arXiv:1809.04430v2 [cs.CV] 20 Oct 2020

Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy

Stanislav Nikolov[1*], Sam Blackwell[1*], Alexei Zverovitch[2*], Ruheena Mendes[3], Michelle Livne[2], Jeffrey De Fauw[1], Yojan Patel[2], Clemens Meyer[1], Harry Askham[2], Bernardino Romera-Paredes[1], Christopher Kelly[2], Alan Karthikesalingam[2], Carlton Chu[1], Dawn Carnell[3], Cheng Boon[4], Derek D'Souza[3], Syed Ali Moinuddin[3], DeepMind Radiographer Consortium[1], Hugh Montgomery[2,5,6], Geraint Rees[2,5], Mustafa Suleyman[1], Trevor Back[1], Cían O. Hughes[2,3+], Joseph R. Ledsam[7+], and Olaf Ronneberger[1+]

[1]DeepMind, London, UK
[2]Google Health, London, UK
[3]University College London Hospitals NHS Foundation Trust, London, UK
[4]Clatterbridge Cancer Centre NHS Foundation Trust, Liverpool, UK
[5]University College London, London, UK
[6]Centre for Human Health and Performance, and Institute for Sports, Exercise and Health, University College London, London, UK
[7]Google AI, Tokyo, Japan
[*]These authors contributed equally to this work
[+]These authors contributed equally to this work

Over half a million individuals are diagnosed with head and neck cancer each year worldwide. Radiotherapy is an important curative treatment for this disease, but it requires manual time consuming delineation of radio-sensitive organs at risk (OARs). This planning process can delay treatment, while also introducing inter-operator variability with resulting downstream radiation dose differences. While auto-segmentation algorithms offer a potentially time-saving solution, the challenges in defining, quantifying and achieving expert performance remain. Adopting a deep learning approach, we demonstrate a 3D U-Net architecture that achieves expert-level performance in delineating 21 distinct head and neck OARs commonly segmented in clinical practice. The model was trained on a dataset of 663 deidentified computed tomography (CT) scans acquired in routine clinical practice and with both segmentations taken from clinical practice and segmentations created by experienced radiographers as part of this research, all in accordance with consensus OAR definitions. We demonstrate the model's clinical applicability by assessing its performance on a test set of 21 CT scans from clinical practice, each with the 21 OARs segmented by two independent experts. We also introduce surface Dice similarity coefficient (surface DSC), a new metric for the comparison of organ delineation, to quantify deviation between OAR surface contours rather than volumes, better reflecting the clinical task of correcting errors in the automated organ segmentations. The model's generalisability is then demonstrated on two distinct open source

datasets, reflecting different centres and countries to model training. With appropriate validation studies and regulatory approvals, this system could improve the efficiency, consistency, and safety of radiotherapy pathways.

1 Introduction

Each year, 550,000 people are diagnosed with cancer of the head and neck worldwide [1]. This incidence is rising [2], more than doubling in certain subgroups over the last 30 years [3, 4, 5]. Where available, most will be treated with radiotherapy which targets the tumour mass and areas at high risk of microscopic tumour spread. However, strategies are needed to mitigate the dose-dependent adverse effects which result from incidental irradiation of normal anatomical structures ('organs at risk', OARs) [6, 7, 8, 9].

The efficacy and safety of head and neck radiotherapy thus depends upon the accurate delineation of OARs and tumour, a process known as segmentation or contouring. However, the fact that this process is predominantly done manually means that results may be both inconsistent and imperfectly accurate [10], leading to large inter- and intra-practitioner variability even amongst experts, and thus variation in care quality [11].

Segmentation is also very time consuming: an expert can spend four hours or more on a single case [12]. The duration of resulting delays to treatment commencement (see Fig. 1) is associated with increased risk both of local recurrence and of overall mortality [13, 14]. Increasing demands for, and shortages of, trained staff already place a heavy burden on healthcare systems which can lead to long delays for patients as radiotherapy is planned [15, 16], and the continued rise in head and neck cancer incidence may make it impossible to maintain even current temporal reporting standards [4]. Such issues also represent a barrier to 'Adaptive Radiotherapy'- the process of repeated scanning, segmentation and radiotherapy planning throughout treatment which maintains the precision of tumour targeting (and OARs avoidance) in the face of treatment-related anatomic changes such as tumour shrinkage [17].

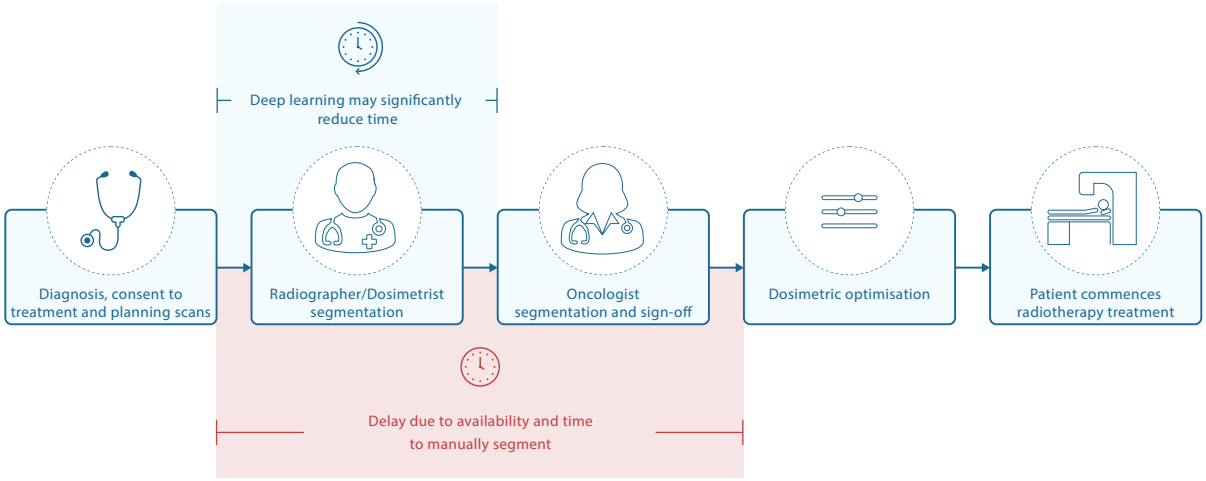

Figure 1 | A typical clinical pathway for radiotherapy. After a patient is diagnosed and the decision is made to treat with radiotherapy, a defined workflow aims to provide treatment that is both safe and effective. In the UK the time delay between decision to treat and treatment delivery should be no greater than 31 days[18]. Time-intensive manual segmentation and dose optimisation steps can introduce delays to treatment.

Automated (i.e. computer-performed) segmentation has the potential to address these challenges. However, most segmentation algorithms in clinical use are atlas-based, producing their segmentations by fitting previously labelled reference images to the new target scan. This might not sufficiently account for either

post-surgical changes, or the variability in normal anatomical structure which exists between patients, particularly when considering the variable effect that tumours may have on local anatomy; they may thus be prone to systematic error. To date, such algorithm-derived segmentations still require significant manual editing, perform at expert levels on only a small number of organs, demonstrate an overall performance in clinical practice inferior to that of human experts, and have failed to significantly improve clinical workflows [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31].

In recent years, deep learning based algorithms have proven capable of delivering substantially better performance than traditional segmentation algorithms. In head and neck cancer segmentation, several deep learning based approaches have been proposed. Some of them use standard convolutional neural network classifiers on patches with tailored pre- and post-processing [32, 33, 34, 35, 36]. However, the U-Net convolutional architecture [37] has shown promise in the area of deep-learning based medical image segmentation [38] and has now also been applied to head and neck radiotherapy segmentation [39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52].

Despite the promise deep learning offers, barriers remain to the application of auto-segmentation to radiotherapy planning. These include the absence of consensus on how 'expert' performance is defined, the lack of available methods by which such human performance can be compared to that delivered by automated segmentation processes, and thus how the clinical acceptability of automated processes can be defined.

Here we address these challenges, and report a deep learning approach that delineates a wide range of important OARs in head and neck cancer radiotherapy scans, to human expert standard. We achieve this using a study design that includes (i) the introduction of a clinically meaningful performance metric for segmentation in radiotherapy planning; (ii) a representative set of images acquired during routine clinical practice; (iii) an unambiguous segmentation protocol for all organs; and (iv) a segmentation of each test set image according to these protocols by two independent experts. In addition to the model's generalisability, as demonstrated on two distinct open source datasets, by achieving performance equal to human experts on previously unseen patients from the same hospital site used for training we demonstrate the clinical applicability of our approach.

2 Results

2.1 Selecting clinically representative datasets

Datasets are described in detail in the Methods section. In brief, the first dataset was a representative sample of CT scans used to plan curative-intent radiotherapy of head and neck cancer for patients at University College London Hospitals NHS Foundation Trust (UCLH), a single high-volume centre. We performed iterative cycles of model development using the UCLH scans ('training' and 'validation' subsets), taking the performance on a previously unseen subset ('test') as our primary outcome.

It is also important to demonstrate a model's generalisability to data from previously unseen demographics and distributions. To do this we curated test and validation datasets of open source CT scans. These were collected from The Cancer Imaging Archive ("TCIA test set") [53, 54, 55] and the "PDDCA": Public Domain Database for Computational Anatomy dataset released as part of the 2015 challenge ("PDDCA test set"; [30]).

Table 2 details the characteristics of these datasets and their patient demographics. Ethnicity and protected-group status is not reported, as this information was not available in the source systems. Twenty-one organs at risk were selected to represent a wide range of anatomical regions throughout the head and neck. To provide a human clinical comparison for the algorithm, each case was manually segmented by a single radiographer with arbitration by a second radiographer. This was compared to our study's 'gold standard' ground truth graded by two further radiographers and arbitrated by one of two independent

specialist oncologists, each with a minimum of four years specialist experience in radiotherapy treatment planning for head and neck patients.

An example of model performance is shown in Fig. 2. We compare our performance (model vs oncologist) to radiographer performance (radiographer vs oncologist). For more information on dataset selection, inclusion and exclusion criteria for patients and OARs please refer to the Methods section.

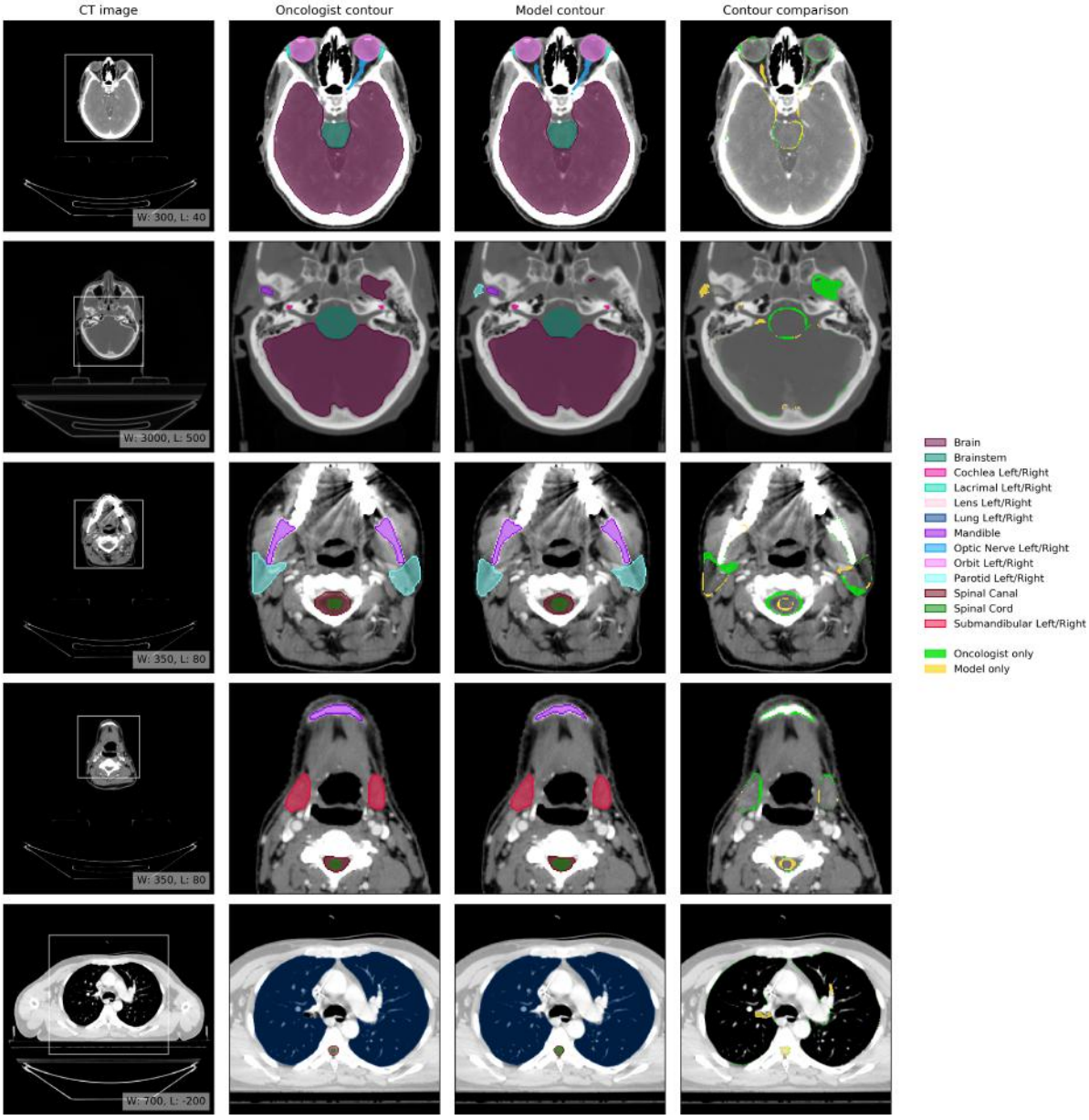

Figure 2 | Example results. (**CT image**) Axial slices at five representative levels from the raw CT scan of a 55-59 year old male patient was selected from the UCLH dataset (patient UCLH-20) were selected to best demonstrate the OARs included in the work. The levels shown as 2D slices have been selected to demonstrate all 21 OARs included in this study. The window levelling has been adjusted for each to best display the anatomy present. (**Oncologist contour**) The ground truth segmentation, as defined by experienced radiographers and arbitrated by a head and neck specialist oncologist. (**Model contour**) Segmentations produced by our model. (**Contour comparison**) Contoured by Oncologist only (green region) or Model only (yellow region). Two further randomly selected UCLH set scans are shown in Fig. 12 and Fig. 13. Best viewed on a display.

2.2 A New Metric for Assessing Clinical Performance

In routine clinical care, algorithm-derived segmentation would be reviewed and potentially corrected by a human expert, just as those created by radiographers currently are. Segmentation performance is thus best assessed by determining the fraction of the surface that needs to be redrawn. The standard volumetric Dice similarity coefficient (volumetric DSC; [56]) is not well suited to this because it weighs all regions of misplaced delineation equally and independently of their distance from the surface. For example, two inaccurate segmentations could have a similar volumetric DSC score if one were to deviate from the correct surface boundary by a small amount in many places while another had a large deviation at a single point. Correcting the former would likely take a considerable amount of time as it would require redrawing almost all of the boundary while the latter could be corrected much faster, potentially with a single edit action.

For quantitative analysis we therefore introduce a new segmentation performance metric, "surface Dice similarity coefficient" (surface DSC) (Fig. 3), which assesses the overlap of two surfaces (at a specified tolerance) instead of the overlap of two volumes. This provides a measure of agreement between the surfaces of two structures, which is where most of the human effort in correcting is usually expended. In doing so, we also address the volumetric DSC's bias towards large OARs, where the large (and mostly trivial) internal volume accounts for a much larger proportion of the score.

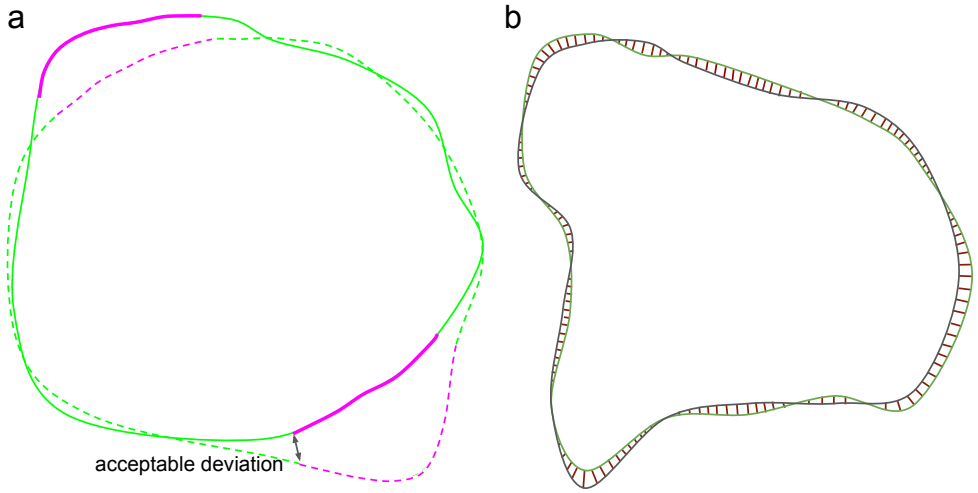

Figure 3 | Surface DSC performance metric. (**a**) Illustration of the computation of the surface DSC. Continuous line: predicted surface. Dashed line: ground truth surface. Black arrow: the maximum margin of deviation which may be tolerated without penalty, hereafter referred to by τ. Note that in our use case each OAR has an independently calculated value for τ. Green: acceptable surface parts (distance between surfaces $\leq \tau$). Pink: unacceptable regions of the surfaces (distance between surfaces $> \tau$). The proposed surface DSC metric reports the good surface parts compared to the total surface (sum of predicted surface area and ground truth surface area). (**b**) Illustration of the determination of the organ-specific tolerance. Green: segmentation of an organ by oncologist A. Black: segmentation by oncologist B. Red: distances between the surfaces. We defined the organ-specific tolerance as the 95th percentile of the distances collected across multiple segmentations from a subset of seven TCIA scans, where each segmentation was performed a radiographer and then arbitrated by an oncologist, neither of whom had seen the scan previously.

When evaluating the surface DSC we must define a threshold within which variation is clinically acceptable. To do this we first defined organ-specific tolerances (in mm) as a parameter of the proposed metric, τ. We computed these acceptable tolerances for each organ by measuring the inter-observer variation in segmentations between three different consultant oncologists (each with over 10 years experience

in OAR delineation) on the validation subset of TCIA images.

To penalise both false negative and false positive parts of the predicted surface, our proposed metrics measures both of the non-symmetric distances between the surfaces and then normalises by the combined surface area. Like the volumetric DSC, the surface DSC ranges from 0 (no overlap) to 1 (perfect overlap).

This means that approximately 95% of the surface was properly outlined (i.e. within τ mm of the correct boundary) while 5% needs to be corrected. There is no consensus as to what constitutes non-significant variation in such segmentation. We thus selected a surface DSC of 0.95 - a stringency which likely far exceeds expert oncologist intra-rater concordance [19, 57]. For a more formal definition and implementation, please refer to the Methods section.

2.3 Model Performance

Model performance was evaluated alongside that of therapeutic radiographers (each with at least 4 years of experience) segmenting the test set of UCLH images independently of the oncologist-reviewed scans (which we used as our ground truth).

The model performed similarly to humans: on all OARs studied there was no clinically meaningful difference between the deep learning model's segmentations and those of the radiographers (Fig. 4 and Table 8).

To investigate the generalisability of our model, we additionally evaluate performance on open source scans ('TCIA test set'). These were collected from sites in the USA, where the patient demographics, the clinical pathways for radiotherapy and the scanner type and parameters differed from our UK training set in meaningful ways. Nevertheless, model performance was preserved and, in 19 of 21 OARs, the model performed within the threshold defined for human variability Fig. 5. The fact that performance in 2 OARs (brainstem and right lens) was less than that in UK data may relate to issues of image quality in several TCIA test set scans.

For more detailed results demonstrating surface DSC and volumetric DSC for each individual patient from the TCIA test set please refer to Table 4 and Table 5 respectively in the appendix.

2.4 Comparison to previous work

An accurate quantitative comparison to previously published literature is difficult due to inherent differences in definitions of ground truth segmentations and varied processes of arbitration and consensus building. Given that the use of surface DSC is novel to this study, we also report the standard volumetric DSC scores achieved by our algorithm (despite the shortcomings of this method) so that direct comparison of our results can be made with those in the existing literature. An overview of past papers which have reported mean volumetric DSC for unedited automatic delineation of head and neck CT OARs can be found in Table 1 and Table 11. Each used different datasets, scanning parameters and labelling protocols, meaning that resulting volumetric DSC results varied significantly. No study other than ours segmented lacrimal glands. We compared these results to those obtained when we applied our model to three different datasets: the TCIA open source test set, an additional test set from the original UCLH dataset ("UCLH test set") and the dataset released by the Public Domain Database for Computational Anatomy (PDDCA) as part of the 2015 MICCAI head and neck radiotherapy OAR segmentation challenge ("PDDCA test set"; [30]). To contextualise the performance of our model, radiographer performance is shown on the TCIA test set, and oncologist inter-observer variation is shown on the UCLH test set.

While not the primary test set, we nevertheless present per-patient surface DSC and volumetric DSC for the PDDCA test set in Table 6 and Table 7 in the appendix.

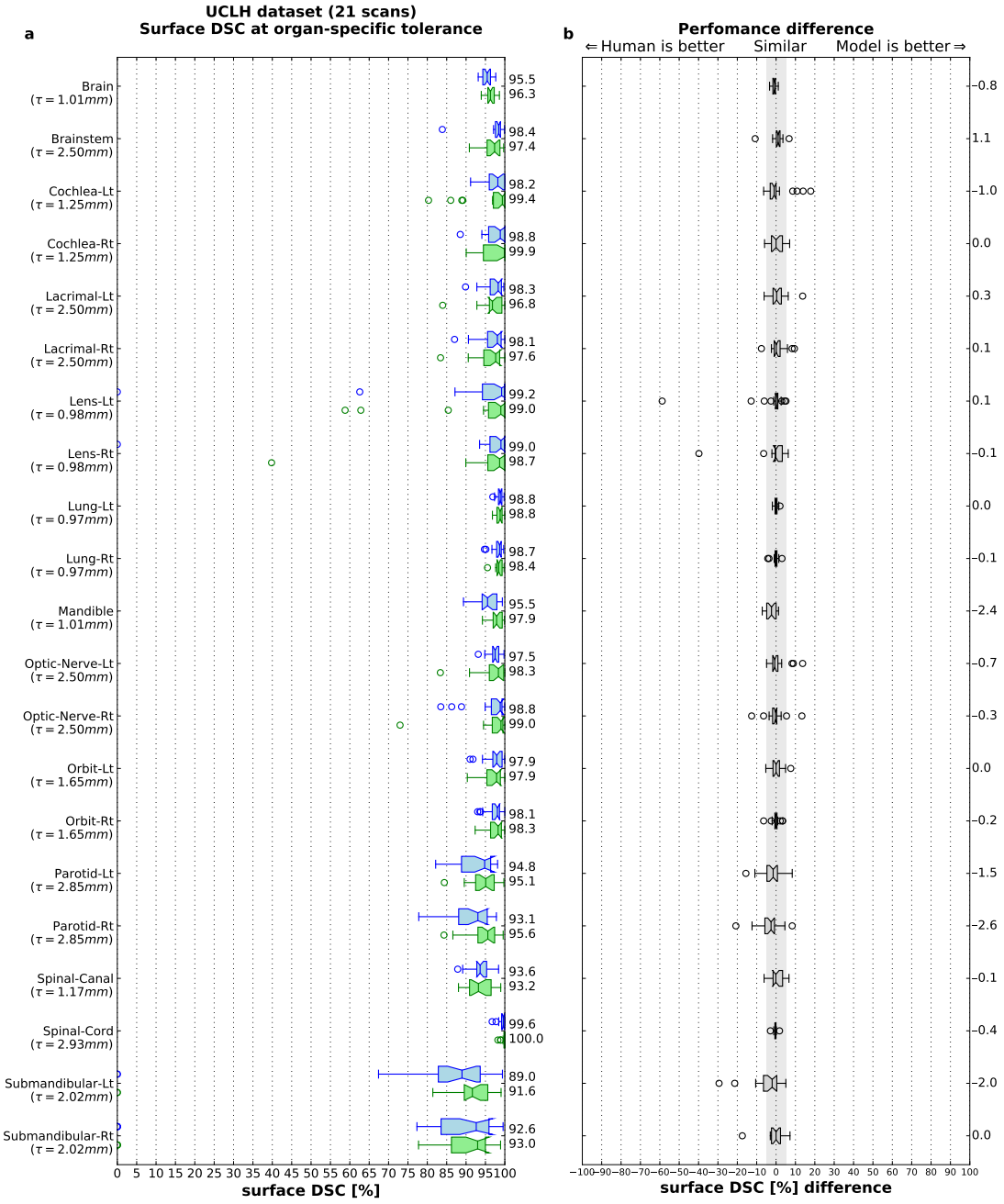

Figure 4 | UCLH test set: Quantitative performance of the model in comparison to radiographers. (**a**) The model achieves a surface DSC similar to humans in all 21 organs at risk (on the UCLH held out test set) when compared to the gold standard for each organ at an organ-specific tolerance τ. Blue: our model, green: radiographers. (**b**) Performance difference between the model and the radiographers. Each blue dot represents a model-radiographer pair. The grey area highlights non-substantial differences (-5% to +5%).
The box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers indicate most extreme, non-outlier data points. Where data lies outside 1.5 $\times$ interquartile range it is represented as a circular flier. The notches represent the 95% confidence interval (CI) around the median.

