Computers in Biology and Medicine 171 (2024) 108216

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Analysis of domain shift in whole prostate gland, zonal and lesions segmentation and detection, using multicentric retrospective data

Nuno Miguel Rodrigues [a,b,*,1], José Guilherme de Almeida [b,1], Ana Sofia Castro Verde [b], Ana Mascarenhas Gaivão [c], Carlos Bilreiro [c], Inês Santiago [c], Joana Ip [c], Sara Belião [c], Raquel Moreno [a], Celso Matos [a], Leonardo Vanneschi [d], Manolis Tsiknakis [e,f], Kostas Marias [f,g], Daniele Regge [h,i], Sara Silva [b,2], ProCAncer-I Consortium [3], Nickolas Papanikolaou [a,j,2]

[a] Computational Clinical Imaging Group, Champalimaud Foundation, Portugal
[b] LASIGE, Faculty of Sciences, University of Lisbon, Portugal
[c] Radiology Department, Champalimaud Clinical Center, Champalimaud Foundation, Lisbon, Portugal
[d] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal
[e] Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), GR 700 13, Heraklion, Greece
[f] Department of Electrical and Computer Engineering, Hellenic Mediterranean University, GR 710 04, Heraklion, Greece
[g] Computational BioMedicine Laboratory (CBML), Institute of Computer Science, Foundation for Research and Technology – Hellas (FORTH), Heraklion, Greece
[h] Department of Radiology, Candiolo Cancer Institute, FPO-IRCCS, Strada Provinciale 142 Km 3.95, Candiolo, Turin 10060, Italy
[i] Department of Surgical Sciences, University of Turin, Turin 10124, Italy
[j] Department of Radiology, Royal Marsden Hospital, Sutton, UK

ARTICLE INFO

Keywords:
ProstateNet
Prostate segmentation
Lesion segmentation
Zone segmentation

ABSTRACT

Despite being one of the most prevalent forms of cancer, prostate cancer (PCa) shows a significantly high survival rate, provided there is timely detection and treatment. Computational methods can help make this detection process considerably faster and more robust. However, some modern machine-learning approaches require accurate segmentation of the prostate gland and the index lesion. Since performing manual segmentations is a very time-consuming task, and highly prone to inter-observer variability, there is a need to develop robust semi-automatic segmentation models. In this work, we leverage the large and highly diverse ProstateNet dataset, which includes 638 whole gland and 461 lesion segmentation masks, from 3 different scanner manufacturers provided by 14 institutions, in addition to other 3 independent public datasets, to train accurate and robust segmentation models for the whole prostate gland, zones and lesions. We show that models trained on large amounts of diverse data are better at generalizing to data from other institutions and obtained with other manufacturers, outperforming models trained on single-institution single-manufacturer datasets in all segmentation tasks. Furthermore, we show that lesion segmentation models trained on ProstateNet can be reliably used as lesion detection models.

1. Introduction

According to the 2022 cancer statistics provided by the American Cancer Society, Prostate Cancer is the most prominent cancer in males, and the second most prominent cancer overall, behind breast cancer [1]. Despite the high prevalence, it has a low mortality rate of ≈ 12%, although early detection is key for optimal treatment outcomes. Several computational methods for prostate cancer have been developed [2–6], but they oftentimes require segmentations of either the prostate or index lesions. While manual segmentations are currently the most common practice, performing them is a highly time-consuming task notably dependent on the expertise of the radiologists, with a very high degree of inter- and intra-observer variability [7–9]. Thus, there is a need to develop reliable and robust automatic segmentation models that can help clinicians deliver timely and accurate diagnoses.

Over the last few years, an extensive collection of models for automatic segmentation of the whole prostate gland, zones and lesions

* Corresponding author at: LASIGE, Faculty of Sciences, University of Lisbon, Portugal.
E-mail addresses: nmrodrigues@fc.ul.pt (N.M. Rodrigues), jose.almeida@research.fchampalimaud.org (J.G.d. Almeida).
[1] Shared authorship.
[2] Shared senior authorship.
[3] www.procancer-i.eu complete list of members in Supplementary Information.

https://doi.org/10.1016/j.compbiomed.2024.108216
Received 12 December 2023; Received in revised form 9 February 2024; Accepted 25 February 2024
Available online 2 March 2024

have been proposed, using a wide variety of deep learning architectures, from standard convolutional neural network variants to vision transformers and object detectors [10–21]. Despite this abundance of models, prior works have shown that several of these different U-Net-based models have very similar capabilities when performing prostate gland and zone segmentation, showing little to no statistical difference [17,22].

However, similarly to many other studies, a common caveat is that these models were trained on a reduced sample size acquired with a single type of scanner. The effect of distribution shifts associated with different scanners on the performance of segmentation and detection models in biomedical images has been previously noted. This is particularly true for brain MRI segmentation [23,24], which may lead to downstream errors in quantifying clinically relevant aspects of the patient [25]. However, when it comes to prostate MRI, little is known in terms of the impact that different scanners may have on the learning and on how well these models generalize.

In this study, we make use of the ProstateNet image archive, which was created under the scope of the ProCAncer-I project, which contains large amounts of retrospective multi-vendor and multi-institution data that were used to train segmentation models for three distinct tasks: whole prostate gland, prostate zones (peripheral and transitional zones) and index lesion segmentation. Additionally, we also make use of two publicly available benchmark datasets: ProstateX, and Prostate158, to compare to and enhance ProstateNet. An interesting and similar work was done by Meglič et al. [26], where they joined the publicly available ProstateX data with private internal data (both obtained from Siemens scanners), to improve whole gland, peripheral and transitional zone segmentation performance. Our experiments expand and complement these experiments by including a significantly larger sample of data, obtained from an extensive range of scanner and institutional providers.

The results show that for whole gland and zone segmentation, the models benefit from data abundance and variability to be able to generalize on out-of-institution data, becoming more robust to domain shifts. For index lesion segmentation, we show that simply having large amounts of data is not enough. Additionally, for index lesion segmentation, we show that more intricate strategies, such as using multiparametric MRI (mpMRI, a technique which uses anatomical and functional MRI sequences) data or multi-resolution models, do not yield better results than regular full-resolution models trained on T2 W (anatomical) data. However, while the segmentation performance is relatively poor, we show that some of these index lesion segmentation models can be reliably used as robust index lesion detection models.

2. Methods

2.1. Data

Four different datasets were used in this study:

- **Prostate158** is a collection of biparametric MRI volumes that include T2 W, DWI and ADC modalities. These volumes were obtained by the German university hospital - Charité University Hospital Berlin, using Siemens 3T MR scanners (VIDA and Skyra). Regarding the acquisition of the images, the following description was provided by the authors of the paper: "T2w sequences were acquired with the following parameters: slice thickness 3 mm, no interslice gap, in-plane resolution 0.47 × 0.47 mm, field of view (FOV) size 180×180 mm, time to echo (TE)/repetition time (TR) 116 ms/4040 ms, turbo factor 25, flip angle 160°, acquisition time 3 min and 56s". [27]
- **ProstateX** is a collection of prostate MRI volumes that include T2 W, DWI and ADC modalities. These volumes were obtained by the Prostate MR Reference Center — Radboud University Medical Centre (Radboudumc) in the Netherlands, using two Siemens 3T MR scanners (MAGNETOM Trio and Skyra). Regarding the acquisition of the images, the following description was provided by the challenge's organizers: "T2-weighted images were acquired using a turbo spin echo sequence and had a resolution of around 0.5 mm in plane and a slice thickness of 3.6 mm" [28]. The manual segmentations used as ground truth for both the prostate gland and the lesions were performed independently by an expert radiologist (M.L., 10 years of experience) on T2 W and DWI sequences separately (153 volumes total for each sequence), while the transition and peripheral zone ground truth masks were obtained from the public dataset repository (139 T2 W volumes).
- **ProstateNet** is a collection of multiparametric MRI volumes, that include T2 W, DWI and ADC modalities. These volumes were obtained by 12 clinical partners of the Procancer-I project. These partners used Siemens (Aera, Skyra, Sola, Avanto, VIDA, Tim, Prisma, Veri, Symphony, Osirix), Philips (Ingenia, Achieva, Multiva) and GE scanners (Optima, Signa, DISCOVERY). Given that each centre has specific acquisition protocols, no single one was used across all mpMRI studies done. Given that labels could be defined automatically (using the QUIBIM-developed ProCAncer-I prostate region segmentation tool), and could either be manually corrected but not validated and manually corrected and validated, we define a hierarchy of annotations, selecting the whichever one is available first: (1) manually corrected and validated ($n = 610$), (2) manually corrected but not validated ($n = 30$), (3) automatically generated ($n = 65$).
- **ProstateAll** is a combination of all the previous datasets. It was created with the purpose of increasing the data pool used to train the models, regarding both the total number of samples, as well as regarding data variability, with the aim of producing more robust models.

Table 1
Stratification of samples by manufacturer for all four segmentation datasets per task. Both the ProstateNet and ProstateAll datasets also include a very residual amount (≤ 5) of Toshiba scanner samples, which were accounted for in the total values.

Gland				
	Total	Siemens	Philips	GE
Prostate158	139	139	–	–
ProstateX	182	182	–	–
ProstateNet	638	152	245	239
ProstateAll	959	473	245	239
Zones				
Prostate158	139	139	–	–
ProstateX	181	181	–	–
ProstateNet	638	152	245	239
ProstateAll	958	472	245	239
Lesions				
Prostate158	82	82	–	–
ProstateX	190	190	–	–
ProstateNet	461	136	184	136
ProstateAll	733	408	184	136
ProstateNet mpMRI	417	131	178	107

Table 1 shows the composition of the different datasets. for each segmentation task (gland, zones and lesion) stratified by scanner manufacturer. From these numbers, 15% of the samples were used as a hold-out test set, and the remaining were used for training, following a 5-fold cross-validation strategy. For the mpMRI study, a total of 417 cases were used (354 for training and 64 for hold-out). Details on the distribution of scanner model per data partition of the ProstateNet dataset are available in the Supplementary Methods (Fig. B.1).

2.2. Experiments

For this study, we focus on three different tasks:

- **Whole prostate gland segmentation** is the simplest task of the three and consists in producing a binary map where the positive values correspond to pixels belonging to the whole prostate gland. For this, exclusively T2 W sequences were used;
- **Prostate zone segmentation** can be seen as a slightly harder subset of the whole gland segmentation and consists of the semantic segmentation of the peripheral zone (PZ) and the transitional zone with the central zone (TZ). Similarly to the previous task, exclusively T2 W sequences were used;
- **Lesion segmentation** is the hardest segmentation task regarding the prostate due to the small size, low visibility and irregular shape of the lesions. For this task, all annotated lesions were predicted for each study. Additionally, for the ProstateNet dataset, apart from using T2 W data, we assessed whether using a combination of T2 W, DWI and ADC (mpMRI) sequences could improve segmentation performance.

2.3. Deep learning model specification

Three distinct 3D deep-learning (DL) segmentation models were trained - a simple U-Net model [29] (U-Net), a U-Net model with deep supervision [30] (U-Net + D.S.), and a full resolution nnUNet model (nnUNet) [31], which also uses deep supervision. Deep supervision is a technique that uses intermediate predictions, generated at each step of the decoder to ensure that the model is learning at all stages of the encoder and decoder rather than taking shortcuts through its skip connections. All networks are implemented in Pytorch [32] and were trained for 1000 epochs (250 mini-batches per epoch) and 200 epochs, for the nnUnet and U-Net/U-Net + D.S., respectively. Training for U-Net and U-Net + D.S. was performed using Lightning [33], a low-code and heavily customizable framework for neural network training and testing in PyTorch.

To train the nnUNet models, we used the provided 3D full resolution architecture. This framework uses stochastic gradient descent with Nesterov momentum ($\mu = 0.99$), a maximum initial learning rate of 0.01, and polynomial [34] learning rate policy which reduces the learning rate by a factor of $(1 - epoch/epoch_{max})^{0.9}$ in each epoch. The loss function is an equal combination of Dice and cross-entropy losses and the batch size was 2 sequences per iteration. nnUNet applies automatic preprocessing based on the dataset fingerprint, and therefore the models for each dataset worked on data with slightly different spatial structures:

- **ProstateX**: spacing = 0.5×0.5×3.0 mm; crop size = 320 × 320 × 16 voxels
- **Prostate158**: spacing = 0.4 × 0.4 × 3.0 mm; crop size = 256 × 256 × 28 voxels
- **ProstateNet**: spacing = 0.5 × 0.5 × 3.0 mm; crop size = 256 × 256 × 28 voxels
- **ProstateAll**: spacing = 0.5 × 0.5 × 3.0 mm; crop size = 256 × 256 × 26 voxels

Details on training the remaining U-Net models are available in the Supplementary Methods (E.1).

Additionally, we perform a self-contained study on whole gland segmentation using transformed-based models (i.e. UNETR [35], Swin-UNETR [36]), showing how they perform worse than nnUNet models (Supplementary Methods (E.2 and E.3)).

2.4. Model evaluation

Each model is evaluated by its Dice score (DS) using 5-fold cross-validation (CV) according to the best observed DS during training, and its generalizability is assessed using the hold-out test set. To assess how models perform on different datasets, we test each model on the hold-out test set of each different dataset. Additionally, since DS only provides an overlap score, we also include the Hausdorff Distance (HD), Average Symmetric Surface distance (ASSD), and Relative Absolute Volume Difference (RAVD) during quality assessment of the model, as these metrics provide a quantitative measure of the spatial accuracy by considering the shape and volume of the segmented regions [37] (both distance metrics were calculated using MedPy [38]). Details on each metric are available in the Supplementary Methods (E.4).

Table 2
Stratification of prospective samples by manufacturer for Gland (PZ and TZ included) and index Lesion segmentation.

	Total	Siemens	Philips	GE
Gland	211	29	176	6
Lesions	19	7	8	4

2.5. Prospective validation

Lastly, to further assess how the developed models perform on contemporary data, a prospective validation is performed on a smaller cohort of patients. The prospective cases were downloaded from the ProstateNet platform on October 11th 2023. An overview of the data, stratified by manufacturer, can be seen in Table 2. Since whole gland masks are generated by merging both Peripheral (PZ) and Transitional+Center (TZ) masks, they share the same data composition.

3. Results

3.1. Data variety and size are associated with better performance in whole prostate and zone segmentation

Whole gland segmentation

For whole gland segmentation, we observe a relatively small range of performances in CV, with both U-Net variants achieving Dice scores between 0.89 and 0.91, and the nnUNet between 0.91 and 0.93 (Fig. 1 and Table 3).

For the hold-out test set, the same behaviour is observed but only when models are trained and tested on data from the same distribution (i.e. same dataset; Fig. 1A and Table 4).

We observe that the performance of nnUNet was almost always better than the one of other U-Net models (Fig. 1B), both for CV and for the hold-out test set. This is on par with previous studies [17,31] and holds up for both in- and out-of-distribution data. Taking this into consideration, we did not use any other models apart from nnUNet for the remaining tasks.

When looking at the performance of the nnUNet models trained on Prostate158 and then evaluated on ProstateX, it is possible to observe that it produces errors (HD) up to 100× larger than all other models (Fig. 2 and Table 4). While unexpected – both datasets were obtained using Siemens scanners – the performance differences may be associated with the specific scanner model or centre. These shifts in performance are considerably smaller when models are trained on ProstateNet, hinting that model-specific effects on performance decrease as the amount and variety of data increases. This effect is extendable from datasets to scanner manufacturers (Fig. 1B right). Indeed, the model-specific effects on performance become increasingly negligible as the amount and variability of data increases. Table 4 highlights the benefits of applying models to data which is similar to the training data — the nnUNet model trained on the ProstateAll data produces far better distance metrics across all datasets. To better understand where these models failed, a qualitative analysis was performed; this showed that most errors were actually poor quality annotations (details in the Supplementary Results C).

