Deformable MRI Sequence Registration for AI-based Prostate Cancer Diagnosis

Alessa Hering[1,2] ALESSA.HERING@RADBOUDUMC.NL
Sarah de Boer[1] SARAH.DEBOER@RADBOUDUMC.NL
Anindo Saha[1] ANINDYA.SHAHA@RADBOUDUMC.NL
Jasper J. Twilt[1] JASPER.TWILT@RADBOUDUMC.NL,
Derya Yakar[3,4] D.YAKAR@UMCG.NL
Maarten de Rooij[1] MAARTEN.DEROOIJ@RADBOUDUMC.NL
Henkjan Huisman[1,5] HENKJAN.HUISMAN@RADBOUDUMC.NL
Joeran S. Bosma[1,6] JOERAN.BOSMA@RADBOUDUMC.NL

[1] *Department of Medical Imaging, Radboudumc, Nijmegen, The Netherlands*
[2] *Fraunhofer MEVIS, Lübeck, Germany*
[3] *Department of Radiology, University Medical Center Groningen, The Netherlands*
[4] *Department of Radiology, Netherlands Cancer Institute, The Netherlands*
[5] *Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Norway*
[6] *Institut für Medizinische Informatik, Universität zu Lübeck, Germany*

Abstract

The PI-CAI (Prostate Imaging: Cancer AI) challenge led to expert-level diagnostic algorithms for clinically significant prostate cancer detection. The algorithms receive biparametric MRI scans as input, which consist of T2-weighted and diffusion-weighted scans. These scans can be misaligned due to multiple factors in the scanning process. Image registration can alleviate this issue by predicting the deformation between the sequences. We investigate the effect of image registration on the diagnostic performance of AI-based prostate cancer diagnosis. First, the image registration algorithm, developed in MeVisLab, is analyzed using a dataset with paired lesion annotations. Second, the effect on diagnosis is evaluated by comparing case-level cancer diagnosis performance between using the original dataset, rigidly aligned diffusion-weighted scans, or deformably aligned diffusion-weighted scans. Rigid registration showed no improvement. Deformable registration demonstrated a substantial improvement in lesion overlap (+10% median Dice score) and a positive yet non-significant improvement in diagnostic performance (+0.3% AUROC, p=0.18). Our investigation shows that a substantial improvement in lesion alignment does not directly lead to a significant improvement in diagnostic performance. Qualitative analysis indicated that jointly developing image registration methods and diagnostic AI algorithms could enhance diagnostic accuracy and patient outcomes.

Keywords: Image Registration, Prostate Cancer, Artificial Intelligence, MRI

1. Introduction

Prostate cancer (PCa) has 1.4 million new cases each year (Sung et al., 2021), a high incidence-to-mortality ratio and risks associated with treatment and biopsy; making non-invasive diagnosis of clinically significant prostate cancer (csPCa) crucial to reduce both

overtreatment and unnecessary (confirmatory) biopsies (Stavrinides et al., 2019). MRI scans provide the best non-invasive diagnosis for prostate cancer (Eldred-Evans et al., 2021), for which a 47% increase in demand is expected by 2040 (Sung et al., 2021). Due to the world-wide shortage of diagnostic personnel (Hricak et al., 2021), workload efficiency optimization is necessary to maintain healthcare accessibility in high-income countries and improve accessibility in low and middle-income countries.

Computer-aided diagnosis (CAD) can assist radiologists to diagnose csPCa and reduce the radiology workload (Winkel et al., 2021), but the observed workload reduction is limited. Larger workload reduction can be achieved through autonomous operation of diagnostic algorithms. Recent advances resulted in expert-level diagnostic performance for csPCa detection algorithms using biparmetric MRI (Saha et al., 2023).

Biparametric MRI (bpMRI) consists of T2-weighted (T2W) and diffusion-weighted imaging (DWI), and the DWI is used to calculate the apparent diffusion coefficient (ADC) and typically also the high b-value (HBV) map. T2W and DWI scans are usually acquired in immediate succession in about 15-30 minutes, but slight patient movement and processes like bladder filling can lead to misalignment between sequences (Kovacs et al., 2023). This misalignment results in lesion image features being misaligned between the sequences. For an accurate csPCa diagnosis, the information of both sequences are necessary to consider (Weinreb et al., 2016), meaning that csPCa detection algorithms have to combine information from different spatial locations when misalignment occurs. Current state-of-the-art csPCa algorithms (see Appendix B for more details) use an early fusion strategy for the combination of the different sequences, which may lead to challenges in accurate lesion detection and characterization when the lesion image features are not well aligned.

In Appendix F, we discuss in more details why such a misalignment results in incorrect predictions. To address this, misalignment in the Prostate Imaging – Cancer Artificial Intelligence (PI-CAI) dataset was manually corrected (85/1000 (8.5%) of the test cases and 54/9107 (0.6%) of training cases), and algorithms were subsequently trained and evaluated on these manually aligned MRI studies. However, manual alignment is labor-intensive, potentially undermining the efficiency gains offered by automated csPCa diagnostic methods when required during inference. Consequently, the efficacy of these algorithms in scenarios where sequences are not manually aligned remains uncertain and might be limited.

During inference, image registration can address the issue of misaligned sequences, by providing a plausible estimation of the patient movement and deformation and thus replaces the manual alignment step. Although the prostate cancer detection research field is vibrant, there has been limited focus on the registration of prostate MRI sequences. To address the issue of global misalignment, (Sanyal et al., 2020) proposed an affine registration approach based on prostate gland segmentation and (De Vente et al., 2020) presented a rigid registration based on Mutual Information. For compensating local deformations, both (Pellicer-Valero et al., 2022) and (Netzer et al., 2021), employed the SimpleITK non-rigid B-Spline registration using Mutual Information. However, the focus of these studies was not on the evaluation of registration performance, resulting in only (Netzer et al., 2021) examining this using the Dice Score of automatically generated prostate segmentations. In contrast, the other studies have assessed registration performance through visual examination of registered ADC images. To the best of our knowledge, only recently (Kovacs et al., 2023) explored the impact of image registration on prostate cancer detection performance

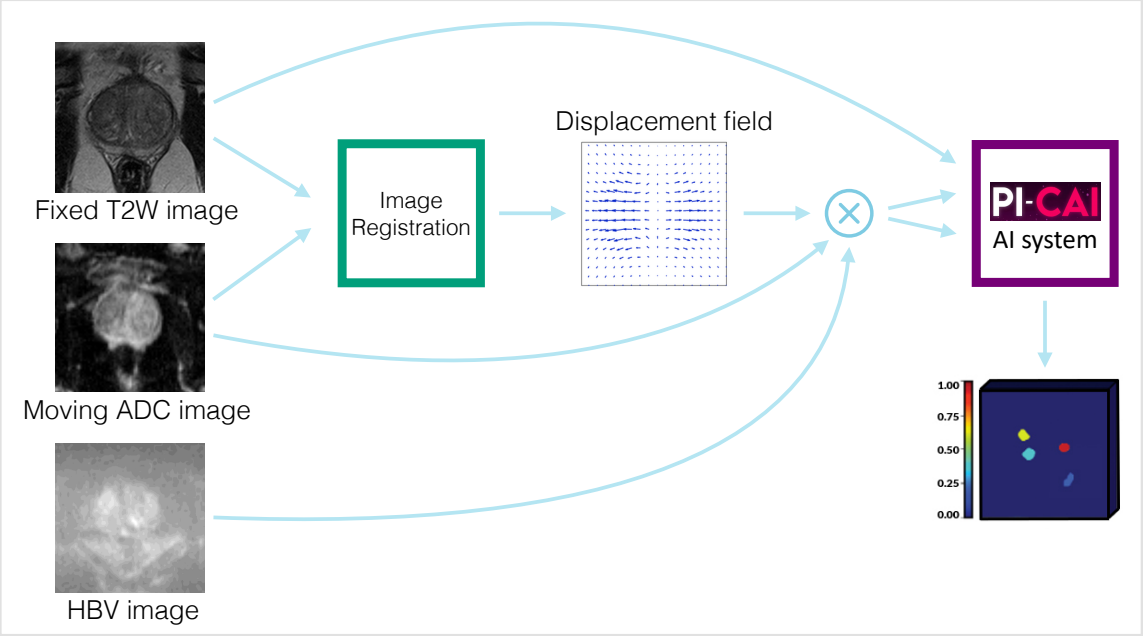

Figure 1: Overview of our method. The T2W scan is used as fixed image and the ADC map as moving image to find the displacement field using the registration method. The displacement field is applied to the ADC and HBV maps. The registered and original scans are used as input for the PI-CAI AI system (see Section 2.3) to detect clinically significant prostate cancer. The case-level diagnosis performance of the end-to-end pipeline is evaluated and used as a measure of effectiveness.

of algorithms using bpMRI. The results show that the B-Spline registration, which is based on (Netzer et al., 2021), improves the overlap of manually annotated lesions as measured by the Dice score. Additionally, the performance of the downstream task of patient-level csPCa diagnosis measured by the AUROC showed a non-significant improvement from 0.76 to 0.79. These results suggest that registration is a useful preprocessing step in an automated prostate cancer diagnosis pipeline. However, due to limited sample size (only 46 positive cases in the test set) and the lack of external testing, the ability to draw definitive conclusions is hindered.