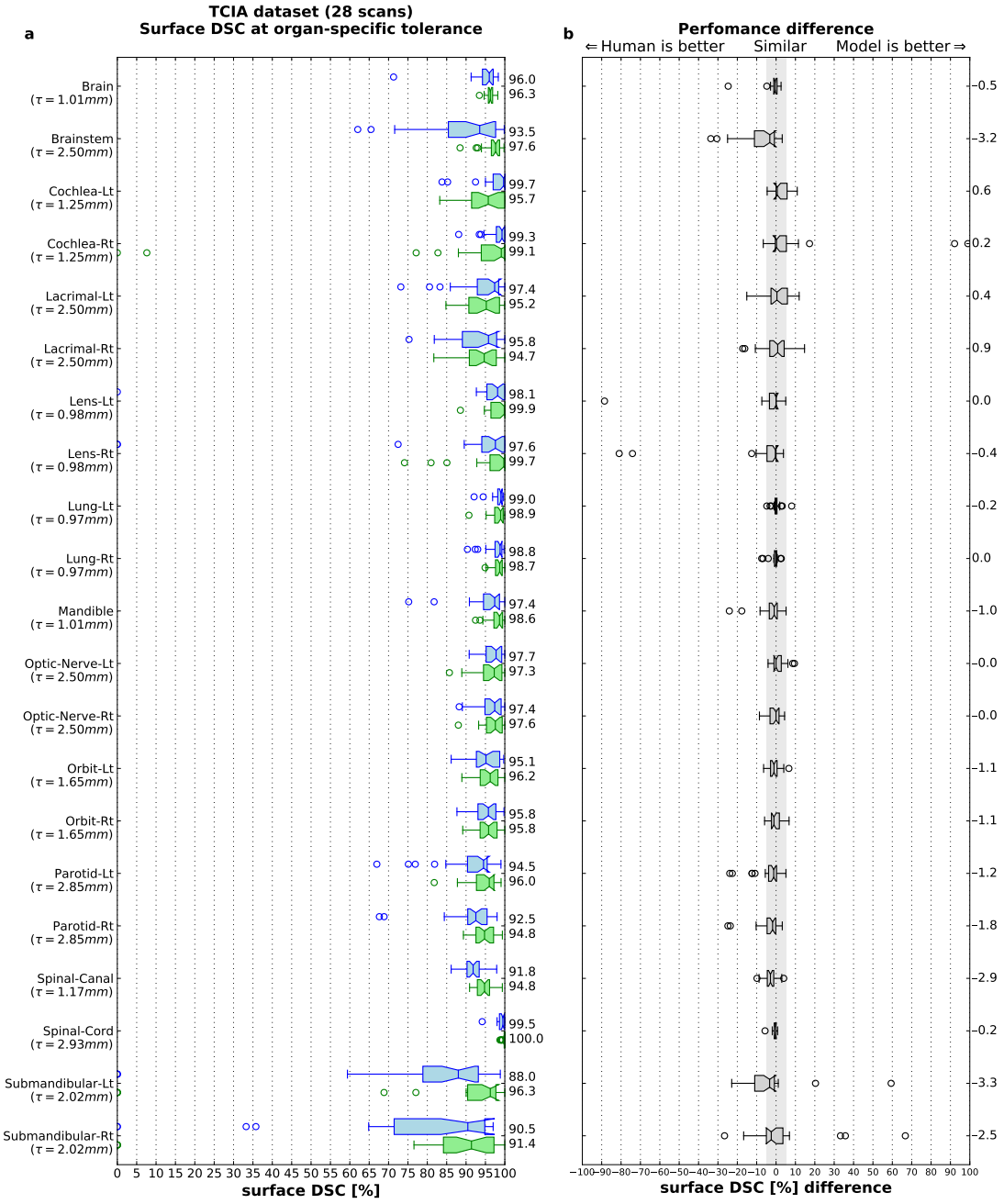

Figure 5 | Model generalisability to an independent test set from TCIA. Quantitative performance of the model on the TCIA test set in comparison to radiographers. (**a**) Surface DSC (on the TCIA open source test set) for the segmentations compared to the gold standard for each organ at an organ-specific tolerance τ. Blue: our model, green: radiographers. (**b**) Performance difference between the model and the radiographers. Each blue dot represents a model-radiographer pair. Red lines show the mean difference. The grey area highlights non-substantial differences (-5% to +5%).
The box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend from the box to show the range of the data, except where data lies outside 1.5 $\times$ interquartile range, which is represented as a circular flier. The notches represent the 95% confidence interval (CI) around the median.

3 Discussion

We demonstrate an automated deep learning-based segmentation algorithm that can perform as well as experienced radiographers for head and neck radiotherapy planning. Our model was developed using CT

Table 1 | Volumetric DSC performance of our model and previously published deep learning models. An overview of previously published deep-learning based automatic segmentation works that reported volumetric DSC for the OARs included in this study on planning CT scans. Due to the large volume of publications, this overview includes only results of deep learning works. For a full literature overview see Table 11. The datasets and ground truths used varied between studies making comparison difficult. Despite this, we show results alongside our evaluation of our model, radiographers and oncologists against our ground truth across multiple datasets. The latter assesses inter-observer variation between oncologists.

Study	Brain	Brainstem	Cochlea		Lacrimal		Lens		Lung		Mandible	Optic Nerve		Orbit		Parotid		Spinal Canal		Spinal Cord	Submandibular	
			lt	rt	lt	rt	lt	rt	lt	rt	rt		lt	rt	lt	rt	lt	rt	lt	rt		lt	rt
Guo (2020) [51]		88									94	72	71			87	86				78	81	
Liang (2020) [58]		92					88	87			94		74	93	93		88			90		81	
Qiu (2020) [59]											95												
Sun (2020) [60]		86									94			90	90	84	81			89	78	77	
van Dijk (2020) [61]		83[1]									95[1]					84[1]	83[1]			87[1]	77[1]	78[1]	
Wong (2020) [62]		83											47				80			79		82	
Chan (2019) [63]		89									91					85	86			87	84	85	
Gao (2019) [64]		86					81	79				64	62	88	91	77	80			87			
Jiang (2019) [65]		88									93					85	86				79	77	
Lei (2019) [66]		87											66				86						
Men (2019) [46]		90									92					86	86						
Rhee (2019) [48]	98	86	65	68			73	70			87	89	90	89	90	83	83			83			
Sun (2019) [67]							85	84				80	82	94	94								
Tang (2019) [49]		86					82	83			93	75	76	92	92	85	85			86	81	83	
Tappeiner (2019) [47]		82									91	64	63			80	81						
Tong (2019) [68]		87									94	66	70			85	86				81	82	
van Rooij (2019) [50]		64														83	83				82	81	
Wang (2019) [45]		88									93	74	74			86	85				76	73	
Xue (2019) [69]		90									96	86	84			89	89				86	85	
Zhong (2019) [36]													89				92						
Hänsch (2018) [39]																	86						
Kodym (2018) [44]		92									95		80				90					88	
Liang (2018) [42]		90					83	84			91	66	72			85	85						
Močnik (2018) [34]																	77						
Nikolov (2018) [70]	99	88	65	75	69	70	81	80	99	99	96	76	77	95	95	85	85	95	88	85	85		
Tong (2018) [41]		87									94	65	69			84	83				76	81	
Ren (2018) [35]												72	70										
Willems (2018) [43]		92	75	73							96					86	90				79	88	
Zhu (2018) [40]		87									93	72	71			88	87				81	81	
Ibragimov (2017) [33]											90	64	65	88	88	77	78			87	70	73	
Fritscher (2016) [32]																	81					65	
Radiographer (TCIA) (28 scans)	99.1	90.0 ±0.2	74.9 ±2.5	69.6 ±10.9	67.3 ±23.1	67.8 ±10.4	87.7 ±11.0	84.5 ±8.0	98.7 ±14.7	98.9 ±0.7	94.2 ±0.5	79.3 ±2.2	78.4 ±4.9	93.3 ±6.2	93.4 ±2.1	87.1 ±1.9	87.4 ±3.4	93.9 ±3.1	84.3 ±1.8	84.7 ±4.6	77.5 ±18.3	±28.5	
Our model (TCIA) (28 scans)	98.8	85.1 ±1.1	80.5 ±7.1	81.0 ±8.8	64.4 ±7.2	63.8 ±11.9	81.6 ±9.0	75.7 ±16.6	98.7 ±24.5	98.8 ±0.6	92.9 ±0.7	77.9 ±3.5	76.3 ±5.0	92.6 ±5.8	93.1 ±2.0	84.1 ±1.8	84.6 ±5.8	91.7 ±4.2	80.3 ±1.6	81.8 ±7.6	77.8 ±8.7	±18.1	
Radiographer (UCLH) (21 scans)	99.2	90.1 ±0.2	77.9 ±2.4	80.3 ±14.0	74.1 ±10.1	71.8 ±7.0	82.7 ±7.8	83.9 ±22.6	98.6 ±23.8	98.6 ±0.9	95.8 ±1.3	80.3 ±1.2	79.4 ±5.2	93.9 ±7.4	94.2 ±1.4	88.1 ±0.9	87.5 ±2.8	93.1 ±3.4	81.6 ±2.0	87.5 ±6.0	86.8 ±4.0	±4.0	
Our model (UCLH) (21 scans)	99	91 ±0.2	81 ±2.2	79 ±8.2	73 ±5.7	72 ±5.6	78 ±5.8	81 ±25.0	98 ±25.8	98 ±1.3	93 ±2.2	77 ±2.0	75 ±4.8	95 ±7.0	95 ±1.3	85 ±1.0	84 ±3.8	93 ±4.5	78 ±1.4	83 ±8.9	86 ±8.4	±4.9	
Oncologist (UCLH) (8 - 75 scans)[2]	99.0[3]	91.9[3]	68.5	75.8 ±14.8	63.3 ±8.5	61.6 ±13.1	86.2 ±14.3	87.6 ±10.1	98.4[3] ±9.9	98.6[3]	95.4[3]	77.1	76.0 ±6.3	94.8[3] ±7.1	94.8[3]	90.1[3]	90.7[3]	94.9[3]	87.7[3]	91.1[3]	90.1[3]		
Our model (PDDCA) (15 scans)		84.2	±5.2								93.8	71.6 ±1.9	69.1 ±6.2	±5.9		88.1	86.6 ±2.0	±3.5		76.5	79.2 ±9.1	±6.5	

Values for volumetric DCS are mean (± standard deviation) unless otherwise stated. "CNN": convolutional neural network. "FCN": fully convolutional network. "GAN": generative adversarial network.

[1] Values estimated from figures; actual values not reported.

[2] Number of scans per organ varies, see Table 10.

[3] Volumetric DSC estimated from sparse labels.

scans derived from routine clinical practice, and therefore should be applicable in a hospital setting for segmentation of OARs, routine Radiation Therapy Quality Assurance (RTQA) peer review and reducing the associated variability between different specialists and radiotherapy centres [71].

Clinical applicability must be supported not only by a high model performance but also by evidence of model generalisability to new external datasets. To achieve this, we present these results on three separate

test sets, one of which (the PDDCA test set) uses a different segmentation protocol. Here, performance in the majority of OARs was maintained when tested on scans taken from a range of previously unseen international sites. Although these scans varied in patient demographics, scanning protocol, device manufacturer and image quality, the model still achieved human performance on 19 of the 21 OARs studied; only the right lens and brainstem were below radiographer performance. For these OARs, the performance of the model might have been lower than expert performance owing to lower image quality. This is particularly evident for the right lens, where the anatomical borders were quite indistinct in some TCIA test set cases, thus preventing full segmentation by the model (Fig. 11). Moreover, a precise CT definition of the brainstem's proximal and distal boundaries is lacking, a factor which might have contributed to labelling variability and thus to decreased model performance. Finally, demographic bias may have resulted from the TCIA data set selecting for cases of more advanced head and neck cancer [53], or from variability in the training data [10].

One major contribution of this article is the presentation of a performance measure that represents the clinical task of OAR correction. In the first pre-print of this work we introduced surface DSC [70], a metric conceived to be sensitive to clinically significant errors in OAR delineation. Surface DSC has recently been shown to be more strongly correlated with the amount of time required to correct a segmentation for clinical use than traditional metrics including volumetric DSC [72, 73]. Small deviations in OAR border placement can have a potentially serious impact, increasing the risk of debilitating side effects for the patient. Misplacement by only a small offset may thus require the whole region to be redrawn and in such cases an automated segmentation algorithm may offer no time-savings at all. Volumetric DSC is relatively insensitive to such small changes for large organs as the absolute overlap is also large. Difficulties identifying the exact borders of smaller organs can result in large differences in volumetric DSC even if these differences are not clinically relevant in terms of their effect on radiotherapy treatment. By strongly penalising border placement outside a tolerance determined by consultant oncologists, the surface DSC metric resolves these issues.

While volumetric DSC is therefore not representative of clinical consequences, it remains the most popular metric for evaluating segmentation models and therefore the only metric that allows comparison to previously published works. In recent years, fully convolutional networks became the most popular and successful methodology for OAR segmentation in head and neck CT for de-novo radiotherapy planning [45, 46, 47, 48, 49, 50, 65, 66, 64, 63, 69, 62, 59, 58, 60]. Although not directly comparable due to different datasets and labelling protocols, our volumetric DSC results compare favourably against the existing published literature for many of the OARs (see Table 1 and Table 11 for more details on this and other prior publications). In OARs with inferior volumetric DSC score compared to the published literature, both our model and the human radiographers achieved similar scores. This suggests that current and previously published results are difficult to compare, either due to the inclusion of more difficult cases than previous studies, or due to different segmentation and scanning protocols. To allow more objective comparisons of different segmentation methods, we make our labelled TCIA datasets freely available to the academic community.[1] At least 11 auto-segmentation software solutions are currently available commercially, with varying claims regarding their potential to lower segmentation time during radiotherapy planning [74]. The principal factor that determines whether or not automatic segmentation is time-saving during the radiotherapy workflow is the degree to which automated segmentations require correction by oncologists.

The wide variability in state-of-the-art and limited uptake in routine clinical practice motivates the need for clinical studies evaluating model performance in practice. Future work will seek to define the clinical acceptability of the segmented OARs produced by our models, and estimating the time-saving that could be achieved during the radiotherapy planning workflow in a real-world setting.

[1]The dataset is available at https://github.com/deepmind/tcia-ct-scan-dataset.

A number of other study limitations should also be addressed in future work. First, we included only planning CT scans since magnetic resonance imaging (MRI) and Positron Emission Tomography (PET) scans were not routinely performed for all patients in the UCLH dataset. Some OAR classes, such as optic chiasm, require co-registration with MR images for optimal delineation and access to additional imaging has been shown to improve the delineation of optic nerves [34]. As a result, certain OAR classes were deliberately excluded from this CT-based project and will be addressed in future work which will incorporate MRI scans. A second limitation regards the classes of OARs in this study. While we present one of the largest sets of reported OARs in the literature [75, 49, 76], some omissions occurred (e.g., oral cavity) due to an insufficient number of examples in the training data that conformed to a standard international protocol. The number of oncologists used in the creation of our ground truth may not have fully captured the variability in OAR segmentation, or may have been biased towards a particular interpretation of the Brouwer Atlas used as our segmentation protocol. Even in an organ as simple as the spinal cord that is traditionally reliably outlined by auto-segmentation algorithms, there is ambiguity between the inclusion of, for example, the nerve roots. Such variation may widen the thresholds of acceptable deviation in favour of the model despite a consistent protocol. Future work will address these deficits, alongside time-consuming lymph node segmentation.

Finally, neither of the test sets used in this paper include the patients' protected-characteristic status. This is a significant limitation as is prevents study of intersectional fairness.

3.1 Conclusion

In conclusion, we demonstrate that deep learning can achieve human expert level performance in the segmentation of head and neck OARs in radiotherapy planning CT scans, using a clinically applicable performance metric designed for this clinical scenario. We provide evidence of the generalisability of this model by testing it on patients from different geographies, demographics and scanning protocols. This segmentation algorithm performed with similar accuracy compared to experts and has the potential to improve the speed, efficiency, and consistency of radiotherapy workflows, with an expected positive influence on patient outcomes. Future work will investigate the impact of our segmentation algorithm in clinical practice.

4 Methods

4.1 Datasets

University College London Hospitals NHS Foundation Trust (UCLH) serves an urban, mixed socioeconomic and ethnicity population in central London, U.K. and houses a specialist centre for cancer treatment. Data were selected from a retrospective cohort of all adult (>18 years of age) UCLH patients who had computed tomography (CT) scans to plan radical radiotherapy treatment for head and neck cancer between 01/01/2008 and 20/03/2016. Both initial CT images and re-scans were included in the training dataset. Patients with all tumour types, stages and histological grades were considered for inclusion, so long as their CT scans were available in digital form and were of sufficient diagnostic quality. The standard CT pixel spacing was 0.976mm by 0.976mm by 2.5mm, and scans with non-standard spacing (with the exception of 1.25mm spacing scans which were subsampled) were excluded to ensure consistent performance metrics during training. Note that for the TCIA test set, below, the in-plane pixel spacing was not used as an exclusion criteria, i ranged from 0.94mm to 1.27mm. For the PDDCA test set we included all scans, and the voxels varied between 2mm - 3mm in height and 0.98mm - 1.27mm in the axial dimension. The wishes of patients who had requested that their data should not be shared for research were respected.

Of the 513 patients who underwent radiotherapy at UCLH within the given study dates, a total of 486 patients (838 scans), mean age 57, male 337, female 146, gender unknown 3, met the inclusion criteria. Of note, no scans were excluded on the basis of poor diagnostic quality. Scans from UCLH were split into a training set (389 patients, 663 scans), validation set (51 patients, 100 scans) and test set (46 patients, 75 scans). From the selected test set 19 patients (21 scans) underwent adjudicated Contouring described below. No patient was included in multiple datasets: in cases where multiple scans were present for a single patient, all were included in the same subset. Where multiple scans were present for a single patient, this reflects CT scans taken for the purpose of re-planning radiotherapy due to anatomical changes during a course of treatment. It is important for models to perform well in both scenarios as treatment naive and post-radiotherapy OAR anatomy can differ. However, to avoid potential correlation between the same organs segmented twice in the same dataset, care was taken to avoid this in the TCIA test set (see below).

Twenty-one organs at risk were selected throughout the head and neck area to represent a wide range of anatomical regions. We used a combination of segmentations sourced from those used clinically at UCLH and additional segmentations performed in-house by trained radiographers.

We divided our UCLH dataset into the following categories: (1) **Training set**: Used to train the model, a combination of UCLH clinical segmentations and in-house segmentations, some of which were only 2D slices [2]. (2) **UCLH Validation set**: Used to evaluate model performance and steer additional dataset priorities, this used in-house segmentations only, as we didn't want to overfit to any clinical bias. (3) **UCLH test set**: Our primary result set, each scan has every OAR labelled and was independently segmented from scratch by two radiographers before one of the pair of scans (chosen arbitrarily) was reviewed and corrected by an experienced radiation oncologist.

As these scans were taken from UCLH patients not present elsewhere, and to consider generalisability, we curated additional open source CT scans available from The Cancer Imaging Archive (TCGA-HNSC and Head-Neck Cetuximab) [53, 54, 55]. The open source (4) **TCIA validation set** and (5) **TCIA test set** were both labelled in the same way as our UCLH test set.

Non-CT planning scans and those that did not meet the same slice thickness as the UCLH scans (2.5mm) were excluded. These were then manually segmented in-house according to the Brouwer Atlas ([77]; the segmentation procedure is described in further detail below). We included 31 scans (22 Head-Neck Cetuximab, 9 TCGA-HNSC) which met these criteria, which we further split into validation (6 patients, 7 scans) and test (24 patients, 24 scans) sets (Fig. 6). The original segmentations from the Head-Neck Cetuximab dataset were not included; a consensus assessment by experienced radiographers and oncologists found the segmentations either non-conformant to the selected segmentation protocol or below the quality that which would be acceptable for clinical care. The original inclusion criteria for Head-Neck Cetuximab were patients with stage III-IV carcinoma of the oropharynx, larynx, and hypopharynx, having Zubrod performance of 0-1, and meeting predefined blood chemistry criteria between 11/2005 to 03/2009. The TCGA-HNSC dataset included patients treated for Head-Neck Squamous Cell Carcinoma, with no further restrictions being apparent. For more information please refer to the specific citations [55, 53].

All test sets were kept separate during model training and validation. Table 2 describes in further detail the demographics and characteristics within the datasets; to obtain a balanced demographic in each of the test, validation and training datasets we sampled randomly stratified splits and selected one that minimised the differences between the key demographics in each dataset.

In addition, the (6) **PDDCA open source dataset** consisted of 15 patients selected from the Head-Neck Cetuximab open source dataset [53]; due to differences in selection criteria and test/validation/training set

[2] Due to the time required to segment larger organs manually, we initially relied heavily on sparse segmentations to make efficient use of the radiographers' time.

allocation there were five scans present in both the TCIA and PDDCA test sets. This dataset was used without further post-processing and only accessed once for assessing volumetric DSC performance. The PDDCA test set differ from the TCIA test set in both segmentation protocol and axial slice thickness. For more details on the dataset characteristics and preprocessing please refer to the work of Raudaschl and colleagues [30].

Table 2 details the characteristics of these datasets and the patient demographics.

