Original Article

Development and acceptability validation of a deep learning-based tool for whole-prostate segmentation on multiparametric MRI: a multicenter study

Lili Xu[1,2], Gumuyang Zhang[1], Daming Zhang[1], Jiahui Zhang[1], Xiaoxiao Zhang[1], Xin Bai[1], Li Chen[1], Ru Jin[1], Li Mao[3], Xiuli Li[3], Hao Sun[1,2], Zhengyu Jin[1,2]

[1]Department of Radiology, State Key Laboratory of Complex Severe and Rare Disease, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China; [2]National Center for Quality Control of Radiology, Beijing, China; [3]AI Lab, Deepwise Healthcare, Beijing, China

Contributions: (I) Conception and design: L Xu, H Sun, Z Jin, L Mao, X Li; (II) Administrative support: H Sun, Z Jin, X Li; (III) Provision of study materials or patients: H Sun, Z Jin; (IV) Collection and assembly of data: L Xu, G Zhang, D Zhang, J Zhang, X Zhang, X Bai, L Chen; (V) Data analysis and interpretation: L Xu, L Mao, X Li, H Sun; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Hao Sun; Zhengyu Jin. Department of Radiology, State Key Laboratory of Complex Severe and Rare Disease, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, No. 1 Shuaifuyuan, Wangfujing Street, Dongcheng District, Beijing 100730, China. Email: sunhao_robert@126.com; jinzy@pumch.cn.

Background: Accurate whole prostate segmentation on magnetic resonance imaging (MRI) is important in the management of prostatic diseases. In this multicenter study, we aimed to develop and evaluate a clinically applicable deep learning-based tool for automatic whole prostate segmentation on T2-weighted imaging (T2WI) and diffusion-weighted imaging (DWI).

Methods: In this retrospective study, 3-dimensional (3D) U-Net-based models in the segmentation tool were trained with 223 patients who underwent prostate MRI and subsequent biopsy from 1 hospital and validated in 1 internal testing cohort (n=95) and 3 external testing cohorts: PROSTATEx Challenge for T2WI and DWI (n=141), Tongji Hospital (n=30), and Beijing Hospital for T2WI (n=29). Patients from the latter 2 centers were diagnosed with advanced prostate cancer. The DWI model was further fine-tuned to compensate for the scanner variety in external testing. A quantitative evaluation, including Dice similarity coefficients (DSCs), 95% Hausdorff distance (95HD), and average boundary distance (ABD), and a qualitative analysis were used to evaluate the clinical usefulness.

Results: The segmentation tool showed good performance in the testing cohorts on T2WI (DSC: 0.922 for internal testing and 0.897–0.947 for external testing) and DWI (DSC: 0.914 for internal testing and 0.815 for external testing with fine-tuning). The fine-tuning process significantly improved the DWI model's performance in the external testing dataset (DSC: 0.275 *vs.* 0.815; P<0.01). Across all testing cohorts, the 95HD was <8 mm, and the ABD was <3 mm. The DSCs in the prostate midgland (T2WI: 0.949–0.976; DWI: 0.843–0.942) were significantly higher than those in the apex (T2WI: 0.833–0.926; DWI: 0.755–0.821) and base (T2WI: 0.851–0.922; DWI: 0.810–0.929) (all P values <0.01). The qualitative analysis showed that 98.6% of T2WI and 72.3% of DWI autosegmentation results in the external testing cohort were clinically acceptable.

Conclusions: The 3D U-Net-based segmentation tool can automatically segment the prostate on T2WI with good and robust performance, especially in the prostate midgland. Segmentation on DWI was feasible, but fine-tuning might be needed for different scanners.

Keywords: Deep learning; prostate segmentation; magnetic resonance imaging (MRI); T2-weighted imaging (T2WI); diffusion-weighted imaging (DWI)

Submitted Oct 05, 2022. Accepted for publication Jan 30, 2023. Published online Mar 16, 2023.
doi: 10.21037/qims-22-1068
View this article at: https://dx.doi.org/10.21037/qims-22-1068

Introduction

Accurate prostate whole-gland segmentation on magnetic resonance imaging (MRI) plays an important role in the management of prostate cancer and benign prostate hyperplasia (1,2). For malignant prostate cancer, accurate and efficient prostate contour identification is critical for magnetic resonance (MR) ultrasound fusion biopsy, cancer staging, and radiation therapy (2,3). Apart from providing the prostate contour, autosegmentation of the whole prostate gland can also calculate the prostate volume more efficiently and accurately, which could facilitate the management of benign prostate hyperplasia in surgery planning and treatment response evaluation (4).

Traditionally, the segmentation of the whole prostate gland is performed manually on T2-weighted imaging (T2WI). However, manual segmentation is a time-consuming process that requires the observer to have a good grasp of the anatomy of the prostate gland and its relevant structures. The recent advancements in the application of deep learning have resulted in outstanding achievements in medical imaging analysis. A recently proposed U-Net architecture has been successfully applied to prostate segmentation (5,6). Previous studies focused on the autosegmentation of the whole prostate gland with T2WI have reported mean Dice scores of 0.825 to 0.940 (6-10). However, most of these studies were based on public or single-center datasets. Although several multicenter and multi-MR vendor studies have been carried out to prove the stability and reliability of these convolutional neural networks (CNNs), these studies lack external validation using nonpublic heterogeneous datasets (5,7). Additionally, with the diagnosis and management of prostate malignancies being considered, diffusion-weighted imaging (DWI) is also vital (11,12). Nevertheless, few studies have reported the feasibility of segmenting the whole prostate gland on DWI. Therefore, a preliminary study is needed to demonstrate the feasibility of this approach, which could serve as the foundation for future intraprostatic lesion segmentation. Furthermore, a more important indicator of these segmentation models—the clinical utility—has only been sparsely reported in the literature.

In this study, we aimed to develop a 3-dimensional (3D) U-Net-based segmentation tool for automatic and accurate segmentation of the whole prostate gland on both T2WI and DWI and to verify the tool's performance in heterogeneous multicenter datasets. In addition, we performed a qualitative analysis of the autosegmentation results to assess the models' clinical practicability. We present the following article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-22-1068/rc).

Methods

Datasets

Patients who were treatment-naive and who underwent prostate multiparametric MRI (mpMRI) and subsequent biopsy between November 2014 and December 2018 at Peking Union Medical College Hospital were retrospectively enrolled. Patients were excluded if the image quality was poor and had severe artifacts or if the normal prostate margins were difficult to identify due to extensive tumor invasion. A total of 318 patients were finally included. These patients were divided into a training-validation cohort (n=223; n=178 in the training group and n=45 in the validation group, in each fold of cross-validation) and an independent internal testing cohort (n=95) based on their MR examination times.

