Accurate segmentation of head and neck radiotherapy CT scans with 3D CNNs: consistency is key

To cite this article: Edward G A Henderson et al 2023 Phys. Med. Biol. 68 085003

View the article online for updates and enhancements.

You may also like
- Self-derived organ attention for unpaired CT-MRI deep domain adaptation based MRI segmentation
  Jue Jiang, Yu-Chi Hu, Neelam Tyagi et al.
- A comparison of methods for fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers
  Aurora Rosvoll Groendahl, Ingerid Skjel Knudtsen, Bao Ngoc Huynh et al.
- A proposed framework for consensus-based lung tumour volume auto-segmentation in 4D computed tomography imaging
  Spencer Martin, Mark Brophy, David Palma et al.
Accurate segmentation of head and neck radiotherapy CT scans with 3D CNNs: consistency is key

Edward G A Henderson, Eliana M Vasquez Osorio, Marcel van Herk, Charlotte L Brouwer, Roel J H M Steenbakkers and Andrew F Green

1 Division of Cancer Sciences, The University of Manchester, M13 9PL Manchester, United Kingdom
2 Department of Radiotherapy Related Research, The Christie NHS Foundation Trust, M20 4BX Manchester, United Kingdom
3 Department of Radiation Oncology, University Medical Center Groningen, 9713 GZ Groningen, The Netherlands

E-mail: edward.henderson@postgrad.manchester.ac.uk

Supplementary material for this article is available online

Abstract

Objective. Automatic segmentation of organs-at-risk in radiotherapy planning computed tomography (CT) scans using convolutional neural networks (CNNs) is an active research area. Very large datasets are usually required to train such CNN models. In radiotherapy, large, high-quality datasets are scarce and combining data from several sources can reduce the consistency of training segmentations. It is therefore important to understand the impact of training data quality on the performance of auto-segmentation models for radiotherapy. Approach. In this study, we took an existing 3D CNN architecture for head and neck CT auto-segmentation and compare the performance of models trained with a small, well-curated dataset (n = 34) and then a far larger dataset (n = 185) containing less consistent training segmentations. We performed 5-fold cross-validations in each dataset and tested segmentation performance using the 95th percentile Hausdorff distance and mean distance-to-agreement metrics. Finally, we validated the generalisability of our models with an external cohort of patient data (n = 12) with five expert annotators. Main results. The models trained with a large dataset were greatly outperformed by models (of identical architecture) trained with a smaller, but higher consistency set of training samples. Our models trained with a small dataset produce segmentations of similar accuracy as expert human observers and generalised well to new data, performing within inter-observer variation. Significance. We empirically demonstrate the importance of highly consistent training samples when training a 3D auto-segmentation model for use in radiotherapy. Crucially, it is the consistency of the training segmentations which had a greater impact on model performance rather than the size of the dataset used.

1. Introduction

One in two people will get cancer in their lifetime (Ahmad et al 2015) and many of them will receive radiotherapy as part of their treatment; this is the case for approximately 80% of head and neck (HN) cancers (Strojan et al 2017). The 3D segmentation of organs-at-risk (OARs) in planning CT images is a crucial step in the radiotherapy pathway as they are used to optimise the radiation dose delivered during treatment. Segmentation or delineation by clinicians is, however, slow and expensive. Moreover, clinicians are prone to inter- and intra-observer variability even when following detailed delineation consensus guidelines (Brouwer et al 2012, 2015, Vinod et al 2016). Any errors in the segmentation can have a serious impact on the patient’s treatment, for example, errors for normal tissues could result in healthy organs receiving unnecessarily high doses, increasing the likelihood of side effects. Alternatively, convolutional neural networks (CNNs) can be used to generate these segmentations. Fully convolutional CNNs are now the state-of-the-art (Cardenas et al 2019, Brouwer et al 2020, Henderson et al 2022) for automated medical image
segmentation and several methods are being actively proposed and implemented to allow segmentation to be performed faster and with higher consistency (Vrtovec et al 2020, Yu et al 2021).

CNNs are a deep learning method that make predictions based on previously seen training data. In general, deep learning models trained with larger datasets will generalise better to unseen data; for classical computer vision tasks, performance scales logarithmically with dataset size (Sun et al 2017). However, in radiotherapy, large, highly consistent datasets (e.g. of a few hundred patients) are scarce (Nikolov et al 2018, van Dijk et al 2020).

An alternative often used is to expand the training dataset with samples from an ever-growing pool of publicly-available sources, such as data repositories (Clark et al 2013, Nikolov et al 2018), segmentation challenges (e.g. Raudasz 2017) and competition websites (e.g. Kaggle). However, this method risks introducing low quality or inconsistent training samples and contaminating the training dataset. It is therefore crucial to understand the attainable performance of segmentation models trained with small datasets and understand how the quality of the training data impacts the performance of 3D segmentation models.

