Inter-site Variability in Prostate Segmentation Accuracy Using Deep Learning

Eli Gibson[1(✉)], Yipeng Hu[1], Nooshin Ghavami[1], Hashim U. Ahmed[2], Caroline Moore[1], Mark Emberton[1], Henkjan J. Huisman[3], and Dean C. Barratt[1]

[1] University College London, London, UK
eli.gibson@ucl.ac.uk
[2] Imperial College, London, UK
[3] Radboud University Medical Center, Nijmegen, The Netherlands

Abstract. Deep-learning-based segmentation tools have yielded higher reported segmentation accuracies for many medical imaging applications. However, inter-site variability in image properties can challenge the translation of these tools to data from '*unseen*' sites not included in the training data. This study quantifies the impact of inter-site variability on the accuracy of deep-learning-based segmentations of the prostate from magnetic resonance (MR) images, and evaluates two strategies for mitigating the reduced accuracy for data from unseen sites: training on multi-site data and training with limited additional data from the unseen site. Using 376 T2-weighted prostate MR images from six sites, we compare the segmentation accuracy (Dice score and boundary distance) of three deep-learning-based networks trained on data from a single site and on various configurations of data from multiple sites. We found that the segmentation accuracy of a single-site network was substantially worse on data from unseen sites than on data from the training site. Training on multi-site data yielded marginally improved accuracy and robustness. However, including as few as 8 subjects from the unseen site, e.g. during commissioning of a new clinical system, yielded substantial improvement (regaining 75% of the difference in Dice score).

Keywords: Segmentation · Deep learning · Inter-site variability
Prostate

1 Introduction

Deep-learning-based medical image segmentation methods have yielded higher reported accuracies for many applications including prostate [8], brain tumors [1] and abdominal organs [7]. Applying these methods in practice, however, remains challenging. Few segmentation methods achieve previously reported accuracies on new data sets. This may be due, in part, to *inter-site* variability in image and

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11073, pp. 506–514, 2018.
https://doi.org/10.1007/978-3-030-00937-3_58

reference segmentation properties at different imaging centres due to different patient populations, clinical imaging protocols and image acquisition equipment.

Inter-site variability has remained a challenge in medical image analysis for decades [9,12]. Data sets used to design, train and validate segmentation algorithms are, for logistical and financial reasons, sampled in clusters from one or a small number of imaging centres. The distribution of images and reference segmentations in this clustered sample may not be representative of the distribution of these data across other centres. Consequently, an algorithm developed for one site may not be optimal for other 'unseen' sites not included in the sample, and reported estimates of segmentation accuracy typically overestimate the accuracy achievable at unseen sites.

Data-driven methods, including deep learning, may be particularly susceptible to this problem because they are explicitly optimized on the clustered training data. Additionally, deep-learning-based methods typically avoid explicit normalization methods, such as bias field correction [12], to mitigate known sources of inter-site variability and high-level prior knowledge, such as anatomical constraints, to regularize models. Instead, normalization and regularization are implicitly learned from the clustered training data. The accuracy of deep-learning-based methods may, therefore, depend more heavily on having training data that is representative of the images to which the method will be applied.

One strategy to mitigate this effect is to use images and reference segmentations sampled from multiple sites to better reflect inter-site variability in the training data. A second approach is to 'commission' the systems: in clinical practice, when introducing new imaging technology, hospital staff typically undertake a commissioning process to calibrate and validate the technology, using subjects or data from their centre. In principle, such a process could include re-training or fine-tuning a neural network using a limited sample of data from that site. These strategies have not been evaluated for deep-learning-based segmentation.

In this study, we aimed to quantify the impact of inter-site variability on the accuracy of deep-learning-based segmentations of the prostate from T2-weighted MRI of three deep-learning-based methods, and to evaluate two strategies to mitigate the accuracy loss at a new site: training on multi-site data and training augmented with limited data from the commissioning site. To identify general trends, we conducted these experiments using three different deep-learning based methods. Specifically, this study addresses the following questions:

1. How accurate are prostate segmentations using networks trained on data from a single site when evaluated on data from the same and unseen sites?
2. How accurate are prostate segmentations using networks trained on data from multiple sites when evaluated on data from the same and unseen sites?
3. Can the accuracy of these prostate segmentations be improved by including a small sample of data from the unseen site?