Additionally, to fully verify the performance of the models in different groups, 3 external testing cohorts (1 public and 2 private) were employed in this study. The public external testing cohort used the testing group of the PROSTATEx Challenge available from *The Cancer Imaging Archive* (PX_{test} cohort, n=141), in which the task was to predict the clinical significance of prostate lesions found in MRI (13,14). Another 2 external testing cohorts included patients from different nonpublic centers with various vendors. These patients received a diagnosis of advanced prostate cancer and were candidates for androgen-deprivation treatment. After patients were excluded following the same criteria mentioned above, the final datasets included 30 patients from Tongji Hospital (TJH cohort) and 29 patients from Beijing Hospital (BJH cohort). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and approved by the institutional review board of Peking Union Medical

College Hospital (No. K22C1922). Individual consent for this retrospective analysis was waived. Figure S1 shows the dataset selection process of this study.

Prostate MR image acquisition and ground truth segmentation

For patients from our institution, a 3.0 T MR scanner (Discovery 750, GE Healthcare, Milwaukee, WI, USA) was used. For patients from the PROSTATEx dataset, 2 different types of Siemens 3.0 T MR scanners, the Magnetom Trio and Skyra, were used. The MR vendors for the TJH cohort and the BJH cohort included Siemens, GE, and Philips, with different magnetic field strengths (1.5 T and 3.0 T). The detailed MRI acquisition parameters of each group are presented in Table S1. Our institution followed the recommendation of Prostate Imaging-Reporting and Data System version 2 (PI-RADS v.2) and acquired DWI images with b values of 0, 100, 150, 200, 500, 800, 1,000, 1,500, and 2,000 s/mm^2. However, in the external testing group (PX_{test} cohort), only 3 b values were acquired (50, 400, and 800 s/mm^2). On the other hand, benign prostate signal suppression increased as the b value increased, which would limit the reorganization of the whole prostate gland (15). Therefore, we chose DWI images with b value =800 s/mm² for analysis in this study.

The patients' axial T2WI and DWI images were collected and manually segmented by 1 radiologist (>1,000 prostate MR images interpreted) to serve as the ground truth. Another senior radiologist (>3,000 prostate MR images interpreted) reviewed the images and modified the contour if necessary. Manual segmentation on MR images was performed on the Deepwise Research Platform (Deepwise Healthcare; http://label.deepwise.com). In the manual segmentation process, the radiologist who was blinded to the automatic segment carefully segmented the prostate nature contour to serve as the ground truth, which meant that when the tumor was beyond the prostate capsule, this portion would not be included in the segmentation. For the TJH cohort and BJH cohort, only T2WI images were collected for testing, and because of the retrospective collecting manner, these patients' DWI images were unavailable.

Segmentation process for the whole prostate gland

Preprocessing

Before the model was trained, the axial images were resampled to a uniform pixel spacing, the median of the pixel spacing of the training cohort, to offset the bias caused by resolution inconsistency (T2WI images: 3.00×0.51×0.51; DWI images: 3.00×1.41×1.41). The input patch size of images was the average nonzero area of all images (T2WI images: 14×352×352; DWI images: 20×174×250). Then, the intensity of the images was normalized using z score normalization.

CNN

The proposed segmentation tool contained 2 CNN models for T2WI and DWI. 3D U-Net-based segmentation models were trained with a self-configuring nnU-Net (16), in which the model structure and the optimal hyperparameters were obtained automatically. For every pixel of an input image, the model outputs a likelihood estimate for being part of the whole prostate gland. The overall framework of the model training procedure is shown in *Figure 1*. The model's structure is elaborated in Appendix 1 and Table S2.

To avoid overfitting, data augmentation methods, including mirroring, scaling, rotation, and translation, were applied. The training was performed using 5-fold cross-validation with 500 epochs for each run and an initial learning rate of 0.01. The batch size of 2 was calculated using the nnU-Net framework based on the graphics processing unit (GPU) memory and the number of parameters of the model. The loss function was the combination of the Dice loss and the binary cross entropy (BCE) loss.

In the inference stage, the ensemble results of the 5 models trained in the 5-fold cross-validation procedure were used as the mode prediction, and the postprocessing was applied to further refine the segmentation results. In the postprocessing procedure, the regions other than the maximum connected domain were filtered out to reduce the false-positive identification of extraprostatic organs or tissues.

Fine-tuning process

Previous studies reported a significantly inferior performance of CNN models for autosegmentation on DWI images in external datasets with different MR vendors (17,18). With the concern of the inferiority of our DWI model in external testing, the fine-tuning method was set as an alternative compensation method for the DWI model in our study. Furthermore, the training group of PROSTATEx (PX_{train} cohort, n=203) was collected for fine-tuning. The original training cohort and the fine-tuning cohort were combined to train another 50 epochs, with an initial learning rate of 0.0001. This procedure is seen as an upper

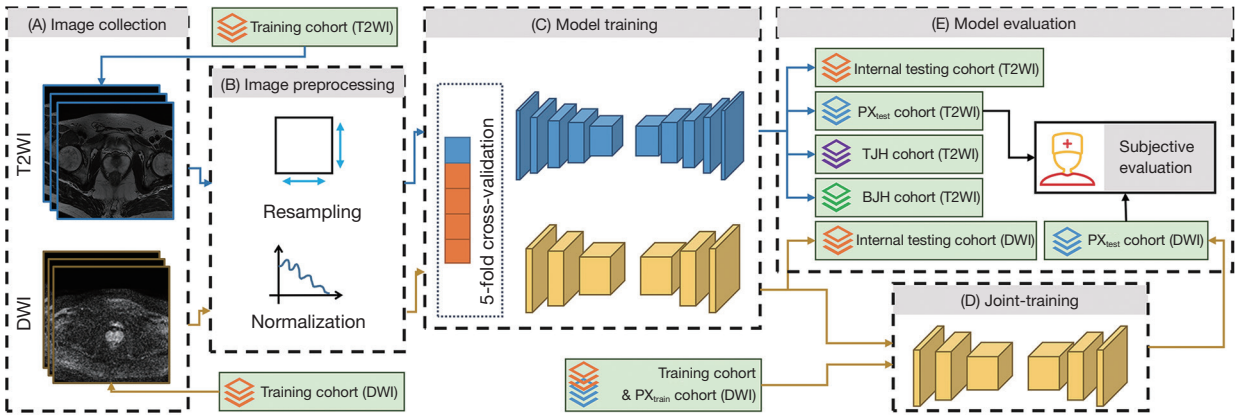

Figure 1 A schematic illustration of the model training and evaluation process of this study. (A) Image collection. T2WI and DWI were included in the study. (B) Image preprocessing. The preprocessing procedure was composed of resampling and normalization. (C) Model training. The T2WI and DWI models were trained using a 5-fold cross-validation procedure. (D) Joint training. The DWI model was further trained through the combination of a training cohort and the PX_{train} cohort. (E) Model evaluation. Our models were evaluated using 1 internal testing cohort and 3 external testing cohorts (PX_{test} cohort, TJH cohort, and BJH cohort), and subjective evaluation was also performed. T2WI, T2-weighted images; DWI, diffusion-weighted images; PX_{train}, the training group of the PROSTATEx Challenge dataset; PX_{test} cohort, testing group of the PROSTATEx Challenge dataset; TJH cohort, Tongji Hospital cohort; BJH cohort, Beijing Hospital cohort.

bound on the possible performance of domain adaption (19). More details about the fine-tuning process are presented in Appendix 2.