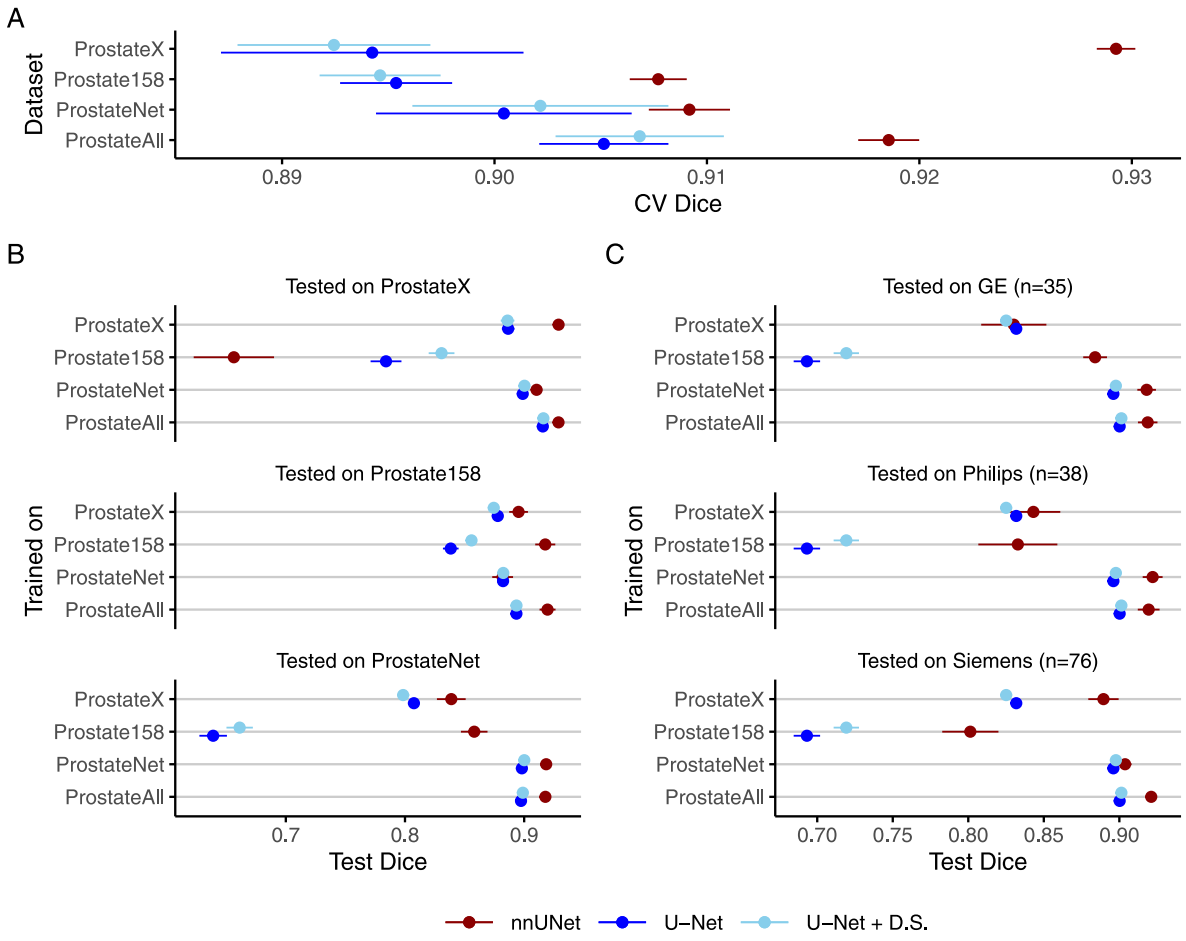

Fig. 1. Whole gland segmentation performance of the nnUNet and U-Net models. (A) CV Dice scores. (B) Hold-out test set Dice scores. Left: performance stratified by dataset. Right: performance stratified by manufacturer.

Prostate zone segmentation. During CV, zonal segmentation models present diverse levels of performance, in terms of both the segmented zone and different sets of data used during training (Fig. 3 and Table 3) — both ProstateX and Prostate158 models yield Dice scores significantly worse than those of ProstateAll in the context of PZ segmentation. Apart from the lower Dice, it can be noted that Prostate158 produces maximum errors (Hausdorf) $5-8$ mm bigger than the remaining models.

Considering the hold-out test performance, and similarly to what was shown for the whole gland segmentation, it can be seen that the performance is similar between CV and hold-out test set (Tables 3 and 5). As expected, based on the previous results, both ProstateX and Prostate158 models fail to generalize to multi-vendor data, suffering a severe drop in performance when tested on ProstateNet/ProstateAll data. On par with what was observed during whole gland segmentation, both ProstateX and Prostate158 models fail to generalize to each other despite having been acquired with scanners produced by the same manufacturer. Additionally, it can be seen that the trend of producing very large maximum errors (HD) - on average 100× (PZ) and 200× (TZ) larger than all other models (Fig. 2) - is kept, when evaluating the Prostate158 models on ProstateX data.

Learning curve analysis. To better understand how data variety and size impact performance, we conducted a simple learning curve analysis by training the nnUNet model on different proportions of the ProstateAll training dataset ($[0.1, 0.5, 0.7, 1.0]$). Initially, our conceptual understanding focused on data variability, rather than on the amount of data — i.e. having data from more diverse sources would lead to improved performance. However, it was also possible that simply increasing the amount of data could lead to better performance. The CV results allow us to validate the expected outcome of this analysis (Fig. 4 **B**) — performance increases as the amount of data increases. Extending this analysis to the hold-out test set and stratifying by testing dataset (Fig. 4 **C**) while comparing the learning curve performance with that of models trained and tested on the same dataset reveals something striking – for starters, even at relatively low amounts of ProstateAll data (0.1 – corresponding to 65 cases) the performance is at least comparable to that of models trained and tested on the same dataset, suggesting that data variability leads to improved performance. The additional insight is that increasing the size of the training set – which inevitably increases the variability of the data as well – also leads to improved performance which is oftentimes better than that observed in models trained and tested on the same data. For this reason, we suggest here that this increase in performance is driven by a combination of increased data size and increased variability, highlighting the importance of initiatives such as ProstateNet in training algorithms that can be clinically deployable and robust to large shifts in performance at test and inference time.

3.2. Increasing dataset size and diversity improves index lesion detection

The performance of index lesion segmentation models is generally worse when compared with that of other prostate segmentation tasks (Fig. 3 and Table 3), mostly due to the small size and irregular shape of the lesions. Using data which is both more abundant and more diverse – ProstateNet/ProstateAll – improves performance when compared with ProstateX and Prostate158, two small, single-institution datasets. This

Table 3
nnUNet CV results stratified by segmentation task. For each dataset, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviations, are presented. p-values for Kruskal–Wallis significance test comparing the Dice score between ProstateAll results and each other model are also shown, with significant results (p-value < 0.01) are marked as **green**.

	Dice	Hausdorf	RAVD	ASSD	Recall	p-value	
	Gland						
ProstateX	0.93 ± 0.02	8.01 ± 4.36	0.01 ± 0.07	0.32 ± 0.1	1.0 ± 0.0	0.1745	
Prostate158	0.91 ± 0.03	14.02 ± 11.81	0.0 ± 0.1	0.35 ± 0.19	1.0 ± 0.0	0.0758	
ProstateNet	0.91 ± 0.09	12.48 ± 23.27	0.09 ± 1.78	0.43 ± 0.72	1.0 ± 0.0	0.2506	
ProstateAll	0.92 ± 0.08	12.71 ± 27.82	0.06 ± 1.42	0.45 ± 1.79	1.0 ± 0.0	–	
	Zones PZ TZ						
ProstateX	0.8 ± 0.08	16.45 ± 9.75	−0.0 ± 0.24	0.64 ± 0.39	1.0 ± 0.0	**0.009**	
	0.88 ± 0.05	14.97 ± 9.31	0.02 ± 0.17	0.55 ± 0.22	1.0 ± 0.0	**0.0758**	
Prostate158	0.76 ± 0.09	20.95 ± 18.13	−0.01 ± 0.26	0.64 ± 1.16	1.0 ± 0.0	**0.009**	
	0.88 ± 0.06	22.09 ± 16.84	0.03 ± 0.14	0.44 ± 0.35	1.0 ± 0.0	**0.0758**	
ProstateNet	0.81 ± 0.11	15.7 ± 22.22	0.08 ± 1.45	0.56 ± 0.8	1.0 ± 0.0	0.4647	
	0.89 ± 0.08	13.13 ± 11.9	0.14 ± 2.2	0.5 ± 0.62	1.0 ± 0.0	0.0758	
ProstateAll	0.82 ± 0.1	15.43 ± 19.97	0.06 ± 1.24	0.5 ± 0.81	1.0 ± 0.0	–	PZ
	0.9 ± 0.08	14.61 ± 13.34	0.09 ± 1.78	0.44 ± 0.54	1.0 ± 0.0		TZ
	Lesions						
ProstateX	0.17 ± 0.24	100.45 ± 89.78	2.22 ± 13.35	32.9 ± 51.24	0.4 ± 0.04	0.009	
Prostate158	0.25 ± 0.27	95.44 ± 87.01	0.04 ± 1.17	24.5 ± 47.58	0.5 ± 0.06	0.1172	
ProstateNet	0.38 ± 0.3	66.45 ± 66.83	0.33 ± 2.79	18.43 ± 35.56	0.7 ± 0.02	0.4647	
ProstateAll	0.36 ± 0.3	77.3 ± 73.68	0.9 ± 9.75	22.74 ± 39.54	0.65 ± 0.02	–	
ProstateAll Cascade	0.36 ± 0.3	81.03 ± 73.59	0.68 ± 8.39	23.98 ± 40.4	0.65 ± 0.02	0.9168	
ProstateNet mpMRI	0.4 ± 0.28	71.3 ± 69.18	0.61 ± 3.16	15.37 ± 27.46	0.76 ± 0.02	0.1172	

Table 4
nnUNet whole gland segmentation hold-out test set results. For each pairwise evaluation, the average Dice, Hausdorf, RAVD and ASSD performance, along with their respective standard deviations, are presented.

		Tested on				
		ProstateX	Prostate158	ProstateNet	ProstateAll	
Trained on	ProstateX	0.93 ± 0.02	0.9 ± 0.04	0.84 ± 0.12	0.86 ± 0.11	Dice
		8.8 ± 4.61	19.12 ± 35.05	47.99 ± 76.07	36.56 ± 65.85	Hausdorf
		−0.03 ± 0.06	0.05 ± 0.11	0.24 ± 0.56	0.17 ± 0.48	RAVD
		0.32 ± 0.08	0.48 ± 0.45	2.91 ± 5.75	2.08 ± 4.86	ASSD
	Prostate158	0.66 ± 0.18	0.92 ± 0.04	0.86 ± 0.11	0.83 ± 0.15	
		364.2 ± 87.76	13.04 ± 12.98	22.37 ± 42.1	85.29 ± 143.78	
		0.95 ± 0.84	0.03 ± 0.08	0.19 ± 0.62	0.31 ± 0.7	
		49.91 ± 24.9	0.31 ± 0.18	1.74 ± 5.76	10.59 ± 22.29	
	ProstateNet	0.91 ± 0.02	0.88 ± 0.04	0.92 ± 0.04	0.91 ± 0.04	
		8.64 ± 3.75	13.23 ± 6.88	9.88 ± 9.94	10.12 ± 8.8	
		−0.07 ± 0.07	−0.03 ± 0.08	−0.0 ± 0.08	−0.02 ± 0.08	
		0.39 ± 0.09	0.42 ± 0.14	0.34 ± 0.15	0.36 ± 0.14	
	ProstateAll	0.93 ± 0.02	0.92 ± 0.03	0.92 ± 0.04	0.92 ± 0.04	
		9.01 ± 4.52	10.71 ± 8.28	9.62 ± 9.76	9.66 ± 8.81	
		−0.03 ± 0.06	0.02 ± 0.07	0.01 ± 0.08	0.0 ± 0.08	
		0.31 ± 0.07	0.27 ± 0.1	0.34 ± 0.14	0.33 ± 0.13	

is true not only of Dice scores, but also of HD and ASSD (both generally lower for both ProstateNet and ProstateAll).

Apart from the standard single modality full-resolution models, we also trained a cascade model for the ProstateAll T2 W data (in theory more capable of capturing small objects) and a full-resolution model using ProstateNet T2 W, DWI and ADC (mpMRI) images (DWI and ADC are more appropriate than T2 W, as they hyper- and hypo-saturate in the areas where index lesions exist). During CV, while the cascade model provided minimal differences apart from slightly smaller errors, the ProstateNet model using the three modalities showed some improvements, most notably in terms of dice and recall, albeit these improvements were not statistically significant.

Similarly to what was shown for whole gland and zone segmentation models, the majority of in-distribution hold-out results (Table D.4) are similar to those obtained during CV. Regarding out-of-distribution performance, both single-manufacturer and institution models (ProstateX and Prostate158) show a performance drop on multi-centric multi-manufacturer data, while ProstateAll provides consistent results all throughout. Additionally, the performance drops significantly when testing the Prostate158 model on the ProstateX dataset. Contrary to what was hypothesized, the mpMRI ProstateNet and the Cascade ProstateAll models showed a poor generalization performance, decreasing more when compared to its T2W-only counterpart, in both Dice and recall. Overall, the best model turned out to be the ProstateAll T2W-only model, particularly when considering the average precision for each model (Fig. 5) — this is further highlighted in Fig. 6, showing the co-localization of index lesions and voxels predicted as having high index lesion probability(see Table 6).

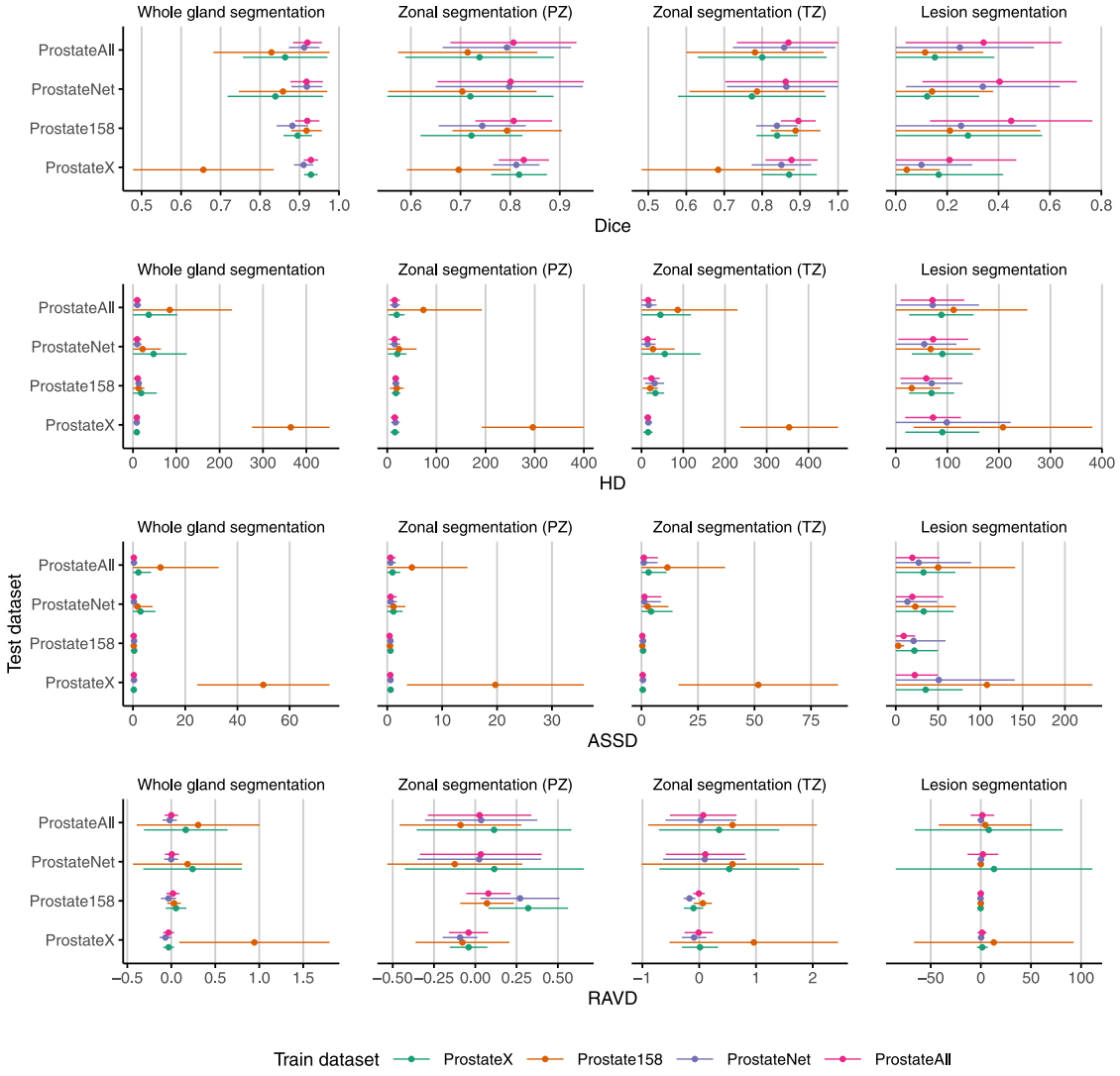

Fig. 2. Multi-metric analysis of the hold-test performance of the nnUNet models, in- and out-of-distribution, stratified by segmentation task.

