Prostate Lesion Estimation using Prostate Masks from Biparametric MRI

Ahmet Karagöz[1*] , Mustafa Ege Şeker[2*], Mert Yergin[3*], Tarkan Atak Kan[1*], Mustafa Said Kartal[4*], Ercan Karaarslan[2*], Deniz Alis[2*] and Ilkay Oksuz[1*]

[1] Computer Engineering, Istanbul Technical University, Istanbul, Turkey
[2] Acibadem Mehmet Ali Aydinlar University School of Medicine, Istanbul, Turkey
[3] Department of Software Engineering and Applied Sciences, Bahcesehir University, Istanbul, Turkey
[4] Faculty of Medicine, Sivas Cumhuriyet University, Sivas, Turkey

Abstract. Biparametric MRI has emerged as an alternative to multi-parametric prostate MRI, which eliminates the need for the potential harms to the patient due to the contrast medium. One major issue with biparametric MRI is difficulty to detect clinically significant prostate cancer (csPCA). Deep learning algorithms ave emerged as an alternative solution to detect csPCA in cohort studies. We present a workflow which predicts csPCA on biparametric prostate MRI PI-CAI 2022 Challenge with over 10,000 carefully-curated prostate MRI exams. We propose to to segment the prostate gland first to the central gland (transition + central zone) and the peripheral gland. Then we utilize these predcitions in combination with T2, ADC and DWI images to train an ensemble nnU-net model. Finally, we utilize clinical indices PSA and ADC intensity distributions of lesion regions to reduce the false positives. Our method achieves top results on open-validation stage with a AUROC of 0.888 and AP of 0.732.

Keywords: Prostate MRI · Lesion Detection · nnU-net · Deep Learning

1 Introduction

Magnetic resonance imaging (MRI) plays an imperative role in prostate cancer (PCa) diagnostics, and the number of prostate MRI scans is expected to significantly increase as the recent evidence suggests that performing pre-biopsy prostate MRI in men with suspicion of PCa [1]. The main objective of prostate MRI is to identify clinically significant PCa (csPCa) (i.e., $GleasonScore \geq 3 + 4$) while sparring men with benign lesions or indolent prostate from unnecessary interventions or treatment. The prostate imaging-reporting and data system (PI-RADS) was introduced in 2012 and updated twice to standardize prostate MRI acquisition and interpretation [13]. The up-to-date multi-parametric prostate MRI (mpMRI) protocol includes axial T2-weighted images, diffusion-weighted images with a high b-value, apparent diffusion coefficient (ADC) maps calculated

---

* These authors contributed equally to this work

from DWI and dynamic contrast-enhanced images. The alternative to prostate mpMRI is bi-parametric prostate, which omits the contrast-enhanced sequence, thereby eliminating the potential hazards from the contrast medium. Though the benefits of the PI-RADS have been well recognized over the years, prostate MRI still suffers from intra-reader and inter-reader differences and non-negligible amounts of false-positive and false-negative results [12].

Deep learning (DL) has shown remarkable performance on a broad spectrum of medical imaging tasks in recent years, with prostate cancer diagnostics no exception. However, despite the promises of DL technology in PCa, most of the proposed DL models have been trained on small single-center proprietary data and were not publicly shared, hindering the reliability and wide-spread adaptation [7]. ProstateX challenge partially addressed this problem, yet it did not have the adequate data size to train and test DL models for prostate MRI effectively [14].

PI-CAI (Prostate Imaging: Cancer AI) is a new grand challenge encompassing over 10.000 carefully annotated prostate MRI scans [8]. The main goal of the challenge is to allow researchers to design, train and test publicly available DL models in large-scale for identifying csPCa on bi-parametric prostate MRI. In this challenge, we used the state-of-the-art medical image segmentation model, nnUnet, as the base model and made several contributions to improve its performance.

2   Method

Studies have shown a relationship between prostate volume, benign and malignant pathologies [5]. In particular, the combination of prostate volume with prostate-specific antigen (PSA) can be used as a biomarker in clinically significant prostate cancer (csPCA) detections [9]. PSA density (PSAd) which is the ratio between PSA and prostate volume, is an important biomarker in today's prostate cancer guidelines such as PI-RADS. Interestingly, it was shown that Transition zone PSA density (TZPSAd) is a more accurate indicator than PSAd in csPCA detection [11]. This shows that prostate cancers, which are located within the prostate areas that host histologically different structures, differ in terms of cancer incidence, prognosis, and outcomes. For these reasons, before creating the prostate lesion segmentation / detection models, the prostate gland segmentation model was created. The central gland (transition + central zone) and the peripheral gland were segmented after.