Quantitative analysis

We calculated the Dice similarity coefficient (DSC), 95% Hausdorff distance (95HD), and average boundary distance (ABD) to evaluate the performance of our 3D U-Net models. The DSC is widely used to quantify the spatial overlap between segmentations, with its value ranging from 0 (indicating no overlap) to 1 (indicating perfect overlap) (20). 95HD and ABD are commonly used to evaluate the boundary errors of segmentations. To investigate how the prostate morphology would affect the model's performance, we calculated the DSC in the apex, midgland, and base of the prostate.

Qualitative analysis

To evaluate the utility of CNN segmentation results in clinical practice, qualitative analysis was performed in the internal testing cohort and PX_{test} cohort. The radiologist (>1,000 prostate MR images interpreted) scored the CNN segmentation results on both T2WI and DWI using the following 5-point scoring system: 5, perfect; 4, acceptable; 3, unacceptable with moderate corrections required; 2, unacceptable with major corrections required; and 1, reject. The soring details are presented in *Table 1*. CNN segmentation results with scores ≥4 were considered clinically acceptable.

Statistical analysis

The software used for data analysis was R version 4.2.0 (The R Foundation for Statistical Computing, https://www.r-project.org) and PyTorch version 1.6.0 (https://pytorch.org/). The major component of our code is available in open-source repositories (nnU-Net: available from https://github.com/MIC-DKFZ/nnUNet). The patients' characteristics and the metrics' distributions are described as the median and interquartile range (IQR) or mean and standard deviation for quantitative characteristics and by the frequency and percentage for qualitative characteristics. The quantitative segmentation performance difference between before and after the fine-tuning process and among the different parts of the prostate was compared using a paired Wilcoxon test. The P values were corrected with the Bonferroni-Holm procedure for the number of combinations. P values <0.05 were considered statistically

Table 1 Criteria for qualitative analysis

Score	Criteria
1	Reject: adjusting the CNN segmentation would take more time than manual segmentation, and manual resegmentation is preferred
2	Unacceptable with major corrections required: the time needed for adjusting CNN segmentation and manual resegmentation is comparable
3	Unacceptable with moderate corrections required: manual adjustment would be faster than manual segmentation but would require several minutes
4	Acceptable: minor corrections are needed (corrections are expected to be completed within 1 minute). Adjusting the CNN segmentation consumes substantially less time than does manual segmentation
5	Perfect: no corrections or slight adjustments are needed (it might require a few seconds). The CNN segmentation is ready for clinical use

CNN, convolutional neural network.

Table 2 Clinicopathological characteristics of patients in the nonpublic datasets

Variable	Training cohort (n=223)	Internal testing cohort (n=95)	TJH cohort (n=30)	BJH cohort (n=29)
Age (years)[†]	65±8	66±8	67±8	71±8
Prostate volume (mL)[‡,§]	49.1 [34.0–70.3]	41.6 [28.3–58.8]	46.2 [38.0–68.8]	43.0 [31.1–66.5]
PI-RADS[¶]				
1–2	28.3 (63/223)	15.8 (15/95)	0.0 (0/30)	0.0 (0/29)
3	9.0 (20/223)	9.5 (9/95)	0.0 (0/30)	24.1 (7/29)
4	29.1 (65/223)	50.5 (48/95)	3.3 (1/30)	27.6 (8/29)
5	33.6 (75/223)	24.2 (23/95)	96.7 (29/30)	48.3 (14/29)
Nonprostate cancer[¶]	53.4 (119/223)	29.5 (28/95)	0.0 (0/30)	0.0 (0/29)
Prostate cancer[¶]	46.6 (104/223)	70.5 (67/95)	100.0 (30/30)	100.0 (29/29)
Clinical T stage[¶]				
T1–2	59.6 (62/104)	70.1 (47/67)	0.0 (0/30)	20.7 (6/29)
T3	38.5 (40/104)	26.9 (18/67)	40.0 (12/30)	37.9 (11/29)
T4	1.9 (2/104)	3.0 (2/67)	60.0 (18/30)	41.4 (12/29)

[†], mean ± standard deviation; [‡], median [interquartile range]; [§], volume was calculated from manual segmentation results; [¶], % (n/N). TJH cohort, Tongji Hospital cohort; BJH cohort, Beijing Hospital cohort; PI-RADS, Prostate Imaging-Reporting and Data System.

significant.

Results

Patients' demographic characteristics

The clinicopathological data for patients in nonpublic datasets are summarized in *Table 2*. The mean age in the training cohort, internal testing cohort, TJH cohort, and BJH cohort was 65±8, 66±8, 67±8, and 71±8 years, respectively, with a median prostate volume of 49.1 (IQR, 34.0–70.3), 41.6 (IQR, 28.3–58.8), 46.2 (IQR, 38.0–68.8), and 43.0 (IQR, 31.1–66.5) mL, respectively.

Model segmentation performance for the whole prostate gland

The DSC of the 3D U-Net model on T2WI images reached 0.922±0.033 in the internal testing cohort, with a

Table 3 Dice similarity coefficient, 95% Hausdorff distance, and average boundary distance of the 3-dimensional U-Net model using T2-weighted and diffusion-weighted images

Sequences	Cohort	DSC	95HD (mm)	ABD (mm)
T2-weighted imaging	Training cohort (n=223)	0.989±0.002 (0.988–0.989)	0.469±0.124 (0.452–0.485)	0.048±0.021 (0.045–0.051)
	Internal testing cohort (n=95)	0.922±0.033 (0.915–0.929)	4.024±3.224 (3.363–4.684)	0.681±0.446 (0.590–0.773)
	PX_{test} cohort (n=141)	0.897±0.051 (0.889–0.906)	4.719±2.768 (4.257–5.182)	1.093±0.697 (0.976–1.209)
	TJH cohort (n=30)	0.902±0.134 (0.851–0.953)	4.447±4.046 (2.910–5.983)	0.948±1.188 (0.497–1.399)
	BJH cohort (n=29)	0.947±0.032 (0.934–0.959)	3.481±2.803 (2.396–4.566)	0.468±0.359 (0.329–0.607)
Diffusion-weighted imaging	Training cohort (n=223)	0.952±0.011 (0.951–0.954)	0.570±0.406 (0.516–0.624)	0.080±0.038 (0.075–0.085)
	Internal testing cohort (n=95)	0.914±0.045 (0.905–0.923)	1.510±1.999 (1.100–1.919)	0.243±0.276 (0.187–0.300)
	PX_{test} cohort (n=141)†	0.275±0.371 (0.213–0.337)	50.823±45.858 (42.988–58.658)	33.385±36.413 (27.163–39.606)
	PX_{test} cohort (n=141)‡	0.815±0.155*** (0.789–0.841)	7.411±13.131*** (5.217–9.605)	2.102±8.264*** (0.721–3.482)

Data are presented as mean ± standard deviation (95% confidence intervals). ***, P<0.001 when compared with the results before fine-tuning; †, the results before fine-tuning; ‡, the results after fine-tuning. DSC, Dice similarity coefficient; 95HD, 95% Hausdorff distance; ABD, average boundary distance; PX_{test} cohort, testing group of the PROSTATEx Challenge dataset; TJH cohort, Tongji Hospital cohort; BJH cohort, Beijing Hospital cohort.