Table 5
nnUNet PZ and TZ segmentation hold-out test set results. For each pairwise evaluation, the average Dice, Hausdorf, RAVD and ASSD performance, along with their respective standard deviations, are presented.

		Tested on								
		ProstateX		Prostate158		ProstateNet		ProstateAll		
Trained on	ProstateX	0.82 ± 0.06	0.87 ± 0.07	0.72 ± 0.1	0.84 ± 0.05	0.72 ± 0.17	0.77 ± 0.19	0.74 ± 0.15	0.8 ± 0.17	Dice
		15.66 ± 8.74	15.73 ± 11.06	17.73 ± 8.4	33.21 ± 20.62	20.46 ± 18.72	56.17 ± 85.67	19.2 ± 16.3	45.56 ± 73.0	Hausdorf
		−0.04 ± 0.11	0.01 ± 0.31	0.32 ± 0.24	−0.1 ± 0.17	0.12 ± 0.54	0.54 ± 1.23	0.12 ± 0.47	0.35 ± 1.06	RAVD
		0.61 ± 0.48	0.57 ± 0.35	0.59 ± 0.3	0.71 ± 0.22	1.14 ± 1.65	4.28 ± 9.5	0.97 ± 1.4	3.11 ± 8.01	ASSD
	Prostate158	0.7 ± 0.1	0.68 ± 0.2	0.79 ± 0.11	0.89 ± 0.06	0.7 ± 0.15	0.79 ± 0.18	0.71 ± 0.14	0.78 ± 0.18	
		296.17 ± 101.81	353.73 ± 115.13	19.48 ± 13.78	20.96 ± 18.05	23.92 ± 35.72	27.65 ± 51.67	73.94 ± 118.74	87.37 ± 143.28	
		−0.08 ± 0.28	0.96 ± 1.46	0.07 ± 0.16	0.06 ± 0.16	−0.12 ± 0.41	0.59 ± 1.6	−0.09 ± 0.37	0.59 ± 1.49	
		19.69 ± 15.82	51.68 ± 34.68	0.51 ± 0.48	0.36 ± 0.15	1.16 ± 2.13	2.78 ± 9.16	4.51 ± 10.11	11.55 ± 25.48	
	ProstateNet	0.81 ± 0.05	0.85 ± 0.08	0.74 ± 0.09	0.84 ± 0.05	0.8 ± 0.15	0.86 ± 0.16	0.79 ± 0.13	0.86 ± 0.13	
		16.73 ± 8.43	16.52 ± 8.18	17.49 ± 6.84	31.48 ± 22.19	15.02 ± 11.23	14.77 ± 19.17	15.67 ± 10.3	17.34 ± 18.97	
		−0.09 ± 0.1	−0.09 ± 0.21	0.27 ± 0.23	−0.17 ± 0.1	0.03 ± 0.37	0.1 ± 0.73	0.04 ± 0.34	0.03 ± 0.62	
		0.59 ± 0.3	0.66 ± 0.41	0.58 ± 0.26	0.66 ± 0.29	0.63 ± 1.15	1.24 ± 7.34	0.62 ± 0.96	1.06 ± 6.06	
	ProstateAll	0.83 ± 0.05	0.88 ± 0.07	0.81 ± 0.08	0.9 ± 0.04	0.8 ± 0.15	0.86 ± 0.16	0.81 ± 0.13	0.87 ± 0.14	
		15.19 ± 7.79	15.17 ± 8.3	17.05 ± 6.94	23.98 ± 19.29	14.69 ± 11.55	14.64 ± 20.05	15.1 ± 10.44	15.99 ± 18.6	
		−0.04 ± 0.12	−0.01 ± 0.25	0.08 ± 0.13	−0.01 ± 0.1	0.03 ± 0.37	0.11 ± 0.7	0.03 ± 0.31	0.07 ± 0.59	
		0.56 ± 0.45	0.55 ± 0.35	0.43 ± 0.25	0.36 ± 0.13	0.62 ± 1.13	1.31 ± 7.51	0.58 ± 0.96	1.04 ± 6.21	
		PZ	TZ							

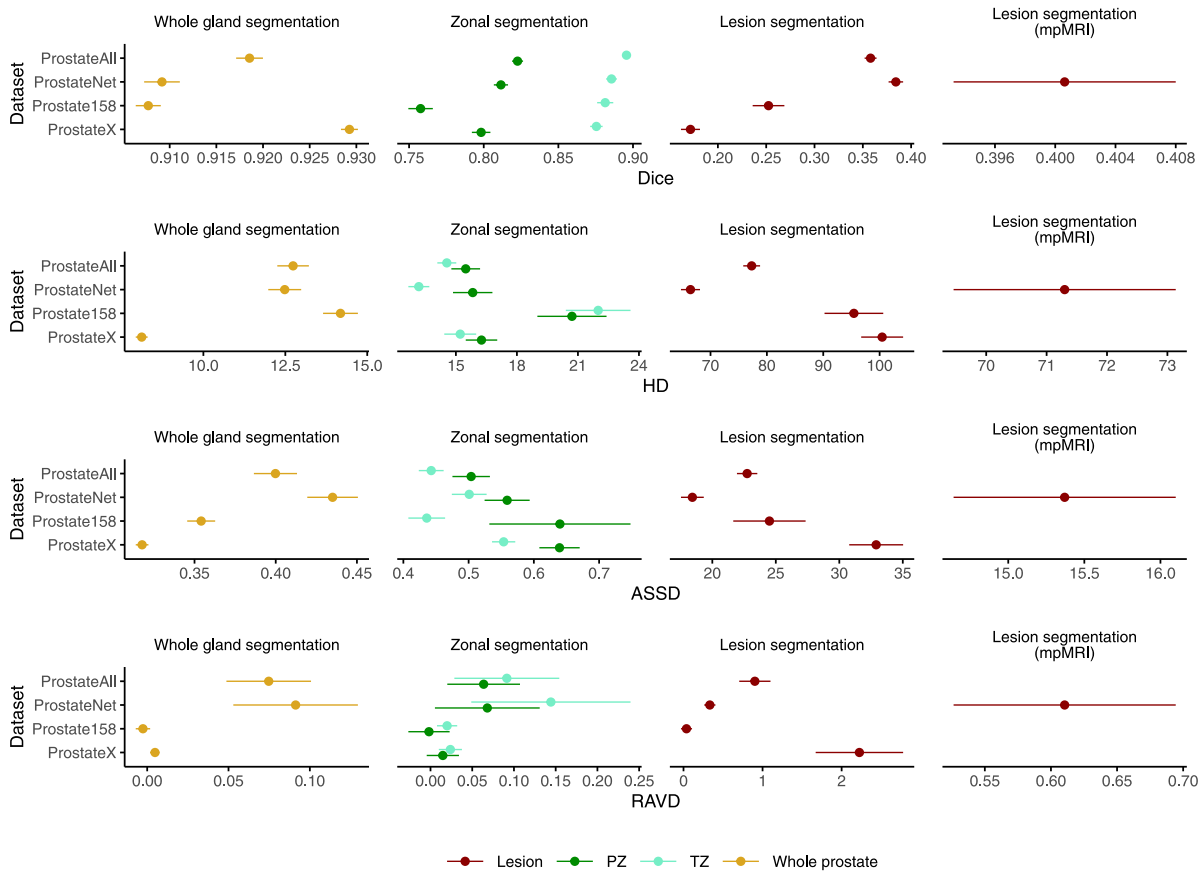

Fig. 3. Multi-metric analysis of the CV performance of the nnUNet models, stratified by segmentation task. Each circle represents the mean and the horizontal lines represent the standard error around the mean.

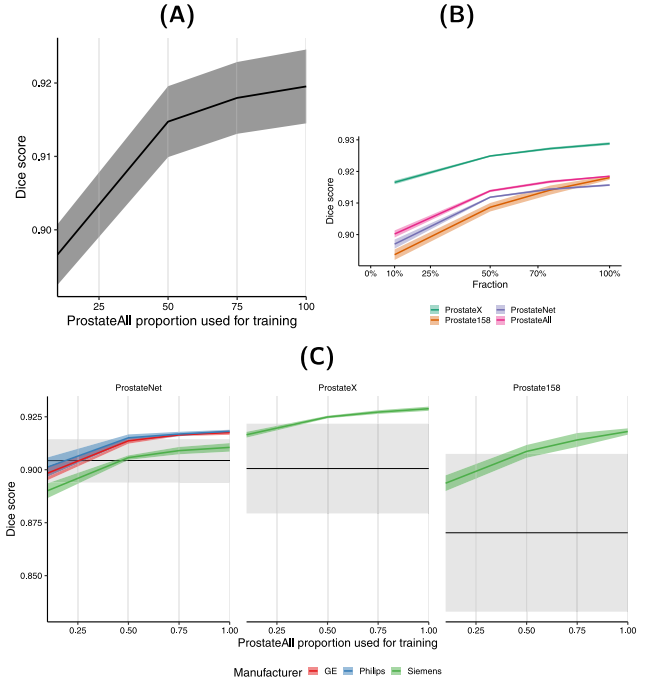

Fig. 4. Learning curves performance scores for the data partitions of ProstateAll. **A**: Cross-validation performance. **B**: Test set performance stratified by dataset.**C**: Test set performance stratified by dataset and manufacturer.

3.3. Data diversity and size may not guarantee generalization

To understand how performance is affected by changes to the testing conditions, we also analyse the differences in performance at the centre level (Figs. 7, B.4, B.2 and B.3). We note that for most cases, there are no large performance drops, even in instances where data comes from centres with a high prevalence of endorectal coil use (UNIPI, FPO). It can also be noted that Philips data produces a very diverse set of results on a wide variety of institutions, unlike other scanners, leading us to believe these differences are most likely dependent on the specific model of the scanners, rather than the institutions. Additionally, it should be noted that models trained on datasets with lower amounts of data variability (Prostate158 or ProstateX, both single-institution datasets with data acquired using Siemens scanners) tend to perform worse across most data centres. However, it should be highlighted that for some datasets, performance is relatively poor for index lesion detection 7.

Indeed, while we observe good performance in both ProstateNet and ProstateAll, it is evident that ProstateNet still does not generalize as well as ProstateAll to ProstateX or Prostate158. Indeed, the fact that ProstateAll models had access to data from all datasets during training is likely what led to their better performance across different tasks. As expected, while we still observe an improvement in out-of-distribution performance when considering ProstateNet models, it is likely that the larger benefit lies in training not only with diverse data, but also with data which is more similar to that used during testing and real-world applications.

3.4. Prospective validation

In order to understand if our segmentation models were capable of generalizing to new, prospective data, we tested models that performed

Table 6
nnUNet lesion segmentation hold-out test set results. For each pairwise evaluation, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviations, are presented.

		Tested on				
		ProstateX	Prostate158	ProstateNet	ProstateAll	
Trained on	ProstateX	0.17 ± 0.25	0.28 ± 0.28	0.12 ± 0.21	0.15 ± 0.23	Dice
		90.33 ± 70.94	69.41 ± 41.86	89.95 ± 59.44	88.11 ± 62.67	Hausdorf
		1.49 ± 5.37	−0.19 ± 0.6	13.98 ± 100.1	8.22 ± 74.71	RAVD
		35.34 ± 43.29	21.98 ± 26.6	33.03 ± 35.23	32.78 ± 37.76	ASSD
		0.39 ± 0.07	0.58 ± 0.14	0.33 ± 0.06	0.38 ± 0.04	Recall
	Prostate158	0.04 ± 0.13	0.21 ± 0.34	0.14 ± 0.23	0.11 ± 0.22	
		207.98 ± 171.3	30.94 ± 53.34	69.03 ± 96.71	114.29 ± 143.71	
		12.97 ± 78.73	0.04 ± 0.65	0.06 ± 2.14	4.6 ± 47.14	
		107.81 ± 123.17	2.97 ± 6.7	24.09 ± 48.69	51.53 ± 91.71	
		0.11 ± 0.05	0.33 ± 0.14	0.3 ± 0.06	0.24 ± 0.04	
	ProstateNet	0.1 ± 0.2	0.25 ± 0.28	0.33 ± 0.3	0.24 ± 0.29	
		98.97 ± 122.66	69.69 ± 57.29	54.52 ± 62.22	71.62 ± 90.4	
		0.31 ± 2.21	−0.25 ± 0.77	0.26 ± 1.59	0.23 ± 1.79	
		50.86 ± 88.74	21.07 ± 36.2	14.32 ± 35.78	27.83 ± 62.42	
		0.23 ± 0.06	0.58 ± 0.14	0.62 ± 0.06	0.48 ± 0.04	
	ProstateAll	0.21 ± 0.26	0.45 ± 0.3	0.39 ± 0.3	0.33 ± 0.3	
		76.23 ± 54.67	65.81 ± 43.74	67.3 ± 65.2	67.55 ± 59.25	
		1.24 ± 4.54	−0.07 ± 0.56	1.39 ± 5.96	1.18 ± 5.22	
		22.91 ± 28.61	11.69 ± 15.09	18.73 ± 34.38	19.15 ± 30.3	
		0.45 ± 0.08	0.83 ± 0.11	0.7 ± 0.06	0.63 ± 0.04	
	ProstateAll Cascade	0.21 ± 0.26	0.45 ± 0.32	0.4 ± 0.3	0.33 ± 0.3	
		68.42 ± 52.75	52.56 ± 53.6	78.81 ± 70.46	75.38 ± 64.22	
		1.21 ± 4.61	0.06 ± 0.71	2.92 ± 21.38	2.06 ± 16.14	
		21.85 ± 25.89	6.97 ± 10.84	22.25 ± 39.89	21.02 ± 34.62	
		0.48 ± 0.08	0.75 ± 0.12	0.7 ± 0.06	0.62 ± 0.04	
	ProstateNet mpMRI	–	–	0.34 ± 0.28	–	
		–	–	80.16 ± 80.62	–	
		–	–	0.16 ± 1.32	–	
		–	–	28.52 ± 47.19	–	
		–	–	0.71 ± 0.03	–	

the best on the retrospective data – Full resolution ProstateAll nnUNets – on 211 cases, for both Whole Gland and Zone segmentations, and 19 cases for index lesions segmentation.

Table 7 and Fig. 8 show the obtained results. As it can be observed, the results are fairly similar to the ones obtained during retrospective evaluation, with Whole Gland showing the largest performance drop of ≈ 4%, Peripheral Zone dropping ≈ 2%, Transitional Zone gaining ≈ 1%, and lastly, Index lesion segmentation showing a huge performance increase, of ≈ 27%. Analysing Fig. 8, it is clear that the performance of the index lesion segmentation model is quite broad, both failing to segment anything in 3 out of 19 cases, and producing Dice scores above 0.8 for 9 out of 19 cases.

When comparing the performance for each of the providers (Fig. 9) it can be observed that overall, regardless of the segmentation task, all providers show similar scores, that is, apart from RMH, that shows considerably worse performance for all metrics regarding index lesion segmentation. All zero and low scores obtained are in cases provided by that institution. Considering the scanner models associated with these images (2 Aera, 1 Vida, 1 Sola, 1 Avanto, 1 Skyra and 1 Ingenia), and looking back at the scanner model distribution (Fig. B.1), one possible cause could be due to the low prevalence of these models during training, however, this turned out to be a false assumption, given that 3/4 cases that failed came from either Aera or Skyra, which were very prevalent in the training set.

