Author/s:
Clark, B;Hardcastle, N;Johnston, LA;Korte, J

Title:
Transfer learning for auto-segmentation of 17 organs-at-risk in the head and neck: Bridging the gap between institutional and public datasets

Date:
2024-07-01

Citation:
Clark, B., Hardcastle, N., Johnston, L. A. & Korte, J. (2024). Transfer learning for auto-segmentation of 17 organs-at-risk in the head and neck: Bridging the gap between institutional and public datasets. Medical Physics, 51 (7), pp.4767-4777. https://doi.org/10.1002/mp.16997.

Persistent Link:
https://hdl.handle.net/11343/345439

Received: 4 May 2023 | Revised: 4 December 2023 | Accepted: 31 January 2024

DOI: 10.1002/mp.16997

RESEARCH ARTICLE

MEDICAL PHYSICS

Transfer learning for auto-segmentation of 17 organs-at-risk in the head and neck: Bridging the gap between institutional and public datasets

Brett Clark[1,2] | **Nicholas Hardcastle**[2,3,4] | **Leigh A. Johnston**[1,5,6] | **James Korte**[1,2]

[1]Department of Biomedical Engineering, University of Melbourne, Melbourne, Australia

[2]Department of Physical Sciences, Peter MacCallum Cancer Centre, Melbourne, Australia

[3]Centre for Medical Radiation Physics, University of Wollongong, Wollongong, Australia

[4]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia

[5]Melbourne Brain Centre Imaging Unit, University of Melbourne, Melbourne, Australia

[6]Graeme Clark Institute, University of Melbourne, Melbourne, Australia

Correspondence
Brett Clark, University of Melbourne, Grattan St, Parkville, VIC 3010, Australia.
Email: baclark@student.unimelb.edu.au

Funding information
Australian Government Research Training Program (RTP) scholarship

Abstract

Background: Auto-segmentation of organs-at-risk (OARs) in the head and neck (HN) on computed tomography (CT) images is a time-consuming component of the radiation therapy pipeline that suffers from inter-observer variability. Deep learning (DL) has shown state-of-the-art results in CT auto-segmentation, with larger and more diverse datasets showing better segmentation performance. Institutional CT auto-segmentation datasets have been small historically ($n < 50$) due to the time required for manual curation of images and anatomical labels. Recently, large public CT auto-segmentation datasets ($n > 1000$ aggregated) have become available through online repositories such as The Cancer Imaging Archive. Transfer learning is a technique applied when training samples are scarce, but a large dataset from a closely related domain is available.

Purpose: The purpose of this study was to investigate whether a large public dataset could be used in place of an institutional dataset ($n > 500$), or to augment performance via transfer learning, when building HN OAR auto-segmentation models for institutional use.

Methods: Auto-segmentation models were trained on a large public dataset (public models) and a smaller institutional dataset (institutional models). The public models were fine-tuned on the institutional dataset using transfer learning (transfer models). We assessed both public model generalizability and transfer model performance by comparison with institutional models. Additionally, the effect of institutional dataset size on both transfer and institutional models was investigated. All DL models used a high-resolution, two-stage architecture based on the popular 3D U-Net. Model performance was evaluated using five geometric measures: the dice similarity coefficient (DSC), surface DSC, 95th percentile Hausdorff distance, mean surface distance (MSD), and added path length.

Results: For a small subset of OARs (left/right optic nerve, spinal cord, left submandibular), the public models performed significantly better ($p < 0.05$) than, or showed no significant difference to, the institutional models under most of the metrics examined. For the remaining OARs, the public models were inferior to the institutional models, although performance differences were small (DSC ≤ 0.03, MSD < 0.5 mm) for seven OARs (brainstem, left/right lens, left/right parotid, mandible, right submandibular). The transfer models performed significantly better than the institutional models for seven OARs (brainstem, right lens, left/right optic nerve, left/right parotid, spinal cord) with a small margin of improvement (DSC ≤ 0.02, MSD < 0.4 mm). When numbers of institutional training samples were limited, public and transfer models outperformed the

institutional models for most OARs (brainstem, left/right lens, left/right optic nerve, left/right parotid, spinal cord, and left/right submandibular).
Conclusion: Training auto-segmentation models with public data alone was suitable for a small number of OARs. Using only public data incurred a small performance deficit for most other OARs, when compared with institutional data alone, but may be preferable over time-consuming curation of a large institutional dataset. When a large institutional dataset was available, transfer learning with models pretrained on a large public dataset provided a modest performance improvement for several OARs. When numbers of institutional samples were limited, using the public dataset alone, or as a pretrained model, was beneficial for most OARs.

KEYWORDS
deep learning, image segmentation, transfer learning

1 | INTRODUCTION

An estimated 74% of head and neck (HN) cancer patients can benefit from radiation therapy (RT),[1] with the main forms of treatment being intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy. Outlining organs-at-risk (OARs), or segmentation, on computed tomography (CT) images is a crucial component of IMRT planning that allows for irradiation of tumors while sparing adjacent OARs. Sparing OARs can reduce the incidence of toxicities, such as xerostomia (dry mouth).[2]

Manual segmentation is susceptible to inter-observer variability (IOV) due to factors such as level of experience.[3] Consensus guidelines for segmentation of HN OARs were created by a panel of Asian, Australian, European, and North American radiation oncologists, to standardize manual segmentation and reduce IOV.[4] Manual segmentation, undertaken by clinicians, is time-consuming and can take up to 3 h for a single HN patient.[5] Auto-segmentation algorithms have the potential to greatly reduce segmentation time, with deep learning (DL) auto-segmentation models able to segment HN OARs in as little as 4 min.[6] DL models display state-of-the-art performance for HN auto-segmentation using CT images,[7,8] with some models achieving scores within the range of IOV,[9] and commercial solutions are now used clinically.