95HD of 4.024±3.224 mm and an ABD of 0.681±0.446 mm. The average DSCs in the 3 external testing groups were all above 0.890, with a range of 0.897–0.947 (*Table 3*).

As for the DWI segmentation model, its performance is also presented in *Table 3*. The DSC reached 0.914±0.045 for the internal testing cohort but declined sharply to 0.275 for the external testing cohort. To clarify the underlying reasons for this finding, the influencing factor analysis in the external testing cohort is shown in Appendix 3 and Table S3. The results showed that scanning parameters, including repetition time, MR scanner, echo train length, and slice thickness, were significant influencing factors for the DWI model in the external testing cohort (all P values <0.05).

After the fine-tuning procedure, the model reached a DSC of 0.815±0.155 for the PX_{test} cohort, and its 95HD and ABD were 7.411±13.131 and 2.102±8.264 mm, respectively. All these parameters were significantly improved compared to the parameters before fine-tuning (all P values <0.001).

Model segmentation performance for the apex, midgland, and base of the prostate

The segmentation results in the apex, midgland, and base of the prostate in both the internal and external testing cohorts are presented in *Figure 2*. In general, the segmentation performance was the best in the midgland of the prostate, followed by the base and the apex.

For the T2WI model, the DSCs in the midgland of the prostate were 0.949–0.976 in the testing cohorts, while the DSCs in the apex (0.833–0.926) and base (0.851–0.922) were significantly lower (all P values <0.01). Using DWI images, the DSCs in the midgland of the prostate were also higher (0.843–0.942) than in the apex (0.755–0.821) and base (0.810–0.929) in the testing cohorts (P<0.01).

Qualitative analysis of segmentation performance

The qualitative analysis results for the internal testing cohort and PX_{test} cohort are presented in *Table 4*. In the internal testing cohort, the subjective scores were greater than or equal to 4 in 100% of cases for T2WI (92.6% with a score of 5 and 7.4% with a score of 4), and the scores were greater than or equal to 4 in 97.9% of cases for DWI (89.5% with a score of 5 and 8.4% with a score of 4). In the PX_{test} cohort, for T2WI, 79.4% of cases were rated as a score of 5, 19.1% of cases were rated as a score of 4, and only 2 cases (1.4%) were rated as a score of 3. Therefore, 98.6% of segmentation results on T2WI were considered clinically acceptable. For DWI results, most cases (102/141, 72.3%) reached satisfying subjective scores. Specifically, 51.8% of cases were rated as a score of 5, and 20.6% of cases were rated as a score of 4. *Figure 3* shows 2 examples of the CNN segmentation results from the PX_{test} cohort with corresponding DSC and subjective scores to illustrate the models' clinical utility with both quantitative and qualitative analyse.

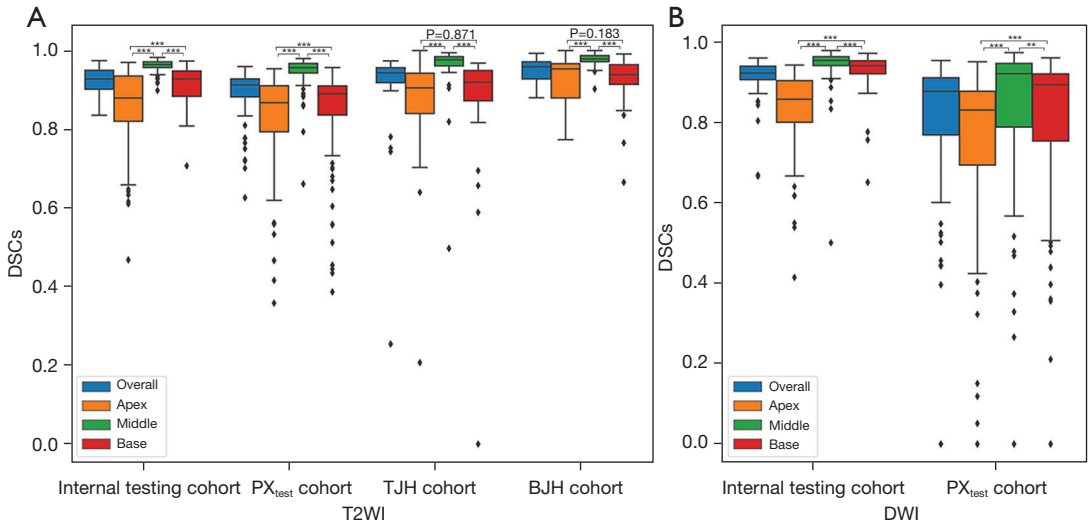

Figure 2 Boxplot of the distribution of DSCs. (A) The DSCs of the overall prostate gland, the apex, midgland, and base of the prostate in the internal testing cohort, the PX_{test} cohort, the TJH cohort, and the BJH cohort using T2WI. (B) The DSCs of the overall prostate gland, the apex, midgland, and base of the prostate in the internal testing cohort and the PX_{test} cohort using DWI. **, P<0.01; ***, P<0.001. DSC, Dice similarity coefficient; T2WI, T2-weighted images; DWI, diffusion-weighted images; PX_{test} cohort, testing group of the PROSTATEx Challenge dataset; TJH, Tongji Hospital; BJH, Beijing Hospital.

Table 4 Qualitative analysis of the segmentation results in the internal testing cohort and PXtest cohort using T2-weighted and diffusion-weighted images

Cohorts	Cases	Score				
		5	4	3	2	1
Internal testing cohort	T2WI	92.6 (88/95)	7.4 (7/95)	0.0 (0/95)	0.0 (0/95)	0.0 (0/95)
	DWI	89.5 (85/95)	8.4 (8/95)	2.1 (2/95)	0.0 (0/95)	0.0 (0/95)
PX_{test} cohort	T2WI	79.4 (112/141)	19.1 (27/141)	1.4 (2/141)	0.0 (0/141)	0.0 (0/141)
	DWI	51.8 (73/141)	20.6 (29/141)	9.9 (14/141)	8.5 (12/141)	9.2 (13/141)

Data are presented as % (n/N). PX_{test} cohort, testing group of the PROSTATEx Challenge dataset; T2WI, T2-weighted imaging; DWI, diffusion-weighted imaging.

