scientific reports

Check for updates

OPEN

Addressing image misalignments in multi-parametric prostate MRI for enhanced computer-aided diagnosis of prostate cancer

Balint Kovacs[1,2,3]✉, Nils Netzer[2,3], Michael Baumgartner[1,4,5], Adrian Schrader[2,3], Fabian Isensee[1,4], Cedric Weißer[2,3], Ivo Wolf[6], Magdalena Görtz[7,8], Paul F. Jaeger[4,9], Victoria Schütz[8], Ralf Floca[1], Regula Gnirs[2], Albrecht Stenzinger[10], Markus Hohenfellner[8], Heinz-Peter Schlemmer[2,11], David Bonekamp[2,3,11,13] & Klaus H. Maier-Hein[1,4,11,12,13]

Prostate cancer (PCa) diagnosis on multi-parametric magnetic resonance images (MRI) requires radiologists with a high level of expertise. Misalignments between the MRI sequences can be caused by patient movement, elastic soft-tissue deformations, and imaging artifacts. They further increase the complexity of the task prompting radiologists to interpret the images. Recently, computer-aided diagnosis (CAD) tools have demonstrated potential for PCa diagnosis typically relying on complex co-registration of the input modalities. However, there is no consensus among research groups on whether CAD systems profit from using registration. Furthermore, alternative strategies to handle multi-modal misalignments have not been explored so far. Our study introduces and compares different strategies to cope with image misalignments and evaluates them regarding to their direct effect on diagnostic accuracy of PCa. In addition to established registration algorithms, we propose 'misalignment augmentation' as a concept to increase CAD robustness. As the results demonstrate, misalignment augmentations can not only compensate for a complete lack of registration, but if used in conjunction with registration, also improve the overall performance on an independent test set.

Diagnosis of prostate cancer (PCa) is one of the most challenging tasks in oncology due to its complex diagnostic chain[1–4]. Magnetic resonance imaging (MRI) is quickly becoming the standard of care for pre-biopsy evaluation to determine whether targets are present for trans-rectal ultrasound-guided stereotactic fusion biopsy. MRI interpretation has been standardized by the Prostate Imaging Reporting and Data System (PI-RADS)[5–7], currently in version 2.1[8]. According to PI-RADS, standard prostate MRI is acquired in a multi-parametric (mp) fashion, including T2-weighted (T2w), diffusion-weighted (DWI) and dynamic contrast-enhanced (DCE) sequences. Regarding localized PCa, the MRI index lesion, which is the most suspicious lesion according to the zone-specific MRI appearance, dictates the overall PI-RADS score. The patient-based assessment both for PI-RADS in terms of the highest PI-RADS score of any lesion and the whole-prostate highest Gleason Grade Group histopathological assessment, convey the most important clinical information for treatment decisions[7].

[1]Division of Medical Image Computing, German Cancer Research Center (DKFZ) Heidelberg, Im Neuenheimer Feld 223, 69120 Heidelberg, Germany. [2]Division of Radiology, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany. [3]Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany. [4]Helmholtz Imaging, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany. [5]Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany. [6]Mannheim University of Applied Sciences, Mannheim, Germany. [7]Junior Clinical Cooperation Unit 'Multiparametric Methods for Early Detection of Prostate Cancer', German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany. [8]Department of Urology, University of Heidelberg Medical Center, Heidelberg, Germany. [9]Interactive Machine Learning Group, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany. [10]Institute of Pathology, University of Heidelberg Medical Center, Heidelberg, Germany. [11]German Cancer Consortium (DKTK), DKFZ, Core Center Heidelberg, Heidelberg, Germany. [12]Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany. [13]These authors contributed equally: David Bonekamp and Klaus H. Maier-Hein. ✉email: balint.kovacs@dkfz-heidelberg.de

While PI-RADS has led to better standardization, its inter-rater variability remains high[9,10] and interpreting the MRI images requires radiologists with a high level of task-specific expertise[11]. Thus, there is an increasing interest in clinically applicable computer-aided diagnosis (CAD) tools to support radiologists in prostate MRI decision making, and especially in convolutional neural network (CNN) based systems[12,13] due to their recent successful application to complex clinical problems[14–16]. Due to its inherent interpretability and clinical value, approaches based on semantic segmentation are the most popular image analysis task in the biomedical community, which have already been successfully applied to the diagnosis and intervention planning of PCa[17–24]. Semantic segmentation has become even more applicable due to the self-configuring framework of nnU-Net[25] for medical images. It allows a standardized state-of-the-art performance and it has already been successfully adopted for patient-level PCa diagnosis[23,26].

Though CAD systems achieve outstanding performances, there is no common consensus on how to handle misalignments that occur between acquisitions. Despite instructions for patients to remain still during acquisition that typically lasts 30 minutes or more, slight movements can still lead to misalignments between sequences. Even in shorter time frames, soft tissue deformation may occur due to involuntary muscle contractions e.g. of the bowel or slow bladder filling, resulting in local elastic misalignments[27]. Additionally, each MRI sequence, in other words modality, responds differently to prostate tissue properties, leading to a unique contrast characteristic and thus in different lesion contours. Importantly, geometric distortions resulting from susceptibility effects can lead to different positioning of tissue depending on the type of imaging sequence[27], especially between the T2w and DWI sequences[28–30]. These misalignments can result in prostate lesions being notably misaligned between modalities, and thereby performing clinical segmentations on one modality does not ensure accurate lesion ground truth segmentation on the other modalities. Consequently, this can potentially limit model performance, particularly in applications that rely on semantic segmentation being sensitive to spatial information.

Co-registration of the input image modalities is usually performed as a pre-processing step to eliminate misalignments between the sequences, thereby matching the ground truth segmentations to all modalities[31]. Some research groups have implemented complex deformable registration algorithms trying to align the anatomical structures on top of each other correcting local elastic misalignments additionally to the global misalignments[26,32]. To avoid implausible registration transformations, of which risk is especially present in local regions around the prostate gland where the similarity metric fails due to image artifacts like rectal gas-induced artifacts[27–29], many studies limit their registration to non-elastic deformation fields[17,22,33]. Other research studies[23,24,34] claim that standardized precautionary measures (namely minimal temporal difference between acquisitions, administration of antispasmodic agents to reduce bowel motility, use of rectal catheter to minimize distension[27,35,36]) are sufficient to eliminate the need for registration[23] supporting these claims with visual observations.

Although the publications above presented valuable insights, the influence of registration on the diagnostic performance of CAD systems remains unexplored. Studies that additionally applied non-rigid transformations also did not investigate whether their downstream task benefited from that complex registration or not. It is known that registration methods are seldomly perfect, particularly in the case of prostate MRI, with its often anisotropic voxel dimensions required to guarantee sufficient signal-to-noise ratio and with the common deformations especially occurring on the susceptible diffusion sequences. Consequently, misalignments may remain and could cause issues in CNN training and inference. Furthermore, the use of alternative strategies to handle multi-modal misalignments in PCa diagnosis has not been studied so far.