In this study, we conduct a comprehensive analysis of the impact of image registration on the clinical downstream task of case-level csPCa diagnosis, utilizing two extensive evaluation datasets. Registration accuracy is assessed through the measurement of lesion alignment across an independent dataset comprising 473 cases, each annotated with paired lesions per modality. Further, we evaluate the downstream diagnostic efficacy on an external testing set consisting of 546 cases.

2. Materials and Methods

2.1. Registration

The aim of the image registration approach is to align the DWI maps (ADC and HBV) with the T2W scan (see Figure 1). Since the csPCa detection algorithms resample the DWI maps to the T2W scan, we chose the T2W scan as the fixed image and the ADC map as the moving image for the registration. The image registration algorithm is developed in the MeVisLab framework using the RegLib. We adopt a two-step approach which consists of a rigid registration and a deformable registration. Hereby, the registration pipeline starts with robust methods with fewer degrees of freedom and moves on to more precise, but less robust methods, which require better starting points due to their higher degrees of freedom. The calculated rigid and deformable transformation are applied to both DWI maps.

Let $\mathcal{F}, \mathcal{M} : \mathbb{R}^3 \to \mathbb{R}$ denote the fixed image and moving image, respectively, and let $\Omega \subset \mathbb{R}^3$ be a domain modeling the field of view of $\mathcal{F}$. The registration method aims to compute a deformation $y : \Omega \to \mathbb{R}^3$ that aligns the fixed image $\mathcal{F}$ and the moving image $\mathcal{M}$ on the field of view Ω such that $\mathcal{F}(x)$ and $\mathcal{M}(y(x))$ are similar for $x \in \Omega$.

Rigid Registration The rigid registration adopts the method of (Rühaak et al., 2017). We use the normalized gradient field distance measure (Haber and Modersitzki, 2006), that focuses on the alignment of image gradients of the fixed image $\mathcal{F}$ and the deformed moving image $\mathcal{M}(y)$. The edge hyper-parameter $\epsilon > 0$ is used to suppress small image noise, without affecting image edges. The optimization problem is solved using a Gauss-Newton optimization scheme and is embedded into a multi-level scheme with two levels.

Deformable Registration We deploy the matrix-free deformable registration of (König et al., 2018). The deformation is defined as a minimizer of the cost function

$$\min_y \mathcal{D}(\mathcal{F}, \mathcal{M}(y)) + \alpha \mathcal{R}(y),$$

with the normalized gradient field distance measure $\mathcal{D}^{\text{NGF}}$. To focus the registration to inside the prostate, we restrict Ω_{NGF} to the support of the prostate segmentation of the fixed image, which is automatically generated with the prostate segmentation algorithm provided by (Saha et al., 2022). The second-order curvature regularizer $\mathcal{R}^{curv}$ (Fischer and Modersitzki, 2003) enforces smooth deformation by penalizing spatial derivatives. The parameter α is a weighting factor of the regularizer. The optimization problem is solved using the limited-memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) optimization scheme (Liu and Nocedal, 1989). Optimization was performed in a multi-level scheme with two levels on images with successively declining levels of smoothing to guide registration from larger structures to smaller refinements. The deformable registration uses the output of the rigid registration as an initial starting point. Hyperparameters of the registration method were experimentally set using the first ten cases of the PI-CAI training dataset.

2.2. Data

Three datasets with bpMRI scans (axial T2W, ADC and HBV (b $\geq$ 1000) imaging) for prostate cancer detection were used. For each dataset, the reference standard was set by histopathology, with clinically significant prostate cancer defined as ISUP 2-5 (intermediate to very high risk) (Epstein et al., 2016). Informed consent was waived, given the retrospective scientific use of deidentified patient data. Scan characteristics are given in Table A.

PI-CAI: For csPCa detection model development, 10,207 cases of 9129 patients from 10 Dutch hospitals and 1 Norwegian hospital were used (Saha et al., 2022). Cases were acquired using 1.5 or 3-Tesla MRI scanners between 2012 and 2021 from patients with suspicion of harboring prostate cancer. Exclusion criteria included prior prostate-specific treatment, prior ISUP $\geq$ 2 findings, incomplete studies, and diagnostically insufficient image quality. Manual voxel-level annotations were available for 1175 positive training cases (1323 lesions) and for an additional 892 positive training cases (1037 lesions) AI-derived voxel-level annotations were provided.

PCNN: For testing of the registration algorithm, cases from the PI-CAI training set with manual voxel-level annotations per modality (T2W and ADC) were included. This selected 473 cases of 438 patients from Prostaat Centrum Noord-Nederland (PCNN) (8 hospitals).

PROMIS: For external testing, 546 cases of 546 patients from 11 United Kingdom hospitals were included (Ahmed et al., 2017). Cases were acquired using 1.5-Tesla MRI scanners between 2012 and 2015 from patients with suspicion of harboring prostate cancer. Exclusion criteria included prior prostate treatment, prior biopsies, incomplete studies, and diagnostically insufficient image quality. No manual voxel-level annotations were available.

2.3. PI-CAI AI system

The PI-CAI AI system was developed in the PI-CAI challenge. The algorithm is the ensemble of the top 5 submissions, selected based on testing with 1000 cases. The models were trained using a dataset of 9107 cases. The algorithm uses the axial T2W, ADC and HBV scans and clinical variables (e.g. PSA density). The U-Net backbone was predominantly used, with early fusion of the scans. See Supplementary Materials Section B for details.

3. Experiments

The aim of this study is to evaluate the effect of image registration on the clinical downstream task of case-level csPCa diagnosis. We first evaluated the registration performance by measuring lesion alignment and the plausibility of the displacement field. In the second experiment, we employed the csPCa detection algorithms developed in the PI-CAI challenge for the diagnostic evaluation. For both experiments, we compare the results on three dataset variants: the original dataset, the dataset with rigidly aligned T2W and DWI scans, and the dataset with deformably aligned T2W and DWI scans.

Registration performance The evaluation of registration performance was conducted using the PCNN validation dataset, chosen for its availability of lesion annotations across both T2W and ADC scans. The hyperparameters for the registration method were manually

fine-tuned using only the first 10 cases from the PI-CAI Public Training and Development dataset, which did not overlap with this PCNN dataset. Therefore, the PCNN dataset serves as an independent evaluation set for assessing the registration performance.

To quantitatively assess the quality of image registration in the absence of reference displacement fields, we employed two surrogate metrics. The Dice coefficient was utilized to quantify the overlap of lesion segmentations between T2W scans and ADC maps. Although we recognize that the Dice coefficient may not be the perfect metric for assessing the registration performance (Rohlfing, 2011), its usage is justified in this context given the critical importance of accurate lesion alignment in T2W scans and ADC maps for the reliable performance of csPCa detection algorithms. The choice of the Dice score, therefore, aligns with our objective to prioritize lesion alignment in the evaluation of registration effectiveness. Additionally, to ensure realistic deformations, we evaluated the plausibility of the displacement field by examining the percentage of voxels exhibiting folding within the prostate region of the predicted deformation field.

csPCa detection performance Diagnostic performance is assessed using the area under the receiver operator characteristic curve (AUROC). For case-level risk estimation of significant cancer, we utilized voxel-level detection maps generated by each of the five PI-CAI AI ensemble algorithms (detailed in Appendix B) on the external PROMIS test dataset.

Additionally, we evaluated diagnostic performance using the PCNN dataset, to facilitate an evaluation of diagnostic performance in relation to the registration accuracy. We note that this dataset is not independent for the diagnostic algorithms, since this is a subset of the training data of the algorithms. Therefore, the results can be found in Appendix E and Appendix F.

