



Data Article

Dataset of prostate MRI annotated for anatomical zones and cancer



Lisa C. Adams^{a,b}, Marcus R. Makowski^c, Günther Engel^{a,d}, Maximilian Rattunde^a, Felix Busch^a, Patrick Asbach^a, Stefan M. Niehues¹, Shankeeth Vinayahalingam^e, Bram van Ginneken^e, Geert Litjens^e, Keno K. Bressen^{a,*}

^a Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Institute for Radiology, Hindenburgdamm 30, 12203 Berlin, Germany

^b Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

^c Technical University of Munich, Department of Diagnostic and Interventional Radiology, Faculty of Medicine, Ismaninger Str. 22, 81675, Munich, Germany

^d Institute for Diagnostic and Interventional Radiology, Georg-August University, Göttingen, Germany

^e Radboud University Medical Center, Nijmegen, GA, The Netherlands

ARTICLE INFO

Article history:

Received 26 July 2022

Revised 3 November 2022

Accepted 4 November 2022

Available online 9 November 2022

Dataset link: [Prostate158 - Training data \(Original data\)](#)

Keywords:

Prostate cancer

Pixel-wise segmentation

T2-weighted imaging

Apparent diffusion coefficient (ADC)

Diffusion-weighted imaging

3.0 Tesla MRI

ABSTRACT

In the present work, we present a publicly available, expert-segmented representative dataset of 158 3.0 Tesla biparametric MRIs [1]. There is an increasing number of studies investigating prostate and prostate carcinoma segmentation using deep learning (DL) with 3D architectures [2–7]. The development of robust and data-driven DL models for prostate segmentation and assessment is currently limited by the availability of openly available expert-annotated datasets [8–10]. The dataset contains 3.0 Tesla MRI images of the prostate of patients with suspected prostate cancer. Patients over 50 years of age who had a 3.0 Tesla MRI scan of the prostate that met PI-RADS version 2.1 technical standards were included. All patients received a subsequent biopsy or surgery so that the MRI diagnosis could be verified/matched with the histopathologic diagnosis. For patients who had undergone

DOI of original article: [10.1016/j.compbiomed.2022.105817](https://doi.org/10.1016/j.compbiomed.2022.105817)

* Corresponding author: Dr. Keno Bressen, Department of Radiology, Charité University, Hindenburgdamm 30, 12203 Berlin, Germany.

E-mail address: keno-kyrill.bressen@charite.de (K.K. Bressen).

Social media: [@k_bressen](#) (K.K. Bressen)

<https://doi.org/10.1016/j.dib.2022.108739>

2352-3409/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

multiple MRIs, the last MRI, which was less than six months before biopsy/surgery, was included. All patients were examined at a German university hospital (Charité Universitätsmedizin Berlin) between 02/2016 and 01/2020. All MRI were acquired with two 3.0 Tesla MRI scanners (Siemens VIDA and Skyra, Siemens Healthineers, Erlangen, Germany). Axial T2W sequences and axial diffusion-weighted sequences (DWI) with apparent diffusion coefficient maps (ADC) were included in the data set.

T2W sequences and ADC maps were annotated by two board-certified radiologists with 6 and 8 years of experience, respectively. For T2W sequences, the central gland (central zone and transitional zone) and peripheral zone were segmented. If areas of suspected prostate cancer (PIRADS score of ≥ 4) were identified on examination, they were segmented in both the T2W sequences and ADC maps.

Because restricted diffusion is best seen in DWI images with high b-values, only these images were selected and all images with low b-values were discarded. Data were then anonymized and converted to NIfTI (Neuroimaging Informatics Technology Initiative) format.