In this work we are investigating multiple strategies for dealing with misalignments for enhanced PCa diagnosis. Importantly, we base our conclusions on the performance of the clinical downstream task of patient-level PCa diagnosis derived from the CAD system as opposed to surrogate alignment or similarity measures. Our results do not only underline the importance of registration, but we propose to tackle the problem of misalignments and remaining registration errors by introducing the new strategy of misalignment augmentation, a data augmentation technique that simulates additional, probabilistic registration errors on-the-fly during training. By substantially augmenting the diversity of misalignments in the training data, the hope is to teach CNNs to become robust and invariant to them to some extent, thus increasing their performance on unknown datasets. Finally, we investigate the complementary value of misalignment augmentation in combination with registration, demonstrating improvements that go beyond pure registration-based techniques.

Results

We evaluated different strategies to deal with misalignments on the clinical task of patient-level PCa diagnosis opposed to surrogate alignment or similarity measures. As the current state-of-the-art method for PCa diagnosis, we derived the diagnosis through semantic segmentation of malignant lesions. Given that semantic segmentation directly depends on spatial information, it was particularly well suited for testing strategies for dealing with misalignments. We evaluated two strategies with different objectives:

- we co-register the MRI sequences using B-spline registration to eliminate misalignments and to match the ground truth segmentations with all image modalities,
- we propose misalignment augmentation, a new strategy to make CNNs robust for misalignments and to tackle remaining registration errors.

Our evaluation approach aimed to disconnect the assessment of strategies for handling misalignments from surrogate alignment measures and focus on their performance in the clinical task. We supported this approach by investigating the relationship between lesion-wise Dice score density functions and patient-wise AUROC values for different registration techniques.

We summarize the results of our systematic analysis of the effect of B-spline registration and misalignment augmentation on an independent multicentric test set consisting of 129 bi-parametric MRI (bpMRI) exams with biopsy-confirmed diagnosis. The following results not only show the importance of registration, but also underline the importance of misalignment augmentation due to its induced high regularization effect. Furthermore, the results reveal their complementary effect.

Achieving the most robust setup by combining registration and misalignment augmentation

To find out the most robust setup against multi-modal image misalignments, the effects of registration and misalignment augmentation are systematically evaluated on the patient-level area under the receiver operating characteristic curve (AUROC). The results are summarized in Table 1.

Addressing misalignments either with registration or with misalignment augmentation increased the AUROC value compared to the unregistered dataset, but neither of these improvements reached statistical significance, with p-values of $p = 0.31$ and $p = 0.11$, respectively. However, combining these strategies reached the highest AUROC value with statistically significant improvement ($p = 0.02$) compared to the unregistered dataset without misalignment augmentation, indicating that this is the most robust setting addressing misalignments.

Furthermore, we evaluated individually our methods on each dataset included for this study (PROSTATEx and in-house dataset) and found that the AUROC values were consistent, indicating that our approach is effective across the two data centers. The AUROC results for each dataset can be found in 'Supplementary Table S4'.

The test results are presented in more detail in the following subsections.

Predictive performance competitive with radiologists

To highlight the practical advantages of techniques addressing misalignments, we compare the performance of the trained models to radiologists' interpretation on our cohort. We calculated the ROC curves for every model and the radiologists' performance for PI-RADS ≥ 3 and ≥ 4, which locate clinically informative performance points on the ROC diagram. Figure 1 illustrates the effect of registration, misalignment augmentation and their combination on the ROC. The radiologists' performance with PI-RADS ≥ 3 and ≥ 4 with specificity and sensitivity of (0.21, 0.98) and (0.56, 0.91), respectively, are marked in Fig. 1 too.

Using either registration or misalignment augmentation increased the sensitivity in the low and high specificity area compared to the unregistered dataset without misalignment augmentation. The performance of the registered dataset closely match, the unregistered dataset with misalignment augmentation sightly exceed the PI-RADS ≥ 3 point, but none of them increased the sensitivity towards PI-RADS ≥ 4. However, their combination was the only setting, which closely matches both clinical PI-RADS performance points and increased the sensitivity over the widest specificity range.

Qualitative results support our findings

To provide a short visual insight, the influence of registration, misalignment augmentation and their combination on the predicted segmentation results are visualized. Figure 2 shows the manual annotations as well as the lesions identified by semantic segmentation in terms of a color scale for a patient exam with one large lesion (PI-RADS 4, GS 7a) and with one puncture lesion (PI-RADS 4, GS 7a) in the left prostate base.

Registration and misalignment augmentation increased the predicted lesion volume for both datasets resulted in a more complete coverage of the pathological lesions. Additionally, misalignment augmentation enabled segmentation of a small punctate lesion in the PZ too, which was not picked up by the other configurations.

Proposed method on par with method that uses ground-truth human segmentation

To support the quality of the used B-spline registration method for our CAD system, we compare it with the reference ground-truth-matching (GT-matching) registration. For that, a lesion overlap metric and ROC of the actual clinical task were used. Figure 3 shows the Dice score probability density functions for our dataset using different registration techniques, as well as without any registration.

The mean and standard deviation of the Dice score for the unregistered dataset of 0.44 ± 0.21 is improved to 0.48 ± 0.19 using the GT-matching and to 0.50 ± 0.18 using the B-spline registration slightly outperforming the reference. The corresponding ROC curves are illustrated in Fig. 4.

Both registration techniques, B-spline registration and GT-matching, increased the AUROC value by 3.2% and 3.5%, respectively, compared to the unregistered dataset. However, the differences in the Dice score density functions between the two registered datasets did not translate to the differences in the corresponding AUROC performances.

AUROC (test set)	Unregistered dataset	Registered dataset
Without misalignment augmentation	75.93% (CI: 67.49–83.68, reference)	79.11% (CI: 70.95–86.93, $p = 0.31$)
With misalignment augmentation	80.13% (CI: 71.57–87.18, $p = 0.11$)	82.07% (CI: 74.18–89.38, $p = 0.02$)

Table 1. Patient-level AUROC results with 95% confidence intervals (CI) and p-values resulted from the DeLong test on the independent test set. Addressing misalignments either with registration or with misalignment augmentation increased the AUROC value compared to the unregistered dataset without misalignment augmentation. The highest AUROC value with statistically significant improvement is reached by combining registration and misalignment augmentation.

