Magnetic Resonance Imaging 131 (2026) 110698

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

journal homepage: www.elsevier.com/locate/mri

Incorporating functional soft tissue deformations in AI model training for spatially accurate prostate cancer detection

Balint Kovacs [a,b,c,d,1], Kevin Sun Zhang [b,1], Michael Baumgartner [a,e,f], Thomas Hielscher [g], Dimitrios Bounias [a,c], Nils Netzer [b,c], Ralf Floca [a], Carolin Eith [b,c], Paul F. Jäger [f,h], Clara Meinzer [b], Fabian Isensee [a,f], Regula Gnirs [b], Magdalena Görtz [i,j], Markus Hohenfellner [j], Albrecht Stenzinger [k], Heinz-Peter Schlemmer [b,l], Ivo Wolf [m], Klaus H. Maier-Hein [a,d,f,l,n,2], David Bonekamp [b,c,l,o,*,2]

[a] German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Heidelberg, Germany
[b] DKFZ Heidelberg, Division of Radiology, Heidelberg, Germany
[c] Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany
[d] HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[e] Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany
[f] DKFZ Heidelberg, Helmholtz Imaging, Heidelberg, Germany
[g] DKFZ Heidelberg, Division of Statistics, Heidelberg, Germany
[h] DKFZ Heidelberg, Interactive Machine Learning Group, Heidelberg, Germany
[i] DKFZ Heidelberg, Junior Clinical Cooperation Unit 'Multiparametric methods for early detection of prostate cancer', Heidelberg, Germany
[j] Department of Urology, University of Heidelberg Medical Center, Germany
[k] Institute of Pathology, University of Heidelberg Medical Center, Germany
[l] German Cancer Consortium (DKTK), DKFZ, core center Heidelberg, Germany
[m] Technical University of Applied Sciences Mannheim, Germany
[n] Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[o] National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg, Germany

ABSTRACT

Purpose: To increase performance and generalization ability of artificial intelligence prostate cancer detection systems by simulating physiological size changes of the bladder and rectum and, thereby, associated deformations of the prostate and its lesions.
Materials and methods: This retrospective study included 1028 bi-parametric MRI examinations of men (age range: 40–90 years) performed between 2014 and 2019, divided into training/test sets (771/257). We integrated an 'anatomy-informed' transformation into the training of nnU-Net, by simulating soft-tissue deformations of the prostate resulting from size changes of the rectum and bladder. The effects of these strategies were evaluated using free-response receiver operating characteristic (FROC) to assess lesion-level performance, along with a variant: weighted alternative FROC (wAFROC), which prioritizes patient-level effects with localization criteria. Change in sensitivity was tested using a clustered McNemar test. Patient-level performance was assessed with standard and localized receiver operating characteristics (ROC/LROC) analysis.
Results: On the independent test set, the anatomy-informed model simulating changes of both rectum and bladder significantly increased lesion-level detection of true positive lesions by 18.8% (from 48 to 57, $p = 0.01$) and demonstrated significantly higher performance in the wAFROC analysis (from 0.597 to 0.639, $p < 0.01$). Patient-level ROC increased slightly (from 0.779 to 0.782, $p = 0.89$), while LROC analysis demonstrated increased performance (from 0.471 to 0.546).
Conclusion: Simulation of rectum and bladder size variations during model training led to significant improvement in lesion detection performance, which may be crucial for diagnostics and therapeutic measures depending on correct lesion localization, e.g. MRI-guided biopsies or focal therapy regimes.

* Corresponding author at: German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany.
E-mail address: d.bonekamp@dkfz-heidelberg.de (D. Bonekamp).
[1] Shared first authorship
[2] Equal contribution

https://doi.org/10.1016/j.mri.2026.110698
Received 16 February 2026; Received in revised form 23 April 2026; Accepted 27 April 2026
Available online 5 May 2026

1. Introduction

Spatially accurate detection of prostate cancer (PCa) on a lesion-level basis on multi-parametric magnetic resonance images (mpMRI) is crucial for guiding subsequent clinical interventions and determination of adequate therapy. Procedures such as biopsy for histologic confirmation and location-dependent treatment methods like transurethral ultrasound ablation [1] or radiotherapy with added focal boost to the tumor bed [2] rely heavily on precise tumor localization. Additionally, correct localization is crucial for prostate cancer staging [3]. Therefore, spatial information of prostate cancer lesions directly influences patient-level diagnosis and the therapeutic success.

Interpretation of prostate MRI remains a challenging task for radiologists due to the considerable variability in lesion size and shape [4], as both influence the assessment of prostate lesions on mpMRI according to the Prostate Imaging-Reporting and Data System (PI-RADS) [5]. This shape variability is not solely due to inherent tumor characteristics but is also influenced by the constant soft tissue deformations resulting from muscle contractions, respiration, and more importantly, variable filling of the directly adjacent organs, the rectum and the bladder [6].

Considering that one MRI examination captures only one (time) point from all the functional states resulting from the continuous change of adjacent organs, the deformation of the prostate will never be the same on any repeat or subsequent examination. Given the limited number of training examples for artificial intelligence (AI) aided PCa detection systems [7], typical results of soft tissue deformations could be underrepresented in the training dataset and thus algorithms may be limited in their ability to accurately assess the full spectrum of lesion characteristics. However, the training of current state-of-the-art AI systems still relies on simple spatial transformations - translation, rotation, cropping, and scaling - by globally augmenting the MRI sequences without substantially increasing the morphological diversity [8–11].

Recently, a training strategy utilizing 'anatomy-informed' augmentation has been proposed [12], which generates size variations of the rectum and the bladder. This approach effectively enables the use of anatomically accurate soft tissue deformations during AI model training, a significant advancement over complex models that do not scale to online data augmentation [13,14]. This strategy has demonstrated the potential of realistic transformations compared to random elastic image transformations by preserving essential image features [12].

In this study, we aim to validate the training strategy extended with soft tissue deformations by assessing its impact on lesion detection performance in an AI model for prostate cancer detection on MRI. We extended the training scheme of the current state-of-the-art segmentation convolutional neural network, nnU-Net [15], with the proposed 'anatomy-informed' augmentation scheme and compared its performance with and without the additional augmentation method.

We hypothesized that by enriching the training data with increased anatomically correct lesion shape variability, the AI system's lesion-level detection performance would increase, which may also benefit patient-level performance.

2. Methods

2.1. Study sample

This retrospective study included consecutive patients, who received mpMRI and MRI/transrectal ultrasound (MRI/TRUS) fusion, as well as extended systematic biopsies between 01/2014–12/2019 at our institutions. The local ethics committee approved the study and waived informed consent.