Additionally, it is also possible to show that the index lesion segmentation models are capable of detecting the index lesions, presenting a very high Recall, of 0.86, for predictions above an IoU of 10%. In fact, if not for the outlier performance solely on RMH data (2 cases with 0 Dice), they would have a perfect Recall of 1.

Table 7
Mean prospective results stratified by segmentation task.

	Gland	PZ	TZ	Lesions
Dice	0.88 ± 0.01	0.78 ± 0.01	0.87 ± 0.01	0.66 ± 0.06
HD	15.7 ± 1.36	19.19 ± 1.35	12.65 ± 0.88	30.15 ± 10.77
ASSD	0.57 ± 0.05	0.68 ± 0.04	0.79 ± 0.2	11.02 ± 8.17
RAVD	0.15 ± 0.02	0.21 ± 0.04	0.63 ± 0.33	0.09 ± 0.16
Recall	1	1	1	0.86

4. Discussion

In this work, we develop and analyse multi-institution and multi-vendor segmentation models for prostate whole gland, zones and index lesion segmentation. We evaluated these models on a wide array of distinct factors, such as total amount of data, scanner manufacturer variability and institutional provider variability.

The more data the better. From the analysis of the whole gland models, in particular the learning curves, it can be seen that for this task the main factor for good performance is the total amount of data, as the model is not fully saturated even with the entire set of samples. Even when provided with ground truths that contain artefacts and are, sometimes, grossly incorrect, these models still benefit from the additional data.

More variability leads to better generalization. As consistently shown throughout all segmentation tasks, where ProstateNet and ProstateAll outperform single-provider single-manufacturer models in- and out-of distribution, more data variability during training makes the models resistant to domain shift. The fairness analysis further highlights this, as both of these models, and in particular the ProstateAll models, not only consistently provide the best dice scores, but also produce far smaller errors.

Bad segmentation models are good detection models. Despite showing a poor segmentation performance, the ProstateAll index lesion segmentation models show a remarkable detection capability. This was first shown by their overall Recall performance, at both 0.1 and 0.5

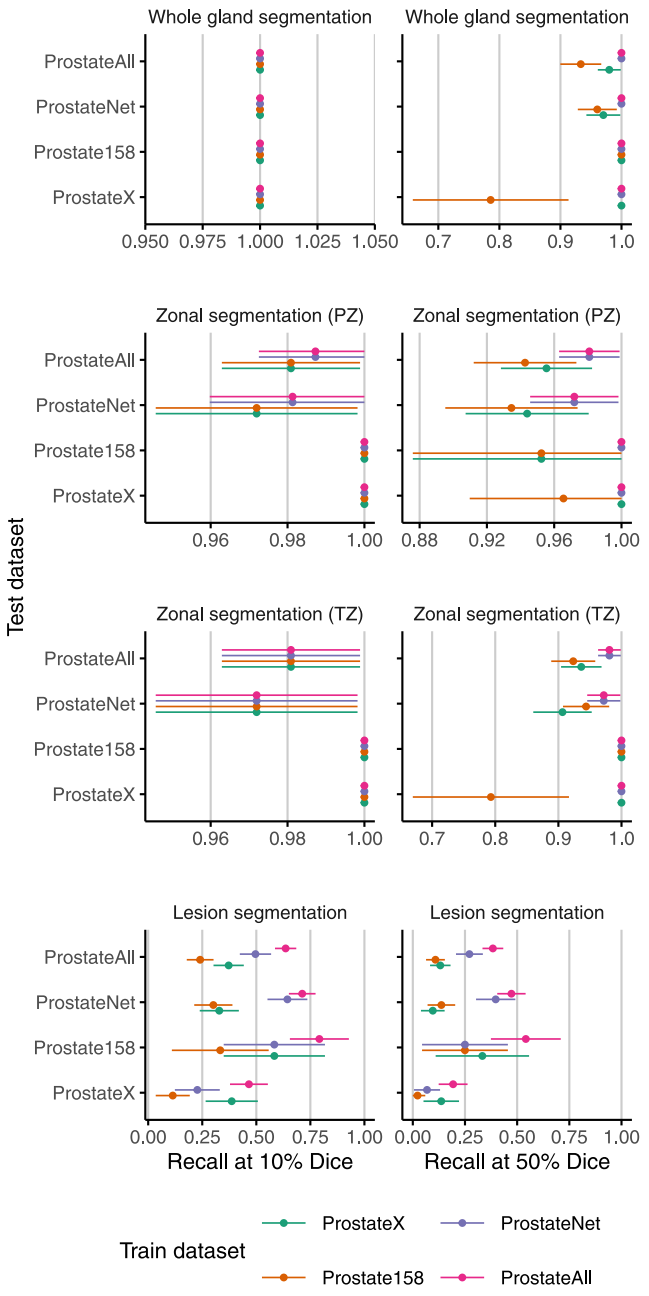

Fig. 5. Test set ROI detection analysis based on Recall at different Dice levels (10% and 50%).

Dice thresholds, when evaluating the segmentation performance, and later by the extracted index lesion candidates and the corresponding heatmaps when focusing solely on detection.

Models show good prospective performance. The results obtained from the prospective validation of the segmentation models show that there is no deterioration in performance, with whole gland and prostate zone models producing results similar to those obtained on retrospective data, and the index lesion segmentation models showing a considerable performance improvement. However, this was done on a very small sample of prospective cases, which raises some concerns about the real applicability of the models.

Limitations and future directions. This work serves the purpose of building new robust models that can be seen as strong domain-invariant baselines for prostate whole gland, zone and index lesion segmentation, as well as lesion detection. While the overall quality of these models was shown, evaluated on both retrospective and prospective data, the prospective cohort was fairly small and did not have a time continuity. It would be interesting to have a prospective cohort obtained throughout a continuous period of time to assess if the models would show a continuous degradation from domain shift. Additionally, following the promising lesion detection results, it would be interesting to further improve these models, and evaluate their capabilities on tumour stage detection [21], as well as differentiate between aggressive and non-aggressive lesions, stratified by ISUP scores. One last possible research direction would be to use semi-supervised learning to mitigate the lack of lesion annotations, and determine how this could improve model performance and generalization.

CRediT authorship contribution statement

Nuno Miguel Rodrigues: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **José Guilherme de Almeida:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Ana Sofia Castro Verde:** Resources, Data curation. **Ana Mascarenhas Gaivão:** Resources, Data curation. **Carlos Bilreiro:** Resources, Data curation. **Inês Santiago:** Resources, Data curation. **Joana Ip:** Resources, Data curation. **Sara Belião:** Resources, Data curation. **Raquel Moreno:** Resources, Data curation. **Celso Matos:** Funding acquisition. **Leonardo Vanneschi:** Writing – review & editing. **Manolis Tsiknakis:** Funding acquisition. **Kostas Marias:** Funding acquisition. **Daniele Regge:** Funding acquisition. **Sara Silva:** Writing – review & editing, Supervision. **ProCAncer-I Consortium:** Funding acquisition. **Nickolas Papanikolaou:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

None Declared

Acknowledgements

This work was partially supported by the Fundação para a Ciência e a Tecnologia, Portugal, through funding of the LASIGE Research Unit refs. UIDB/00408/2020 (https://doi.org/10.54499/UIDB/00408/2020), UIDP/00408/2020 (https://doi.org/10.54499/UIDP/00408/2020) and UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Nuno M. Rodrigues was supported by PhD Grant 2021/05322/BD. All authors except Nuno Rodrigues, Leonardo Vanneschi and Sara Silva, were supported by the European Union H2020: ProCAncer-I project (EU grant 952159).

Appendix A. Supplementary information: ProCAncer-I members

- Manolis Tsiknakis (FORTH - Institute of Computer Science - Computational BioMedicine Lab, Greece)
- Kostas Marias (FORTH - Institute of Computer Science - Computational BioMedicine Lab, Greece)
- Stelios Sfakianakis (FORTH - Institute of Computer Science - Computational BioMedicine Lab, Greece)
- Varvara Kalokyri (FORTH - Institute of Computer Science - Computational BioMedicine Lab, Greece)
- Eleftherios Trivizakis (FORTH - Institute of Computer Science - Computational BioMedicine Lab, Greece)
- Grigorios Kalliatakis (FORTH - Institute of Computer Science - Computational BioMedicine Lab, Greece)

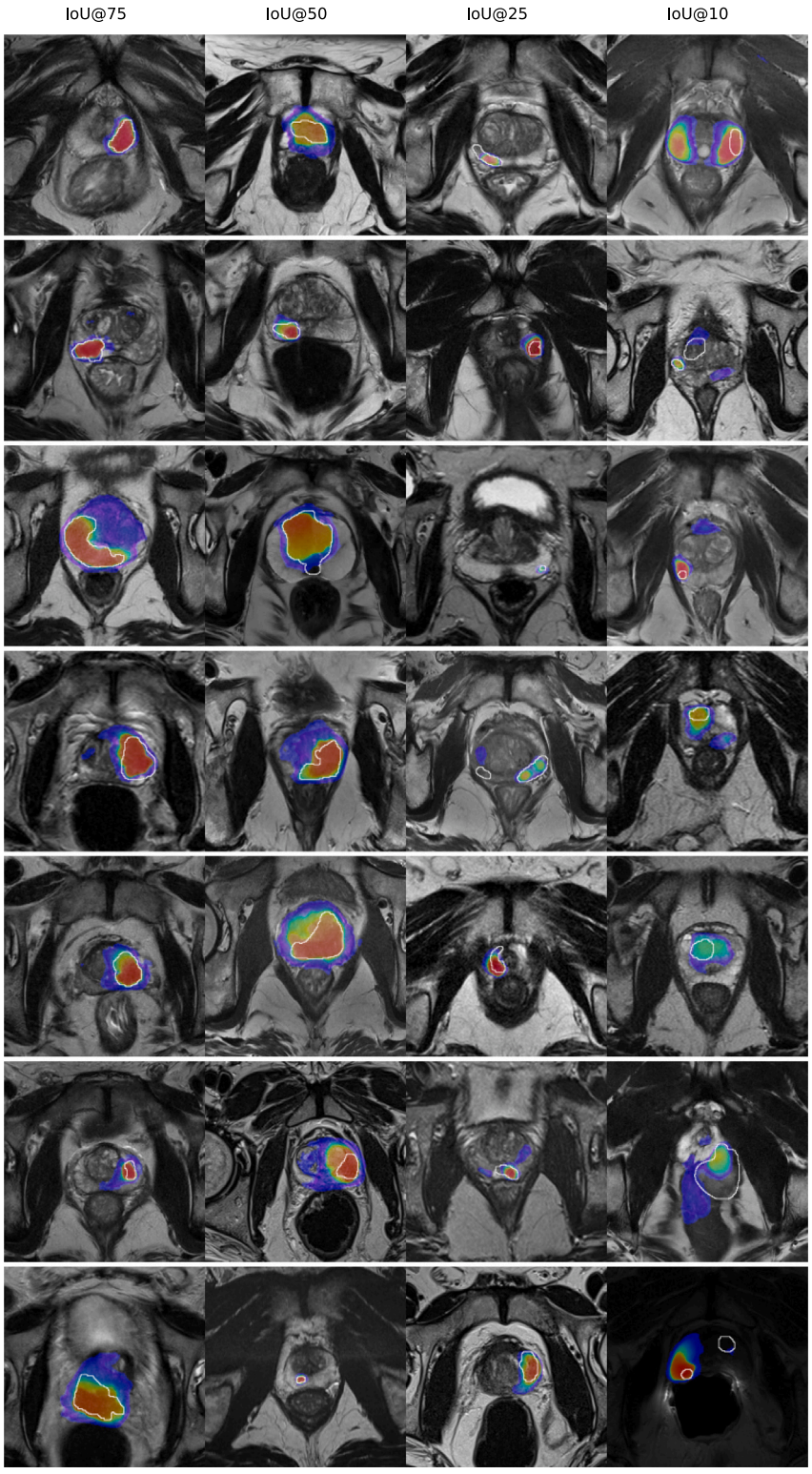

Fig. 6. Index lesion detection probability maps at different IoU thresholds. Colder hues (blue/green) denote zones of lower probability, while hotter hues (red/orange) denote areas of higher probability. The ground truth is marked as the solid white line.

- Avtantil Dimitriadis (FORTH - Institute of Computer Science - Computational BioMedicine Lab, Greece)
- Dimitris Fotiadis (FORTH - Institute of Molecular Biology and Biotechnology (FORTH-IMBB/BR), Greece)
- Nikolaos Tachos (FORTH - Institute of Molecular Biology and Biotechnology (FORTH-IMBB/BR), Greece)
- Eugenia Mylona (FORTH - Institute of Molecular Biology and Biotechnology (FORTH-IMBB/BR), Greece)
- Dimitris Zaridis (FORTH - Institute of Molecular Biology and Biotechnology (FORTH-IMBB/BR), Greece)
- Charalampos Kalantzopoulos (FORTH - Institute of Molecular Biology and Biotechnology (FORTH-IMBB/BR), Greece)
- Nikolaos Papanikolaou (Champalimaud Foundation, Portugal)

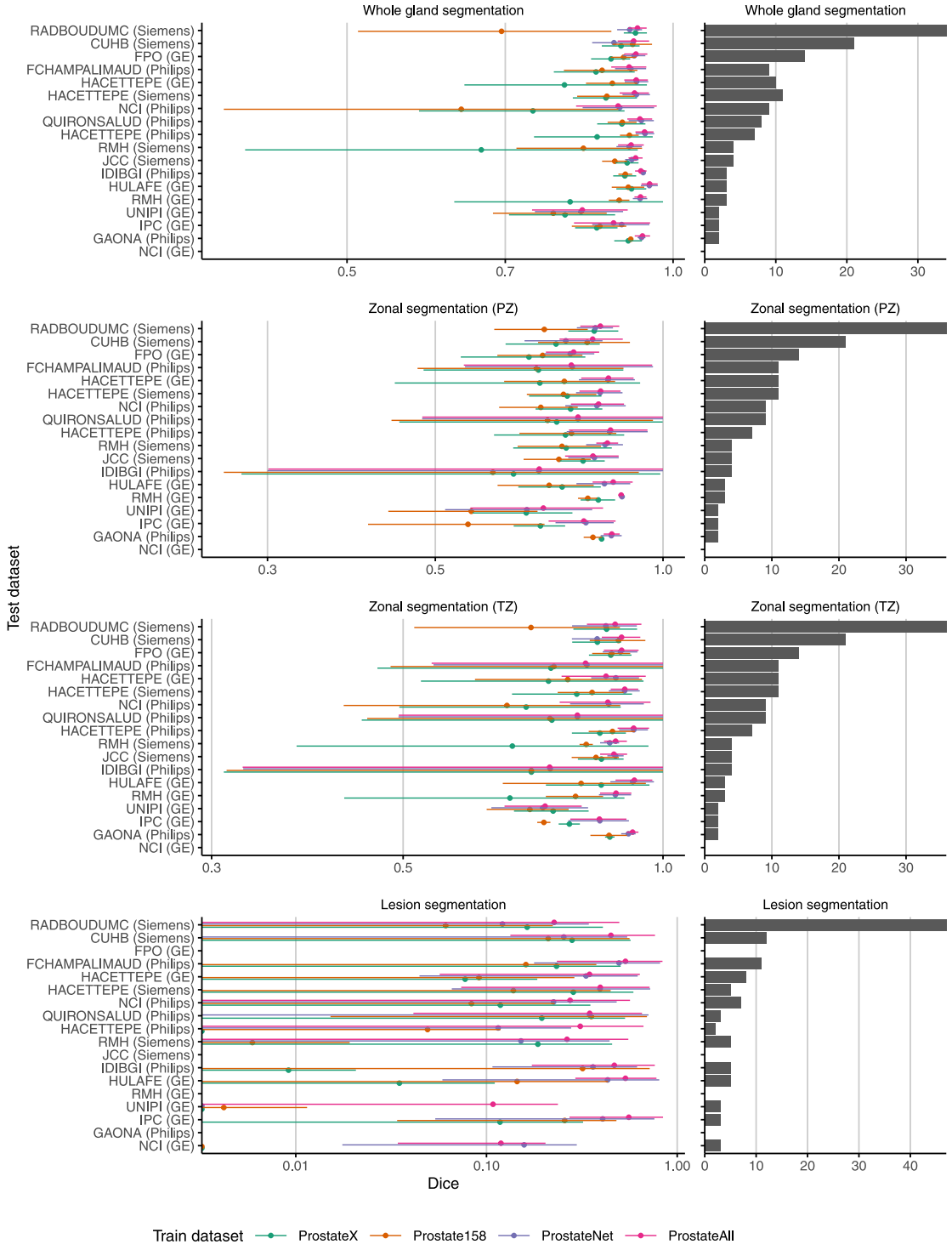

Fig. 7. Fairness analysis showing the Dice scores when testing the different nnUNet models, in- and out-of-distribution, stratified by manufacturer and provider.