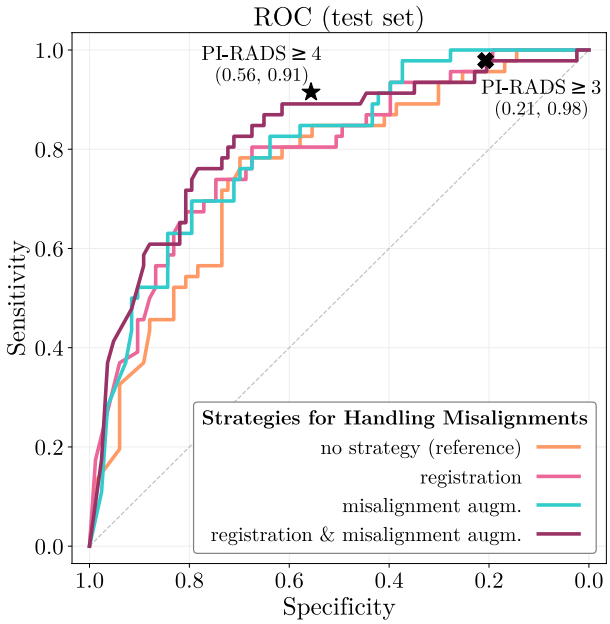

Figure 1. Predictive performance comparison of the trained models and the radiologists on the ROC. The radiologists' performance with PI-RADS ≥ 3 and ≥ 4 are marked to locate clinically informative performance points. The sensitivity in the low and high specificity area is increased using either registration or misalignment augmentation compared to the unregistered dataset. Their performance closely match and slightly exceed the PI-RADS ≥ 3 point, respectively, but none of them improved the sensitivity towards PI-RADS ≥ 4. Combining registration and misalignment augmentation closely matches both PI-RADS performance points with the highest sensitivity improvement over the widest specificity range.

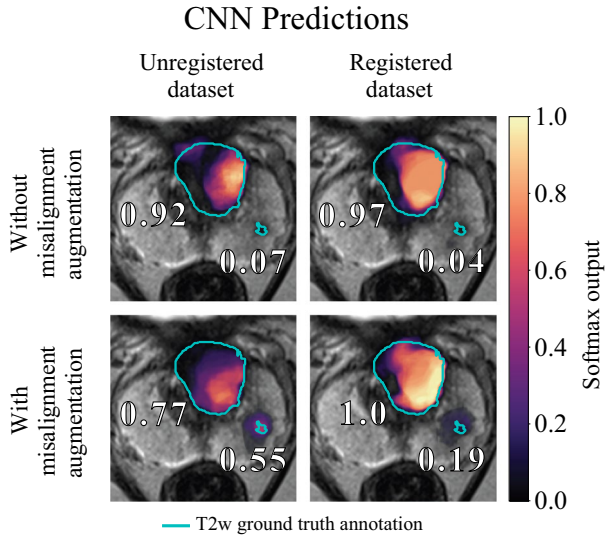

Figure 2. Comparison of manual annotations and segmentation predictions for a patient exam with one large lesion (PI-RADS 4, GS 7a) and with one puncture lesion (PI-RADS 4, GS 7a) in the left prostate base. On a representative T2w slice of the index lesion manual annotations given as a cyan outline as well as the probability map from the different settings superimposed using a colormap, with the transparency of probability values raised to the power of 0.2. CNN probabilities are given in the first row for the unregistered dataset and for the registered CAD dataset. The bottom row shows the same datasets in the first row with misalignment augmentation. The highest probability values for the individual lesions are marked lesion-wise too. All of the CNNs predicted the large index lesion with high confidence. However, the highest predictions for the second punctate lesion belong to settings with misalignment augmentations.

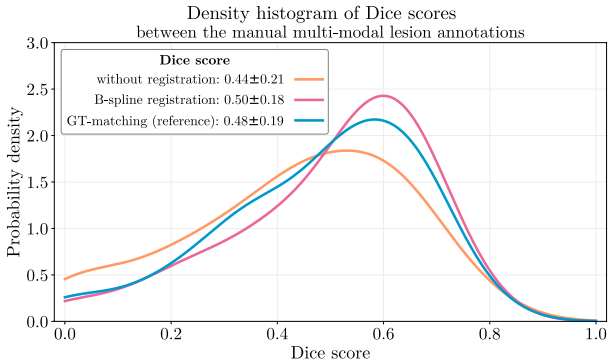

Figure 3. Dice score probability density functions between the manual lesion annotations in the T2w and ADC modalities for our dataset without registration, with B-spline registration, and with the reference GT-matching. Mean and standard deviation values are calculated for all settings and located in the corresponding legend entries. The best Dice score distribution belongs to the B-spline registration sightly outperforming the reference GT-matching.

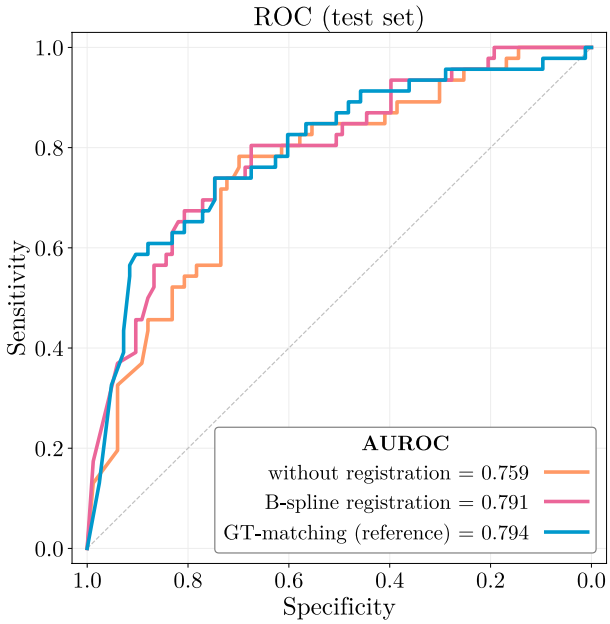

Figure 4. The influence of the different registration techniques on the predictive performance. Using B-spline registration and the reference GT-matching, increased the AUROC value of the unregistered dataset by 3.2% and 3.5%, respectively, especially the sensitivity in the low and high specificity ranges.

Discussion

This paper systematically evaluates and highlights the importance of handling multi-modal image misalignments both by registration and our proposed misalignment augmentation for enhanced patient-level prostate cancer diagnosis on a multicentric prostate bpMRI dataset.

Registering the MRI sequences increased the patient-level AUROC compared to the unregistered dataset. The results indicate that aligning the anatomical structures on top of each other helped the CAD system to couple multi-modal information, which leads to improved downstream performance. Furthermore, the lack of the performance difference in the AUROC values of the different registration techniques compared to the observable differences in the Dice score density functions indicates that the surrogate measures for evaluating registration performance do not necessarily translate to the clinical downstream task performance. This underlines the importance of our study evaluating the quality of registration on the clinical task of patient-level PCa diagnosis. Despite the observed improvement in the AUROC value resulting from registering the MRI sequences, the improvement was not statistically significant.