All men had suspicion for PCa or participated in our active surveillance program and were included, if they met the following criteria: (a) MRI performed on institutional MRI scanners, (b) MRI/TRUS fusion biopsy performed at our institutions with biopsy results available. Exclusion criteria were a) previous therapy for PCa (prostatectomy, ablation or radiation therapy, anti-hormonal treatment, b) severe imaging artifacts, e.g. due hip implants, c) biopsy within 2 months prior to MRI or time to biopsy after MRI exceeding 6 months, d) uncommon/rare pathological diagnoses with irregular imaging manifestations (i.e., one patients with smooth muscle cell tumor of uncertain malignant potential) e) exams in which no suspicious lesion was visible on MRI but clinically significant prostate cancer was detected during systematic biopsy – exclusion of these exams was necessary to enable lesion-level performance assessment requiring ground truth lesion segmentations as control. This, consequently, does not imply that all patients without MRI-visible lesions were excluded.

The exams reported in this manuscript have been part of previously reported studies analyzing performance of different AI models, most recently [8,11,16]. Nevertheless, the incorporation of soft tissue deformations into the training strategy has not been investigated yet. The technical and practical advantages of realistic transformations compared to global and/or random transformations have been reported during the initial proposal of the 'anatomy-informed' augmentation scheme at the International Conference on Medical Image Computing and Computer-Assisted Intervention 2023 [12]. Compared to the results of the conference proceedings, this study contained a larger patient cohort, included detailed cohort characterization, and detailed lesion-level analysis with advanced statistical testing.

2.2. MRI protocol

Multiparametric prostate MRI examinations were performed on three scanners: Magnetom Prisma and Biograph mMR, both 3 T (Siemens Healthineers); and Magnetom Aera, 1.5 T (Siemens Healthineers) based on PI-RADS recommendations and ESUR guidelines at the time of acquisition [17]. Standard multichannel body coils and integrated spine-phased array coils were used for image acquisition. Detailed MRI acquisition parameters of utilized T2- and diffusion-weighted sequences are given in Supplementary Table 1.

2.3. PI-RADS assessment, data annotation and histopathological ground truth

PI-RADS [18] assessment was performed during clinical routine by board-certified radiologists. All exams were discussed in interdisciplinary conferences prior to biopsy. Histopathological ground truth was established for suspicious lesions by targeted transperineal MRI/TRUS fusion biopsy followed by extended systematic biopsies according to the Ginsburg protocol [19]. Histopathology was performed in clinical routine under supervision of a dedicated uropathologist (A.S., >15 years of experience). Clinically significant PCa (csPCa) was defined as International Society of Urological Pathology (ISUP) Gleason Grade Group≥2.

Preexisting lesion segmentations were utilized with data processing described previously [8,11]. Briefly, segmentations of prostate lesions harboring csPCa were performed with reference to clinical written reports and corresponding pictograms using the Medical Imaging Interaction Toolkit [20] under supervision of a board-certified, fellowship-trained radiologist (D.B., >10 years of experience in prostate MRI). Histopathological lesion ground truth (GT) was established from targeted-biopsy results supplemented with data from systematic biopsies, i.e. if the targeted biopsy of a MRI-visible lesion showed no csPCa, but the systematic biopsy of this prostate sextant (see Ginsburg protocol) demonstrated csPCa, then the lesion was still considered true positive. The combined biopsy approach (targeted & systematic) has previously proved to be a reliable histological assessment with sensitivity comparable to radical prostatectomy specimen [21].

Rectum and bladder segmentations were generated in an iterative process: A medical student (C.E.) generated segmentations of the bladder and rectum on a subcohort; the subcohort was then used to train a nnU-Net model and predict further bladder and rectum segmentations

on a larger partition of the data, which were manually refined by two radiology residents (C.M. and K.S.Z.) if necessary. After visual inspection / refinement, the larger annotated data set was then used to train the final in-house model used for the prediction of this study's rectum and bladder segmentations. The quality of the predicted segmentations was confirmed by the fellowship-trained radiologist (D.B.).

2.4. Anatomy-informed data augmentation

To integrate soft tissue deformation in the human pelvis resulting from physiological processes into AI model training, we use the 'anatomy-informed' transformation based on a biomechanical model proposed previously [12]. In this model, soft tissue deformations are applied to the rectum and the bladder, which are the two organs closest to the prostate and whose morphological changes due to their filling status lead to direct deformations of the prostate and its lesions [6,22]. In the model, simplifications are made by setting multiple boundary conditions for the biomechanical properties of the pelvis: isotropic transformation perpendicular to the organ surfaces due to the isotropic mechanical behavior of both the rectum and bladder [23], assumed homogeneous material composition for the prostate and surrounding muscle tissues due to their similar Young's modulus and Poisson's ratio [24], no resistance for the transformation as solid shape preserving solid tissues like bone are rather distant from the specific organs. The deformation reflecting the elastic behavior of soft tissues is approximated by a nonlinear displacement field which decays proportionally with distance from the rectum and bladder surfaces, implemented using Grid sampling and interpolation, to generate realistic, hyperelastic deformations [25]. Viscoelasticity was disregarded as time effects were not relevant.

The biomechanical model enables on-the-fly data augmentation in contrast to computationally more complex methods [13,14], while producing realistic distension/evacuation of the bladder or rectal spaces, see Fig. 1. To prioritize rectal transformations in line with clinical insights (zonal lesion distribution [5], strength of deformation [6]), we introduced an additional probability factor that controls the transformation likelihood between the rectum and bladder in our model.

2.5. Model training

We used a bi-parametric model with co-registration of MRI sequences. Registration employed B-spline transformation, calculated via mutual information between the sequences of T2w and DWI with the lowest b-values to match the GT segmentations to all modalities [16]. The transformation parameters were then applied to DWIs with the highest b-values and the ADC maps. We crop the images preserving anatomical context (offsets of $\pm$9 mm axial to the prostate and $\pm$ 11.25 mm in the axial plane to the rectum and bladder) to be able to use the anatomy-informed augmentation and to avoid training instability [16,26].

We trained 3D nnU-Net models [15] using 5-fold cross-validation, extending its standard augmentation scheme (used as the reference model) with organ deformations (anatomy-informed model). Training settings were consistent with previous reports, employing balanced sampling for clinically significant prostate cancer (csPCa) prevalence [12], and reducing the number of epochs to 600 to prevent overfitting. During model development, the Free-response Receiver Operating Characteristic (FROC) score, which is integrated logarithmically within the range of 0.25–1.0 average false positives per scan (avgFPs/scan), was optimized. This was guided by radiologists' lesion-level avgFPs/scan at PI-RADS $\geq$4 and 3 (0.38 and 0.82, respectively). Final models were ensembled and evaluated on the independent test set.

