Author/s:
Korte, JC;Hardcastle, N;Ng, SP;Clark, B;Kron, T;Jackson, P

Title:
Cascaded deep learning-based auto-segmentation for head and neck cancer patients: Organs at risk on T2-weighted magnetic resonance imaging

Date:
2021-12-01

Citation:
Korte, J. C., Hardcastle, N., Ng, S. P., Clark, B., Kron, T. & Jackson, P. (2021). Cascaded deep learning-based auto-segmentation for head and neck cancer patients: Organs at risk on T2-weighted magnetic resonance imaging. Medical Physics, 48 (12), pp.7757-7772. https://doi.org/10.1002/mp.15290.

Persistent Link:
https://hdl.handle.net/11343/299147

Title:

Cascaded deep-learning based auto-segmentation for head and neck cancer patients: organs at risk on T2 weighted magnetic resonance imaging

Running Title:

MRI DL auto-segmentation of OAR in HNC

Authors & Affiliations:

James C Korte[1,2], Nicholas Hardcastle[1,3,4], Sweet Ping Ng,[5,6], Brett Clark[1,2], Tomas Kron[1,4], Price Jackson[1,4]

[1]Department of Physical Science, Peter MacCallum Cancer Centre, Melbourne, Australia
[2]Department of Biomedical Engineering, University of Melbourne, Melbourne, Australia
[3]Centre for Medical Radiation Physics, University of Wollongong, Wollongong, Australia
[4]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia
[5]Department of Radiation Oncology, Peter MacCallum Cancer Centre, Melbourne, Australia
[6]Department of Radiation Oncology, Olivia Newton-John Cancer and Wellness Centre, Austin Health, Melbourne, Australia

Corresponding Author:

Dr James C Korte

James.Korte@petermac.org

Peter MacCallum Cancer Centre

305 Grattan Street

Melbourne, Victoria

3000 Australia

Previous publication of manuscript text or data:

Early results of the low-resolution auto-segmentation method from this paper were presented at the Australian Magnetic Resonance in Radiation Therapy annual meeting in 2019.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.15290.

Abstract (323/500 words)

Purpose: To investigate multiple deep learning methods for automated segmentation (auto-segmentation) of the parotid glands, submandibular glands, level II and level III lymph nodes on magnetic resonance imaging (MRI). Outlining radiosensitive organs on images used to assist radiation therapy (radiotherapy) of patients with head and neck cancer (HNC) is a time-consuming task, in which variability between observers may directly impact on patient treatment outcomes. Auto-segmentation on computed tomography imaging has been shown to result in significant time reductions and more consistent outlines of the organs at risk.

Methods: Three convolutional neural network (CNN) based auto-segmentation architectures were developed using manual segmentations and T2 weighted MRI images provided from the AAPM RT-MAC challenge dataset (n=31). Auto-segmentation performance was evaluated with segmentation similarity and surface distance metrics on the RT-MAC dataset with institutional manual segmentations (n=10). The generalisability of the auto-segmentation methods was assessed on an institutional MRI dataset (n=10).

Results: Auto-segmentation performance on the RT-MAC images with institutional segmentations was higher than previously reported MRI methods for the parotid glands (dice: 0.860±0.067, mean surface distance: 1.33±0.40 mm) and the first report of MRI performance for submandibular glands (dice: 0.830±0.032, mean surface distance: 1.16±0.47 mm). We demonstrate that high-resolution auto-segmentations with improved geometric accuracy can be generated for the parotid and submandibular glands by cascading a localiser CNN and a cropped high-resolution CNN. Improved mean surface distances were observed between automatic and manual segmentations of the submandibular glands when a low-resolution auto-segmentation was used as prior knowledge in the second stage CNN. Reduced auto-segmentation performance was observed on our institutional MRI dataset when trained on external RT-MAC images; only the parotid gland auto-segmentations were considered clinically feasible for manual correction (dice: 0.775±0.105, mean surface distance: 1.20±0.60 mm).

Conclusion: This work demonstrates that CNNs are a suitable method to auto-segment the parotid and submandibular glands on MRI images of patients with HNC, and that cascaded CNNs can generate high resolution segmentations with improved geometric accuracy. Deep learning methods may be suitable for auto-segmentation of the parotid glands on T2 weighted MRI images from different scanners, but further work is required to improve the performance and generalisability of these methods for auto-segmentation of the submandibular glands and lymph nodes.

KEYWORDS (3-5):

Magnetic Resonance Imaging, Image Segmentation, Head and Neck Cancer, Organs at Risk, Convolutional Neural Networks

Introduction

Head and neck cancer (HNC) is diagnosed in 887,659 patients globally every year [1], with a potential 74% of these patients able to benefit from radiation therapy [2]. Many of these patients are treated with intensity modulated radiotherapy (IMRT) as it allows the precise delivery of radiation to tumour while minimising the dose to surrounding healthy tissue or organs at risk (OARs). In particular, IMRT is beneficial for reducing the radiation delivered to the parotid glands, which reduces the incidence of xerostomia (dry mouth) [3], a common toxicity that affects the quality of life of HNC patients.

The quality of a radiotherapy plan in HNC is known to impact overall survival [4]. A key step in the planning process is to segment a patient image; spatially outlining tumour regions to target with radiation and defining OARs to spare healthy tissue and reduce the associated side effects. Segmentation of HNC patient images is time consuming, the average reported time taken from 2.7 to 3.0 hours [5] with high inter-observer variability (IOV) reported between segmentations defined by multiple clinicians [6-8]. Computational methods have been proposed to automate the process, generating an automatic segmentation (auto-segmentation) directly from patient imaging. Head and neck auto-segmentation studies have demonstrated significant time savings in comparison to manual segmentation [9-12] and reduced inter-observer variability [11,13].

Auto-segmentation has been widely explored in HNC using computed tomography (CT) imaging, the majority of studies applying atlas-based methods, some improving performance with a machine learning based refinement or directly through deep learning methods [5]. Magnetic resonance imaging (MRI) has demonstrated superior soft tissue visualisation to CT in the head and neck region [14] and is taking a larger role in radiotherapy with the advent of MRI only radiotherapy planning and treatment systems with integrated MRI [15,16]. Adaptive radiotherapy [17], where the treatment plan is adapted throughout treatment to account for anatomical changes, has been recommended to make full use of the enhanced soft-tissue visualisation these systems provide at treatment time. MRI based adaptive workflows have reported segmentation as one of the most time consuming steps [18], making

it an attractive candidate for auto-segmentation methods which may reduce adaptive treatment times and increase patient access to this new technology. There are limited MRI auto-segmentation studies for OARs in HNC, the majority focusing on the parotid glands with atlas-based methods [19-22] or deep learning methods [23,24]. Previous deep learning studies have demonstrated improved auto-segmentation performance with multi-modality methods (CT & MRI images) [23], no significant performance difference for multi-contrast MRI methods (T1 & T2 weighted MRI images) and no significant performance difference between two dimensional (2D) and three dimensional (3D) methods [24].

In this study we explore deep learning based methods to auto-segment HNC organs at risk on T2 weighted MRI, including the parotid glands, submandibular glands, level II and level III lymph nodes. We investigate the relative performance of three 3D convolutional neural network (CNN) based architectures (Figure 1); a single low-resolution CNN, cascading multiple CNNs to generate higher resolution auto-segmentations, and a similar cascade of CNNs with a low-resolution auto-segmentation as prior knowledge to improve accuracy. The auto-segmentation networks are developed using an open-source dataset (RT-MAC) [25], made available as part of an American Association of Physicists in Medicine (AAPM) grand challenge, and accuracy is reported with segmentation similarity and surface distance metrics. As the RT-MAC dataset does not include publicly available testing segmentations, we assessed the auto-segmentation performance of the networks against manual segmentations defined by an expert radiation oncologist from our institute. Auto-segmentation performance was compared to inter-observer variability (IOV) between RT-MAC manual segmentation and institutional manual segmentations of the RT-MAC validation dataset. To assess the potential use of the auto-segmentation methods at our institute, we explore the generalisability of the proposed methods to auto-segment an institutional dataset of T2 weighted MRI images and manual segmentations.

Materials and Methods

Study Cohorts

Two annotated head and neck MRI datasets were used in this study, a publicly available dataset to develop and assess the performance of three auto-segmentation methods and an institutional dataset to evaluate the generalisability of the auto-segmentation methods. The publicly available dataset consists of a training set of 31 patients, with MRI images and manual segmentations, and a test set of 12 patients with MRI images only. A radiation oncologist from our institute manually segmented 5 of the public training set and 10 of the public test set. Our institutional dataset consists of 10 patients and includes MRI images and manual segmentations. A summary of the study data is shown in Figure 2.

The publicly available dataset, the radiotherapy MRI auto-contouring (RT-MAC) dataset [25], was released for an AAPM grand challenge during the 2019 AAPM annual meeting. The dataset includes 31 training cases (RT-MAC Train*), that include T2 weighted MRI of head and neck patients in a radiotherapy treatment position and associated manual segmentations for eight organs at risk; parotid glands (left/right), submandibular glands (left/right), level II and level III lymph nodes (left/right). Manual segmentation was performed by a radiation oncologist with over 10 years of clinical experience [25]. For our study, the RT-MAC Train* dataset was split into a training dataset (RT-MAC Train) to train the CNNs and a validation dataset (RT-MAC Validate) used to select the best performing networks from all epochs of training. We define the function of the datasets as follows; a training dataset for training the model, a validation dataset for selecting the best performing model during training, a test dataset which is not seen by the model during training and is used to evaluate the performance of the selected model. The RT-MAC dataset has 12 testing cases (RT-MAC Test*) which include a T2 weighted MRI but have no publicly available segmentations. An experienced radiation oncologist, from our institute, manually segmented the eight organs at risk on 10 cases of the RT-MAC Test* dataset (cases 3-12) to form a testing dataset (RT-MAC Test) to assess auto-

segmentation performance. To measure the IOV between public and institutional manual segmentations, the radiation oncologist from our institute manually segmented the 5 cases of the RT-MAC Validate dataset.