- José Guilherme de Almeida (Champalimaud Foundation, Portugal)
- Ana Castro Verde (Champalimaud Foundation, Portugal)
- Ana Carolina Rodrigues (Champalimaud Foundation, Portugal)
- Nuno Rodrigues (Champalimaud Foundation, Portugal)
- Miguel Chambel (Champalimaud Foundation, Portugal)
- Henkjan Huisman (Radboud, Netherlands)
- Maarten de Rooij (Radboud, Netherlands)
- Anindo Saha (Radboud, Netherlands)
- Jasper J. Twilt (Radboud, Netherlands)
- Jurgen Futterer (Radboud, Netherlands)
- Luis Martí-Bonmatí(HULAFE - Biomedical Imaging Research Group, Instituto de Investigación Sanitaria La Fe; Medical Imaging Department, Hospital Universitari i Politècnic La Fe, Spain)
- Leonor Cerdá-Alberich (HULAFE - Biomedical Imaging Research Group. Instituto de Investigación Sanitaria La Fe, Spain)
- Gloria Ribas (HULAFE - Biomedical Imaging Research Group. Instituto de Investigación Sanitaria La Fe, Spain)

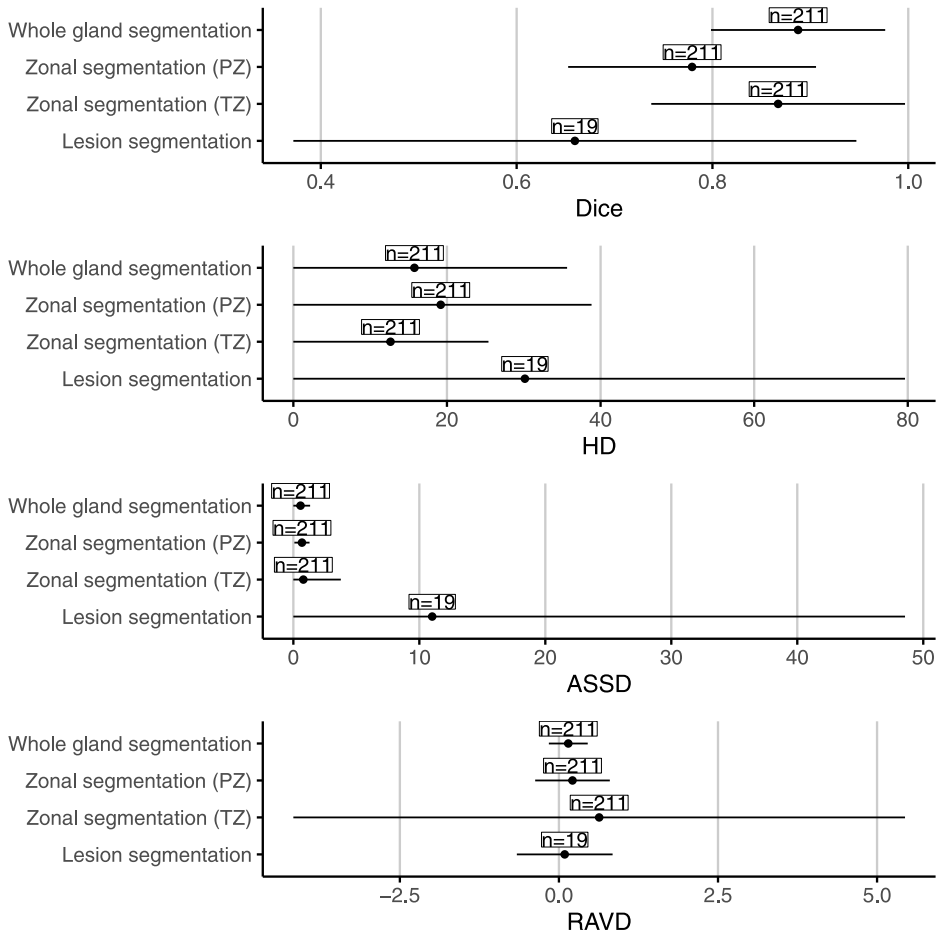

Fig. 8. Distribution of the prospective results stratified by segmentation task.

- Silvia Navarro (HULAFE - Biomedical Imaging Research Group. Instituto de Investigación Sanitaria La Fe, Spain)
- Manuel Marfil (HULAFE - Biomedical Imaging Research Group. Instituto de Investigación Sanitaria La Fe, Spain)
- Emanuele Neri (Academic Radiology, Department of Translational Research, University of Pisa, Italy)
- Giacomo Aringhieri (Academic Radiology, Department of Translational Research, University of Pisa, Italy)
- Lorenzo Tumminello (Academic Radiology, Department of Translational Research, University of Pisa, Italy)
- Vincenzo Mendola (Academic Radiology, Department of Translational Research, University of Pisa, Italy)
- nan (Institut Paoli-Calmettes, France)
- Deniz Akata (Hacettepe - Department of Radiology, Turkey)
- Mustafa Özmen (Hacettepe - Department of Radiology, Turkey)
- Ali Devrim Karaosmanoglu (Hacettepe - Department of Radiology, Turkey)
- Firat Atak (Hacettepe - Department of Radiology, Turkey)
- Musturay Karcaaltincaba (Hacettepe - Department of Radiology, Turkey)
- Joan C. Vilanova (Institute of Biomedical Research of Girona Dr. Josep Trueta (IDIBGI), Department of Radiology (IDI), Girona, Spain)
- Jurgita Usinskiene (National cancer institute, Vilnius, Lithuania)
- Ruta Briediene (National cancer institute, Vilnius, Lithuania)
- Audrius Untanas (National cancer institute, Vilnius, Lithuania)
- Kristina Slidevska (National cancer institute, Vilnius, Lithuania)
- Katsaros Vasilis (General Anti-Cancer and Oncological Hospital of Athens, Greece)
- Georgiou Georgios (General Anti-Cancer and Oncological Hospital of Athens, Greece)
- Dow-Mu Koh (Radiology & AI Research Hub, The Royal Marsden NHS Foundation Trust, London; Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, UK, UK)
- Robby Emsley (Radiology & AI Research Hub, The Royal Marsden NHS Foundation Trust, London; Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, UK, UK)
- Sharon Vit (Radiology & AI Research Hub, The Royal Marsden NHS Foundation Trust, London; Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, UK, UK)
- Ana Ribeiro (Radiology & AI Research Hub, The Royal Marsden NHS Foundation Trust, London; Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, UK, UK)
- Simon Doran (Radiology & AI Research Hub, The Royal Marsden NHS Foundation Trust, London; Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, UK, UK)
- Tiaan Jacobs (Radiology & AI Research Hub, The Royal Marsden NHS Foundation Trust, London; Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, UK, UK)
- Gracián García-Martí(Quirónsalud Hospital/CIBERSAM, Valencia, Spain)
- Daniele Regge (Candiolo Cancer Institute, FPO-IRCCS, Str. Prov.le 142 km 3.95, 10060 Candiolo, Turin, Italy)
- Valentina Giannini (Candiolo Cancer Institute, FPO-IRCCS, Str. Prov.le 142 km 3.95, 10060 Candiolo, Turin, Italy)
- Simone Mazzetti (Candiolo Cancer Institute, FPO-IRCCS, Str. Prov.le 142 km 3.95, 10060 Candiolo, Turin, Italy)
- Giovanni Cappello (Candiolo Cancer Institute, FPO-IRCCS, Str. Prov.le 142 km 3.95, 10060 Candiolo, Turin, Italy)

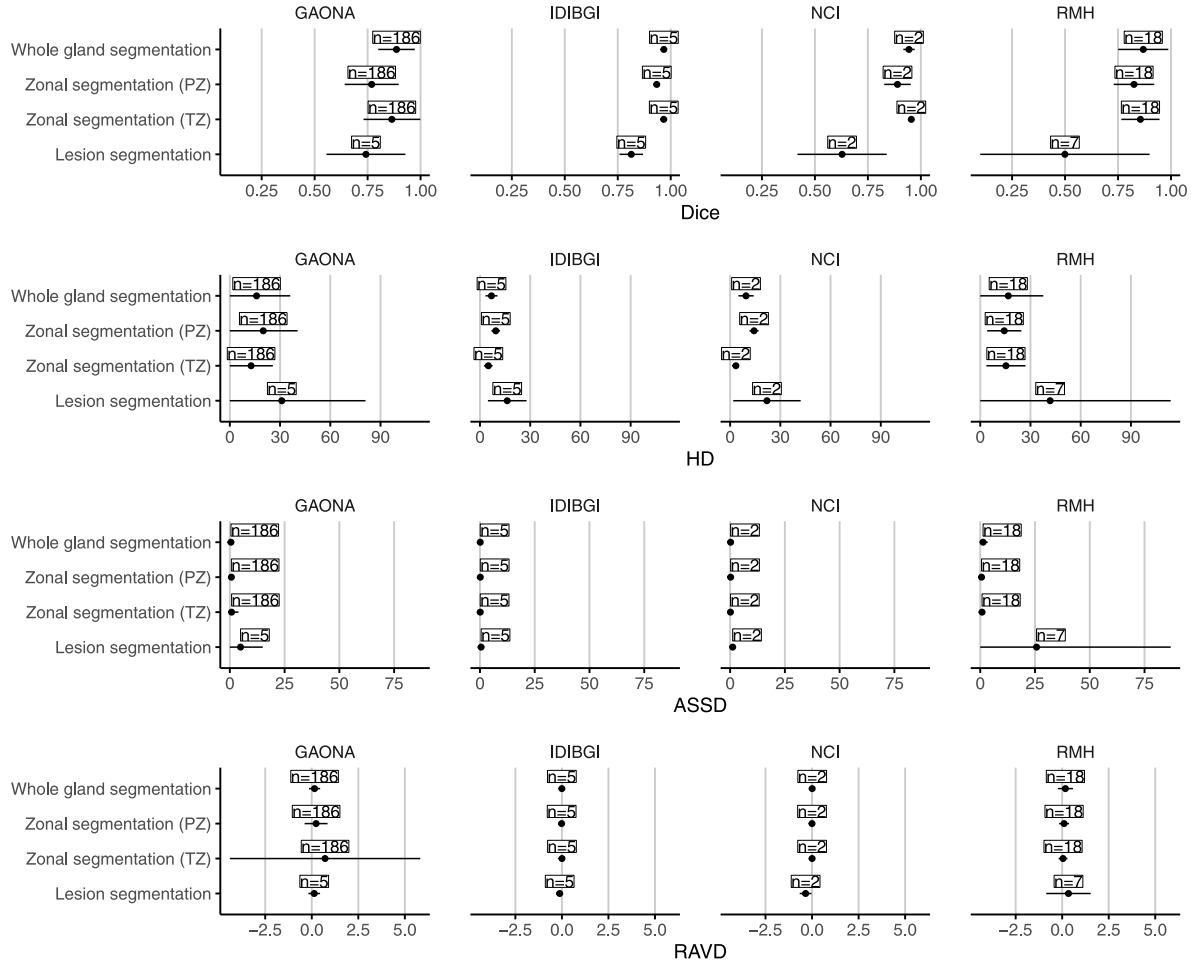

Fig. 9. Distribution of the prospective results stratified by data provider.

- Giovanni Maimone (Candiolo Cancer Institute, FPO-IRCCS, Str. Prov.le 142 km 3.95, 10060 Candiolo, Turin, Italy)
- Valentina Napolitano (Candiolo Cancer Institute, FPO-IRCCS, Str. Prov.le 142 km 3.95, 10060 Candiolo, Turin, Italy)
- Sara Colantonio (Institute of Information Science and Technologies of the National Reserch Council of Italy, Italy)
- Maria Antonietta Pascali (Institute of Information Science and Technologies of the National Reserch Council of Italy, Italy)
- Eva Pachetti (Institute of Information Science and Technologies of the National Reserch Council of Italy, Italy)
- Giulio del Corso (Institute of Information Science and Technologies of the National Reserch Council of Italy, Italy)
- Danila Germanese (Institute of Information Science and Technologies of the National Reserch Council of Italy, Italy)
- Andrea Berti (Institute of Information Science and Technologies of the National Reserch Council of Italy, Italy)
- Gianluca Carloni (Institute of Information Science and Technologies of the National Reserch Council of Italy, Italy)
- Jayashree Kalpathy-Cramer (Mass General Hospital, Boston MA, USA)
- Christopher Bridge (Mass General Hospital, Boston MA, USA)
- Joao Correia (B3D, UK)
- Walter Hernandez (B3D, UK)
- Zoi Giavri (Advantis, Greece)
- Christos Pollalis (Advantis, Greece)
- Dimitrios Agraniotis (Advantis, Greece)
- Ana Jiménez Pastor (Quibim, S.L., Valencia, Spain)
- Jose Munuera Mora (Quibim, S.L., Valencia, Spain)
- Clara Saillant (Univie, Austria)
- Theresa Henne (Univie, Austria)
- Rodessa Marquez (Univie, Austria)

Appendix B. Supplementary figures

See Figs. B.1–B.4.

Appendix C. Qualitative analysis

See Figs. C.1–C.3.

Appendix D. ProstateNet vendor-specific analysis

In order to assess the individual contribution and overall performance of each vendor, we trained additional models on subsets of data stratified by manufacturer (vendor) — GE, Philips and Siemens. We perform both a CV and hold-out test set analysis of these models and compare the results between manufacturer and to those obtained by the master models described previously.

During CV results, Table D.1 shows two very distinct scenarios: For whole gland and zones segmentation, we can see that the performance is very similar across all metrics, between all manufacturers – with the exception of GE for the peripheral zone – and that these are also similar to the ones obtained by the master models; On the other hand, the lesion segmentation models show large discrepancies between manufacturer, with the Siemens model clearly outperforming the other two,

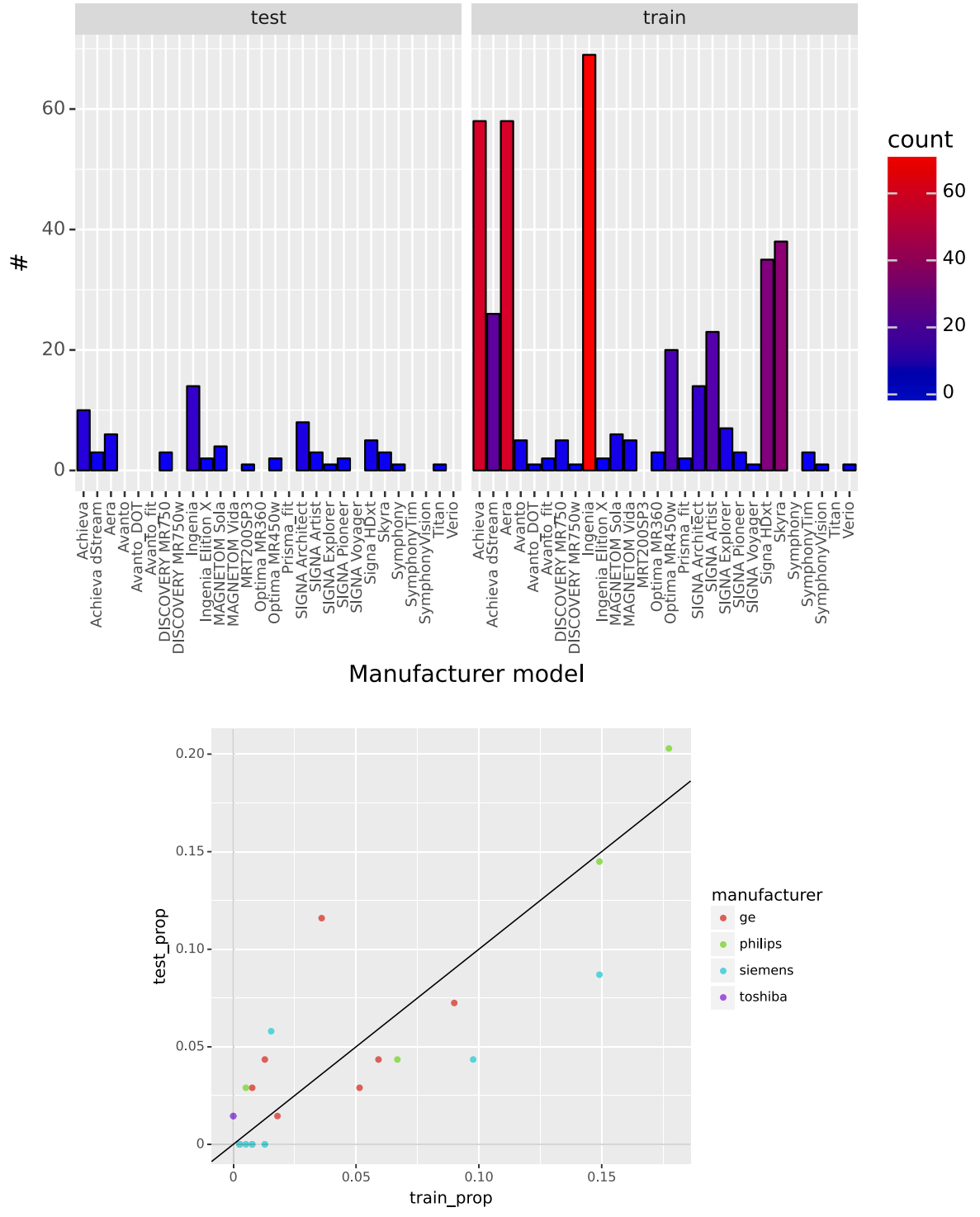

Fig. B.1. Top: Scanner models distribution per data partition. Bottom: Proportion comparison per scanner model per data partition.