2.6. AI performance evaluation and statistical testing

The CNN model was trained on radiologist-report based ground truth lesions harboring csPCa verified by biopsy. The task was to detect and perform segmentations of csPCa lesions on a voxel-wise basis. Predicted lesion instances were identified by thresholding the probability maps generated by the model at 0.5, followed by connected component analysis. Prediction scores under this value were set to zero. Intersection over Union (IoU) of 0.1 or higher between the AI-predicted and the GT lesion segmentation based on radiologist reports were classified as True Positives (TPs) via Greedy-matching [27]. Other model-predicted lesions were considered false positives.

Patient-level evaluation is based on whether a patient is correctly identified as possessing csPCa, i.e. ISUP Grade Group$\geq$2, or having no cancer / insignificant prostate cancer. Therefore, on a per patient basis, a patient with csPCa and at least one model-predicted lesion was considered true positive. The overlap of the predicted and the GT segmentation was no necessary condition. This approach is reflected in the area under the curve (AUC) of the standard receiver operating characteristic (ROC) analysis; statistical testing was performed using the DeLong test [28]. As this approach does not account for correct localization of a prediction, a case might be considered 'true positive', when the model predicts csPCa at the wrong location while not detecting any lesion at the factual location of csPCa within the prostate. To partly account for this effect, localization ROC (LROC) analysis was performed, which plots the probability of a correct localization (PCL) using the predicted lesion with the highest score vs. the false-positive fraction (1- Specificity). However, LROC only considers one prediction (with the highest probability) per case [29]. Therefore, it can partially account for the localization blind spot of traditional ROC analysis, however, prostate cancer is known for multifocality, and the possibility of a patient having multiple lesions is not considered.

These restrictions were addressed by lesion-level analysis and metrics: Free-response receiver operating characteristic (FROC) analysis was performed, which compares lesion-level sensitivity (with correct localization) to the average number of false positives per scan (avgFPs/scan) [29]. A variant, the weighted alternative FROC (wAFROC), was also calculated: the wAFROC compares the fraction of correctly localized true-positives to the patient-level false-positive fraction [29], with the weights to adjust for multiple lesions per patient. Therefore, it is a combined lesion-level / patient-level metric. Differences in wAFROC metric were tested with ANOVA approach for dependent observations according to Obuchowski [30] as implemented in the R package RJafroc [31].

Furthermore, lesion-level sensitivity differences at the detection threshold were compared using a clustered McNemar test to account for patient effects [32]. To our knowledge, there is no common way to test the FROC and LROC metrics for significant differences.

3. Results

3.1. Study sample

The final study cohort comprised 1028 exams. An inclusion and exclusion diagram is given in Fig. 2. After balanced allocation regarding cancer rate (36.3%), the Median patient exam age was 64 years, interquartile range (IQR) [58–69], and 65 years, IQR [59–70] years for the training and test cohort, respectively. The training set contained 771 (75%) and the test set 257 exams (25%), and 271 and 90 exams with csPCa, corresponding to 372 and 121 csPCa lesions, respectively. Detailed demographic and characteristics of the cohort are given in Table 1.

3.2. Training cohort

During development using 5-fold cross-validation, the model incorporating both rectal and bladder changes demonstrated higher performance compared to single-organ models and was therefore used for subsequent analysis. The parameters for the probability and amplitude changes for the bladder and rectum were based on results during model