Auto-segmentation for medical imaging has been extremely well studied for different architectures, anatomical sites and training methodologies (Cicék et al 2016, Nikolov et al 2018, Wong et al 2018, Cardenas et al 2019, van Dijk et al 2020) however few have investigated the role of the quality of the training data. Those that do include the studies of Barragán-Montero et al (Barragán-Montero et al 2021) and Wesemeyer et al (Wesemeyer et al 2021). In their study, Barragán-Montero and colleagues investigated the effect of data quality and quantity on the performance of CNN models for radiotherapy dose prediction (Barragán-Montero et al 2021). They found their model performance improved when trained with a high-quality database, and degraded when the database was extended with low quality data. Wesemeyer and colleagues studied the impact of dataset size versus quality for a 2D segmentation task with RGB images of the eyeball (Wesemeyer et al 2021). Crowd-sourced annotations from non-expert observers were combined with a consensus algorithm (STAPLE (Warfield et al 2004)) to simulate different quality gold-standard segmentations. Increasing annotation quality level in the training data was shown to improve the performance of the 2D segmentation model. These studies agree that the quality of the training data heavily impacts model performance. However, to the best of our knowledge, ours is the first study to compare identical CNN models trained with datasets of differing annotation quality for 3D segmentation of medical images.

In this paper we have taken an existing 3D HN auto-segmentation method shown to be capable of producing high-quality OAR contours (Henderson et al 2022) and evaluated the impact of the consistency of the data used to train the CNN model on its performance. To do this, separate CNN models were trained using either a small dataset (n = 34) with highly curated segmentations produced by a single observer and audited by multiple highly-trained clinicians, or with a much larger dataset (n = 185) which contained less consistent segmentations. Each method was validated in an external cohort of patients with segmentations from multiple expert observers and including consideration of inter-observer variation.

2. Materials and methods

2.1. Datasets

We trained CNN models for 3D auto-segmentation of head and neck (HN) CT scans. The HN OARs included in this segmentation study were the left and right parotid glands, the spinal cord and the brainstem. Three distinct datasets were used in this work. We describe each of them here:

Dataset A consists of 34 HN planning CT scans sourced from a public open data repository (https://github.com/deepmind/tcia-ct-scan-dataset, Nikolov et al 2018).4 The size of these CT images range between 119 × 512 × 512 and 437 × 512 × 512 voxels, and have a voxel resolution of 1 × 1 × 2.5 mm. This particular dataset contains two sets of volumetric segmentations of the OARs for each CT image. One of these sets was manually segmented by a single radiographer, according to the consensus guidelines defined by Brouwer et al (Brouwer et al 2015), followed by arbitration from three further radiographers and a final check from a specialist oncologist with a minimum of 4 years experience planning head and neck cancer treatments. This set was selected as the gold-standard segmentations for dataset A in our study. The second set are the segmentations of the original radiographer, prior to arbitration, and were only used for an extended analysis in the supplementary materials.

Dataset B contains 185 planning CTs sourced from The Cancer Imaging Archive (TCIA), HNSCC dataset (Clark et al 2013, Elhalawani et al 2017, Grossberg et al 2018, 2020). All images in this dataset originate from a single institution (MD Anderson) and are accompanied by volumetric segmentations. These delineations were

4 One of the original 35 CT images was omitted as the patient had had the left portion of the mandible surgically removed. This procedure is unrepresentative in such a small dataset.
used in the course of routine treatment, but no explicit information is given on the particular observer’s experience or the guidelines followed (Grossberg et al 2018). As such, these segmentations are expected to be less consistent than those in dataset A. The images in this dataset ranged in size from $85 \times 512 \times 512$ and $348 \times 512 \times 512$ voxels and were resampled to a voxel resolution of $1 \times 1 \times 2.5$ mm if required. This meant resampling roughly 17% of the images which had an original slice thickness of 3 mm. CT scans were resampled using third order spline interpolation with anti-aliasing and the volumetric segmentations were resampled using nearest-neighbour interpolation.

**Dataset C** contains an external cohort of six patients from a single institution (Brouwer et al 2012). Each patient had CT imaging at two time points for a total of 12 CT images with sizes ranging from $143 \times 512 \times 512$ to $190 \times 512 \times 512$ voxels. These CT images had axial-plane resolution ranging from 0.94 to 1.27 mm and an original slice thickness of 2 mm which was resampled to 2.5 mm using the same methods as described for dataset B. Volumetric segmentations of the OARs, except the brainstem, were produced by five different expert observers. For each OAR, we created a baseline contour from the five expert observers by applying the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm (Warfield et al 2004).

As we used open data for training our models (dataset A and dataset B) and data external to our institution to assess model generalisation (dataset C), we could not guarantee consistency on tumour location, contrast, immobilisation strategies amongst the datasets. Please refer to section 4 for a more detailed discussion on this.

2.2. Data preprocessing

In order to train a 3D CNN model without exceeding our GPUs memory capacity, full field-of-view CT images require downsizing (either by cropping or down-sampling). Preliminary experiments showed that downsampling the CT images resulted in decreased segmentation performance, therefore we decided to crop the images prior to segmentation, see figure 1, so the CNN could be trained with full-resolution CT scans ($\sim 1 \times 1 \times 2.5$ mm voxels). A sub-volume of $56 \times 200 \times 200$ voxels was cropped symmetrically around the centre-of-mass of the parotid glands and brainstem. The position of the crop centre was predicted with an in-house localisation CNN model, which has been shown to predict soft-tissue anatomical locations in CT scans to a 3D Euclidean distance of $5.8 \pm 5.0$ mm ($\sim 2–3$ voxels) (Henderson et al 2022). The preprocessing applied in these experiments was fully-automated, requiring no manual input, and was performed in the same way during the training and inference stages for all the datasets.