2.1   Prostate Gland Segmentation

The prostate gland segmentation model was trained on 203 patients of the ProstateX [3] dataset. Only the T2 modality was used in this model, because it was the modality in which anatomical structures can be best distinguished. nnUNet is a self-configuring framework that selects the best architecture depending on the input dataset by adhering to a set of principles [6] (as detailed

in Figure 1). The sum of crossentropy and soft Dice loss is commonly applied by the nnU-Net framework, and the loss is applied at different resolutions (deep supervision). A standardized data preparation and augmentation pipeline is included with nnU-Net. In addition, extreme data augmentation techniques (such as Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution and gamma augmentation, elastic deformation, scale, flip, mirror, and rotate) were used to maintain model performance at its highest level despite variations in acquisition parameters, resolution, image size, and quality. Five models, each a 3D nnUNet model, were trained over the course of a thousand epochs using the training data that was split into five folds. On the PICAI dataset with 1500 Patients, it was found that there were no statistically significant differences between the performances of the single model and the ensemble of 5 models. In order to lessen the overall burden and shorten the prediction time, it was decided to employ a single model.
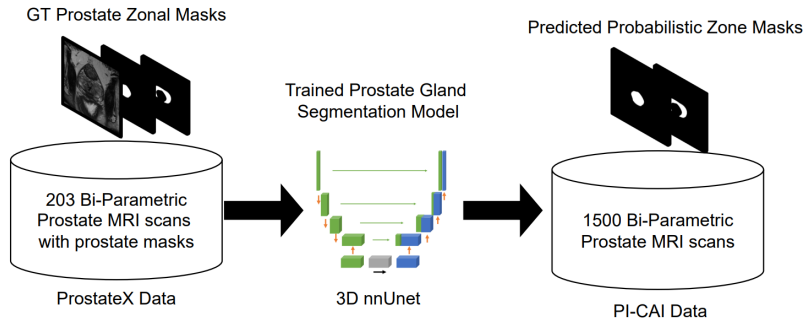

Fig. 1: Prostate gland segmentation model

2.2 Data Preparation

All images were converted from MHA (.mha) to nifti format (.nii.gz) as a preprocessing step for the lesion segmentation / detection models because nnUNet requires nifti images within its own structure. To create images of the same size and resolution, each patient's ADC and DWI images were resampled, cropped, or padded. T2 was used as the primary modality within each patient, and the DWI and ADC images were adjusted to match T2 in terms of size and resolution. ADC images were normalized with complete z-score normalization with respect to the entire dataset while T2 and DWI images were normalized with instance-wise z-score normalization since ADC images are more robust than T2 and DWI images. Lesion segmentation was carried out using the NnUNet framework, but

instead of the default combination of dice and cross entropy loss, Focal Loss and cross entropy loss were employed, as in [2].

2.3   Lesion Detection

To begin with, a nnUNet model was trained using T2, ADC, and DWI images in order to establish a baseline model and track the effects of upcoming techniques on model performance. Throughout each experiment, the train and test splits were kept in order to completely remove the impact of changing data distribution on performance. Despite the dataset being divided into five folds, the initial experiments were only employed on one-fold. The main goal is to use the limited computational resources as effectively as possible while also reducing the workload required to train five models for each new method. Additionally, test-time augmentation was used to improve the model's performance by making predictions on many augmented images. To create a detection map that is required by the PICAI evaluation step, we acquired unique lesion candidates, as done in [2], using the voxel-level confidence maps that were produced. By starting at the voxel with the highest degree of confidence and encompassing all related voxels (in 3D) with at least 40% of the peak's degree of confidence, we specifically produced a lesion candidate. The candidate lesion is then eliminated from the softmax (confidence map), and the procedure is repeated up to the extraction of 5 lesions, if there are still remaining candidates. Small candidates with 10 voxels or fewer (0.009 cm3) are eliminated.
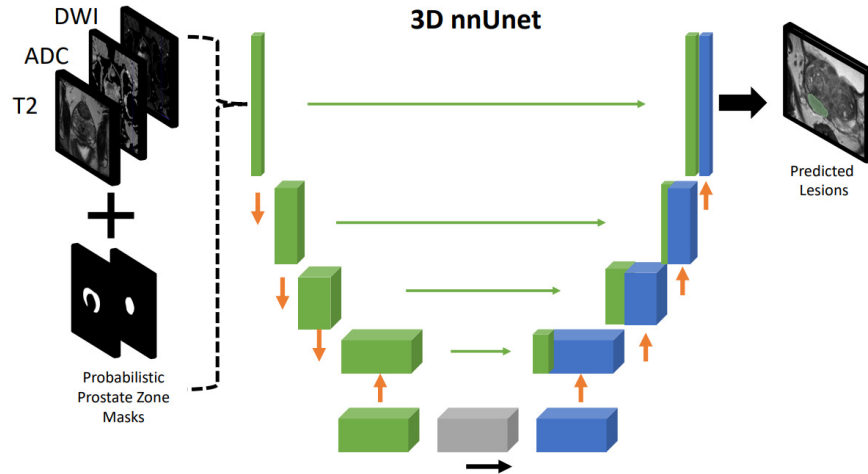