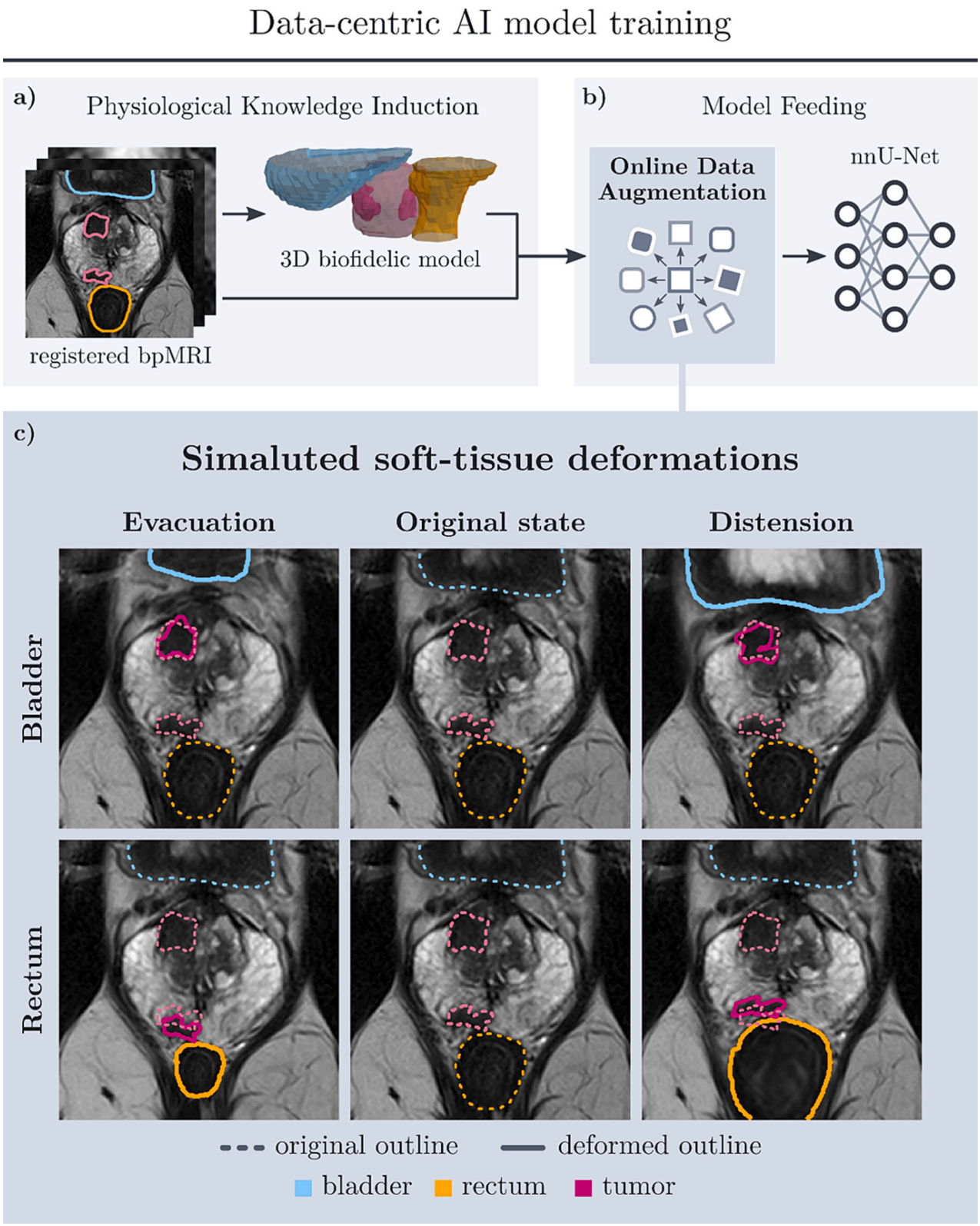

Fig. 1. Simulation of soft tissue deformations due to functional changes in the shape of the rectum and bladder during AI model training. a) Transformation parameters for the soft tissue transformations are calculated using knowledge about the biomechanical properties of the pelvis and the anatomy segmentations. b) In every training step, the AI model is fed with different physiological states and, thus, a broader spectrum of lesion characteristics. c) Resulting simulation of bladder (top row) and rectal (bottom row) functions like distension (right column) and evacuation (left column). The middle column provides the original MRI image with the unchanged rectum (orange), bladder (blue) and lesion (purple) outline. The morphological organ changes result in localized tissue deformations affecting primarily the adjacent lesions. Dotted lines represent the original and the solid lines the deformed outlines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

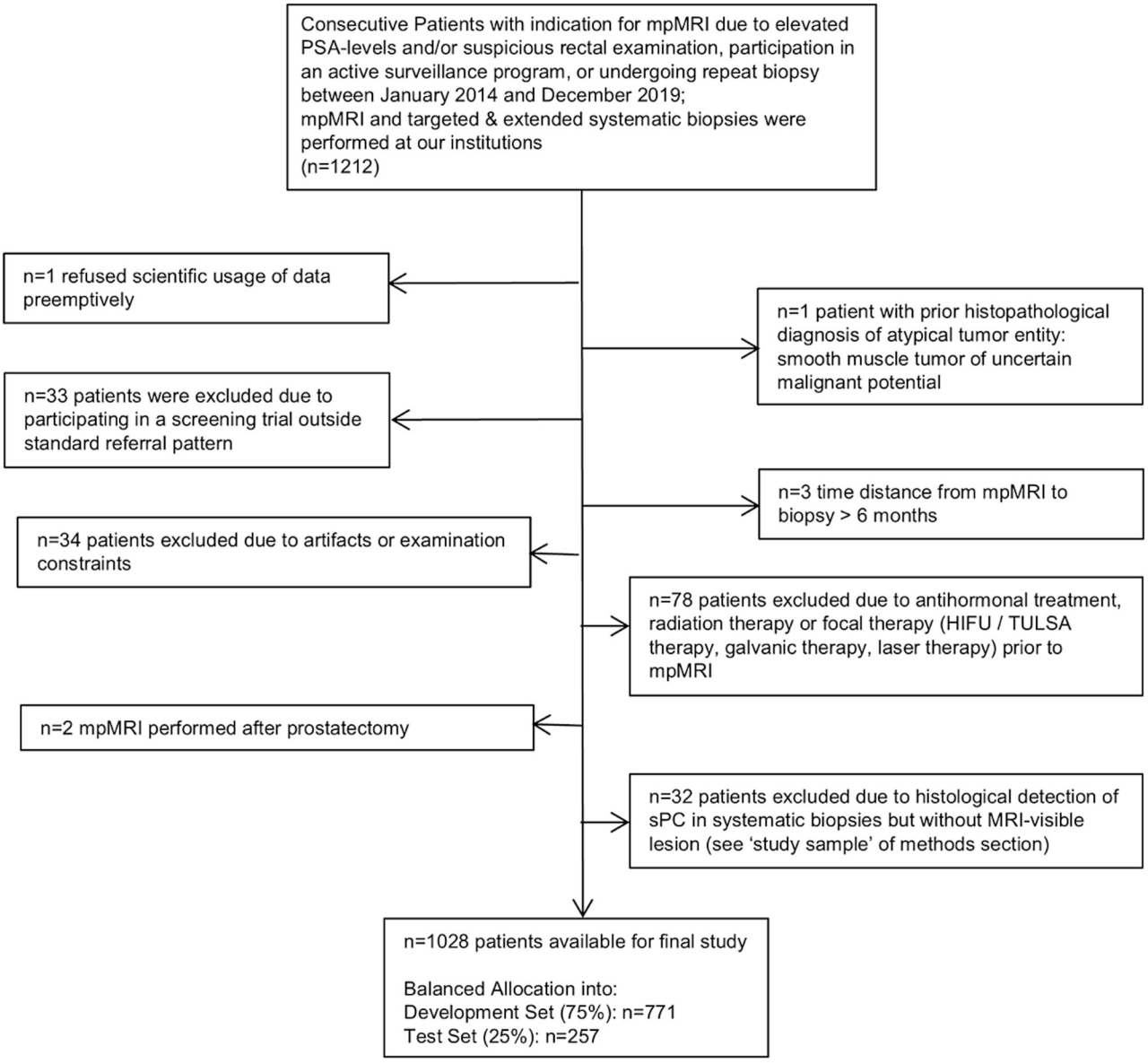

Fig. 2. Inclusion and exclusion diagram of patient population. mpMRI = multiparametric MRI, HIFU = high-intensity focused ultrasound, sPC = significant prostate cancer, TULSA = transurethral ultrasound ablation.

training (Supp. S2).

On the lesion level, compared to the reference model, the FROC curve of the anatomy-informed model suggested slightly higher performance with the latter's curve running above the reference model for all data points (Supp. S3): FROC AUC for the reference model was 0.43 [CI 0.39, 0.48] and for anatomy-informed model 0.47 [CI 0.43, 0.51]. Clustered McNemar test demonstrated an increase in sensitivity close to significance level ($p = 0.07$) with 9 additional detected lesions (Suppl. S5).

On the patient level, the anatomy-informed augmentation scheme also increased the ROC AUC from 79.7% [CI 0.77, 0.83] to 81.8% [CI 0.79, 0.85] ($p = 0.08$, Supp. S4).

3.3. Independent test cohort

The final evaluation of the optimized AI model was performed on the independent test set. On the lesion level, utilization of the augmented model resulted in an increase in sensitivity by 18.8%, from 48 to 57 true-positive lesions, statistically significant according to clustered McNemar test ($p = 0.01$, Supp. S5). FROC analysis also suggested slightly increased performance of the augmented model, except for the low sensitivity range; reference vs. augmented model: 0.33 [CI 0.26, 0.42] vs. 0.37 [CI 0.30, 0.45]; see Fig. 3.

