Physics and Imaging in Radiation Oncology 15 (2020) 1–7

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

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

A convolutional neural network for contouring metastatic lymph nodes on diffusion-weighted magnetic resonance images for assessment of radiotherapy response

Oliver J. Gurney-Champion[a,1,*], Jennifer P. Kieselmann[a], Kee H. Wong[b], Brian Ng-Cheng-Hin[c], Kevin Harrington[c], Uwe Oelfke[a]

[a] Joint Department of Physics, The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, London, United Kingdom
[b] The Royal Marsden NHS Foundation Trust, London, United Kingdom
[c] Targeted Therapy Team, The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, London, United Kingdom

ARTICLE INFO

Keywords:
Diffusion magnetic resonance imaging
Radiotherapy
Deep learning
Neural networks, Computer
Head and neck neoplasms
Lymph nodes
Magnetic resonance imaging
Contouring
Convolutional neural networks
MR-Linac
MR-guided radiotherapy

ABSTRACT

Background and purpose: Retrieving quantitative parameters from magnetic resonance imaging (MRI), e.g. for early assessment of radiotherapy treatment response, necessitates contouring regions of interest, which is time-consuming and prone to errors. This becomes more pressing for daily imaging on MRI-guided radiotherapy systems. Therefore, we trained a deep convolutional neural network to automatically contour involved lymph nodes on diffusion-weighted (DW) MRI of head and neck cancer (HNC) patients receiving radiotherapy.
Materials and methods: DW-images from 48 HNC patients (18 induction-chemotherapy + chemoradiotherapy; 30 definitive chemoradiotherapy) with 68 involved lymph nodes were obtained on a diagnostic 1.5 T MR-scanner prior to and 2–3 timepoints throughout treatment. A radiation oncologist delineated the lymph nodes on the b = 50 s/mm^2 images. A 3D U-net was trained to contour involved lymph nodes. Its performance was evaluated in all 48 patients using 8-fold cross-validation and calculating the Dice similarity coefficient (DSC) and the absolute difference in median apparent diffusion coefficient (ΔADC) between the manual and generated contours. Additionally, the performance was evaluated in an independent dataset of three patients obtained on a 1.5 T MR-Linac.
Results: In the definitive chemoradiotherapy patients (n = 96 patients/lymphnodes/timepoints) the DSC was 0.87 (0.81–0.91) [median (1st-3rd quantiles)] and ΔADC was 1.9% (0.8–3.4%) and both remained stable throughout treatment. The network performed worse in the patients receiving induction-chemotherapy (n = 65), with DSC = 0.80 (0.71–0.87) and ΔADC = 3.3% (1.6–8.0%). The network performed well on the MR-Linac data (n = 8) with DSC = 0.80 (0.75–0.82) and ΔADC = 4.0% (0.6–9.1%).
Conclusions: We established accurate automatic contouring of involved lymph nodes for HNC patients on diagnostic and MR-Linac DW-images.

1. Introduction

By studying the tumour microenvironment throughout radiotherapy (RT) treatment, we might be able to determine an optimal tumour-specific dose depending on the treatment's efficacy and update treatment accordingly [1]. One way of studying the tumour microenvironment is by diffusion-weighted (DW) magnetic resonance (MR) imaging (MRI) [2–5]. However, the exact predictive and prognostic value of DW MRI's quantitative parameter (apparent diffusion coefficient; ADC) in the context of RT remains to be defined.

The ideal system for obtaining regular DW MRI of RT patients is an MR-guided RT system, such as the MR-Linac. Studies assessing longitudinal DW MRI on such systems are currently underway [6], and, ultimately, DW MRI can be performed daily throughout treatment. Such an approach would deliver a wealth of information, enabling a full evaluation of the relation between treatment response and ADC.

However, to retrieve ADC values, regions of interest (ROIs) need to be drawn within the images. Currently, an expert clinician places these

* Corresponding author at: 15 Cotswold Road, Sutton, London SM2 5NG, United Kingdom
E-mail address: o.j.gurney-champion@amsterdamumc.nl (O.J. Gurney-Champion).
[1] Currently at Amsterdam UMC, University of Amsterdam, Department of Radiology and Nuclear Medicine, Cancer Center Amsterdam, Meibergdreef 9, Amsterdam, Netherlands.

https://doi.org/10.1016/j.phro.2020.06.002
Received 13 February 2020; Received in revised form 9 June 2020; Accepted 9 June 2020
Available online 21 June 2020
2405-6316/ © 2020 The Author(s). Published by Elsevier B.V. on behalf of European Society of Radiotherapy & Oncology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

ROIs manually. Such a process is labour-intensive, which will become a major issue when DW MRI is obtained daily (30 fractions; 30 contour sets). Furthermore, clinicians do not agree upon the precise ROI boundaries [7], resulting in contour variations.

Automation of contouring could substantially decrease the workload while increasing contour consistency [8]. In recent years, computer vision has greatly improved, especially due to the introduction of convolutional neural networks [9]. A promising commonly used network for biomedical image contouring is the U-net [10,11], which has been successfully employed for contouring on head and neck cancer (HNC) patient computed tomography (CT) images [12] and T2-weighted MR-images [13].

We hypothesise that a 3D U-net can be utilized for automatic and accurate contouring of metastatic lymph nodes in DW-images from HNC patients. We used a database with diagnostic MR-images from 48 patients with metastatic lymph nodes who underwent MRI at different timepoints throughout treatment to train and evaluate our network. We further assessed the network's performance on a fully independent dataset from the MR-Linac.