Our institutional MRI dataset is composed of diagnostic T2 weighted MRI of 10 patients with head and neck cancer. Approval to conduct this retrospective study was provided by our local ethics committee. An experienced radiation oncologist manually segmented the same eight organs at risk as the RT-MAC dataset.

MRI Imaging Protocols

The RT-MAC MRI images (Figure 3a) were acquired on a Siemens 1.5T Aera MRI scanner (Siemens Healthcare, Erlangen, Germany) with the patient in a radiotherapy position. T2 weighted images were acquired with a 2D turbo-spin echo sequence with parameters: flip angle = 90°, refocusing angle = 180°, echo time = 80 ms, repetition time = 4800 s, echo train length = 15, averages = 1, slice thickness = 2 mm, 120 slices, in-plane resolution = 0.5 mm, field of view = 256 mm x 256 mm, bandwidth = 300 Hz/Pixel.

Our institutional MRI images (Figure 3b) were acquired with two Siemens 1.5T MRI scanners (Siemens Healthcare, Erlangen, Germany), an Aera and a Sola. On both scanners, T2 weighted images were acquired with a Dixon turbo-spin echo sequence: flip angle = 90°, refocusing angle = 140-150°, echo time = 55-71 ms, repetition time = 4980-5220 s, echo train length = 8-19, averages = 2, slice thickness = 4 mm, 40-42 slices, in-plane resolution = 0.5625 mm, field of view = 180.0 mm x 157.5-180.0 mm, bandwidth = 275-300 Hz/Pixel. The in-phase images were used in this study and were retrospectively selected for similar image contrast to the RT-MAC data but have different spatial parameters, such as a smaller field of view and thicker slices, see Supplementary Table 1 for a detailed comparison of imaging parameters.

Auto-Segmentation Models

The U-net architecture [26] was selected as it has shown good performance for segmentation of medical images, including the auto-segmentation of organs at risk on CT images of head and neck cancer patients [27]. A three dimensional (3D) variant of the architecture, or 3D U-net [28], was chosen as 3D CNNs have demonstrated improved performance over 2D or 2.5D CNNs for segmentation tasks on MRI images [29]. A limitation of 3D networks is their large graphical processing unit (GPU) memory requirement, which is often met by downsampling the input image, with a corresponding loss of high frequency information from the original image which may degrade the segmentation performance. To overcome this, multi-stage or cascaded CNNs have been used, first identifying a bounding box [30,31] or candidate region [32,33] on a reduced resolution image or a lower dimension image, then performing the segmentation task at the original image resolution over a reduced spatial extent. To explore these methods for MRI segmentation, we developed three auto-segmentation models as shown in Figure 1; a low-resolution 3D U-net, a cascaded high-resolution 3D U-net and a cascaded high-resolution 3D U-net with prior knowledge (the label prediction from the low-resolution 3D U-net). For all three models, organ specific auto-segmentation networks were developed for each of the eight organs at risk.

Convolutional neural network architectures

All networks were implemented, trained and tested in Python (v3.5) using Keras (v2.1) [34] with a Tensorflow backend (v1.8) [35] using CUDA (v9). We used a modified 3D U-net architecture developed at our institute for segmenting the kidneys on CT images[36] which was previously adjusted to fit the memory specification of the GPUs available to our department (NVIDIA Tesla P100/12 Gb). Convolution layers consisted of a 3x3x3 kernel, followed by batch normalization [37] and rectified linear unit activation layers [38]. On the encoding arm of the U-net, downsampling to lower resolutions was achieved with 2x2x2 max pooling layers. On the decoding arm of the U-net, concatenation and upsampling operations were performed to return to the input image resolution. The final layer of

the network is a softmax activation layer that generates a voxel-wise probability of two classes, an organ at risk or background. See Supplementary Figure 1 and Supplementary Table 2 for details of the network architecture.

Image pre-processing

MRI images were spatially and intensity normalised to account for acquisition differences between the RT-MAC dataset and our institutional dataset, and to enable future application of the auto-segmentation models to datasets from other sources. Whilst the image contrast in the two datasets was similar, intensity normalisation was required as the absolute range of intensities differed between datasets. Image pre-processing was applied to all MRI data and was achieved with the following ordered image operations; datatype casting to 64 bit floating point numbers, intensity normalisation to a zero mean with unity variance, N4 bias correction (10 iterations per level, 5 resolution levels) [39] using a non-background mask, intensity normalisation to a zero mean with unity variance, intensity rescaling (from [-3.0, 15.0], to [0, 65000]), datatype casting to a 16 bit unsigned integer and linear spatial interpolation to match the resolution and field of view of the RT-MAC dataset. The non-background mask used in the N4 bias correction was constructed by intensity thresholding the first normalised MRI image (intensity>0.5) followed by a binary hole filling operation. The second intensity normalisation step accounted for intensity changes of the N4 bias correction and provided a repeatable intensity range for the intensity rescaling operation. The image pre-processing was performed with SimpleITK [40] on a desktop computer (Intel i7-4770S, 8 Cores, 3.10 GHz, 32 GB Ram).

Network training strategy

The networks were trained on MRI image and segmentation pairs from the RT-MAC Train dataset (n=26), using the RT-MAC Validation dataset (n=5) to select the highest performing network. All 3D U-nets were trained with a dice loss function [41] and Adam optimisation (learning rate= 0.0001, β_1=0.9, β_2=0.999) [42] over 300 epochs with a single volumetric image batch size. The dice loss was

calculated between the predicted organ at risk probability map, $p_i \in P$, and the ground truth binary segmentation, $g_i \in G$, and is defined as,

$$D_{loss} = \frac{-2 \sum_i^N p_i g_i + \epsilon}{\sum_i^N p_i + \sum_i^N g_i + \epsilon},$$

where N is the number voxels in the segmentation volume and ϵ is a smoothing factor set to 1. To reduce overfitting in the training process, data augmentation was performed and the network from the epoch with the best validation dice loss was selected. The pre-processed MRI data was augmented during training with random intensity rescaling (85% - 115% of pre-processed intensity) and random spatial rotations and translations in the sagittal (± 22.5°, ±2.5mm), coronal (± 7.5°, ±2.5mm) and transverse (± 7.5°, ±2.5mm) planes. We used a larger sagittal rotation range to reflect the neck tilt angles observed between the RT-MAC immobilised radiotherapy position and the institutional diagnostic MRI position. An augmented image intensity range was selected to account for variability in the intensity normalisation during pre-processing, due to differences in water and lipid composition in patient images.

Low-resolution Networks

The low-resolution networks required downsampled MRI data as the 3D volume at native resolution would exceed the available GPU memory during training, and were trained on downsampled MRI image and segmentation pairs. The native RT-MAC voxel size of 0.5 x 0.5 x 2.0 mm was downsampled to 2.0 x 2.0 x 2.5 mm, using linear interpolation for the image data and a nearest neighbour interpolation for the segmentations. The dice loss, used to train the low-resolution networks, was calculated between the low-resolution predicted probability and the downsampled ground truth segmentation. Conversely, the similarity and distance metrics described in section 3.3 and reported in the results section were calculated between the original ground truth segmentation and an upsampled predicted segmentation from the low-resolution network.

Localiser Networks

The localiser networks are the first stage of the cascaded networks, and locate the centre of an OAR of interest. These networks were trained in isolation with a different split of the RT-MAC Train* dataset (train=18, validate=13) to introduce some imperfection in the localisation data used during the training of the second stages in the cascade networks. Similar to the low-resolution networks, the localiser networks were trained on downsampled MRI image and segmentation pairs. As the extent of the predicted segmentation is more important than edge accuracy, the segmentations used during training were dilated with a six voxel radius. The centroid of the crop region, to be used by the cascaded networks, was calculated as the centre of the extent of the predicted localiser segmentation.

High-resolution Cascaded Networks

The first cascaded networks were trained on MRI image and segmentation pairs at the original resolution, cropped to the region defined by a localiser network. The predicted high-resolution cropped segmentation was zero-padded to match the spatial extent of the original image. The extent of the crop region for each organ at risk was defined as the maximum extent observed in the RT-MAC Train* dataset and then increased in size within the memory limits of the GPU (see Supplementary Table 2)

High-resolution Cascaded Networks with Prior Knowledge

To incorporate global information, the second cascaded networks were trained on similar data to the first cascaded networks, with the addition of a predicted segmentation from a low-resolution network, which was cropped to a region defined by a localiser network. The predicted high-resolution cropped segmentation was zero-padded to match the spatial extent of the original image.