Table 2 | Dataset Characteristics

		UCLH			TCIA		PDDCA
		Train	Validation	Test	Validation	Test	Test
Total scans (patients)		663 (389)	100 (51)	21 (19)	7 (6)	24 (24)	15 (15)
Average patient age		57.1	57.5	59.6	56.5	59.9	58.6
Sex	Female	207 (115)	36 (19)	7 (6)	2 (2)	2 (2)	2 (2)
	Male	450 (271)	64 (32)	14 (13)	5 (4)	20 (20)	9 (9)
	Unknown	6 (3)	0	0	0	2 (2)	4 (4)
Tumour site	Oropharynx	145 (86)	27 (15)	7 (6)	0	8 (8)	2 (2)
	Lip, oral cavity and pharynx	80 (52)	20 (8)	4 (4)	1 (1)	3 (3)	0
	Tongue	53 (26)	8 (5)	1 (1)	2 (2)	7 (7)	0
	Larynx	46 (31)	8 (3)	2 (2)	2 (2)	4 (4)	0
	Nasopharynx	48 (24)	5 (3)	0	0	0	0
	Head, face and neck	37 (23)	8 (3)	1 (1)	0	0	0
	Nasal Cavity	32 (19)	2 (1)	1 (1)	0	0	0
	Connective and soft tissue	37 (18)	2 (1)	1 (1)	0	0	0
	Hypopharynx	17 (10)	1 (1)	0	2 (1)	1 (1)	0
	Accessory sinus	10 (7)	2 (1)	0	0	0	0
	Oesophagus	6 (2)	1 (1)	0	0	0	0
	Other	33 (20)	0	0	0	1 (1)	0
	Unknown	119 (71)	16 (9)	4 (3)	0	0	13 (13)
Source	TCGA	-	-	-	2 (2)	7 (7)	0
	HN_Cetux	-	-	-	5 (4)	17 (17)	15 (15)
Site	UCLH	663 (389)	100 (51)	21 (19)	0	0	0
	MD Anderson Cancer Clinic	0	0	0	2 (2)	7 (7)	0
	Unknown (US)	0	0	0	5 (4)	17 (17)	15 (15)

Tumour sites are derived from ICD codes. Numbers show number of scans with the number of unique patients in parenthesis. "TCGA": The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma[55], an open source dataset hosted on TCIA. "HN_Cetux": Head-Neck Cetuximab, an open source dataset hosted on TCIA[53]. "PDDCA": Public Domain Database for Computational Anatomy dataset released as part of the 2015 challenge in the segmentation of head and neck anatomy at the International Conference On Medical Image Computing & Computer Assisted Intervention (MICCAI).

4.2 Clinical taxonomy

In order to select which OARs to include in the study, we used the Brouwer Atlas (consensus guidelines for delineating OARs for head and neck radiotherapy, defined by an international panel of radiation oncologists; [77]). From this, we excluded those regions which required additional magnetic resonance imaging for segmentation, were not relevant to routine head and neck radiotherapy, or that were not used clinically at UCLH. This resulted in a set of 21 organs at risk; see Table 3.

4.3 Clinical labelling & annotation

Due to the large variability of segmentation protocols used and annotation quality in the UCLH dataset, all segmentations from all scans selected for inclusion in the training set were manually reviewed by a

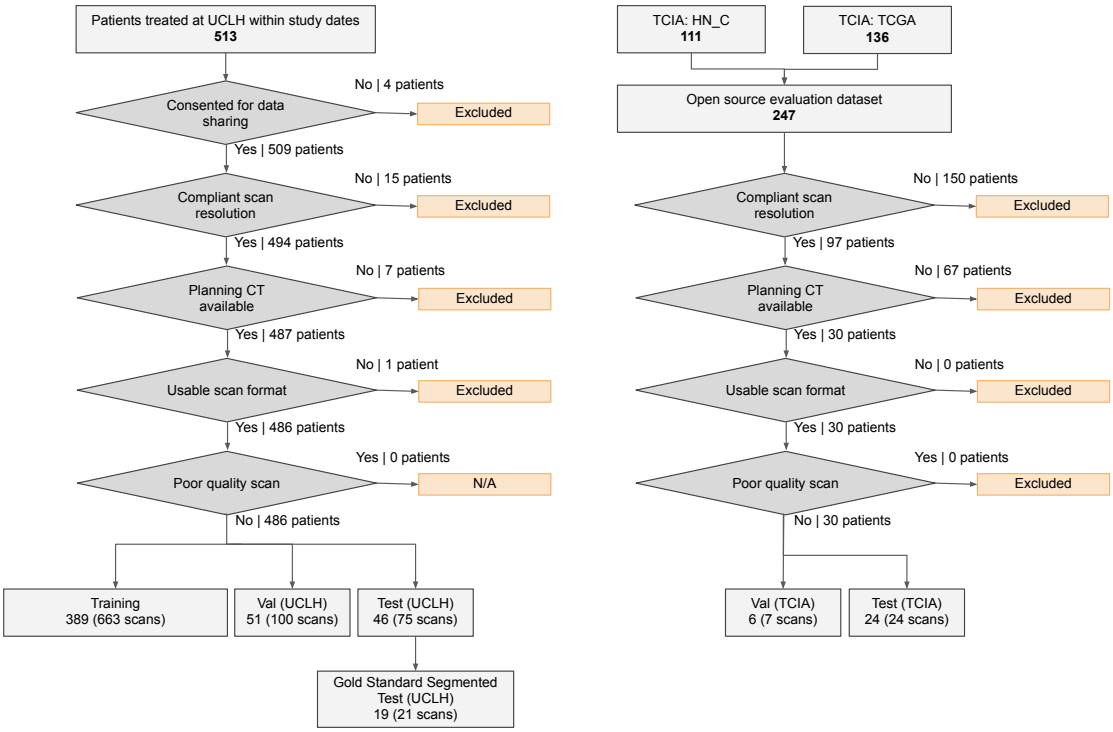

Figure 6 | Case selection from UCLH and TCIA CT datasets. A consort-style diagram demonstrating the application of inclusion and exclusion criteria to select the training, validation (val) and test sets used in this work.

radiographer with at least 4 years experience in the segmentation of head and neck OARs. Volumes that did not conform to the Brouwer Atlas were excluded from training. In order to increase the number of training examples, additional axial slices were randomly selected for further manual OAR segmentations to be added based on model performance or perceived imbalances in the dataset. These were then produced by a radiographer with at least 4 years experience in head and neck radiotherapy, arbitrated by a second radiographer with the same level of experience. The total number of examples from the original UCLH segmentations and the additional slices added are provided in Table 3.

For the TCIA test and validation sets, the original dense segmentations were not used due to poor adherence to the chosen study protocol. To produce the ground truth labels, the full volumes of all 21 OARs included in the study were segmented. This was done initially by a radiographer with at least four years experience in the segmentation of head and neck OARs and then arbitrated by a second radiographer with similar experience. Further arbitration was then performed by a radiation oncologist with at least five years post-certification experience in head and neck radiotherapy. The same process was repeated with two additional radiographers working independently but after peer arbitration these segmentations were not reviewed by an oncologist; rather they became the human reference to which the model was compared. This is shown schematically in Fig. 7. Prior to participation all radiographers and oncologists were required to study the Brouwer Atlas for head and neck OAR segmentation [77] and demonstrate competence in adhering to these guidelines.

Table 3 | Taxonomy of segmentation regions.

OAR	Total number of labelled slices included	Anatomical Landmarks and Definition
Brain	11476	Sits inside the cranium and includes all brain vessels excluding the brainstem and optic chiasm.
Brainstem	34794	The posterior aspect of the brain including the midbrain, pons and medulla oblongata. Extending inferior from the lateral ventricles to the tip of the dens at C2. It is structurally continuous with the spinal cord.
Cochlea-Lt Cochlea-Rt	4526 4754	Embedded in the temporal bone and lateral to the internal auditory meatus.
Lacrimal-Lt Lacrimal-Rt	17186 17788	Concave shaped gland located at the superolateral aspect of the orbit.
Lens-Lt Lens-Rt	3006 3354	An oval structure that sits within the anterior segment of the orbit. Can be variable in position but never sitting posterior beyond the level of the outer canthus.
Lung-Lt Lung-Rt	8340 9158	Encompassed by the thoracic cavity adjacent to the lateral aspect of the mediastinum, extending from the 1st rib to the diaphragm excluding the carina.
Mandible	25074	The entire mandible bone including the temporomandibular joint, ramus and body, excluding the teeth. The mandible joins to the inferior aspect of the temporal bone and forms the entire lower jaw.
Optic-Nerve-Lt Optic-Nerve-Rt	3458 3012	A 2-5mm thick nerve that runs from the posterior aspect of the eye, through the optic canal and ends at the lateral aspect of the optic chiasm.
Orbit-Lt Orbit-Rt	8538 8242	Spherical organ sitting within the orbital cavity. Includes the vitreous humor, retina, cornea and lens with the optic nerve attached posteriorly.
Parotid-Lt Parotid-Rt	8984 11752	Multi lobed salivary gland wrapped around the mandibular ramus. Extends medially to styloid process and parapharyngeal space. Laterally extending to subcutaneous fat. Posteriorly extending to sternocleidomastoid muscle. Anterior extending to posterior border of mandible bone and masseter muscle. In cases where retromandibular vein is encapsulated by parotid this is included in the segmentation.
Spinal-Canal	37000	Hollow cavity that runs through the foramen of the vertebrae, extending from the base of skull to the end of the sacrum.
Spinal-Cord	37096	Sits inside the Spinal Canal and extends from the level of the foramen magnum to the bottom of L2.
Submandibular-Lt Submandibular-Rt	10652 10716	Sits within the submandibular portion of the anterior triangle of the neck, making up the floor of the mouth and extending both superior and inferior to the posterior aspect of the mandible and is limited laterally by the mandible and medially by the hypoglossal muscle.

4.4 Model architecture

We used a residual 3D U-Net architecture with 8 levels (see Fig. 8). Our network takes in a CT volume (single channel) and outputs a segmentation mask with 21 channels, where each channel contains the binary segmentation mask for a specific OAR. The network consists of 7 residual convolutional blocks in the downward path, a residual fully connected block at the bottom, and 7 residual convolutional blocks in the upward path. A 1x1x1 convolution layer with sigmoidal activation produces the final output in the

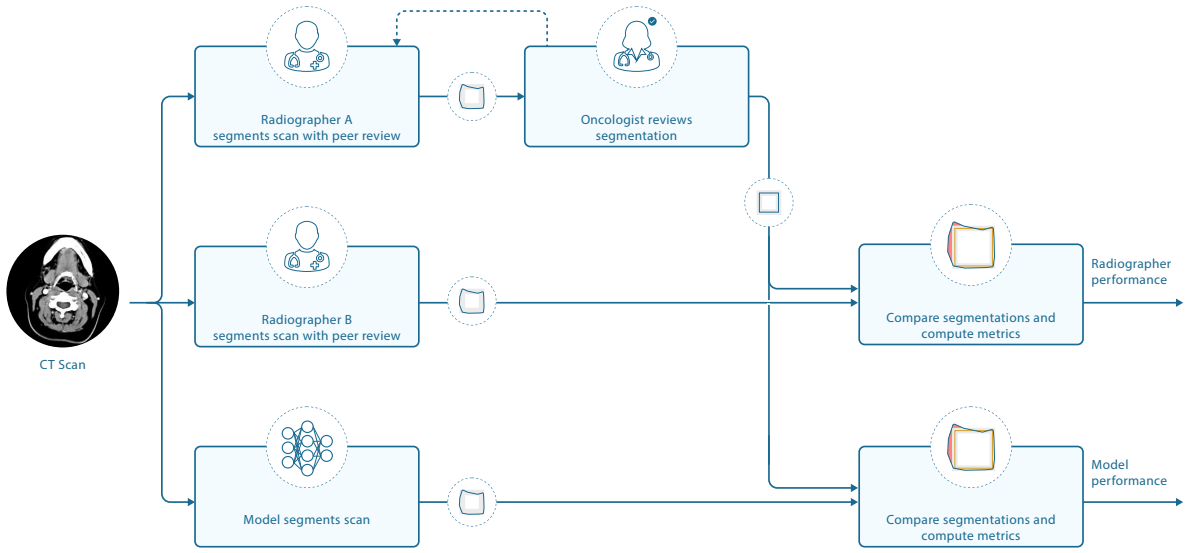

Figure 7 | Process for segmentation of ground truth and radiographer OAR volumes. The flowchart illustrates how the ground truth segmentations were created and compared with independent radiographer segmentations and the model. For the ground truth each CT scan in the TCIA test set was segmented first by a radiographer and peer reviewed by a second radiographer. This then went through one or more iterations of review and editing with a specialist oncologist before creating a ground truth used to compare with the segmentations produced by both the model and additional radiographers.

original resolution of the input image. Each predicted slice has 21 slices of context[3].

We trained our network with a regularised top-k-percent pixel-wise binary cross-entropy loss [78]: for each output channel, the top-k loss selects only the k% most difficult pixels (those with the highest binary cross-entropy), and only adds their contribution to the total loss. This speeds up training and helps the network to tackle the large class imbalance and to focus on difficult examples.

We regularised the model using standard L2 weight regularisation with scale 10^{-6} and extensive data augmentation: we used random in-plane (i.e. in x- and y- directions only) translation, rotation, scaling, shearing, mirroring, elastic deformations, and pixel-wise noise. We used uniform translations between -32 and 32 pixels; uniform rotations between -9 and 9 degrees; uniform scaling factors between 0.8 and 1.2; and uniform shear factors between -0.1 and 0.1. We mirrored images (and adjusted corresponding left and right labels) with a probability of 0.5. We performed elastic deformations by placing random displacement vectors (standard deviation: 5mm, in-plane displacements only) on a control point grid with 100mm x 100mm x 100mm spacing and by deriving the dense deformation field using cubic b-spline interpolation. In the implementation all spatial transformations are first combined to a dense deformation field, which is then applied to the image using bilinear interpolation and extrapolation with zero padding. We added zero mean Gaussian intensity noise independently to each pixel with a standard deviation of 20 Hounsfield Units.

We trained the model with the Adam optimiser [79] for 120,000 steps and a batch size of 32 (32 GPUs) using synchronous SGD. We used an initial learning rate of 10^{-4} and scaled the learning rate by 1/2, 1/8, 1/64, and 1/256 at timesteps 24,000, 60,000, 108,000, and 114,000, respectively.

We used the validation set to select the model which performed at over 95% for the most OARs according to our chosen surface DSC performance metric, breaking ties by preferring better performance on

[3]The 21 slices context (i.e. 21×2.5mm = 52.5mm) were found to provide the optimal context. It has nothing to do with the 21 OARs used in this study.

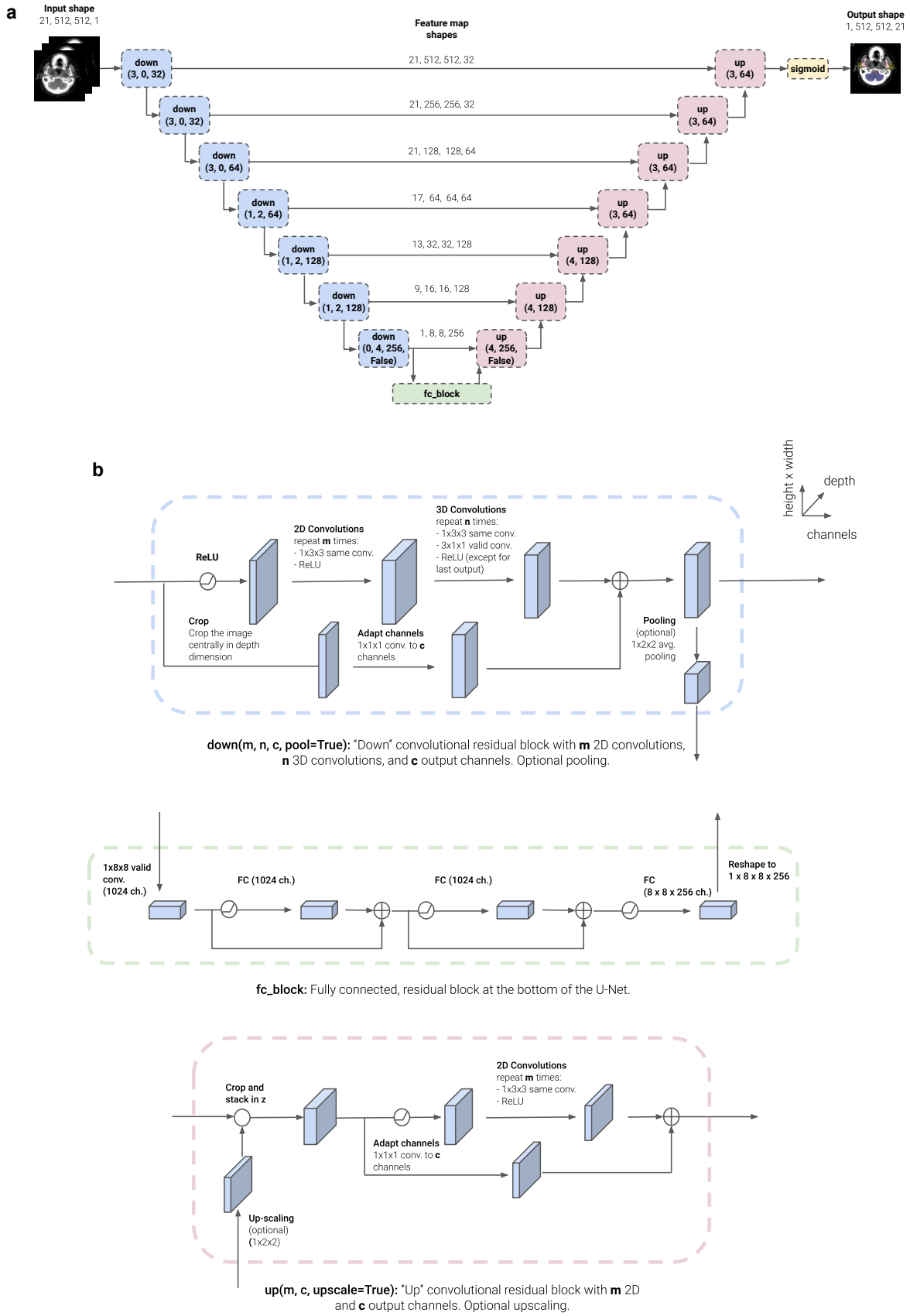

Figure 8 | 3D U-Net model architecture. (**a**) At training time, the model receives 21 contiguous CT slices, which are processed through a series of "down" blocks, a fully connected block, and a series of "up" blocks to create a segmentation prediction. (**b**) A detailed view of the convolutional residual down and up blocks, and the residual fully connected block.

more clinically impactful OARs and the absolute performance obtained.

4.5 Performance metrics

All performance metrics are reported for each organ independently (e.g. separately for just the left parotid), so we only need to deal with binary masks (e.g. a left parotid voxel and a non left-parotid voxel). Masks are defined as a subset of $\mathbb{R}^3$, i.e. $\mathcal{M} \subset \mathbb{R}^3$ (see Fig. 9).

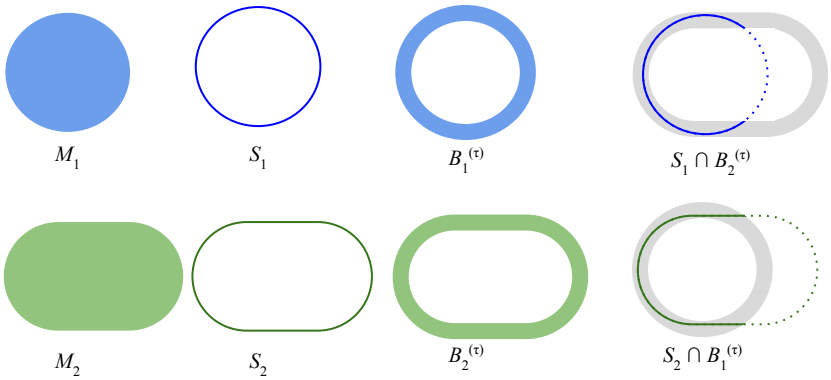

Figure 9 | Illustrations of masks, surfaces, border regions, and the "overlapping" surface at tolerance τ

The volume of a mask is denoted as $|\cdot|$, with

$$|\mathcal{M}| = \int_{\mathcal{M}} d\mathbf{x}\,.$$

With this notation the standard (volumetric) DSC for two given masks $\mathcal{M}_1$ and $\mathcal{M}_2$ can be written as

$$C_{\text{DSC}} = \frac{2|\mathcal{M}_1 \cap \mathcal{M}_2|}{|\mathcal{M}_1| + |\mathcal{M}_2|}\,.$$

In the case of sparse ground truth segmentations (i.e. only a few slices of the CT scan are labelled), we estimate the volumetric DSC by aggregating data from labelled voxels across multiple scans and patients as

$$C_{\text{DSC, est}} = \frac{2\sum_p |\mathcal{M}_{1,p} \cap \mathcal{M}_{2,p} \cap \mathcal{L}_p|}{\sum_p |\mathcal{M}_{1,p} \cap \mathcal{L}_p| + |\mathcal{M}_{2,p} \cap \mathcal{L}_p|}\,,$$

where the mask $\mathcal{M}_{1,p}$ and the labelled region $\mathcal{L}_p$ represent the sparse ground truth segmentation for a patient p and the mask $\mathcal{M}_{2,p}$ is the full volume predicted segmentation for patient p.

Due to the shortcomings of the volumetric DSC metric for the presented radiotherapy use case, we introduce the "surface DSC" metric, which assesses the overlap of two surfaces (at a specified tolerance) instead of the overlap of two volumes (see Results section). A surface is the border of a mask, $\mathcal{S} = \partial\mathcal{M}$, the area of a surface is denoted as

$$|\mathcal{S}| = \int_{\mathcal{S}} d\boldsymbol{\sigma}$$

where $\boldsymbol{\sigma} \in \mathcal{S}$ is a point on the surface, using an arbitrary parameterisation. The mapping from this parameterisation to a point in $\mathbb{R}^3$ is denoted as $\boldsymbol{\xi} : \mathcal{S} \to \mathbb{R}^3$, i.e. $\mathbf{x} = \boldsymbol{\xi}(\boldsymbol{\sigma})$. With this we can define the

border region $\mathcal{B}_i^{(\tau)} \subset \mathbb{R}^3$, for the surface $\mathcal{S}_i$, at a given tolerance τ as

$$\mathcal{B}_i^{(\tau)} = \left\{ \mathbf{x} \in \mathbb{R}^3 \mid \exists \boldsymbol{\sigma} \in \mathcal{S}_i, \|\mathbf{x} - \boldsymbol{\xi}(\boldsymbol{\sigma})\| \leq \tau \right\},$$

(see Fig. 9 for an example). Using these definitions we can write the "surface DSC at tolerance τ" as

$$R_{i,j}^{(\tau)} = \frac{|\mathcal{S}_i \cap \mathcal{B}_j^{(\tau)}| + |\mathcal{S}_j \cap \mathcal{B}_i^{(\tau)}|}{|\mathcal{S}_i| + |\mathcal{S}_j|},$$

using an informal notation for the intersection of the surface with the boundary, i.e.:

$$|\mathcal{S}_i \cap \mathcal{B}_j^{(\tau)}| := \int_{\mathcal{S}_i} \mathbb{1}_{\mathcal{B}_j^{(\tau)}} \left(\boldsymbol{\xi}(\boldsymbol{\sigma})\right) d\boldsymbol{\sigma}$$

4.6 Implementation of surface DSC

The computation of surface integrals on sampled images is not straightforward, especially for medical images, where the voxel spacing is usually not equal in all three dimensions. The common approximation of the integral by counting surface voxels can lead to substantial systematic errors.

Another common challenge is the representation of the surface with voxels. As the surface of a binary mask is located between voxels, a definition of "surface voxels" in the raster-space of the image introduces a bias: using foreground voxels to represent the surface leads to an underestimation of the surface, while the use of background voxels leads to an overestimation.