2. Materials and Methods

2.1. Data

We used two datasets: the diagnostic MRI set and the MR-Linac set (Table 1). Our local ethics committee approved both studies and all patients gave written informed consent. Our exclusion criteria were: 1) lymph nodes smaller than 100 voxels (< 0.8 cm^3) because our evaluation metric, the Dice similarity coefficient (DSC), was not suitable for small volumes; 2) retropharyngeal lymph nodes, as only one metastatic retropharyngeal lymph node was visible in our dataset; 3) images with large artefacts, e.g. due to dental implants, as clinicians were also unable to accurately contour.

Clinical results from the diagnostic set were previously published [2,14]. It contained 60 patients receiving chemoradiotherapy (CRT). After the exclusion criteria listed above were applied, the dataset consisted of 124 DW-images of 68 metastatic lymph nodes from 48 patients. Eighteen patients received a course of induction chemotherapy (IC) prior to CRT and the remainder received definitive CRT alone. CRT consisted of six weeks of RT with concomitant chemotherapy (100 mg/m^2 cisplatin or carboplatin AUC 5 on days 1 and 29), whereas the IC consisted of two additional cycles of three-weekly TPF chemotherapy prior to RT (day 1: 75 mg/m^2 docetaxel and 75 mg/m^2 cisplatin; days1–4: 1000 mg/m^2 5-fluorouracil). For the IC + CRT group, MR-images were obtained at baseline, during IC (three weeks and six weeks into treatment) and one week into CRT. For the CRT-only patients, MR-images were obtained at baseline, and one week and two weeks into CRT.

Table 1
MRI scan parameters.

	Diagnostic dataset	MR-Linac dataset
Patients	48	3
lymph nodes	68	8
Scanner	1.5 T Magnetom Aera*	1.5 T Unity†
Coils	Large flex (8-channel) and spine (32-channel)	Posterior (4-channel) and anterior (4-channel)
Sequence	Axial 2D multi-slice EPI	Axial 2D multi-slice EPI
Diffusion-weighting	Mono-polar diffusion gradients	Mono-polar diffusion gradients
Field of view	200 × 200 mm^2	400 × 240 mm^2
Resolution	2 × 2 mm^2	3.2 × 3.2 mm^2 (1.2 × 1.2 mm^2 reconstruction)
Slices	40	39
Slice thickness	2 mm	4.5 mm
TR/TE	13,400/61 ms	5,000/63 ms
Bandwidth	1,000 Hz	2,053 Hz
b-values	50, 400 and 800 s/mm^2	0, 50, 400 s/mm^2
Averages	5, 5, 5	5, 5, 15

* Siemens Healthineers, Erlangen, Germany; † Elekta, Stockholm, Sweden.
Abbreviations: EPI: echo-planar imaging; TR: repetition time; TE: echo time.

The MR-Linac set consisted of DW-images from three patients with a total of eight metastatic lymph nodes. The images were taken at baseline (three patients) and two weeks into treatment (one patient).

For both datasets, patients were imaged in RT positioning, using a flat tabletop, a headrest with 5-point thermoplastic shell immobilisation (i.e. Fig. 2 from [14]). Table 1 shows further acquisition details.

An expert clinician (KW; 6 years of experience) contoured the metastatic lymph nodes (including necrotic regions) on the b = 50 s/mm^2 image with guidance of the other available images (T2-weighted and dynamic contrast-enhanced images for the diagnostic set; T2-weighted and Dixon images for the MR-Linac set) using the treatment planning system RayStation (RaySearch Laboratories AB, Stockholm, Sweden). The clinicians felt most confident contouring on the b = 50 s/mm^2 as it had a good trade-off between signal-to-noise ratio and visibility of the involved lymph nodes and surrounding tissue. The contours were drawn for the purpose of evaluating the ADC values within the lymph nodes. These contours were used to train and evaluate the network. To enable the evaluation of interobserver variation as a reference benchmark, a second expert clinician (BN; 7 years of experience) contoured the metastatic lymph nodes on 15 randomly selected baseline scans. ADC-maps were calculated by the vendor-provided software using all b-values.

2.2. Network

In a clinical workflow, one is interested in the ADC of a given lymph node. Therefore, we envisioned a clinical workflow in which a clinician selects a metastatic lymph node (mouse click) on the image to initiate the network. In this workflow, a bounding box (64 × 64 × 32 voxels) is placed centred at the selected voxel and used as input for the U-net.

We implemented a 3D U-net [11] in Python (version 3.6.6) using Keras (version 2.2.2) [15] and Tensorflow (version 1.10) [16]. The network built upon an earlier implementation by Kieselmann et al [17]. The input consisted of a single-channel image of 64 × 64 × 32 voxels. Our 3D U-net was similar to the original 3D U-net [11], except that we used zero-padding, had 5 resolution steps (similar to 2D U-net [10]), instead of 4 and added a local bias layer (LocalBias from neurons toolkit [18]) before each ReLu layer. The bias layer allowed the network to have spatial awareness and, hence, to focus on the central lymph node. At full resolution, the convolutions consisted of 64 feature channels and at each subsequent resolution level, the convolution doubled the number of features up to 1024 at the bottleneck. Our final layer consisted of a 1 × 1 × 1 convolution followed by a local bias layer and sigmoid activation function.

