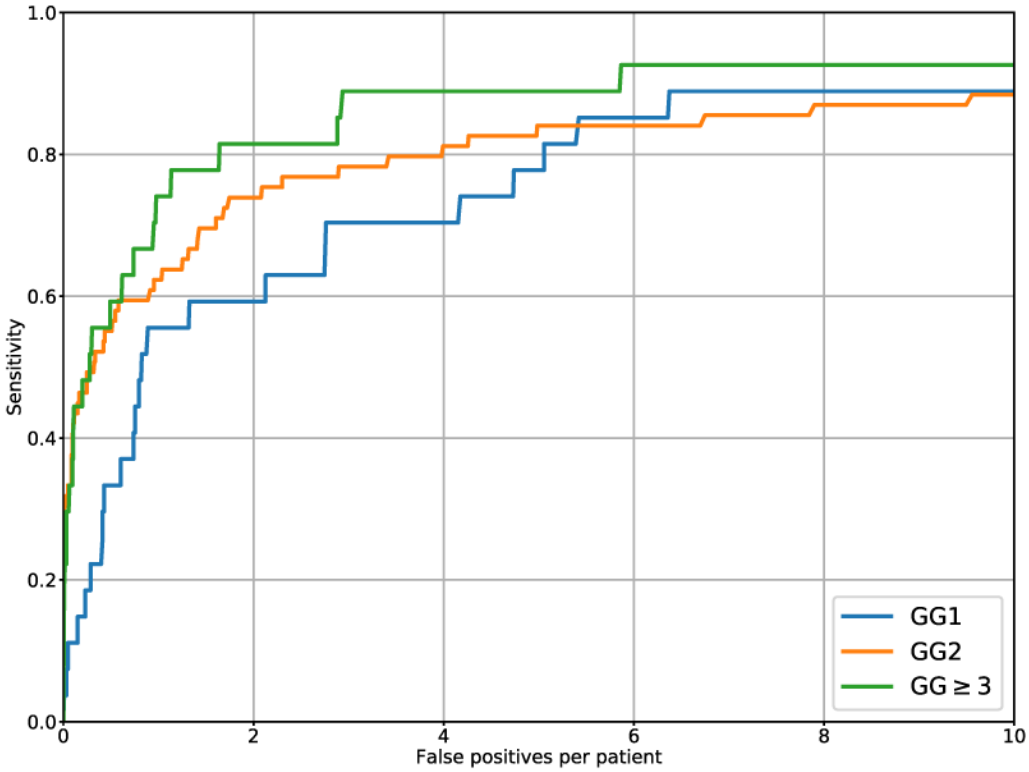
Figure 5: Free-response receiver operating characteristics (FROC) analysis for clinically significant lesion detection. The shaded area surrounding the FocalNet curve (blue) shows the 95% confidence interval for detection sensitivity by bootstrapping the patient population. Dots indicate each radiologist performance at suspicion score thresholds.

Supplement Table S1. MRI-invisible prostate cancer lesion characteristics.