Fig. 2: Prostate lesion segmentation model with prostate zone masks

In the second experiment, the prostate gland segmentation model's output, Peripheral and Central gland predictions (probabilistic), were given to the model

as separate modalities in addition to T2, ADC, and DWI (as detailed in Figure 2). By doing this, the effects of prostate glands on lesion segmentation performance were examined. The third experiment did not involve any model training. Prediction performance of the model that had been trained in the second experiment were evaluated on cropped images. In this experiment we center cropped overly large pictures to remove the tissues that did not contain the prostate area in order to evaluate the effects of these deleted tissues on model performance. The largest and most dislocated prostatic regions are contained inside the cropped images of 81 mm x 192 mm x 192 mm size. In the fourth experiment, models were trained for five folds using T2, ADC, and DWI images, just as in experiment 1. Model performances were evaluated using fivefold cross validation, and an ensemble of five models was tested on 100 patients (PICAI open validation dataset) as detailed in Figure 3. To assess how the models performed on cross validation and as an ensemble of five models on the test set, experiment 5 employed the same setup as experiment 2 for 5 folds. The third experiment did not involve any model training, but center-cropped images were used to evaluate the prediction capacities of models that were trained in experiment 5 (similar to experiment 3).
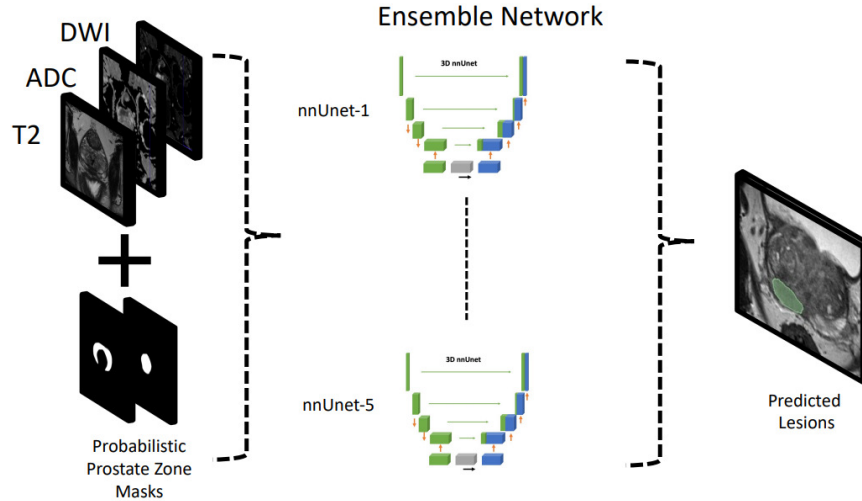