© 2022 The Author(s). Published by Elsevier Inc.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

Specifications Table

Subject	Radiography and radiology
Specific subject area	Magnetic resonance imaging for diagnosing prostate cancer
Type of data	Image
How the data were acquired	MRI images were acquired with Siemens VIDA and Skyra clinical 3.0 T scanners (Siemens Healthineers, Erlangen, Germany) according to an acquisition protocol that complies with PIRADS version 2.1 guidelines. Examinations contained T2w sequences and diffusion-weighted imaging (DWI) with apparent diffusion coefficient (ADC) maps. T2w sequences were acquired with a slice thickness of 3 mm, no interslice gap, an in-plane resolution of 0.47×0.47 mm, a field of view (FOV) of 180×180 mm, TE / TR 116 ms / 4040 ms, turbo factor 25, flip angle 160° , and an acquisition time of 3 minutes and 56 seconds. DWI sequences were acquired with a slice thickness of 3 mm, no interslice gap, in-plane resolution of 1.4×1.4 mm, field of view (FOV) of 220×220 mm, TE / TR 61 ms / 4400 ms, b values of 5, 100, 500, and 1000 s/mm ² , total acquisition time 5 minutes and 5 seconds. The software pre-installed on the MRI scanner (version VE11A) was used to calculate an ADC map from b-values of 50 to 1000 s/mm ² .
Data format	Filtered
Description of data collection	Men over 50 years who underwent MRI of the prostate at a 3.0 Tesla scanner and then underwent subsequent surgery or biopsy of the prostate were included in this study. All patients were examined at a German university hospital (Charité Universitätsmedizin Berlin) between 02/2016 and 01/2020. The dataset contains 102 patients with histologically verified PCa, and 56 patients who served as controls. Presence of pathologically significant prostate cancer was defined as at least Gleason grade 6 [Gleason Grade Group (GGG) 1] and tumor volume ≥ 0.1 ml. Cancer was confirmed on a patient level. <ul style="list-style-type: none"> • Institution: Charité - Universitätsmedizin Berlin • City/Town/Region: Berlin • Country: Germany
Data source location	

(continued on next page)

Data accessibility	Repository name: https://zenodo.org Data identification number: 10.5281/zenodo.6481141 Access at: zenodo.org/record/6481141 Instructions for accessing the data: Please register an account at Zenodo.org, you can then request access to the data. You will need to enter your name and a valid institutional email address. Please note that sharing of the data is prohibited and each person working with the data needs to request access zenodo.org individually. Access to the data will usually be granted within less than 24 hours.
Related research article	L.C. Adams, M.R. Makowski, G. Engel, M. Rattunde, F. Busch, P. Asbach, S.M. Niehues, S. Vinayahalingam, B. van Ginneken, G. Litjens, Prostate158-An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection, Computers in Biology and Medicine (2022) 105817.

Characteristics of the Patient Cohort

	Entire dataset	Training set	Test set
Numbers	158	139	19
Age, years (\pm SD, range)	69 \pm 9, 35–84	66 \pm 9, 35–84	71 \pm 10, 53–83
Verified PCa (%)	102/139 (73.4)	83/139 (60)	19/19 (100)
PSA (IQR)	7.5 (6.3)	7.6 (6.7)	7.4 (7.1)
PI-RADS scores : number of patients	4: 55 5: 47	4: 45 5: 38	4: 10 5: 9
Gleason grade groups (GGG) : number of patients	1: 11 2: 37 3: 24 4: 21 5: 9	1: 9 2: 29 3: 19 4: 18 5: 8	1: 2 2: 8 3: 5 4: 3 5: 1
BMI, kg/m² (IQR)	25.9 (6.0)	25.8 (6.3)	25.7 (5.9)

Value of the Data

- Expert annotated MRI datasets of confirmed prostate cancer are rare [8,12,13], but are needed to develop sufficient deep learning algorithms to assist radiologists and urologists in the detection and treatment of prostate cancer.
- This dataset provides segmentations from experts in urologic radiology which were extensively reviewed to ensure high quality segmentations. Researchers in computer vision, who do not have access to medical data and/or radiologic expertise, can therefore use it to develop new algorithms for prostate cancer detection.
- The dataset can serve either as a training dataset for developing new algorithms or as a test set for trained algorithms to check their performance and generalizability.
- Part of the dataset remains as a hidden test dataset to be able to check the performance of the trained algorithms on <https://prostate158.grand-challenge.org>. Researchers can compare trained algorithms under realistic conditions. As the test set remains hidden, overfitting on test data is not possible.