Since the PROMIS dataset contains scans with very large field-of-views with anatomical structures not present in the PI-CAI training dataset, we filtered out lesion predictions further than 3 mm away from the prostate segmentation. Following the approach used in the PI-CAI challenge, each algorithm's case-level prediction was the maximum lesion-level prediction, and the AI system's case-level prediction was the equally-weighted prediction of each algorithm.

Statistical analysis The diagnostic performance differences on the external testing set were statistically analyzed. The performance with the deformably and rigidly aligned images are compared against the performance with the original dataset. To determine the probability of one configuration outperforming another, we performed DeLong's test (DeLong et al., 1988). Multiplicity was corrected for using the Holm-Bonferroni method, with a base alpha value of 0.05. Details are in Appendix D.

4. Results

Registration performance The median Dice score improved to 0.58 with deformable registration, compared to 0.48 for the original dataset and 0.47 with rigid registration.

For one case, 1% of voxels in the prostate were folded. For all other cases, no foldings occurred in the deformation field. Quantitative results are summarized in Figure 3 and Appendix Figure B. One example is shown in Figure 2. Additionally, in Appendix F, the

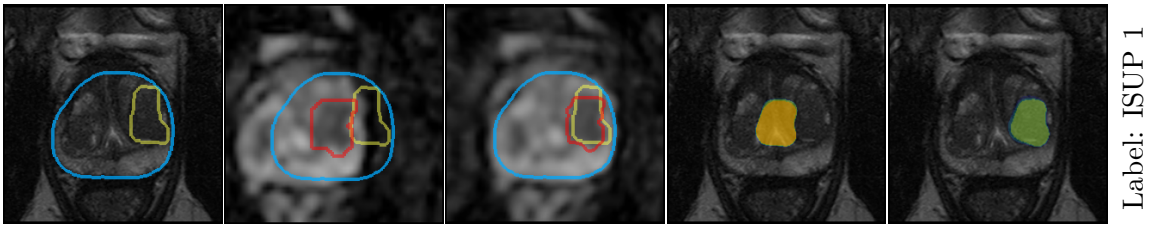

Figure 2: Qualitative registration results showing an exemplary case with prostate gland, lesion annotated on T2, lesion annotated on ADC. In the last two column, the prediction maps (PM) generated with the original dataset and the deformably aligned dataset are overlayed on the T2W scan.

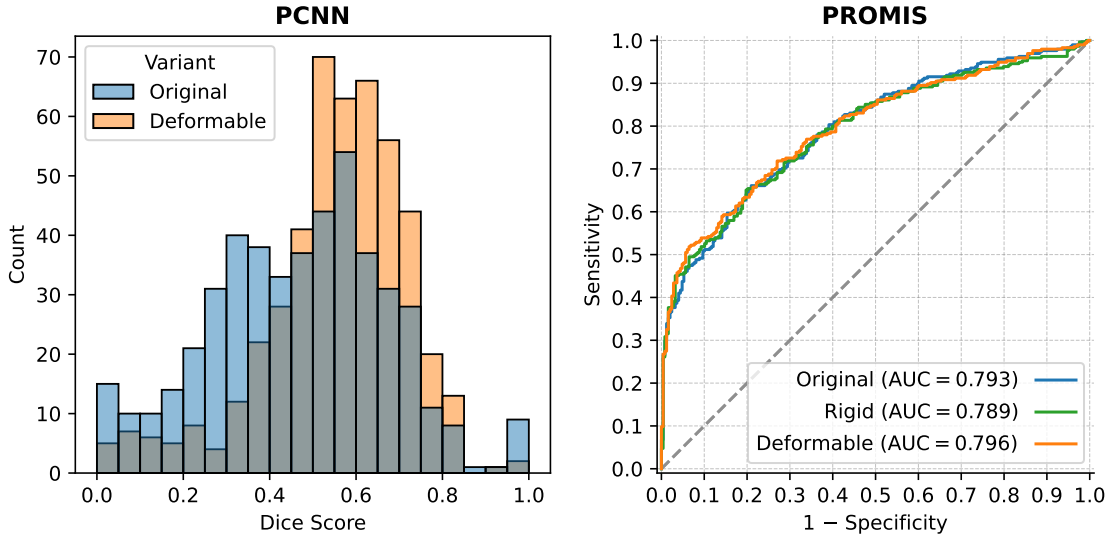

Figure 3: Quantitative registration results. *(left)* Distribution of Dice scores between the lesion annotation on the T2W and ADC scans for the original and deformably aligned PCNN datasets. *(right)* Model performance for the PI-CAI AI system with the original, rigidly aligned and deformably aligned PROMIS datasets.

cases with largest improvement, median improvement, and the largest decrease in Dice score are visualized, alongside the clinical interpretation for each case.

csPCa diagnosis performance For the PROMIS external testing dataset, the PI-CAI AI system showed a positive yet non-significant improvement in diagnostic performance (+0.3% AUROC, p=0.18) with deformably aligned scans compared to the original dataset. A comprehensive qualitative analysis of representative cases is given in Appendix F.

5. Discussion and Conclusion

In this study, we investigated the effect of image registration on the clinical downstream task of case-level csPCa diagnosis when integrated at the inference stage. Deformable registration demonstrated a substantial improvement in lesion overlap on the validation dataset (+9% average Dice score) which is even slightly more than the one reported in (Kovacs et al., 2023) (+6% average Dice score). However, since different datasets were used, a direct comparison is not possible. Additionally, we showed a positive yet non-significant improvement in diagnostic performance on the PROMIS test dataset (+0.3% AUROC, p=0.18) with deformably aligned scans. Our investigation shows that a substantial improvement in lesion alignment does not directly equal a significant improvement in diagnostic performance. To illustrate the impact of misalignment on the algorithmic results, we present detailed visualizations and analyses of several PCNN and PROMIS cases in Appendix F. These results showed that the PI-CAI AI system demonstrated robustness to minor misalignments, particularly when these misalignments did not result in lesions being misrepresented in incorrect zones. Additionally, we anticipate a comparable number of misaligned cases in the PROMIS dataset as observed in the PI-CAI dataset, where the incidence was low. Therefore, the expected improvement in AUROC is limited. The positive yet non-significant improvement in diagnostic performance might be the result from those cases.

Our method had limitations. The deformable registration method potentially introduced unrecognized artifacts into the images which might result in worse diagnostic performance. Addressing this through retraining the csPCa algorithms to adapt to registration-induced image variations represents a promising strategy. It is crucial to note that the registration method avoided the generation of physiological unrealistic deformations. This is achieved by applying a high regularization weight to obtain smooth and plausible displacement fields. Another critical aspect is the choice of resampling strategy. This factor considerably impacts the smoothing of ADC values, especially for small lesions, and influences the diagnostic quality of images through the effects of multiple resamples. Merging all resampling steps into one would visibly increase the quality, but is only possible in an end-to-end approach.

The relevance of the PROMIS dataset in present-day analyses has been a subject of debate, particularly among radiologists. The diagnostic quality of MRI scans has markedly improved since the trial finished in 2015. Additionally, the PROMIS dataset contains acquired high b-value scans, while contemporary protocols calculate this based on acquired lower b-value scans, which results in less noise and better diagnostic quality. Despite these limitations, the relevance of the PROMIS dataset should not be understated. This dataset can serve as a benchmark for scenarios where access to high-end, expensive scanners is limited. This situation is a common reality for many institutions, highlighting the importance of developing algorithms that can perform well across a range of image acquisition methods.

In conclusion, our study shows that while image registration can substantially improve lesion overlap in csPCa diagnosis, it does not directly lead to a significant improvement in diagnostic performance. However, the qualitative analysis showed promising results and indicate that joint development of image registration methods and diagnostic AI algorithms could enhance diagnostic accuracy and patient outcome.

Acknowledgments

This research is funded by the European Union HORIZON-HLTH-2022: COMFORT (101079894), Health Holland (LSHM20103), European Union H2020: ProCAncer-I project (952159), European Union H2020: PANCAIM project (101016851), and Siemens Healthineers (CID: C00225450). The PROMIS dataset was provided by the PROMIS study (see Appendix H for details).

References

Hashim U Ahmed, Ahmed El-Shater Bosaily, Louise C Brown, Rhian Gabe, Richard Kaplan, Mahesh K Parmar, Yolanda Collaco-Moraes, Katie Ward, Richard G Hindley, Alex Freeman, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *The Lancet*, 389 (10071):815–822, 2017.