2 Methods

2.1 Imaging

This study used T2-weighted 3D prostate MRI from 6 sites (256 from one site [SITE1] and 24 from 5 other sites [SITE2–SITE6]), drawn from publicly available data sets and clinical trials requiring manual prostate delineation. Reference standard manual segmentations were performed at one of 3 sites: SITE1, SITE2 or SITE5. Images were acquired with anisotropic voxels, with in-plane voxel spacing between 0.5 and 1.0 mm, and out-of-plane slice spacing between 1.8 and 5.4 mm. All images, without intensity normalization, and reference standard segmentations were resampled from their full field of view ($12 \times 12 \times 5.7\,\text{cm}^3$ – $24 \times 24 \times 17.2\,\text{cm}^3$) to $256 \times 256 \times 32$ voxels before automatic segmentation.

2.2 Experimental Design

We evaluated the segmentation accuracies (Dice score and the symmetric boundary distance (BD)) of networks in three experiments with training data sets taken (1) from a single site, (2) with the same sample size from multiple sites, or (3) from multiple sites but with fewer samples from one 'commissioned' site. Segmentation accuracy was evaluated with 'same-site' test data from sites included in training data, 'unseen-site' test data from sites excluded from the training data, and 'commissioned-site' test data from the commissioned site. No subject was included in both training and test data for the same trained network. Three network architectures (Sect. 2.3) were trained and tested for each data partition.

Experiment 1: Single-site Networks. To evaluate the segmentation accuracy of networks trained on data from one site (referred to as *single-site* hereafter), we trained them on 232 subjects from SITE1, and evaluated them on the remaining 24 subjects from SITE1 and all subjects from the other sites.

Experiment 2: Multi-site Networks. To evaluate the segmentation accuracy of networks trained on data from multiple sites, we used two types of data partitions. First, we conducted a patient-level 6-fold cross-validation (referred to as *patient-level* hereafter) where, in each fold, 16 subjects from each site were used for training, and 8 subjects from each site were used for same-site testing. This same-site evaluation has been used in public challenges, such as the PROMISE12 segmentation challenge [8]. Because this may overestimate the accuracy at a site that has not been seen in training, we conducted a second site-level 6-fold cross-validation (referred to as *site-level* hereafter) where, in each fold, 24 subjects from each of 5 sites were used for training, and 24 subjects from the remaining site were used for unseen-site testing.

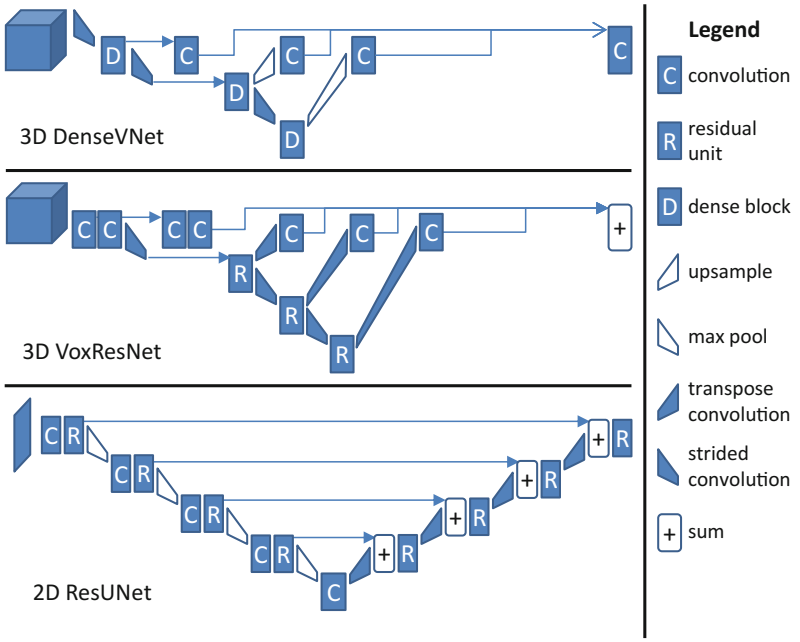

Fig. 1. Architectures of the neural networks.

Experiment 3: Commissioned Networks. To evaluate the utility of commissioning segmentation methods at new imaging centres, we conducted a 6×6-fold hierarchical cross-validation where the 6 outer folds correspond to selecting one site as the commissioned site and the 6 inner folds correspond to selecting a subset of subjects from the commissioned site (3 subsets with 8 subjects and 3 subsets with 16). Each network was trained with the 8 or 16 selected subjects from the commissioned site and 24 subjects from each of the other 5 sites (referred to as commission-8 and commission-16, hereafter). In each fold, the remaining subjects from the commissioned site that were excluded from training were used for commissioned-site testing.