2.3. Training

Our network was trained to contour on the b = 50 s/mm^2 DW-images (no additional channels). Networks were trained on a Tesla V100-PCIE-16 GB GPU with 112 TFLOPS (NVIDIA, CA, USA). MR-images were normalized by dimming the 0.5% brightest voxels to the 0.5% percentile intensity and then normalizing all intensities to a value between 0 and 255. We used a Dice loss as loss function [19]. The network was trained using an Adam optimiser [20] with a learning rate of 2×10^{-4} and a batch size of 6. Dropout [21] of 20% was introduced throughout the network, as well as batch normalisation [22]. Once the performance of the network on the validation dataset did not improve over the past 20 epochs, the training was stopped and the best performing model was saved. We used data augmentation. By mirroring in left–right direction pre-training all data was doubled. On-the-fly data-augmentation was used to simulate the clinician's click by selecting a random voxel from the lymph node contour as centre for our input patch.

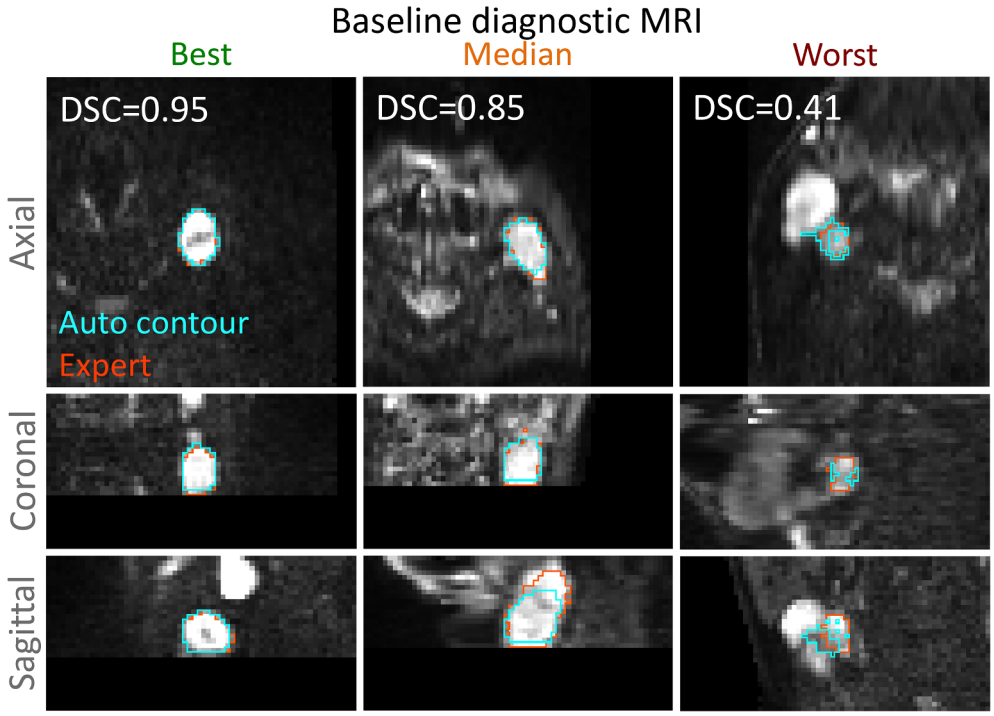

Fig. 1. DW-images (b = 50 s/mm^2) with the best, median and worst performing auto-contours of the baseline data.

2.4. Evaluation

We validated the network on all patients by making use of 8-fold cross-validation at patient level (repeated scans were in the same group). Eight networks were trained separately. For each network, six different patients were removed for independent testing and not shown to that network. The remaining patients were used to train that network and were split such that 80% of the lymph nodes were used to train the system and update the weights, whereas 20% were used as the validation dataset for determining stopping criteria. Once trained, the network was evaluated on the six independent test patients that the network had not seen nor used for validation. This way, every patient could be used as independent test for one of the eight networks and the networks were validated on a total of 48 patients to extensively evaluate it performance.

All voxels for which the network was 50% certain of being metastatic lymph nodes were included in the predicted contours. For some patients, the lymph nodes were close to each other and multiple nodes would be present within the input patch. For automatic evaluation, a post-processing toolkit was developed that selected the central lymph node of interest. This toolkit used a distance transform on the predicted lymph node map, followed by a watershed algorithm (scikit's skimage.morphology.watershed; compactness = 0.15) [23] originating from the different selected lymph node locations (simulated clicks).

Quantitative evaluation of data was done separately for the IC + CRT and CRT-only patients. The DSC between the manual contours and the contours generated by the network was used as the main evaluation criteria (1 is full overlap, 0 is no overlap). We also calculated the DSC between the manual contours of the two expert clinicians in the subset of 15 patients in which we had obtained repeated contours. For comparison, the median DSC between the auto-contour and the expert clinician was also recalculated using only these 15 patients. After testing for normality (Shapiro-Wilk test at significance level $\alpha = 0.05$), a paired samples Wilcoxon signed-rank test was performed to identify any significant differences (significance level $\alpha = 0.05$).

One of the clinically interesting parameters is the median ADC value from within the ROI which can potentially be used as biomarker to personalise treatment or for treatment response monitoring. Therefore, we compared the median ADC value from within the auto-contour ROI to the one from the clinician. Due to the low sample size, it was hard to guarantee normality, and hence we used a paired samples Wilcoxon signed-rank test to test for any significant systematic differences (significance level $\alpha = 0.05$). We also reported the absolute difference of median ADC over the patient group as ΔADC.

To investigate how acquisition at a lower resolution would affect the performance of our network, we repeated training and validation of the diagnostic MRI data while decreasing the simulated acquisition resolution from 2.0 mm to 5.0 mm in steps of 0.5 mm. This was done by downscaling the image to the desired resolution and upscaling back to a blurred 2.0 mm.