Coen De Vente, Pieter Vos, Matin Hosseinzadeh, Josien Pluim, and Mitko Veta. Deep learning regression for prostate cancer detection and grading in bi-parametric MRI. *IEEE Transactions on Biomedical Engineering*, 68(2):374–383, 2020.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

Zhangfu Dong, Yuting He, Xiaoming Qi, Yang Chen, Huazhong Shu, Jean-Louis Coatrieux, Guanyu Yang, and Shuo Li. MNet: rethinking 2D/3D networks for anisotropic medical image segmentation. In *Thirty-First International Joint Conference on Artificial Intelligence {IJCAI-22}*, pages 870–876. International Joint Conferences on Artificial Intelligence Organization, 2022.

David Eldred-Evans, Paula Burak, Martin John Connor, Emily Day, Martin Evans, Francesca Fiorentino, Martin Gammon, Feargus Hosking-Jervis, Natalia Klimowska-Nassar, W. McGuire, Anwar R. Padhani, Andrew Toby Prevost, Derek Price, Heminder Sokhi, Henry H. Tam, Mathias Winkler, and Hashim Uddin Ahmed. Population-based prostate cancer screening with magnetic resonance imaging or ultrasonography. *JAMA Oncology*, 7:395 – 402, 2021.

Jonathan I. Epstein, Lars Egevad, Mahul B. Amin, Brett Delahunt, John R. Srigley, and Peter A. Humphrey. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *The American Journal of Surgical Pathology*, 40: 244–252, 2016.

Bernd Fischer and Jan Modersitzki. Curvature based image registration. *Journal of Mathematical Imaging and Vision*, 18(1):81–85, 2003.

Eldad Haber and Jan Modersitzki. Intensity gradient based registration and fusion of multi-modal images. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, volume 3216, pages 591–598, 2006.

Hedvig Hricak, May Abdel-Wahab, Rifat Atun, Miriam Mikhail Lette, Diana Paez, James A Brink, Lluís Donoso-Bach, Guy Frija, Monika Hierath, Ola Holmberg, et al. Medical imaging and nuclear medicine: a lancet oncology commission. *The Lancet Oncology*, 22 (4):e136–e172, 2021.

Hongyu Kan, Jun Shi, Minfan Zhao, Zhaohui Wang, Wenting Han, Hong An, Zhaoyang Wang, and Shuo Wang. ITUNet: Integration of transformers and unet for organs-at-risk segmentation. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2123–2127. IEEE, 2022.

Lars König, Jan Rühaak, Alexander Derksen, and Jan Lellmann. A matrix-free approach to parallel and memory-efficient deformable image registration. *SIAM Journal on Scientific Computing*, 40(3):B858–B888, 2018.

Balint Kovacs, Nils Netzer, Michael Baumgartner, Adrian Schrader, Fabian Isensee, Cedric Weißer, Ivo Wolf, Magdalena Görtz, Paul F Jaeger, Victoria Schütz, et al. Addressing image misalignments in multi-parametric prostate MRI for enhanced computer-aided diagnosis of prostate cancer. *Scientific Reports*, 13(1):19805, 2023.

Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Nils Netzer, Cedric Weißer, Patrick Schelb, Xianfeng Wang, Xiaoyan Qin, Magdalena Görtz, Viktoria Schütz, Jan Philipp Radtke, Thomas Hielscher, Constantin Schwab, et al. Fully automatic deep learning in bi-institutional prostate magnetic resonance imaging: effects of cohort size and heterogeneity. *Investigative radiology*, 56(12):799–808, 2021.

Oscar J Pellicer-Valero, Jose L Marenco Jimenez, Victor Gonzalez-Perez, Juan Luis Casanova Ramon-Borja, Isabel Martin Garcia, Maria Barrios Benito, Paula Pelechano Gomez, José Rubio-Briones, María José Rupérez, and José D Martín-Guerrero. Deep learning for fully automatic detection, segmentation, and gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. *Scientific reports*, 12 (1):2975, 2022.

Torsten Rohlfing. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE transactions on medical imaging*, 31(2): 153–163, 2011.

Jan Rühaak, Lars König, Florian Tramnitzke, Harald Köstler, and Jan Modersitzki. A matrix-free approach to efficient affine-linear image registration on CPU and GPU. *Journal of Real-Time Image Processing*, 13:205–225, 2017.

Anindo Saha, Jasper J. Twilt, Joeran S. Bosma, Bram van Ginneken, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen Fütterer, Maarten de Rooij, and Henkjan Huisman. Artificial intelligence and radiologists at prostate cancer detection in MRI: The PI-CAI challenge (study protocol). *Zenodo*, 2022. doi: 10.5281/zenodo.6667655.

Anindo Saha, Joeran Bosma, Jasper Twilt, Bram van Ginneken, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen Fütterer, Maarten de Rooij, et al. Artificial intelligence

and radiologists at prostate cancer detection in MRI—the PI-CAI challenge. In *Medical Imaging with Deep Learning, short paper track*, 2023.

Josh Sanyal, Imon Banerjee, Lewis Hahn, and Daniel Rubin. An automated two-step pipeline for aggressive prostate lesion detection from multi-parametric MR sequence. *AMIA Summits on Translational Science Proceedings*, 2020:552, 2020.

Arun Seetharaman, Indrani Bhattacharya, Leo C Chen, Christian A Kunder, Wei Shao, Simon JC Soerensen, Jeffrey B Wang, Nikola C Teslovich, Richard E Fan, Pejman Ghanouni, et al. Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging. *Medical Physics*, 48(6):2960–2972, 2021.

Vasilis Stavrinides, Francesco Giganti, Mark Emberton, and Caroline M Moore. MRI in active surveillance: a critical review. *Prostate Cancer and Prostatic Diseases*, 22(1):5–15, 2019.

Xu Sun and Weichao Xu. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014.

Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

Jeffrey C Weinreb, Jelle O Barentsz, Peter L Choyke, Francois Cornud, Masoom A Haider, Katarzyna J Macura, Daniel Margolis, Mitchell D Schnall, Faina Shtern, Clare M Tempany, et al. Pi-rads prostate imaging–reporting and data system: 2015, version 2. *European urology*, 69(1):16–40, 2016.

David J Winkel, Angela Tong, Bin Lou, Ali Kamen, Dorin Comaniciu, Jonathan A Disselhorst, Alejandro Rodríguez-Ruiz, Henkjan Huisman, Dieter Szolar, Ivan Shabunin, 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. *Investigative radiology*, 56(10): 605–613, 2021.

Appendix A. Scan characteristics

Table A: Scan characteristics showing the median, (95% confidence interval) and [min-max] in voxels or mm/voxel.

	PI-CAI	PCNN	PROMIS
T2W in-plane size	640 (320, 1024) [256, 1078]	1024 (296, 1024) [256, 1024]	512 (256, 512) [256, 640]
T2W number of slices	21 (19, 29) [15, 45]	27 (20, 35) [15, 45]	26 (23, 38) [15, 94]
T2W in-plane resolution	0.3 (0.3, 0.6) [0.2, 0.8]	0.3 (0.2, 0.7) [0.2, 0.8]	0.4 (0.4, 0.8) [0.4, 0.9]
T2W slice thickness	3.6 (3.0, 3.6) [1.3, 5.0]	3.0 (3.0, 4.8) [2.2, 4.8]	3.3 (3.3, 3.6) [0.8, 6.5]
ADC in-plane size	128 (102, 256) [70, 336]	240 (114, 256) [108, 336]	172 (128, 172) [126, 256]
ADC number of slices	21 (19, 29) [11, 41]	27 (11, 33) [11, 41]	13 (11, 19) [11, 24]
ADC in-plane resolution	2.0 (1.4, 2.0) [0.9, 2.6]	1.4 (0.9, 1.9) [0.9, 2.0]	1.5 (1.5, 1.7) [1.1, 2.0]
ADC slice thickness	3.6 (3.0, 3.6) [3.0, 5.8]	3.0 (3.0, 5.5) [3.0, 5.8]	5.0 (5.0, 5.5) [4.0, 6.0]

Appendix B. PI-CAI AI system

The PI-CAI Challenge involved a multi-step development of the PI-CAI AI system. In the development phase, participants could access an annotated dataset of 1500 MRI cases, and create AI models incorporating MRI scans and various clinical variables (e.g. patient age, PSA level, prostate volume, scanner manufacturer) for csPCa detection and diagnosis. During the development phase, teams could periodically submit their algorithm to be validated on 100 tuning cases. At the end of this phase, each team could submit a single algorithm for blind testing on 1000 cases. The training methods from the top five teams, after minor optimizations for computational efficiency, were trained on a larger dataset of 9107 cases. Once trained, all five algorithms were ensembled with equal weighting into the single PI-CAI AI system. Each algorithm's test performance is available on the leaderboard (https://pi-cai.grand-challenge.org/evaluation/closed-testing-phase-final-ranking/leaderboard/). Each algorithm is described in more detail in the "Supplementary File" which is linked on the same leaderboard.