When training DL models for auto-segmentation, use of larger and more diverse datasets improves model performance.[10–12] However, these models may see decreased performance when applied to an independent test dataset from a different institution.[13] Transfer learning is a technique commonly applied in the absence of a large training dataset,[10] in which knowledge learned from a large source dataset (pretraining) is applied to a similar task on a small target dataset (fine-tuning).[14] Several works have demonstrated the benefit of transfer learning for CT segmentation in the presence of small target datasets, including cross-task transfer from liver to pancreas segmentation[15] and creating patient-specific models during adaptive HN treatment.[16]

In this work we investigated whether a large public dataset could be used in place of, or to augment, an institutional dataset for auto-segmentation of HN OARs. We trained separate CT auto-segmentation models per OAR on a large public dataset and a smaller institutional dataset. We investigated the generalizability of the models trained on public data alone, relative to models trained on institutional data alone, when evaluated on our institutional dataset. Fine-tuning was then applied to the public models using institutional data, to investigate if segmentation performance improved with transfer learning. Additionally, we explored the effect of the number of institutional training samples on the performance of institutional and transfer learning models.

2 | MATERIALS

A large public dataset of 1144 patients was created by aggregating the Head-Neck-Radiomics-HN1[17] (125 patients, prior to June 2014), Head-Neck-PET-CT[18] (288 patients, 2006–2014), HNSCC[19] (140 patients, 2003–2013), and OPC-Radiomics[20] (591 patients, 2005–2010) datasets from The Cancer Imaging Archive.[21] This dataset contained CT images and associated OAR segmentation labels for 17 OARs in the head and neck, with a range of 83–1053 labels per OAR. CT images from the public dataset were acquired at North American and European institutions using 15 different CT scanners with a median voxel spacing and field-of-view of $0.977 \times 0.977 \times 396$ and $500 \times 500 \times 396$ mm, respectively. The public dataset patient demographics contained an age range of 18–91 years, 20.3/79.7% female to male split, and most common primary tumor sites: the oropharynx (83.8%), larynx (7.5%), and nasopharynx (2.7%).

An institutional dataset of 571 patients (PMCC) was collected retrospectively from clinical practice at the

Peter MacCallum Cancer Centre (Melbourne, Australia; 2018–2021). This dataset contained labels for the same 17 OARs as the public dataset and was segmented and reviewed by radiation oncologists at an academic centre with a high volume of patients with HN cancer. The institutional dataset contained a range of 130–500 labels per OAR. Images were acquired predominantly using a Philips Brilliance Big Bore CT scanner with median voxel spacing and field-of-view of $1.172 \times 1.172 \times 2$ and $600 \times 600 \times 456$ mm, respectively.

CT images were excluded from the PMCC dataset for inconsistent spacing between axial slices. Incorrect OAR labels were excluded from the PMCC dataset due to missing axial slices, or incorrect lateral tagging (e.g., left parotid labelled as right parotid). No curation was applied to the public dataset as this study was concerned with the utility of public datasets without the need for time-consuming curation. Additionally, a level of public data curation was assumed as all public datasets were included in previous studies. Approval to conduct this retrospective study was given by our local ethics committee.

Details of CT images from all data sources are provided in Supplementary Figure 1, while patient demographics are given in Supplementary Figure 2. Supplementary Tables 1, 2, and 3 show: OAR label numbers per dataset, dataset geographical location, and CT scanner makes/models, respectively.

3 | METHODS

3.1 | Experiments

3.1.1 | Public model generalizability

To explore the generalizability of the public models (Figure 1, green) to the institutional dataset, we trained separate public models per OAR on the public dataset alone, using all available samples (e.g., n = 1051 for left parotid, see Supplementary Table 1 for all OAR numbers). We then evaluated the performance of the public models on each of the five test folds of the institutional dataset (20% of available samples each, e.g., n = 85 for left parotid). Separate institutional models (Figure 1, blue) were trained per OAR for comparison with the public models.

3.1.2 | Transfer versus institutional models

A comparison of transfer (Figure 1, orange) and institutional models was performed. Separate transfer models were trained per OAR using parameter-based transfer learning,[22] in which parameters of the pretrained public models were used to initialize training (fine-tuning) on the institutional dataset. Localizer stages and the first half (encoder) of each segmenter stage were frozen during fine-tuning. Institutional models were trained with randomly initialized parameters (no pretraining). Transfer and institutional models were trained and evaluated using five-fold cross validation on the institutional dataset with four folds reserved for training (80% of available samples, e.g., n = 340 for left parotid) and one for testing.

3.1.3 | Effect of institutional dataset size

To investigate the effect of the number of institutional training samples (n) on transfer and institutional model performance, we varied the size of the institutional training dataset during five-fold cross validation from five samples up to the maximum number of samples available per OAR (e.g., n = 5, 10, 20, 50, 100, 200, and 340 for left parotid).

3.2 | Auto-segmentation models

3.2.1 | Model architecture

A two-stage model was employed, consisting of sequential low-resolution localizer and high-resolution segmenter stages (Figure 2), with both stages based on the 3D U-Net architecture (Supplementary Figure 3).[23] This two-stage configuration allowed for high-resolution segmentation of CT images without exceeding the graphics processing unit (GPU) memory constraint of 12GB (NVIDIA P100 GPU). Segmenter stage patch sizes were calculated per OAR using maximum extent statistics from the public datasets (Supplementary Table 4).

3.2.2 | Data preprocessing

Spatial normalisation of CT images and binary labels was applied using the SimpleITK library (v2.1.0).[24] Interpolation schemes were: trilinear interpolation for CT images, and nearest neighbour interpolation for binary labels. Localizer stage training data was downsampled to 4 mm isotropic to allow for segmentation of the whole CT volume without exceeding the GPU memory constraint. Binary labels for smaller OARs (left/right brachial plexus, left/right cochlea, left/right lens, and left/right optic nerve) were dilated (three iterations) to improve localizer performance. Segmenter stage training data was resampled to $1 \times 1 \times 2$ mm, a high enough resolution to retain the spatial information present in PMCC dataset images (Supplementary Figure 1). No intensity normalization was applied to CT images as voxel values, in Hounsfield units (HU), are inherently normalized through their relationship to the density and atomic number of the imaged tissue.