2.5. MR-Linac

To assess the performance of our network in a different independent dataset, we applied our network to MR-Linac data. Note that this dataset had a substantially different image acquisition protocol. We sampled down the diagnostic MRI data to 3.2 × 3.2 × 4.5 mm^3 (acquisition resolution from MR-Linac data) and then sampled up both datasets to 2 × 2 × 2 mm^3 resolution. The network was retrained using all resampled diagnostic MRI data with an 80/20% split between training/validation. Once trained, its performance was evaluated on the MR-Linac dataset, without ever having seen MR-Linac data.

3. Results

3.1. Diagnostic dataset

The network took an average of 245 min to train (range 221–265 min), whereas inference only took 55 ms (range 52–58 ms). Fig. 1 illustrates the contours on the baseline lymph nodes where the network had best, median and worst performance. In the worst performing case, the network contoured the lymph node properly, but the post-processing attributed the contours to its neighbouring lymph node, instead. The contours of the CRT-only patients showed a median DSC of

Table 2
The median (1st quantile-3rd quantile) DSC and ΔADC for the CRT-only (top) and IC + CRT (bottom) patients.

CRT-only	Baseline		Week 1		Week 2		Overall	
n*	41		29		25		96	
DSC	0.89	(0.82–0.92)	0.85	(0.8–0.89)	0.84	(0.79–0.89)	0.87	(0.81–0.91)
ΔADC (%)	1.4	(0.57–3.4)	2.1	(1.0–4.8)	1.8	(0.8–2.5	1.9	(0.8–3.4)

IC + CRT			Induction chemo				Radiotherapy			
	Baseline		**Week 3**		**Week 6**		**Week 1**		**Overall**	
n*	26		18		12		9		65	
DSC	0.82	(0.78–0.87)	0.82	(0.71–0.87)	0.72	(0.63–0.84)	0.71	(0.40–0.79)	0.80	(0.71–0.87)
ΔADC (%)	3.0	(1.2–7.3)	2.7	(1.5–7.4)	4.0	(2.2–7.8)	5.7	(3.3–11.4)	3.3	(1.6–8.0)

* As patients responded to treatment, fewer metastasized lymph nodes were observed throughout treatment. Abbreviations: n is the number of metastasized lymph nodes analysed, DSC = Dice similarity coefficient, ΔADC = the percentage of absolute change in ADC between expert observer and auto-contour.

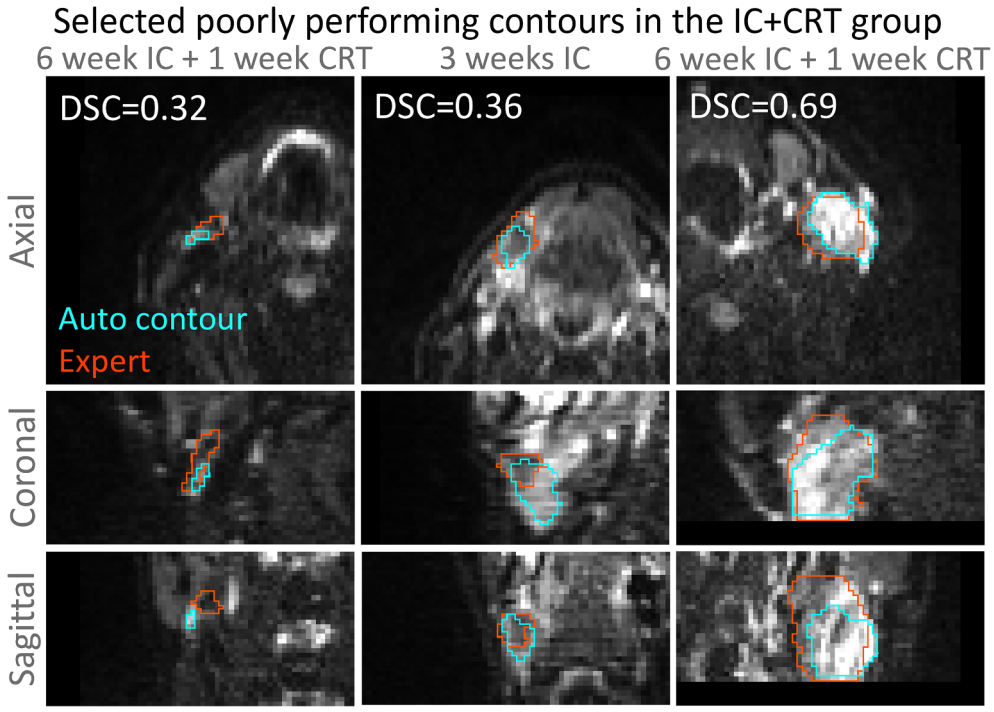

Fig. 2. Selected poorly performing contours in DW-images (b = 50 s/mm^2) from different timepoints throughout IC + CRT.