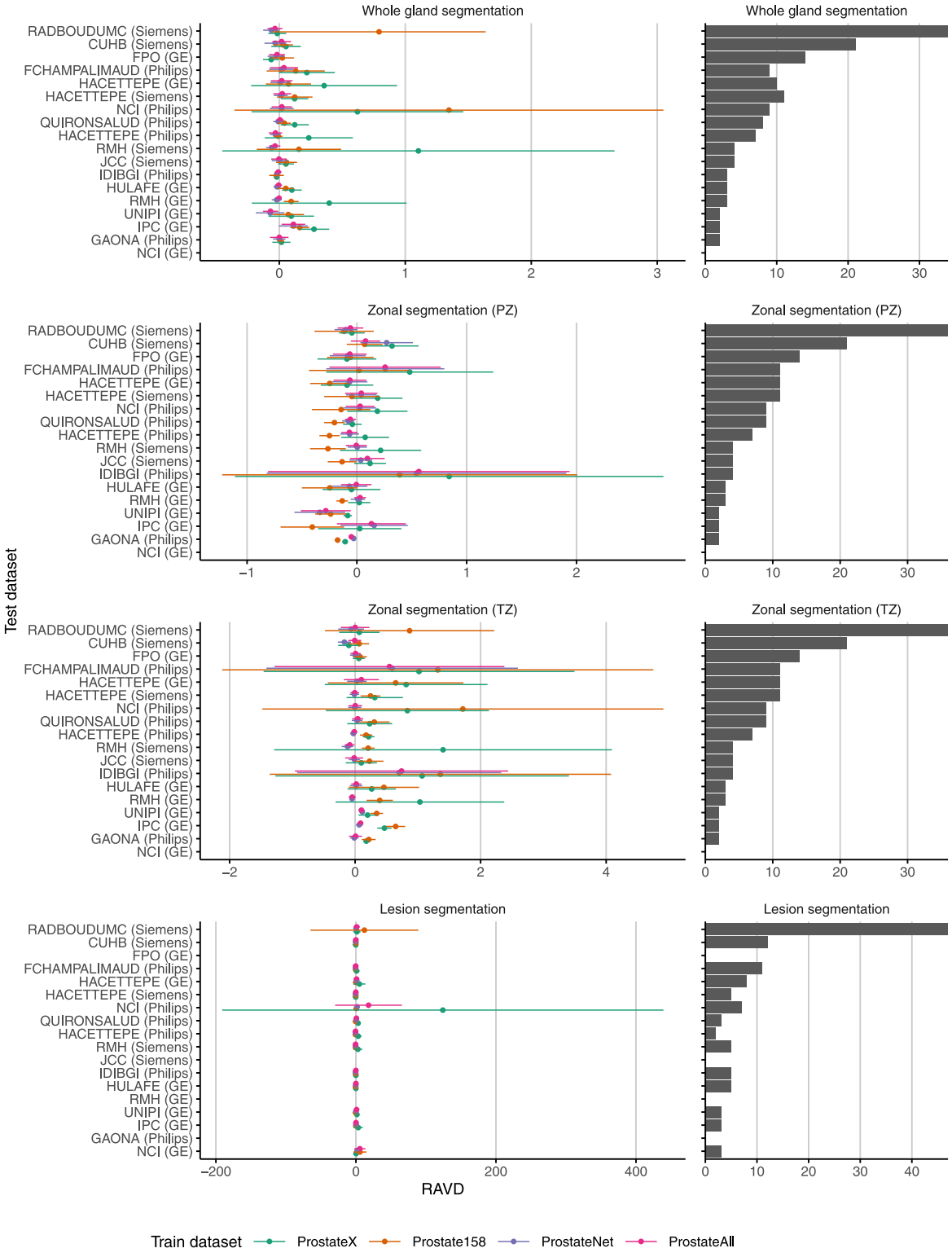

Fig. B.2. Fairness analysis showing how much the predicted volumes differ from the ground truths when testing the different nnUNet models, in- and out- of distribution, stratified by manufacturer and provider.

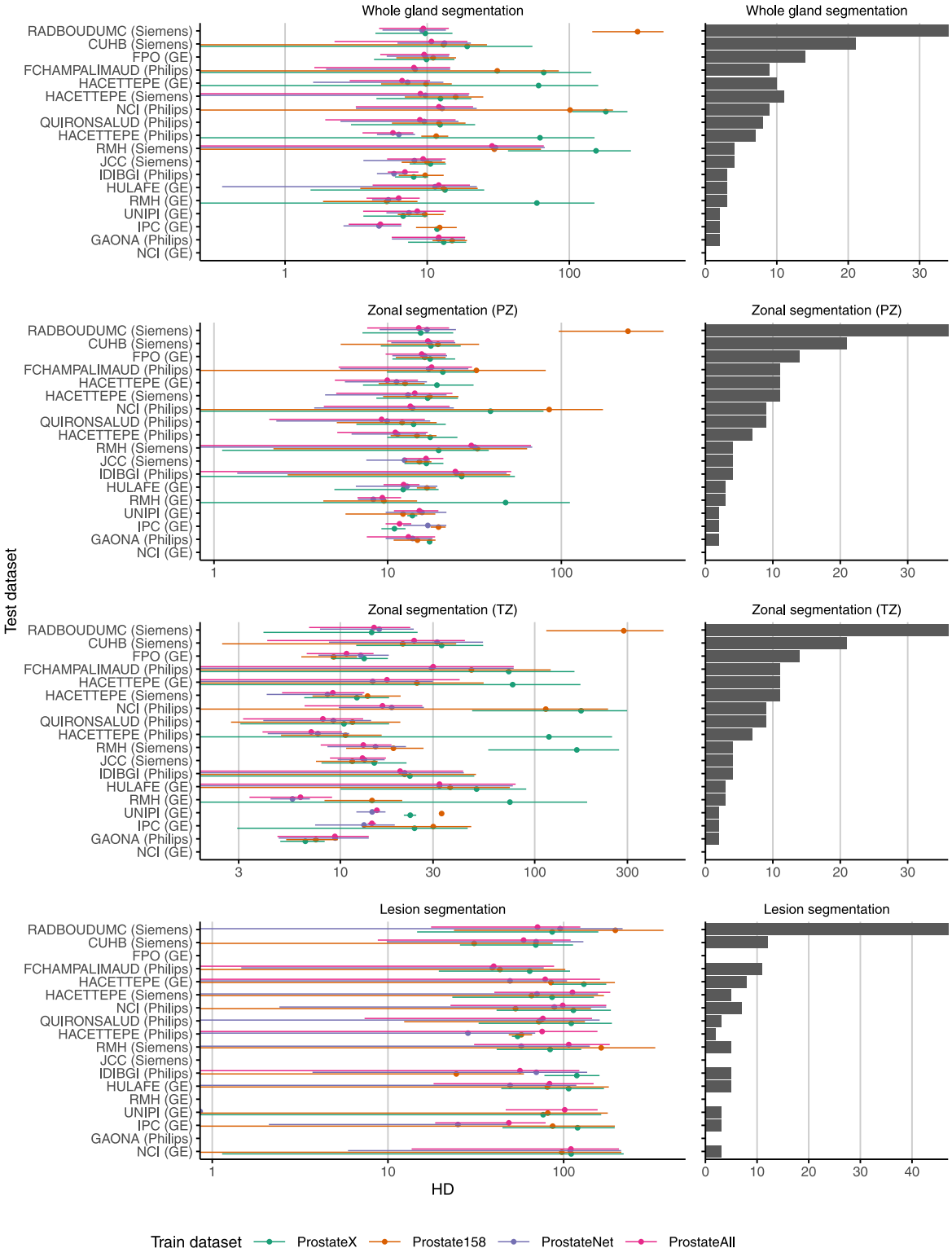

Fig. B.3. Fairness analysis showing the maximum errors of the predictions when testing the different nnUNet models, in- and out- of distribution, stratified by manufacturer and provider.

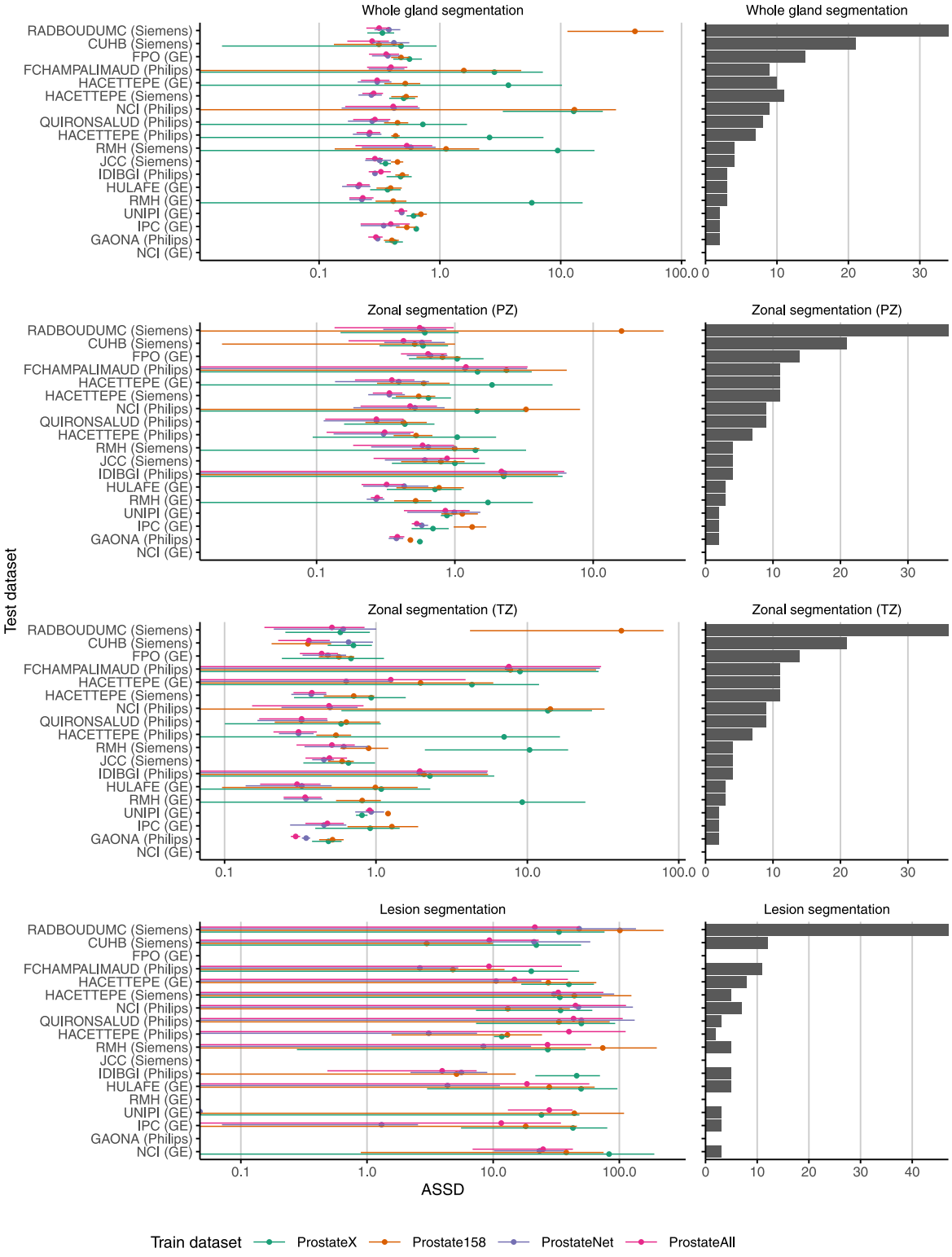

Fig. B.4. Fairness analysis showing how much the surface of the predicted volumes differ from that of the ground truths when testing the different nnUNet models, in- and out-of distribution, stratified by manufacturer and provider.

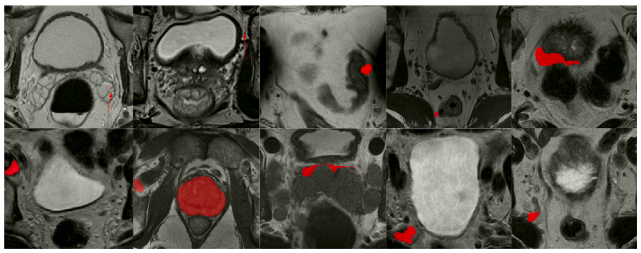

Fig. C.1. Sample of ground truth errors detected during the largest connect component analysis, showing random artefacts and incorrect gland masks.

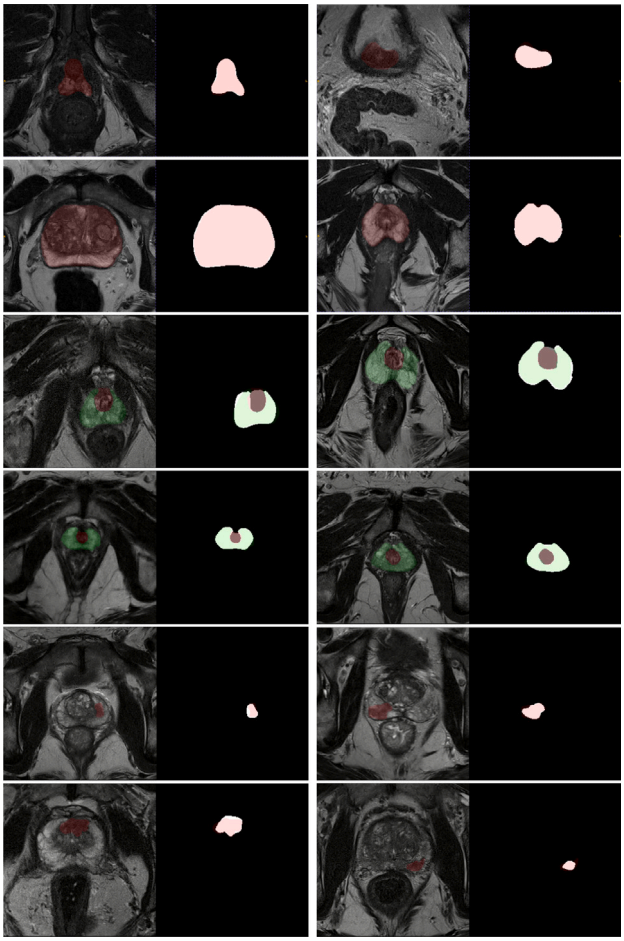

Fig. C.2. Set of examples showcasing the quality of the developed segmentation models, for whole gland, zones and lesion segmentation, on various areas of the prostate. Predicted segmentations are white for the whole gland and lesions, and white+brown for zones, while ground truth masks are red for whole gland and lesions, and green+red for zones.

and being the only one producing results similar to those obtained by the master models.