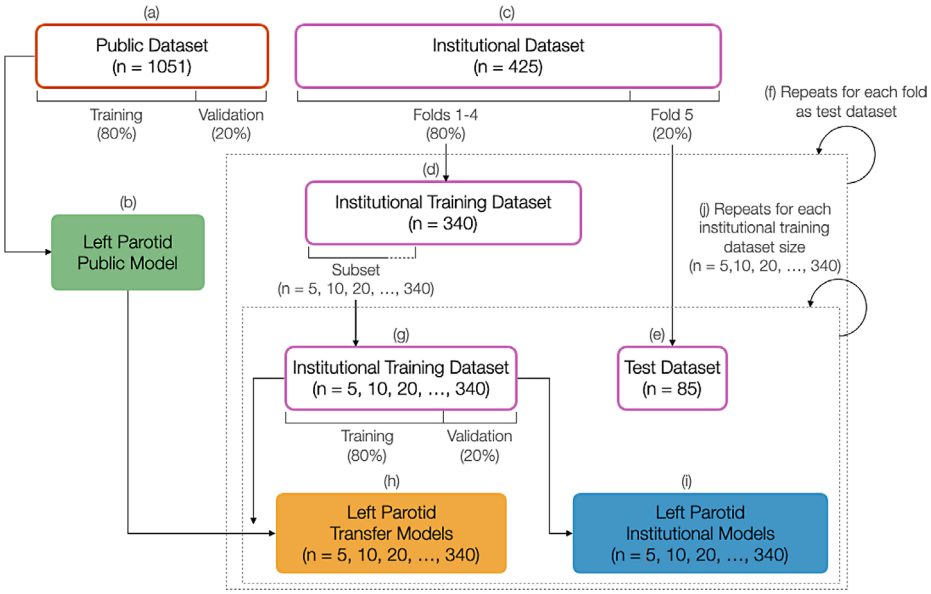

FIGURE 1 Training workflow for public (green), transfer (orange), and institutional (blue) left parotid auto-segmentation models. Separate models were trained for each of 17 OARs. The public dataset (a, n = 1051 samples) was used to train a public model (b). The institutional dataset (c) was split into five folds, with four folds reserved for training (d) and one for testing (e). Five-fold cross-validation was then performed (f) with a different fold used for testing on each iteration. Institutional training datasets (g) of increasing size were simulated by selecting a subset of the institutional training samples (e.g., n = 5, 10, 20, 50, 100, 200, and 340). Transfer (h) and institutional models (i) were trained for each dataset size (j), with transfer model parameters initialized from the public models. Public, institutional, and transfer models were evaluated on the test dataset (e).

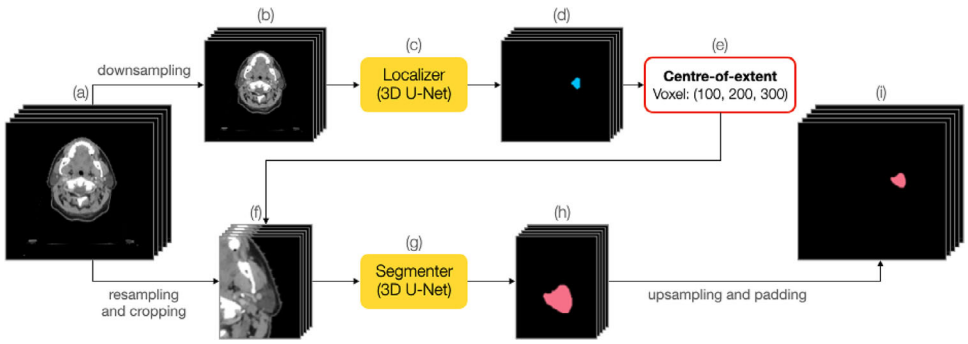

FIGURE 2 Inference pipeline for two-stage auto-segmentation model. CT image (a) at scanning spatial resolution was downsampled to 4 mm isotropic (b) and passed to the localizer stage (c). The localizer stage produced a 3D binary prediction (d) with values indicating presence or absence of the OAR. The centre-of-extent (e) of the prediction was calculated. The CT image (a) was resampled to $1 \times 1 \times 2$ mm and a 3D patch (f) was extracted surrounding the localizer centre-of-extent. The CT patch was passed to the segmenter stage (g) and produced a high-resolution 3D binary prediction (h) of the OAR location. The prediction (h) was resampled and padded to produce final prediction (i) at spatial resolution and size of image (a). Connected component analysis was applied to (i) to select the largest contiguous structure.

3.2.3 | Data augmentation

Training data was augmented to prevent overfitting to the training dataset by applying affine transformations to CT images and OAR labels. Augmentation consisted of random translation (−50 to 50 mm), rotation (−5 to 5 degrees), and scaling (80–120%) along all axes, with new voxels assigned the minimum value of existing voxels (i.e., HU ≅ −1000 for CT images, 0 for OAR labels). When training the segmenter stage, variation in localizer prediction center-of-extent was simulated by applying uniform random translation of the OAR label within segmenter patch boundaries. Data augmentation was implemented using the torchio library (v0.18.84).[25]

3.2.4 | Model training

Models were trained using a supervised learning approach with paired CT images and OAR labels in batches of size one. Model parameters were updated using the stochastic gradient descent optimizer

(learning rate = 0.001, momentum = 0.9) and the dice loss objective.[26] Mixed precision training (16-bit) was applied to reduce the GPU memory footprint.

For each cross-validation fold, the training dataset was further split into a training dataset (80%) and a validation dataset (20%). After each training epoch, model performance on the validation dataset was calculated using the mean dice similarity coefficient (DSC). Models were trained for 150 epochs, except when training with smaller dataset sizes: n = 5, 10, and 20, which were trained for 900, 450, and 300 epochs, respectively, as smaller dataset sizes required more epochs for validation loss convergence. Final model parameters were selected using the highest validation score over all training epochs. Example training and validation curves are shown in Supplementary Figures 4 and 5.

Models were implemented in python (v3.8.2) and trained using the pytorch-lightning framework (v1.7.6),[27] with torch (v1.12.1).[28] Models were trained using the source code at https://github.com/clarkbab/hn-segmentation-with-transfer-learning.