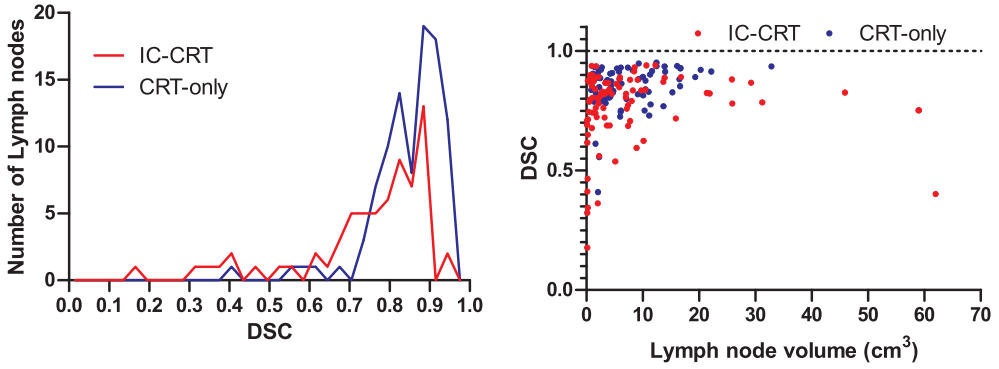

Fig. 3. Histograms of the Dice similarity coefficients (DSCs) for both patient groups (left) as well as the relation between DSC and mask size (right).

0.87, which did not change considerably throughout treatment (Table 2). For the IC + CRT patients, the network performed similarly at baseline, however, its performance substantially decreased throughout treatment (Table 2). The decrease in performance was partially attributed to the fact that metastatic lymph nodes got poorly defined diffuse borders during IC-CRT (Fig. 2). Fig. 3 shows that most DSCs were skewed towards the high end of the spectrum with some outliers to lower values for both groups. The lowest DSCs were found in smaller lymph nodes (Fig. 3).

The median (1st-3rd quartiles; notation used throughout paper) DSC

between both expert observers was 0.92 (0.87–0.93) in the 15 patients that had two sets of contours. The network had a median DSC of 0.89 (0.83–0.93) in these patients. This subset of patients happened to include the one patient where the post-processing step from the neural network failed (Fig. 1). When this data point was considered an outlier due to malfunctioning of the post-processing, the median DSC of the network increased to 0.90 (0.84–0.93). In both cases, the difference between manual and automatic contouring was not significant ($p = 0.27$; $p = 0.44$).

For the CRT-only patients, the median ΔADC was 1.9% (0.8–3.4%) and it remained stable throughout treatment (Table 2). For IC + CRT patients, the median ΔADC was 3.3% (1.6–8.0%) and increased during treatment. The difference between ADC from the automatically generated contour and that from the manual contour was significantly not normally distributed ($p < 0.001$ for CRT-only and $p = 0.009$ for IC + CRT data), justifying the use of the signed-rank Wilcoxon test. There was no significant difference between the ADCs obtained by the network and the expert observers for the CRT-only patients, with $p = 0.20$ and median difference of −0.2%. For IC + CRT patients, the ADCs were significantly ($p < 0.001$) lower, with a median decrease of 2.4% compared to the expert observer.

We found that the median DSC (over all patients and time-points) decreased as function of resolution, with DSCs of 0.83 at 2.0 mm throughout, 0.81 (2.5 mm), 0.82 (3.0 mm), 0.81 (3.5 mm), 0.79 (4.0 mm) and 0.78 (4.5 mm) to 0.77 at 5.0 mm.

3.2. MR-Linac dataset

In the fully independent MR-Linac test dataset, the DSC was slightly lower at 0.80 (1st-3rd quantile: 0.75–0.82), with ΔADC of 4.0% (0.6–9.1%). Fig. 4 highlights the best, median and worst contour in these patients, respectively. Note that the network had not seen any MR-Linac data before this evaluation and that none of the network parameters were tweaked.

4. Discussion

We have successfully trained a 3D convolutional neural network to automatically contour metastatic lymph nodes on DW-images of HNC patients throughout RT. There was no significant difference between the performance of our algorithm and expert observers. Furthermore, we demonstrated the success of our network on an independent and highly relevant dataset of DW-images obtained on an MR-Linac.

We found that for the CRT-only patients, the contouring remained stable throughout the first two weeks of treatment, during which treatment-induced changes are considerable [24]. This would indicate that our auto-contouring framework will be accurate throughout treatment, which is essential when studying treatment response.

The worst performing contour, depicted in Fig. 1, third column, seemed to only contour the edges of the lymph node. On closer inspection, it showed the network had accurately contoured the lymph node but that the post-processing step had failed, as the randomly selected seed point was selected close to the lymph node's edge. We felt further fine-tuning of the post-processing kit might lose generalizability. Instead, we believe this case can easily be noticed by an observer and can be corrected for by repeating the contour while using a different seed point. In a rerun, we found that selecting a more sensible seed point resulted in a DSC of 0.77 instead of 0.41.

Clinically relevant changes in the ADC throughout the treatment of lymph nodes are in the order of 15–19% [4], depending on the time of assessment. It is promising to see that the difference in ADC between the auto-contour and an observer (medians of 1.9% for CRT-only, 3.3% for IC + CRT) was substantially smaller than these clinically relevant changes. Note that both our ADC and the ADC from [4] were calculated using b-values from < 150 s/mm^2, and hence both ADCs can include some intravoxel incoherent motion effects.

Our network performed poorly on the IC + CRT patients, with lower DSCs, larger ΔADCs and a significant bias. It would appear that IC caused the boundaries of tumours to be less well detectable/more diffuse, as depicted in Fig. 2. Potentially, a network that only trains on post-IC patients would perform better in this subgroup. However, we were not able to test this hypothesis with the limited number of IC + CRT patients in our dataset.