Analysing the hold-out test results, for the whole gland segmentation models (Table D.2) we can see that on all manufacturer models show the same generalization degree to Siemens and Philips data, whilst GE shows a 2/3% improvement over the others on GE data. It can also be noted that these results match those of the ProstateNet master models. For the zone segmentation (Table D.3), the key finding is how well all models seem to generalize to Philips data. Both Siemens and GE models generalize better on Philips data than on their own vendor-specific data. Regarding the lesion segmentation models (Table D.4) there are several interesting aspects. Similar to what was observed during CV, Siemens models are the ones that overall generalize the

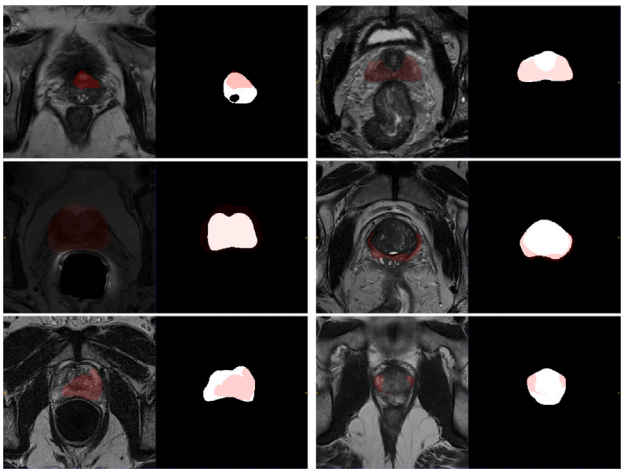

Fig. C.3. Examples where the predicted segmentation (white) was evaluated with a low dice score due to incorrect ground truth masks (red) for whole gland segmentation, evaluated by an expert radiologist.

better, however, it can also be noted that the data on which all models better generalize is that of Philips. Even models such as the GE, which show a very poor in-distribution performance, show higher results on Philips data. When evaluating the lesion detection capabilities (Recall) of each models, we can see that the Siemens model produces very good results, in particular when tested in-distribution (on Siemens data). When comparing these results to those produced by the master models, we can observe that, overall, only the Siemens model – and also the Philips model trained on Philips data – come close regarding Dice and Recall performance, whilst both the GE model and GE data wield very poor results.

Appendix E. Supplementary methods

E.1. Training the U-Net and U-Net + D.S. models

To train the U-Net and U-Net + D.S. models, we used a stochastic gradient descent optimizer with 0.99 momentum with a maximum learning rate of 0.01, weight decay of 0.005 [39] and cosine decay with a minimum learning rate of 0. The encoder is composed of five regular convolutional layers with kernel size $[3, 3, 3]$ and increasing depths $(32, 64, 128, 256, 320)$ intercalated with 2x2 max-pooling operations with strides $[2, 2, 1]$, $[2, 2, 1]$, $[2, 2, 1]$, $[2, 2, 2]$, $[2, 2, 2]$, similar to what is used in nnUNet [31], while the decoder replicates the encoder but replaces the max-pooling operations with transpose convolutions. In any case, Swish activation functions [40] and instance normalization operations are used after each convolution, with a dropout [41] probability of 0.1. The anysotropic max-pooling allows for the preservation of a minimal resolution of 4 in the slice dimension. Using a batch-size of 2, we sampled $256 \times 256 \times 16$ patches from the image such that patches with and without positive samples (i.e. voxels belonging to the prostate gland) are sampled equally. A combo loss [42] — the addition of the generalized Dice [43] and weighted focal losses [44] with alpha=0.5 — was used to train both U-Net and U-Net + D.S. models.

Deep supervision [30] for U-Net + D.S. was implemented using an additional classifier after each layer decoder layer that classifies voxels at a decreased resolution. To calculate the loss for each deep supervision output, the ground truth was downsampled to match the resolution at each decoder layer and the loss was calculated. All losses (for the original resolution and for the deep supervision) were combined using a weighted average where the weight is parameterized as $1/2^{ds-1}$, where ds is the downsampling level. For instance, for the full resolution, this weight evaluates to 1, whereas for the lowest resolution (ds=4) this evaluates to 1/8. During U-Net and U-Net + D.S. training, we augment

Table D.1
nnUNet CV results for all segmentation tasks, stratified by manufacturer. For each dataset, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviations, are presented.

	Dice	HD	ASSD	RAVD	Recall	
	Gland					
Siemens	0.9 ± 0.01	12.37 ± 1.49	0.04 ± 0.02	0.52 ± 0.11	1.0 ± 0.0	
Philips	0.9 ± 0.01	12.43 ± 1.13	0.04 ± 0.02	0.47 ± 0.07	0.99 ± 0.01	
GE	0.91 ± 0.01	11.23 ± 1.33	0.21 ± 0.21	0.42 ± 0.03	1.0 ± 0.0	
	Zones					
Siemens	0.8 ± 0.01	13.99 ± 0.77	0.06 ± 0.03	0.52 ± 0.03	1.0 ± 0.0	
	0.88 ± 0.01	15.45 ± 1.42	0.02 ± 0.02	0.55 ± 0.04	1.0 ± 0.0	
Philips	0.82 ± 0.01	15.15 ± 1.0	0.07 ± 0.03	0.56 ± 0.07	0.99 ± 0.01	
	0.87 ± 0.01	13.85 ± 1.05	0.19 ± 0.13	0.9 ± 0.36	0.98 ± 0.01	
GE	0.78 ± 0.01	17.08 ± 1.94	0.22 ± 0.19	0.68 ± 0.08	1.0 ± 0.0	PZ
	0.88 ± 0.01	13.47 ± 0.73	0.24 ± 0.21	0.54 ± 0.05	1.0 ± 0.0	TZ
	Lesions					
Siemens	0.36 ± 0.03	74.32 ± 6.17	0.08 ± 0.11	15.11 ± 2.4	0.7 ± 0.04	
Philips	0.24 ± 0.02	64.66 ± 5.32	0.82 ± 0.37	22.62 ± 3.37	0.5 ± 0.04	
GE	0.29 ± 0.03	74.39 ± 6.7	1.17 ± 0.45	17.92 ± 2.67	0.58 ± 0.05	

Table D.2
nnUNet whole gland segmentation hold-out test set results. For each pairwise evaluation, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviations, are presented.

		Tested on			
		Siemens	Philips	GE	
Trained on	Siemens	0.92 ± 0.03	0.91 ± 0.03	0.88 ± 0.09	Dice
		8.81 ± 5.8	21.59 ± 43.48	17.87 ± 45.61	HD
		0.02 ± 0.09	0.01 ± 0.07	0.01 ± 0.12	ASSD
		0.32 ± 0.11	0.81 ± 2.38	0.62 ± 0.64	RAVD
		1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	Recall
	Philips	0.92 ± 0.03	0.92 ± 0.02	0.89 ± 0.1	
		8.28 ± 5.38	14.12 ± 24.47	18.21 ± 45.59	
		0.01 ± 0.08	0.01 ± 0.07	−0.01 ± 0.11	
		0.33 ± 0.11	0.36 ± 0.31	0.6 ± 0.65	
		1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	
	GE	0.92 ± 0.03	0.92 ± 0.03	0.91 ± 0.09	
		9.08 ± 5.56	17.39 ± 27.73	17.12 ± 46.21	
		−0.0 ± 0.08	−0.01 ± 0.07	0.0 ± 0.09	
		0.33 ± 0.1	0.45 ± 0.49	0.52 ± 0.65	
		1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	

data in real time to increase the variability of observed data by our model using MONAI [45], with each transform having a 0.1 probability of being applied. Particularly we use:

- Random contrast adjustments (gamma multiplier between 0.5 and 1.5)
- Random intensity standard deviation shift (multiplier between 0.9 and 1.1)
- Random intensity shift (multiplier between 0.9 and 1.1)
- Random addition of Rician noise (with a standard deviation of 0.02)
- Random addition of Gibbs noise (alpha between 0.0 and 0.6 and standard deviation of 0.25)
- Random affine transform (translation range in voxels $[(0,4),(0,4),(0,1)]$, rotation range in radians $[(0,\pi/16),(0,\pi/16),(0,\pi/16)]$
- Random shearing (shearing factor between 0.9 and 1.1 for all axes)
- Gaussian blurring (sigma between 0.25 and 1.5)

E.2. Training the UNETR and Swin-UNETR models

UNETR [35] and Swin-UNETR [36] are segmentation models which make use of transformers [46] in the feature encoder layer. Both of these were trained using the exact same hyperparameters as those used to train the standard U-Net and U-Net + D.S. models described in Appendix E.1 with few exceptions — the use of a linear learning rate warm-up of 25 epochs (10% of the total number of training epochs) followed by a cosine decay for the remainder of the training, a maximum learning rate of 0.001, and a weight decay of 0.05 and 0.005 for Swin-UNETR and UNETR, respectively. Additionally, a batch size of 2 was used for both UNETR and Swin-UNETR, and each sample was padded (if necessary) to a minimum size of 256 × 256 × 32 voxels and the randomly cropped to have the same size. The reason for the difference in padding/crop sizes between the standard U-Net/nnUNet and UNETR and Swin-UNETR is that the latter require a pre-specified number of down-scaling operations during the encoding.

The UNETR architecture was implemented using an in-house library[4] with a patch size of 16 × 16 × 16 and composed by 12 vision transformer blocks, skip connections at blocks 3, 6 and 9 and 4 encoder stages with depths of 16, 32, 64 and 128 features; each block had an embedding size of 768 with a multilayer perceptron with 3072 hidden units. The Swin-UNETR architecture is available as a part of the MONAI package for Python [45], which is identical to the one used in the original publication [36].

E.3. UNETR and Swin-UNETR results

To further solidify the motivation for using nnUNet and regular UNet models as opposed to transformer-based segmentation architectures, we trained a UNETR and Swin-UNETR models for whole gland segmentation and compared the results to those of nnUNet and both UNet variations. Comparing the CV results (Table E.1) it is possible to observe that the results are considerably worse than those of nnUNet and both UNet variants, being statistically significantly worse in all datasets when compared with nnUNet (Kruskall-wallis p-test < 0.01). When analysing the hold-out test set results (Table E.2), not only can it be observed that the results are worse, but that there is clear domain drift, in particular between ProstateX and ProstateNet models.

Considering these underwhelming results and the cost/carbon impact of training these large models, which take up to 14 h per fold on 2x A6000 GPUs, we concluded that training these models for the remaining tasks was neither relevant nor beneficial.

[4] Available at https://github.com/CCIG-Champalimaud/adell-mri under `adell_mri/modules/segmentation/unetr.py`.

Table D.3
nnUNet zones segmentation hold-out test set results. For each pairwise evaluation, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviations, are presented.

		Tested on						
		Siemens		Philips		GE		
Trained on	Siemens	0.79 ± 0.1	0.84 ± 0.14	0.8 ± 0.08	0.89 ± 0.04	0.73 ± 0.18	0.86 ± 0.09	Dice
		16.55 ± 17.51	16.18 ± 17.64	21.32 ± 24.26	15.76 ± 12.46	14.56 ± 6.55	11.17 ± 6.01	HD
		0.02 ± 0.17	0.2 ± 1.02	−0.02 ± 0.16	0.02 ± 0.12	0.11 ± 0.37	0.05 ± 0.17	ASSD
		0.53 ± 0.26	1.29 ± 3.46	0.61 ± 0.41	0.5 ± 0.2	0.75 ± 0.57	0.53 ± 0.22	RAVD
		1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	Recall
	Philips	0.79 ± 0.09	0.85 ± 0.08	0.83 ± 0.07	0.91 ± 0.04	0.75 ± 0.16	0.87 ± 0.09	
		16.25 ± 15.99	12.83 ± 7.5	17.75 ± 23.77	13.2 ± 13.32	14.48 ± 7.54	11.5 ± 5.67	
		0.06 ± 0.16	−0.0 ± 0.22	0.02 ± 0.15	−0.0 ± 0.1	0.11 ± 0.27	0.03 ± 0.16	
		0.5 ± 0.21	0.61 ± 0.24	0.5 ± 0.34	0.4 ± 0.2	0.75 ± 0.63	0.48 ± 0.22	
		1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	
	GE	0.79 ± 0.08	0.85 ± 0.08	0.81 ± 0.06	0.89 ± 0.04	0.78 ± 0.15	0.87 ± 0.1	
		16.43 ± 16.78	12.87 ± 7.03	20.61 ± 24.5	15.02 ± 12.83	12.45 ± 5.29	9.81 ± 5.18	
		0.06 ± 0.18	−0.02 ± 0.23	−0.01 ± 0.15	0.01 ± 0.12	0.01 ± 0.18	0.1 ± 0.17	
		0.53 ± 0.23	0.6 ± 0.26	0.57 ± 0.36	0.47 ± 0.21	0.57 ± 0.41	0.46 ± 0.24	
		1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	
		PZ	TZ					

Table D.4
nnUNet lesion segmentation hold-out test set results. For each pairwise evaluation, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviations, are presented.

		Tested on			
		Siemens	Philips	GE	
Trained on	Siemens	0.33 ± 0.28	0.33 ± 0.31	0.19 ± 0.26	Dice
		40.84 ± 50.38	79.66 ± 69.08	63.53 ± 71.39	HD
		−0.3 ± 0.42	−0.14 ± 1.28	0.21 ± 2.17	ASSD
		11.23 ± 32.49	28.32 ± 44.85	20.31 ± 34.75	RAVD
		0.7 ± 0.1	0.63 ± 0.09	0.4 ± 0.11	Recall
	Philips	0.15 ± 0.22	0.34 ± 0.33	0.23 ± 0.3	
		34.68 ± 51.57	47.94 ± 51.18	26.72 ± 49.08	
		−0.37 ± 0.4	−0.04 ± 1.15	0.05 ± 0.92	
		5.78 ± 10.95	12.45 ± 27.82	3.42 ± 8.13	
		0.4 ± 0.11	0.56 ± 0.1	0.4 ± 0.11	
	GE	0.22 ± 0.29	0.29 ± 0.31	0.21 ± 0.28	
		45.14 ± 65.89	49.02 ± 53.59	51.54 ± 59.9	
		−0.31 ± 0.36	0.09 ± 1.97	−0.26 ± 0.71	
		8.7 ± 15.1	12.53 ± 31.32	7.56 ± 9.24	
		0.4 ± 0.11	0.56 ± 0.1	0.45 ± 0.11	

Table E.1
Whole gland CV results for the transformer models. For each dataset the average Dice and associated standard deviation are presented.

	ProstateX	Prostate158	ProstateNet	ProstateAll
UNETR	0.84 ± 0.02	0.77 ± 0.03	0.8 ± 0.01	0.84 ± 0.01
Swin-UNETR	0.89 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.89 ± 0.0

E.4. Validation metrics

The **Dice Similarity Coefficient** (DSC) is a widely used metric for segmentation tasks, and it measures the spatial overlap between the voxels of the ground truth and predicted masks. The produces score is in the range of $[0, 1]$, where one indicates a perfect segmentation.

The **Average Symmetric Surface Distance** (ASSD) measures, in millimetres, the difference between the surface voxels of the predicted mask and the ground truth mask. For each surface voxel of both images, the Euclidean distance to the closest surface voxel of the opposite image is calculated using the approximate nearest neighbour technique. All measurements are averaged, with the final score indicating the average distance, where a value of zero indicates a perfect segmentation. In essence, the ASSD provides information about the spatial accuracy of the segmentation (how closely the prediction boundary matches the ground truth boundary).

The **Hausdorff Distance** (HD), also known as Maximum Symmetric Surface Distance, measures, in millimetres, the maximum difference between the surface voxels of the predicted mask and the ground truth mask. The Euclidean distance is measured in the same way as the ASSD, however, the final score will be the maximum distance, or error, between both masks, where a value of zero indicates a perfect segmentation.