The lesion-level increase in performance was not reflected in the patient-level ROC analysis – ROC AUC increase was marginal ($p = 0.88$, Fig. 5). However, taking the LROC and wAFROC metrics, which consider the localization criterion, i.e., the matching of the model prediction with the ground truth segmentation, the higher performance could be visualized by LROC and wAFROC analysis (Fig. 5). ANOVA testing for the wAFROC metric demonstrated significantly higher performance ($p < 0.01$).

Further investigation of the exams with differences in true positive lesions revealed that the anatomy-informed model correctly predicted lesions in cases where the reference model missed csPCa lesions (5 cases) or where the reference model predicted lesions with insufficient overlap to ground truth (5 cases); except in one case where the augmented model missed a lesion detected by the reference model (see Fig. 4 for examples and Supp. S6 for the detailed list of cases).

Table 1
Demographic and Clinical Characteristics of 1028 Included Exams. PI-RADS = Prostate Imaging Reporting and Data System; IQR = interquartile range; PSA = prostate-specific antigen; ISUP = International Society of Urological Pathology Grade Group.

Characteristic	Development set (75%)	Test Set (25%)
No. exams (Total: 1028)		
	771	257
• Without csPCa	500	167
• With csPCa	271	90
Median age	64 (IQR 58–69)	65 (IQR 59–70)
Mean PSA	8.94 ± 8.79	9.21 ± 7.83
No. exams		
• w/o MRI-detected lesion	26	5
• MRI-detected index lesions		
o PI-RADS 2	79	24
o PI-RADS 3	201	68
o PI-RADS 4	308	113
o PI-RADS 5	157	47
No. csPCa lesions per exam		
• 1 lesion	188	64
• 2 lesions	66	21
• 3 lesions	16	5
• 4 lesions	1	0
csPCa location		
• Transition zone	86	31
• Peripheral zone	268	87
• Multi zone	18	3
ISUP Grade		
• no PCa	327	129
• ISUP 1	173	38
• ISUP 2	156	55
• ISUP 3	48	10
• ISUP 4	22	15
• ISUP 5	45	10

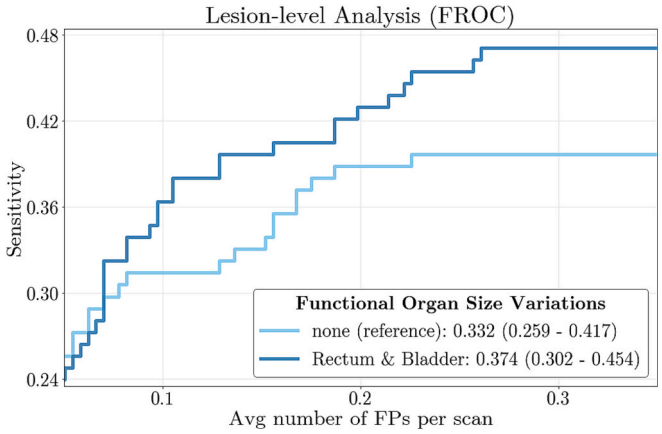

Fig. 3. Free-response receiver operating characteristic (FROC) curves comparing the lesion detection performance of the AI models on the independent test set. The model trained with additional simulated soft-tissue deformations demonstrated slightly improved sensitivity, detecting nine additional true-positive lesions compared to the reference model (10 additionally detected lesions, but 1 missed lesion), with a statistically significant improvement (clustered McNemar test, $p = 0.01$; see Supplement S5). The FROC area under the curve was calculated for the range of 0–0.35 false positives (FPs) per scan and the confidence intervals were calculated by case-based bootstrap.

4. Discussion

In recent years, there has been a growing trend in medical image analysis among researchers to integrate data-specific knowledge into AI model training to improve performance and reliability. This approach has shown promise in various applications, such as providing prostate zonal information for csPCa detection [9,10], generating artificial misalignments between the MRI sequences during training to make the network robust for them by accounting for misalignments and registration errors [16], extending the loss function with topology-preserving term for vessel/nerve segmentation [33,34], or with size-related risk factor for breast cancer detection [35], and using statistical shape and intensity models to augment the geometry of hip bones [36]. Following this trend, our study demonstrated the value of accounting for variations in the functional states of the rectum and bladder in AI model training for csPCa detection, contributing to the ongoing evolution of data-centric AI models.

We evaluated nnU-Net, the current neural network-based AI framework for medical semantic segmentation, which already contains the current augmentation schemes, and compared its performance with and without the proposed 'anatomy-informed' training strategy. We focused on lesion- and patient-level PCa detection performance based on ground truth segmentations of malignant lesions. Given the method's reliance on spatial information, semantic segmentation proved particularly well-suited for testing strategies involving 'anatomy-informed' transformations, addressing localized soft tissue deformations of the prostate.

Incorporating artificial variations in rectal and bladder sizes during model training led to a statistically significant improvement on the lesion-level detection performance in the test set. This improvement was not reflected on the patient-level by the commonly used ROC AUC analysis, which does not incorporate spatial lesion information, e.g. a true positive patient is considered correctly classified if the model predicts a lesion somewhere in the prostate, i.e., the prediction and the ground truth segmentation do not need to overlap (sufficiently). We, therefore, highlighted an important aspect in patient-level model evaluation, namely making the right diagnosis based on the right reason: LROC and wAFROC, metrics consider the spatial information of the model predictions and demonstrate an increase in performance due to the increased number of spatially-correct predictions.

Correct tumor localization is crucial in the outcome of subsequent clinical management and interventions. Correct local staging of prostate cancer enables patients to be included in active surveillance programs [37]. MRI-guided ultrasound fusion biopsies, depending on lesion localization information, have demonstrated significantly increased detection rates compared to standard biopsy alone [38,39]. Traditional treatment options use a holistic approach towards the prostate, i.e. prostatectomy or whole-gland radiation. However, recently location-dependent options are gaining traction. In a completed phase II clinical trial by Ghai et al., ultrasound cancer ablation demonstrated good locoregional control without decline in life quality after treatment [1]. Another approach has added radiation boost to the visible tumor volume when performing whole-gland radiation, which increased significantly the biochemical disease-free survival time without impacting late toxicity or health-related quality of life [2]. In addition, correct localization and determination of tumor extent are essential for TNM classification and risk assessment [3].