We are unaware of any other CNNs being used for automated contouring of metastatic lymph nodes of HNC patients using DW MRI data, impairing a direct comparison of our results to literature. In the past, neural networks were used to contour nasopharyngeal carcinoma on T2-weighted MR-image, where a median DSC of 0.79 was reported [13], which is lower than in our study. Atlas-based attempts were reported, particularly for contouring of organs at risk (e.g. [25,26]), which achieved DSCs in the order of 0.74–0.85 on MR-images. Note that such approaches are more challenging in metastatic lymph nodes due to the huge variation of potentially involved nodal levels (although this has been done for CT [27–29]). Many automated contouring algorithms were developed for contouring organs at risk (e.g. [12,17,30–33]), tumours (e.g. [28,34]) and lymph nodes (e.g. [27–29,34]).

Clinicians had all the available MRI information present for contouring, whereas our network only saw the b = 50 s/mm^2 image. It is promising to see that, despite not seeing the additional images, our network was similarly effective at contouring. Potentially, additional channels containing these images could be added to the network [35]. However, these images were not aligned to each other (e.g. motion, deformations due to field heterogeneity) and in exploratory work (data not show), this reduced the network's performance compared to a single channel. Furthermore, adding modalities reduces the network's flexibility, as it requires the additional images, or needs strategies to deal with missing data [36].

Our network performed slightly worse (median DSC 8.0% lower) in the independent MR-Linac dataset compared to the diagnostic data. This is not fully explained by the lower resolution alone (which showed 2.4% decrease at this resolution). After closer examination, we observed that the test dataset consisted of a particularly large number of neighbouring lymph nodes. The network mainly made errors at borders between those neighbouring lymph nodes. The blurry edges (due to lower resolution) of such neighbouring lymph nodes caused our algorithm to predict both nodes as a single lymph node. Subsequently, the automated post-processing step developed to separate neighbouring lymph nodes failed due to a large overlap of the borders, which was not typically seen in the diagnostic MR-images. The only two lymph nodes without neighbouring lymph nodes were contoured accurately, with DSCs of 0.89 and 0.91. Finally, we only used a limited amount of data augmentation (flipping and shifting). In a preliminary study, we found that additional augmentation did not improve the performance of the U-net in our diagnostic MRI dataset (results not shown). As the MR-Linac data was an independent dataset, we wanted to evaluate the U-net without further optimisation. However, it is known [37] that data augmentation can increase the generalizability of networks and, hence, we also retrained the network with on-the-fly augmentation of the diagnostic dataset (scaling, sheering, rotation, mirroring, shifting and blurring), which at first attempt already substantially improved the contours with DSC of 0.82 (0.81–0.88) and ΔADC of 2.1% (1.5–4.0%) when tested on the MR-Linac data.

Neural networks often lack generalisability [38]. Typically, when the MRI acquisition protocol changes, one will have to retrain the network using new data obtained with the new protocol. In the current study, we instead modified the diagnostic MRI data to mimic MR-Linac data, by blurring. In other work [17], we have shown that when the new imaging protocol is substantially different (i.e. MRI instead of CT), one can use cycle-GANs for this, too. However, these approaches still require retraining the neural network for the newly generated data.

Our dataset consisted of repeated scans throughout treatment. In our approach, we contoured from scratch on each repeated scan. As the

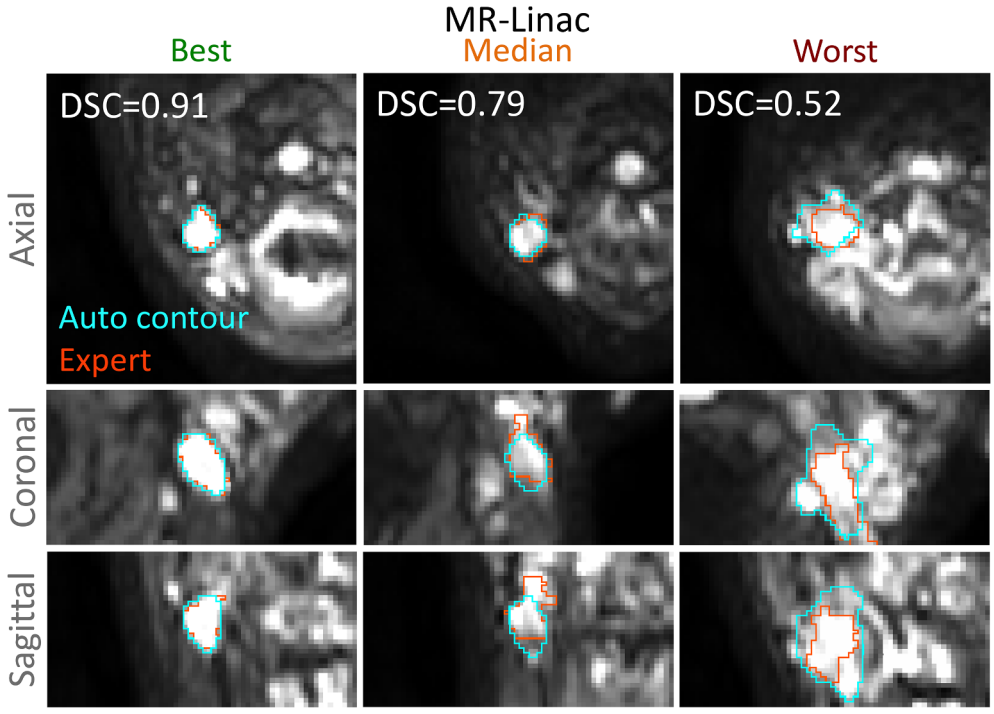

Fig. 4. DW-images (b = 50 s/mm^2) with the best, median and worst performing auto-contours of the MR-Linac data.

baseline scans were contoured most accurately one could focus on harnessing these contours for improving subsequent scans in future research.