Auto-Segmentation Performance

The performance of the auto-segmentation methods was evaluated with the dice similarity coefficient [43], the Hausdorff distance [44] and the mean surface distance, also referred to as the

average symmetric surface distance [45]. To transform the predicted voxel-wise probability map, P, which is output from the softmax layer of each CNN into a binary OAR segmentation, $\hat{P}$, we used a probability threshold of greater than 0.45; a value adapted from our previous work[36]. The dice similarity coefficient is defined as,

$$DSC = 2\frac{|\hat{P} \cap G|}{|\hat{P}| + |G|},$$

where G is the ground truth binary OAR segmentation. The Hausdorff distance is defined as,

$$HD = \max\left(\max_{p \in \hat{P}} h(p, G), \max_{g \in G} h(g, \hat{P})\right),$$

where, $h(a, B) = \min_{b \in B}\|a - b\|$, and $\|\cdot\|$ is a Euclidean norm. The mean surface distance is defined as,

$$MSD = \frac{\sum_{p \in \partial \hat{P}} h(p, \partial G) + \sum_{g \in \partial G} h(g, \partial \hat{P})}{|\partial \hat{P}| + |\partial G|},$$

where $\partial \hat{P}$ is the set of voxels on the boundary of the predicted OAR segmentation and ∂G is the set of voxels on the boundary of the ground truth OAR segmentation. All performance metrics were calculated on the largest connected structure of the predicted binary segmentation in Python 3.6.3 with SimpleITK 1.1.0 [40]. As per the recommendation of [46] to use complimentary metrics, we selected an overlap similarity metric (dice coefficient) to assess general accuracy and two surface based metrics to assess the average (mean surface distance) and worst case (Hausdorff distance) boundary accuracy of the segmentations.

Network architecture comparison

The time efficiency and segmentation performance of the three auto-segmentation architectures were compared. The time taken to pre-process the MRI images and to auto-segment each organ at risk was measured programmatically. The auto-segmentation performance of each network architecture was calculated on MRI images from the RT-MAC Validation and RT-MAC Test datasets,

with all performance metrics assessed against institutional manual segmentations. To test for significant performance differences ($p < 0.05$) between the three auto-segmentation methods, we used a two-tailed Wilcoxon signed-rank test on the RT-MAC Test dataset. The Wilcoxon test was calculated with SciPy (v1.7)[47] for each organ at risk, including both the left and right segmentations in the sample (n=20).

Generalisability of auto-segmentation methods

The generalisability of the proposed auto-segmentation methods were assessed on the institutional dataset. Here we define generalisability as the method's ability to segment previously unseen MRI data, specifically data from a different MRI scanner with different image acquisition parameters than the MRI data used to train the auto-segmentation method. The auto-segmentation performance of each network architecture was calculated on MRI images from the institutional datasets, with all performance metrics assessed against institutional manual segmentations. Surface based metrics were only calculated on an auto-segmentations if the organ at risk was detected (see Supplementary Table 3 for details of undetected OARs). The networks with the highest performance on the institutional dataset were compared to the highest performing networks for the RT-Mac Test dataset and against the inter-observer variability for manual segmentation. To test for significant differences ($p < 0.05$) between the model performance on MRI images from the RT-MAC Test and institutional datasets, we used a two-tailed Mann-Whitney U test as calculated with SciPy (v1.7)[47] for each organ at risk, including both the left and right segmentations in the sample (n=20).

RESULTS

Auto-Segmentation Performance

All network architectures were successfully trained to detect the specified organs at risk on the RT-MAC dataset, an example of auto-segmentations generated on the validation dataset is shown in Figure 4. Performance metrics for all networks and datasets are reported in Supplementary Table 4

(dice coefficient), Supplementary Table 5 (mean surface distance) and Supplementary Table 6 (Hausdorff distance).The expected trend of higher performance across all metrics in the training set as compared to the validation set was observed. A general trend across the OARs was observed for the dice performance of auto-segmentations generated on the RT-MAC test dataset (Figure 5); dice performance in order from highest to lowest was the parotid glands, submandibular glands, level II lymph nodes and level III lymph nodes. A similar trend was observed in the surface distance metrics, in order from highest to lowest performing; submandibular glands, parotid glands, level II lymph nodes and level III lymph nodes.

Auto-segmentation network performance on the RT-MAC test dataset, and the inter-observer variability on the RT-MAC validation dataset is shown in Table 1. The auto-segmentation performance was similar to the inter-observer variability across all metrics and for all organs at risk (dice coefficients within 0.035 of IOV, mean surface distances within 0.61 mm of IOV and Hausdorff distances within 2.0 mm of IOV). The level II lymph node auto-segmentations performed slightly higher than the IOV for the dice coefficient and mean surface distance. The Hausdorff distance performance was slightly higher than IOV for auto-segmentations of the right parotid gland. The trend in inter-observer variability across manual segmentations of OARs was identical to that observed in the auto-segmentation of the validation and training datasets for both the dice coefficient and the surface distance metrics (see Supplementary Figure 2).

The highest dice coefficient, of any network, on the RT-MAC test dataset when comparing auto-segmentations to institutional manual segmentations for each organ at risk was 0.860±0.067 (left parotid), 0.857±0.063 (right parotid), 0.830±0.032 (left submandibular), 0.785±0.123 (right submandibular), 0.708±0.053 (left level 2 lymph nodes), 0.715±0.071 (right level 2 lymph nodes), 0.561±0.100 (left level 3 lymph nodes) and 0.573±0.105 (right level 3 lymph nodes).

The lowest mean surface distance, of any network, on the RT-MAC test dataset when comparing auto-segmentations to institutional manual segmentations for each organ at risk was 1.41±0.33 mm

(left parotid), 1.33±0.39 mm (right parotid), 1.16±0.47 mm (left submandibular), 1.19±0.50 mm (right submandibular), 2.44±0.41 mm (left level 2 lymph nodes), 2.33±0.49 mm (right level 2 lymph nodes), 3.70±0.80 mm (left level 3 lymph nodes) and 3.53±1.22 mm (right level 3 lymph nodes).

The lowest Hausdorff distance, of any network, on the RT-MAC test dataset when comparing auto-segmentations to institutional manual segmentations for each organ at risk was 11.31±6.09 mm (left parotid), 9.68±4.37 mm (right parotid), 6.83±3.29 mm (left submandibular), 7.62±4.34 mm (right submandibular), 16.61±3.47 mm (left level 2 lymph nodes), 17.72±3.66 mm (right level 2 lymph nodes), 20.26±4.24 mm (left level 3 lymph nodes) and 17.91±5.98 mm (right level 3 lymph nodes).

Network Architecture Comparison

Auto-segmentation performance

A comparison of auto-segmentation performance on the RT-MAC test dataset for the low-resolution, high-resolution and high-resolution with prior knowledge methods is shown in Figure 5. The high-resolution network improved performance in comparison to the low-resolution network, with significantly higher dice coefficients for the submandibular glands and level III lymph nodes. The high-resolution network with prior knowledge improved performance in comparison to the low-resolution network, with significantly higher dice coefficients for the parotid and submandibular glands and significantly smaller mean surface distances for the submandibular glands. In the high-resolution networks, the addition of prior knowledge changed auto-segmentation performance, with improved dice coefficients and mean surface distances for the submandibular glands, but a lower performing Hausdorff distance for the level II lymph nodes.

Time efficiency

The time taken to auto-segment the eight organs at risk on MRI of a head and neck cancer patient increased with network architecture complexity, taking on average 0.7 minutes (low-resolution

network), 2.6 minutes (high-resolution network) and 3.5 minutes (high-resolution network with prior knowledge) as detailed in Table 2.

Generalisability of Auto-Segmentation Methods

The methods developed on the RT-MAC dataset were able to auto-segment the majority of organs at risk on the institutional MRI images, an example of auto-segmentation is shown in Figure 6. The quality of the auto-segmentations generated on the institutional data (Table 1) was poorer across the majority of metrics in comparison to auto-segmentations generated on the RT-MAC test dataset (Figure 7). The dice performance on institutional data was significantly poorer for all auto-segmentation models across all organs-at-risk. The mean surface distance was significantly larger on institutional data for the majority of cases, with no significant difference observed for the submandibular glands with a low-resolution method, no significant difference for the parotid glands with a high-resolution network and no significant difference for the level II lymph nodes with either of the high-resolution networks. The Hausdorff distance was significantly larger on institutional data for the majority of cases, with no significant difference observed for the level II and level III lymph nodes with the low-resolution method. Improved distance performance was observed in one case, with significantly smaller means surface distance in the level II lymph nodes with the low-resolution method. On the institutional MRI data, only the parotid gland auto-segmentation may be considered suitable for clinical use with manual correction, with a mean dice score of 0.730/0.775 (left/right), a mean surface distance 1.64/1.20 mm and a Hausdorff distance of 12.88/14.00 mm.

Discussion

A performance comparison of the best results from our methods on the RT-MAC testing dataset and previously published HNC studies for each OAR are detailed in Table 3. The surface distance metrics are reported in millimetres to two significant figures to allow direct comparison with previous studies. For auto-segmentation of the parotid glands on MRI, our methods have the highest

performance with respect to dice coefficient (0.860±0.067), mean surface distance (1.33±0.40 mm) and Hausdorff distance (9.68±4.37 mm). One MRI based auto-segmentation method [19] was excluded from this comparison as it has an additional requirement, a planning MRI and manual segmentation to train a patient specific atlas and a patient specific support vector machine. Our proposed parotid auto-segmentation methods have comparable surface distance performance and slightly lower dice performance than those reported for CT [48,49]. Our submandibular auto-segmentation has higher dice performance than two CT studies [48,50], is comparable to a CT atlas-based method [51] and has slightly lower performance than a CT deep learning method [49]. For level II lymph nodes our auto-segmentation methods have better mean surface distance performance but lower dice performance than a CT and MRI study [21]. For level III lymph nodes our methods have higher dice and surface distance performance than a previous CT atlas-based method [52].

The observed trend in auto-segmentation dice performance, on the RT-MAC test dataset across organs at risk from highest to lowest (parotid glands, submandibular glands, level 2 lymph nodes, level 3 lymph nodes) was previously reported using a single atlas based method on CT [53]. Auto-segmentation of the parotid glands having a higher dice coefficient than auto-segmentation of the submandibular glands has been observed in CT based deep learning studies [27,49,54]. These trends are mirrored in the IOV of manual segmentations (Supplementary Figure 2) which suggests that the parotid glands may be inherently less challenging to reproducibly segment than the submandibular glands. The lymph node levels have the poorest auto-segmentation performance and highest IOV of the investigated OARs, this may be due to the challenging task of separating the individual nodal levels and is the focus of ongoing efforts to standardise practise [55,56].