2.3 Neural Networks: Architectures and Training

To distinguish general trends from network-specific properties, three different neural network architectures, illustrated in Fig. 1 were used in this study: DenseVNet [4], ResUNet [3], and VoxResNet [2]. Like many recent medical image segmentation networks, these networks are all variants of U-Net architectures [11] comprising a downsampling subnetwork, an upsampling subnetwork and skip connections. ResUNet segments 2D axial slices using a 5-resolution U-Net with residual units [5], max-pooling, and additive skip connections. DenseVNet segments 3D volumes using a 4-resolution V-Net with dense blocks [6] with batch-wise spatial dropout, and convolutional skip connections concatenated prior to

a final segmentation convolution. VoxResNet segments 3D volumes using a 4-resolution V-Net with residual units [5], transpose-convolution upsampling, and deep supervision to improve gradient propagation. It is important to note that this study is not designed to compare the absolute accuracy of these networks; accordingly, the network dimensionality and features, hyperparameter choices, and training regimen were not made equivalent, and, apart from setting an appropriate anisotropic input shape, no hyperparameter tuning was done.

For each fold of each experiment, the network was trained by minimizing the Dice loss using the Adam optimizer for 10000 iterations. The training data set was augmented using affine perturbations. Segmentations were post-processed to eliminate spurious segmentations by taking the largest connected component.

3 Results

The described experiments generated more than 2000 segmentations across various data partitioning schemes: single-site networks trained on data from one site, patient-level networks trained on data from all sites, site-level networks trained on data from all sites except the testing site, and commissioned networks trained on 8 or 16 subjects from the commissioned site and all subjects from all other sites. The segmentation accuracies for DenseVNet, VoxResNet and ResUNet are detailed in Table 1, illustrated in Fig. 2 and summarized below.

For single-site networks, the mean accuracy on unseen-site test data was lower than on same-site test data and varied substantially between sites, confirming the same-site evaluation overestimated the unseen-site accuracy due to inter-site variability. The mean Dice score decreased by 0.12 ± 0.15 [0.00–0.47] (mean $\pm$ SD [range]) and the mean boundary distance increased by 2.0 ± 2.6 [0.1–6.9] mm.

For the multi-site training, the mean accuracies generally improved as more training data from the testing site was included, best illustrated in Fig. 2. The patient-level and site-level cross-validations yield two notable observations. First, for the *patient-level* networks, the same-site mean accuracies (Dice: 0.88, 0.84, 0.85; BD: 1.6 mm, 2.0 mm, 1.9 mm) were nearly identical to the same-site testing of the single-site networks (Dice 0.88, 0.85, 0.87; BD: 1.6 mm, 2.0 mm, 1.7 mm), suggesting that it was not inherently more difficult to train the networks on multi-site data than on single-site data. Second, for the *site-level* VoxResNet and ResUNet networks (those with worse generalization), the unseen-site accuracies for multi-site training (Dice: 0.75, 0.75; BD: 4.5 mm, 3.5 mm) were better and less variable than for single-site training (Dice: 0.68, 0,71; BD: 4.9 mm, 4.1 mm), suggesting that training on multi-site data alone yields improvements in generalization. This effect was not observed for DenseVNet, however.

For commissioned networks (with some training data from the testing site), segmentation accuracies on commissioned-site test data regained most of the difference between the *same-site* patient-level and *unseen-site* site-level cross-validations. With only 8 subjects used as commissioning data, segmentation accuracies regained $75 \pm 21\%$ [28–97%] (mean $\pm$ SD [range]) of the Dice score difference (averaged Dice: 0.87, 0.84, 0.83; BD: 1.7 mm, 2.1 mm, 2.3 mm) when

the Dice score discrepancy was >0.02. With 16 subjects used as commissioning data, segmentation accuracies regained a $90 \pm 12\%$ [66–100%] of the Dice score difference (averaged Dice: 0.87, 0.85, 0.84; BD: 1.7 mm, 1.9 mm, 2.0 mm) when the Dice score discrepancy was >0.02.