Our proposed implementation uses a surface representation that provides less biased estimates but still allows us to compute the performance metrics with linear complexity ($\mathcal{O}(N)$, with N : number of voxels). We place the surface points between the voxels on a raster that is shifted by half of the raster spacing on each axis (see Fig. 10 for a 2D illustration). For 3D images, each point in this raster has 8 neighbouring

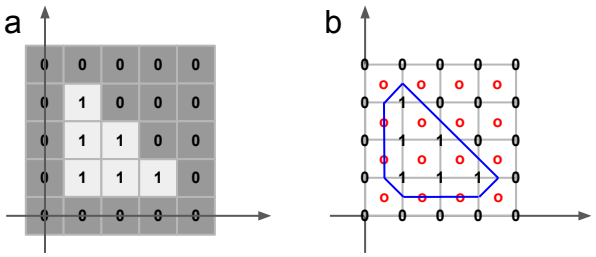

Figure 10 | 2D illustration of the implementation of the surface DSC. (**a**) A binary mask displayed as image. The origin of the image raster is (0,0). (**b**) The surface points (red circles) are located in a raster that is shifted half of the raster spacing on each axis. Each surface point has 4 neighbours in 2D (8 neighbours in 3D). The local contour (blue line) assigned to each surface point (red circle) depends on the neighbour constellation.

voxels. As we analyse binary masks, there are only $2^8 = 256$ possible neighbour constellations. For each of these constellations we compute the resulting triangles using the marching cube triangulation [80] and store the surface area of the triangles (in mm^2) in a look-up table. With this look-up table we then create a surface image (on the above mentioned raster) that contains zeros at positions that have 8 identical neighbours or the local surface area at all positions that have both foreground and background neighbours. These surface images are created for the masks $\mathcal{M}_1$ and $\mathcal{M}_2$. Additionally we create a distance map from each of these surface images using the distance transform algorithm [81]. Iterating over the non-zero elements in the first surface image and looking up the distance from the other surface in

the corresponding distance map allows to create a list of tuples (surface element area, distance from other surface). From this list we can easily compute the surface area by summing up the area of the surface elements that are within the tolerance. To account for the quantised distances – there is only a discrete set $\mathcal{D} = \left\{ \sqrt{(n_1 d_1)^2 + (n_2 d_2)^2 + (n_3 d_3)^2} \mid n_1, n_2, n_3 \in \mathbb{N} \right\}$ of distances between voxels in a 3D raster with spacing (d_1, d_2, d_3) – we also round the tolerance to the nearest neighbour in the set $\mathcal{D}$ for each image before computing the surface DSC. For more details, please have a look at our open source implementation of the surface DSC, available from https://github.com/deepmind/surface-distance.

5 Code availability

The codebase for the deep learning framework makes use of proprietary components and we are unable to publicly release this code. However, all experiments and implementation details are described in sufficient detail in the methods section to allow independent replication with non-proprietary libraries. The surface DSC performance metric code is available at https://github.com/deepmind/surface-distance.

6 Data availability

The clinical data used for training and validation sets were collected and de-identified at University College London Hospitals NHS Foundation Trust. Data were used with both local and national permissions. They are not publicly available and restrictions apply to their use. The data, or a subset, may be available from UCLH NHS Foundation Trust subject to local and national ethical approvals. The released test/validation set data was collected from two datasets hosted on The Cancer Imaging Archive (TCIA). The subset used, along with the ground truth segmentations added is available at https://github.com/deepmind/tcia-ct-scan-dataset.

7 Acknowledgement

We thank the patients treated at UCLH whose scans were used in the work, A. Zisserman, D. King, D. Barrett, V. Cornelius, C. Beltran, J. Cornebise, R. Sharma, J. Ashburner, J. Good and N. Haji for discussions, M. Kosmin for his review of the published literature, A. Warry, U. Johnson, V. Rompokos and the rest of the UCLH Radiotherapy Physics team for work on the data collection, R. West for work on the visuals, C. Game, D. Mitchell and M. Johnson for infrastructure and systems administration, A. Paine at Softwire for engineering support at UCLH, A. Kitchener and the UCLH Information Governance team for support, J. Besley and M. Bawn for legal assistance, K. Ayoub, K. Sullivan and R. Ahmed for initiating and supporting the collaboration, the DeepMind Radiographer Consortium made up of B. Hatchard, Y. McQuinlan, K. Hampton, S. Ireland, K. Fuller, H. Frank, C. Tully, A. Jones and L. Turner, and the rest of the DeepMind team for their support, ideas and encouragement.

G.R. and H.M. were supported by University College London and the National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

8 Author contributions

M.S., T.B., O.R., J.L., R.M., H.M., S.A.M., D.D'S., C.C., & K.S. initiated the project
S.B., R.M., D.C., C.B., D.D'S., C.C. & J.L., created the dataset

S.B, S.N., J.D.F., A.Z., Y.P., C.H., H.A. & O.R. contributed to software engineering
S.N., J.D.F., B.R.P. & O.R. designed the model architectures
D.R.C. manually segmented the images
R.M., D.C., C.B., D.D'S., S.A.M., H.M., G.R., C.H., A.K. & J.L. contributed clinical expertise
C.M., J.L., T.B., S.A.M., K.S. & O.R. managed the project
C.H., C.K., M.L., J.L., S.N., S.B., J.D.F., H.M., G.R. & O.R. wrote the paper

9 Competing financial interests

G.R., H.M. and the D.R.C. were paid contractors of DeepMind and/or Google Health. The authors have no other competing interests to disclose.

References

[1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA Cancer J. Clin.*, vol. 61, no. 2, pp. 69–90, Mar. 2011. Available: http://dx.doi.org/10.3322/caac.20107

[2] Cancer Research UK, "Head and neck cancers incidence statistics," https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/head-and-neck-cancers/incidence#heading-Two, Feb. 2018, accessed: 2018-2-8.

[3] National Cancer Intelligence Network, "NCIN data briefing: Potentially HPV-related head and neck cancers," http://www.ncin.org.uk/publications/data_briefings/potentially_hpv_related_head_and_neck_cancers, May 2012.

[4] Oxford Cancer Intelligence Unit, "Profile of head and neck cancers in england: Incidence, mortality and survival," National Cancer Intelligence Network, Tech. Rep., 2010. Available: http://www.ncin.org.uk/view?rid=69

[5] D. M. Parkin, L. Boyd, and L. C. Walker, "16. the fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010," *Br. J. Cancer*, vol. 105, no. S2, pp. S77–81, 2011. Available: http://dx.doi.org/10.1038/bjc.2011.489

[6] K. Jensen, K. Lambertsen, and C. Grau, "Late swallowing dysfunction and dysphagia after radiotherapy for pharynx cancer: frequency, intensity and correlation with dose and volume parameters," *Radiother. Oncol.*, vol. 85, no. 1, pp. 74–82, Oct. 2007. Available: http://dx.doi.org/10.1016/j.radonc.2007.06.004

[7] P. Dirix, S. Abbeel, B. Vanstraelen, R. Hermans, and S. Nuyts, "Dysphagia after chemoradiotherapy for head-and-neck squamous cell carcinoma: Dose–effect relationships for the swallowing structures," *Int J Radiat Oncol Biol Phys*, vol. 75, no. 2, pp. 385–392, Oct. 2009. Available: https://doi.org/10.1016/j.ijrobp.2008.11.041

[8] J. J. Caudell, P. E. Schaner, R. A. Desmond, R. F. Meredith, S. A. Spencer, and J. A. Bonner, "Dosimetric factors associated with long-term dysphagia after definitive radiotherapy for squamous cell carcinoma of the head and neck," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 76, no. 2, pp. 403–409, Feb. 2010. Available: http://dx.doi.org/10.1016/j.ijrobp.2009.02.017

[9] C. M. Nutting, J. P. Morden, K. J. Harrington, T. G. Urbano, S. A. Bhide, C. Clark, E. A. Miles, A. B. Miah, K. Newbold, M. Tanay, F. Adab, S. J. Jefferies, C. Scrase, B. K. Yap, R. P. A'Hern, M. A. Sydenham, M. Emson, E. Hall, and PARSPORT trial management group, "Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial," *Lancet Oncol.*, vol. 12, no. 2, pp. 127–136, Feb. 2011. Available: http://dx.doi.org/10.1016/S1470-2045(10)70290-4

[10] B. E. Nelms, W. A. Tomé, G. Robinson, and J. Wheeler, "Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 82, no. 1, pp. 368–378, Jan. 2012. Available: http://dx.doi.org/10.1016/j.ijrobp.2010.10.019

[11] P. W. J. Voet, M. L. P. Dirkx, D. N. Teguh, M. S. Hoogeman, P. C. Levendag, and B. J. M. Heijmen, "Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? a dosimetric analysis," *Radiother. Oncol.*, vol. 98, no. 3, pp. 373–377, Mar. 2011. Available: http://dx.doi.org/10.1016/j.radonc.2010.11.017

[12] P. M. Harari, S. Song, and W. A. Tomé, "Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 77, no. 3, pp. 950–958, Jul. 2010. Available: http://dx.doi.org/10.1016/j.ijrobp.2009.09.062

[13] Z. Chen, W. King, R. Pearcey, M. Kerba, and W. J. Mackillop, "The relationship between waiting time for radiotherapy and clinical outcomes: a systematic review of the literature," *Radiother. Oncol.*, vol. 87, no. 1, pp. 3–16, Apr. 2008. Available: http://dx.doi.org/10.1016/j.radonc.2007.11.016

[14] J. S. Mikeljevic, R. Haward, C. Johnston, A. Crellin, D. Dodwell, A. Jones, P. Pisani, and D. Forman, "Trends in postoperative radiotherapy delay and the effect on survival in breast cancer patients treated with conservation surgery," *Br. J. Cancer*, vol. 90, no. 7, pp. 1343–1348, Apr. 2004. Available: http://dx.doi.org/10.1038/sj.bjc.6601693

[15] C. E. Round, M. V. Williams, T. Mee, N. F. Kirkby, T. Cooper, P. Hoskin, and R. Jena, "Radiotherapy demand and activity in england 2006-2020," *Clin. Oncol.*, vol. 25, no. 9, pp. 522–530, Sep. 2013. Available: http://dx.doi.org/10.1016/j.clon.2013.05.005

[16] Z. E. Rosenblatt E, "Radiotherapy in cancer care: Facing the global challenge," International Atomic Energy Agency, Tech. Rep., 2017. Available: https://www-pub.iaea.org/MTCD/Publications/PDF/P1638_web.pdf

[17] C. Veiga, J. McClelland, S. Moinuddin, A. Lourenço, K. Ricketts, A. J, M. Modat, O. S, D. D'Souza, and G. Royle, "Toward adaptive radiotherapy for head and neck patients: Feasibility study on using ct-to-cbct deformable registration for 'dose of the day' calculations," *Med. Phys*, vol. 41, no. 3, p. 031703 (12pp.), 2014. Available: https://doi.org/10.1118/1.4864240

[18] Department of Health, "The NHS Cancer Plan, Chapter 5," 2000.

[19] J.-F. Daisne and A. Blumhofer, "Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation," *Radiat. Oncol.*, vol. 8, p. 154, Jun. 2013. Available: http://dx.doi.org/10.1186/1748-717X-8-154

[20] V. Fortunati, R. F. Verhaart, F. van der Lijn, W. J. Niessen, J. F. Veenland, M. M. Paulides, and T. van Walsum, "Tissue segmentation of head and neck CT images for treatment planning: a multiatlas approach combined with intensity modeling," *Med. Phys.*, vol. 40, no. 7, p. 071905, Jul. 2013. Available: http://dx.doi.org/10.1118/1.4810971

[21] H. Duc, K. Albert, G. Eminowicz, R. Mendes, S.-L. Wong, J. McClelland, M. Modat, M. J. Cardoso, A. F. Mendelson, C. Veiga, and Others, "Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer," *Med. Phys.*, vol. 42, no. 9, pp. 5027–5034, 2015. Available: https://doi.org/10.1118/1.4927567

[22] M. S. Hoogeman, X. Han, D. Teguh, P. Voet, P. Nowak, T. Wolf, L. Hibbard, B. Heijmen, and P. Levendag, "Atlas-based auto-segmentation of CT images in head and neck cancer: What is the best approach?" *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 72, no. 1, p. S591, Sep. 2008. Available: https://doi.org/10.1016/j.ijrobp.2008.06.196

[23] P. C. Levendag, M. Hoogeman, D. Teguh, T. Wolf, L. Hibbard, O. Wijers, B. Heijmen, P. Nowak, E. Vasquez-Osorio, and X. Han, "Atlas based auto-segmentation of CT images: clinical evaluation of using auto-contouring in high-dose, high-precision radiotherapy of cancer in the head and neck," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 72, no. 1, p. S401, Sep. 2008. Available: https://doi.org/10.1016/j.ijrobp.2008.06.1285

[24] A. A. Qazi, V. Pekar, J. Kim, J. Xie, S. L. Breen, and D. A. Jaffray, "Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach," *Med Phys*, vol. 38, no. 11, pp. 6160–6170, 2011. Available: https://doi.org/10.1118/1.3654160

[25] R. Sims, A. Isambert, V. Grégoire, F. Bidault, L. Fresco, J. Sage, J. Mills, J. Bourhis, D. Lefkopoulos, O. Commowick, M. Benkebil, and G. Malandain, "A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck," *Radiother. Oncol.*, vol. 93, no. 3, pp. 474–478, Dec. 2009. Available: http://dx.doi.org/10.1016/j.radonc.2009.08.013

[26] D. N. Teguh, P. C. Levendag, P. W. J. Voet, A. Al-Mamgani, X. Han, T. K. Wolf, L. S. Hibbard, P. Nowak, H. Akhiat, M. L. P. Dirkx, B. J. M. Heijmen, and M. S. Hoogeman, "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.*, vol. 81, no. 4, pp. 950–957, Nov. 2011. Available: http://dx.doi.org/10.1016/j.ijrobp.2010.07.009

[27] D. Thomson, C. Boylan, T. Liptrot, A. Aitkenhead, L. Lee, B. Yap, A. Sykes, C. Rowbottom, and N. Slevin, "Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk," *Radiat. Oncol.*, vol. 9, p. 173, Aug. 2014. Available: http://dx.doi.org/10.1186/1748-717X-9-173

[28] G. V. Walker, M. Awan, R. Tao, E. J. Koay, N. S. Boehling, J. D. Grant, D. F. Sittig, G. B. Gunn, A. S. Garden, J. Phan, W. H. Morrison, D. I. Rosenthal, A. S. R. Mohamed, and C. D. Fuller, "Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer," *Radiother. Oncol.*, vol. 112, no. 3, pp. 321–325, Sep. 2014. Available: http://dx.doi.org/10.1016/j.radonc.2014.08.028

[29] S. J. Gacha and S. A. G. León, "Segmentation of mandibles in computer tomography volumes of patients with foam cells carcinoma," in *2018 IX International Seminar of Biomedical Engineering (SIB)*, May 2018, pp. 1–7.

[30] P. F. Raudaschl, P. Zaffino, G. C. Sharp, M. F. Spadea, A. Chen, B. M. Dawant, T. Albrecht, T. Gass, C. Langguth, M. Lüthi, and Others, "Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015," *Med. Phys.*, vol. 44, no. 5, pp. 2020–2036, 2017. Available: https://doi.org/10.1002/mp.12197

[31] X. Wu, J. K. Udupa, Y. Tong, D. Odhner, G. V. Pednekar, C. B. Simone, D. McLaughlin, C. Apinorasethkul, O. Apinorasethkul, J. Lukens, D. Mihailidis, G. Shammo, P. James, A. Tiwari, L. Wojtowicz, J. Camaratta, and D. A. Torigian, "Aar-rt - a system for auto-contouring organs at risk on ct images for radiation therapy planning: Principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases," *Medical Image Analysis*, 2019. Available: http://www.sciencedirect.com/science/article/pii/S1361841518305668

[32] K. Fritscher, P. Raudaschl, P. Zaffino, M. F. Spadea, G. C. Sharp, and R. Schubert, "Deep neural networks for fast segmentation of 3D medical images," in *Med Image Comput Comput Assist Interv*. Springer International Publishing, 2016, pp. 158–165. Available: http://dx.doi.org/10.1007/978-3-319-46723-8_19

[33] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks," *Med. Phys.*, vol. 44, no. 2, pp. 547–557, 2017. Available: https://doi.org/10.1002/mp.12045

[34] D. Močnik, B. Ibragimov, L. Xing, P. Strojan, B. Likar, F. Pernuš, and T. Vrtovec, "Segmentation of parotid glands from registered CT and MR images," *Phys. Med.*, vol. 52, pp. 33–41, Aug. 2018. Available: https://doi.org/10.1016/j.ejmp.2018.06.012

[35] X. Ren, L. Xiang, D. Nie, Y. Shao, H. Zhang, D. Shen, and Q. Wang, "Interleaved 3D-CNNs for joint segmentation of small-volume structures in head and neck CT images," *Med. Phys.*, vol. 45, no. 5, pp. 2063–2075, May 2018. Available: http://dx.doi.org/10.1002/mp.12837

[36] T. Zhong, X. Huang, F. Tang, S. Liang, X. Deng, and Y. Zhang, "Boosting-based cascaded convolutional neural networks for the segmentation of ct organs-at-risk in nasopharyngeal carcinoma," *Medical physics*, vol. 46, no. 12, pp. 5602–5611, 2019. Available: https://doi.org/10.1002/mp.13825

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Med Image Comput Comput Assist Interv*. Springer International Publishing, 2015, pp. 234–241. Available: http://dx.doi.org/10.1007/978-3-319-24574-4_28

[38] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.*, vol. 24, pp. 1342—1350, Aug. 2018. Available: http://dx.doi.org/10.1038/s41591-018-0107-6

[39] A. Hänsch, M. Schwier, T. Gass, T. Morgas, B. Haas, J. Klein, and H. K. Hahn, "Comparison of different deep learning approaches for parotid gland segmentation from CT images," in *Med Imaging Comp-Aided Diag*, vol. 10575. International Society for Optics and Photonics, Feb. 2018, p. 1057519. Available: https://doi.org/10.1117/12.2292962

[40] W. Zhu, Y. Huang, H. Tang, Z. Qian, N. Du, W. Fan, and X. Xie, "AnatomyNet: Deep 3D squeeze-and-excitation U-Nets for fast and fully automated whole-volume anatomical segmentation," Aug. 2018, arXiv preprint arXiv:1808.05238.

[41] N. Tong, S. Gou, S. Yang, D. Ruan, and K. Sheng, "Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks," *Medical Physics*, vol. 0, no. ja, 2018. Available: https://doi.org/10.1002/mp.13147

[42] S. Liang, F. Tang, X. Huang, K. Yang, T. Zhong, R. Hu, S. Liu, X. Yuan, and Y. Zhang, "Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning," *European Radiology*, Oct 2018. Available: https://doi.org/10.1007/s00330-018-5748-9

[43] S. Willems, W. Crijns, A. La Greca Saint-Esteven, J. Van Der Veen, D. Robben, T. Depuydt, S. Nuyts, K. Haustermans, and F. Maes, "Clinical implementation of deepvoxnet for auto-delineation of organs at risk in head and neck cancer patients in radiotherapy," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, D. Stoyanov, Z. Taylor, D. Sarikaya, J. McLeod, M. A. González Ballester, N. C. Codella, A. Martel, L. Maier-Hein, A. Malpani, M. A. Zenati, S. De Ribaupierre, L. Xiongbiao, T. Collins, T. Reichl, K. Drechsler, M. Erdt, M. G. Linguraru, C. Oyarzun Laura, R. Shekhar, S. Wesarg, M. E. Celebi, K. Dana, and A. Halpern, Eds. Cham: Springer International Publishing, 2018, pp. 223–232. Available: https://link.springer.com/chapter/10.1007/978-3-030-01201-4_24

[44] O. Kodym, M. Španěl, and A. Herout, "Segmentation of head and neck organs at risk using cnn with batch dice loss," Dec. 2018, arXiv preprint arXiv:1812.02427.

[45] Y. Wang, L. Zhao, M. Wang, and Z. Song, "Organ at risk segmentation in head and neck ct images using a two-stage segmentation framework based on 3d u-net," *IEEE Access*, vol. 7, pp. 144 591–144 602, 2019. Available: https://doi.org/10.1109/ACCESS.2019.2944958

[46] K. Men, H. Geng, C. Cheng, H. Zhong, M. Huang, Y. Fan, J. P. Plastaras, A. Lin, and Y. Xiao, "Technical note: More accurate and efficient segmentation of organs-at-risk in radiotherapy with convolutional neural networks cascades," *Medical Physics*, vol. 46, no. 1, pp. 286–292, 2019. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13296

[47] E. Tappeiner, S. Pröll, M. Hönig, P. F. Raudaschl, P. Zaffino, M. F. Spadea, G. C. Sharp, R. Schubert, and K. Fritscher, "Multi-organ segmentation of the head and neck area: an efficient hierarchical neural networks approach," *International journal of computer assisted radiology and surgery*, vol. 14, no. 5, pp. 745–754, 2019. Available: https://doi.org/10.1007/s11548-019-01922-4

[48] D. J. Rhee, C. E. Cardenas, H. Elhalawani, R. McCarroll, L. Zhang, J. Yang, A. S. Garden, C. B. Peterson, B. M. Beadle, and L. E. Court, "Automatic detection of contouring errors using convolutional neural networks," *Medical physics*, vol. 46, no. 11, pp. 5086–5097, 2019. Available: https://doi.org/10.1002/mp.13814

[49] H. Tang, X. Chen, Y. Liu, Z. Lu, J. You, M. Yang, S. Yao, G. Zhao, Y. Xu, T. Chen *et al.*, "Clinically applicable deep learning framework for organs at risk delineation in ct images," *Nature Machine Intelligence*, vol. 1, no. 10, pp. 480–491, 2019. Available: https://doi.org/10.1038/s42256-019-0099-z

[50] W. van Rooij, M. Dahele, H. R. Brandao, A. R. Delaney, B. J. Slotman, and W. F. Verbakel, "Deep learning-based delineation of head and neck organs at risk: Geometric and dosimetric evaluation," *International Journal of Radiation Oncology* Biology* Physics*, vol. 104, no. 3, pp. 677–684, 2019. Available: https://doi.org/10.1016/j.ijrobp.2019.02.040

[51] S. Gou, N. Tong, S. X. Qi, S. Yang, R. K. Chin, and K. Sheng, "Self-channel-and-spatial-attention neural network for automated multi-organ segmentation on head and neck ct images," *Physics in Medicine & Biology*, 2020. Available: https://doi.org/10.1088/1361-6560/ab79c3

[52] R. H. Mak, M. G. Endres, J. H. Paik, R. A. Sergeev, H. Aerts, C. L. Williams, K. R. Lakhani, and E. C. Guinan, "Use of Crowd Innovation to Develop an Artificial Intelligence–Based Solution for Radiation Therapy Targeting," *JAMA Oncology*, vol. 5, no. 5, pp. 654–661, 05 2019. Available: https://doi.org/10.1001/jamaoncol.2019.0159

[53] W. R. Bosch, W. L. Straube, J. W. Matthews, and J. A. Purdy, "Head-neck cetuximab - the cancer imaging archive," 2015. Available: https://wiki.cancerimagingarchive.net/display/Public/Head-Neck+Cetuximab

[54] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): maintaining and operating a public information repository," *J Digit Imaging*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013. Available: http://dx.doi.org/10.1007/s10278-013-9622-7

[55] M. L. Zuley, R. Jarosz, S. Kirk, L. Y., R. Colen, K. Garcia, and N. D. Aredes, "Radiology data from the cancer genome atlas head-neck squamous cell carcinoma [TCGA-HNSC] collection," 2016. Available: http://dx.doi.org/10.7937/K9/TCIA.2016.LXKQ47MS

[56] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945. Available: http://doi.wiley.com/10.2307/1932409

[57] T. Hong, W. Tome, R. Chappell, and P. Harari, "Variations in target delineation for head and neck imrt: An international multi-institutional study," *International Journal of Radiation Oncology*Biology*Physics*, vol. 60, no. 1, Supplement, pp. S157 – S158, 2004. Available: http://www.sciencedirect.com/science/article/pii/S0360301604011307

[58] S. Liang, K.-H. Thung, D. Nie, Y. Zhang, and D. Shen, "Multi-view spatial aggregation framework for joint localization and segmentation of organs at risk in head and neck ct images," *IEEE Transactions on Medical Imaging*, 2020. Available: https://doi.org/10.1109/TMI.2020.2975853