Discussion

In this study, we developed a 3D U-Net-based segmentation tool to automatically segment the whole prostate gland on T2WI and DWI images. We tested the tool's reliability using various internal and external testing groups. In general, the model's performance for segmenting the whole prostate gland on T2WI images was outstanding and remained stable across different vendors with various magnetic field strengths and scanning parameters, especially in the midgland of the prostate. The segmentation performance on DWI images was also excellent in the internal testing group but was impaired by the scanning parameter variance in the external testing group. Nevertheless, the fine-tuning method was demonstrated to be useful in solving this problem and improving the model's performance. Apart from this quantitative analysis, our study also performed a qualitative analysis of the autosegmentation results, which supported the clinical utility of the automated CNN segmentation results.

Numerous studies have proposed specialized architectures and training scheme modifications to achieve competitive segmentation (21-23), but it remains difficult to improve upon the basic U-Net if the corresponding training procedure is designed adequately (16). Thus, U-Net

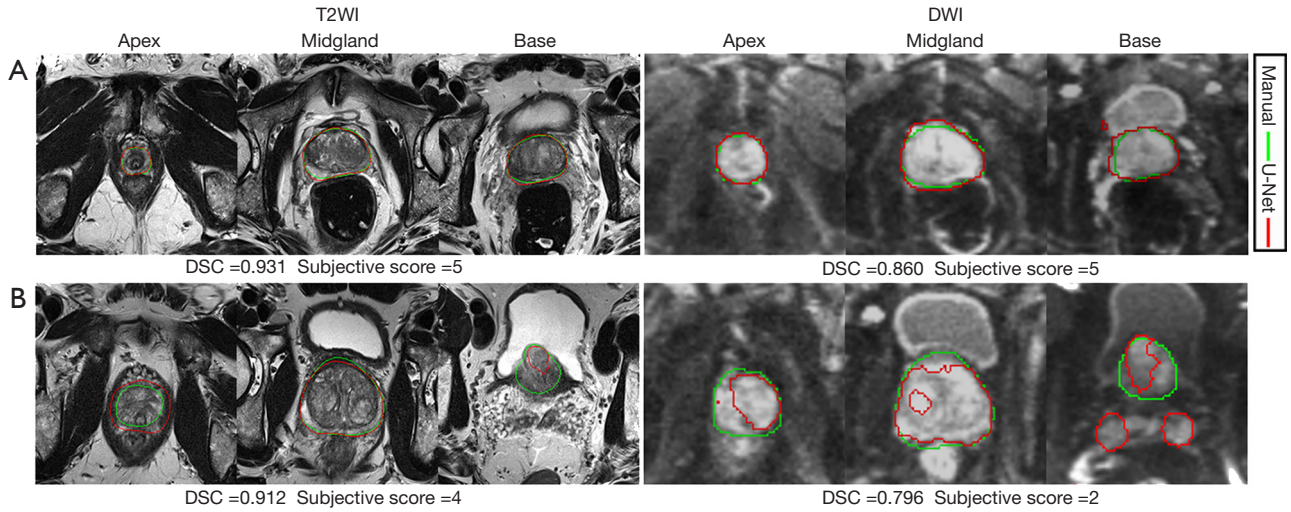

Figure 3 Examples of segmentation results. (A) In this case, the 3-dimensional U-Net model (red line) showed a good overlap with the manual segmentation (green line) in the apex, midgland, and base (from left to right) of the prostate on both T2WI and DWI, with an overall DSC of 0.931 for T2WI and 0.860 for DWI. The subjective scores for both T2WI and DWI segmentation results were 5. (B) In this case, the overall DSC was 0.912 on T2WI and 0.796 for DWI. The subjective score for T2WI was 4, which required minor modification in the apex and base of the prostate. As for DWI images, the model showed a poor overlap with the manual segmentation, with a subjective score of 2. T2WI, T2-weighted images; DWI, diffusion-weighted images; DSC, Dice similarity coefficient.

was chosen as the foundation of our model. Studies have reported DSCs ranging from 0.825 to 0.940 for CNN-based autosegmentation on T2WI (10). However, performances in nonpublic heterogenous external testing datasets have rarely been reported. Zavala-Romero *et al.* (7) developed a 3D U-Net-based segmentation algorithm for prostate segmentation using T2WI across multiple MRI vendors and demonstrated that training with heterogeneous data could achieve a more stable performance. Soerensen *et al.* (5) used multicenter images to train their deep learning model, ProGNet, for prostate whole gland segmentation, which showed good performance in public datasets for external validation. Comparatively, our segmentation models, which were trained using images from 1 MR scanner, also showed stable and robust performance in various external testing cohorts regardless of the differences in MR scanners, magnetic field strengths, and MR acquisition parameters. In addition, even though the 2 nonpublic external testing cohorts contained more advanced-stage patients and shared different patient characteristics with the training group, our T2WI model also showed good performance in these datasets. The possible reasons for the excellent performance of our model on T2WI could be the following: first, the scanning parameters of T2WI in the training cohort were heterogeneous, and the number of patients was sufficient, which would make our model well adapted to various cohorts; second, the training procedure in the nnU-Net framework was designed adequately, a large variety of data augmentation techniques were adopted to prevent overfitting, and the ensemble method was used to improve the robustness.

DWI is indispensable for evaluating prostate malignancies according the PI-RADS. In addition, DWI is considered valuable in the dose escalation to the dominant intraprostatic lesions in radiotherapy (11). Accurate whole prostate gland segmentation on DWI could serve as a foundation for future intraprostatic lesion segmentation on the same images. However, the feasibility has not been demonstrated in previous studies. Some studies have used DWI as a complementary sequence and built CNN models for the combination of T2WI and DWI (10,24,25). Clark *et al.* (26) designed a model for segmenting DWI images (b=0 s/mm^2) of the prostate and found good training results. In general, DWI images with a b value =0 s/mm^2 share a great similarity with T2WI, and provide no additional information for the diagnosis, while DWI images with higher b values are more frequently used in clinical studies. We trained our model to segment the whole prostate gland on DWI images with b=800 s/mm^2 and found excellent performance in the internal testing cohort. In addition, we

tested the model using an external cohort. Although the segmentation performance sharply decreased when the model was directly applied to the external testing cohort with different MR scanners, this finding was consistent with previous studies that segmented acute ischemic lesions and lymph nodes on DWI (17,18). The possible reason for the decline of the DWI model's performance in the external testing cohort could be the scanning parameter difference between these images. The multivariate logistic regression analysis showed that the repetition time, MR scanner, echo train length, and slice thickness were significant influencing factors for autosegmentation in the heterogeneous external testing group. Although increasing the cohort size and heterogeneity of the training group is a possible solution, the fine-tuning procedure—which was demonstrated to be useful in our study—could be a more feasible method to make up for this deficiency.

We also analyzed the model's performance in different anatomical structures of the prostate: the apex, midgland, and base. The results showed that this model performed the best in the midgland of the prostate, followed by the base and the apex. Segmentation difficulties in the apex and base of the prostate were also reported by other studies. Nai *et al.* (10) reported the highest segmenting DSC in the midgland of the prostate, but the DSC in the base was quite low, with a value of 0.576. Montagne *et al.* (27) found radiologists' manual segmentation of the whole prostate gland showed a significantly higher DSC for the midgland (0.95) and a lower DSC for the apex (0.90) and the base (0.87). In the apex and base of the prostate, the prostate outlines are difficult to differentiate from adjacent structures, which increases the difficulty for automatic segmentation algorithms. Further improvements of the model should be made in the segmentation of the apex and base of the prostate.