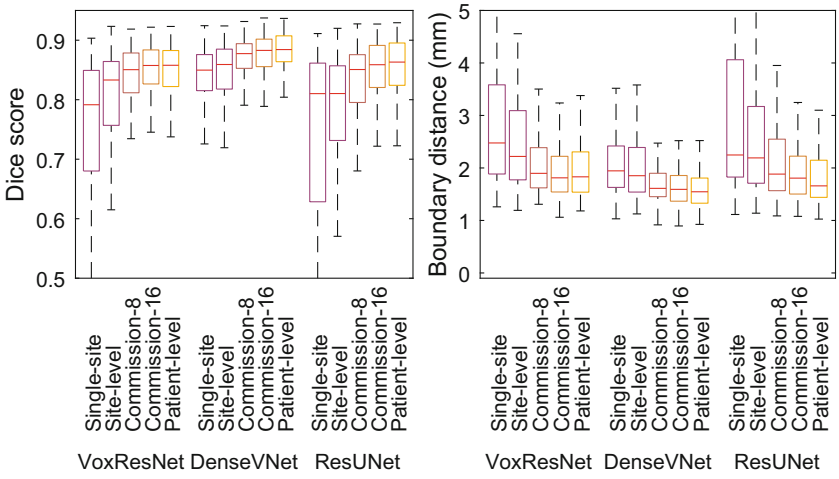

Fig. 2. Box and whisker plots of segmentation accuracies.

4 Discussion

In this work, we demonstrated that multiple deep-learning-based segmentation networks have poor accuracy when applied to data from unseen sites. This challenges the translation of segmentation tools based on these networks to other research sites and to clinical environments.

As illustrated in our study, different medical image analysis methods have different capacities to generalize to new sites. Since this is important for their clinical and research impact, methods' generalization ability should become a metric evaluated by our community. This will require the creation of multi-site datasets, such as PROMISE12 [8] and ADNI [10], to design and evaluate methods. Standardized evaluation protocols, in independent studies and in MICCAI challenges, should include unseen sites in the test set to evaluate generalizability. This will promote the development of methods that generalize better, using established techniques, e.g. dropout as in DenseVNet, or new innovations.

For both single- and multi-site training data set, some sites consistently yielded poorer accuracy when no data from that site was included in training. SITE5 yielded low accuracies in many analyses, likely due to site-specific differences in prostate MRI protocol: for example, the median inter-slice spacing at SITE5 was 4.7 mm compared to 2.8 mm across the other sites. One solution

Table 1. Segmentation accuracies for DenseVNet, VoxResNet and ResUNet.

DenseVNet		SITE1	SITE2	SITE3	SITE4	SITE5	SITE6	Pooled
Training	Testing	Dice coefficient (0–1)						
Single-site	same-site	0.88						
Single-site	unseen-site		0.88	0.84	0.83	0.77	0.85	0.83
Patient-level	same-site	0.87	0.90	0.87	0.88	0.86	0.88	0.88
Site-level	unseen-site	0.87	0.88	0.85	0.74	0.78	0.85	0.83
Commission-8	commissioned-site	0.87	0.89	0.87	0.86	0.86	0.88	0.87
Commission-16	commissioned-site	0.87	0.89	0.88	0.86	0.86	0.88	0.87
		Boundary distance (mm)						
Single-site	same-site	1.6						
Single-site	unseen-site		1.8	2.0	2.2	3.4	2.0	2.3
Patient-level	same-site	1.7	1.5	1.5	1.5	2.0	1.6	1.6
Site-level	unseen-site	1.8	1.7	1.8	4.2	3.2	2.0	2.4
Commission-8	commissioned-site	1.8	1.6	1.6	1.7	2.0	1.7	1.7
Commission-16	commissioned-site	1.8	1.6	1.5	1.7	2.1	1.6	1.7
VoxResNet		SITE1	SITE2	SITE3	SITE4	SITE5	SITE6	Pooled
Training	Testing	Dice coefficient (0–1)						
Single-site	same-site	0.85						
Single-site	unseen-site		0.81	0.83	0.58	0.37	0.80	0.68
Patient-level	same-site	0.84	0.87	0.86	0.84	0.80	0.86	0.84
Site-level	unseen-site	0.83	0.83	0.85	0.66	0.50	0.83	0.75
Commission-8	commissioned-site	0.85	0.86	0.85	0.83	0.79	0.84	0.84
Commission-16	commissioned-site	0.85	0.88	0.86	0.85	0.82	0.85	0.85
		Boundary distance (mm)						
Single-site	same-site	2.0						
Single-site	unseen-site		2.7	2.1	8.1	8.9	2.6	4.9
Patient-level	same-site	2.1	1.9	1.7	1.9	2.7	1.9	2.0
Site-level	unseen-site	2.2	2.2	1.8	5.8	6.6	2.3	3.5
Commission-8	commissioned-site	2.0	1.9	1.8	2.1	2.9	2.0	2.1
Commission-16	commissioned-site	2.0	1.7	1.7	1.8	2.5	1.9	1.9
ResUNet		SITE1	SITE2	SITE3	SITE4	SITE5	SITE6	Pooled
Training	Testing	Dice coefficient (0–1)						
Single-site	same-site	0.87						
Single-site	unseen-site		0.84	0.77	0.48	0.63	0.82	0.71
Patient-level	same-site	0.85	0.88	0.87	0.87	0.81	0.84	0.85
Site-level	unseen-site	0.83	0.84	0.83	0.71	0.51	0.80	0.75
Commission-8	commissioned-site	0.84	0.85	0.86	0.84	0.74	0.82	0.83
Commission-16	commissioned-site	0.84	0.86	0.85	0.86	0.78	0.85	0.84
		Boundary distance (mm)						
Single-site	same-site	1.7						
Single-site	unseen-site		2.0	2.4	8.2	5.9	2.2	4.1
Patient-level	same-site	2.0	1.7	1.6	1.6	2.4	2.1	1.9
Site-level	unseen-site	2.1	2.0	1.9	3.9	8.4	2.5	3.5
Commission-8	commissioned-site	2.1	2.0	1.6	2.0	3.7	2.3	2.3
Commission-16	commissioned-site	2.1	1.8	1.7	1.7	2.8	2.0	2.0