[59] B. Qiu, J. Guo, J. Kraeima, H. H. Glas, R. J. Borra, M. J. Witjes, and P. van Ooijen, "Recurrent convolutional neural networks for mandible segmentation from computed tomography," *arXiv preprint arXiv:2003.06486*, 2020. Available: https://arxiv.org/abs/2003.06486

[60] S. Sun, Y. Liu, N. Bai, H. Tang, X. Chen, Q. Huang, Y. Liu, and X. Xie, "Attentionanatomy: A unified framework for whole-body organs at risk segmentation using multiple partially annotated datasets," *arXiv preprint arXiv:2001.04446*, 2020. Available: https://arxiv.org/abs/2001.04446

[61] L. V. van Dijk, L. Van den Bosch, P. Aljabar, D. Peressutti, S. Both, R. J. Steenbakkers, J. A. Langendijk, M. J. Gooding, and C. L. Brouwer, "Improving automatic delineation for head and neck organs at risk by deep learning contouring," *Radiotherapy and Oncology*, vol. 142, pp. 115–123, 2020. Available: https://doi.org/10.1016/j.radonc.2019.09.022

[62] J. Wong, A. Fong, N. McVicar, S. Smith, J. Giambattista, D. Wells, C. Kolbeck, J. Giambattista, L. Gondara, and A. Alexander, "Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning," *Radiotherapy and Oncology*, vol. 144, pp. 152–158, 2020. Available: https://doi.org/10.1016/j.radonc.2019.10.019

[63] J. W. Chan, V. Kearney, S. Haaf, S. Wu, M. Bogdanov, M. Reddick, N. Dixit, A. Sudhyadhom, J. Chen, S. S. Yom *et al.*, "A convolutional neural network algorithm for automatic segmentation of head and neck organs at risk using deep lifelong learning," *Medical physics*, vol. 46, no. 5, pp. 2204–2213, 2019. Available: https://doi.org/10.1002/mp.13495

[64] Y. Gao, R. Huang, M. Chen, Z. Wang, J. Deng, Y. Chen, Y. Yang, J. Zhang, C. Tao, and H. Li, "Focusnet: Imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck ct images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 829–838. Available: https://doi.org/10.1007/978-3-030-32248-9_92

[65] J. Jiang, E. Sharif, H. Um, S. Berry, and H. Veeraraghavan, "Local block-wise self attention for normal organ segmentation," *arXiv preprint arXiv:1909.05054*, 2019. Available: https://arxiv.org/abs/1909.05054

[66] W. Lei, H. Wang, R. Gu, S. Zhang, S. Zhang, and G. Wang, "Deepigeos-v2: Deep interactive segmentation of multiple organs from head and neck images with lightweight cnns," in *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*. Springer, 2019, pp. 61–69. Available: https://link.springer.com/content/pdf/10.1007/978-3-030-33642-4.pdf#page=76

[67] Y. Sun, H. Shi, S. Zhang, P. Wang, W. Zhao, X. Zhou, and K. Yuan, "Accurate and rapid ct image segmentation of the eyes and surrounding organs for precise radiotherapy," *Medical physics*, vol. 46, no. 5, pp. 2214–2222, 2019. Available: https://doi.org/10.1002/mp.13463

[68] N. Tong, S. Gou, S. Yang, M. Cao, and K. Sheng, "Shape constrained fully convolutional densenet with adversarial training for multiorgan segmentation on head and neck ct and low-field mr images," *Medical physics*, vol. 46, no. 6, pp. 2669–2682, 2019. Available: https://doi.org/10.1002/mp.13553

[69] Y. Xue, H. Tang, Z. Qiao, G. Gong, Y. Yin, Z. Qian, C. Huang, W. Fan, and X. Huang, "Shape-aware organ segmentation by predicting signed distance maps," *arXiv preprint arXiv:1912.03849*, 2019. Available: https://arxiv.org/abs/1912.03849

[70] S. Nikolov, S. Blackwell, R. Mendes, J. De Fauw, C. Meyer, C. Hughes, H. Askham, B. Romera-Paredes, A. Karthikesalingam, C. Chu *et al.*, "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," *arXiv preprint arXiv:1809.04430*, 2018. Available: https://arxiv.org/abs/1809.04430v1

[71] E. J. Wuthrick, Q. Zhang, M. Machtay, D. I. Rosenthal, P. F. Nguyen-Tan, A. Fortin, C. L. Silverman, A. Raben, H. E. Kim, E. M. Horwitz, N. E. Read, J. Harris, Q. Wu, Q.-T. Le, and M. L. Gillison, "Institutional clinical trial accrual volume and survival of patients with head and neck cancer," *Journal of Clinical Oncology*, vol. 33, no. 2, pp. 156–164, 2015, pMID: 25488965. Available: https://doi.org/10.1200/JCO.2014.56.5218

[72] F. Vaassen, C. Hazelaar, A. Vaniqui, M. Gooding, B. van der Heyden, R. Canters, and W. van Elmpt, "Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy," *Physics and Imaging in Radiation Oncology*, vol. 13, pp. 1–6, Jan 2020. Available: https://doi.org/10.1016/j.phro.2019.12.001

[73] K. Kiser, A. Barman, S. Stieb, C. D. Fuller, and L. Giancardo, "Novel autosegmentation spatial similarity metrics capture the time required to correct segmentations better than traditional metrics in a thoracic cavity segmentation workflow," *medRxiv*, 2020. Available: https://www.medrxiv.org/content/early/2020/05/18/2020.05.14.20102103

[74] G. Sharp, K. D. Fritscher, V. Pekar, M. Peroni, N. Shusharina, H. Veeraraghavan, and J. Yang, "Vision 20/20: perspectives on automated image segmentation for radiotherapy," *Med. Phys.*, vol. 41, no. 5, p. 050902, May 2014. Available: http://dx.doi.org/10.1118/1.4871620

[75] M. Kosmin, J. Ledsam, B. Romera-Paredes, R. Mendes, S. Moinuddin, D. de Souza, L. Gunn, C. Kelly, C. Hughes, A. Karthikesalingam *et al.*, "Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer," *Radiotherapy and Oncology*, vol. 135, pp. 130–140, 2019.

[76] D. Guo, D. Jin, Z. Zhu, T.-Y. Ho, A. P. Harrison, C.-H. Chao, J. Xiao, A. Yuille, C.-Y. Lin, and L. Lu, "Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search," *arXiv preprint arXiv:2004.08426*, 2020.

[77] C. L. Brouwer, R. J. H. M. Steenbakkers, J. Bourhis, W. Budach, C. Grau, V. Grégoire, M. van Herk, A. Lee, P. Maingon, C. Nutting, B. O'Sullivan, S. V. Porceddu, D. I. Rosenthal, N. M. Sijtsema, and J. A. Langendijk, "CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG oncology and TROG consensus guidelines," *Radiother. Oncol.*, vol. 117, no. 1, pp. 83–90, Oct. 2015. Available: http://dx.doi.org/10.1016/j.radonc.2015.07.041

[78] Z. Wu, C. Shen, and A. v. D. Hengel, "Bridging category-level and instance-level semantic image segmentation," May 2016, arXiv preprint arXiv:1605.06885v1.

[79] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," Dec. 2014, arXiv preprint arXiv:1412.6980.

[80] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *Comp Graph*, vol. 21, no. 4, pp. 163–169, 1987. Available: http://doi.acm.org/10.1145/37401.37422

[81] P. F. Felzenszwalb and D. P. Huttenlocher, "Distance transforms of sampled functions," *Theory Comput.*, vol. 8, no. 19, pp. 415–428, 2012. Available: http://dx.doi.org/10.4086/toc.2012.v008a019

[82] K. D. Fritscher, M. Peroni, P. Zaffino, M. F. Spadea, R. Schubert, and G. Sharp, "Automatic segmentation of head and neck ct images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours," *Medical physics*, vol. 41, no. 5, p. 051910, 2014. Available: https://doi.org/10.1118/1.4871623

[83] C. M. Tam, X. Yang, S. Tian, X. Jiang, J. J. Beitler, and S. Li, "Automated delineation of organs-at-risk in head and neck CT images using multi-output support vector regression," in *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10578. International Society for Optics and Photonics, Mar. 2018, p. 1057824. Available: https://doi.org/10.1117/12.2292556

[84] Z. Wang, L. Wei, L. Wang, Y. Gao, W. Chen, and D. Shen, "Hierarchical vertex regression-based segmentation of head and neck CT images for radiotherapy planning," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 923–937, Feb. 2018. Available: http://dx.doi.org/10.1109/TIP.2017.2768621

[85] N. Torosdagli, D. K. Liberton, P. Verma, M. Sincan, J. Lee, S. Pattanaik, and U. Bagci, "Robust and fully automated segmentation of mandible from ct scans," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 1209–1212. Available: https://doi.org/10.1109/ISBI.2017.7950734

[86] Z. Wang, X. Liu, and W. Chen, "Head and neck ct atlases alignment based on anatomical priors constraint," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 9, pp. 2004–2011, 2019. Available: https://doi.org/10.1166/jmihi.2019.2844

[87] A. Ayyalusamy, S. Vellaiyan, S. Subramanian, A. Ilamurugu, S. Satpathy, M. Nauman, G. Katta, and A. Madineni, "Auto-segmentation of head and neck organs at risk in radiotherapy and its dependence on anatomic similarity," *Radiation oncology journal*, vol. 37, no. 2, p. 134, 2019. Available: https://doi.org/10.3857/roj.2019.00038

[88] R. Haq, S. L. Berry, J. O. Deasy, M. Hunt, and H. Veeraraghavan, "Dynamic multiatlas selection-based consensus segmentation of head and neck structures from ct images," *Medical physics*, vol. 46, no. 12, pp. 5612–5622, 2019. Available: https://doi.org/10.1002/mp.13854

[89] R. E. McCarroll, B. M. Beadle, P. A. Balter, H. Burger, C. E. Cardenas, S. Dalvie, D. S. Followill, K. D. Kisling, M. Mejia, K. Naidoo *et al.*, "Retrospective validation and clinical implementation of automated contouring of organs at risk in the head and neck: A step toward automated radiation treatment planning for low-and middle-income countries," *Journal of global oncology*, vol. 4, pp. 1–11, 2018. Available: https://doi.org/10.1200/JGO.18.00055

[90] Q. Liu, A. Qin, J. Liang, and D. Yan, "Evaluation of atlas-based auto-segmentation and deformable propagation of organs-at-risk for head-and-neck adaptive radiotherapy," *Recent Patents Top Imaging*, vol. 5, pp. 79–87, 2016. Available: https://www.researchgate.net/profile/An_Qin2/publication/304143072_Evaluation_of_Atlas-Based_Auto-Segmentation_and_Deformable_Propagation_of_Organs-at-Risk_for_Head-and-Neck_Adaptive_Radiotherapy/links/5bd8b8fda6fdcc3a8db1722c/Evaluation-of-Atlas-Based-Auto-Segmentation-and-Deformable-Propagation-of-Organs-at-Risk-for-Head-and-N pdf

[91] A. K. Hoang Duc, G. Eminowicz, R. Mendes, S.-L. Wong, J. McClelland, M. Modat, M. J. Cardoso, A. F. Mendelson, C. Veiga, T. Kadir *et al.*, "Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer," *Medical physics*, vol. 42, no. 9, pp. 5027–5034, 2015. Available: https://doi.org/10.1118/1.4927567

[92] C.-J. Tao, J.-L. Yi, N.-Y. Chen, W. Ren, J. Cheng, S. Tung, L. Kong, S.-J. Lin, J.-J. Pan, G.-S. Zhang *et al.*, "Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: a multi-institution clinical study," *Radiotherapy and Oncology*, vol. 115, no. 3, pp. 407–411, 2015. Available: https://doi.org/10.1016/j.radonc.2015.05.012

[93] C. Wachinger, K. Fritscher, G. Sharp, and P. Golland, "Contour-driven atlas-based segmentation," *IEEE transactions on medical imaging*, vol. 34, no. 12, pp. 2492–2505, 2015. Available: https://doi.org/10.1109/TMI.2015.2442753

[94] M. Zhu, K. Bzdusek, C. Brink, J. G. Eriksen, O. Hansen, H. A. Jensen, H. A. Gay, W. Thorstad, J. Widder, C. L. Brouwer *et al.*, "Multi-institutional quantitative evaluation and clinical validation of smart probabilistic image contouring engine (spice) autosegmentation of target structures and

normal tissues on computer tomography images in the head and neck, thorax, liver, and male pelvis areas," *International Journal of Radiation Oncology* Biology* Physics*, vol. 87, no. 4, pp. 809–816, 2013. Available: https://doi.org/10.1016/j.ijrobp.2013.08.007

[95] X. Han, L. S. Hibbard, N. P. O'Connell, and V. Willcut, "Automatic segmentation of parotids in head and neck ct images using multi-atlas fusion," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 297–304, 2010. Available: https://www.researchgate.net/profile/Lyndon_Hibbard/publication/228519091_Automatic_Segmentation_of_Parotids_in_Head_and_Neck_CT_Images_using_Multi-atlas_Fusion/links/0deec516d54dfccb97000000.pdf

[96] R. Sims, A. Isambert, V. Grégoire, F. Bidault, L. Fresco, J. Sage, J. Mills, J. Bourhis, D. Lefkopoulos, O. Commowick *et al.*, "A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck," *Radiotherapy and Oncology*, vol. 93, no. 3, pp. 474–478, 2009. Available: https://doi.org/10.1016/j.radonc.2009.08.013

[97] X. Han, M. S. Hoogeman, P. C. Levendag, L. S. Hibbard, D. N. Teguh, P. Voet, A. C. Cowen, and T. K. Wolf, "Atlas-based auto-segmentation of head and neck ct images," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2008, pp. 434–441. Available: https://link.springer.com/chapter/10.1007/978-3-540-85990-1_52

[98] C. Huang, M. Badiei, H. Seo, M. Ma, X. Liang, D. Capaldi, M. Gensheimer, and L. Xing, "Atlas based segmentations via semi-supervised diffeomorphic registrations," *arXiv preprint arXiv:1911.10417*, 2019. Available: https://arxiv.org/abs/1911.10417

[99] N. Hardcastle, W. A. Tomé, D. M. Cannon, C. L. Brouwer, P. W. Wittendorp, N. Dogan, M. Guckenberger, S. Allaire, Y. Mallya, P. Kumar *et al.*, "A multi-institution evaluation of deformable image registration algorithms for automatic organ delineation in adaptive head and neck radiotherapy," *Radiation oncology*, vol. 7, no. 1, p. 90, 2012. Available: https://link.springer.com/article/10.1186/1748-717X-7-90

[100] M. La Macchia, F. Fellin, M. Amichetti, M. Cianchetti, S. Gianolini, V. Paola, A. J. Lomax, and L. Widesott, "Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer," *Radiation Oncology*, vol. 7, no. 1, p. 160, 2012. Available: https://link.springer.com/article/10.1186/1748-717X-7-160

[101] T. Zhang, Y. Chi, E. Meldolesi, and D. Yan, "Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy," *International Journal of Radiation Oncology* Biology* Physics*, vol. 68, no. 2, pp. 522–530, 2007. Available: https://doi.org/10.1016/j.ijrobp.2007.01.038

10 Appendix

Table 4 | Surface DSC on TCIA data set

Organ	M/H	0522c 0017	0522c 0057	0522c 0161	0522c 0226	0522c 0248	0522c 0251	0522c 0331	0522c 0416	0522c 0419	0522c 0427	0522c 0457	0522c 0479	0522c 0629	0522c 0659	0522c 0667	0522c 0669	0522c 0708	0522c 0768	0522c 0770	0522c 0773	0522c 0845	TCGA-CV-7236	TCGA-CV-7243	TCGA-CV-7245	TCGA-CV-A6JO	TCGA-CV-A6JY	TCGA-CV-A6K0	TCGA-CV-A6K1	mean, stddev	diff.
Brain (84054.9 mm^2)	(M)	95	94	96	94	96	94	96	98	94	71	98	95	97	94	96	97	97	98	91	94	97	97	97	97	97	96	96	94	94.9±4.8	-1.3
	(H)	97	96	97	95	93	96	96	97	95	96	98	96	97	97	97	96	96	98	96	97	97	98	96	96	97	96	95	95	96.3±1.0	
Brainstem (6531.0 mm^2)	(M)	84	72	92	79	86	65	92	94	87	62	99	93	97	99	99	100	99	98	72	82	95	97	98	87	98	96	94	94	89.5±10.5	-7.5
	(H)	96	97	88	96	93	99	94	97	97	93	100	96	97	100	100	99	98	100	97	97	99	98	98	98	99	98	98	98	97.1±2.5	
Cochlea-Lt (93.3 mm^2)	(M)	98	99	97	99	100	100	100	100	97	96	95	100	100	85	84	98	97	100	100	100	92	100	100	100	98	100	95	100	97.6±4.1	2.4
	(H)	100	100	90	94	94	91	100	90	100	94	100	99	94	85	83	87	97	100	99	100	92	93	91	100	100	100	92	100	95.2±5.1	
Cochlea-Rt (85.5 mm^2)	(M)	100	99	99	97	100	100	97	100	94	95	98	93	100	98	100	100	98	100	99	100	100	100	88	99	95	100	100	100	98.2±2.7	8.6
	(H)	100	100	0	100	95	88	90	100	100	94	100	100	100	99	8	100	95	100	98	95	83	93	77	95	100	100	100	99	89.6±24.5	
Lacrimal-Lt (535.1 mm^2)	(M)	(99)	(92)	(97)	(89)	(73)	(86)	(97)	(97)	(83)	(87)	(98)	(81)	(93)	(97)	(96)	(99)	(98)	(99)	(96)	(98)	(100)	(99)	(98)	(100)	(98)	(100)	(96)	(98)	94.4±6.6	-0.1
	(H)	(100)	(91)	(85)	(98)	(88)	(100)	(92)	(90)	(86)	(100)	(100)	(95)	(100)	(98)	(99)	(99)	(99)	(95)	(91)	(88)	(93)	(93)	(89)	(96)	(98)	(94)	(95)	(98)	94.6±4.7	
Lacrimal-Rt (553.9 mm^2)	(M)	(97)	(99)	(99)	(92)	(75)	(86)	(88)	(87)	(96)	(82)	(89)	(96)	(88)	(99)	(95)	(100)	(85)	(97)	(100)	(98)	(90)	(98)	(100)	(96)	(91)	(97)	(93)	(92)	93.1±6.1	-0.2
	(H)	(100)	(98)	(96)	(98)	(91)	(95)	(84)	(82)	(97)	(99)	(100)	(88)	(96)	(98)	(99)	(100)	(83)	(96)	(94)	(94)	(89)	(99)	(93)	(92)	(93)	(82)	(93)	(83)	93.3±5.8	
Lens-Lt (193.5 mm^2)	(M)	100	96	93	100	96	100	100	94	100	100	100	0	100	93	95	100	100	94	95	97	97	100	98	99	96	99	94	100	94.1±18.3	-4.2
	(H)	100	96	100	95	100	100	100	100	96	100	100	89	100	100	96	100	100	100	96	97	95	100	98	100	99	97	100	99	98.3±2.6	
Lens-Rt (193.4 mm^2)	(M)	100	0	95	100	100	100	93	0	100	96	100	72	90	92	100	94	100	96	100	96	100	92	96	100	100	100	94	99	89.4±25.4	-7.2
	(H)	100	74	97	100	100	96	96	81	99	100	100	85	100	97	100	100	100	96	100	98	100	93	96	100	100	99	100	100	96.7±6.2	
Lung-Lt (56292.2 mm^2)	(M)	(100)	(99)	(100)	(99)	(98)	(99)	(92)	(97)	(99)	(99)	(99)	(97)	(94)	(97)	(99)	(100)	(99)	(99)	(99)	(98)	(98)	(99)	(99)	(99)	(99)	(99)	(100)	(99)	(99)	98.4±1.6	0.2
	(H)	(99)	(99)	(100)	(97)	(97)	(96)	(97)	(95)	(99)	(98)	(100)	(98)	(98)	(100)	(100)	(100)	(100)	(99)	(100)	(97)	(98)	(99)	(99)	(99)	(99)	(99)	(96)	(100)	(91)	98.2±1.9	
Lung-Rt (58043.6 mm^2)	(M)	(99)	(99)	(92)	(99)	(98)	(99)	(90)	(95)	(99)	(99)	(99)	(97)	(93)	(99)	(98)	(100)	(99)	(99)	(99)	(95)	(99)	(99)	(99)	(98)	(99)	(99)	(100)	(97)	97.8±2.4	-0.5
	(H)	(99)	(99)	(100)	(97)	(98)	(99)	(97)	(95)	(98)	(98)	(100)	(98)	(97)	(100)	(99)	(100)	(100)	(99)	(99)	(95)	(99)	(99)	(99)	(98)	(98)	(97)	(100)	(96)	98.3±1.4	
Mandible (20867.9 mm^2)	(M)	97	95	95	94	98	98	93	99	96	75	100	91	99	99	98	93	100	98	82	95	98	99	98	99	91	96	95	99	95.4±5.3	-2.6
	(H)	99	100	96	97	100	94	98	99	96	99	100	99	100	98	98	97	100	99	99	94	92	98	97	98	99	99	99	98	98.0±2.0	
Optic-Nerve-Lt (717.6 mm^2)	(M)	92	100	98	99	98	95	98	91	95	99	100	99	93	98	95	100	93	96	94	100	97	99	97	98	100	99	99	95	97.0±2.6	0.9
	(H)	89	99	100	99	91	99	99	95	86	90	98	99	96	99	96	98	94	96	91	100	100	100	100	92	95	96	98	95	96.1±3.8	
Optic-Nerve-Rt (719.9 mm^2)	(M)	89	99	97	97	99	92	99	96	98	99	95	100	89	99	93	88	96	100	99	100	96	100	97	95	100	99	99	92	96.4±3.5	-0.7
	(H)	88	99	95	97	95	100	98	99	95	100	99	99	95	100	97	97	99	99	97	100	99	100	95	93	96	95	98	96	97.2±2.7	
Orbit-Lt (2320.5 mm^2)	(M)	93	99	93	99	95	86	96	94	98	99	100	96	99	99	99	97	99	94	94	86	97	92	90	89	95	91	92	95	94.9±3.8	-1.0
	(H)	97	98	95	92	94	93	95	96	98	95	98	99	98	95	98	100	100	97	98	90	100	94	89	89	99	96	94	96	95.9±3.1	
Orbit-Rt (2360.3 mm^2)	(M)	96	93	94	97	97	93	92	91	98	95	100	96	96	98	100	99	99	93	93	91	99	91	95	96	96	88	98	94	95.3±3.0	-0.5
	(H)	100	94	92	95	98	95	98	97	95	97	96	93	100	100	98	99	100	95	95	91	96	93	97	89	99	91	96	91	95.7±3.0	
Parotid-Lt (7991.9 mm^2)	(M)	91	82	91	96	92	92	77	95	75	95	97	95	95	95	95	93	99	97	87	89	95	94	95	94	85	67	97	97	91.1±7.5	-3.3
	(H)	94	82	93	94	96	98	88	91	99	96	98	93	97	95	97	97	97	97	99	89	90	93	97	94	97	90	97	97	94.4±3.9	
Parotid-Rt (8322.3 mm^2)	(M)	96	91	89	93	97	84	93	90	68	94	97	93	91	90	97	95	98	96	90	95	93	91	92	95	69	95	90	93	91.2±7.0	-3.5
	(H)	96	90	91	89	98	95	93	98	92	95	99	99	93	96	98	98	97	97	94	95	95	94	97	92	92	97	90	95	94.8±2.8	
Spinal-Canal (18036.4 mm^2)	(M)	(93)	(92)	(93)	(91)	(92)	(89)	(87)	(98)	(94)	(87)	(93)	(91)	(91)	(98)	(92)	(86)	(92)	(91)	(88)	(94)	(88)	(96)	(87)	(93)	(93)	(91)	(93)	(96)	91.8±3.1	-2.9
	(H)	(95)	(96)	(96)	(91)	(94)	(92)	(96)	(95)	(96)	(97)	(99)	(94)	(95)	(96)	(92)	(91)	(94)	(95)	(97)	(98)	(93)	(97)	(95)	(93)	(93)	(94)	(96)	(92)	94.7±2.0	
Spinal-Cord (8623.7 mm^2)	(M)	99	99	99	99	100	99	100	100	94	98	100	98	99	100	100	100	100	98	99	100	99	100	99	100	100	100	98	100	99.1±1.1	-0.6
	(H)	99	100	100	100	99	100	100	100	100	100	100	99	99	100	100	100	100	99	100	100	100	100	100	100	100	100	100	100	99.8±0.3	
Submandibular-Lt (3167.6 mm^2)	(M)	(75)	(86)	(89)	(88)	(97)	(96)	(66)	(96)	(99)	(83)	(94)	(89)	(86)	(98)	(80)	(83)	(91)	(93)	(72)	(98)	(89)	(-)	83	(59)	(-)	(-)	(88)	(89)	86.7±9.9	-3.7
	(H)	(91)	(98)	(98)	(94)	(98)	(97)	(77)	(98)	(98)	(96)	(100)	(90)	(96)	(99)	(96)	(99)	(96)	(97)	(94)	(100)	(69)	(-)	97	(0)	(-)	(-)	(91)	(91)	90.4±19.7	
Submandibular-Rt (3156.2 mm^2)	(M)	(67)	(83)	(65)	(82)	(93)	(95)	(86)	(95)	(97)	(95)	(95)	(91)	(96)	(97)	(96)	(90)	(92)	(92)	(93)	(83)	(93)	(33)	(36)	(73)	(-)	(89)	(-)	(67)	83.6±17.1	1.6
	(H)	(83)	(99)	(91)	(84)	(92)	(98)	(89)	(96)	(91)	(100)	(100)	(90)	(92)	(100)	(89)	(100)	(97)	(96)	(99)	(97)	(87)	(0)	(0)	(77)	(-)	(86)	(-)	(0)	82.0±30.2	
aggr. surface DSC[1]	(M)	95.2	93.0	95.0	94.1	96.0	92.7	94.9	97.0	91.3	77.2	98.5	94.7	96.8	95.5	96.7	96.5	97.9	97.5	89.2	94.0	96.9	97.1	96.3	96.5	93.4	94.5	95.4	95.2		
	(H)	97.1	95.9	95.6	95.1	95.2	96.5	95.9	97.1	95.4	96.6	98.5	96.6	97.3	97.3	97.4	96.8	97.1	98.3	96.8	96.2	95.8	97.2	96.3	96.2	97.4	96.1	96.1	96.1		
difference		-1.8	-2.9	-0.6	-1.0	0.8	-3.7	-0.9	-0.1	-4.1	-19.4	-0.0	-1.9	-0.4	-1.7	-0.6	-0.3	0.8	-0.8	-7.6	-2.2	1.1	-0.2	-0.0	0.3	-4.0	-1.6	-0.7	-0.9		