1. Objective

The development of robust computational models for prostate imaging is limited by the availability of high-quality datasets for prostate cancer (PCa) and zonal anatomy. Many deep learning papers published in the last years were based on the same open-source patient data from the widely known PROSTATEx challenge dataset, while most studies used non-publicly available single-center datasets. The limited data availability is accompanied by limited availability

of open source code. To increase the availability of open data and code for prostate and PCa segmentation, we here provide a comparatively large, publicly available, expert-segmented representative dataset of 158 3T bpMRIs that overall conforms to PI-RADSTM v2.1 guidelines, along with an open-source baseline U-Net algorithm.

2. Data Description

The dataset consists of 158 3.0 Tesla MRIs from patients with suspected prostate cancer annotated by two board-certified radiologists. Inclusion criteria for patients were a minimum age over 50 years, an MRI examination with a biparametric imaging protocol meeting PI-RADS v2 technical standards, a subsequent biopsy or surgery, and an available pathology report confirming a diagnosis of prostate cancer. In patients with multiple MRIs, the last examination which was less than six months before pathology was considered. All examinations were performed at a German university hospital (Charité University Hospital Berlin) between 02/2016 and 01/2020. [11]

2.1. Structure of the data archive

Folder:

Prostate158_train:

→ train

→ 020

→ adc.nii.gz # ADC sequence in NiftI format

→ adc_tumor_reader1.nii.gz # Tumor segmentation of reader 1

→ adc_tumor_reader2.nii.gz # Tumor segmentation of reader 2

→ dwi.nii.gz # DWI sequence in NiftI format

→ t2.nii.gz # T2W sequence in NiftI format

→ t2_anatomy_reader1.nii.gz # Anatomy segmentation of reader 1

→ t2_tumor_reader1.nii.gz # Tumor segmentation of reader 1

→ 021

...

→ 158

train.csv # file names with relative paths to be used for training
valid.csv # file names with relative paths to be used for model validation during training

All examinations in the dataset are in axial plane and resampled to the same pixel-spacing and orientation, as well as center-cropped for easier handling. Therefore, the resolution of the original sequences, as given in the below table, is not preserved. Example slices for all sequences with corresponding annotations are provided by Fig. 1.

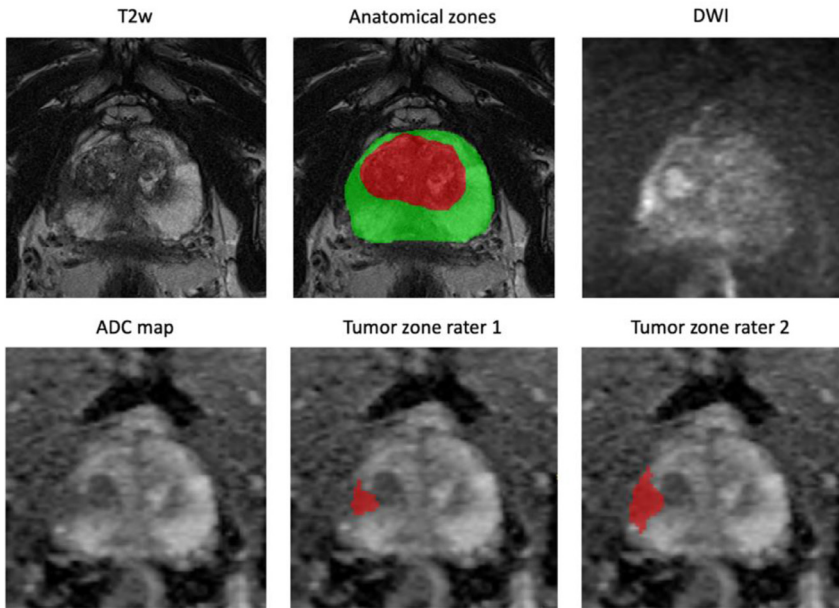


Fig. 1. Example axial slices of resampled sequences in the Prostate158 dataset with expert annotations for anatomical zones and tumor zones.

3. Experimental Design, Materials and Methods

3.1. Image acquisition and extraction

The data set consists of 158 MRI examinations, all performed on Siemens VIDA and Skyra (Siemens Healthineers, Erlangen, Germany) clinical 3.0-T scanners according to an acquisition protocol that meets current guidelines and using B1 shimming.