to this problem would be to adjust clinical imaging at this site to be more consistent with other sites; however, such a solution could be very disruptive. Note that this effect almost disappears in the patient-level cross-validation suggesting that these cases are probably not substantially harder to segment, as long as they are represented in the training data to some extent. This suggests that the more practical solution of retraining the segmentation network with some data from each site during the commissioning process may be effective.

The conclusions of this study should be considered in the context of its limitations. Our study focused exclusively on prostate segmentation, where deep-learning-based segmentation methods have become dominant and multi-site data sets are available. Reproducing our findings on other segmentation problems, once appropriate data are available, will be valuable. We observed variability between networks in their generalization to new sites; while we evaluated three different networks, we cannot conclude that all networks will need commissioning with data from each new site. Evaluating each network required training 49 networks, so a more exhaustive evaluation was not feasible for this work.

Our analysis confirmed that the accuracy of deep-learning-based segmentation networks trained and tested on data from one or more sites can overestimate the accuracy at an unseen site. This suggests that segmentation evaluation and especially segmentation challenges should include data from one or more completely unseen sites in the test data to estimate how well methods generalize, and promote better generalization. This also suggests that commissioning segmentation methods at a new site by training networks with a limited number of additional samples from that site could effectively mitigate this problem.

Acknowledgements. This publication presents independent research supported by Cancer Research UK (Multidisciplinary C28070/A19985).

References

1. Bakas, S., Menze, B., Davatzikos, C., Reyes, M., Farahani, K. (eds.): International MICCAI BraTS Challenge (2017)
2. Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A.: VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage (2017)
3. Ghavami, N., et al.: Automatic slice segmentation of intraoperative transrectal ultrasound images using convolutional neural networks. In: SPIE Medical Imaging, February 2018
4. Gibson, E., et al.: Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE TMI (2018)
5. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. arXiv:1603.05027 (2016)
6. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. arXiv:1608.06993 (2016)
7. Landman, B., Xu, Z., Igelsias, J.E., Styner, M., Langerak, T.R., Klein, A.: MICCAI Multi-atlas Labeling Beyond the Cranial Vault - Workshop and Challenge (2015)
8. Litjens, G., et al.: Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med. Image Anal. **18**(2), 359–373 (2014)

9. Mirzaalian, H., et al.: Harmonizing diffusion MRI data across multiple sites and scanners. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 12–19. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_2
10. Mueller, S.G., et al.: The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. **15**(4), 869–877 (2005)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. Styner, M.A., Charles, H.C., Park, J., Gerig, G.: Multisite validation of image analysis methods: assessing intra-and intersite variability. In: Medical Imaging 2002: Image Processing, vol. 4684, pp. 278–287. SPIE (2002)