Here, we provide a brief summary of the methods. The ensemble of the top 5 teams from the PI-CAI challenge involved 50 models. The architectures were configured by the nnU-Net (21x) and nnDetection (5x) frameworks, or by the teams: ITUNet (5x) (Kan et al., 2022), two variants of the SPCNet (2x5x) (Seetharaman et al., 2021), a UNet (5x) and the Z-SSMNet (5x) architecture which was based on (Dong et al., 2022). 49 of these models used early fusion for the T2-weighted and diffusion-weighted inputs, where scans are resampled to the same spacing and concatenated as channels before processing by the network. One model, which was used to segment the prostate, only used the T2-weighted scan.

None of the teams leveraged affine or non-rigid deformations between the input sequences during training (such as in (Kovacs et al., 2023)). The data augmentation schemes did include e.g. rotation, scaling and translation, but applied this to the T2-weighted and diffusion-weighted images equally.

Appendix C. Computational Resources

For image registration, we used an NVIDIA 1080 Ti with a total wall clock time of 8 hours. For inference on the PROMIS and PCNN datasets, we used several GPUs (NVIDIA GTX Titan X, GTX 1080 Ti, RTX 2080 Ti or A100) with a total wall clock time of 200 hours.

Appendix D. Statistical Analysis Plan

The csPCa detection performance on the external PROMIS testing dataset is statistically evaluated. The study objectives are described in a predefined hierarchical tree and the tests are performed accordingly (see Figure A). Multiplicity is corrected for at each stage using the Holm-Bonferroni method, considering a base alpha value of 0.05. The hierarchical structure is formed following the research question as proposed in this study. Our aim is to investigate if image registration can boost model performance, in family 1A and 1B we test whether either rigid or deformable registration is better than using no registration. If one or both of these registration methods turn out to boost the performance of the AI model then we move on to testing whether model performance is better using deformable registration relative to rigid registration (family 2A) and/or whether model performance is better using rigid registration relative to deformable registration (family 2B).

Calculation of p-values - DeLong's test The p-value for a superiority test comparing AUROC scores can be calculated using DeLong's test (DeLong et al., 1988). To test the null hypotheses as described in Figure A, we use the DeLong's formulas to estimate the variance of and the covariance between the AUROC scores. In our experiments we use the faster implementation of the DeLong's algorithm, as proposed by (Sun and Xu, 2014). We can calculate the z score using the following formula:

$$z = \frac{\hat{\theta}^{(1)} - \hat{\theta}^{(2)}}{\sqrt{\mathbb{V}[\hat{\theta}^{(1)}] + \mathbb{V}[\hat{\theta}^{(2)}] - 2\mathbb{C}[\hat{\theta}^{(1)}, \hat{\theta}^{(2)}]}} \tag{1}$$

The P value is then calculated from the z score using a one-tailed test.

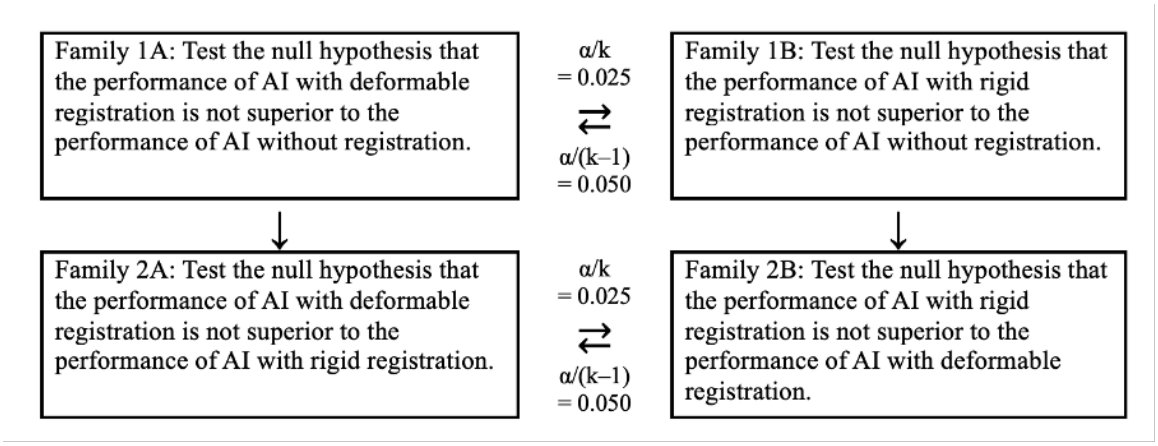

Figure A: Flowchart illustrating the statistical plan to test the study objectives. The significance thresholds used for family 1A and 1B are adjusted using the Holm-Benferroni method, considering a base alpha of 0.05. If the null hypothesis of family 1A is rejected, but that of family 1B is not rejected, then we move on to testing family 2A with an alpha value of 0.05 but do not test family 2B. Conversely, if the null hypothesis for family 1B is rejected but that of family 1A is not rejected, then we move on to testing family 2B with an alpha value of 0.05, but do not test family 2A. If the null hypothesis for both family 1A and 1B are rejected, then we move on to testing both family 2A and 2B, where the significance threshold is once again adjusted using the Holm-Benferroni method considering a base alpha value of 0.05. If neither family 1A nor 1B are rejected, neither family 2A nor 2B will be tested.

Appendix E. Quantitative Results

Quantitative results for the registration methods are shown in Figure B. We evaluated the diagnostic performance of the PI-CAI AI algorithms on the PCNN dataset, which was part of the training data for these algorithms. For this evaluation, we also included 205 cases that did not have lesion annotations per modality, including all 678 PCNN cases from the PI-CAI dataset.

For the PCNN dataset, three algorithms showed a small increase in diagnostic performance (+0.4%, +0.8%, +0.1% AUROC) and two algorithms showed decreased performance (-8.1%, -1.2% AUROC). The ensemble of the first three algorithms demonstrated a small increase in diagnostic performance (+0.5% AUROC), as shown in Figure B. The first three algorithms were ensembled in this analysis because we observed that the latter two algorithms demonstrated strong overfitting to the training cases. Their performance on these training cases was much higher compared to the other three algorithms, but all algorithms performed comparably on the testing leaderboard.

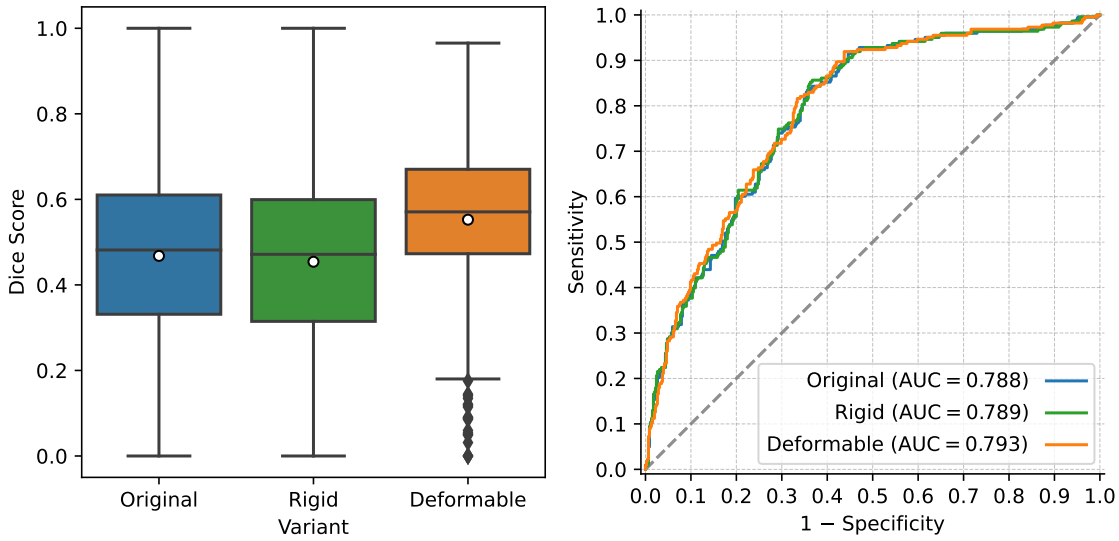

Figure B: *(left)* Distribution of Dice scores between the lesion annotation on the T2W and ADC scans for the original, rigidly and deformably aligned PCNN dataset. The median Dice score for the original dataset was 0.48, with 2.5% and 97.5% quantiles of the distribution of Dice scores being 0.03 and 0.90. For the rigidly aligned dataset these metrics were 0.47 [0.01, 0.82], and for the deformably aligned dataset 0.58 [0.10, 0.81]. *(right)* Model performance for the Ensemble of 3 PI-CAI algorithms with the original, rigidly aligned and deformably aligned PCNN datasets. AUROC = area under the receiver operator characteristic curve.