A limitation was that we did not evaluate the performance in small (< 0.8 cm^3) lymph nodes. Our evaluation metric, DSC, is not very reliable for small volumes [39], as they are less likely to overlap (see e.g. Fig. 3 right). Including these nodes introduced several outliers that greatly biased the results to mainly reflect those outliers. However, we believe the network can still perform equally well as expert observers in such lymph nodes.

It would also be interesting to evaluate the ADC in the primary tumour. However, HNC has a large variety of locations and shapes and we felt our dataset was insufficient to train for contouring of the primary tumour. We believe that with this work, we have shown the capability of deep neural networks to contour relevant pathologies in HNC patients and we are convinced that once a larger dataset becomes available, the network should be able to learn the contouring of primary tumours in the future.

Our network required a seed point to determine a bounding box. This is, to some extent, similar to radiation oncologists, who often rely upon additional medical information such as cytology or radiology report to determine which lymph nodes are involved. We, therefore, interpret the seed selection by a click as a translation from the medical terminology to a numerical input for the network. Note that in repeated measures, the click could potentially be replaced by registration to the previous acquisition.

We believe that clinical implementation of automated contouring to obtain quantitative parameters from an image should be relatively straight forward. Despite the network being a black box, the resulting contours are easily visually checked. Even if all contours would initially require visual quality assurance, this would still be a substantial time saver compared to full contouring from scratch.

In conclusion, we have trained a deep neural network that can accurately contour metastatic lymph nodes on DW-images. The network can reduce the workload in DW MRI studies and potentially improve contouring consistently. This will particularly be beneficial for longitudinal studies that collect multiple DW-images, such as daily imaging on an MR-Linac.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The Institute of Cancer Research and the Royal Marsden NHS Foundation Trust are part of the MR-Linac consortium. Dr. Harrington received personal fees (disclosed travel payments and honoraria) from Elekta outside the submitted work. Since finishing this work, Jennifer Kieselmann has become an employee at Varian Medical Systems, Inc. There are no other potential conflicts of interest to declare.

Acknowledgements

We would like to thank Dualta McQuaid for his helpful scripts for exporting MRI data from Raystation.

Funding

This work was supported by the Cancer Research UK Programme (grants C7224/A23275, and C33589/ A19727) and Oracle Cancer Trust. We acknowledge CRUK and EPSRC support to the Cancer Imaging Centre at ICR and RMH in association with MRC and Department of Health C1060/A10334, C1060/A16464 and NHS funding to the NIHR Biomedical Research Centre and the Clinical Research Facility in Imaging. This report is independent research funded by the National Institute for Health Research. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

References

[1] Gurney-Champion OJ, Mahmood F, van Schie M, Julian R, George B, Philippens MEP, et al. Quantitative imaging for radiotherapy purposes. Radiother Oncol 2020;146:66–75. https://doi.org/10.1016/j.radonc.2020.01.026.

[2] Wong KH, Panek R, Dunlop A, Mcquaid D, Riddell A, Welsh LC, et al. Changes in multimodality functional imaging parameters early during chemoradiation predict treatment response in patients with locally advanced head and neck cancer. Eur J Nucl Med Mol Imaging 2018;45:759–67. https://doi.org/10.1007/s00259-017-