3.3 | Performance metrics

Several common overlap and distance metrics were selected to evaluate model performance, including the DSC, 95th percentile Hausdorff distance (95HD), and mean surface distance (MSD). Additionally, the surface DSC[29] and added path length (APL)[30] were included as they show higher correlation with clinically-relevant measures, such as time saved during manual correction of auto-segmentations, than traditional segmentation metrics.[30,31] Metrics were aggregated over patients within the test fold using the mean, or lower/upper quartile (Q1/Q3) statistics. Q1 and Q3 were included as they provide an estimate of worst-case performance for overlap and distance metrics, respectively. For metric definitions see Appendix A.

3.4 | Computation platform and times

GPU hardware consisted of NVIDIA P100 GPUs with 12 GB VRAM, using CUDA Toolkit v10.1. Inference times were measured programmatically and ranged from 5.7 to 6.6 s for the localizer stages and from 7.6 to 9.6 s for the segmenter stages (Supplementary Table 5).

4 | RESULTS

4.1 | Public model generalizability

The public models performed significantly worse ($p < 0.05$) than the institutional models (trained on all institutional training samples) under most mean metrics for 13 OARs (Figure 3, Supplementary Table 6). For eight of these OARs (brain, brainstem, left/right lens, mandible, left/right parotid, right submandibular), the performance differences between public and institutional models were small (DSC $\leq$ 0.03, MSD < 0.5 mm). The public models performed significantly better than, or showed no significant difference to, the institutional models under most mean metrics for four OARs (left/right optic nerve, spinal cord, left submandibular). When using Q1/Q3 statistics instead of the mean, the public models performed significantly worse than the institutional models under most metrics for 12 OARs (Supplementary Table 7). The public models showed significantly better performance than, or no significant difference to, the institutional models under most metrics for five OARs (right lens, left/right optic nerve, spinal cord, left submandibular).

4.2 | Transfer versus institutional models

The transfer models (trained on all institutional samples) performed significantly better ($p < 0.05$) than the institutional models (trained on all institutional samples) under most mean metrics for seven OARs (brainstem, right lens, left/right optic nerve, left/right parotid, spinal cord) (Figure 3, Supplementary Table 8). For these seven OARs, the performance differences between transfer and institutional models were small (DSC $\leq$ 0.02, MSD < 0.4 mm). For a further nine OARs, there were no significant differences in performance between transfer and institutional models (left/right brachial plexus, brain, left/right cochlea, left lens, mandible, left/right submandibular). The transfer model performed significantly worse than the institutional model for the oral cavity. patient metrics were aggregated using Q1/Q3 instead of the mean, the transfer models performed significantly better than, or showed no significant difference to, the institutional models under most metrics for all OARs except the right brachial plexus and oral cavity (Supplementary Table 9).

4.3 | Effect of institutional dataset size

For institutional models to perform significantly better ($p < 0.05$) than the public models under most mean metrics, large numbers of institutional training samples (n, mean = 97 samples, range = 69–197 samples) were required for six OARs (brainstem, left/right lens, left/right parotid, right submandibular) (Figure 4, Supplementary Table 10). The highest numbers of samples were required for the left lens and left/right parotid (86, 86, and 197 samples, respectively). Small n (mean = 13, range = 5–31) were required for the institutional models to perform significantly better than the public models

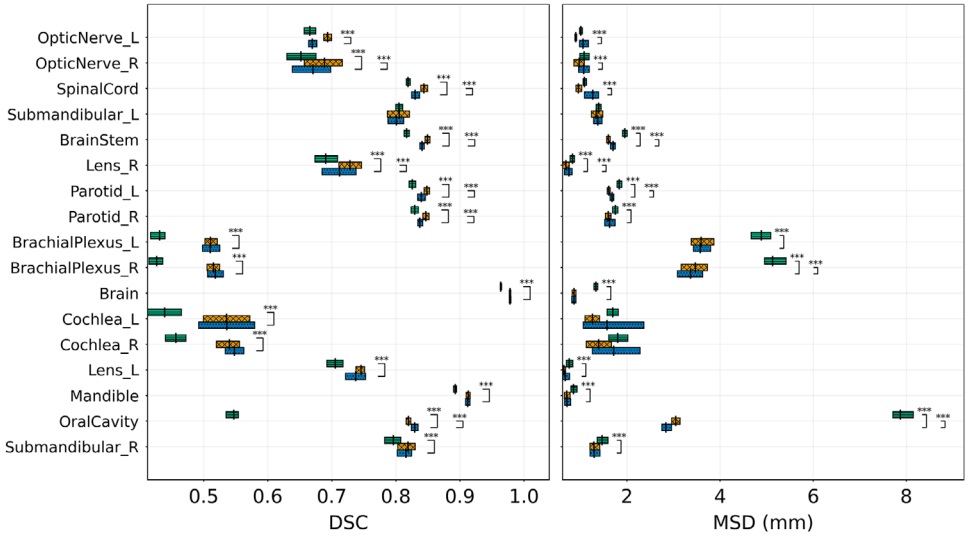

FIGURE 3 Mean (black vertical) and 95% confidence interval (bar) of public (green lined), transfer (orange crossed), and institutional (blue dotted) model performance for 17 OARs using DSC and MSD metrics. Transfer and institutional models were trained using all institutional training samples. OARs are grouped by best public model performance (left/right optic nerve, spinal cord, left submandibular), best transfer model performance (brainstem, right lens, left/right parotid), and all others. Significant differences in model performance are shown at levels: $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***). Significant differences between public and transfer models are not shown.