Appendix F. Qualitative Results

In this section, we present qualitative results of the image registration and subsequent csPCa detection algorithms. Results for the PCNN dataset are shown in Figure C, showing the case with the largest improvement in Dice score (first row), the median improvement in Dice score (second row) and the largest decrease in Dice score (third row) for the deformably aligned dataset, compared to the original datasets.

The first row shows the images for a 58-year-old man with a PSA level of 16 ng/mL and a PSA density of 0.32 ng/mL/cc. The imaging shows mild benign prostatic hyperplasia (BPH) in the transition zone. BPH is a benign condition, which grows over time. A typical transition zone with BPH shows so-called 'organized chaos', with multiple nodules with variable imaging appearance, often with diffusion restriction and enhancement. In transition zone tumors, this typical encapsulation is lost, and the organized aspect changes to a homogeneous low T2W signal with marked diffusion restriction and vivid enhancement. In the left transition zone of this patient (appearing on the right in the images), an encapsulated BPH nodule is annotated in yellow on T2W, with low T2W signal intensity. On the ADC map an area with marked diffusion restriction is annotated in red. A notable discrepancy

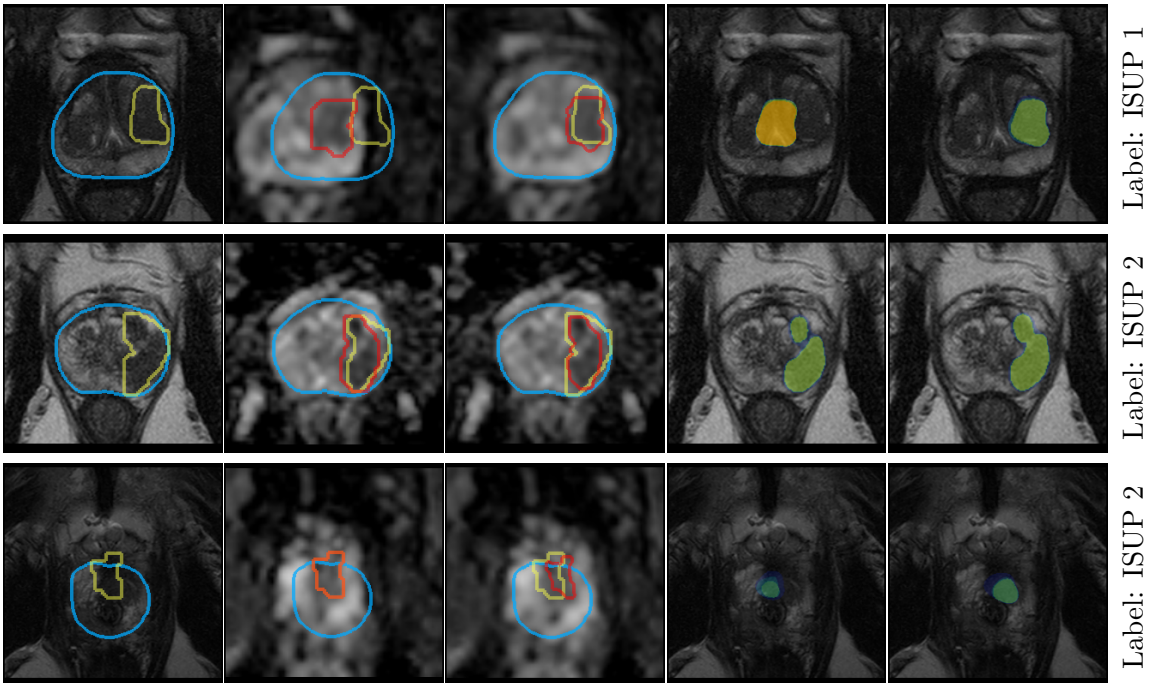

Figure C: Qualitative results on the PCNN datset showing the case with the largest improvement in Dice score (first row), the median improvement in Dice score (second row) and the largest decrease in Dice score (third row). The T2, ADC, and deformably aligned ADC are shown with ■ prostade gland, ■ lesion annotated on T2, ■ lesion annotated on ADC. In the last two column, the prediction maps (PM) generated with the original dataset and the deformbly aligned dataset are overlayed on the T2W image. The label shows the ISUP grade, where 1 is indolent cancer (negative), and ≥ 2 is intermediate to high-risk cancer (positive).

is observed in the alignment between the T2W and ADC imaging, leading to misalignment between the lesion's features on the T2W and ADC scans. In the PI-RADS scoring system, T2W is the dominant sequence in the transitions zone, while in the peripheral zone DWI is the most important sequence. The encapsulated nature on T2W of this nodule is a non-suspicious sign in the transition zone. However, a lesion with similar diffusion restriction in the peripheral zone would be a suspicious sign. Consequently, the PI-CAI AI system with the original scans classifies the lesion in the middle of the transition zone instead of within an encapsulated nodule more laterally due to the misalignment, which suggests a higher risk level (prediction=0.63). The deformable image registration method aligned the two modalities, and the PI-CAI AI system with the deformably aligned scans assigned a lower risk level (prediction=0.47). Targeted biopsies revealed ISUP 1 in the left transition zone, which is an indolent prostate cancer that is often invisible on prostate MRI.

The second row shows the images for a 76-year-old man with a PSA level of 0.94 ng/mL and a PSA density of 0.03 ng/mL/cc. The images show a suspicious area in the left pe-

ripheral zone with low T2W signal intensity (annotated in yellow) and low signal intensity on ADC map (annotated in red), classified as a PI-RADS 5 lesion. Both variants of the dataset (original and deformably aligned) show very similar efficacy (predictions of 0.48 for both) due to similar alignment of the lesion's features across scans. Targeted biopsy of this area revealed ISUP 2 prostate cancer.

The third row shows the images for a 66-year-old man with a PSA level of 13 ng/mL and a PSA density of 0.11 ng/mL/cc. The images show a tumor suspicious area ventral in the apex of the prostate close to the anterior fibromusclar stroma (AFMS) ventral to the transition zone of the prostate. The delineation of the lesion mask was guided by the image features observed in the ADC scan and was subsequently adopted for the T2W scan as well. Upon reconsideration of the lesion segmentation with two radiologists, it appears that the extension into the AFMS is due to oversegmentation, rather than the lesion infiltrating the AFMS. As such, the model predictions capture the lesion extent very well. The prediction with the original dataset had a bit higher confidence (0.62 vs 0.56) for this positive case. Targeted biopsy of this area revealed ISUP 2 prostate cancer.

Results for the PROMIS dataset are shown in Figure D. The first two cases were selected to have the largest prediction increase and decrease for the deformably aligned dataset, compared to the original datasets, for cases with a case-level prediction above 0.3, respectively. The third case was a failure case with the deformably aligned scans. The last two cases were selected to have the largest prediction increase and decrease for the deformably aligned dataset, compared to the original datasets, respectively.

The first row shows the images for a 74-year-old man with a PSA level of 12 ng/mL and a PSA density of 0.27 ng/mL/cc. The images show a well-defined lesion in the left peripheral zone midprostate, with low signal intensity on T2W, and low signal intensity on the ADC map, consistent with a suspicious lesion (PI-RADS 4). The T2W and ADC map are misaligned, both in-plane and through-plane, resulting in the diffusion restriction on the original ADC map to be misaligned with the lesion features on the T2W sequence. The PI-CAI AI system identified the lesion with both variants of the dataset. With the deformably aligned dataset the algorithm confidence increased to 0.51, from a prediction of 0.36 before. Histopathological evaluation confirmed the aggressive nature of this lesion (ISUP 2).