Network architecture impacts auto-segmentation performance. Our high-resolution cascaded networks had significantly improved dice performance for the majority of organs-at-risk and significantly smaller mean surface distances for the submandibular glands than a low-resolution single stage network. The lack of significant improvement in mean surface distance for the parotid

glands, when comparing low and high resolution networks, may indicate that low-resolution methods are sufficient when auto-segmenting organs with a less complex surface geometry. Whilst performance improvements due to a similar coarse-fine method have been reported for OAR segmentation on CT of HNC [33], those improvements were unclear in the parotid glands as there was improved surface distance performance (95% Hausdorff distance) but a reduction in dice performance. Adding prior knowledge to our high-resolution network significantly improved the dice coefficient and mean surface distance when segmenting the submandibular glands, but had significantly larger Hausdorff distances when segmenting the level II lymph nodes. The majority of cases saw no significant change due to prior knowledge, which may indicate that textural information in the cropped region of the organ at risk (local high-resolution information) may be of equal importance as the general location of the organ at risk (global low-resolution information). To our knowledge, this is the first report on the effect of cascaded CNNs on auto-segmentation performance on MRI for OARs in HNC. A similar MRI auto-segmentation study explores the performance of CNNs with different spatial dimensionality or multiple MRI contrasts, and found no significant performance difference between 2D and 3D methods, and no benefit from multiple MRI contrasts [24]. Whilst cascaded methods allow auto-segmentation on a 3D patch of MRI data at native image resolution, when using a GPU similar to this study, a GPU with larger memory may allow auto-segmentation on a full volumetric MRI image at native image resolution with a single 3D U-net.

In the exploration of the generalisability of our auto-segmentation methods on slightly different MRI images, we observed significantly reduced performance in the majority of cases, with only the parotid gland segmentation (metric = left/right, dice = 0.730/0.775, mean surface distance = 1.64/1.20 mm) being considered suitable for clinical use with manual correction. In a similar CNN based auto-segmentation study, a comparable drop in parotid auto-segmentation performance was reported [24] where networks were trained on a 3T MRI dataset and tested on 1.5T MRI dataset. The observed drop in performance suggests that the spatial and intensity normalisation applied as a pre-processing step is not adequate to remove the relevant variability between these MRI datasets prior

to inference with a CNN. Lack of generalisability of MRI auto-segmentation with CNN based methods has previously been reported in the brain [57], and appears to be less of an issue with CT based external validation studies showing good performance in parotid and submandibular glands [58]. With the increasing adoption of MR guided radiotherapy, larger annotated datasets from different MRI scanners and differing imaging protocols may help address the generalisability issues of current MRI auto-segmentation methods. Another approach may be to train models for a specific use, using single site or scanner specific data, such as the machine specific data being collected by the MOMENTUM collaboration [59].

A limitation of this study is the assessment of segmentation performance with purely geometric metrics. The dice coefficient is a measure of overlap, but good segmentation overlap does not guarantee a clinical outcome such as delivered radiation dose[22]. The probability threshold of 0.45 used to create binary labels in this study was selected in previous work to account for systematic volume difference between segmentations[36]. Whilst a threshold of 0.5 is more commonly used, we observed minimal change in all performance metrics on the RT-MAC validation dataset over a threshold range from 0.45 to 0.55 (Supplementary Figure 3). The manual segmentation of 10 of the 12 available RT-MAC Test* cases may limit the comparison of our results with future studies. Many radiotherapy patients will have additional diagnostic imaging which may provide valuable anatomic information that was not considered in this work, a multi-modality network for HNC auto-segmentation has demonstrated improved performance when combining MRI and CT images over CT images alone [23]. The high performance of a parotid auto-segmentation method that incorporates an existing clinically approved segmentation [19] suggests that the integration of planning segmentations and images may benefit auto-segmentation performance in the context of adaptive radiotherapy workflows or longitudinal studies. Whilst this study provides some insight into the performance of deep learning auto-segmentation of organs at risk on MRI of HNC patients, larger studies are needed to validate these methods before clinical implementation. To address the current

issue of small annotated MRI datasets, general adversarial networks are being explored to create synthetic MRI training datasets from larger available annotated CT datasets [60].

Conclusion

We have demonstrated the application of 3D CNNs to auto-segment the parotid and submandibular glands in HNC patients on T2 weighted MRI images, with dice and surface distance performance similar to previous MRI and CT studies. Cascading a localiser and cropped high-resolution 3D CNN can generate higher resolution auto-segmentations with significantly improved accuracy, in comparison to a low-resolution 3D CNN, for both the parotid and submandibular glands. The mean surface distance of auto-segmentations of the submandibular glands can be significantly improved by using a low-resolution predicted label as prior knowledge in the high-resolution stage of a cascaded CNN. The developed methods performed significantly worse on T2 weighted images from a different source than the training dataset, but may be suitable for auto-segmenting the parotid glands. Further work is required to improve the generalisability of the proposed automated segmentation methods for application with the other considered organs at risk.

Acknowledgements

This project was supported by funding from the Peter MacCallum Cancer Foundation. This research was undertaken using the LIEF HPC-GPGPU Facility established with the assistance of LIEF Grant LE170100200 and hosted at the University of Melbourne.

Competing Interests

TK has a Research Collaboration Agreement with Varian Medical Systems. NH has a Clinical Research Collaborations Program Grant from Varian Medical Systems. SPN was funded by the Australian Postgraduate Award, the Royal Australian and New Zealand College of Radiologists (RANZCR) Research Grant and the Radiological Society of North America (RSNA) Fellow Grant. JK, PJ and BC have no conflicts to disclose.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394-424.
2. Delaney G, Jacob S, Barton M. Estimation of an optimal external beam radiotherapy utilization rate for head and neck carcinoma. *Cancer.* 2005;103(11):2216-2227.
3. Nutting CM, Morden JP, Harrington KJ, et al. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): A phase 3 multicentre randomised controlled trial. *Lancet Oncol.* 2011;12(2):127-136.
4. Peters LJ, O'Sullivan B, Giralt J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: Results from TROG 02.02. *J Clin Oncol.* 2010;28(18):2996-3001.
5. Kosmin M, Ledsam J, Romera-Paredes B, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiother Oncol.* 2019;135:130-140.
6. Aliotta E, Nourzadeh H, Siebers J. Quantifying the dosimetric impact of organ-at-risk delineation variability in head and neck radiation therapy in the context of patient setup uncertainty. *Phys Med Biol.* 2019;64(13).
7. Brouwer CL, Steenbakkers RJHM, van den Heuvel E, et al. 3D Variation in delineation of head and neck organs at risk. *Radiat Oncol.* 2012;7(1).
8. Hong TS, Tomé WA, Harari PM. Heterogeneity in head and neck IMRT target design and clinical practice. *Radiother Oncol.* 2012;103(1):92-98.
9. Daisne JF, Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: A clinical validation. *Radiat Oncol.* 2013;8(1).
10. Hoang Duc AK, Eminowicz G, Mendes R, et al. Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. *Med Phys.* 2015;42(9).
11. Stapleford LJ, Lawson JD, Perkins C, et al. Evaluation of Automatic Atlas-Based Lymph Node Segmentation for Head-and-Neck Cancer. *Int J Radiat Oncol Biol Phys.* 2010;77(3):959-966.

12. Teguh DN, Levendag PC, Voet PWJ, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. *Int J Radiat Oncol Biol Phys.* 2011;81(4):950-957.
13. Sims R, Isambert A, Grégoire V, et al. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. *Radiother Oncol.* 2009;93(3):474-478.
14. Noel CE, Parikh PJ, Spencer CR, et al. Comparison of onboard low-field magnetic resonance imaging versus onboard computed tomography for anatomy visualization in radiotherapy. *Acta Oncol.* 2015;54(9):1474-1482.
15. Mutic S, Dempsey JF. The ViewRay System: Magnetic Resonance-Guided and Controlled Radiotherapy. *Semin Radiat Oncol.* 2014;24(3):196-199.
16. Winkel D, Bol GH, Kroon PS, et al. Adaptive radiotherapy: The Elekta Unity MR-linac concept. *Clin Transl Radiat Oncol.* 2019;18:54-59.
17. Yan D, Vicini F, Wong J, Martinez A. Adaptive radiation therapy. *Phys Med Biol.* 1997;42(1):123-132.
18. Güngör G, Serbez İ, Temur B, et al. Time Analysis of Online Adaptive Magnetic Resonance–Guided Radiation Therapy Workflow According to Anatomical Sites. *Practical Radiation Oncology.* 2020.
19. Yang X, Wu N, Cheng G, et al. Automated segmentation of the parotid gland based on atlas registration and machine learning: A longitudinal mri study in head-and-neck radiation therapy. *Int J Radiat Oncol Biol Phys.* 2014;90(5):1225-1233.
20. Cheng G, Yang X, Wu N, et al. Multi-atlas-based segmentation of the parotid glands of MR images in patients following head-and-neck cancer radiotherapy. 2013.
21. Wardman K, Prestwich RJD, Gooding MJ, Speight RJ. The feasibility of atlas-based automatic segmentation of MRI for H&N radiotherapy planning. *J Appl Clin Med Phys.* 2016;17(4):146-154.
22. Kieselmann JP, Kamerling CP, Burgos N, 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(14):17.
23. Močnik D, Ibragimov B, Xing L, et al. Segmentation of parotid glands from registered CT and MR images. *Phys Med.* 2018;52:33-41.
24. Kieselmann JP, Fuller CD, Gurney-Champion OJ, Oelfke U. Auto-segmentation of the parotid glands on MR images of head and neck cancer patients with deep learning strategies. *medRxiv.* 2020.
25. Cardenas CE, Mohamed ASR, Yang J, et al. Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations. *Med Phys.* 2020;47(5):2317-2322.
26. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Frangi AF, Navab N, Hornegger J, Wells WM, eds. Vol 9351: Springer Verlag; 2015:234-241.
27. Zhu W, Huang Y, Zeng L, et al. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys.* 2019;46(2):576-589.
28. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. In: Unal G, Ourselin S, Joskowicz L, Sabuncu MR, Wells W, eds. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* Vol 9901 LNCS: Springer Verlag; 2016:424-432.
29. Milletari F, Ahmadi S-A, Kroll C, et al. Hough-CNN: Deep learning for segmentation of deep brain regions in MRI and ultrasound. *Computer Vision and Image Understanding.* 2017;164:92-102.