Numbers below the organ name show the average surface area of this organ in the test set.
M: our model performance
H: human (radiographer) performance
numbers in brackets indicate that this organ for this patient would not be segmented in current clinical practise
[1]: aggregated only over organs that would be segmented for this patient in current clinical practise. I.e. numbers in brackets were excluded.
Colours indicate the performance difference:

< -10%	(model is worse)
-10% to -5%	(model is slightly worse)
-5% – +5%	(model and human are on par)
+5% to +10%	(model is slightly better)
> +10%	(model is better)

Table 5 | Volumetric DSC on TCIA data set

Organ	M/H	TCIA test set patient ID: 0522c 0017	0522c 0057	0522c 0161	0522c 0226	0522c 0248	0522c 0251	0522c 0331	0522c 0416	0522c 0419	0522c 0427	0522c 0457	0522c 0479	0522c 0629	0522c 0659	0522c 0667	0522c 0669	0522c 0708	0522c 0768	0522c 0770	0522c 0773	0522c 0845	TCGA-CV-7236	TCGA-CV-7243	TCGA-CV-7245	TCGA-CV-A6JO	TCGA-CV-A6JY	TCGA-CV-A6K0	TCGA-CV-A6K1	mean, stddev	diff.
Brain (1386616.3 mm^3)	(M)	99	99	99	99	99	99	99	99	99	93	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99±1	-0.3
	(H)	99	99	99	99	99	99	99	99	99	99	100	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99.1±0.24	
Brainstem (27236.8 mm^3)	(M)	82	73	87	77	83	73	85	90	82	65	93	84	90	92	93	94	90	91	73	81	88	90	89	85	89	90	91	85±7	-4.9
	(H)	90	90	87	86	86	92	87	89	89	87	95	87	91	95	95	91	91	93	89	92	91	89	90	89	91	87	89	90.0±2.53	
Cochlea-Lt (70.9 mm^3)	(M)	75	80	84	92	81	81	83	87	77	80	91	89	89	65	58	81	74	77	85	89	85	79	70	90	75	96	66	76	80±9	5.6
	(H)	79	74	76	82	74	70	71	69	89	72	100	76	76	48	50	60	67	91	77	82	70	79	63	88	79	81	72	80	74.9±10.93	
Cochlea-Rt (66.3 mm^3)	(M)	84	90	93	88	89	83	75	87	75	73	81	81	77	68	77	81	76	89	80	81	85	84	58	86	77	87	84	79	81±7	11.4
	(H)	71	92	0	82	62	60	51	87	97	76	100	84	72	68	0	80	74	88	74	62	55	67	52	73	93	73	84	72	69.6±23.12	
Lacrimal-Lt (599.8 mm^3)	(M)	72	57	76	54	36	55	47	73	57	49	75	35	69	60	72	70	74	69	54	66	72	73	72	81	70	78	70	66	64±12	-2.9
	(H)	83	53	52	68	58	80	48	63	60	75	89	56	71	81	80	74	74	62	63	50	74	67	66	58	76	67	70	65	67.3±10.44	
Lacrimal-Rt (633.8 mm^3)	(M)	60	64	80	59	36	56	50	58	66	51	60	58	59	71	68	79	59	69	69	70	65	70	78	66	63	69	66	66	64±9	-4.0
	(H)	79	68	69	66	55	76	52	62	65	77	99	55	75	81	80	85	56	63	59	67	68	79	61	58	70	53	65	57	67.8±11.02	
Lens-Lt (200.2 mm^3)	(M)	89	85	75	89	87	94	88	75	86	86	84	0	92	75	80	91	77	80	76	86	79	93	85	87	84	87	84	91	82±17	-6.1
	(H)	90	81	96	89	94	91	90	88	84	80	100	54	93	95	92	94	82	86	86	87	80	94	87	88	90	87	89	88	87.7±8.04	
Lens-Rt (201.1 mm^3)	(M)	90	0	65	89	88	91	78	0	89	85	95	27	79	71	76	72	88	85	80	79	90	79	83	91	89	87	80	94	76±25	-8.7
	(H)	90	44	87	89	96	90	83	46	85	92	100	42	90	90	93	90	92	83	92	90	82	84	83	87	96	84	91	93	84.5±14.72	
Lung-Lt (656382.6 mm^3)	(M)	99	99	99	99	99	99	98	99	99	99	99	99	97	98	98	98	98	99	99	99	98	99	99	99	99	99	99	99	99±1	0.0
	(H)	99	99	99	98	99	98	98	98	99	99	100	99	98	99	99	99	99	99	99	98	98	99	99	99	99	98	99	97	98.7±0.66	
Lung-Rt (722139.3 mm^3)	(M)	99	99	98	99	99	99	97	99	99	99	99	99	97	98	98	98	98	99	99	99	99	99	99	99	99	99	99	99	99±1	-0.1
	(H)	99	99	99	98	99	99	99	98	99	99	100	99	98	99	99	99	99	99	99	98	99	99	99	99	99	98	99	98	98.9±0.47	
Mandible (59383.2 mm^3)	(M)	92	94	92	92	96	95	91	94	95	79	96	89	94	96	94	93	96	91	88	92	94	95	93	95	90	95	95	96	93±4	-1.3
	(H)	96	97	93	94	94	90	92	95	92	95	100	96	94	96	96	95	96	92	95	90	90	95	95	92	94	95	94	94	94.2±2.21	
Optic-Nerve-Lt (757.1 mm^3)	(M)	64	80	78	77	82	70	72	69	83	76	79	80	81	78	72	80	73	77	72	83	82	79	79	83	84	84	84	80	78±5	-1.4
	(H)	67	80	76	80	74	89	80	75	77	71	83	83	81	85	80	84	81	79	70	83	83	83	81	78	81	84	81	73	79.3±4.86	
Optic-Nerve-Rt (747.0 mm^3)	(M)	69	73	78	77	79	68	72	71	81	81	79	82	71	74	70	71	80	69	79	86	82	84	80	74	78	81	85	63	76±6	-2.0
	(H)	73	82	77	74	69	82	75	69	77	77	97	81	76	83	86	83	80	73	67	78	87	81	70	76	80	81	84	78	78.4±6.25	
Orbit-Lt (8520.3 mm^3)	(M)	90	94	92	95	93	90	94	92	95	94	94	93	95	96	95	95	93	90	92	88	93	92	91	91	93	91	92	92	93±2	-0.7
	(H)	93	94	93	91	92	93	92	93	93	94	96	95	92	94	96	96	96	93	95	87	95	92	92	90	95	95	92	94	93.3±2.05	
Orbit-Rt (8706.5 mm^3)	(M)	93	93	92	94	95	93	92	90	95	94	94	94	94	96	96	95	96	91	92	89	94	91	94	91	93	91	94	93	93±2	-0.3
	(H)	95	92	91	93	94	93	93	94	93	96	95	91	95	96	96	94	96	92	94	88	93	92	94	90	94	93	93	92	93.4±1.85	
Parotid-Lt (29887.7 mm^3)	(M)	81	78	82	88	82	85	72	88	75	88	88	86	86	86	89	84	89	87	85	82	88	84	89	88	83	64	90	90	84±6	-3.0
	(H)	83	79	83	86	86	90	84	85	90	90	96	83	88	90	89	89	90	88	91	85	86	84	90	87	89	83	90	88	87.1±3.37	
Parotid-Rt (31237.3 mm^3)	(M)	85	84	86	89	87	80	84	86	69	87	91	86	82	82	90	86	89	86	85	88	86	83	83	86	76	85	87	84	85±4	-2.7
	(H)	84	85	86	85	88	88	83	90	85	89	98	88	82	89	90	90	91	88	89	88	87	85	89	84	88	86	85	86	87.4±3.08	
Spinal-Canal (63887.1 mm^3)	(M)	92	91	94	93	92	90	91	94	92	89	92	92	91	94	90	88	90	91	90	93	89	94	91	91	94	92	92	94	92±2	-2.3
	(H)	93	95	94	92	94	93	95	90	94	95	99	94	93	95	94	93	95	94	95	94	92	96	96	91	94	92	95	92	93.9±1.77	
Spinal-Cord (19029.9 mm^3)	(M)	75	83	64	86	88	81	85	84	57	78	89	76	86	88	90	83	79	79	82	88	86	68	80	70	78	83	83	80	80±8	-4.0
	(H)	81	86	75	85	87	85	85	83	70	83	92	83	85	87	88	88	80	80	87	90	88	79	85	78	85	88	87	88	84.3±4.63	
Submandibular-Lt (9339.5 mm^3)	(M)	67	82	80	83	90	88	62	88	93	82	90	83	81	91	79	78	88	88	68	91	85	–	82	60	–	–	84	83	82±9	-2.8
	(H)	81	91	89	86	91	89	74	90	93	91	100	83	90	93	92	92	91	92	88	94	70	–	90	0	–	–	85	85	84.7±18.32	
Submandibular-Rt (9226.6 mm^3)	(M)	59	78	61	78	86	88	82	89	92	89	90	85	89	88	88	82	86	87	86	82	88	19	25	78	–	83	–	65	78±18	0.3
	(H)	78	91	86	78	86	91	86	89	89	93	100	83	87	93	87	95	92	89	91	91	82	0	0	78	–	79	–	0	77.5±28.49	

Numbers below the organ name show the average volume of this organ in the test set.
M: our model performance
H: human (radiographer) performance
Colors indicate performance differences: green: model is better, red: model is worse

Table 6 | Surface DSC on PDDCA data set

	PDDCA test set patient ID															
	off-site test set										on-site test set					
Organ	0522c 0555	0522c 0576	0522c 0598	0522c 0659	0522c 0661	0522c 0667	0522c 0669	0522c 0708	0522c 0727 c	0522c 0746	0522c 0788	0522c 0806	0522c 0845	0522c 0857	0522c 0878	mean, stddev
Brainstem (5042.8 mm^2)	84.4	85.4	87.4	89.9	98.4	79.4	98.9	99.9	98.1	72.3	98.2	89.7	91.2	95.0	71.5	89.3±9.0
Mandible (17215.2 mm^2)	96.1	98.5	97.9	97.5	96.6	98.4	96.4	99.8	98.6	98.3	97.0	93.7	97.6	96.2	94.1	97.1±1.6
Optic-Nerve-Lt (524.6 mm^2)	95.5	99.2	95.7	88.3	86.0	92.3	99.9	94.4	86.6	89.0	95.6	82.4	98.7	98.3	91.5	92.9±5.2
Optic-Nerve-Rt (480.7 mm^2)	95.0	95.6	95.5	93.2	89.6	93.2	95.2	96.0	96.3	79.8	97.2	83.7	96.7	97.0	91.3	93.0±4.9
Parotid-Lt (6710.1 mm^2)	96.4	96.6	99.1	95.7	97.5	95.5	97.4	99.2	89.8	95.1	98.6	92.1	98.1	98.6	96.6	96.4±2.5
Parotid-Rt (6630.9 mm^2)	93.2	94.0	97.7	91.3	98.2	98.1	96.7	96.0	93.8	74.5	97.1	93.2	98.4	97.4	85.5	93.7±6.1
Submandibular-Lt (2258.0 mm^2)	64.2	60.5	85.9	80.9	87.8	76.0	89.2	84.8	97.0	61.3	98.0	77.4	74.0	95.9	80.2	80.9±11.8
Submandibular-Rt (2296.7 mm^2)	81.2	73.4	93.6	85.3	85.2	92.9	86.9	85.8	99.7	68.0	98.9	85.5	79.6	78.0	80.7	85.0±8.5
aggr. surface dice	92.3	91.1	95.8	93.4	95.7	93.5	96.1	97.2	96.2	86.0	97.6	91.4	94.7	95.8	88.9	

Numbers below the organ name show the average surface area of this organ in the PDDCA test set.
Colours indicate the performance difference:

< -10%	(model is worse)
-10% to -5%	(model is slightly worse)
-5% – +5%	(model and human are on par)
+5% to +10%	(model is slightly better)
> +10%	(model is better)

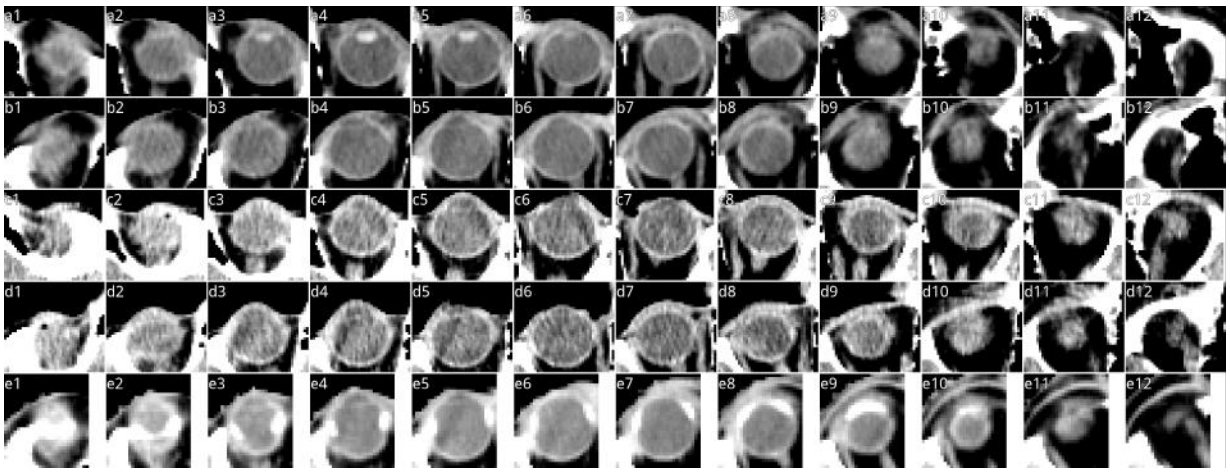

Figure 11 | **Missed lens predictions across the TCIA test set**. Consecutive axial slices of eyes showing both a typical lens and the four cases where the model predictions omitted the lens. The window level is at a constant W:140 L:0. (**a1-a12**) 12 slices through a single eye in which the model was able to detect the lens, which is clearly visible in (a3-a6). (a1) is the upper most slice, (a12) the lower most. (**b1-e12**) Similar to the first row, but these four cases are those for which the model was unable to differentiate the lens from the rest of the eye. Note that all four cases are considerably more challenging than for the first row.

Table 7 | Volumetric DSC on PDDCA data set

	PDDCA test set patient ID															
	off-site test set										on-site test set					
Organ	0522c 0555	0522c 0576	0522c 0598	0522c 0659	0522c 0661	0522c 0667	0522c 0669	0522c 0708	0522c 0727 c	0522c 0746	0522c 0788	0522c 0806	0522c 0845	0522c 0857	0522c 0878	mean, stddev
Brainstem (19778.8 mm^3)	82.0	82.7	83.0	84.9	88.8	76.4	88.5	92.8	86.7	76.3	89.9	85.3	86.5	85.5	73.7	84.2±5.2
Mandible (44477.1 mm^3)	94.2	92.1	95.8	94.9	90.4	94.9	94.6	96.3	96.0	95.4	92.8	90.8	92.5	94.7	91.6	93.8±1.9
Optic-Nerve-Lt (449.1 mm^3)	71.2	85.2	70.4	66.8	64.7	72.6	79.6	64.0	71.3	64.3	70.2	64.3	73.7	76.6	78.9	71.6±6.2
Optic-Nerve-Rt (384.3 mm^3)	69.8	68.6	75.6	63.1	61.9	69.3	62.7	73.8	78.7	63.0	69.1	62.5	65.4	80.5	72.9	69.1±5.9
Parotid-Lt (23677.4 mm^3)	87.9	89.0	91.1	85.6	90.1	89.2	88.7	88.7	84.0	87.0	90.1	84.6	88.9	88.8	87.8	88.1±2.0
Parotid-Rt (23828.3 mm^3)	87.2	88.1	90.8	82.4	89.4	90.2	87.0	86.9	87.1	76.8	87.6	86.5	88.8	88.0	82.6	86.6±3.5
Submandibular-Lt (5522.9 mm^3)	66.1	60.9	81.1	76.0	82.8	76.9	82.0	79.5	87.2	60.6	89.2	75.1	66.8	88.8	74.3	76.5±9.1
Submandibular-Rt (5660.5 mm^3)	80.6	75.8	83.8	76.6	76.7	86.8	83.1	77.2	89.6	66.7	89.8	82.4	74.9	72.5	71.6	79.2±6.5

Numbers below the organ name show the average volume of this organ in the PDDCA test set.
Colours indicate the performance difference:

< -10%	(model is worse)
-10% to -5%	(model is slightly worse)
-5% – +5%	(model and human are on par)
+5% to +10%	(model is slightly better)
> +10%	(model is better)

Table 8 | Surface DSC on UCLH data set

Organ	M/H	UCLH test set patient ID																				mean, stddev	diff.
		UCLH-01	UCLH-02	UCLH-03	UCLH-04	UCLH-05	UCLH-06	UCLH-07	UCLH-08	UCLH-09	UCLH-10	UCLH-11	UCLH-12	UCLH-13	UCLH-14	UCLH-15	UCLH-16	UCLH-17	UCLH-18	UCLH-19	UCLH-20	UCLH-21		
Brain (81738.4 mm^2)	(M)	98	97	94	95	94	96	96	94	94	96	95	95	94	96	94	97	97	95	93	97	96	95.5±1.3	-0.9
	(H)	97	99	94	96	98	96	97	96	96	96	94	97	95	95	95	99	98	97	96	97	98	96.4±1.3	
Brainstem (6555.3 mm^2)	(M)	100	99	84	97	99	98	99	99	97	99	98	98	99	98	98	97	99	99	98	97	100	97.7±3.2	0.7
	(H)	98	99	95	95	98	98	96	97	97	97	95	98	95	91	97	99	99	99	100	95	99	97.0±2.1	
Cochlea-Lt (83.8 mm^2)	(M)	100	95	96	98	100	100	94	99	99	100	100	98	91	96	98	98	97	94	96	100	100	97.6±2.5	0.8
	(H)	100	100	100	89	100	100	100	99	100	98	86	100	98	97	100	80	99	97	99	89	100	96.7±5.5	
Cochlea-Rt (84.5 mm^2)	(M)	95	100	96	100	94	89	96	100	98	100	100	100	96	99	100	96	99	98	100	100	100	97.8±2.9	0.3
	(H)	100	100	96	93	100	93	100	99	100	98	93	100	100	100	100	90	95	100	94	94	100	97.4±3.2	
Lacrimal-Lt (671.4 mm^2)	(M)	99	100	97	99	99	95	97	99	96	93	99	100	98	90	99	98	95	96	98	100	99	97.4±2.5	0.8
	(H)	93	99	99	100	93	99	96	97	97	94	100	97	93	96	98	99	100	99	84	97	97	96.5±3.6	
Lacrimal-Rt (658.2 mm^2)	(M)	100	99	87	98	97	93	98	99	91	98	99	99	100	98	100	94	96	99	99	96	95	96.9±3.3	0.9
	(H)	100	99	95	98	92	83	96	91	93	98	100	96	100	99	99	96	96	99	98	91	98	96.0±4.0	
Lens-Lt (222.7 mm^2)	(M)	100	100	100	98	100	100	0	100	94	63	96	91	100	100	97	99	100	99	90	87	100	91.1±22.0	-3.0
	(H)	95	100	100	96	100	99	59	99	97	63	95	97	100	100	97	100	100	96	85	100	100	94.1±11.3	
Lens-Rt (218.4 mm^2)	(M)	100	100	100	99	100	100	0	95	98	100	100	95	97	99	98	100	100	96	100	93	94	93.5±21.0	-1.2
	(H)	100	100	100	96	100	100	40	96	100	94	96	96	99	99	94	97	100	97	100	100	90	94.8±12.6	
Lung-Lt (44876.4 mm^2)	(M)	100	99	97	98	99	98	100	98	98	97	99	99	97	99	99	99	100	99	98	98	99	98.7±0.8	0.0
	(H)	100	99	99	99	99	99	100	97	98	98	97	100	98	99	99	99	99	98	100	98	99	98.6±0.9	
Lung-Rt (45978.6 mm^2)	(M)	99	99	95	98	99	98	98	98	95	99	99	99	97	99	99	99	100	99	99	98	98	98.3±1.3	-0.2
	(H)	100	98	99	98	98	98	99	98	98	99	96	99	98	98	99	100	99	98	99	98	98	98.5±0.9	
Mandible (21268.1 mm^2)	(M)	95	98	96	95	98	96	99	93	96	94	89	98	98	93	90	97	98	99	94	95	95	95.6±2.7	-2.4
	(H)	95	98	99	100	97	98	100	96	98	100	94	97	99	99	97	100	97	98	97	97	99	97.9±1.5	
Optic-Nerve-Lt (723.6 mm^2)	(M)	98	97	97	98	99	98	97	98	97	95	100	100	99	93	97	97	100	97	97	95	98	97.5±1.6	0.7
	(H)	97	99	83	98	99	100	100	100	96	92	91	99	100	98	96	98	92	99	100	94	100	96.8±4.1	
Optic-Nerve-Rt (722.3 mm^2)	(M)	99	100	86	99	100	96	97	99	100	100	99	99	99	83	98	89	95	98	98	98	100	96.7±4.6	-0.4
	(H)	96	100	73	99	99	98	99	100	100	95	99	99	100	96	99	95	99	98	100	97	100	97.2±5.6	
Orbit-Lt (2553.3 mm^2)	(M)	91	99	92	99	97	100	97	100	99	97	94	98	99	97	100	98	99	95	97	99	98	97.4±2.4	0.3
	(H)	94	100	97	95	98	100	96	100	98	94	97	90	99	99	98	94	95	99	98	100	96	97.1±2.5	
Orbit-Rt (2547.3 mm^2)	(M)	94	98	98	100	94	100	96	94	100	97	93	97	100	98	99	97	98	98	98	98	100	97.4±2.1	-0.2
	(H)	96	99	98	98	95	100	92	100	99	97	95	96	100	97	99	98	95	98	100	99	99	97.6±2.0	
Parotid-Lt (7779.0 mm^2)	(M)	93	90	97	98	92	84	89	95	98	96	95	97	91	96	86	95	95	82	89	97	87	92.4±4.7	-2.2
	(H)	95	91	98	90	96	84	100	97	97	95	93	97	98	96	93	95	99	98	92	91	92	94.6±3.7	
Parotid-Rt (7714.8 mm^2)	(M)	88	93	78	98	93	90	90	95	96	93	82	97	90	98	88	93	93	85	84	93	96	91.1±5.2	-3.3
	(H)	95	96	99	97	97	89	96	96	98	94	87	89	97	100	84	98	95	93	97	96	91	94.4±4.1	
Spinal-Canal (16014.9 mm^2)	(M)	93	96	89	95	93	97	95	94	92	96	95	88	93	95	98	94	93	93	93	93	98	93.8±2.4	0.4
	(H)	99	97	91	89	98	97	96	91	96	93	96	88	89	96	93	92	94	93	89	88	97	93.4±3.5	
Spinal-Cord (7660.0 mm^2)	(M)	99	100	100	100	100	99	100	100	97	100	99	100	99	100	98	100	98	99	100	99	100	99.4±0.8	-0.4
	(H)	100	100	100	98	100	100	100	100	100	100	100	100	100	100	99	100	99	100	100	100	100	99.7±0.5	
Submandibular-Lt (3478.8 mm^2)	(M)	67	74	79	83	-	98	-	87	86	83	90	93	87	89	90	89	95	94	97	94	99	88.1±8.0	-4.5
	(H)	97	96	81	90	-	93	-	93	90	94	96	91	88	89	92	96	91	97	98	91	99	92.6±4.2	
Submandibular-Rt (3279.1 mm^2)	(M)	77	97	84	92	-	95	-	93	96	-	-	95	90	96	85	89	100	98	94	94	99	92.4±5.8	-0.3
	(H)	95	93	86	86	-	94	-	95	94	-	-	96	93	97	78	92	96	93	93	97	99	92.7±5.0	
aggr. surface DSC[1]	(M)	96.6	97.4	93.2	95.8	96.5	96.3	96.9	95.3	95.2	96.8	96.4	96.8	95.3	97.4	96.2	96.8	97.9	96.8	95.5	97.1	97.2		
	(H)	97.7	98.1	95.0	95.6	97.9	96.8	98.2	96.2	97.1	97.3	95.2	96.9	96.1	97.1	96.5	98.0	97.8	97.3	97.0	96.8	98.2		
difference		-1.1	-0.7	-1.8	0.2	-1.5	-0.6	-1.3	-0.9	-1.8	-0.5	1.1	-0.1	-0.8	0.3	-0.3	-1.2	0.0	-0.5	-1.5	0.3	-1.0		