Sequence parameters were as follows:

	T2w	DWI/ADC
Slice Thickness	3 mm	3 mm
Interslice gap	None	None
Inplane resolution	0.47×0.47 mm	1.4×1.4 mm
FOV size	180×180 mm	220×200 mm
TE	116 ms	61 ms
TR	4040 ms	4400 ms
Turbo factor	25	-
Flip angle	160°	-
Acquisition times	3:56 min	5:05 min
B-values	-	0, 100, 500, 1000 s/mm ³

The ADC map was calculated from b-values ranging from 100 to 1000 s/mm³, using the software pre-installed on the MRI scanner (version VE11A).

Images were then extracted from the PACS and converted to NIfTI (Neuroimaging Informatics Technology Initiative) format for further processing.

3.2. Pixel-wise segmentations of anatomical zones and tumor zones

All pixel-wise segmentations were performed with the open-source software ITK-Snap (version 3.8.0; www.itksnap.org). One board-certified radiologist with 6 years of experience segmented the central gland (central zone and transitional zone) and the peripheral zone. Images were then reviewed by a second board-certified radiologist with 8 years of experience.

For prostate carcinoma, both radiologists created independent segmentations of suspected tumor zones. Only lesions with PI-RADS scores of 4 or 5 were included. After a four-week interval, all segmentations were again assessed by the radiologists.

Anatomical zones were segmented in the T2w sequences. Segmentation of PCa lesions was performed in the ADC map and then adjusted in case of over- or underestimation of the tumor extent. This was achieved by creating an overlay of the ADC segmentation mask on T2w and DWI to account for anatomical details and localization of restricted diffusion. Additional segmentations of PCa lesions for T2-weighted images were only performed by reader 1. Segmentations for anatomical zones and tumor zones were saved in different files, each in NiftI format.

3.3. MRI pre-processing

All MRIs were pre-processed using intensity resampling, and centre cropping. In addition, a non-parametric, non-uniform intensity normalization was performed for T2w images to remove distortion fields using the N4ITK algorithm (13). For DWI, only images with the highest b-value (1000) were extracted and the remaining images were discarded.

3.4. Resampling

Although the sequences were acquired on the same patient, the FOV, resolution and spacing of the sequences differ. Since each sequence is also manually planned, the orientation of the sequences of a single examination may also be different. For this reason, identical coordinates in the sequences of the same examination do not correspond to the same anatomical region. Therefore, all examinations were resampled so that (1) the size and resolution were identical and (2) the voxel coordinates in different sequences of an MRI corresponded to the same anatomical region. Resampling was performed using SimpleITK 2.0 with the following command:

```
import SimpleITK as sitk
reference = sitk.ReadImage('path/to/reference')
target = sitk.ReadImage('path/to/target')
resampled = sitk.Resample(target, reference)
```

The T2w sequence was used as reference for all other examinations. For the segmentation mask, the mode for interpolations was set as SimpleITK.sitkNearestNeighbor; for images, the mode was SimpleITK.sitkLinear

3.5. Cropping

Prostate MRIs always contain a large part of the pelvis, which is irrelevant for prostate zone segmentation or prostate cancer detection. Because the prostate is reliably located in the center of the examination, images were centered by removing 25% of the image margins along the anterior-posterior and lateral-medial axes.

```
def _flatten(t):
    return [item for sublist in t for item in sublist]

def crop(image, margin, interpolator):
    """
    Crops a sitk.Image while retaining correct spacing. Negative margins will lead to zero padding
    Args:
        image: a sitk.Image
        margin: margins to crop. Single integer or float (percentage crop), lists of int/float or
              nested lists are supported.
        Interpolator: the method of interpolation to be used
    """
    if isinstance(margin, (list, tuple)):
        assert len(margin) == 3, f'expected margin to be of length 3'
    else:
        assert isinstance(margin, (int, float)), f'expected margin to be a float value'
        margin = [margin, margin, margin]

    margin = [m if isinstance(m, (tuple, list)) else [m,m] for m in margin]
    old_size = image.GetSize()
    old_origin = image.GetOrigin()

    # calculate new origin and new image size
    if all([isinstance(m, float) for m in _flatten(margin)]):
        assert all([m >= 0 and m < 0.5 for m in _flatten(margin)]), f'margins must be between 0 and 0.5'
        to_crop = [[int(sz*_m) for _m in m] for sz, m in zip(old_size, margin)]
    elif all([isinstance(m, int) for m in _flatten(margin)]):
        to_crop = margin
    else:
        raise ValueError('wrong format of margins')

    new_size = [sz - sum(c) for sz, c in zip(old_size, to_crop)]

    # origin has Index (0,0,0)
    # new origin has Index (to_crop[0][0], to_crop[2][0], to_crop[2][0])
    new_origin = image.TransformIndexToPhysicalPoint([c[0] for c in to_crop])

    # create reference plane to resample image
    ref_image = sitk.Image(new_size, image.GetPixelIDValue())
    ref_image.SetSpacing(image.GetSpacing())
    ref_image.SetOrigin(new_origin)
    ref_image.SetDirection(image.GetDirection())
    return sitk.Resample(image, ref_image, interpolator=interpolator)

image = crop(image, (0.25, 0.25, 0.0), sitk.sitkLinear)
```