Fig. 3: Ensemble of prostate lesion segmentation models

2.4 False positive elimination with clinical markers

To determine if the provided clinical information is useful for csPCA detection or not, PSAd thresholding was performed in addition to the analysis carried

out in the first six experiments using only images. PSAd was derived by dividing the PSA value from the clinical report by the prostate volume, which was determined using the segmentation model's predictions for the prostate gland. Although the probability of having csPCA is low in those with PSAd below the threshold value of 0.1 in a study [4], the threshold value was determined as 0.04 in this experiment, thus reducing the possibility of false negative predictions. In other words, in Experiment 7, in addition to the setup in Experiment 6, PSAd thresholding was performed and those with PSAd value below 0.04 were directly counted as non csPCA. Another point to emphasize is the intensity distribution of the lesions in the ADC images. As previous studies have shown [10], the ADC intensity distributions of lesions differ between benign and malignant cancers, so this difference can be used as a significant variable for lesion differentiation. In the final experiment, we investigated whether performance could be improved by thresholding the ADC intensity distributions of lesions. First of all, the ADC intensity distributions of the lesions of csPCA patients in the train dataset were examined to find the ideal threshold value. While deciding the threshold value as a result of histogram analysis, it is aimed to prevent possible false negative errors by adding a certain margin value and as a result, the value of 1200 was determined. Since it was not reasonable to use the lowest ADC intensity value in the predicted lesions as the threshold value, the threshold was made by calculating the average of the 10% percentile. According to this threshold value, those with a value above 1200 have been determined as non csPCA for FP elimination, since the probability of malignancy is very low.

3   Results

The prostate gland segmentation model was the first trained model. In this model, a 3D nnUNet structure was implemented. 203 patients from the ProstateX dataset were used to train the 3D nNUnet model in order to get probabilistic prostate gland segmentation. Only the T2 modality was considered in the prostate gland segmentation model since it offered the finest anatomical structure distinction among the T2, ADC, and DWI modalities. Five folds of the ProstateX dataset were created, and a model was trained for each fold. On the PICAI dataset of 1500 patients, prediction was done using a single model and an ensembled model to compare the performances. Since the ground truth prostate gland masks were not available for the 1500 patients of the PICAI dataset, it was not possible to determine the model performances directly; instead, the performances of the single model and the ensembled model were determined based on the differences in the predictions. Dice scores were calculated in order to compare the predictions of the single model and the predictions of the ensembled model. As a result, the dice scores obtained 95.6% for the central zone and 92.4% for the peripheral zone. Since the predictions of both models are fairly similar, it was decided to deploy only one model in order to cut down on workload and prediction time. Because the main objective of this model was to properly seg-

Table 1: Validation scores of the compared models for significant prostate cancer detection. RS: Ranking Score, AUROC: Area Under the Receiver Operating Characteristic Curve, AP: Average Precision, M: Prostate gland mask, C: Cropped prediction

| | Experiments | Validation (100 Patients) | | |
		RS	AUROC	AP
Single Model	**nnUNet semi**	0.734	0.817	0.65
	nnUNet semi, M	0.755	0.851	0.658
	nnUNet semi, M, C	0.77	0.864	0.677
Ensemble Models	**nnUNet semi**	0.712	0.821	0.602
	nnUNet semi, M	0.777	0.87	0.684
	nnUNet semi, M, C	0.807	0.885	0.729
	nnUNet semi, M, C, PSA th.)	0.81	0.888	0.732
	nnUNet semi, M, C, PSA th., ADC th)	0.81	0.888	0.732

ment prostate zones, the loss function was set to Dice + CE loss during model training.

Fivefold dataset splits were predetermined for the experiments, and they persisted throughout all of the experiments. As a result, it was possible to correctly assess how training and postprocessing methods used in studies affected performances. The first trained lesion segmentation model was the 3D nNUnet model created as the base model. In this experiment, which was done as a single model, first fold data was included (1200 training, 300 validation) and T2, ADC and DWI modalities were used. In external validation (PICAI online validation with 100 patients), it received an area under the receiver operating characteristics curve (AUROC) of 0.817, an average precision (AP) of 0.65, and a ranking score (RS) of 0.734, which is the average of AUROC and AP. In the second experiment, the peripheral and central gland masks, which are the outputs of the prostate gland segmentation model, were fed into the model as two separate modalities in addition to T2, ADC, and DWI. In external validation, this model achieved AUROC of 0.851, AP of 0.658, and RS of 0.755. Based on these results, it was concluded that giving prostate masks as separate modalities to the model had a positive effect on performance. The second experiment's trained model was evaluated on cropped images in the third experiment, which did not include any model training. In this experiment, images having a wide field of view (FOV)