3890-2.
[3] Matoba M, Tuji H, Shimode Y, Toyoda I, Kuginuki Y, Miwa K, et al. Fractional change in apparent diffusion coefficient as an imaging biomarker for predicting treatment response in head and neck cancer treated with chemoradiotherapy. AJNR Am J Neuroradiol 2014;35:379–85. https://doi.org/10.3174/ajnr.A3706.
[4] Vandecaveye V, Dirix P, De Keyzer F, Op De Beeck K, Vander Poorten V, Roebben I, et al. Predictive value of diffusion-weighted magnetic resonance imaging during chemoradiotherapy for head and neck squamous cell carcinoma. Eur Radiol 2010;20:1703–14. https://doi.org/10.1007/s00330-010-1734-6.
[5] Paudyal R, Oh JH, Riaz N, Venigalla P, Li J, Hatzoglou V, et al. Intravoxel incoherent motion diffusion-weighted MRI during chemoradiation therapy to characterize and monitor treatment response in human papillomavirus head and neck squamous cell carcinoma. J Magn Reson Imaging 2017;45:1013–23. https://doi.org/10.1002/jmri.25523.
[6] Yang Y, Cao M, Sheng K, Gao Y, Chen A, Kamrava M, et al. Longitudinal diffusion MRI for treatment response assessment: Preliminary experience using an MRI-guided tri-cobalt 60 radiotherapy system. Med Phys 2016;43:1369–73. https://doi.org/10.1118/1.4942381.
[7] van der Veen J, Gulyban A, Nuyts S. Interobserver variability in delineation of target volumes in head and neck cancer. Radiother Oncol 2019;137:9–15. https://doi.org/10.1016/j.radonc.2019.04.006.
[8] Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. Br J Radiol 2019;92:1–12. https://doi.org/10.1259/bjr.20190001.
[9] LeCun Y, Bengio Y. Convolutional Networks for Images, Speech, and Time-Series. Handb. brain theory neural networks, vol. 3361, 1995. https://doi.org/10.1109/IJCNN.2004.1381049.
[10] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 2015. https://doi.org/10.1007/978-3-319-24574-4_28.
[11] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 2016. https://doi.org/10.1007/978-3-319-46723-8_49.
[12] Zhu W, Huang Y, Zeng L, Chen X, Liu Y, Qian Z, et al. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. Med Phys 2019;46:576–89. https://doi.org/10.1002/mp.13300.
[13] Lin L, Dou Q, Jin YM, Zhou GQ, Tang YQ, Chen WL, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. Radiology 2019;291:677–86. https://doi.org/10.1148/radiol.2019182012.
[14] Welsh L, Panek R, McQuaid D, Dunlop A, Schmidt M, Riddell A, et al. Prospective, longitudinal, multi-modal functional imaging for radical chemo-IMRT treatment of locally advanced head and neck cancer: The INSIGHT study. Radiat Oncol 2015;10:112. https://doi.org/10.1186/s13014-015-0415-7.
[15] Chollet F. Keras 2015. available at https://github.com/fchollet/keras.
[16] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning. Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016, 2016, p. 265–83.
[17] Kieselmann J, Gurney-Champion OJ, Hin B, Nill S, Fuller CD. Oelfke U. Cross-modality deep learning: contouring of MRI data from annotated CTs only. MR in RT 2019:16https://www.mrinrt2019.com/.
[18] Dalca AV, Guttag J. Sabuncu MR. Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2018:9290–9. https://doi.org/10.1109/CVPR.2018.00968.
[19] Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 2017. https://doi.org/10.1007/978-3-319-67558-9_28.
[20] Kingma DP, Adam BaJ. A Method for Stochastic Optimization. ArXiv Prepr ArXiv14126980 2014.
[21] Strivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res 2014;15:1929–58.
[22] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 32nd Int. Conf. Mach. Learn. ICML 2015;1(2015):448–56.
[23] Van Der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. Scikit-image: Image processing in python. PeerJ 2014;2014(2):e453https://doi.org/10.7717/peerj.453.
[24] Sanguineti G, Ricchetti F, Wu B, Agrawal N, Gourin C, Agbahiwe H, et al. Volumetric change of human papillomavirus-related neck lymph nodes before, during, and shortly after intensity-modulated radiation therapy. Head Neck 2012;34:1640–7. https://doi.org/10.1002/hed.21981.
[25] Kieselmann JP, Kamerling CP, Burgos N, Menten MJ, Fuller CD, Nill S, et al. Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region. Phys Med Biol 2018;63.. https://doi.org/10.1088/1361-6560/aacb65.
[26] Cheng G, Yang X, Wu N, Xu Z, Zhao H, Wang Y, et al. Multi-atlas-based segmentation of the parotid glands of MR images in patients following head-and-neck cancer radiotherapy. Med. Imaging 2013 Comput. Diagnosis, vol. 8670, 2013, p. 86702Q. https://doi.org/10.1117/12.2007783.
[27] Stapleford LJ, Lawson JD, Perkins C, Edelman S, Davis L, McDonald MW, et al. Evaluation of Automatic Atlas-Based Lymph Node Segmentation for Head-and-Neck Cancer. Int J Radiat Oncol Biol Phys 2010;77:959–66. https://doi.org/10.1016/j.ijrobp.2009.09.023.
[28] Zhang T, Chi Y, Meldolesi E, Yan D. Automatic Delineation of On-Line Head-And-Neck Computed Tomography Images: Toward On-Line Adaptive Radiotherapy. Int J Radiat Oncol Biol Phys 2007;68:522–30. https://doi.org/10.1016/j.ijrobp.2007.01.038.
[29] Chen A, Deeley MA, Niermann KJ, Moretti L, Dawant BM. Combining registration and active shape models for the automatic segmentation of the lymph node regions in head and neck CT images. Med Phys 2010;37:6338–46. https://doi.org/10.1118/1.3515459.
[30] Raudaschl PF, Zaffino P, Sharp GC, Spadea MF, Chen A, Dawant BM, et al. Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. Med Phys 2017;44:2020–36. https://doi.org/10.1002/mp.12197.
[31] Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys 2017;44:547–57. https://doi.org/10.1002/mp.12045.
[32] Tsuji SY, Hwang A, Weinberg V, Yom SS, Quivey JM, Xia P. Dosimetric Evaluation of Automatic Segmentation for Adaptive IMRT for Head-and-Neck Cancer. Int J Radiat Oncol Biol Phys 2010;77:707–14. https://doi.org/10.1016/j.ijrobp.2009.06.012.
[33] van der Veen J, Willems S, Deschuymer S, Robben D, Crijns W, Maes F, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. Radiother Oncol 2019;138:68–74. https://doi.org/10.1016/j.radonc.2019.05.010.
[34] Street E, Hadjiiski L, Sahiner B, Gujar S, Ibrahim M, Mukherji SK, et al. Automated volume analysis of head and neck lesions on CT scans using 3D level set segmentation. Med Phys 2007;34:4399–408. https://doi.org/10.1118/1.2794174.
[35] Aerts HJWL, Lahaye MJ, Parmar C, Lambregts DMJ, Peters NHGM, Trebeschi S, et al. Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. Sci Rep 2017;7:1–9. https://doi.org/10.1038/s41598-017-05728-9.
[36] Sharma A, Hamarneh G. Missing MRI Pulse Sequence Synthesis using Multi-Modal Generative Adversarial Network. IEEE Trans Med Imaging 2019. https://doi.org/10.1109/tmi.2019.2945521.
[37] Perez L, Wang J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. Convolutional Neural Networks Vis. Recognit 2017:11.
[38] Azulay A, Weiss Y. Why do deep convolutional networks generalize so poorly to small image transformations? J Mach Learn Res 2019;20:1–25.
[39] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging 2015;15:29. https://doi.org/10.1186/s12880-015-0068-x.