After cropping, the images were saved in Nifti format. Fig. 2 provides a visual overview of the preprocessing steps.

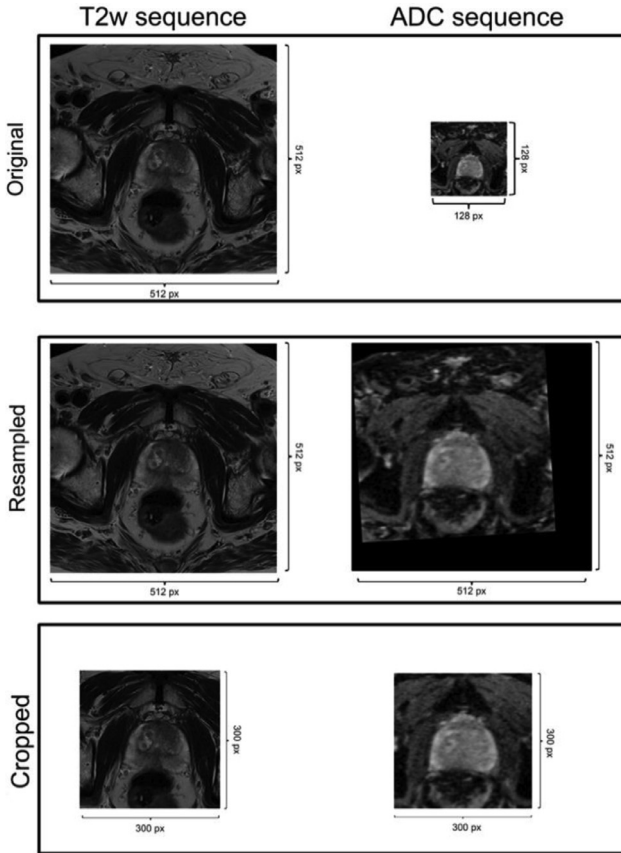


Fig. 2. Overview of preprocessing steps taken.

Ethics Statements

All contributions to this study were approved by the Institutional Review Board (EA4/062/20), including a waiver of informed consent, and performed in accordance with data protection policy. All research has been carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

[Prostate158 - Training data \(Original data\)](#) (Zenodo).

CRedit Author Statement

Lisa C. Adams: Conceptualization, Methodology, Data curation, Investigation, Writing – original draft; **Marcus R. Makowski:** Conceptualization, Methodology, Data curation, Resources, Writing – review & editing; **Günther Engel:** Methodology, Investigation; **Maximilian Rattunde:** Writing – review & editing; **Felix Busch:** Writing – review & editing; **Patrick Asbach:** Writing – review & editing; **Stefan M. Niehues:** Writing – review & editing, Resources; **Shankeeth Vinayahalingam:** Writing – review & editing, Methodology; **Bram van Ginneken:** Writing – review & editing, Methodology; **Geert Litjens:** Writing – review & editing, Methodology; **Keno K. Bresslem:** Conceptualization, Methodology, Software, Validation, Visualization, Writing – original draft.