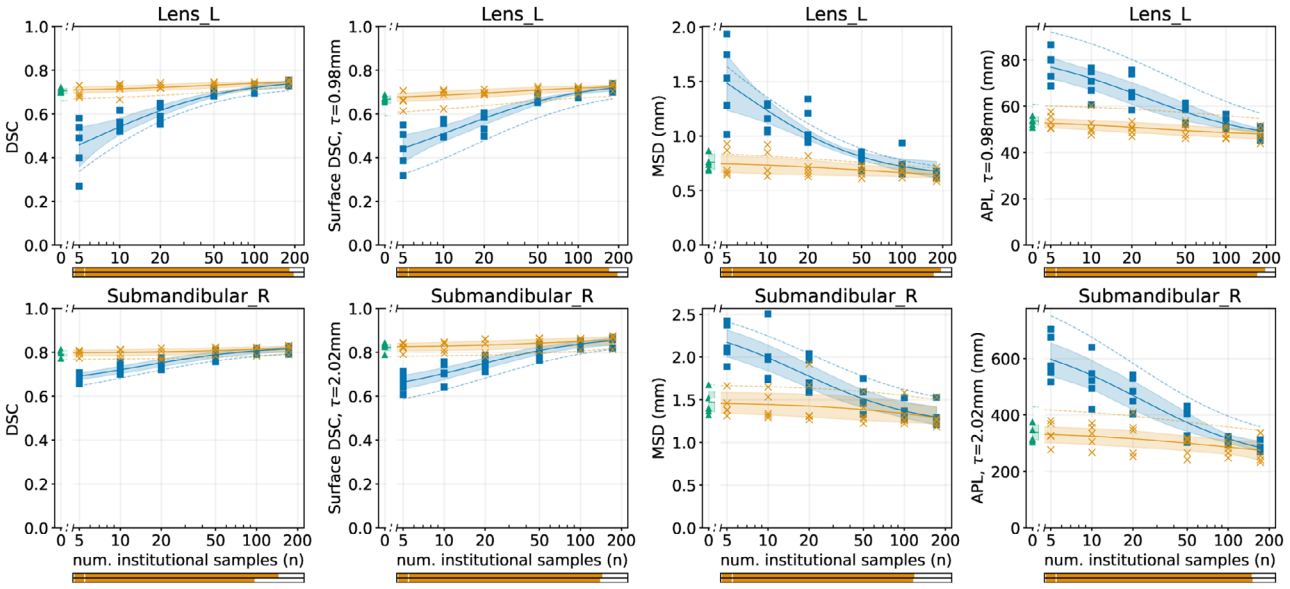

FIGURE 4 Example OARs for which the public (green triangle) and transfer (orange cross) models performed significantly better ($p < 0.05$) than the institutional models (blue square) for small numbers of institutional training samples (n). Scatter plots show mean test fold performance for each of five test folds against log(n). Solid curves show mean fitted values with 95% confidence intervals (shaded regions). Dashed curves show Q1 (for overlap metrics) and Q3 (for distance metrics) as a measure of worst-case performance. Significance markers are shown below each line plot to indicate significant differences ($p < 0.05$) between transfer and institutional models. Significant differences in mean performance (top marker) and Q1/Q3 performance (bottom marker) are shown, with orange indicating better transfer model performance, blue indicating better institutional model performance, and white indicating no significant difference.

under most metrics for seven OARs (left/right brachial plexus, brain, left/right cochlea, mandible, oral cavity). For results for all OARs, see Supplementary Figure 6. The lowest required numbers of institutional samples were for the left/right brachial plexus, brain, and oral cavity (11, 7, 6, and 5 samples, respectively). When patient metrics were aggregated using Q1/Q3 instead of the mean, large n were required (mean = 112 samples, range = 66–232 samples) for the institutional models to surpass public model performance for five

OARs (brainstem, left lens, left/right parotid, right submandibular) (Supplementary Table 11). The highest n were observed for the left/right parotid and right submandibular (90, 232, and 90 samples, respectively). Lowest n were observed for the brain, right cochlea, and oral cavity (five samples each).

For the institutional models to match (show no significant difference to, $p < 0.05$) transfer model performance under most mean metrics, large numbers of institutional training samples (mean = 150, range = 123–196) were required for five OARs (brain, left lens, mandible, left/right submandibular) (Figure 4, Supplementary Table 12). Largest n were required for the left lens, mandible, and right submandibular (169, 196, and 134, respectively). Small n (mean = 20, range = 5–42) were required for five OARs (left/right brachial plexus, left/right cochlea, oral cavity). The smallest n were observed for the left/right cochlea and oral cavity (6, 5, and 5 samples, respectively). Using Q1/Q3, large n were required (mean = 153, range = 69–334) for the institutional models to match transfer model performance under most metrics for seven OARs (brain, left/right lens, mandible, right optic nerve, left/right submandibular) (Supplementary Table 13). The largest n were observed for the left lens, mandible, and right parotid (159, 166, and 334 samples, respectively). Lowest n were observed for the left/right cochlea and oral cavity (five samples each).

5 | DISCUSSION

When developing auto-segmentation models for institutional use, models trained with public data alone were sufficient for four OARs: the left/right optic nerve, spinal cord, and left submandibular, thereby eliminating the need for curation of a large institutional training dataset. Performance differences between the public and institutional models may be small enough to permit use of the public models for a further eight OARs: the brain, brainstem, left/right lens, mandible, left/right parotid, and right submandibular. Curation of an institutional training dataset was necessary for the left/right brachial plexus, left/right cochlea, and oral cavity due to poor public model performance when compared with the institutional models.

When an institutional dataset was available, using transfer learning with pretrained public models showed performance improvements over random initialisation for seven OARs: the brainstem, right lens, left/right optic nerve, left/right parotid, and spinal cord. However, the magnitudes of the improvements were small and may not necessitate the use of pretrained public models.

When numbers of institutional training samples were limited, use of the public models was beneficial for most (ten) OARs: the brainstem, left/right lens, left/right optic nerve, left/right parotid, spinal cord, and left/right submandibular. Use of transfer models was beneficial with limited institutional samples for most (12) OARs: the brain, brainstem, left/right lens, mandible, left/right optic nerve, left/right parotid, spinal cord, and left/right submandibular.

Similarly to Chen et al.,[10] we found that transfer learning can improve model performance when models are pretrained on a large and diverse dataset. Chen et al. saw a large increase in model performance (from 0.71 to 0.94 mean DSC) when applying transfer learning to the task of lung segmentation, in comparison to random initialization of model parameters. Karimi et al.[15] found no significant improvement over a randomly-initialized model when employing transfer learning for magnetic resonance imaging segmentation; however, their models were pretrained on a single source dataset that may not have captured enough variation in imaging or anatomy. We also observed, in agreement with the results of Ghafoorian et al.,[32] that model performance improves with the number of target samples used during fine-tuning.