30. Roth HR, Lu L, Lay N, et al. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Med Image Anal.* 2018;45:94-107.
31. Zhou Y, Xie L, Shen W, Fishman E, Yuille A. Pancreas segmentation in abdominal CT scan: a coarse-to-fine approach. *arXiv preprint arXiv:161208230.* 2016.
32. Roth HR, Oda H, Hayashi Y, et al. Hierarchical 3D fully convolutional networks for multi-organ segmentation. *arXiv preprint arXiv:170406382.* 2017.
33. Tappeiner E, Pröll S, Hönig M, et al. Multi-organ segmentation of the head and neck area: an efficient hierarchical neural networks approach. *Int J Comput Assisted Radiol Surg.* 2019;14(5):745-754.
34. *Keras* [computer program]. https://keras.io2015.
35. Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:160304467.* 2016.
36. Jackson P, Hardcastle N, Dawe N, Kron T, Hofman MS, Hicks RJ. Deep learning renal segmentation for fully automated radiation dose estimation in unsealed source therapy. *Front Oncol.* 2018;8(JUN).
37. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:150203167.* 2015.
38. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Paper presented at: ICML2010.
39. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 bias correction. *IEEE Trans Med Imaging.* 2010;29(6):1310-1320.
40. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The design of simpleITK. *Front Neuroinformatics.* 2013;7(DEC).
41. Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 2016.
42. Kingma DP, Ba JL. Adam: A method for stochastic optimization. 2015.
43. Dice LR. Measures of the amount of ecologic association between species. *Ecology.* 1945;26(3):297-302.
44. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing Images Using the Hausdorff Distance. *IEEE Trans Pattern Anal Mach Intell.* 1993;15(9):850-863.
45. Yeghiazaryan V, Voiculescu I. Family of boundary overlap metrics for the evaluation of medical image segmentation. *J Med Imaging.* 2018;5(1).
46. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging.* 2015;15(1):29.
47. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17(3):261-272.
48. Walker GV, Awan M, Tao R, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiother Oncol.* 2014;112(3):321-325.
49. Willems S, Crijns W, La Greca Saint-Esteven A, et al. Clinical implementation of deepvoxnet for auto-delineation of organs at risk in head and neck cancer patients in radiotherapy. In: Malpani A, Zenati MA, Oyarzun Laura C, et al., eds. Vol 11041 LNCS: Springer Verlag; 2018:223-232.
50. Thomson D, Boylan C, Liptrot T, et al. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. *Radiat Oncol.* 2014;9(1).
51. Qazi AA, Pekar V, Kim J, Xie J, Breen SL, Jaffray DA. Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach. *Med Phys.* 2011;38(11):6160-6170.

52. Gorthi S, Duay V, Houhou N, et al. Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration. *IEEE J Sel Top Sign Proces.* 2009;3(1):135-147.
53. Han X, Hoogeman MS, Levendag PC, et al. Atlas-based auto-segmentation of head and neck CT images. In. Vol 5242 LNCS2008:434-441.
54. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys.* 2017;44(2):547-557.
55. Grégoire V, Ang K, Budach W, et al. Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. *Radiother Oncol.* 2014;110(1):172-181.
56. Grégoire V, Levendag P, Ang KK, et al. CT-based delineation of lymph node levels and related CTVs in the node-negative neck: DAHANCA, EORTC, GORTEC, NCIC, RTOG consensus guidelines. *Radiother Oncol.* 2003;69(3):227-236.
57. Le Berre A, Kamagata K, Otsuka Y, et al. Convolutional neural network-based segmentation can help in assessing the substantia nigra in neuromelanin MRI. *Neuroradiology.* 2019;61(12):1387-1395.
58. Brunenberg EJL, Steinseifer IK, van den Bosch S, et al. External validation of deep learning-based contouring of head and neck organs at risk. *Phy Imaging Radiat Oncol.* 2020;15:8-15.
59. de Mol van Otterloo SR, Christodouleas JP, Blezer ELA, et al. The MOMENTUM Study: An International Registry for the Evidence-Based Introduction of MR-Guided Adaptive Therapy. *Front Oncol.* 2020;10.
60. Kieselmann JP, Fuller CD, Gurney-Champion OJ, Oelfke U. Cross-modality deep learning: Contouring of MRI data from annotated CT data only. *Med Phys.* 2020.

FIGURE AND TABLE LEGENDS

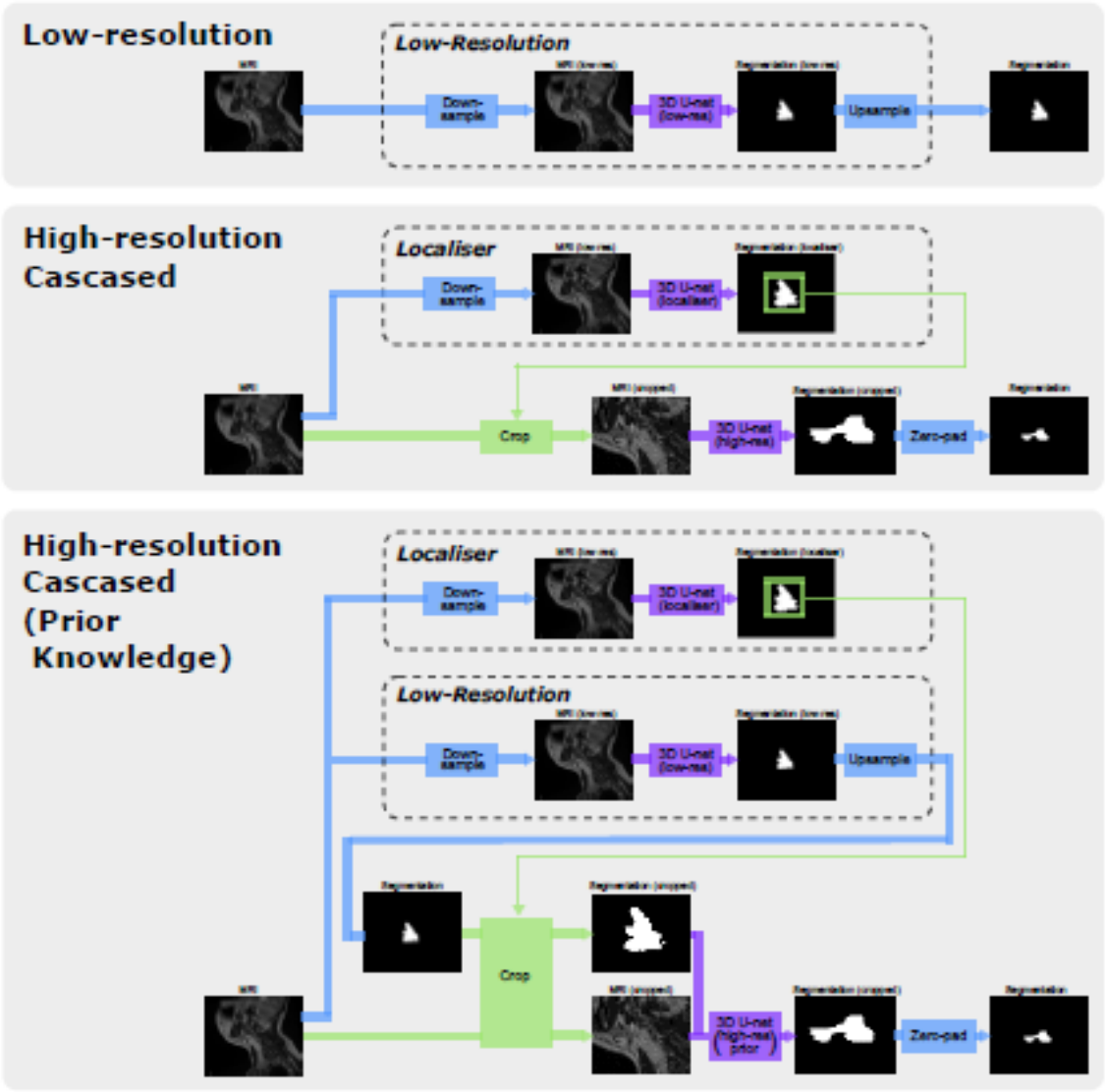

Figure 1: Auto-segmentation models with complexity increasing from top to bottom. A low-resolution model downsamples the entire 3D MRI volume and calculates a low-resolution segmentation. The high-resolution cascaded model uses a localiser network to crop a 3D MRI region at the native image resolution, and calculates a segmentation at the native image resolution. The high-resolution cascaded model with prior knowledge uses a localiser network to crop a 3D region of both the MRI and a label predicted by the low-resolution network, and calculates a segmentation at the native image resolution. The localiser network is a variation of the low-resolution network, which was trained with dilated segmentations and is used to define a crop region in the two cascaded networks.