The **Relative Absolute Volume Difference** (RAVD) measures, in millimetres, the absolute size difference between the volume of the predicted mask and the ground truth mask. RAVD can be either negative or positive, with negative values denoting smaller and positive values a larger predicted volume compared to the ground truth. Values close to zero indicate both volumes are similar.

Appendix F. Results

F.1. Qualitative analysis

To better understand failure cases in whole prostate gland segmentation – here defined as cases where the Dice score was inferior to 90% – we individually inspected each case that fit this criterion in ProstateNet with the assistance of a radiologist with 6.5 years of experience (RM). Interestingly, the outcome of this analysis is not associated with the failure of the model — rather, it is associated with low quality labels as shown in Figs. C.1 and C.3. Particularly, this is associated with cases where labels were automatically generated using the ProCAncer-I tool or when the corrections provided by clinicians contained mistakes. Additionally, through the analysis of large (≈> 20) Hausdorf errors, an issue became apparent — some of the errors stemmed from the existence of more than one connected component (given that the prostate gland is a single continuous object in 3 dimensions, there can be no more than one component corresponding to the prostate gland). To understand this quantitatively, we isolated the largest connected component for all masks and calculated the IoU score between the largest connected component and the totality of the ground truth (if there is no more than one connected component, the IoU score should be 100%). As shown in Table F.1, approximately 1% of ground truths have a large spurious object not belonging to the prostate gland, while most detected abnormalities (74%) are relatively small. In other words, there are cases where the calculated IoU score will be relatively worse than expected due to the quality of the labels as shown by our visual inspection and annotation. Taking the aforementioned aspects into account, it becomes evident that this approach for prostate gland segmentation – training a nnUNet on the ProstateAll dataset – is of high value and can be safely deployed across several different centres.

Table E.2
Whole gland segmentation hold-out test set results for both transformer-based segmentation models. For each pairwise evaluation, the average Dice and associated standard error are presented.

		Tested on				
		ProstateX	Prostate158	ProstateNet	ProstateAll	
Trained on	ProstateX	0.85 ± 0.06 0.89 ± 0.03	0.66 ± 0.15 0.83 ± 0.05	0.45 ± 0.24 0.67 ± 0.18	0.56 ± 0.26 0.73 ± 0.18	UNETR Swin-UNETR
	Prostate158	0.64 ± 0.14 0.44 ± 0.17	0.79 ± 0.07 0.86 ± 0.05	0.54 ± 0.22 0.67 ± 0.18	0.59 ± 0.21 0.65 ± 0.20	
	ProstateNet	0.63 ± 0.16 0.35 ± 0.19	0.78 ± 0.06 0.85 ± 0.05	0.82 ± 0.09 0.89 ± 0.05	0.78 ± 0.13 0.78 ± 0.23	
	ProstateAll	0.87 ± 0.06 0.9 ± 0.04	0.83 ± 0.05 0.88 ± 0.04	0.84 ± 0.09 0.89 ± 0.05	0.85 ± 0.08 0.89 ± 0.05	

Table F.1
Number of ground truths for different IoU scores between the largest connected component and the entire ground truth.

IoU interval	[0.0,0.9[	[0.9,0.99[	[0.99,1.0[	1.0	Total
No. of cases	7	21	82	524	637

Conclusively, the segmentations inferred by our model were of considerable quality (Fig. C.2) and the failure cases were typically associated with poor annotation.

References

[1] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer statistics, 2022, CA Cancer J. Clin. 72 (1) (2022) 7–33.

[2] A. Rodrigues, J. Santinha, B. Galvão, C. Matos, F.M. Couto, N. Papanikolaou, Prediction of prostate cancer disease aggressiveness using bi-parametric MRI radiomics, Cancers 13 (23) (2021) http://dx.doi.org/10.3390/cancers13236065, URL: https://www.mdpi.com/2072-6694/13/23/6065.

[3] A. Rodrigues, N. Rodrigues, J. Santinha, M.V. Lisitskaya, A. Uysal, C. Matos, I. Domingues, N. Papanikolaou, Value of handcrafted and deep radiomic features towards training robust machine learning classifiers for prediction of prostate cancer disease aggressiveness, Sci. Rep. 13 (1) (2023) http://dx.doi.org/10.1038/s41598-023-33339-0.

[4] E. Pachetti, S. Colantonio, 3D-vision-transformer stacking ensemble for assessing prostate cancer aggressiveness from T2w images, Bioengineering 10 (9) (2023) http://dx.doi.org/10.3390/bioengineering10091015, URL: https://www.mdpi.com/2306-5354/10/9/1015.

[5] F. Midiri, F. Vernuccio, P. Purpura, P. Alongi, T.V. Bartolotta, Multiparametric MRI and Radiomics in Prostate Cancer: A Review of the Current Literature, Diagnostics (Basel) 11 (10) (2021).

[6] S. Bernatz, J. Ackermann, P.C. Mandel, B. Kaltenbach, Y. Zhdanovich, P.N. Harter, C. Döring, R.M. Hammerstingl, B. Bodelle, K. Smith, A.M. Bucher, M.H. Albrecht, N. Rosbach, L.M. Basten, I. Yel, M. Wenzel, K. Bankov, I. Koch, F.K.H. Chun, J. Köllermann, P.J. Wild, T.J. Vogl, Comparison of machine learning algorithms to predict clinically significant prostate cancer of the peripheral zone with multiparametric MRI using clinical assessment categories and radiomic features, Eur. Radiol. 30 (2020) 6757–6769.

[7] M.Y. Chen, M.A. Woodruff, P. Dasgupta, N.J. Rukin, Variability in accuracy of prostate cancer segmentation among radiologists, urologists, and scientists, Cancer Med. 9 (19) (2020) 7172–7182, http://dx.doi.org/10.1002/cam4.3386.

[8] P. Steenbergen, K. Haustermans, E. Lerut, R. Oyen, L. De Wever, L. Van den Bergh, L.G. Kerkmeijer, F.A. Pameijer, W.B. Veldhuis, J.R. van der Voort van Zyp, F.J. Pos, S.W. Heijmink, R. Kalisvaart, H.J. Teertstra, C.V. Dinh, G. Ghobadi, U.A. van der Heide, Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation, Radiother. Oncol. 115 (2) (2015) 186–190.

[9] M.R. Liechti, U.J. Muehlematter, A.F. Schneider, D. Eberli, N.J. Rupp, A.M. tker, O.F. Donati, A.S. Becker, Manual prostate cancer segmentation in MRI: interreader agreement and volumetric correlation with transperineal template core needle biopsy, Eur. Radiol. 30 (9) (2020) 4806–4815.

[10] O. Zavala-Romero, A.L. Breto, I.R. Xu, Y.C. Chang, N. Gautney, A. Dal Pra, M.C. Abramowitz, A. Pollack, R. Stoyanova, Segmentation of prostate and prostate zones using deep learning: A multi-MRI vendor analysis, Strahlenther Onkol 196 (10) (2020) 932–942.

[11] O.J. Pellicer-Valero, J.L.M. Jiménez, V. Gonzalez-Perez, J.L.C. Ramón-Borja, I.M. García, M.B. Benito, P.P. Gómez, J. Rubio-Briones, M.J. Rupérez, J.D. Martín-Guerrero, Deep learning for fully automatic detection, segmentation, and gleason grade estimation of prostate cancer in multiparametric magnetic resonance images, 2022, arXiv:2103.12650.

[12] Z. Dai, E. Carver, C. Liu, J. Lee, A. Feldman, W. Zong, M. Pantelic, M. Elshaikh, N. Wen, Segmentation of the prostatic gland and the intraprostatic lesions on multiparametic magnetic resonance imaging Using Mask Region-based convolutional neural networks, Adv. Radiat. Oncol. 5 (3) (2020) 473–481.

[13] R. Cao, X. Zhong, S. Shakeri, A.M. Bajgiran, S.A. Mirak, D. Enzmann, S.S. Raman, K. Sung, Prostate cancer detection and segmentation in multi-parametric MRI via CNN and conditional random field, in: 2019 IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, 2019, pp. 1900–1904, http://dx.doi.org/10.1109/ISBI.2019.8759584.

[14] M. Bardis, R. Houshyar, C. Chantaduly, K. Tran-Harding, A. Ushinsky, C. Chahine, M. Rupasinghe, D. Chow, P. Chang, Segmentation of the prostate transition zone and peripheral zone on MR images with deep learning, Radiol. Imag. Cancer 3 (3) (2021) e200024, http://dx.doi.org/10.1148/rycan.2021200024, PMID: 33929265.

[15] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, R. Strand, F. Malmberg, Y. Ou, C. Davatzikos, M. Kirschner, F. Jung, J. Yuan, W. Qiu, Q. Gao, P.E. Edwards, B. Maan, F. van der Heijden, S. Ghose, J. Mitra, J. Dowling, D. Barratt, H. Huisman, A. Madabhushi, Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge, Med. Image Anal. 18 (2) (2014) 359–373, http://dx.doi.org/10.1016/j.media.2013.12.002, URL: https://www.sciencedirect.com/science/article/pii/S1361841513001734.

[16] L. Rundo, C. Han, J. Zhang, R. Hataya, Y. Nagano, C. Militello, C. Ferretti, M.S. Nobile, A. Tangherloni, M.C. Gilardi, S. Vitabile, H. Nakayama, G. Mauri, CNN-based prostate zonal segmentation on T2-weighted MR images: A cross-dataset study, in: A. Esposito, M. Faundez-Zanuy, F.C. Morabito, E. Pasero (Eds.), Neural Approaches to Dynamics of Signal Exchanges, Springer Singapore, Singapore, 2020, pp. 269–280, http://dx.doi.org/10.1007/978-981-13-8950-4_25.

[17] N.M. Rodrigues, S. Silva, L. Vanneschi, N. Papanikolaou, A comparative study of automated deep learning segmentation models for prostate MRI, Cancers 15 (5) (2023) http://dx.doi.org/10.3390/cancers15051467, URL: https://www.mdpi.com/2072-6694/15/5/1467.

[18] J.S. Bosma, A. Saha, M. Hosseinzadeh, I. Slootweg, M. de Rooij, H. Huisman, Semisupervised Learning with Report-guided Pseudo Labels for Deep Learning-based Prostate Cancer Detection Using Biparametric MRI, Radiol. Artif. Intell. 5 (5) (2023) e230031.

[19] M. Bardis, R. Houshyar, C. Chantaduly, K. Tran-Harding, A. Ushinsky, C. Chahine, M. Rupasinghe, D. Chow, P. Chang, Segmentation of the prostate transition zone and peripheral zone on MR images with deep learning, Radiol. Imag. Cancer 3 (3) (2021) e200024, http://dx.doi.org/10.1148/rycan.2021200024, PMID: 33929265.

[20] P. Hambarde, S.N. Talbar, N. Sable, A. Mahajan, S.S. Chavan, M. Thakur, Radiomics for peripheral zone and intra-prostatic urethra segmentation in MR imaging, Biomed. Signal Process. Control 51 (2019) 19–29, http://dx.doi.org/10.1016/j.bspc.2019.01.024.

[21] P. Hambarde, S. Talbar, A. Mahajan, S. Chavan, M. Thakur, N. Sable, Prostate lesion segmentation in MR images using radiomics based deeply supervised U-net, Biocybern. Biomed. Eng. 40 (4) (2020) 1421–1435, http://dx.doi.org/10.1016/j.bbe.2020.07.011.

[22] M.R.S. Sunoqrot, K.M. Selnæs, E. Sandsmark, S. Langørgen, H. Bertilsson, T.F. Bathen, M. Elschot, The reproducibility of deep learning-based segmentation of the prostate gland and zones on T2-weighted MR images, Diagnostics 11 (9) (2021) http://dx.doi.org/10.3390/diagnostics11091690, URL: https://www.mdpi.com/2075-4418/11/9/1690.

[23] K. Niu, X. Li, L. Zhang, Z. Yan, W. Yu, P. Liang, Y. Wang, C.-P. Lin, H. Zhang, C. Guo, K. Li, T. Qian, Improving segmentation reliability of multi-scanner brain images using a generative adversarial network, Quant. Imaging Med. Surg. 12 (3) (2022) 1775–1786.

[24] M. Svanera, M. Savardi, A. Signoroni, S. Benini, L. Muckli, Fighting the scanner effect in brain MRI segmentation with a progressive level-of-detail network trained on multi-site data, 2022, arXiv:2211.02400.

[25] J.L. Gunter, H.J. Wiste, K. Kantarci, S.D. Weigand, P. Vemuri, C.G. Schwarz, M.M. Mielke, J. Graff-Radford, D.S. Knopman, R.C. Petersen, C.R. Jack Jr., Effects of protocol and scanner changes on segmentation volume estimates in a dedicated crossover data set, Alzheimers. Dement. 17 (S1) (2021).

[26] J. Meglič, M.R.S. Sunoqrot, T.F. Bathen, M. Elschot, Label-set impact on deep learning-based prostate segmentation on MRI, Insights Imag. 14 (1) (2023) http://dx.doi.org/10.1186/s13244-023-01502-w.

[27] L.C. Adams, M.R. Makowski, G. Engel, M. Rattunde, F. Busch, P. Asbach, S.M. Niehues, S. Vinayahalingam, B. van Ginneken, G. Litjens, K.K. Bressem, Prostate158 - An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection, Comput. Biol. Med. 148 (2022) 105817, http://dx.doi.org/10.1016/j.compbiomed.2022.105817, URL: https://www.sciencedirect.com/science/article/pii/S0010482522005789.

[28] S.G. Armato, H. Huisman, K. Drukker, L. Hadjiiski, J.S. Kirby, N. Petrick, G. Redmond, M.L. Giger, K. Cha, A. Mamonov, J. Kalpathy-Cramer, K. Farahani, PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images, J. Med. Imag. (Bellingham) 5 (4) (2018) 044501.

[29] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Lecture Notes in Computer Science, Springer International Publishing, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[30] Q. Zhu, B. Du, B.I. Turkbey, P.L. Choyke, P. Yan, Deeply-supervised CNN for prostate segmentation, in: 2017 International Joint Conference on Neural Networks, IJCNN, 2017, pp. 178–184.

[31] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2) (2020) 203–211, http://dx.doi.org/10.1038/s41592-020-01008-z.

[32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035, URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[33] W. Falcon, The PyTorch Lightning team, PyTorch Lightning, 2019, http://dx.doi.org/10.5281/zenodo.3828935, URL: https://github.com/Lightning-AI/lightning.

[34] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, 2017, arXiv:1606.00915.

[35] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H.R. Roth, D. Xu, Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, 2022, pp. 272–284.

[36] A. Hatamizadeh, D. Yang, H. Roth, D. Xu, UNETR: Transformers for 3D medical image segmentation, in: Proc. IEEE Workshop Appl. Comput. Vis., 2021, pp. 1748–1758.

[37] V. Yeghiazaryan, I. Voiculescu, Family of boundary overlap metrics for the evaluation of medical image segmentation, J. Med. Imag. (Bellingham) 5 (1) (2018) 015006.

[38] O. Maier, A. Rothberg, P.R. Raamana, R. Bèges, F. Isensee, M. Ahern, mamrehn, VincentXWD, J. Joshi, loli/medpy: MedPy 0.4.0, Zenodo, 2019, http://dx.doi.org/10.5281/zenodo.2565940.

[39] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019, arXiv:1711.05101.

[40] P. Ramachandran, B. Zoph, Q.V. Le, Searching for activation functions, 2017, arXiv:1710.05941.

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (56) (2014) 1929–1958, URL: http://jmlr.org/papers/v15/srivastava14a.html.

[42] S.A. Taghanaki, Y. Zheng, S. Kevin Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, G. Hamarneh, Combo loss: Handling input and output imbalance in multi-organ segmentation, Comput. Med. Imaging Graph. 75 (2019) 24–33, http://dx.doi.org/10.1016/j.compmedimag.2019.04.005, URL: https://www.sciencedirect.com/science/article/pii/S0895611118305688.

[43] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, Y. Pan, Rethinking dice loss for medical image segmentation, in: 2020 IEEE International Conference on Data Mining, ICDM, 2020, pp. 851–860, http://dx.doi.org/10.1109/ICDM50108.2020.00094.

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 318–327, http://dx.doi.org/10.1109/TPAMI.2018.2858826.

[45] MONAI Consortium, MONAI: Medical open network for AI, 2022.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, arXiv:1706.03762.