Numbers below the organ name show the average surface area of this organ in the UCLH test set.
M: our model performance
H: human (radiographer) performance
Colours indicate the performance difference:

< -10%	(model is worse)
-10% to -5%	(model is slightly worse)
-5% – +5%	(model and human are on par)
+5% to +10%	(model is slightly better)
> +10%	(model is better)

Table 9 | Volumetric DSC on UCLH data set

		UCLH test set patient ID																						
Organ	M/H	UCLH-01	UCLH-02	UCLH-03	UCLH-04	UCLH-05	UCLH-06	UCLH-07	UCLH-08	UCLH-09	UCLH-10	UCLH-11	UCLH-12	UCLH-13	UCLH-14	UCLH-15	UCLH-16	UCLH-17	UCLH-18	UCLH-19	UCLH-20	UCLH-21	mean, stddev	diff.
Brain ($1316891.7\ mm^3$)	(M)	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99± 0	-0.1
	(H)	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99.2±0.17	
Brainstem ($26422.5\ mm^3$)	(M)	93	93	83	91	92	90	93	90	89	89	91	89	91	90	89	91	92	92	90	91	94	91± 2	0.5
	(H)	90	93	86	88	92	92	91	90	92	87	88	91	87	85	88	92	92	93	91	90	93	90.1±2.40	
Cochlea-Lt ($62.4\ mm^3$)	(M)	94	82	71	72	98	84	73	80	67	81	81	72	68	81	88	80	82	82	83	95	81	81± 8	2.9
	(H)	84	92	65	52	94	87	90	69	69	82	59	89	68	81	89	44	83	87	82	73	98	77.9±13.97	
Cochlea-Rt ($61.3\ mm^3$)	(M)	74	86	72	87	75	73	77	80	73	85	82	86	68	85	81	72	81	79	88	81	82	79± 6	-1.0
	(H)	84	88	73	68	88	80	80	71	79	88	68	95	77	91	84	52	83	96	83	73	88	80.3±10.14	
Lacrimal-Lt ($785.6\ mm^3$)	(M)	77	82	66	76	73	74	79	75	68	68	81	67	68	60	72	69	71	73	75	80	79	73± 6	-1.1
	(H)	73	86	71	81	78	74	78	72	71	67	85	65	69	71	72	82	85	76	57	77	68	74.1±7.03	
Lacrimal-Rt ($768.1\ mm^3$)	(M)	82	79	59	76	69	62	79	75	63	76	77	77	75	76	75	70	69	70	74	70	67	72± 6	0.6
	(H)	82	80	74	78	69	52	72	64	57	80	81	74	75	78	77	75	71	70	67	62	71	71.8±7.77	
Lens-Lt ($244.1\ mm^3$)	(M)	81	86	87	82	89	89	0	88	79	56	84	76	88	83	89	76	87	85	79	68	81	78±19	-5.0
	(H)	83	87	83	87	93	84	27	83	85	58	86	84	91	90	85	88	86	93	84	91	89	82.7±14.16	
Lens-Rt ($237.6\ mm^3$)	(M)	93	87	90	78	89	88	0	81	83	85	90	83	89	84	88	84	86	85	86	75	74	81±19	-3.3
	(H)	90	90	90	86	88	88	18	85	91	85	83	85	89	85	81	93	85	91	91	87	81	83.9±14.98	
Lung-Lt ($510340.2\ mm^3$)	(M)	99	98	93	97	99	99	99	98	98	99	99	98	98	99	99	99	99	99	99	99	99	98± 1	-0.1
	(H)	99	99	95	97	99	99	99	99	99	99	98	99	99	99	99	99	98	98	99	99	99	98.6±0.89	
Lung-Rt ($561923.9\ mm^3$)	(M)	99	99	89	96	99	99	99	98	98	99	100	98	98	99	99	98	99	99	99	99	98	98± 2	-0.4
	(H)	99	99	94	95	99	99	99	99	99	99	98	99	99	99	99	99	99	99	99	99	99	98.6±1.22	
Mandible ($67811.7\ mm^3$)	(M)	94	95	92	89	94	94	95	92	94	91	89	94	95	92	92	94	96	95	93	95	91	93± 2	-2.7
	(H)	95	96	96	95	96	96	98	94	93	97	94	96	97	97	97	97	95	97	96	96	96	95.8±1.23	
Optic-Nerve-Lt ($781.3\ mm^3$)	(M)	68	77	75	81	80	80	80	78	75	72	78	80	84	75	62	76	80	78	81	79	81	77± 5	-3.1
	(H)	81	80	67	77	83	83	84	83	78	67	77	87	86	79	82	76	80	84	82	80	86	80.3±5.22	
Optic-Nerve-Rt ($792.4\ mm^3$)	(M)	78	83	63	83	77	71	79	78	82	85	73	79	77	70	57	70	78	70	70	69	81	75± 7	-4.4
	(H)	76	80	52	83	82	81	82	84	84	70	78	80	87	76	78	75	81	84	89	82	84	79.4±7.40	
Orbit-Lt ($9813.0\ mm^3$)	(M)	93	96	92	96	95	95	95	95	96	92	95	94	95	96	95	95	96	94	92	95	93	95± 1	0.6
	(H)	92	96	94	94	94	94	94	96	95	92	95	91	94	95	94	92	94	95	91	96	94	93.9±1.41	
Orbit-Rt ($9906.5\ mm^3$)	(M)	94	96	94	96	95	95	95	93	96	93	93	95	95	95	94	94	95	95	93	96	95	95± 1	0.5
	(H)	94	95	94	95	93	95	93	96	95	93	94	94	94	94	93	94	93	94	93	96	95	94.2±0.90	
Parotid-Lt ($27542.6\ mm^3$)	(M)	83	82	89	89	84	80	83	87	91	88	88	89	86	85	82	88	88	78	81	88	77	85± 4	-3.1
	(H)	87	85	91	87	88	83	92	91	92	89	87	88	91	88	87	89	92	90	85	84	85	88.1±2.75	
Parotid-Rt ($27663.6\ mm^3$)	(M)	82	86	75	89	85	85	83	89	91	88	81	89	83	89	78	87	86	77	77	84	83	84± 5	-3.4
	(H)	88	88	89	89	89	85	88	90	93	90	85	82	90	91	78	90	89	86	89	86	84	87.5±3.35	
Spinal-Canal ($56388.6\ mm^3$)	(M)	91	94	93	94	92	95	94	91	90	94	94	91	93	94	95	91	93	92	92	93	95	93± 1	-0.3
	(H)	96	95	92	92	96	94	95	94	95	92	93	89	90	94	93	93	93	94	89	92	95	93.1±1.98	
Spinal-Cord ($15607.7\ mm^3$)	(M)	84	82	88	89	76	70	88	85	68	78	74	86	70	70	68	82	58	84	86	64	87	78± 9	-3.6
	(H)	90	90	81	78	85	82	84	86	78	84	84	88	83	84	82	85	66	68	83	74	78	81.6±6.00	
Submandibular-Lt ($10197.2\ mm^3$)	(M)	60	68	68	80	–	90	–	82	84	82	83	88	82	84	82	86	89	90	91	88	92	83± 8	-4.8
	(H)	89	88	74	85	–	87	–	88	88	90	90	85	84	86	85	92	88	91	92	88	92	87.5±3.97	
Submandibular-Rt ($9295.9\ mm^3$)	(M)	75	90	77	84	–	88	–	87	91	–	–	86	84	89	79	85	95	90	89	87	87	86± 5	-0.8
	(H)	86	87	78	83	–	89	–	89	89	–	–	85	88	90	76	88	91	87	89	90	90	86.8±4.02	

Numbers below the organ name show the average volume of this organ in the UCLH test set.
M: our model performance
H: human (radiographer) performance

Colors indicate performance differences: green: model is better, red: model is worse

Table 10 | Number of labelled scans in UCLH test set

		Brain	Brainstem	Cochlea		Lacrimal		Lens		Lung		Mandible	Optic Nerve		Orbit		Parotid		Spinal-Canal	Spinal-Cord	Submandibular	
				lt	rt	lt	rt	lt	rt	lt	rt		lt	rt	lt	rt	lt	rt			lt	rt
Number of scans		75	45	8	8	75	73	75	73	71	72	74	17	15	19	16	33	32	23	24	64	65
Dense segmentation				✓	✓	✓	✓	✓	✓				✓	✓								
Number of labelled slices	axial	309	225							265	275	300			95	75	165	160	345	350	250	260
	coronal	374	225							355	360	375			95	80	165	160	0	0	320	325
	sagittal	374	225							355	360	375			95	80	165	160	0	0	320	325

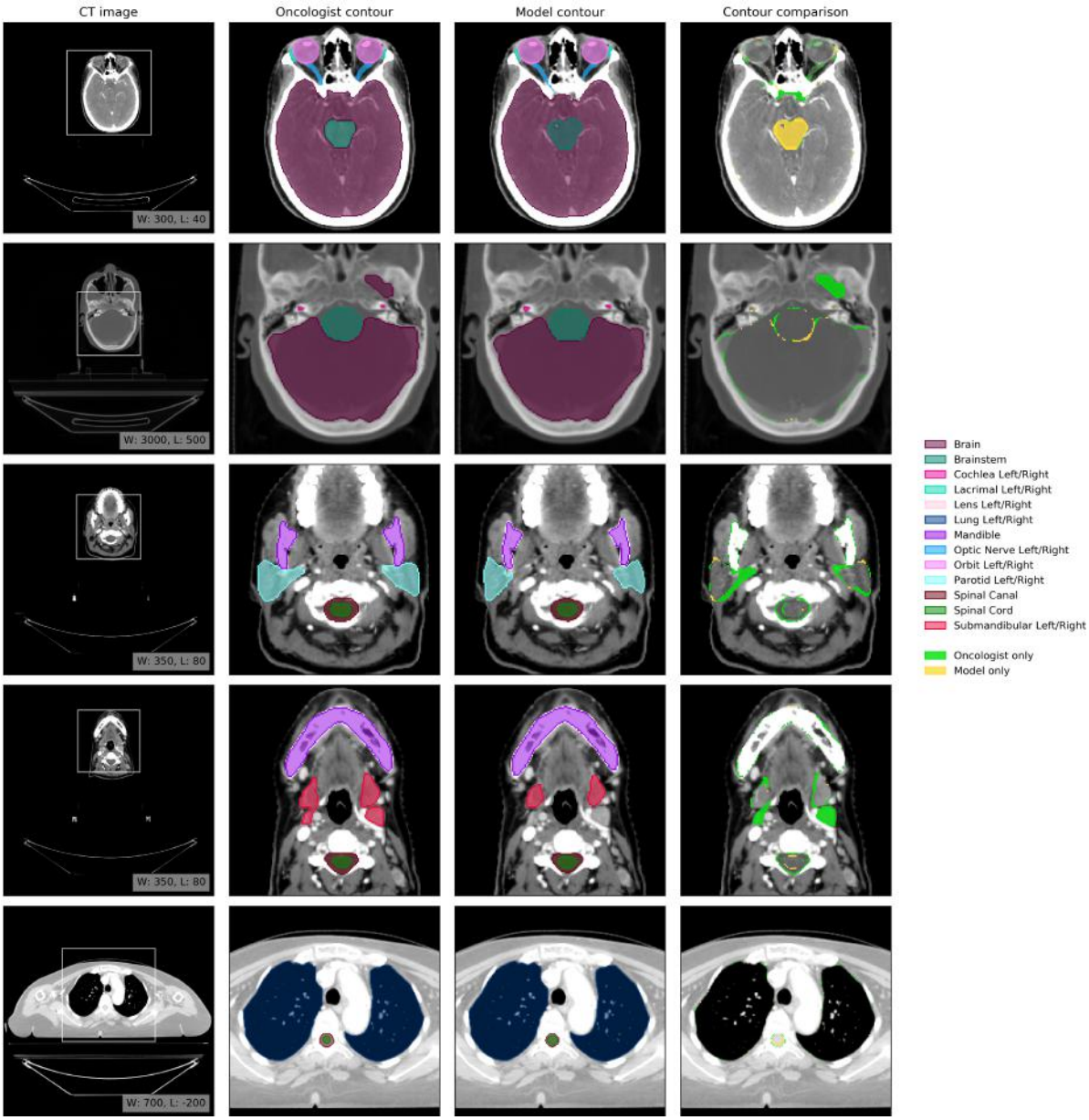

Figure 12 | Example results. Axial slices at five representative levels from the raw CT scan of 70-74 year old female patient from the UCLH test set. The levels shown as 2D slices have been selected to demonstrate all 21 OARs included in this study. The window levelling has been adjusted for each to best display the anatomy present. (**Oncologist contour**) The ground truth segmentation, as defined by experienced radiographers and arbitrated by a head and neck specialist oncologist. (**Model contour**) Segmentations produced by our model. (**Contour comparison**) Contoured by Oncologist only (green region) or Model only (yellow region). Two further randomly selected UCLH set scans are shown in Fig. 12 and Fig. 13. Best viewed on a display.

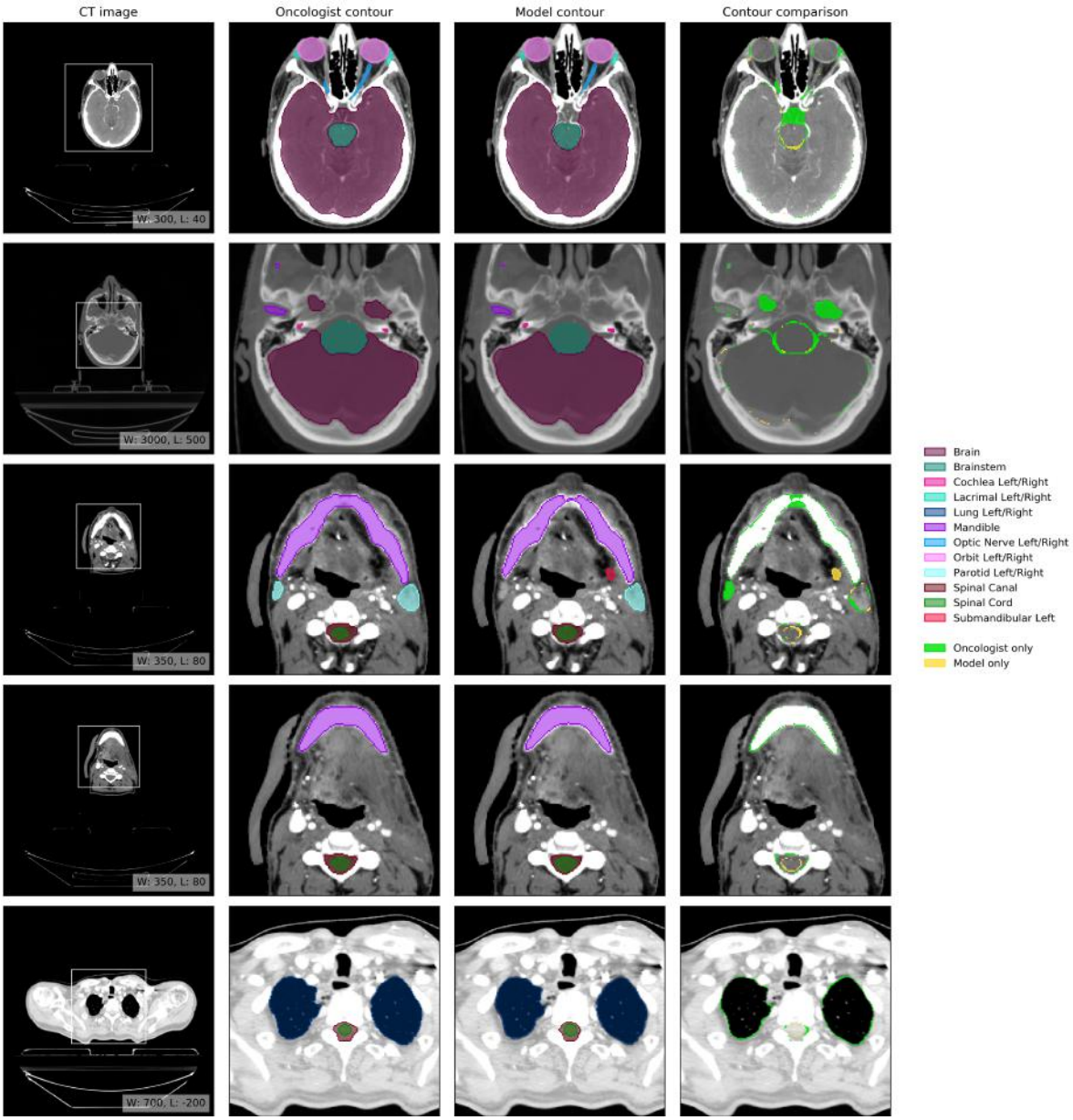

Figure 13 | Example results. Axial slices at five representative levels from the raw CT scan of 70-74 year old male patient from the UCLH test set. The levels shown as 2D slices have been selected to demonstrate all 21 OARs included in this study. The window levelling has been adjusted for each to best display the anatomy present. (**Oncologist contour**) The ground truth segmentation, as defined by experienced radiographers and arbitrated by a head and neck specialist oncologist. (**Model contour**) Segmentations produced by our model. (**Contour comparison**) Contoured by Oncologist only (green region) or Model only (yellow region). Two further randomly selected UCLH set scans are shown in Fig. 12 and Fig. 13. Best viewed on a display.

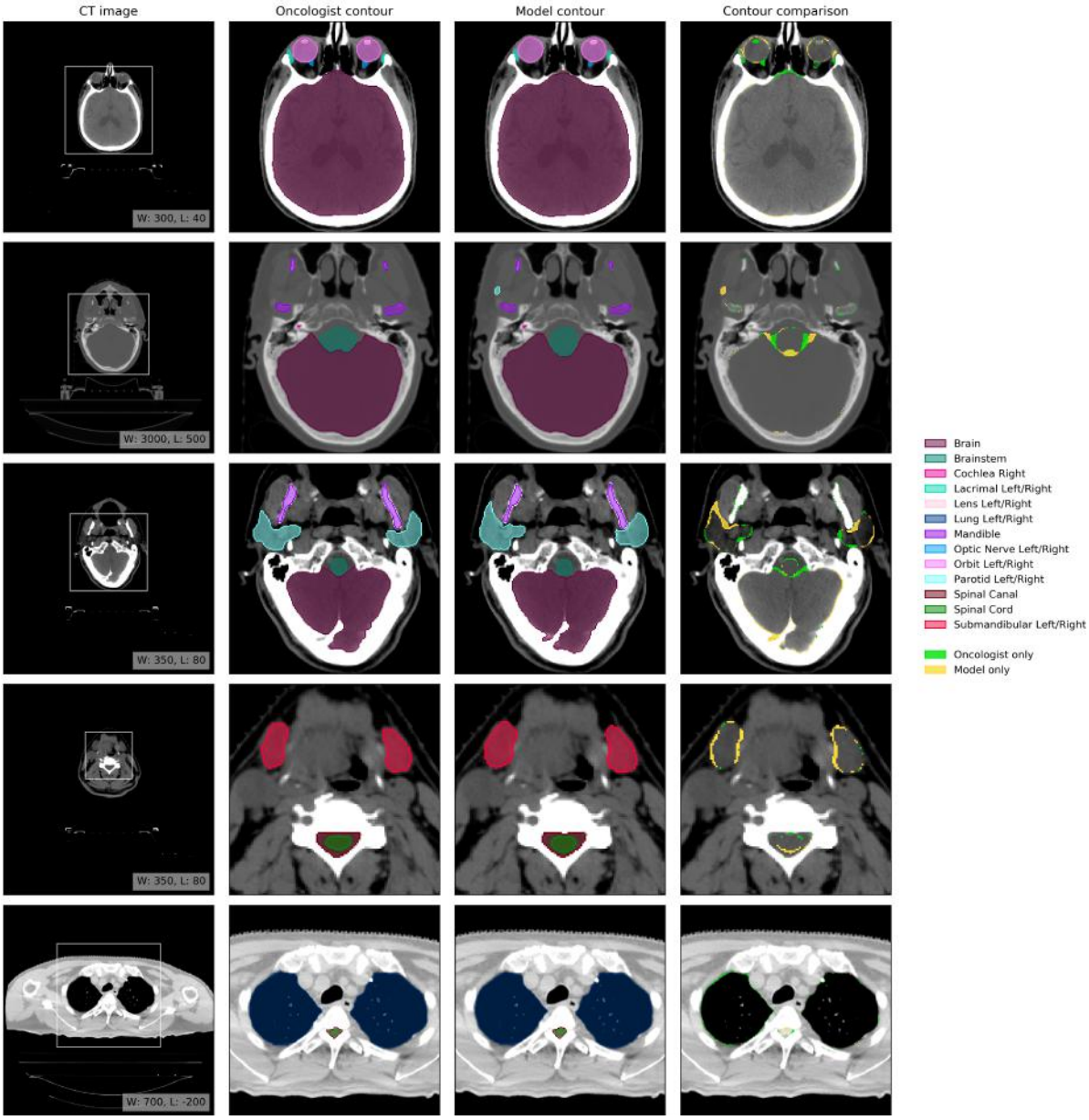

Figure 14 | Example results. (**a1-e1**) Axial slices at five representative levels from the raw CT scan of a 66 year old male patient with a right base of tongue cancer and bilateral lymph node involvement selected from the Head-Neck Cetuximab TCIA dataset (patient 0522c0057; [53]) were selected to best demonstrate the OARs included in the work. The levels shown as 2D slices have been selected to demonstrate all 21 OARs included in this study. The window levelling has been adjusted for each to best display the anatomy present. (**a2-e2**) The ground truth segmentation, as defined by experienced radiographers and arbitrated by a head and neck specialist oncologist. (**a3-e3**) Segmentations produced by our model. (**a4-e4**) Overlap between the model (yellow line) and the ground truth (blue line). Two further randomly selected TCIA set scans are shown in Fig. 15 and Fig. 16. Best viewed on a display.