The public models displayed particularly poor performance when segmenting the oral cavity on the institutional dataset. A qualitative evaluation of the public training data showed that 53% of labels were segmented contrary to guidelines[4] employed at our institution (Supplementary Figure 7). Public training data was segmented in 2015, prior to the creation of these guidelines, emphasising the importance of data age and segmentation guidelines when curating a training dataset.

Our model performance was validated by comparison with two HN auto-segmentation reviews (Supplementary Figure 8).[29,33] Performance was within the DSC range of DL model performance for all OARs except the left/right cochlea. For the left/right cochlea, model performance was below the given range for DL models; however, this range was drawn from three studies alone. Figure 5 shows example predictions for the institutional models trained on all institutional samples. Due to the number of models trained in this work, a qualitative evaluation of model performance was not performed. For example, model predictions see Supplementary Figures 9 and 10.

While this work focused on a single model architecture and transfer learning approach, it may be worthwhile to apply different architectures and approaches to this problem. Several modifications have been proposed to the 3D U-Net architecture,[23] including attention gating of high-resolution encoder features[34] and improving the interaction of global features through self-attention layers within the encoder.[35] Other potential transfer learning methods include instance-based methods, in which samples are weighted according to their relevance to the target domain, or feature-based methods, in which feature representations learned from a source domain are utilised for a related task.[22]

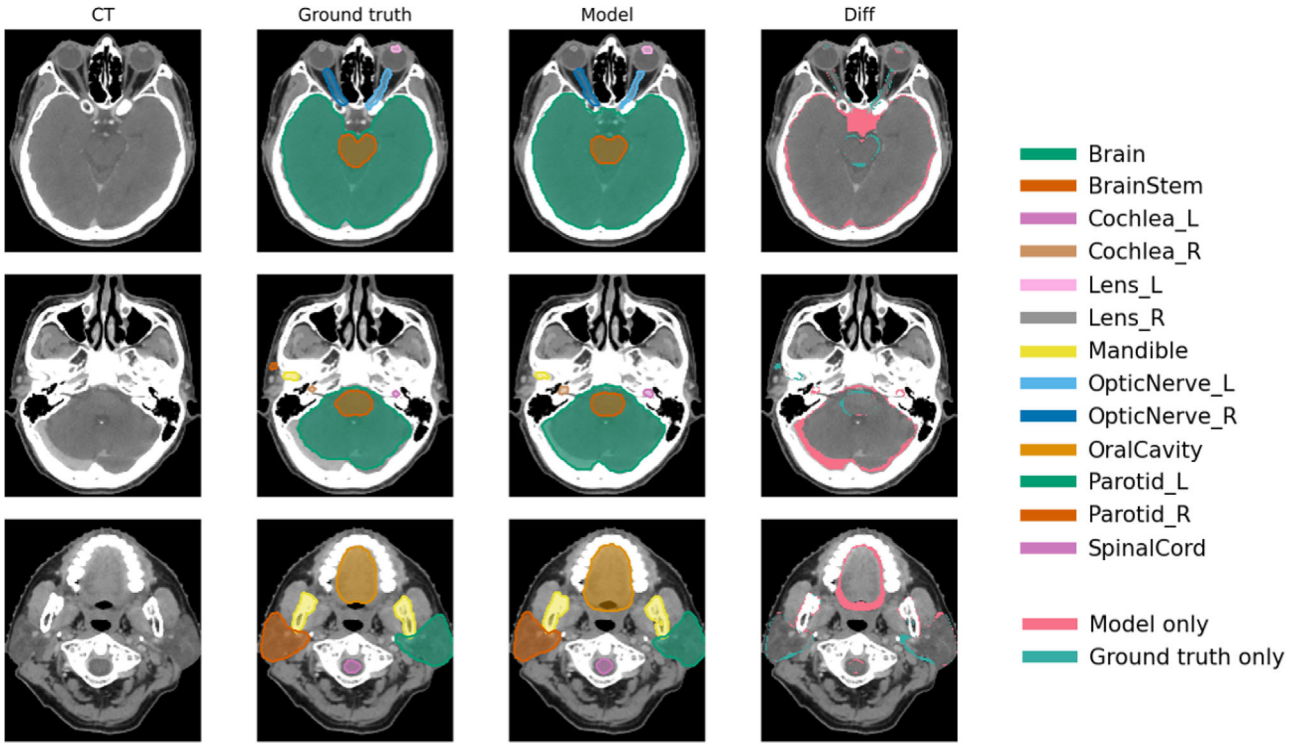

FIGURE 5 Example institutional model (trained on all institutional samples) predictions for a patient selected at random from the PMCC dataset. Rows: three axial slices are shown to display 13 OARs. Columns: the patient CT image, ground truth OAR label, model prediction, and difference between prediction and ground truth are shown for each axial slice. The difference column displays no difference for voxels where differences are present for multiple OARs but with different colours.

6 | CONCLUSION

We investigated the utility of large public datasets when applied to the problem of building CT auto-segmentation models for use in head and neck RT planning. For a small number of OARs, the use of public data in place of institutional data was suitable (left/right optic nerve, spinal cord, left submandibular). Using public data in place of an institutional dataset incurred a small performance deficit for most other OARs (brain, brainstem, left/right lens, left/right parotid, mandible, right submandibular), but may be preferable over curation of a large institutional dataset. In the presence of a large institutional dataset, transfer learning with pretrained public models provided a modest performance improvement for several OARs (brainstem, right lens, left/right optic nerve, left/right parotid, spinal cord). When numbers of institutional samples were limited, using the public dataset alone, or with transfer learning, was beneficial for most OARs (brainstem, left/right lens, left/right optic nerve, left/right parotid, spinal cord, and left/right submandibular). These results suggest that public data alone may be sufficient for developing many deep learning auto-segmentation models, but access to institutional training data may provide minor performance improvements for specific organs. Future work should assess the clinical impact of performance differences between public and institutional models through radiation dose and clinician acceptance rates.

ACKNOWLEDGMENTS

This research was supported by an Australian Government Research Training Program (RTP) scholarship.