Acknowledgments

LCA is grateful for her participation in the BIH Charité Junior Clinician and Clinician Scientist Program and KKB is grateful for his participation in the BIH Charité Digital Clinician Scientist Program all funded by the Charité – Universitätsmedizin Berlin, the Berlin Institute of Health, and the Deutsche Forschungsgemeinschaft. We also thank the BIH for providing access to their HPC for Research cluster for our analysis.

We acknowledge financial support from the Open Access Publication Fund of Charité – Universitätsmedizin Berlin and the German Research Foundation (DFG).

References

- [1] L.C. Adams, Makowski, M.R., Engel, G., Rattunde, M., Busch, F., Vinayahalingam, S., van Ginneken, B., Bresslem, K.K., Prostate158 - An annotated MRI dataset and algorithm for prostate cancer detection, *Computers in Biology and Medicine* (In Press).
- [2] A. Meyer, G. Chlebus, M. Rak, D. Schindele, M. Schostak, B. van Ginneken, A. Schenk, H. Meine, H.K. Hahn, A. Schreiber, Anisotropic 3D multi-stream CNN for accurate prostate segmentation from multi-planar MRI, *Computer Methods and Programs in Biomedicine* 200 (2021) 105821.
- [3] B. Wang, Y. Lei, S. Tian, T. Wang, Y. Liu, P. Patel, A.B. Jani, H. Mao, W.J. Curran, T. Liu, Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation, *Medical physics* 46 (4) (2019) 1707–1718.
- [4] B. Wang, Y. Lei, J.J. Jeong, T. Wang, Y. Liu, S. Tian, P. Patel, X. Jiang, A.B. Jani, H. Mao, Automatic MRI prostate segmentation using 3D deeply supervised FCN with concatenated atrous convolution, *Medical Imaging 2019: Computer-Aided Diagnosis, SPIE* (2019) 988–995.
- [5] Y. Jin, G. Yang, Y. Fang, R. Li, X. Xu, Y. Liu, X. Lai, 3D PBV-Net: an automated prostate MRI data segmentation method, *Computers in Biology and Medicine* 128 (2021) 104160.
- [6] J. Chen, Z. Wan, J. Zhang, W. Li, Y. Chen, Y. Li, Y. Duan, Medical image segmentation and reconstruction of prostate tumor based on 3D AlexNet, *Computer methods and programs in biomedicine* 200 (2021) 105878.
- [7] O.J. Pellicer-Valero, J.L. Marengo Jiménez, V. Gonzalez-Perez, J.L. Casanova Ramón-Borja, I. Martín García, M. Barrios Benito, P. Pelechano Gómez, J. Rubio-Briones, M.J. Rupérez, J.D. Martín-Guerrero, Deep Learning for fully automatic detection, segmentation, and Gleason Grade estimation of prostate cancer in multiparametric Magnetic Resonance Images, *Scientific reports* 12 (1) (2022) 1–13.
- [8] A. Fedorov, M. Schwier, D. Clunie, C. Herz, S. Pieper, R. Kikinis, C. Tempny, F. Fennessy, An annotated test-retest collection of prostate multiparametric MRI, *Scientific data* 5 (1) (2018) 1–13.
- [9] G. Litjens, *The PROMISE12 segmentation challenge*, 2012.
- [10] R. Cucolo, A. Stanzione, A. Castaldo, D.R. De Lucia, M. Imbriaco, Quality control and whole-gland, zonal and lesion annotations for the PROSTATEx challenge public dataset, *European Journal of Radiology* 138 (2021) 109647.
- [11] L.C. Adams, M.R. Makowski, G. Engel, M. Rattunde, F. Busch, P. Asbach, S.M. Niehues, S. Vinayahalingam, B. van Ginneken, G. Litjens, Prostate158-An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection, *Computers in Biology and Medicine* (2022) 105817.
- [12] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge, *Medical image analysis* 18 (2) (2014) 359–373.
- [13] S.G. Armato, H. Huisman, K. Drukker, L. Hadjiiski, J.S. Kirby, N. Petrick, G. Redmond, M.L. Giger, K. Cha, A. Mamonov, PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images, *Journal of Medical Imaging* 5 (4) (2018) 044501.