Generating augmented misalignments between the modalities during the training of CNNs made the diagnostic task harder and forced the network to become more invariant for misalignments. We showed that using these augmentations on the unregistered dataset could replace registration due to its induced regularization effect.

Thus, misalignment augmentations have been shown to be able to address image misalignments and potentially simplify preprocessing by replacing complex non-rigid registration techniques in cases, where the anatomical structures are already partially overlapped. It is important to note that the ability of the method to cope with misalignments is limited depending on the initial overlap between the image modalities, and the type and amplitude of the augmentations. Thus, its parameters have to be adapted to the exact clinical need, and an initial registration is still needed. However, the improvement was not statistically significant, similarly to registration.

Although we addressed the same problem with registration and misalignment augmentation their combination resulted in further improvement in the ROC with a statistically significant difference compared to the unregistered dataset. A potential explanation for this complementary behavior and for the statistically significant improvement could be that these strategies extended each other's limitations. While registration provides anatomical matching for learning more complex features, misalignment augmentation could further increase the robustness of the CNNs against remaining registration errors. It is worth mentioning that in the case of larger datasets, the model's capability to autonomously learn to cope with misalignments without assistance remains an open question. The inherent size and diversity of the data might enable the model to naturally adjust to the broad spectrum of misalignments to some extent. However, in the context of medical datasets with limited size, registration and in particular augmentation techniques, like misalignment augmentation play a vital role in inducing application-specific knowledge, effectively enhancing the model's ability to adapt to various misalignments and maintain robust performance. This configuration not only reached the highest AUROC value on the test set, but it also pushed the ROC curve closer to the radiologists' performance closely matching PI-RADS $\geq$ 4, which was not reached by any other configurations. Our experiments indicate that combining registration with misalignment augmentation is the optimal configuration for the training of multi-modal prostate MRI datasets. Furthermore, our results suggest integrating the paradigm of misalignment augmentation into the standard repertoire of CNN design for prostate MRI studies and utilizing this data augmentation as a blueprint for other multi-modal applications.

Methods

Prostate MRI cohort

We included 625 examinations from two cohorts in this study: 204 exams from the publicly available PROSTATEx[37] challenge dataset and an in-house cohort consisting of 421 consecutive examinations acquired during clinical routine from 2014 to 2016. Exams from the in-house cohort were previously published and included in this study following the same inclusion criteria as in[18,19,26,38]. The ethics committee of the Medical Faculty Heidelberg approved the study and waived informed consent (institutional ethics approval number S-164/2019) to enable analysis of a consecutive cohort. All experiments were performed in accordance with the declaration of Helsinki (64th WMA General Assembly, Fortaleza, Brazil, October 2013) and relevant data privacy regulations. More details about our cohort can be seen in 'Supplementary Tables S1 and S2'.

Although the PI-RADS manual suggests mpMRI including DCE for prostate MRI assessment, bi-parametric MRI (bpMRI) abbreviated protocols, consisting only of T2-weighted imaging (T2w) and DWIs, are being discussed to be of acceptable diagnostic quality. bpMRI may allow a minimal trade-off in diagnostic quality at the advantage of sparing contrast agent administration and gaining faster image acquisition[39], properties which also have made bpMRI an attractive design choice for promising pioneering deep learning[40,41] applications in prostate MRI. Therefore, T2w, DWIs with high b-value, and ADC maps were used for this study.

For all patients, PI-RADS (either v1, v2[5] or v2.1[8]) interpretation was performed by board-certified radiologists during clinical routine. Where only PI-RADS v1 was available in the clinical report, a retrospective PI-RADS v2 read was performed by a board-certified radiologist. These clinical reports are used for the lesion annotation process and the evaluation of the human performance.

Based on the clinical reports, the lesions of the in-house cohort were segmented retrospectively on both modalities by multiple in-house investigators using the Medical Imaging Interaction Toolkit (MITK)[42,43] under the supervision of a board-certified radiologist with more than 13 years of experience in prostate MRI interpretation (D.B.), as previously described in[19,26,38]. Lesions in the PROSTATEx training dataset were manually segmented by the same investigators using the publicly provided lesion coordinates. The lesion segmentations were used as the ground truth for the semantic segmentation task. Additionally, in a subset of the cohort the prostate gland was also segmented manually, then prostate segmentation CNNs trained and segmentation proposals generated for the remainder of the examinations, which were reviewed and manually edited. The prostate masks were not only used for the co-registration of the modalities for creating a reference dataset (see later in section "Registration methods and annotations for training"), but also for selecting the clinically relevant lesions by combining the different biopsy approaches, see later in this section.

In-house patients underwent MRI trans-rectal ultrasound-fusion transperineal biopsy using real-time registration between the T2w and ultrasound images. The evaluation of the histopathological samples was performed according to the International Society of Urological Pathology (ISUP) standards under the supervision of a dedicated uropathologist with more than 19 years of experience (A.S.). Clinically significant prostate cancer (csPCa) was defined as ISUP grade 2 or higher[44,45]. Findings for the PROSTATEx cohort are also biopsy confirmed where csPCa was defined by Gleason score (GS) 7 or higher.

The patient-wise ground truth is determined as the maximum ISUP grade of all available biopsies taken during the biopsy session based one of the MRI examinations for the in-house dataset, or from Gleason score provided as ground truth for the PROSTATEx challenge. The in-house dataset demonstrates 34.44% and the PROSTATEx dataset demonstrates 34.31% prevalence for csPCa. The exact distribution can be seen in Table 2. The patient-wise ground truth was used for the patient-wise evaluation (see section "Evaluation of the clinical downstream task").

Exams	Without csPCa	Without csPCa	Sum
PROSTATEx	134 (65.7%)	70 (34.3%)	204
In-house dataset	276 (65.6%)	145 (34.4%)	421
Sum of exams	410 (65.6%)	215 (34.4%)	625

Table 2. Patient distribution through the datasets with respect to csPCa. The datasets are highly imbalanced, which is handled with a balanced trainer (see section "Training protocol").

For every lesion in the in-house dataset, the maximum ISUP grade is determined from a systematic biopsy-enhanced lesion histopathological ground truth (SELGT) that is determined from a histopathological sextant mapping integrating targeted and systematic biopsies according to the Ginsburg protocol[19,46]. This extended biopsy scheme ascertains that most of the csPCa is correctly diagnosed, as previously shown by comparison of this biopsy scheme to radical prostatectomy specimen[47,48]. In PROSTATEx, the Gleason scores were available for every lesion. The lesion-wise ground truth is used to generate the masks for the training, only lesions containing csPCa were selected.