At present, most CNN-based segmentation models use quantitative methods for model evaluation. Nevertheless, integrating models into clinical practice and analyzing their utility is also important for clinicians. Soerensen *et al.* (5) tested their model in a prospective cohort and integrated their deep learning model into the clinical workflow as part of fusion biopsy. This model outperformed radiology technicians for segmenting the whole prostate gland and took significantly less time (35 seconds/case *vs.* 10 minutes/case) to yield a clinically useable segmentation file. However, their prospective cohort was quite small, with only 11 patients. To evaluate the clinical utility of our model, we proposed a 5-point subjective scoring system based on that of Tang *et al.* (28). This scoring system's criteria were mainly based on the comparison of the estimated time consumption for modification and manual resegmentation. Autosegmentation results that required minor or even no corrections were considered acceptable. Subjective analysis showed most segmentation results on T2WI images were clinically acceptable, and under DWI, most cases were considered clinically acceptable. These automatic segmentation results with high scores were considered ready to be used in clinical scenarios with minor corrections or even without modification.

There were some limitations to this study. First, although we analyzed the model's performance on DWI images, we were limited by the number of external testing cohorts. A thorough analysis of related influence factors and further improvements is needed. Second, although a qualitative analysis was performed to verify the models' clinical practicability, future work integrating them into clinical practice and analyzing the performance is needed. Finally, the acquisition process for manually segmented ground truth data was time-consuming. Considering the outstanding performance of our present model, using autosegmentation with subsequent radiologists' manual adjustment to serve as the ground truth is feasible in future studies. Doing so could save time generating a ground truth dataset and make enlarging the training cohort possible.

Conclusions

We found that the 3D U-Net-based segmentation tool could automatically segment the prostate using T2WI images with a good and robust performance in all testing datasets. Autosegmentation on DWI images was also feasible, while fine-tuning might be needed in external testing groups with heterogenous MR scanning parameters. Regardless of the MR sequences, the segmentation results for the midgland of the prostate were consistently better than for the other parts. These clinically acceptable autosegmentation results can help the management of prostate diseases in the future.

Acknowledgments

Funding: This study was supported by the CAMS Innovation Fund for Medical Sciences (CIFMS; grant No. 2022-I2M-C&T-B-019), the National High-Level Hospital Clinical Research Funding (grant Nos. 2022-PUMCH-A-033, 2022-PUMCH-A-035, and 2022-PUMCH-B-069), and the

2021 Key Clinical Specialty Program of Beijing.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-22-1068/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-1068/coif). LM and XL are employees of Deepwise Healthcare, which is not related to the current study. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the institutional review board of Peking Union Medical College Hospital (No. K22C1922), and individual consent for this retrospective analysis was waived.

References

1. Ghose S, Oliver A, Martí R, Lladó X, Vilanova JC, Freixenet J, Mitra J, Sidibé D, Meriaudeau F. A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images. Comput Methods Programs Biomed 2012;108:262-87.
2. Almeida G, Tavares JMRS. Deep Learning in Radiation Oncology Treatment Planning for Prostate Cancer: A Systematic Review. J Med Syst 2020;44:179.
3. Sonn GA, Margolis DJ, Marks LS. Target detection: magnetic resonance imaging-ultrasound fusion-guided prostate biopsy. Urol Oncol 2014;32:903-11.
4. Diaz TA, Benson B, Clinkenbeard A, Long JR, Kawashima A, Yano M. MRI Evaluation of Patients Before and After Interventions for Benign Prostatic Hyperplasia: An Update. AJR Am J Roentgenol 2022;218:88-99.
5. Soerensen SJC, Fan RE, Seetharaman A, Chen L, Shao W, Bhattacharya I, Kim YH, Sood R, Borre M, Chung BI, To'o KJ, Rusu M, Sonn GA. Deep Learning Improves Speed and Accuracy of Prostate Gland Segmentations on Magnetic Resonance Imaging for Targeted Biopsy. J Urol 2021;206:604-12.
6. Ushinsky A, Bardis M, Glavis-Bloom J, Uchio E, Chantaduly C, Nguyentat M, Chow D, Chang PD, Houshyar R. A 3D-2D Hybrid U-Net Convolutional Neural Network Approach to Prostate Organ Segmentation of Multiparametric MRI. AJR Am J Roentgenol 2021;216:111-6.
7. Zavala-Romero O, Breto AL, Xu IR, Chang YC, Gautney N, Dal Pra A, Abramowitz MC, Pollack A, Stoyanova R. Segmentation of prostate and prostate zones using deep learning : A multi-MRI vendor analysis. Strahlenther Onkol 2020;196:932-42.
8. Cuocolo R, Comelli A, Stefano A, Benfante V, Dahiya N, Stanzione A, Castaldo A, De Lucia DR, Yezzi A, Imbriaco M. Deep Learning Whole-Gland and Zonal Prostate Segmentation on a Public MRI Dataset. J Magn Reson Imaging 2021;54:452-9.
9. Bardis M, Houshyar R, Chantaduly C, Tran-Harding K, Ushinsky A, Chahine C, Rupasinghe M, Chow D, Chang P. Segmentation of the Prostate Transition Zone and Peripheral Zone on MR Images with Deep Learning. Radiol Imaging Cancer 2021;3:e200024.
10. Nai YH, Teo BW, Tan NL, Chua KYW, Wong CK, O'Doherty S, Stephenson MC, Schaefferkoetter J, Thian YL, Chiong E, Reilhac A. Evaluation of Multimodal Algorithms for the Segmentation of Multiparametric MRI Prostate Images. Comput Math Methods Med 2020;2020:8861035.
11. Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, Tempany CM, Choyke PL, Cornud F, Margolis DJ, Thoeny HC, Verma S, Barentsz J, Weinreb JC. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. Eur Urol 2019;76:340-51.
12. Yuan J, Poon DMC, Lo G, Wong OL, Cheung KY, Yu SK. A narrative review of MRI acquisition for MR-guided-radiotherapy in prostate cancer. Quant Imaging Med Surg 2022;12:1585-607.
13. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P,

Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26:1045-57.
14. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. ProstateX Challenge data. The Cancer Imaging Archive 2017. doi: 10.7937/K9TCIA.2017.MURS5CL.
15. Rosenkrantz AB, Parikh N, Kierans AS, Kong MX, Babb JS, Taneja SS, Ream JM. Prostate Cancer Detection Using Computed Very High b-value Diffusion-weighted Imaging: How High Should We Go? Acad Radiol 2016;23:704-11.
16. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203-11.
17. Zhao X, Xie P, Wang M, Li W, Pickhardt PJ, Xia W, Xiong F, Zhang R, Xie Y, Jian J, Bai H, Ni C, Gu J, Yu T, Tang Y, Gao X, Meng X. Deep learning-based fully automated detection and segmentation of lymph nodes on multiparametric-mri for rectal cancer: A multicentre study. EBioMedicine 2020;56:102780.
18. Alis D, Yergin M, Alis C, Topel C, Asmakutlu O, Bagcilar O, Senli YD, Ustundag A, Salt V, Dogan SN, Velioglu M, Selcuk HH, Kara B, Oksuz I, Kizilkilic O, Karaarslan E. Inter-vendor performance of deep learning in segmenting acute ischemic lesions on diffusion-weighted imaging: a multicenter study. Sci Rep 2021;11:12434.
19. Li K, Yu L, Heng PA. Domain-incremental Cardiac Image Segmentation with Style-oriented Replay and Domain-sensitive Feature Whitening. IEEE Trans Med Imaging 2022. [Epub ahead of print]. doi: 10.1109/TMI.2022.3211195.
20. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 1945;26:297-302.
21. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018) 2018;11045:3-11.
22. Siddique N, Paheding S, Alom MZ, Devabhaktuni V. Recurrent residual U-Net with EfficientNet encoder for medical image segmentation. Pattern Recognition and Tracking XXXII 2021;2021;19. doi: 10.1117/12.2591343.
23. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge. In: Crimi A, Bakas S, Kuijf H, Menze B, Reyes M. editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries 2018; Springer International Publishing; 2018;287-97.
24. Lai CC, Wang HK, Wang FN, Peng YC, Lin TP, Peng HH, Shen SH. Autosegmentation of Prostate Zones and Cancer Regions from Biparametric Magnetic Resonance Images by Using Deep-Learning-Based Neural Networks. Sensors (Basel) 2021;21:2709.
25. Rezaeijo SM, Jafarpoor Nesheli S, Fatan Serj M, Tahmasebi Birgani MJ. Segmentation of the prostate, its zones, anterior fibromuscular stroma, and urethra on the MRIs and multimodality image fusion using U-Net model. Quant Imaging Med Surg 2022;12:4786-804.
26. Clark T, Zhang J, Baig S, Wong A, Haider MA, Khalvati F. Fully automated segmentation of prostate whole gland and transition zone in diffusion-weighted MRI using convolutional neural networks. J Med Imaging (Bellingham) 2017;4:041307.
27. Montagne S, Hamzaoui D, Allera A, Ezziane M, Luzurier A, Quint R, Kalai M, Ayache N, Delingette H, Renard-Penna R. Challenge of prostate MRI segmentation on T2-weighted images: inter-observer variability and impact of prostate morphology. Insights Imaging 2021;12:71.
28. Tang X, Jafargholi Rangraz E, Coudyzer W, Bertels J, Robben D, Schramm G, Deckers W, Maleux G, Baete K, Verslype C, Gooding MJ, Deroose CM, Nuyts J. Whole liver segmentation based on deep learning and manual adjustment for clinical use in SIRT. Eur J Nucl Med Mol Imaging 2020;47:2742-52.

Cite this article as: Xu L, Zhang G, Zhang D, Zhang J, Zhang X, Bai X, Chen L, Jin R, Mao L, Li X, Sun H, Jin Z. Development and acceptability validation of a deep learning-based tool for whole-prostate segmentation on multiparametric MRI: a multicenter study. Quant Imaging Med Surg 2023;13(5):3255-3265. doi: 10.21037/qims-22-1068

Supplementary

Appendix 1 The detailed description of the model architecture

The network architecture, including the encoder path and the decoder path, was automatically determined by the nnU-Net framework via the data properties. In our case, for the T2-weighted images (T2WI), U-Net architecture consisted of 5 downsampling blocks and 5 upsampling blocks. In contrast, for the diffusion-weighted images (DWI) with a lower resolution, the model consisted of 3 downsampling blocks and 3 upsampling blocks. The detailed information is elaborated upon in *Table S2*.

Appendix 2 The detailed fine-tuning processes

The original training and fine-tuning cohorts were combined to train another 50 epochs, with a smaller initial learning rate of 0.0001. The parameters of the entire network were all updated. Similar to the original training procedure, the batch size of 2 was calculated with the nnU-Net framework based on the graphic processing unit (GPU) memory and the number of parameters of the model. The loss function was the combination of the Dice loss and the binary cross entropy (BCE) loss.

Appendix 3 Statistical analysis of the influencing factors for the Dice similarity coefficient on DWI images in the PXtest cohort without fine-tuning

A multivariate logistic regression analysis was performed to analyze the potential influencing factors in the decline of the DWI model's performance in the external testing cohort. The candidate influencing factors were echo time, repetition time, MR scanner, echo train length, and slice thickness.