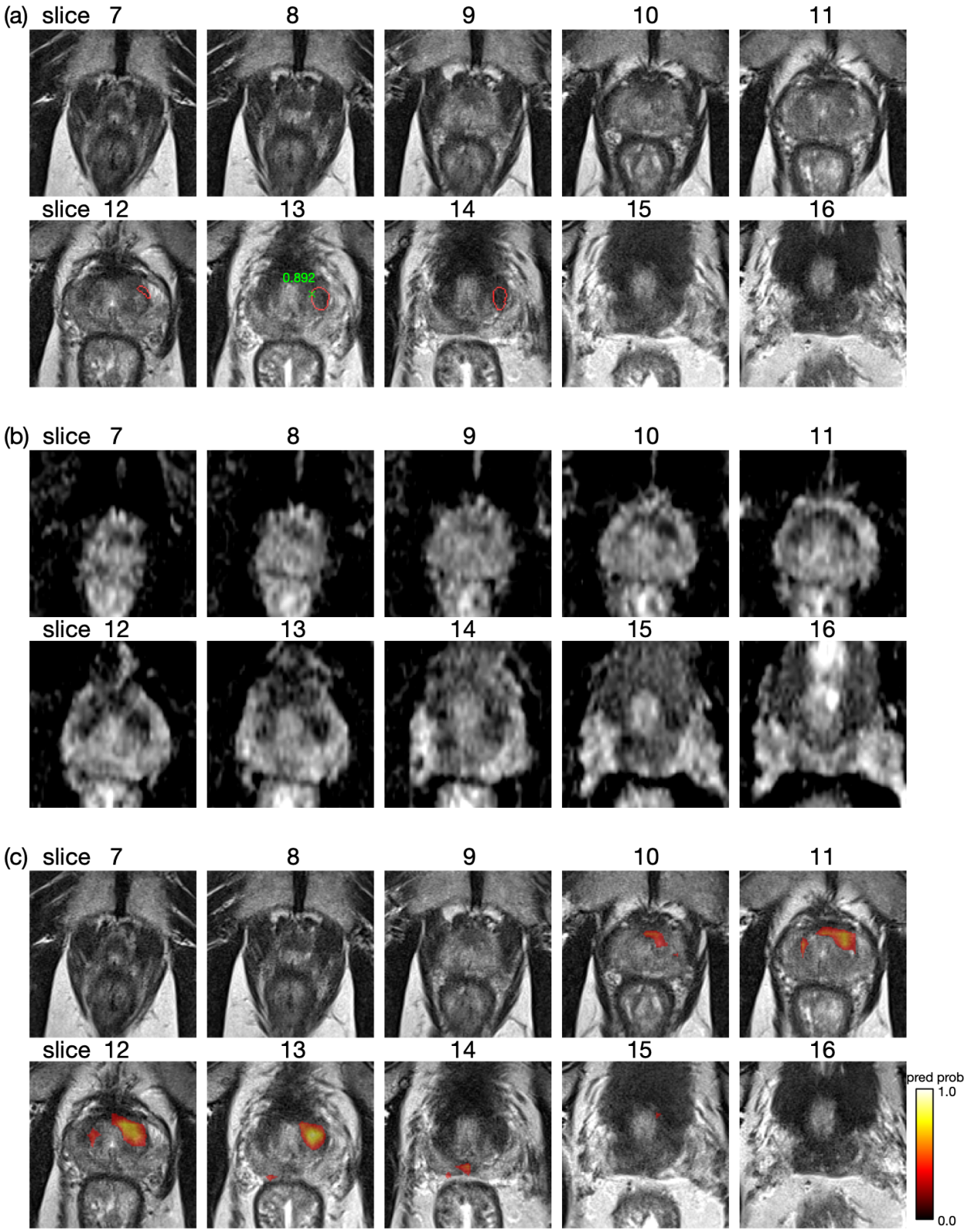
Variable	Development Cohort (n=174)	Evaluation Cohort (n=39)
Gleason Group (Gleason Score)		
1 (3+3)	126	29
2 (3+4)	36	9
3 (4+3)	8	0
4 and 5 (>=8)	4	1



Supplement Figure S1. Free response receiver operating characteristics analysis for the FocalNet's detection of clinically significant prostate cancer lesions grouped by their pathological sizes in diameter.



Supplement Figure S2. Free response receiver operating characteristics analysis for the FocalNet's detection of clinically significant prostate cancer lesions grouped by their Gleason Group.



Supplement Figure S3: An example case in evaluation set. A GG1 (GS 3+3) lesion is on the

transition zone with a pathological size of 20mm. (a) 10 out of 20 slices of T2WI with FocalNet's detection points (green cross) and the corresponding confidence values and the groundtruth lesion annotation (red contours). (b) The corresponding ADC slices. (c) FocalNet's predicted lesion probability map on T2WI.