This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

The authors would like to acknowledge Cameron Patrick (University of Melbourne, Australia) for providing advice on statistical methods, Femke Vaassen (Maastricht University Medical Centre, The Netherlands) for correspondence regarding the calculation of the APL metric, and Stanslav Nikolov (DeepMind, UK) for correspondence regarding calculation of the surface DSC metric.

CONFLICT OF INTEREST STATEMENT

Nicholas Hardcastle receives research grant support from Varian Medical Systems and Reflexion Medical for unrelated research.

DATA AVAILABILITY STATEMENT

Public datasets are available for download from The Cancer Imaging Archive. The institutional dataset is unavailable due to patient privacy restrictions.

REFERENCES

1. Delaney G, Jacob S, Barton M. Estimation of an optimal external beam radiotherapy utilization rate for head and neck carcinoma. *Cancer*. 2005;103(11):2216-2227. doi:10.1002/cncr.21084
2. Nutting CM, Morden JP, Harrington KJ, et al. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial. *Lancet Oncol*. 2011;12(2):127-136. doi:10.1016/S1470-2045(10)70290-4
3. Boero IJ, Paravati AJ, Xu B, et al. Importance of radiation oncologist experience among patients with head-and-neck cancer treated with intensity-modulated radiation therapy. *J Clin Oncol*. 2016;34(7):684-690. doi:10.1200/JCO.2015.63.9898
4. Brouwer CL, Steenbakkers RJHM, Bourhis J, et al. 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*. 2015;117(1):83-90. doi:10.1016/j.radonc.2015.07.041
5. Kosmin M, Ledsam J, Romera-Paredes B, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiother Oncol*. 2019;135:130-140. doi:10.1016/j.radonc.2019.03.004
6. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys*. 2017;44(2):547-557. doi:10.1002/mp.12045
7. Brunenberg EJL, Steinseifer IK, van den Bosch S, et al. External validation of deep learning-based contouring of head and neck organs at risk. *Phys Imaging Radiat Oncol*. 2020;15:8-15. doi:10.1016/j.phro.2020.06.006
8. Hänsch A, Schwier M, Gass T, et al. Evaluation of deep learning methods for parotid gland segmentation from CT images. *J Med Imaging*. 2019;6(1). doi:10.1117/1.JMI.6.1.011005
9. Wong J, Fong A, McVicar N, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother Oncol*. 2020;144:152-158. doi:10.1016/j.radonc.2019.10.019
10. Chen S, Ma K, Zheng Y. Med3D: Transfer Learning for 3D Medical Image Analysis. *ArXiv190400625 Cs*. Published online July 17, 2019. Accessed January 27, 2021. http://arxiv.org/abs/1904.00625
11. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *ArXiv170702968 Cs*. Published online August 3, 2017. Accessed March 12, 2021. http://arxiv.org/abs/1707.02968
12. Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *Eur Radiol Exp*. 2020;4(1):50. doi:10.1186/s41747-020-00173-2
13. van Dijk LV, Van den Bosch L, Aljabar P, et al. Improving automatic delineation for head and neck organs at risk by deep learning contouring. *Radiother Oncol*. 2020;142:115-123. doi:10.1016/j.radonc.2019.09.022
14. Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. *Proc IEEE*. 2021;109(1):43-76. doi:10.1109/JPROC.2020.3004555
15. Karimi D, Warfield SK, Gholipour A. Transfer learning in medical image segmentation: new insights from analysis of the dynamics of model parameters and learned representations. *Artif Intell Med*. 2021;116:102078. doi:10.1016/j.artmed.2021.102078
16. Chen Q, Bernard ME, Duan J, Feng X. A transfer learning approach for improving OAR segmentation in the adaptive therapy or retreatment of head and neck cancer. *Int J Radiat Oncol Biol Phys*. 2021;111(3):e125-e126. doi:10.1016/j.ijrobp.2021.07.550
17. Wee L, Dekker A. Data from Head-Neck-Radiomics-HN1. Published online 2019. doi:10.7937/TCIA.2019.8KAP372N
18. Vallières M, Kay-Rivest E, Perrin L, et al. Data from Head-Neck-PET-CT. Published online 2017. doi:10.7937/K9/TCIA.2017.8OJE5Q00
19. Grossberg A, Mohamed A, El Halawani H, et al. Data from Head and Neck Cancer CT Atlas. Published online 2017. doi:10.7937/K9/TCIA.2017.UMZ8DV6S
20. Kwan JYY, Su J, Huang SH, et al. Data from Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in Oropharyngeal Carcinoma. Published online 2019. doi:10.7937/TCIA.2019.8DHO2GLS
21. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a Public Information Repository. *J Digit Imaging*. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7
22. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22(10):1345-1359. doi:10.1109/TKDE.2009.191
23. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T. Ronneberger O, 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *ArXiv160606650 Cs*. Published online June 21, 2016. Accessed November 4, 2020. http://arxiv.org/abs/1606.06650
24. Yaniv Z, Lowekamp BC, Johnson HJ, Beare R. SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. *J Digit Imaging*. 2018;31(3):290-303. doi:10.1007/s10278-017-0037-8
25. Pérez-García F, Sparks R, TorchIO OurselinS. A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Programs Biomed*. 2021;208:106236. doi:10.1016/j.cmpb.2021.106236
26. Milletari F, Navab N, Ahmadi S. V-Net: fully convolutional neural networks for volumetric medical image segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*. 2016:565-571. doi:10.1109/3DV.2016.79
27. Falcon WA. The PyTorch Lightning team. Pytorch lightning. *GitHub*. Published online 2019. https://github.com/Lightning-AI/lightning
28. Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst*. 2019:32.
29. Nikolov S, Blackwell S, Zverovitch A, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *ArXiv180904430 Phys Stat*. Published online October 20, 2020. Accessed November 3, 2020. http://arxiv.org/abs/1809.04430
30. Vaassen F, Hazelaar C, Vaniqui A, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol*. 2020;13:1-6. doi:10.1016/j.phro.2019.12.001
31. Kiser KJ, Barman A, Stieb S, Fuller CD, Giancardo L. Novel autosegmentation spatial similarity metrics capture the time required to correct segmentations better than traditional metrics in a thoracic cavity segmentation workflow. *J Digit Imaging*. 2021;34(3):541-553. doi:10.1007/s10278-021-00460-3
32. Ghafoorian M, Mehrtash A, Kapur T. Transfer learning for domain adaptation in MRI: application in brain lesion segmentation. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*. Springer International Publishing; 2017:516-524. doi:10.1007/978-3-319-66179-7_59. Lecture Notes in Computer Science.

33. Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods. *Med Phys*. 2020;47(9):e929-e950. doi:10.1002/mp.14320
34. Oktay O, Schlemper J, Folgoc LL, et al. Attention U-Net: Learning Where to Look for the Pancreas. Published online May 20, 2018. doi:10.48550/arXiv.1804.03999
35. Chen J, Lu Y, Yu Q, et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. Published online February 8, 2021. doi:10.48550/arXiv.2102.04306
36. Podobnik G, Strojan P, Peterlin P, Ibragimov B, HaN-Seg VrtovecT. The head and neck organ-at-risk CT and MR segmentation dataset. *Med Phys*. 2023;50(3):1917-1927. doi:10.1002/mp.16197
37. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26(3):297-302. doi:10.2307/1932409

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Clark B, Hardcastle N, Johnston LA, Korte J. Transfer learning for auto-segmentation of 17 organs-at-risk in the head and neck: Bridging the gap between institutional and public datasets. *Med Phys*. 2024;1-11. https://doi.org/10.1002/mp.16997

APPENDICES

A. Performance Metrics

a. Definitions

The dice similarity coefficient (DSC)[37] measures the overlap between sets of voxels A (prediction) and B (ground truth).

$$DSC = \frac{2\,|A \cap B|}{|A| + |B|} \tag{A.1}$$

The surface dice similarity coefficient (surface DSC) measures the overlap between the sets of surface voxels of A and B, with acceptable deviation of the surfaces encapsulated by parameter τ. Given the set of voxels within an acceptable deviation of the surface of S

$$S_\tau = \{x \in \Omega :\ \exists s \in \delta S,\ \|x - s\| \le \tau\} \tag{A.2}$$

where Ω is the set of voxels in a 3D binary image and δS is the set of surface voxels of S, the surface DSC was calculated using

$$surface\ DSC = \frac{|\delta A \cap B_\tau| + |\delta B \cap A_\tau|}{|\delta A| + |\delta B|} \tag{A.3}$$

Distance metrics: 95th percentile Hausdorff distance (95HD) and mean surface distance (MSD) measure the discrepancy between sets of surface voxels δA and δB. The added path length (APL) measures the length in mm of the path that must be added to δA (prediction) to produce δB (ground truth) with a tolerance of τ (mm). These metrics require the calculation of the sets of minimum surfaces distances from A to B

$$D_{AB} = \left\{ \forall a \in \delta A : \min_{b \in \delta B} \|b - a\| \right\} \tag{A.4}$$

Distance metrics were subsequently calculated using

$$95HD = \max\{P_{95}(D_{AB}),\ P_{95}(D_{BA})\} \tag{A.5}$$

$$MSD = \frac{1}{2}\left(\frac{\sum_{d \in D_{AB}} d}{|D_{AB}|} + \frac{\sum_{d \in D_{BA}} d}{|D_{BA}|} \right) \tag{A.6}$$

$$APL = S_{sag} * |\{d \in D_{BA} : d > \tau\}| \tag{A.7}$$

where P_{95} is the 95th percentile of the set of minimum distances, and S_{sag} is the CT spacing along the sagittal axis, assuming isotropic voxel spacing in the axial plane.

All metrics were calculating using the size and voxel spacing of the original CT image. APL, DSC, HD_{95}, and MSD metrics were calculated using the SimpleITK library (python; v2.1.0). Surface DSC was calculated using the implementation at https://github.com/deepmind/surface-distance. Values for τ were taken from Nikolov et al.[29] and were calculated per OAR using the 95th percentile of minimum surface distances between segmentations created by two expert observers.

b. Curve Fitting

Curves were fitted to model performance data (evaluated on five test folds) against number of institutional training samples (n) using non-linear least squares (scipy; v1.6.0) with the following model:

$$f(x) = \frac{-a}{x - b} + c$$

where a is a slope parameter, b is the location of the vertical asymptote and c is the location of the horizontal asymptote.

95% confidence intervals were calculated by repeating the curve fitting procedure for 10 000 bootstrap samples of the model performance data (drawn with replacement). The confidence interval was then calculated using the 2.5th/97.5th percentiles of the fitted values for each n.

c. Statistical Tests

Significant difference ($p < 0.05$, $p < 0.01$, or $p < 0.001$) in performance between two models for a given number

of institutional training samples (n) was determined by calculating the distribution of paired difference (10 000 differences) between each model's fitted values. N could be different for each model, e.g., when comparing public (n = 0) and institutional (n = 340) model performance. Performance was significantly different ($p < 0.05$) when 95% of differences lay above or below zero (99% and 99.9% of differences for $p < 0.01$ and $p < 0.001$, respectively).

B. Model Training Times

Model training times varied with number of training samples and input/label data size (Supplementary Table 5). The left/right parotid OARs had the highest training times (7.5/7.4 days) amongst public localizer stages due to the large number of training samples, whilst amongst the public segmenter stages, the spinal cord had the highest time (8.7 days) due to large input patch size. For both institutional and transfer segmentation stages, the mandible had the highest training times when using all available institutional samples (2.5/3.4 days).

C. Spinal Cord

Due to the large axial length of the spinal cord, localizer stage predictions were cropped at the caudal end to a maximum extent of 290 mm. This ensured that the second stage patch covered the cranial end of the spinal cord. 290 mm was chosen as it provided the second stage spinal cord patch (330 mm axial length) with a 20 mm margin on either end of the cropped localizer prediction to account for variation in localizer prediction centre-of-extent.

Spinal cord predictions were cropped to the lowest axial slice (caudal end) of the ground truth before calculating metrics to avoid penalising predictions that extended further in the caudal direction than the ground truth segmentation.