https://dx.doi.org/10.21037/qims-22-1068

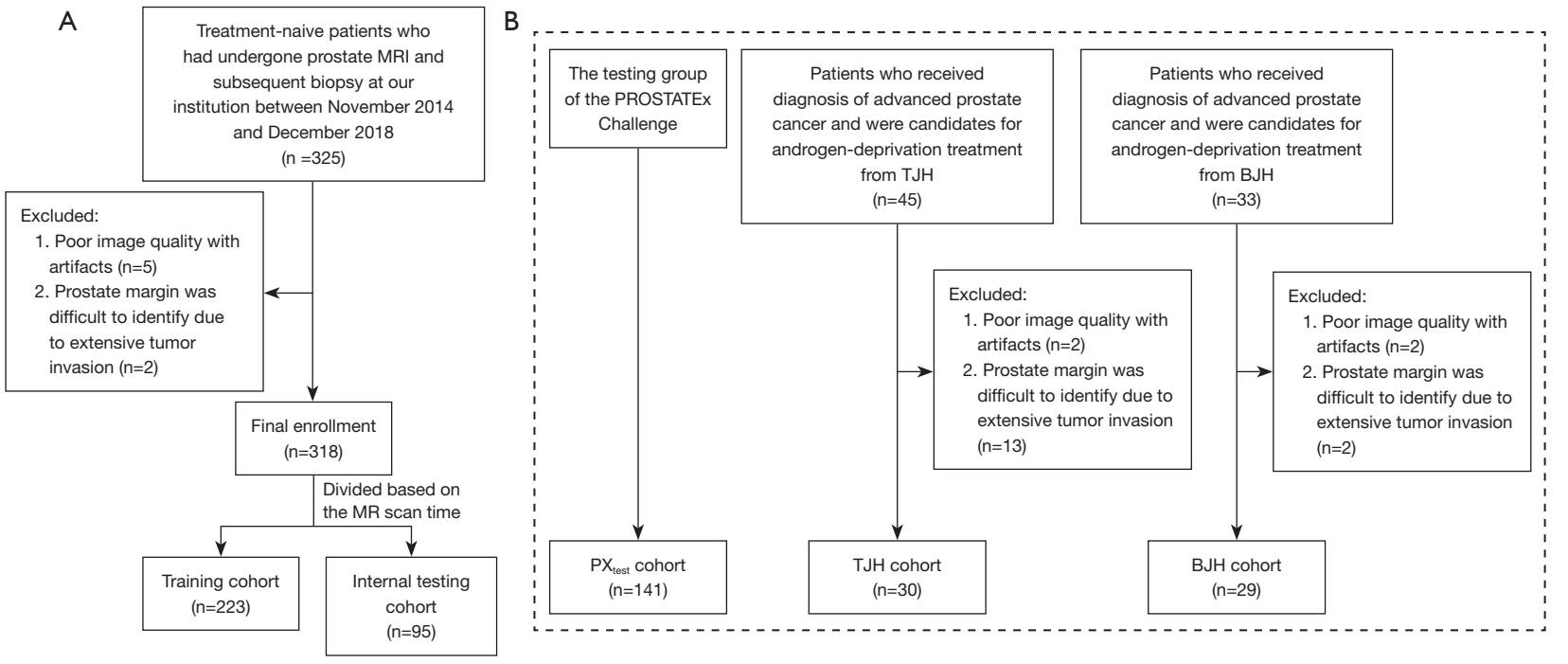

Figure S1 Flowchart of the selection process of the datasets in this study. PX_{test} cohort, testing group of the PROSTATEx Challenge dataset; TJH cohort, Tongji Hospital cohort; BJH cohort, Beijing Hospital cohort.

Table S1 MRI acquisition parameters for the axial T2-weighted imaging and diffusion-weighted imaging sequences [median (range)]

Sequences	Datasets		Cases	Scanner	MR field strength (T)	TR (msec)	TE (msec)	Matrix	Slice thickness (mm)	Pixel spacing
T2WI	Training cohort	Our institution	223	GE Discovery MR 750	3.0	4,422 (2,672–5,367)	108 (97–116)	512×512	3 (3–6)	0.51 (0.43–0.78)
	Internal testing cohort	Our institution	95	GE Discovery MR 750	3.0	4,424 (2,672–5,534)	108 (86–116)	512×512	3 (3–4)	0.51 (0.51–0.53)
	External testing cohort 1	PX_{test} cohort	90	Siemens Magnetom Skyra	3.0	5,660 (5,660–8,624)	104 (101–104)	384×384–640×640	3 (3–3.5)	0.5 (0.3–0.5)
			51	Siemens Magnetom TrioTim	3.0	4,494 (4,480–5,870)	103 (102–104)	256×256–320×320	3 (3–5)	0.56 (0.56–0.70)
	External testing cohort 2	TJH cohort	29	Siemens Magnetom Skyra	3.0	6,750 (6,130–7,970)	104	384×384	3 (3–3.5)	0.47
			1	Siemens Magnetom Aera	1.5	4,920	90	640×640	4	0.38
	External testing cohort 3	BJH cohort	10	Siemens Magnetom Espree	1.5	4,650 (3,800–5,040)	115 (102–118)	256×256–320×320	4 (3–4)	0.75 (0.75–0.78)
			3	GE Optima MR360	1.5	3,682 (3,660–3,682)	1101 (110–111)	512×512	4	0.47 (0.47–0.59)
			1	GE Signa EXCITE	1.5	4,120	122	512×512	5	0.625
			9	Philips Achieva	3.0	4,265 (2,500–5,183)	100 (80–100)	480×480–672×672	4	0.42 (0.30–0.5)
			5	GE Discovery MR 750	3.0	4,525 (4,103–4,775)	87 (87–89)	512×512	4	0.47
			1	GE SIGNA Pioneer	3.0	4,952	86	512×512	4	0.47
DWI	Training cohort	Our institution	223	GE Discovery MR 750	3.0	2,800 (2,000–4,000)	65 (62–70)	256×256	3 (3–6)	1.41 (0.94–1.56)
	Internal testing cohort	Our institution	95	GE Discovery MR 750	3.0	2,800	70 (69–70)	256×256	3 (3–4)	1.41
	Fine-tuning cohort	PX_{train} cohort	203	Siemens Magnetom Skyra, and TrioTim	3.0	2,700 (2,500–3,300)	63 (63–81)	128×84–128×120	3 (3–4.5)	2
	External testing cohort	PX_{test} cohort	90	Siemens Magnetom Skyra	3.0	2,700 (2,700–3,200)	63	128×84–128×120	3 (3–5)	2
			51	Siemens Magnetom TrioTim	3.0	2,800 (2,500–3,224)	70 (64–81)	106×128–128×88	3 (3–4)	2

T2WI, T2-weighted imaging; DWI, diffusion-weighted imaging; TR, repetition time; TE, echo time; PX_{test} cohort, testing group of the PROSTATEx Challenge dataset; TJH cohort, Tongji Hospital cohort; BJH cohort, Beijing Hospital cohort; PXtrain cohort, training group of the PROSTATEx Challenge dataset.

https://dx.doi.org/10.21037/qims-22-1068

Table S2 The architecture of the U-Net models for T2-weighted imaging and diffusion-weighted imaging

Block type	Model for T2WI		Model for DWI	
	Conv kernel	Pooling	Conv kernel	Pooling
Downsample 1	[1, 3, 3] ×2	[1, 2, 2]	[1, 3, 3] ×2	[2, 1, 1]
Downsample 2	[1, 3, 3] ×2	[1, 2, 2]	[1, 3, 3] ×2	[2, 1, 1]
Downsample 3	[3, 3, 3] ×2	[2, 2, 2]	[3, 3, 3] ×2	[1, 2, 2]
Downsample 4	[3, 3, 3] ×2	[1, 2, 2]	-	-
Downsample 5	[3, 3, 3] ×2	[1, 2, 2]	-	-
Bridge	[3, 3, 3] ×2	-	[3, 3, 3] ×2	-
Upsample 1	[3, 3, 3] ×2	[1, 2, 2]	-	-
Upsample 2	[3, 3, 3] ×2	[1, 2, 2]	-	-
Upsample 3	[3, 3, 3] ×2	[2, 2, 2]	[3, 3, 3] ×2	[1, 2, 2]
Upsample 4	[1, 3, 3] ×2	[1, 2, 2]	[1, 3, 3] ×2	[2, 1, 1]
Upsample 5	[1, 3, 3] ×2	[1, 2, 2]	[1, 3, 3] ×2	[2, 1, 1]
Output	[1, 1, 1]	-	[1, 1, 1]	-

T2WI, T2-weighted imaging; DWI, diffusion-weighted imaging.

Table S3 Multivariate regression analyses of factors affecting the Dice similarity coefficient on DWI images in the PX_{test} cohort without fine-tuning

Factors	Coefficients	Lower (2.5%)	Upper (97.5%)	P value
Echo time	0.004	–0.021	0.029	0.7633
Repetition time	0.001	0.001	0.002	<0.001
MR scanner	–0.729	–1.323	–0.136	0.016
Echo train length	–0.028	–0.041	–0.015	<0.001
Slice thickness	–0.273	–0.405	–0.140	<0.001

DWI, diffusion-weighted imaging; PX_{test} cohort, testing group of the PROSTATEx Challenge dataset.

https://dx.doi.org/10.21037/qims-22-1068