The main limitation of this study is its cohort size and single-center character; additionally, most studies were performed on one 3 T scanner. These factors may limit generalizability. However, the homogeneous data quality with expected lower variability of measurements in combination with the high histopathological sampling density obtained in this single-center cohort enhances the focus on biological and anatomical influences on detectability. This likely allowed to more sensitively examine the statistically significant improvements gained from the developed augmentation scheme. Another possible limitation is that this augmentation scheme's beneficial effect may decline with increasing training cohort size, e.g. large cohort like in the PI-CAI challenge [40] may already demonstrate sufficient natural variability. However, there are not sufficient data at the current time to prove or

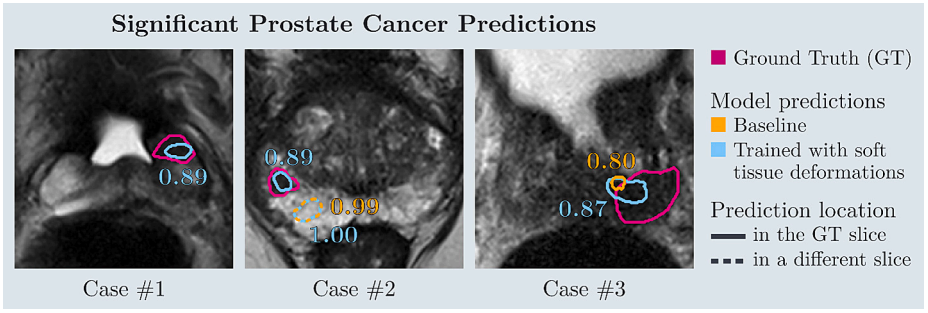

Fig. 4. Ground truth (GT) annotations (pink solid lines), AI model prediction outlines (orange: reference model, blue: anatomy-informed model) and corresponding confidence scores, in the same slice with the GT and false positives in a different slice (dotted line). In the shown cases the anatomy-informed model predicted the GT lesion correctly, meanwhile the reference model either missed it (case #1), predicted it at an incorrect location with high confidence (case #2), or predicted it adjacent to the right location with insufficient overlap (case #3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

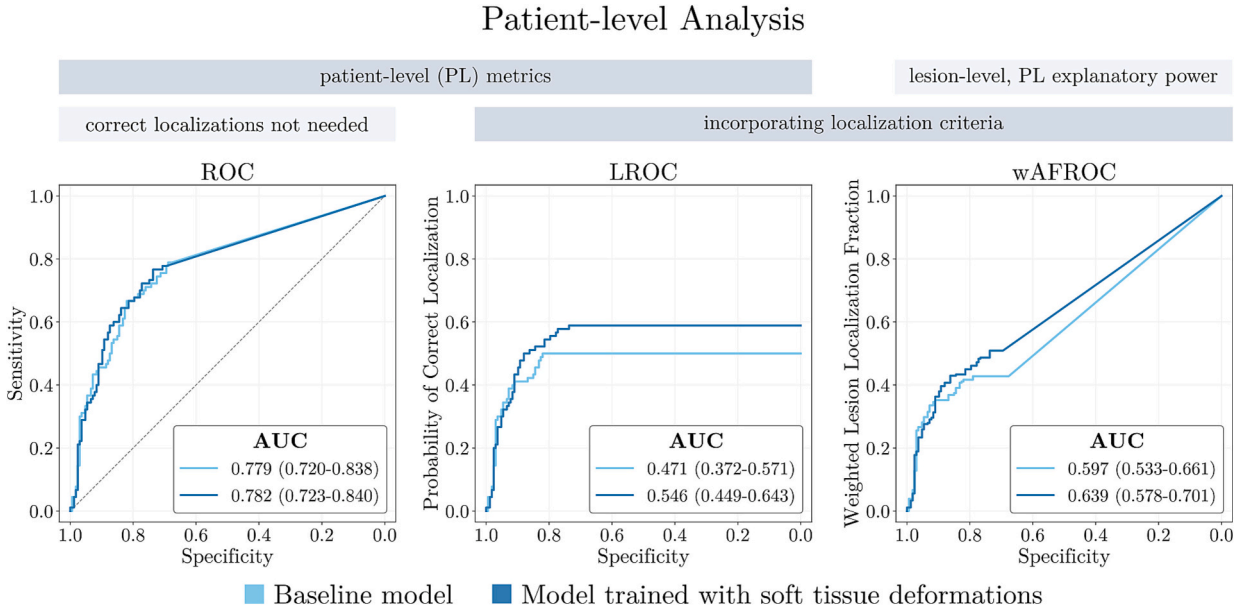

Fig. 5. Patient-level performance comparison of AI models trained without (light blue curves) and with simulated soft tissue deformations (dark blue curves) using receiver operating characteristic (ROC), localization receiver operating characteristic (LROC), and weighted alternative free-response receiver operating characteristic (wAFROC) analyses. The ROC curve shows a marginal increase in area under the curve (AUC) between models ($p = 0.88$, DeLong test). The LROC analysis, which includes the probability of correct localization (PCL) for the highest-scoring lesion, visualizes improved performance when localization is considered, though only one prediction per case is included. The wAFROC analysis addresses the potential multifocality of prostate cancer, providing a lesion-level metric with patient-level interpretability. Respective area under the curve (AUC) values are given in the legends with 95% confidence intervals in brackets. ANOVA testing on the wAFROC score indicated a statistically significant improvement in performance ($p < 0.01$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

disprove this hypothesis. Thus, our results remain valuable given the scale of our study, but the method's full potential is to be further explored in larger and multi-centric data sets.

In conclusion, our results suggest potential benefit of adding an 'anatomy-informed' strategy into standard model training, which may be crucial for location-dependent diagnostic and therapeutic measures in prostate cancer. This approach not only addresses soft-tissue variations effectively but also has the potential be a blueprint for developing data-centric AI models for other applications. In parallel, we highlight the spatial/localization limitation of the commonly used ROC analysis for AI model evaluation.

CRediT authorship contribution statement

Balint Kovacs: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Kevin Sun Zhang:** Writing – review & editing, Writing – original draft, Investigation, Data curation, Conceptualization. **Michael Baumgartner:** Writing – review & editing, Methodology. **Thomas Hielscher:** Writing – review & editing, Visualization, Validation, Methodology, Formal analysis. **Dimitrios Bounias:** Writing – review & editing, Supervision, Conceptualization. **Nils Netzer:** Writing – review & editing, Visualization, Data curation. **Ralf Floca:** Writing – review & editing, Supervision, Software, Conceptualization. **Carolin Eith:** Writing – review & editing, Data curation. **Paul F. Jäger:** Writing – review & editing, Software. **Clara Meinzer:** Writing – review & editing, Data curation. **Fabian Isensee:** Writing – review & editing, Software, Conceptualization. **Regula Gnirs:** Writing – review & editing, Methodology, Data curation. **Magdalena Görtz:** Writing – review & editing, Resources, Data curation. **Markus Hohenfellner:** Writing – review & editing, Supervision, Resources. **Albrecht Stenzinger:** Writing – review & editing, Supervision, Resources, Data curation. **Heinz-Peter Schlemmer:** Writing – review &

editing, Supervision, Resources. **Ivo Wolf:** Writing – review & editing, Supervision, Resources. **Klaus H. Maier-Hein:** Writing – review & editing, Supervision, Resources. **David Bonekamp:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Data curation, Conceptualization.