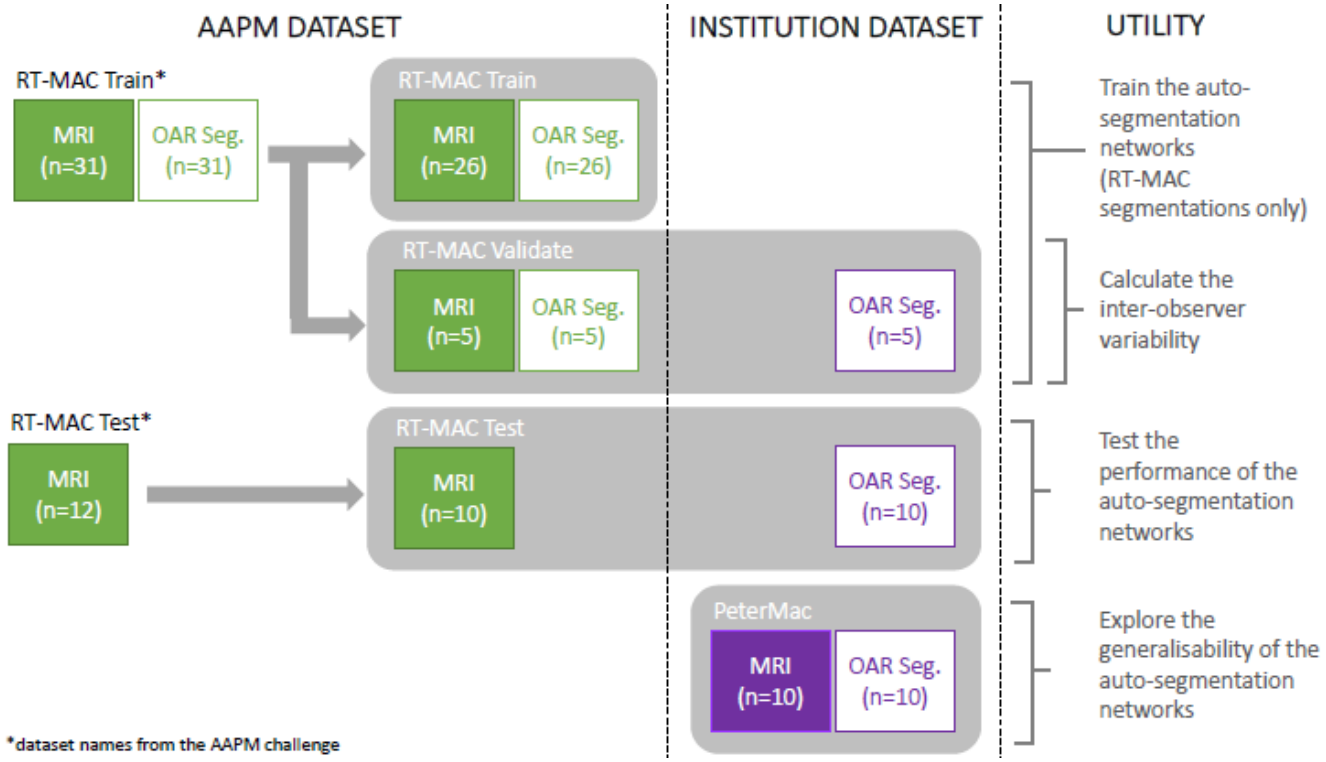

Figure 2: Summary of the RT-MAC and institutional datasets which include magnetic resonance images (MRI) and organ at risk manual segmentations (OAR Seg.). The original RT-MAC Train* dataset was split into the RT-MAC Train dataset for training the auto-segmentation networks and the RT-MAC Validate dataset for selecting the best performing network over the 300 training epochs. To measure inter-observer variability, an experienced radiation oncologist from our institute manually created organ at risk segmentations on the RT-MAC validate dataset. As the original RT-MAC Test* dataset does not include publicly available segmentations, the performance of the auto-segmentation networks on the RT-MAC data was tested using institutional segmentations of the RT-MAC Test dataset, which were created for 10 or the 12 RT-MAC Test* cases. To explore the generalisability of the auto-segmentation networks, their performance was measured on the institutional dataset.

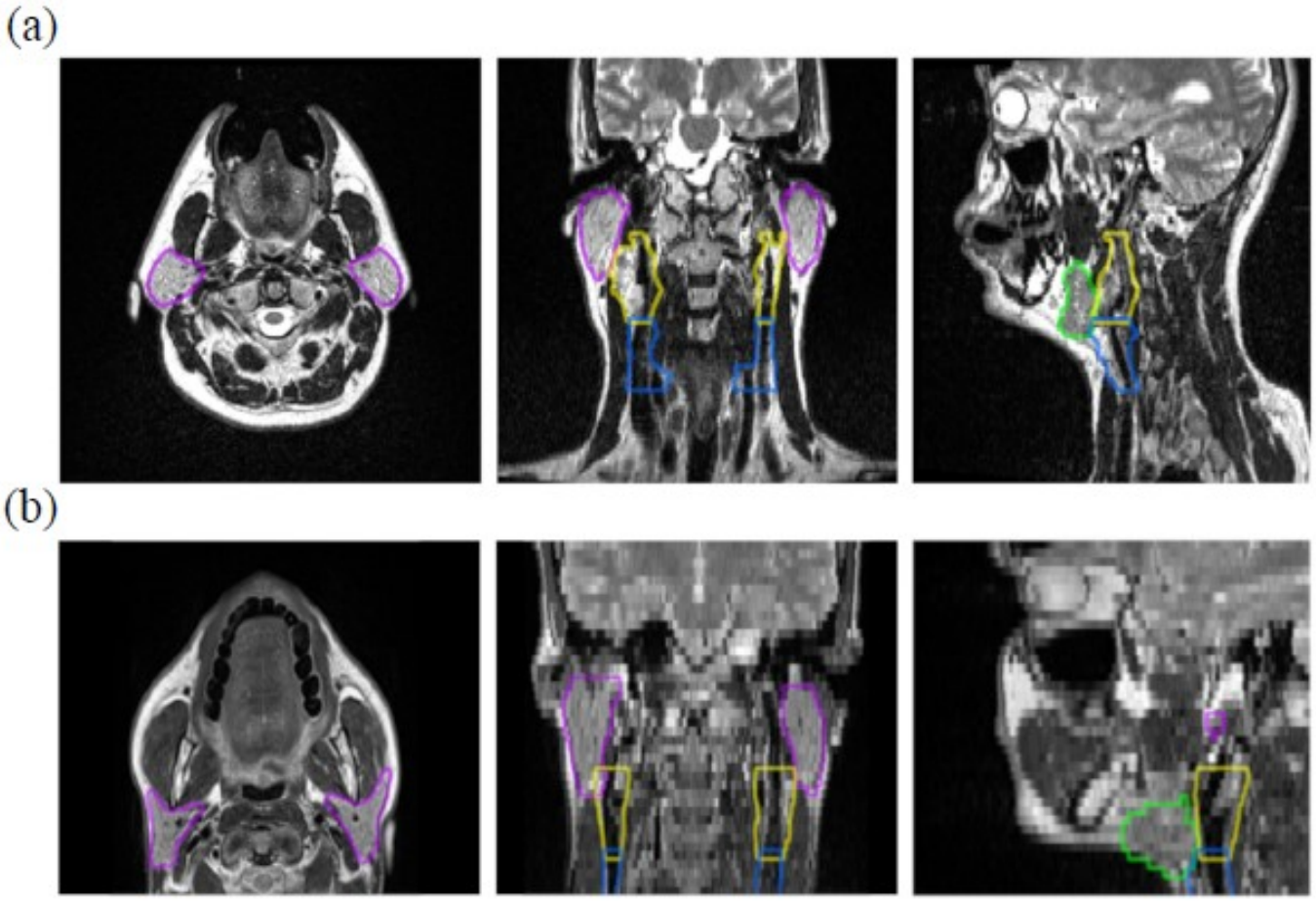

Figure 3: Comparison of T2 weighted magnetic resonance images and organ at risk segmentation from the (a) RT-MAC challenge dataset and (b) our institutional dataset. The organ at risk segmentations are defined for the (purple) parotid glands, (green) submandibular glands, (yellow) level two lymph nodes and (blue) level three lymph nodes. The images from each dataset have similar T2 weighted contrast but differ in spatial coverage and voxel size.