were cropped, and it was attempted to eliminate as much of the regions from the images that were unlikely to contain lesions. The results of the validation on external dataset are: AUROC of 0.864, AP of 0.677, and RS of 0.77. Because the model predicted on smaller images rather than extremely broad images, the prediction result on cropped images may have been greater than that of earlier models. The Figure 4 shows a sample image before and after cropping.
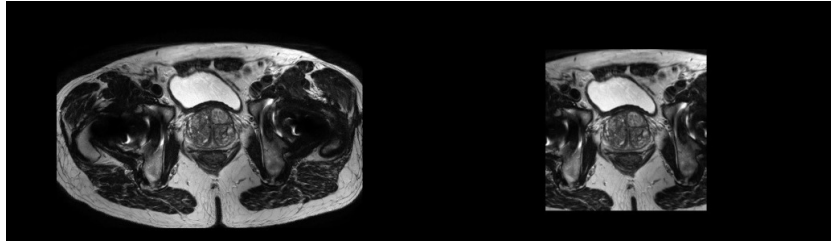

Fig. 4: Sample image before and after crop

After the experiments carried out as a single model, ensemble experiments with 5 models were carried out. These experiments were started in order to prevent possible bias that may occur due to data splitting and to create stronger, more generalizable models by training 5 models for 5 folds by combining their power with ensemble. As a first of the ensemble experiments, 5 models were trained for 5 folds with T2, ADC and DWI modalities, so that a base model was created for the ensemble experiments, and it was possible to compare it with the single model trained with the same setup. The results of the validation on the external dataset are: AUROC of 0.861, AP of 0.602, and RS of 0.712. It was discovered that the single model performed better than the ensembled model when their respective performances were compared. The possible reason behind this was bias due to data distribution. In the next experiment, 5 models were trained for 5 folds with the dataset created by adding peripheral and central zones to T2, ADC and DWI modalities, and the performance of the ensemble model was tested on the external dataset. The resulting metrics are AUROC of 0.87, AP of 0.684 and RS of 0.777. These results highlight that prostate gland segmentations have a beneficial impact on performance in experiments applying ensemble models. Furthermore, the ensemble model outperformed the single model in the experiment with masks whereas the single model outperformed the ensemble model in the experiment without masks. These findings suggest that the performance that prostate gland masks add to the model allows the masked models to avoid biases brought on by distributions. The performance of ensemble models with prostate masks trained in Experiment 5 was assessed on cropped images in Experiment 6. The resulting metrics are RS of 0.807, AUROC of 0.885, and AP of 0.729. Predicting on cropped images has improved performance in ensemble models just as it did in single model.
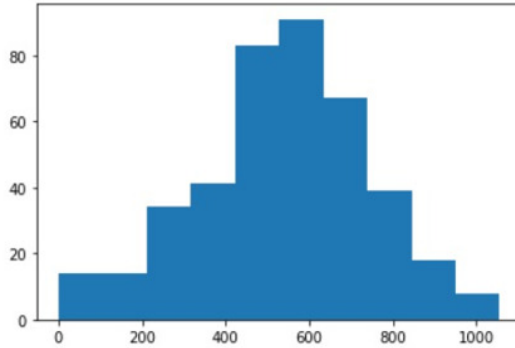
Fig. 5: Histogram of ADC values

PSA thresholding was performed to analyze whether the model performance could be improved by combining the clinical data with the best-performing model obtained as a result of the experiments with DL methods. Although a threshold of 0.1 is considered appropriate in the literature for the differentiation of malignant and benign, in this experiment, 0.04 was chosen as the threshold value to prevent possible false negative predictions, therefore patients below this value were directly considered non-csPCA. The resulting metrics are RS of 0.81, AUROC of 0.888, and AP of 0.732. As the result shows, PSA thresholding did not significantly improve performance, but this was due to the threshold that was kept low to prevent possible false negatives. It should also be noted that validation was performed on a small dataset of 100 patients. In the last experiment, ADC intensity thresholding was applied. First, the ADC intensity distributions of the lesion areas of the patients with csPCA in the train data were found and a histogram was created by taking the average of the 10% percentiles of those distributions. The histogram is as shown in the Figure 5. The threshold value determined according to this histogram is 1200, this is because margin is added to prevent possible false negatives. Patients who exceeded this threshold were directly deemed to not have csPCA. The resulting metrics are RS of 0.81, AUROC of 0.888, and AP of 0.732. From what can be understood, there is no difference between the previous experiment. All results are summarized in Table 1

4   Discussion

We proposed a nnU-net based model for clinically significant prostate cancer detection. We included two key components to our model design: 1) introduction of probabilistic prostate gland segmentations as input for lesion detection, 2) Use of clinical markers such as PSAd and ADC lesion histograms to eliminate false positive predictions. Our proposed framework achieved top performance in open validation phase of PICAI challenge. One limitation of our approach is the use of clinical indices as a post-processing step. We believe utilizing clinically

critical information as a additional input to the model can increase the clinically significant lesion detection similar to the daily clinical practice of the radiologisy. In conclusion, we proposed a framework to detect clinically significant lesions on biparamteric MRI, which can aid in avoiding contrast agent usage in current clinical setups.