Registration methods and annotations for training

To enable methodological comparison, from the mentioned primary prostate MRI cohort (see section "Prostate MRI cohort"), we generated several multi-modal datasets resulting from different image registration techniques (see Fig. 5a,f):

1. We used an unregistered multi-modal dataset with all modalities (T2-w, DWI with the highest b-value, ADC), which we hypothesize to be the lower performance bound in our study.
2. We created a dataset using B-spline registration the same way as Netzer et al.[26], where the registration is based on voxel intensities using mutual information[49] as the similarity metric. As the DWIs with the lowest b-values are the most similar to the original T2w images, it was beneficial to use them to calculate the transformation parameters for the co-registration of the two modalities. These parameters are then applied to the DWIs with the highest b-values and the ADC maps. This registration technique enables to interpret the prediction of the CAD on both modalities without any post processing. We use this dataset for our main experiments.
3. We also created as a reference dataset a segmentation based registration similarly to Sanyal et al.[17] being sure that the segmentations of the lesions and the prostate in the modalities overlap each other. By correcting the inconsistent ground-truths between the modalities resulted by misalignments, we provide a steady reference performance to be able to get information about the quality of our B-spline registration. We call this registration technique the '*ground-truth-matching*' (GT-matching).

Lesion segmentation masks belonging to the T2w modalities were used due to their higher resolution compared to the ADC maps and because they are annotated by also taking information from the ADC maps into consideration.

Misalignment augmentation

Data augmentation is intended to be a solution for the problem of overfitting by enhancing the size of the training dataset[50,51]. It also helps CNNs to cope with transformations, for which they are not invariant and which occur in the natural distribution of the data, such as rotation and scaling. This is essentially a way of introducing inductive biases into the model training to improve performance and robustness. Misalignments between image modalities could be addressed by using data augmentation as well. To become invariant to them to some extent, we introduce misalignment augmentation, a data augmentation technique generating additional, probabilistic misalignments on the fly during training and which can be dropped into any augmentation pipeline. Here we simply add misalignment augmentation to the standard nnU-Net data augmentation pipeline (see Fig. 5b). The amplitude of each transformation is sampled with uniform distribution constrained by a maximum amplitude value in positive and negative directions following the batchgenerators framework[52]. The three applied transformations for generating misalignments are:

- translations with the maximal amplitudes of $(10, 10, 6)$mm into the x,y,z directions, respectively,
- rotations with the maximal amplitude of 15° only in the x-y plane due to the highly anisotropic spacing,
- and a small amount of affine squeezing with the maximal ratio of 0.1 only in the dorsal-ventral direction which is similar to the naturally occurring image distortion between the T2-w and DWIs due to magnetic field inhomogeneities[53].

We make our proposed misalignment augmentations publicly available as part of the batchgenerators framework[52] https://github.com/MIC-DKFZ/batchgenerators and integrate it into a nnU-Net[25] trainer https://github.com/MIC-DKFZ/misalignment_DA.

Identifying prostate zones are crucial not only in PI-RADS[8], but also in the segmentation of lesions with neural networks[34,54]. The T2w images contain rich structural information with high resolution compared to the