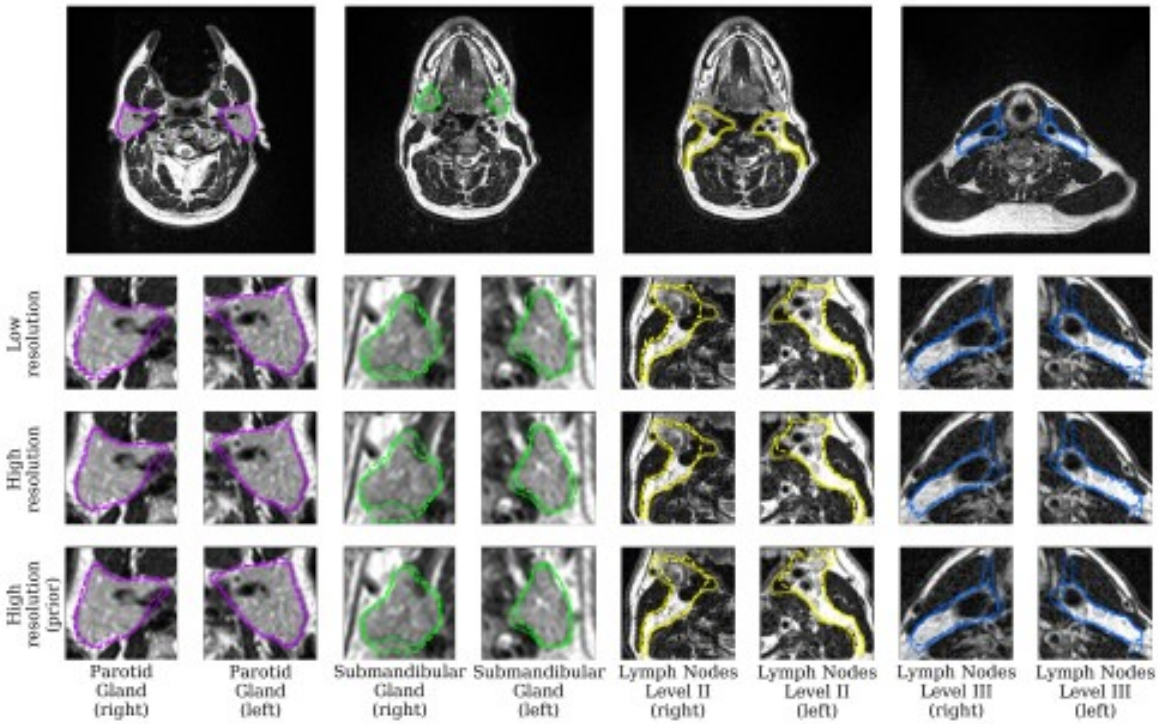

Figure 4: Transverse T2 weighted MRI of a patient with head and neck cancer from the RT-MAC Validate dataset with (solid line) manual institutional segmentations and (dashed line) auto-segmentations of organs at risk; (purple) parotid glands, (green) submandibular glands, (yellow) level two lymph nodes and (blue) level three lymph nodes. The zoomed in panels (bottom 3 rows) highlight the differences between the three auto-segmentation networks.

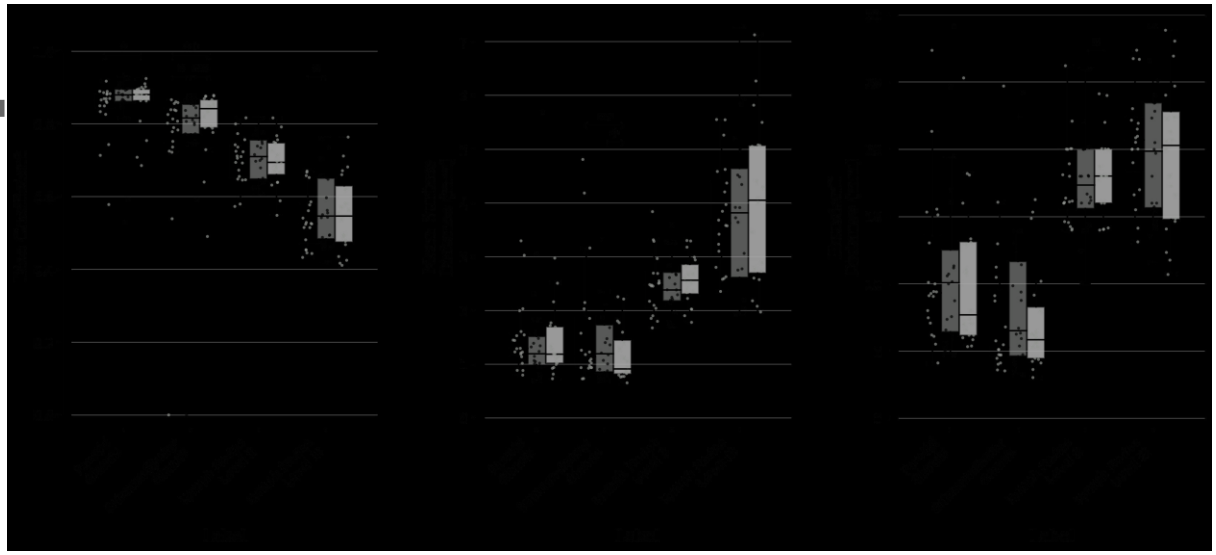

Figure 5: Comparison of the auto-segmentation performance of the different network architectures, as assessed with a similarity metric (dice coefficient) and surface distance metrics (mean surface distance & Hausdorff distance). Significant differences between the (dark grey) low-resolution network, (light grey) high-resolution network and (white) high-resolution network with prior knowledge are marked with a single asterisk*, $p < 0.05$, and a double asterisk, $p < 0.01$.. The general trend in auto-segmentation performance across organs at risk, from highest to lowest dice performance, is the parotid glands, submandibular glands, level two lymph nodes and level three lymph nodes. The trend is slightly different for surface distance metrics with the submandibular glands auto-segmentations having the highest performance. High-resolution methods showed significant improvements over the low-resolution method, with higher dice coefficients observed for all OARs other than the level II lymph nodes and smaller mean surface distances observed for the submandibular glands. The addition of prior knowledge to a high-resolution network significantly improved both the dice coefficient and the mean surface distance for the submandibular glands, but led to significantly larger Hausdorff distances for the level II lymph nodes.**

OAR	Dice Coefficient			Mean Surface Distance (mm)			Hausdorff Distance (mm)		
	RT-MAC Test (institute segmentation)	PeterMac (institute segmentation)	IOV	RT-MAC Test (institute segmentation)	PeterMac (institute segmentation)	IOV	RT-MAC Test (institute segmentation)	PeterMac (institute segmentation)	IOV
Parotid_L	0.860±0.067	0.730±0.078	0.870±0.023	1.41±0.33	1.64±0.95	1.16±0.22	11.31±6.09	12.88± 4.51	9.59±3.13
Parotid_R	0.857±0.063	0.775±0.105	0.882±0.011	1.33±0.39	1.20±0.59	1.13±0.19	9.68±4.37	14.00± 5.46	10.41±4.81
Submand_L	0.830±0.032	0.537±0.303	0.831±0.036	1.16±0.47	5.04±7.90	0.96±0.26	6.83±3.29	14.26±11.16	5.46±1.19
Submand_R	0.785±0.123	0.435±0.284	0.817±0.022	1.19±0.50	3.52±3.22	1.08±0.19	7.62±4.34	13.84± 6.71	5.63±0.93
LN Lvl II_L	0.708±0.053	0.553±0.192	0.703±0.033	2.44±0.41	2.31±1.58	2.56±0.56	16.62±3.47	19.34± 6.06	15.60±3.29
LN Lvl II_R	0.715±0.071	0.525±0.194	0.709±0.047	2.33±0.49	3.73±2.74	2.94±1.40	17.72±3.66	22.91±14.19	16.21±3.97
LN Lvl III_L	0.561±0.100	0.304±0.195	0.577±0.057	3.70±0.80	6.91±5.82	3.65±0.76	20.26±4.24	22.60± 8.81	20.56±3.62
LN Lvl III_R	0.573±0.105	0.189±0.183	0.588±0.069	3.53±1.22	8.22±6.37	3.52±0.69	17.91±5.98	22.35± 7.41	18.24±2.06

Table 1: Auto-segmentation network performance as compared to manual institutional segmentations, for the highest performing network architecture on each OAR and for each performance metric. Auto-segmentation performance is reported for auto-segmentations generated on the RT-MAC Test dataset and the PeterMac institutional dataset and compared to the inter-observer variability (IOV) as measured on the RT-MAC Validate dataset between RT-MAC manual segmentations and institutional manual segmentations.

Network Architecture	MRI Normalisation (seconds)	Auto-segmentation per organ-at-risk (seconds)	Auto-segmentation per patient (seconds)
Low-resolution	15.0	3.3±1.1	41.8± 8.9
High-resolution	15.0	17.5±1.0	155.8± 8.0
High-resolution (prior knowledge)	15.0	24.5±1.3	210.8±10.7

Table 2: Time efficiency of the different auto-segmentation network architectures. The spatial and intensity normalisation applied to the MRI images was identical for all networks, total auto-segmentation time increased as function of network complexity.