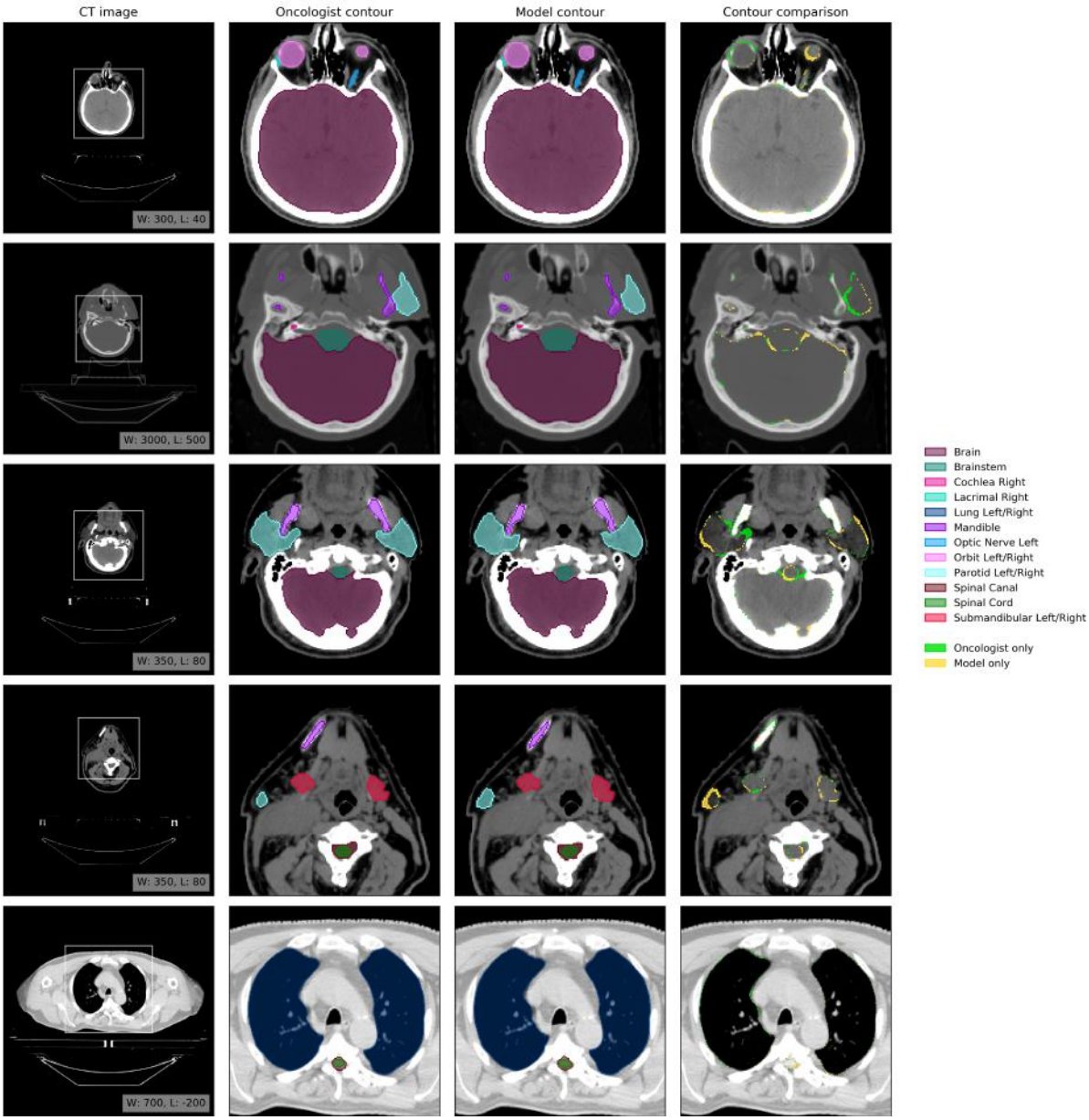

Figure 15 | Example results from a second randomly selected case from the TCIA test set. Five axial slices from the scan of a 58 year old male patient with a cancer of the right tonsil selected from the Head-Neck Cetuximab TCIA dataset (patient 0522c0416; [53]). (a1-e1) The raw CT scan slices at five representative levels were selected to best demonstrate the OARs included in the work. The window levelling has been adjusted for each to best display the anatomy present. (a2-e2) The ground truth segmentation was defined by experienced radiographers and arbitrated by a head and neck specialist oncologist. (a3-e3) The model produced segmentations of the same structures. Overlap between the model (yellow line) and the ground truth (blue line) is shown in (a4-e4). Best viewed on a display.

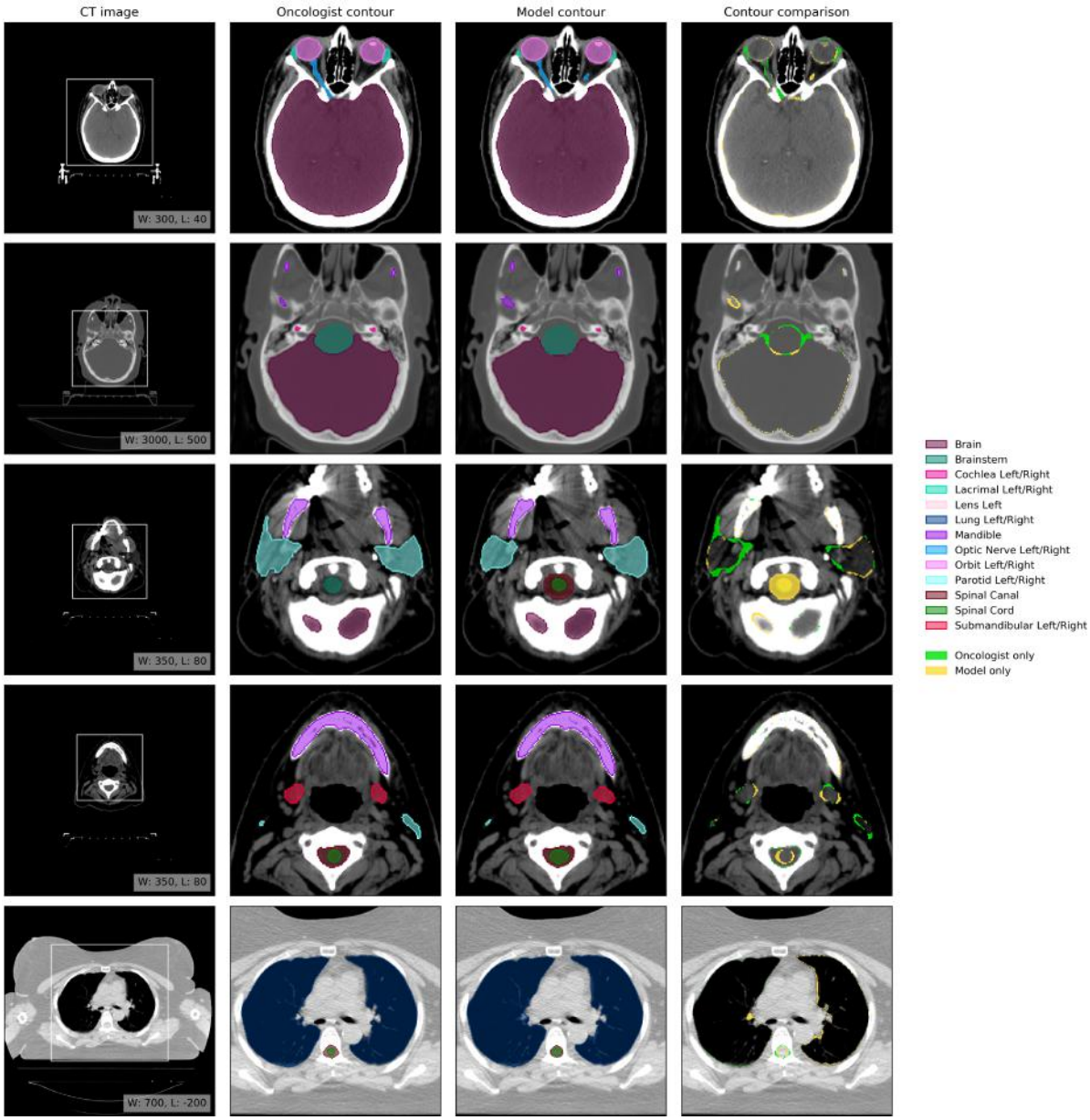

Figure 16 | Example results from a third randomly selected case from the TCIA test set. Five axial slices from the scan of a 53 year old female patient with a left oropharyngeal cancer with base of tongue invasion included selected from the Head-Neck Cetuximab TCIA dataset (patient 0522c0251; [53]). (a1-e1) The raw CT scan slices at five representative levels were selected to best demonstrate the OARs included in the work. The window levelling has been adjusted for each to best display the anatomy present. (a2-e2) The ground truth segmentation was defined by experienced radiographers and arbitrated by a head and neck specialist oncologist. (a3-e3) The model produced segmentations of the same structures. Overlap between the model (yellow line) and the ground truth (blue line) is shown in (a4-e4). Best viewed on a display.

Table 11 | Volumetric DSC performance of our model and previously published results. An overview of previously published automatic segmentation works that reported volumetric DSC for the OARs included in this study on planning head and neck CT scans. The datasets and ground truths used varied between studies making comparison difficult. Despite this, we show results alongside our evaluation of our model, radiographers and oncologists against our ground truth across multiple datasets. The latter assesses inter-observer variation between oncologists.

Study	Method	Brain	Brainstem	Cochlea		Lacrimal		Lens		Lung		Mandible	Optic Nerve		Orbit		Parotid		Spinal Canal	Spinal Cord	Submandibular	
				lt	rt	lt	rt	lt	rt	lt	rt		lt	rt	lt	rt	lt	rt			lt	rt
van Dijk (2020) [61]	CNN		83[2]									95[2]					84[2]	83[2]		87[2]	77[2]	78[2]
Zhong (2019) [36]	CNN												89				92					
Močnik (2018) [34]	CNN																77					
Ren (2018) [35]	CNN												72	70								
Ibragimov (2017) [33]	CNN											90	64	65	88	88	77	78		87	70	73
Fritscher (2016) [32]	CNN																81				65	
Guo (2020) [51]	FCN		88									94	72	71			87	86			78	81
Qiu (2020) [59]	FCN											95										
Sun (2020) [60]	FCN		86									94			90	90	84	81		89	78	77
Wong (2020) [62]	FCN		83										47				80			79	82	
Liang (2020) [58]	FCN		92					88	87			94	74		93	93	88			90	81	
Xue (2019) [69]	FCN		90									96	86	84			89	89			86	85
Chan (2019) [63]	FCN		89									91					85	86		87	84	85
Gao (2019) [64]	FCN		86					81	79				64	62	88	91	77	80		87		
Lei (2019) [66]	FCN		87										66				86					
Sun (2019) [67]	FCN							85	84				80	82	94	94						
Jiang (2019) [65]	FCN		88									93					85	86			79	77
van Rooij (2019) [50]	FCN		64														83	83			82	81
Tang (2019) [49]	FCN		86					82	83			93	75	76	92	92	85	85		86	81	83
Rhee (2019) [48]	FCN	98	86	65	68			73	70			87	89	90	89	90	83	83		83		
Tappeiner (2019) [47]	FCN		82									91	64	63			80	81				
Men (2019) [46]	FCN		90									92					86	86				
Wang (2019) [45]	FCN		88									93	74	74			86	85			76	73
Nikolov (2018) [70]	FCN	99	88	65	75	69	70	81	80	99	99	96	76	77	95	95	85	85	95	88	85	85
Kodym (2018) [44]	FCN		92									95	80				90				88	
Tong (2018) [41]	FCN		87									94	65	69			84	83			76	81
Zhu (2018) [40]	FCN		87									93	72	71			88	87			81	81
Willems (2018) [43]	FCN		92	75	73							96					86	90			79	88
Hänsch (2018) [39]	FCN																86					
Liang (2018) [42]	FCN		90					83	84			91	66	72			85	85				
Tong (2019) [68]	GAN		87									94	66	70			85	86			81	82
Gacha (2018) [29]	HAS											80										
Raudashl (2017) [30]	HAS		88									93	62				84				78	
Fritcher (2014) [82]	HAS		87[2,3]														84[2,3]	83[2,3]				
Walker (2014) [28]	HAS		97	56								98	71				89			90	73	
Thomson (2014) [27]	HAS			30[2,3]													79[3]				80[3]	
Fortunati (2013) [20]	HAS		78					67					62		81					85		
Qazi (2011) [24]	HAS		91									93										
Wu (2019) [31]	Machine learning											89					74	75			73	73
Tam (2018) [83]	Machine learning		91	67	72			75	74			85			94	94	83	82		83	87	87
Wang (2017) [84]	Machine learning		90									94					82	83				
Torosdagli (2017) [85]	Machine learning											94										
Wang (2019) [86]	Multi-ABAS		84									75	56	53			75	74			74	72
Ayyalusamy (2019) [87]	Multi-ABAS		83[2]									85[2]					81[2]			84[2]		
Haq (2019) [88]	Multi-ABAS		76									85					76	76		84	60	60
McCarroll (2018) [89]	Multi-ABAS	98	81	47						48		84			85		78			71		
Liu (2016) [90]	Multi-ABAS		92									90					87	85		80	83	80
Hoang Duc (2015) [91]	Multi-ABAS		82[2,3]												68[2,3]		70[2,3]	71[2,3]		79[2,3]		
Tao (2015) [92]	Multi-ABAS		86	43	42														77			
Wachinger (2015) [93]	Multi-ABAS																78[2,3]	77[2,3]				
Zhu (2013) [94]	Multi-ABAS	95[2]	72[2]									90[2]					72[2]			72[2]	70[2]	
Teguh (2011) [26]	Multi-ABAS		78[1]														79			78[1]	70	
Han (2010) [95]	Multi-ABAS																85	86				
Sims (2009) [25]	Multi-ABAS		77									82					84	86				
Sims (2009) [96]	Multi-ABAS		58									78					69	66				
Han (2008) [97]	Multi-ABAS		84[2,3]									91[2,3]					83[2,3]			75[2,3]	70[2,3]	
Hoogeman (2008) [22]	Multi-ABAS		71[1]																	71[1]		
Huang (2019) [98]	Single-ABAS											84	73	71			84	84		82	75	78
Daisne (2013) [19]	Single-ABAS		75[2]														72[2]					
Hardcastle (2012) [99]	Single-ABAS		86[2]														80[2]	80[2]		83[2]		
La Macchia (2012) [100]	Single-ABAS		81	69	63							86					78	79	81			
Zhang (2007) [101]	Single-ABAS		80[2]									85[2]					81[2]	80[2]		83[2]		
Radiographer (TCIA) (28 scans)	Manual	99.1 ±0.2	90.0 ±2.5	74.9 ±10.9	69.6 ±23.1	67.3 ±10.4	67.8 ±11.0	87.7 ±8.0	84.5 ±14.7	98.7 ±0.7	98.9 ±0.5	94.2 ±2.2	79.3 ±4.9	78.4 ±6.2	93.3 ±2.1	93.4 ±1.9	87.1 ±3.4	87.4 ±3.1	93.9 ±1.8	84.3 ±4.6	84.7 ±18.3	77.5 ±28.5
Our model (TCIA) (28 scans)	Deep Learning	98.8 ±1.1	85.1 ±7.1	80.5 ±8.8	81.0 ±7.2	64.4 ±11.9	63.8 ±9.0	81.6 ±16.6	75.7 ±24.5	98.7 ±0.6	98.8 ±0.7	92.9 ±3.5	77.9 ±5.0	76.3 ±5.8	92.6 ±2.0	93.1 ±1.8	84.1 ±5.8	84.6 ±4.2	91.7 ±1.6	80.3 ±7.6	81.8 ±8.7	77.8 ±18.1
Radiographer (UCLH) (21 scans)	Manual	99.2 ±0.2	90.1 ±2.4	77.9 ±14.0	80.3 ±10.1	74.1 ±7.0	71.8 ±7.8	82.7 ±22.6	83.9 ±23.8	98.6 ±0.9	98.6 ±1.3	95.8 ±1.2	80.3 ±5.2	79.4 ±7.4	93.9 ±1.4	94.2 ±0.9	88.1 ±2.8	87.5 ±3.4	93.1 ±2.0	81.6 ±6.0	87.5 ±4.0	86.8 ±4.0
Our model (UCLH) (21 scans)	Deep Learning	99 ±0.2	91 ±2.2	81 ±8.2	79 ±5.7	73 ±5.6	72 ±5.8	78 ±25.0	81 ±25.8	98 ±1.3	98 ±2.2	931 ±1.9	77 ±4.8	75 ±7.0	95 ±1.3	95 ±1.0	85 ±3.8	84 ±4.5	93 ±1.4	78 ±8.9	83 ±8.4	86 ±4.9
Oncologist (UCLH) (8 - 75 scans)[4]	Manual	99.0[5]	91.9[5]	68.5 ±14.8	75.8 ±8.5	63.3 ±13.1	61.6 ±14.3	86.2 ±10.1	87.6 ±9.9	98.4[5]	98.6[5]	95.4[5]	77.1 ±6.3	76.0 ±7.1	94.8[5]	94.8[5]	90.1[5]	90.7[5]	94.9[5]	87.7[5]	91.1[5]	90.1[5]
Our model (PDDCA) (15 scans)	Deep Learning		84.2 ±5.2									93.8 ±1.9	71.6 ±6.2	69.1 ±5.9			88.1 ±2.0	86.6 ±3.5			76.5 ±9.1	79.2 ±6.5

Values for volumetric DCS are mean (± standard deviation) unless otherwise stated. "ABAS": atlas based auto segmentation. "CNN": convolutional neural network. "FCN": fully convolutional network. "GAN": generative adversarial network. "HAS": hybrid atlas-based segmentation.

[1] merged brainstem and spinal cord. [2] Values estimated from figures; actual values not reported. [3] Median; mean not reported. [4] Number of scans per organ varies, see Table 10. [5] Volumetric DSC estimated from sparse labels.

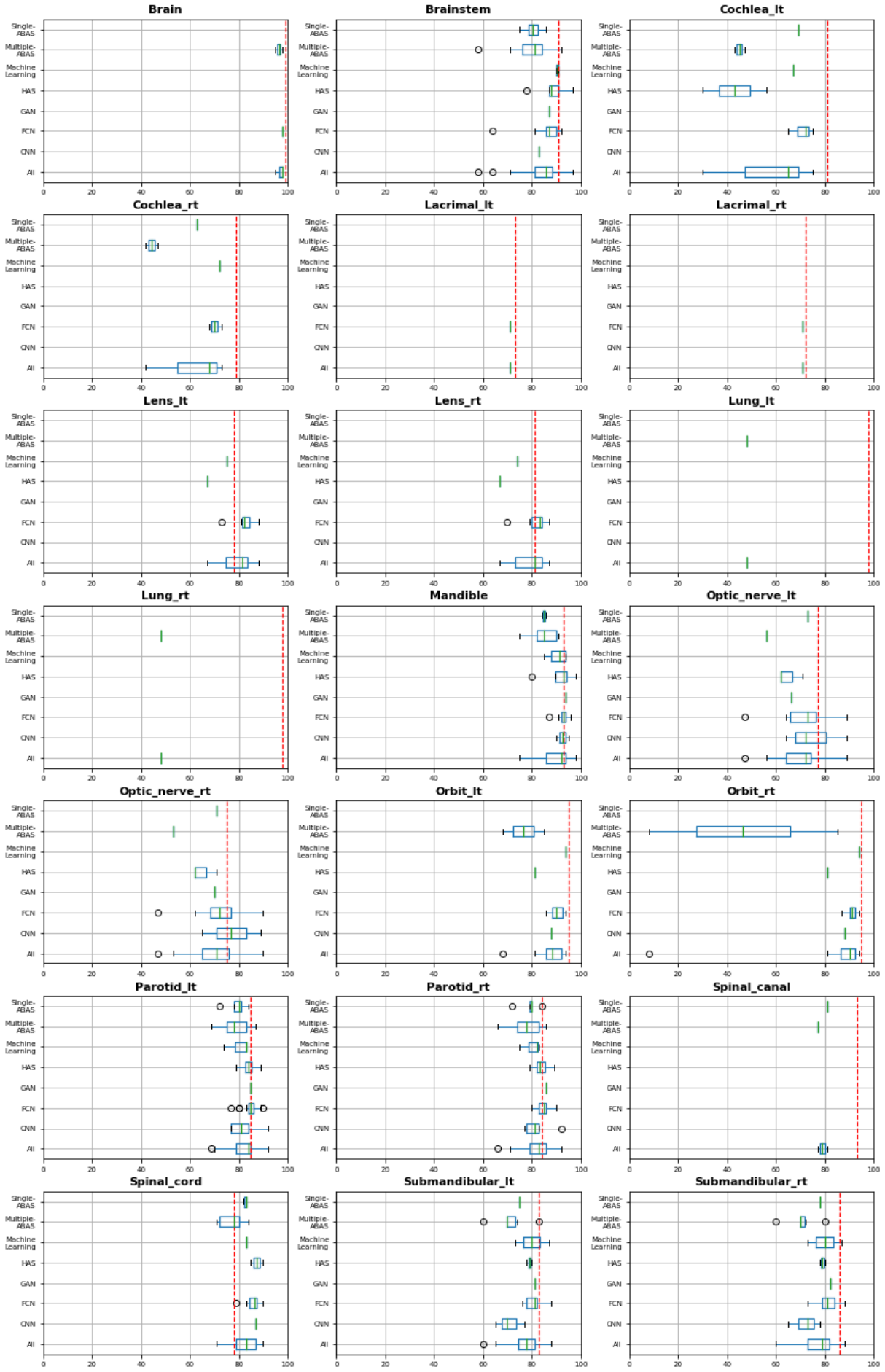

Figure 17 | Comparison of volumetric DSC performance or our model and previously published results. The volumetric-DSC performance distribution is shown for each OAR. The performance distribution is shown for each method family and for all methods collectively. The blue boxes indicate the 1st and 3rd quartiles around the median (marked in green). The whiskers indicate most extreme, non-outlier data points. The red vertical lines indicate the performance of our model on the UCLH data.