The second row shows the images for a 67-year-old man with a PSA level of 5.2 ng/mL and a PSA density of 0.10 ng/mL/cc. The T2W and ADC map are misaligned in-plane. Consequently, a substantial part of the prostate on the ADC map appears outside of the prostate region of the T2W sequence. The ADC map shows diffusion restriction (low signal intensity; darker appearance) in the right transition zone midprostate. Due to the misalignment, this darker area on the original ADC map appears to be in the right peripheral zone on the T2W scan, and therefore misclassification can occur. After deformable alignment, the darker area on the ADC map aligns with the transition zone instead of peripheral zone. This is reflected in the lesion detection of the PI-CAI AI system, which predicts a lesion with confidence of 0.44 with the original scans and with a confidence of 0.18 with the deformably aligned scans. Targeted biopsies revealed ISUP 1 in the right transition zone, which is an indolent prostate cancer that is often invisible on prostate MRI. No aggressive PCa was detected.

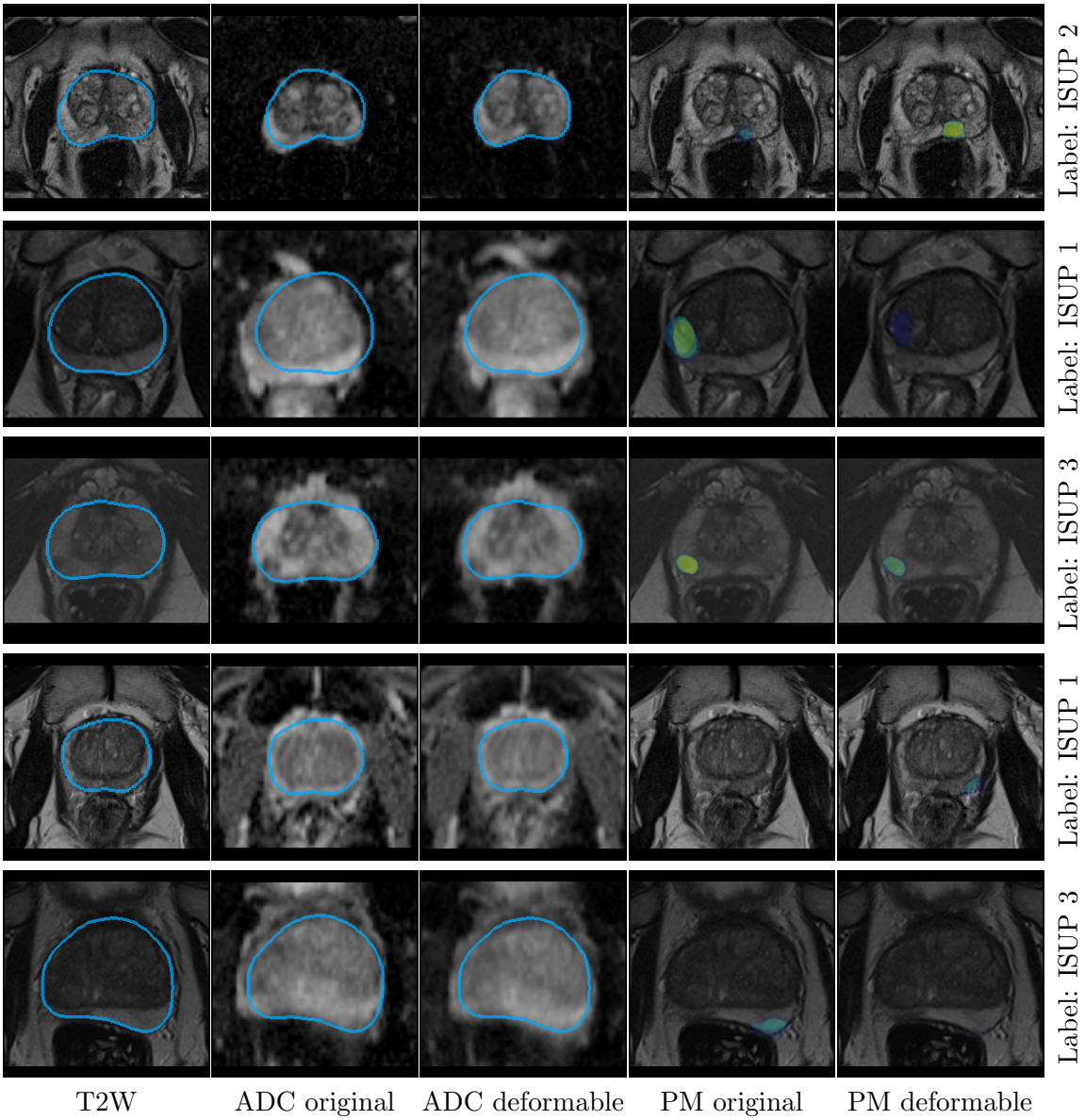

Figure D: Qualitative results on the PROMIS datset. The T2, ADC, and deformably aligned ADC are shown with ■ prostade gland. In the last two column, the prediction maps (PM) generated with the original dataset and the deformbly aligned dataset are overlayed on the T2W image. The label shows the ISUP grade, where 1 is indolent cancer (negative), and ≥ 2 is intermediate to high-risk cancer (positive). The first two cases were selected to have the largest prediction increase and decrease for the deformably aligned dataset, compared to the original datasets, for cases with a case-level prediction above 0.3, respectively. The third case was a failure case with the deformably aligned scans. The last two cases were selected to have the larges prediction increase and decrease for the deformably aligned dataset, compared to the original datasets, respectively.

The third row shows the images for a 55-year-old man with a PSA level of 12 ng/mL and a PSA density of 0.30 ng/mL/cc. The images show a small lesion in the right peripheral zone midprostate. The lesion appears as a well circumscribed area with low signal intensity (dark) on T2W images and the ADC map, suspicious for clinically significant cancer (PI-RADS 4). For this case, the deformable image registration slightly misaligned the T2W and ADC image features of the lesion, which resulted in the detection algorithm to decrease its lesion prediction from 0.49 to 0.39. Histopathological evaluation confirmed the aggressive nature of this lesion (ISUP 3).

The fourth row shows the images for a 61-year-old man with a PSA level of 7.7 ng/mL and a PSA density of 0.12 ng/mL/cc. The images show ill-defined areas of low signal intensity in the peripheral zone on both sides. On the left side there is a better defined area with low signal intensity, however, without apparent restriction on the ADC map. This was evaluated as PI-RADS 2 lesion, suggestive of (post)inflammatory changes or indolent prostate cancer. The PI-CAI AI system with the original dataset assigned a very low prediction of 0.07 to this case. The same algorithm with the deformably aligned imaging assigned a prediction of 0.33 to this lesion. This adjustment indicates a nuanced increase in the estimated risk, although it remains relatively low. Pathological evaluation classified this lesion as indolent cancer, with an ISUP grade of 1, indicating a low-risk profile.

The fifth row presents the images for a 75-year-old man with a PSA level of 7.2 ng/mL and a PSA density of 0.08 ng/mL/cc. Template mapping biopsy as part of the PROMIS trial identified an aggressive lesion in the left side of the prostate (ISUP 3). This lesion was not found during prospective radiological assessment. Retrospective consultation with an expert urogenital radiologist also did not reveal image features that indicate the location of this lesion.