Declaration of competing interest

Balint Kovacs reports financial support, equipment, drugs, or supplies, and travel were provided by HIDSS4Health - Helmholtz Information and Data Science School for Health. **Michael Baumgartner** reports financial support was provided by Helmholtz Imaging, a platform of the Helmholtz Incubator on Information and Data Science. **Clara Meinzer** reports financial support was provided by Bundesministerium für Wirtschaft und Klimaschutz (BMWK): 01MT21004B. **Magdalena Görtz** reports a relationship with Bayer Vital GmbH that includes: consulting or advisory and speaking and lecture fees. **Magdalena Görtz** reports a relationship with AstraZeneca that includes: consulting or advisory and speaking and lecture fees. **Albrecht Stenzinger** reports a relationship with AstraZeneca that includes: board membership, consulting or advisory, speaking and lecture fees, and travel reimbursement. **Albrecht Stenzinger** reports a relationship with Novartis that includes: board membership, consulting or advisory, speaking and lecture fees, and travel reimbursement. **Albrecht Stenzinger** reports a relationship with Illumina Inc that includes: consulting or advisory and speaking and lecture fees. **Albrecht Stenzinger** reports a relationship with Bristol Myers Squibb that includes: board membership, consulting or advisory, speaking and lecture fees, and travel reimbursement. **Albrecht Stenzinger** reports a relationship with Roche that includes: consulting or advisory and speaking and lecture fees. **Albrecht Stenzinger** reports a relationship with Thermo Fischer that includes: board membership, consulting or advisory, speaking and lecture fees, and travel reimbursement. **Heinz-Peter Schlemmer** reports a relationship with Bracco Imaging Deutschland GmbH that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. **Heinz-Peter Schlemmer** reports a relationship with Siemens that includes: speaking and lecture fees. **Heinz-Peter Schlemmer** reports a relationship with Bayer Vital GmbH that includes: speaking and lecture fees and travel reimbursement. **Klaus Maier-Hein** reports financial support, equipment, drugs, or supplies, and travel were provided by HIDSS4Health - Helmholtz Information and Data Science School for Health. **David Bonekamp** reports a relationship with Bayer Vital GmbH that includes: speaking and lecture fees. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mri.2026.110698.

References

[1] Ghai S, Finelli A, Corr K, Lajkosz K, McCluskey S, Chan R, et al. MRI-guided focused ultrasound focal therapy for intermediate-risk prostate cancer: final results from a 2-year phase II clinical trial. Radiology 2024;310(3):e231473. https://doi.org/10.1148/radiol.231473.

[2] Kerkmeijer LGW, Groen VH, Pos FJ, Haustermans K, Monninkhof EM, Smeenk RJ, et al. Focal boost to the intraprostatic tumor in external beam radiotherapy for patients with localized prostate cancer: results from the FLAME randomized phase III trial. J Clin Oncol 2021;39(7):787–96. https://doi.org/10.1200/jco.20.02873.

[3] Lin Y, Johnson LA, Fennessy FM, Turkbey B. Prostate Cancer local staging with magnetic resonance imaging. Radiol Clin North Am 2024;62(1):93–108. https://doi.org/10.1016/j.rcl.2023.06.010.

[4] Turkbey B, Choyke PL. Future perspectives and challenges of prostate MR imaging. Radiol Clin North Am 2018;56(2):327–37. https://doi.org/10.1016/j.rcl.2017.10.013.

[5] Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. Eur Urol 2019;76(3):340–51. https://doi.org/10.1016/j.eururo.2019.02.033.

[6] Boubaker MB, Ganghoffer J-F. Chapter 14 - bladder/prostate/rectum: Biomechanical models of the mobility of pelvic organs in the context of prostate radiotherapy. In: Payan Y, Ohayon J, editors. Biomechanics of living organs. Oxford: Academic Press; 2017. p. 307–24.

[7] Twilt JJ, van Leeuwen KG, Huisman HJ, Fütterer JJ, de Rooij M. Artificial intelligence based algorithms for prostate Cancer classification and detection on magnetic resonance imaging: a narrative review. Diagnostics (Basel) 2021;11(6). https://doi.org/10.3390/diagnostics11060959.

[8] Schrader A, Netzer N, Hielscher T, Görtz M, Zhang KS, Schütz V, et al. Prostate cancer risk assessment and avoidance of prostate biopsies using fully automatic deep learning in prostate MRI: comparison to PI-RADS and integration with clinical data in nomograms. Eur Radiol 2024. https://doi.org/10.1007/s00330-024-10818-0.

[9] Hosseinzadeh M, Saha A, Brand P, Slootweg I, de Rooij M, Huisman H. Deep learning-assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. Eur Radiol 2022;32(4):2224–34. https://doi.org/10.1007/s00330-021-08320-y.

[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. Med Image Anal 2021;73:102155. https://doi.org/10.1016/j.media.2021.102155.

[11] Netzer N, Weisser C, Schelb P, Wang X, Qin X, Gortz M, et al. Fully automatic deep learning in bi-institutional prostate magnetic resonance imaging: effects of cohort size and heterogeneity. Investig Radiol 2021;56(12):799–808. https://doi.org/10.1097/RLI.0000000000000791.

[12] Kovacs B, Netzer N, Baumgartner M, Eith C, Bounias D, Meinzer C, et al. Anatomy-informed data augmentation for enhanced prostate cancer detection. In: Medical image computing and computer assisted intervention – MICCAI 2023: 26th international conference, Vancouver, BC, Canada, October 8–12, 2023, proceedings, part VII. Vancouver, BC, Canada: Springer-Verlag; 2023. p. 531–40.

[13] Hu Y, Gibson E, Ghavami N, Bonmati E, Moore CM, Emberton M, et al. Adversarial deformation regularization for training image registration neural networks. In: *Medical image computing and computer assisted intervention – MICCAI* 2018: 21st international conference, Granada, Spain, September 16-20, 2018, proceedings, part I. Granada, Spain: Springer-Verlag; 2018. p. 774–82.