Acknowledgements

References

1. "Ahmed, H.U., Bosaily, A.E.S., Brown, L.C., Gabe, R., Kaplan, R., Parmar, M.K., Collaco-Moraes, Y., Ward, K., Hindley, R.G., Freeman, A., others": Diagnostic accuracy of multi-parametric mri and trus biopsy in prostate cancer (promis): a paired validating confirmatory study. The Lancet **389**(10071), 815–822 (2017) 1

2. Bosma, J.S., Saha, A., Hosseinzadeh, M., Slootweg, I., de Rooij, M., Huisman, H.: Report-guided automatic lesion annotation for deep learning-based prostate cancer detection in bpmri. arXiv preprint arXiv:2112.05151 (2021) 4

3. Cuocolo, R., Stanzione, A., Castaldo, A., De Lucia, D.R., Imbriaco, M.: Quality control and whole-gland, zonal and lesion annotations for the prostatex challenge public dataset. European Journal of Radiology **138**, 109647 (2021) 2

4. Girometti, R., Giannarini, G., Panebianco, V., Maresca, S., Cereser, L., De Martino, M., Pizzolitto, S., Pecoraro, M., Ficarra, V., Zuiani, C., et al.: Comparison of different thresholds of psa density for risk stratification of pi-radsv2. 1 categories on prostate mri. The British journal of radiology **95**(1131), 20210886 (2022) 6

5. Heidler, S., Drerup, M., Lusuardi, L., Bannert, U., Bretterbauer, K., Bures, J., Dietersdorfer, F., Dlouhy-Schütz, E., Hessler, C., Karpf, R., et al.: The correlation of prostate volume and prostate-specific antigen levels with positive bacterial prostate tissue cultures. Urology **115**, 151–156 (2018) 2

6. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021) 2

7. van Leeuwen, K.G., Schalekamp, S., Rutten, M.J., van Ginneken, B., de Rooij, M.: Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. European radiology **31**(6), 3797–3804 (2021) 2

8. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al.: International evaluation of an ai system for breast cancer screening. Nature **577**(7788), 89–94 (2020) 2

9. Nordström, T., Akre, O., Aly, M., Grönberg, H., Eklund, M.: Prostate-specific antigen (psa) density in the diagnostic algorithm of prostate cancer. Prostate cancer and prostatic diseases **21**(1), 57–63 (2018) 2

10. Saha, A., Hosseinzadeh, M., Huisman, H.: End-to-end prostate cancer detection in bpmri via 3d cnns: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. Medical image analysis **73**, 102155 (2021) 6

11. Schneider, A.F., Stocker, D., Hötker, A.M., Eberli, D., Rupp, N.J., Donati, O.F., Becker, A.S.: Comparison of psa-density of the transition zone and whole gland for risk stratification of men with suspected prostate cancer: A retrospective mri-cohort study. European Journal of Radiology **120**, 108660 (2019) 2

12. Sonn, G.A., Fan, R.E., Ghanouni, P., Wang, N.N., Brooks, J.D., Loening, A.M., Daniel, B.L., To'o, K.J., Thong, A.E., Leppert, J.T.: Prostate magnetic resonance imaging interpretation varies substantially across radiologists. European urology focus **5**(4), 592–599 (2019) 2

13. Turkbey, B., Rosenkrantz, A.B., Haider, M.A., Padhani, A.R., Villeirs, G., Macura, K.J., Tempany, C.M., Choyke, P.L., Cornud, F., Margolis, D.J., et al.: Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. European urology **76**(3), 340–351 (2019) 1

14. Westphalen, A.C., McCulloch, C.E., Anaokar, J.M., Arora, S., Barashi, N.S., Barentsz, J.O., Bathala, T.K., Bittencourt, L.K., Booker, M.T., Braxton, V.G., 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 **296**(1), 76–84 (2020) 2