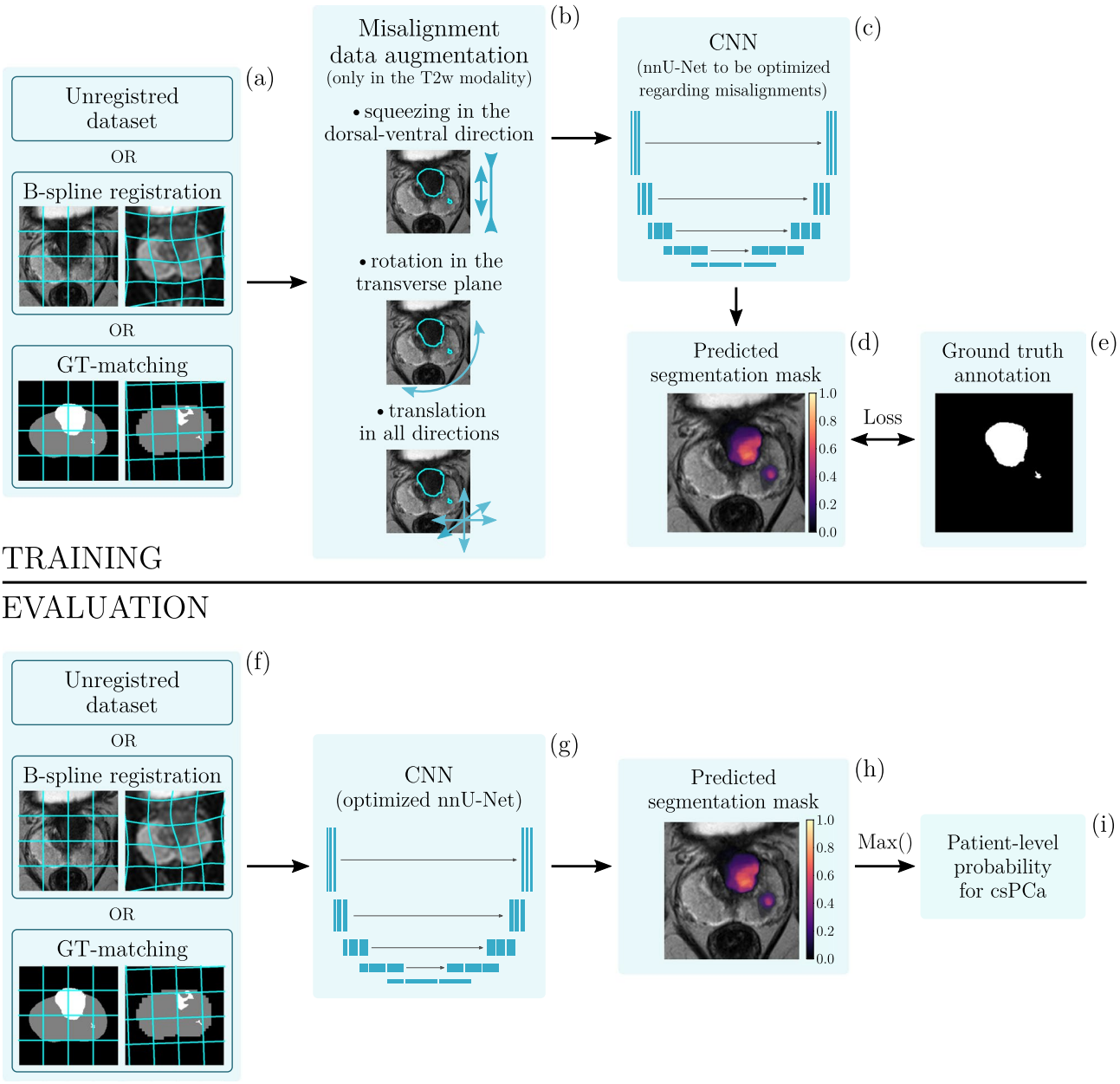

Figure 5. Overview of the training procedure with misalignment augmentation and the evaluation of the trained CNN. Training: (**a**) Three multi-modal datasets with different registration methods were created: an unregistered dataset as a lower bound reference, a dataset using B-spline registration, and a reference GT-matched dataset using the ground-truth segmentations. (**b**) Before any global image transformation, an optional misalignment data augmentation is implemented generating additional plausible misalignments between the image modalities. (**c**–**e**) We optimized the nnU-Net framework with the proposed modifications for the training of Netzer et al.[26], which was already successfully applied on the same data source. Application: (**f**) The GT-matched dataset is used only as an additional reference dataset for the evaluation of our B-spline registration. (**g**–**i**) Following the same evaluation scheme as Netzer et al.[26], the patient-wise PCa probability is calculated by taking the maximum value of the predicted lesion masks by the optimized CNN.

DWI sequences. Therefore, misalignments are generated by displacing the T2w modality with the corresponding T2w lesion annotations relative to the DWI sequences. To minimize transformation artifacts, the misalignment augmentation is implemented before any global data augmentation.

Though misalignments can be significantly reduced by registration, applying misalignment augmentation for the registered datasets could make CNNs more robust for remaining registration errors.

Evaluation of the clinical downstream task

Despite recent improvements in the field of registration techniques[55,56], registration depends on surrogate measures, which do not guarantee that high measurement scores translate into improved clinical predictions[57]. In contrast, we evaluate the effects of registration techniques (see section "Registration methods and annotations for training") and misalignment augmentation (see section "Misalignment augmentation") on the clinical downstream task: patient-level csPCa diagnosis.

To get quantitative measures about the quality of the registered CAD dataset, the Dice score probability density functions for every datasets are calculated by using the manual lesion annotations from the T2w and ADC modalities. Calculating this surrogate registration overlap metric not only provides information about the quality of the datasets regarding misalignments, but it also gives information about the correlation between the overlap measures and the clinical performances.

We choose patient-wise whole image PCa diagnosis derived from semantic segmentation for the clinical downstream task (see Fig. 5c–e). Their predictions are not just strongly relying on spatial information, but semantic segmentation is clinically interpretable task by providing spatial localization[14–16]. For assessing the performance of the trained models (Fig. 5g), we are using the area under the receiver operating characteristic curve (AUROC) as a discrimination measure. Since the clinical diagnosis in case of PCa is based on the whole image, we are evaluating the results of the downstream task predictions as patient-wise AUC from the whole 3D images. For the patient-wise PCa prediction, we are taking the maximum value of the predicted lesion masks the same way as it is previously published in[26] on the same data source (see Fig. 5h,i).

To be able to compare the performance of the trained models to radiologists' interpretation, we also calculate the radiologists' performance using the PI-RADS scores for the clinical index lesion as predictions and the maximum Gleason score of the systematic and targeted biopsy as the ground truth. According to PI-RADS, index lesions are scored on a Likert scale from 1 to 5 with higher scores indicating a higher risk of csPCa. The category of PI-RADS 3 has equivocal and PI-RADS 4 has high risk for csPCa[27], which make these two categories the most informative area on the ROC curves. Thus, we calculate the sensitivity and specificity for PI-RADS $\geq$ 4 and PI-RADS $\geq$ 3. Calculating the performance of the radiologists during clinical practice provides a fixed reference point.

Training protocol

The 625 data records were split into a training and hold out independent test set by stratifying exams by institution, PROSTATEx or local, and by the prevalence of csPC. The training set was used for 5-fold cross-validation. The exact distribution can be seen in Table 3. As there is an imbalance in the dataset, a balanced data loader is applied during the training.

Small prostate lesion sizes compared to the size of the entire image introduce significant noise during the training[17], especially as we have a limited number of exams with csPCa. Therefore, we are performing the same prostate-cropping mechanism based on the form of the prostate as Netzer et al.[26] to increase the training stability.

The images are preprocessed by the automated image preprocessing algorithm of nnU-Net[25]. Specifically, the image modalities were linearly resampled to a common resolution with the spacing of 0.3125 mm in the transverse plane. The slice distance of 3 mm on the longitudinal axis was common in both datasets and remained unchanged. In addition to the spatial resampling, the T2-weighted and DWIs with high B-value were normalized patient-wise, but the ADC maps were normalized by intensity statistics across the training dataset as the voxel intensities can be ordered into a physical quantity. The resulting input patch size was 320x256x20 in the x-y-z axes, respectively.

Compared to the standard nnU-Net settings, we implemented a balanced sampling of patients regarding the prevalence of csPC. We used Mish activation function instead of Leaky ReLU, Ranger instead of SGD optimizer, a cosine anneal instead of Poly learning rate scheduler, and an initial learning rate of 0.001 instead of 0.01 following Netzer et al.[26], which resulted in more stable training and better validation results.

We trained the 3D nnU-Net instance—consisting of 5 models—for each configuration of different dataset preprocessing techniques (see section "Registration methods and annotations for training") and the use of misalignment data augmentation (see section "Misalignment augmentation"). We optimized the instances using

Exams	Without csPCa	With csPCa	Sum
Training set	327	169	496
Test set	83	46	129
Sum	410	215	625

Table 3. Patient distribution in the training and in the test set regarding csPCa.

early stopping and tuned the misalignment augmentation probability with respect to the 5-fold-cross-validation AUROC. To reduce the hyperparameter space for the probability of misalignment augmentation, we applied misalignment augmentation with the probabilities of $P = \{0.0, 0.1, 0.2, 0.4\}$ for the transformations to occur at all. The cross-validation results can be seen in 'Supplementary Table S3'. The final models are then ensembled and evaluated on the independent test set using bootstrapping with 1000 replications to provide 95% confidence intervals. Additionally, we calculate p-values using the DeLong test[58] to determine the statistical significance of the performance increase. We consider differences in the performance with $p < 0.05$ statistically significant.

Data availability

MRI exams from the PROSTATEx challenge[37] are available at https://doi.org/10.7937/K9TCIA.2017.MURS5CL. MRI exams from our in-house cohort can not be made publicly available, due to data protection requirements. We make our proposed misalignment augmentations publicly available as part of the batchgenerators framework[52] https://github.com/MIC-DKFZ/batchgenerators and integrate it into a nnU-Net[52] trainer https://github.com/MIC-DKFZ/misalignment_DA.