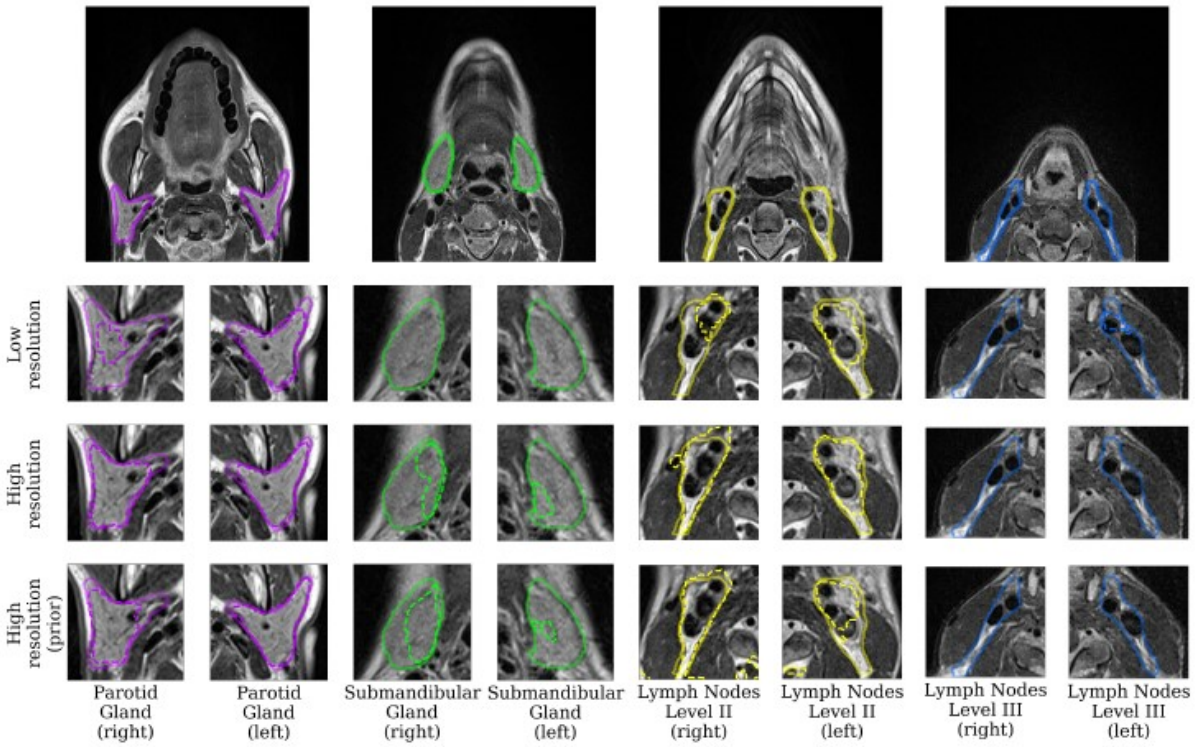

Figure 6: Transverse T2 weighted MRI of a patient with head and neck cancer from the institutional MRI dataset with (solid line) manual segmentations and (dashed line) auto-segmentations of organs at risk; (purple) parotid glands, (green) submandibular glands, (yellow) level two lymph nodes and (blue) level three lymph nodes. The zoomed in panels (bottom 3 rows) highlight the differences between the three auto-segmentation network architectures.

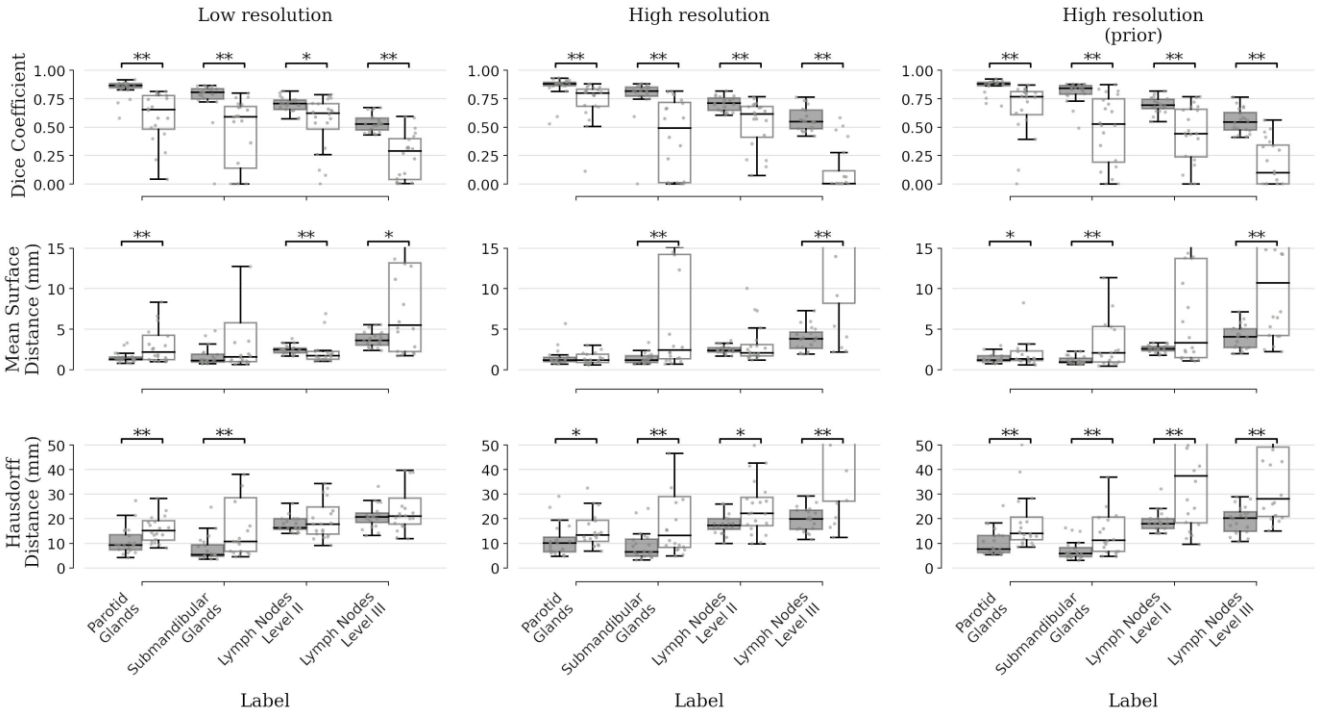

Figure 7: Comparison of auto-segmentation performance of three models on the (dark grey) RT-MAC test dataset and (white) institutional dataset. Dice performance is significantly poorer on the institutional dataset for all three models across all organs-at-risk. Surface distance metric performance is significantly poorer for all three models on the majority of organs-at-risk. There was no significant performance difference on the institutional MRI data for some models with respect to mean surface distance in the parotid glands, submandibular glands and level II lymph nodes, and with respect to Hausdorff distance in the level II and level III lymph nodes. The only improved performance was observed with the low-resolution network, with significantly smaller mean surface distance for the level II lymph nodes.

Publication	Imaging Modality	Magnetic Field Strength	Method	OAR	Images		Dice Coefficient	Mean Surface Distance (mm)	Hausdorff Distance (mm)
					Train	Test			
Yang 2014 [19]	MRI (T2w)	1.5T	Patient specific Atlas + SVM	Parotid	15	42	0.911±0.016	0.29±0.11	3.46 ± 1.22
Willems 2018 [48]	CT	-	3D FCN	Parotid	70	20	0.897	1.05	10.06
Walker 2014 [47]	CT	-	Atlas based	Parotid	-	40	0.89 ±0.11		
Korte 2021	MRI (T2w)	1.5T	3D U-net	Parotid	31	10	0.860±0.067	1.33±0.40	9.677± 4.37
Cheng 2013 [20]	MRI (T2w)	NR	Multi-atlas Hybrid + SVM	Parotid	4	1(5)	0.853		
Kieselmann 2020 [24]	MRI (T1w+T2w)	3T	Multi-atlas	Parotid	24	3(27)	0.85 ±0.03	1.39±0.54	14.98 ± 6.88
Kieselmann 2020 [24]	MRI (T1w+T2w)	3T	2D/2.5D/3D Multimodal U-net	Parotid	24	3(27)	0.85 ±0.08	1.63±1.27	15.19 ± 8.09
Kieselmann 2018 [22]	MRI (T1w)	3T	Multi-atlas	Parotid	11	1(12)	0.83 ±0.03	1.35±0.40	12.13 ± 3.91
Wardman 2016 [21]	MRI (T1w)	1.5T	Multi-atlas	Parotid	13	1(14)	0.79	4.79	
Močnik 2018 [23]	CT+MRI (T1w)	3T	CNN	Parotid	34-35	8-9(43)	0.788	1.57	
Močnik 2018 [23]	CT	-	CNN	Parotid	34-35	8-9(43)	0.765	1.57	
Wardman 2016 [21]	CT	-	Multi-atlas	Parotid	13	1(14)	0.76	6.23	
Willems 2018 [48]	CT	-	3D FCN	Submand	70	20	0.877	0.83	5.54
Qazi 2011 [50]	CT	-	Atlas+Feature Refinement	Submand	15	10	0.84		3.52
Korte 2021	MRI (T2w)	1.5T	3D U-net	Submand	31	10	0.830±0.032	1.16±0.47	6.83 ± 3.30
Thomson 2014 [49]	CT	-	Atlas+Intensity Shape Model	Submand	-	10	0.8	0.6	5.6
Walker 2014 [47]	CT	-	Atlas based	Submand	-	40	0.73 ±0.25		
Wardman 2016 [21]	MRI (T1w)	1.5T	Multi-atlas	LN Lvl II	8	1(9)	0.8	3.95	
Wardman 2016 [21]	CT	-	Multi-atlas	LN Lvl II	8	1(9)	0.78	5.57	
Korte 2021	MRI (T2w)	1.5T	3D U-net	LN Lvl II	31	10	0.715±0.071	2.33±0.49	16.61± 3.47
Gorthi 2009 [51]	CT	-	Atlas + Active Contours	LN Lvl II	1	9(10)	0.58 ±0.08		14.41± 6.11
Korte 2021	MRI (T2w)	1.5T	3D U-net	LN Lvl III	31	10	0.573±0.105	3.53±1.22	17.91± 5.98
Gorthi 2009 [51]	CT	-	Atlas + Active Contours	LN Lvl III	1	9(10)	0.48 ±0.18		18.45±11.85

Table 3: Comparison of auto-segmentation performance against previously reported methods. If multiple structures (i.e. left/right) or multiple network architectures are reported, the best performing result has been selected. For each organ at risk the studies are ordered by auto-segmentation performance, as defined by the dice coefficient. The MRI studies are conducted on T1 weighted (T1w) or T2 weighted (T2w) images and the magnetic field strength is noted, apart from one study where it was not reported (NR). There are atlas based methods with refinement steps such as support vector machines (SVM) and a range of deep learning methods such as fully connected networks (FCN), convolutional neural networks (CNN). The number of images used to train and test the auto-segmentation methods is listed, in papers where cross validation was used the number of train and test images for one fold of validation is reported, with the total number of test images in brackets. The number of patients per study is the combined number of test and training images, apart from Yang 2014 which includes 15 patients imaged at multiple timepoints.