Appendix G. Diagnostic performance

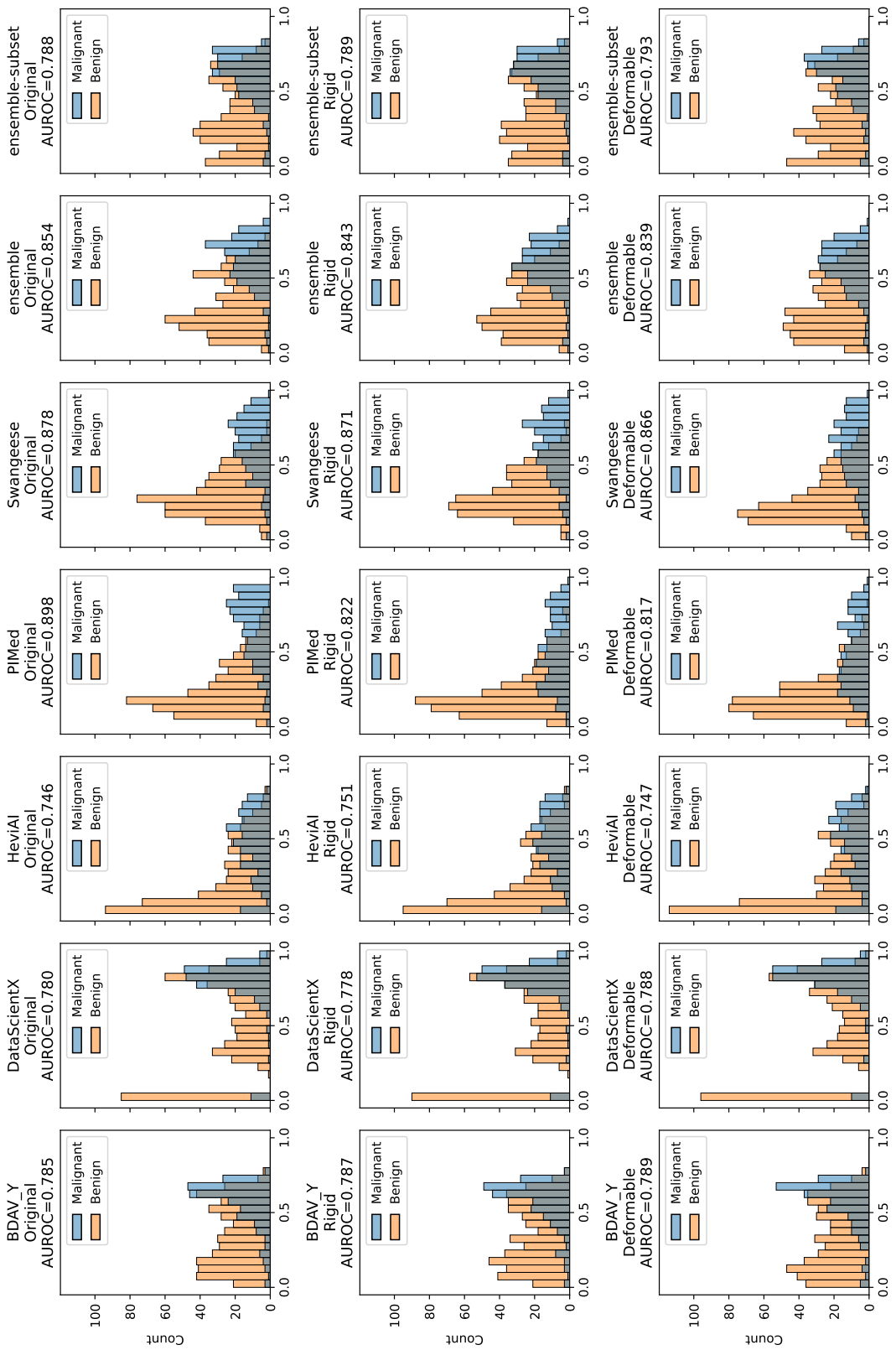

Figure E: The prediction distribution for the PCNN dataset of each PI-CAI algorithm, their ensemble, and the ensemble of 3 PI-CAI algorithms (BDAV_Y, DataScientX and HeviAI) "ensemble-subset".

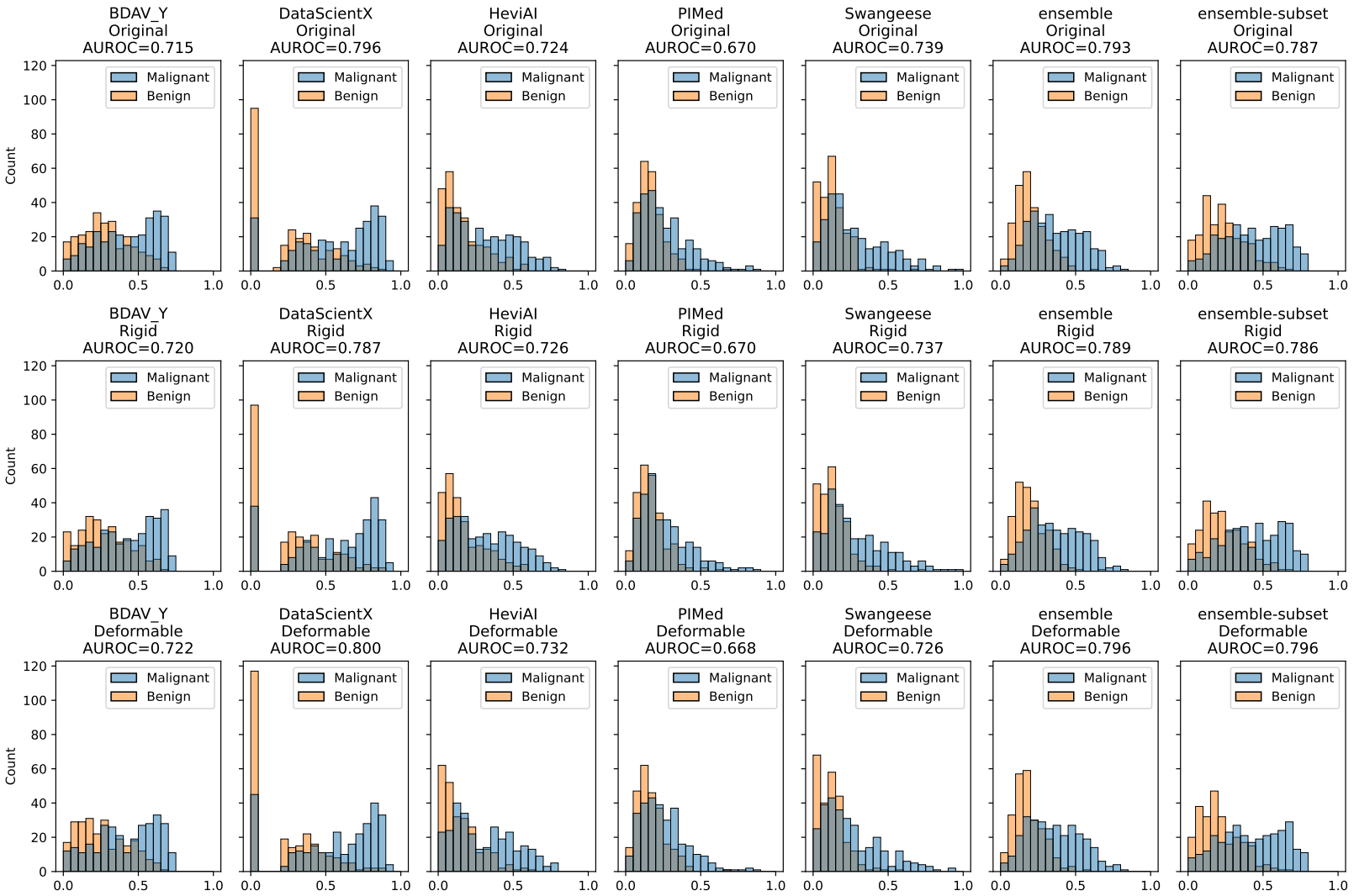

Figure F: The prediction distribution for the PROMIS dataset of each PI-CAI algorithm, their ensemble, and the ensemble of 3 PI-CAI algorithms (BDAV_Y, DataScientX and HeviAI) "ensemble-subset".

Appendix H. PROMIS Dataset acknowledgements

The PROMIS data used in the analysis for this manuscript were provided from the PROMIS study, led by University College London (UCL). PROMIS was funded by the UK Government Department of Health, National Institute of Health Research–Health Technology Assessment Programme, (Project number 09/22/67). Support was also provided by National Institute for Health Research (NIHR) UCLH/UCL Biomedical Research Centre, National Institute for Health Research (NIHR) The Royal Marsden and Institute for Cancer Research Biomedical Research Centre and National Institute for Health Research (NIHR) Imperial Biomedical Research Centre. The original PROMIS study was coordinated by the Medical Research Council Clinical Trials Unit (MRC CTU) at UCL and sponsored by UCL. The PROMIS Biobank was funded by Prostate Cancer UK (PG10-17). The PROMIS dataset and the biobank is under the research governance of the ReIMAGINE Risk Trial Management Group (funded by Medical Research Council (UKRI) and Cancer Research UK: MR/R014043/1).