Received: 18 October 2022; Accepted: 4 November 2023
Published online: 13 November 2023

References

1. Lomas, D. J. & Ahmed, H. U. All change in the prostate cancer diagnostic pathway. *Nat. Rev. Clin. Oncol.* **17**, 372–381. https://doi.org/10.1038/s41571-020-0332-z (2020).
2. Tewari, A. K., Whelan, P. & Graham, J. D. *Prostate Cancer: Diagnosis and Clinical Management* (Wiley, 2013).
3. Vilanova, J. C., Catalá, V., Algaba, F. & Laucirica, O. *Atlas of Multiparametric Prostate MRI: With PI-RADS Approach and Anatomic-MRI-Pathological Correlation* (Springer, 2018).
4. Rosenkrantz, A. B. *MRI of the Prostate: A Practical Approach* (Thieme, 2016).
5. Weinreb, J. C. *et al.* PI-RADS prostate imaging-reporting and data system: 2015, version 2. *Eur. Urol.* **69**, 16–40. https://doi.org/10.1016/j.eururo.2015.08.052 (2016).
6. Panebianco, V. *et al.* An update of pitfalls in prostate mpMRI: A practical approach through the lens of PI-RADS v. 2 guidelines. *Insights Imaging* **9**, 87–101. https://doi.org/10.1007/s13244-017-0578-x (2018).
7. Israël, B. *et al.* Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: What urologists need to know. part 2: interpretation. *Eur. Urol.* **77**, 469–480. https://doi.org/10.1016/j.eururo.2019.10.024 (2020).
8. Turkbey, B. *et al.* Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *Eur. Urol.* **76**, 340–351. https://doi.org/10.1016/j.eururo.2019.02.033 (2019).
9. Drost, F.-J.H. *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.* **77**, 78–94. https://doi.org/10.1016/j.eururo.2019.06.023 (2020).
10. Westphalen, A. C. *et al.* Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: Experience of the society of abdominal radiology prostate cancer disease-focused panel. *Radiology* **296**, 76–84. https://doi.org/10.1148/radiol.2020190646 (2020).
11. Venderink, W. *et al.* Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: What urologists need to know. part 3: targeted biopsy. *Eur. Urol.* **77**, 481–490. https://doi.org/10.1016/j.eururo.2019.10.009 (2020).
12. Winkel, D. J. *et al.* A novel deep learning based computer-aided diagnosis system improves the accuracy and efficiency of radiologists in reading biparametric magnetic resonance images of the prostate: results of a multireader, multicase study. *Investig. Radiol.* **56**, 605–613. https://doi.org/10.1097/RLI.0000000000000780 (2021).
13. Bosma, J. S. *et al.* Report-guided automatic lesion annotation for deep learning-based prostate cancer detection in bpMRI. arXiv preprint arXiv:2112.05151 (2021).
14. De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350. https://doi.org/10.1038/s41591-018-0107-6 (2018).
15. Bernard, O. *et al.* Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?. *IEEE Trans. Med. Imaging* **37**, 2514–2525. https://doi.org/10.1109/TMI.2018.2837502 (2018).
16. Nikolov, S. *et al.* Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. *J. Med. Internet Res.* **23**, e26151. https://doi.org/10.2196/26151 (2021).
17. Sanyal, J., Banerjee, I., Hahn, L. & Rubin, D. An automated two-step pipeline for aggressive prostate lesion detection from multi-parametric MR sequence. *AMIA Summits Transl. Sci. Proc.* **2020**, 552 (2020).
18. Kohl, S. *et al.* Adversarial networks for the detection of aggressive prostate cancer. In *Workshop on Machine Learning for Health (NIPS ML4H 2017)* (2017).
19. Schelb, P. *et al.* Simulated clinical deployment of fully automatic deep learning for clinical prostate MRI assessment. *Eur. Radiol.* **31**, 302–313. https://doi.org/10.1007/s00330-020-07086-z (2021).
20. Arif, M. *et al.* Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric MRI. *Eur. Radiol.* **30**, 6582–6592. https://doi.org/10.1007/s00330-020-07008-z (2020).
21. Alkadi, R., Taher, F., El-Baz, A. & Werghi, N. A deep learning-based approach for the detection and localization of prostate cancer in T2 magnetic resonance images. *J. Digit. Imaging* **32**, 793–807. https://doi.org/10.1007/s10278-018-0160-1 (2019).
22. De Vente, C., Vos, P., Hosseinzadeh, M., Pluim, J. & Veta, M. Deep learning regression for prostate cancer detection and grading in bi-parametric MRI. *IEEE Trans. Biomed. Eng.* **68**, 374–383. https://doi.org/10.1109/TBME.2020.2993528 (2020).
23. 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.* **73**, 102155. https://doi.org/10.1016/j.media.2021.102155 (2021).
24. Saha, A., Bosma, J., Linmans, J., Hosseinzadeh, M. & Huisman, H. Anatomical and diagnostic bayesian segmentation in prostate MRI—Should different clinical objectives mandate different loss functions? arXiv preprint arXiv:2110.12889 (2021).
25. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211. https://doi.org/10.1038/s41592-020-01008-z (2021).
26. Netzer, N. *et al.* Fully automatic deep learning in bi-institutional prostate magnetic resonance imaging: Effects of cohort size and heterogeneity. *Investig. Radiol.* **56**, 799–808. https://doi.org/10.1097/RLI.0000000000000791 (2021).
27. Engels, R. R., Israël, B., Padhani, A. R. & Barentsz, J. O. Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: What urologists need to know. part 1: acquisition. *Eur. Urol.* **77**, 457–468. https://doi.org/10.1016/j.eururo.2019.09.021 (2020).