[14] Khallaghi S, Sánchez CA, Rasoulian A, Nouranian S, Romagnoli C, Abdi H, et al. Statistical biomechanical surface registration: application to MR-TRUS fusion for prostate interventions. IEEE Trans Med Imaging 2015;34(12):2535–49. https://doi.org/10.1109/tmi.2015.2443978.

[15] 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(2):203–11. https://doi.org/10.1038/s41592-020-01008-z.

[16] Kovacs B, Netzer N, Baumgartner M, Schrader A, Isensee F, Weißer C, et al. Addressing image misalignments in multi-parametric prostate MRI for enhanced computer-aided diagnosis of prostate cancer. Sci Rep 2023;13(1):19805. https://doi.org/10.1038/s41598-023-46747-z.

[17] Barentsz JO, Richenberg J, Clements R, Choyke P, Verma S, Villeirs G, et al. ESUR prostate MR guidelines 2012. Eur Radiol 2012;22(4):746–57. https://doi.org/10.1007/s00330-011-2377-y.

[18] Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, et al. PI-RADS prostate imaging - reporting and data system: 2015, version 2. Eur Urol 2016;69(1):16–40. https://doi.org/10.1016/j.eururo.2015.08.052.

[19] Kuru TH, Wadhwa K, Chang RT, Echeverria LM, Roethke M, Polson A, et al. Definitions of terms, processes and a minimum dataset for transperineal prostate biopsies: a standardization approach of the Ginsburg study Group for Enhanced Prostate Diagnostics. BJU Int 2013;112(5):568–77. https://doi.org/10.1111/bju.12132.

[20] Wolf I, Vetter M, Wegner I, Böttger T, Nolden M, Schöbinger M, et al. The medical imaging interaction toolkit. Med Image Anal 2005;9(6):594–604. https://doi.org/10.1016/j.media.2005.04.005.

[21] Radtke JP, Schwab C, Wolf MB, Freitag MT, Alt CD, Kesch C, et al. Multiparametric magnetic resonance imaging (MRI) and MRI-transrectal ultrasound fusion biopsy for index tumor detection: correlation with radical prostatectomy specimen. Eur Urol 2016;70(5):846–53. https://doi.org/10.1016/j.eururo.2015.12.052.

[22] Engels RRM, Israël B, Padhani AR, Barentsz JO. Multiparametric magnetic resonance imaging for the detection of clinically significant prostate Cancer: what urologists need to know. Part 1: acquisition. Eur Urol 2020;77(4):457–68. https://doi.org/10.1016/j.eururo.2019.09.021.

[23] Rubod C, Brieu M, Cosson M, Rivaux G, Clay J-C, de Landsheere L, et al. Biomechanical properties of human pelvic organs. Urology 2012;79(4). https://doi.org/10.1016/j.urology.2011.11.010. 968.e17-.e22.

[24] Qasim M, Puigjaner D, Herrero J, López JM, Olivé C, Fortuny G, et al. Biomechanical modelling of the pelvic system: improving the accuracy of the location of neoplasms in MRI-TRUS fusion prostate biopsy. BMC Cancer 2022;22(1):338. https://doi.org/10.1186/s12885-022-09432-4.

[25] Wang Y, Ni D, Qin J, Xu M, Xie X, Heng P-A. Patient-specific deformation modelling via elastography: application to image-guided prostate interventions. Sci Rep 2016;6(1):27386. https://doi.org/10.1038/srep27386.

[26] Sanyal J, Banerjee I, Hahn L, Rubin D. An automated two-step pipeline for aggressive prostate lesion detection from multi-parametric MR sequence. AMIA Jt Summits Transl Sci Proc 2020;2020:552–60.

[27] Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods 2024;21(2):195–212. https://doi.org/10.1038/s41592-023-02151-z.
[28] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–45.
[29] He X, Frey E. ROC, LROC, FROC, AFROC: an alphabet soup. J Am Coll Radiol 2009; 6(9):652–5. https://doi.org/10.1016/j.jacr.2009.06.001.
[30] Obuchowski Jr NA, Rockette Jr HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations. Commun Statist - Simulat Comput 1995;24(2):285–308. https://doi.org/10.1080/03610919508813243.
[31] Dev Chakraborty XZ. RJafroc: Artificial intelligence systems and observer performance. R package version 2.1.3. 2023.
[32] Durkalski VL, Palesch YY, Lipsitz SR, Rust PF. Analysis of clustered matched-pair data. Stat Med 2003;22(15):2417–28. https://doi.org/10.1002/sim.1438.
[33] Shit S, Paetzold JC, Sekuboyina A, Ezhov I, Unger A, Zhylka A, et al. clDice - a novel topology-preserving loss function for tubular structure segmentation. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE Computer Society; 2021. p. 16555–64.
[34] Kirchhoff Y, Rokuss MR, Roy S, Kovacs B, Ulrich C, Wald T, et al. Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures. arXiv preprint arXiv:240403010. 2024.
[35] Bounias D, Baumgartner M, Neher P, Kovacs B, Floca R, Jaeger PF, et al. Risk-adjusted training and evaluation for medical object detection in breast cancer MRI. In: ICML 3rd workshop on interpretable machine learning in healthcare (IMLH); 2023.
[36] Schmid J, Assassi L, Chênes C. A novel image augmentation based on statistical shape and intensity models: application to the segmentation of hip bones from CT images. Eur Radiol Exp 2023;7(1):39. https://doi.org/10.1186/s41747-023-00357-6.
[37] Bolenz C, Grimm MO, Heidenreich A, Kristiansen G, Schimmöller L, Schmidt S, et al. Active surveillance of prostate cancer. Urologie 2025;64(6):595–605. https://doi.org/10.1007/s00120-025-02555-z.
[38] Ahdoot M, Wilbur AR, Reese SE, Lebastchi AH, Mehralivand S, Gomella PT, et al. MRI-targeted, systematic, and combined biopsy for prostate Cancer diagnosis. N Engl J Med 2020;382(10):917–28. https://doi.org/10.1056/NEJMoa1910038.
[39] Drost FH, Osses D, Nieboer D, Bangma CH, Steyerberg EW, Roobol MJ, et al. Prostate magnetic resonance imaging, with or without magnetic resonance imaging-targeted biopsy, and systematic biopsy for detecting prostate Cancer: A Cochrane systematic review and meta-analysis. Eur Urol 2020;77(1):78–94. https://doi.org/10.1016/j.eururo.2019.06.023.
[40] Saha A, Bosma JS, Twilt JJ, van Ginneken B, Bjartell A, Padhani AR, et al. Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. Lancet Oncol 2024;25 (7):879–87. https://doi.org/10.1016/s1470-2045(24)00220-1.