28. Plodeck, V. *et al.* Rectal gas-induced susceptibility artefacts on prostate diffusion-weighted MRI with epi read-out at 3.0 T: Does a preparatory micro-enema improve image quality?. *Abdom. Radiol.* **45**, 4244–4251. https://doi.org/10.1007/s00261-020-02600-9 (2020).
29. van Griethuysen, J. J. *et al.* Gas-induced susceptibility artefacts on diffusion-weighted MRI of the rectum at 1.5 T—Effect of applying a micro-enema to improve image quality. *Eur. J. Radiol.* **99**, 131–137. https://doi.org/10.1016/j.ejrad.2017.12.020 (2018).
30. Kim, C. K., Park, B. K. & Kim, B. Diffusion-weighted MRI at 3 T for the evaluation of prostate cancer. *Am. J. Roentgenol.* **194**, 1461–1469 (2010).
31. Wang, S., Burtt, K., Turkbey, B., Choyke, P. & Summers, R. M. Computer aided-diagnosis of prostate cancer on multiparametric MRI: A technical review of current research. *BioMed Res. Int.*https://doi.org/10.1155/2014/789561 *(2014).*
32. Pellicer-Valero, O. J. *et al.* Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. *Sci. Rep.* **12**, 1–13. https://doi.org/10.1038/s41598-022-06730-6 (2022).
33. Cao, R. *et al.* Joint prostate cancer detection and Gleason score prediction in mp-MRI via focalnet. *IEEE Trans. Med. Imaging* **38**, 2496–2506. https://doi.org/10.1109/TMI.2019.2901928 (2019).
34. Hosseinzadeh, M. *et al.* Deep learning-assisted prostate cancer detection on bi-parametric MRI: Minimum training data size requirements and effect of prior knowledge. *Eur. Radiol.*https://doi.org/10.1007/s00330-021-08320-y *(2021).*
35. Caglic, I. & Barrett, T. Optimising prostate mpMRI: Prepare for success. *Clin. Radiol.* **74**, 831–840. https://doi.org/10.1016/j.crad.2018.12.003 (2019).
36. Reischauer, C., Cancelli, T., Malekzadeh, S., Froehlich, J. M. & Thoeny, H. C. How to improve image quality of DWI of the prostate-enema or catheter preparation?. *Eur. Radiol.* **31**, 6708–6716. https://doi.org/10.1007/s00330-021-07842-9 (2021).
37. Armato, S. G. *et al.* Prostatex challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *J. Med. Imaging* **5**, 044501. https://doi.org/10.1117/1.JMI.5.4.044501 (2018).
38. Schelb, P. *et al.* Classification of cancer at prostate MRI: Deep learning versus clinical PI-RADS assessment. *Radiology* **293**, 607–617 (2019).
39. van der Leest, M. *et al.* High diagnostic performance of short magnetic resonance imaging protocols for prostate cancer detection in biopsy-naive men: The next step in magnetic resonance imaging accessibility. *Eur. Urol.* **76**, 574–581. https://doi.org/10.1016/j.eururo.2019.05.029 (2019).
40. Kuhl, C. K. *et al.* Abbreviated biparametric prostate MR imaging in men with elevated prostate-specific antigen. *Radiology* **285**, 493–505. https://doi.org/10.1148/radiol.2017170129 (2017).
41. Tavakoli, A. A. *et al.* Contribution of dynamic contrast-enhanced and diffusion MRI to PI-RADS for detecting clinically significant prostate cancer. *Radiology*https://doi.org/10.1148/radiol.212692 *(2022).*
42. Wolf, I. *et al.* The medical imaging interaction toolkit. *Med. Image Anal.* **9**, 594–604. https://doi.org/10.1016/j.media.2005.04.005 (2005).
43. Maleike, D., Nolden, M., Meinzer, H.-P. & Wolf, I. Interactive segmentation framework of the medical imaging interaction toolkit. *Comput. Methods Programs Biomed.* **96**, 72–83. https://doi.org/10.1016/j.cmpb.2009.04.004 (2009).
44. Egevad, L., Delahunt, B., Srigley, J. R. & Samaratunga, H. International society of urological pathology (ISUP) grading of prostate cancer—An ISUP consensus on contemporary grading. https://doi.org/10.1111/apm.12533 (2016).
45. Epstein, J. I. *et al.* The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *Am. J. Surg. Pathol.* **40**, 244–252. https://doi.org/10.1097/PAS.0000000000000530 (2016).
46. Kuru, T. H. *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.* **112**, 568–577. https://doi.org/10.1111/bju.12132 (2013).
47. Radtke, J. P. *et al.* Multiparametric magnetic resonance imaging MRI and MRI-transrectal ultrasound fusion biopsy for index tumor detection: correlation with radical prostatectomy specimen. *Eur. Urol.* **70**, 846–853 (2016).
48. Drost, F.-J.H. *et al.* Prostate MRI, with or without MRI-targeted biopsy, and systematic biopsy for detecting prostate cancer. *Cochrane Database Syst. Rev.*https://doi.org/10.1002/14651858.CD012663.pub2 *(2019).*
49. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G. & Suetens, P. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging* **16**, 187–198. https://doi.org/10.1109/42.563664 (1997).
50. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 1–48. https://doi.org/10.1186/s40537-019-0197-0 (2019).
51. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324. https://doi.org/10.1109/5.726791 (1998).
52. Isensee, F. *et al. Batchgenerators—A python framework for data augmentation*https://doi.org/10.5281/zenodo.3632567 *(2020).*
53. Koolstra, K., O'Reilly, T., Börnert, P. & Webb, A. Image distortion correction for MRI in low field permanent magnet systems with strong b 0 inhomogeneity and gradient field nonlinearities. *Magn. Reson. Mater. Phys. Biol. Med.*https://doi.org/10.1007/s10334-021-00907-2 *(2021).*
54. Duran, A. *et al.* Prostattention-net: A deep attention model for prostate cancer segmentation by aggressiveness in MRI scans. *Med. Image Anal.*https://doi.org/10.1016/j.media.2021.102347 *(2022).*
55. Fu, Y. *et al.* Deep learning in medical image registration: A review. *Phys. Med. Biol.* **65**, 20TR01. https://doi.org/10.1088/1361-6560/ab843e (2020).
56. Haskins, G., Kruger, U. & Yan, P. Deep learning in medical image registration: A survey. *Mach. Vis. Appl.* **31**, 1–18. https://doi.org/10.1007/s00138-020-01060-x (2020).
57. Rohlfing, T. Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable. *IEEE Trans. Med. Imaging* **31**, 153–163. https://doi.org/10.1109/TMI.2011.2163944 (2011).
58. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 837–845 (1988).

Acknowledgements

We thank the research groups for their work in this highly interdisciplinary collaboration. Furthermore, we thank Sergio Mukherjee for his writing style suggestions.

Author contributions

The study design was set up by B.K., M.B., F.I., P.F.J., R.F., D.B., and K.H.M.; Acquisition, interpretation, and annotation of clinical, pathological, and radiological data was performed by N.N., A.S., C.W., M.G., V.S., R.G., A.S., M.H., H.S., and D.B; N.N. and D.B. provided their published software frameworks[26] for reusing their B-Spline registration algorithm and for optimizing their CAD system regarding misalignments; B.K. implemented the proposed augmentations and integrated them into the CAD system, designed and conceived the experiments using registration and misalignment augmentations, analyzed the results, created the figures, wrote the original

draft under the supervision of M.B., I.W., D.B., and K.H.M.; All authors reviewed, edited, and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. Bálint Kovács is supported by the DKFZ International Scholarship Program. This research received research support from the Bundesministerium für Wirtschaft und Klimaschutz (BMWK)—01MT21004B. Part of this work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—410981386 and the Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-46747-z.

Correspondence and requests for materials should be addressed to B.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.