2.3. Consistency of the datasets

To determine the consistency of the gold-standard segmentations of the training datasets and to identify possible bias in the testing dataset we completed the following analysis.

2.3.1. Segmentation consistency in the training datasets

Measuring the observer segmentation consistency across a dataset where only a single observer has delineated a given structure is a challenging question in itself, as anatomical differences may dominate the observed variations. We therefore first measured and compared the volumes of each of the delineated OARs in each dataset.

While this method gives a hint into the observer segmentation consistency, it is affected by inter-patient anatomical differences. To discern the observer segmentation consistency amidst the anatomical variations we directly compared the CT scans to obtain a measure of the inter-patient anatomical consistency. For this comparison, we calculated the normalised cross-correlation (NCC) for each pair of images within dataset A and for each pair within dataset B after automated cropping. NCC is a widely used measure of similarity between images with an ideal value of 1 (Brock et al 2017). By comparing the spread of the OAR volume distributions and

---

**Figure 1.** A schematic of the automated preprocessing performed on the CT images in all the datasets. Using a localisation model to identify the position of the crop centre means cropped sub-volumes are anatomically consistent.
the inter-patient anatomical consistency for each dataset, we can gain insight into the relative observer segmentation consistency for datasets A and B.

### 2.3.2. Evaluating potential bias between the training and testing datasets

A factor which may affect the performance of segmentation models is whether the training and test datasets are in- versus out-of-distributions. For example, if dataset A was more similar to dataset C than dataset B, models trained with dataset A would gain an advantage when tested on our external dataset.

Dataset C is small ($n = 12$), therefore we focused on evaluating whether the imaging of our testing dataset is biased toward dataset A or B. We calculated the NCC for each pair of images within datasets A and C, and for each pair of images in datasets B and C. We then compared the resulting distributions of NCC values with a two-sample Kolmogorov–Smirnov (KS) test to assess whether the samples were drawn from different distributions.

For both evaluations 2.3.1 and 2.3.2 we calculated the NCC for the anatomically consistent image sub-volumes, obtained after preprocessing, section 2.2. This ensured the robustness of our consistency evaluation as the NCC measure is sensitive to misalignment.

### 2.4. Segmentation model

The segmentation CNN used in this study was previously developed and fully-described by Henderson et al (Henderson et al 2022). This CNN is based on a 3D UNet, with a symmetric structure of repeated convolutions and down-sampling, followed by convolutions and up-sampling (Cicek et al 2016). Residual skip connections smooth the training process (He et al 2016) and deep supervision accelerates model convergence. Group normalisation as described by Wu and He is applied after each convolutional layer (Wu and He 2018). A figure showing the CNN architecture is included in the supplementary materials.

Windowing (or grey-level mapping) is used to enhance the contrast of different tissues and highlight particular structures, while also normalising the image from Hounsfield units (HU) into the range $[0, 1]$ prior to inputting into the CNN. The CT image volumes are processed simultaneously by the CNN model in different windowing settings, chosen according to the OARs to be segmented (Henderson et al 2022). The selected window and level ($W, L$) contrast settings are commonly used to specifically view general soft tissue (350 HU, 40 HU), bony anatomy (2800 HU, 600 HU) and brain tissue (160 HU, 20 HU) (Hoang et al 2010).

### 2.5. Implementation details

All CNN models are implemented in PyTorch 1.10.1. All network training was performed using a single NVIDIA GeForce RTX 3090 GPU with 24 GB of memory.

We used the same implementation as previously described in Henderson et al (2022), including extensive data augmentation and the ‘Exponential Logarithmic Loss’ function of Wong et al (2018). The implementation details are specified in full in section SI.B of the supplementary materials.

### 2.6. Segmentation performance metrics

Overlap metrics, such as the Dice similarity coefficient (DSC), are commonly reported in auto-segmentation literature but can hide clinically relevant differences between structure boundaries as they are volume-biased and insensitive to fine details (Sharp et al 2014, Vrtovec et al 2020). Distance metrics are preferred in radiotherapy, where small deviations in the borders of segmentations can have a potentially serious impact (Peroni et al 2013). The distance metrics applied in this study were the mean distance-to-agreement (mDTA) and the 95th percentile maximum distance or 95th percentile Hausdorff distance (HD95) (Vrtovec et al 2020). The mDTA provides an assessment of the overall segmentation quality and HD95 highlights large errors. These metrics were calculated using the same methodology as described in Henderson et al (2022).

When evaluating the segmentation results in dataset C, in addition to calculating the HD95 and mDTA, we performed a 3D analysis of variation, as described by Brouwer et al called global 3D SD (Brouwer et al 2012). The global 3D SD provides an indication of the global consistency between observers. To compute the global 3D SD, the closest distance to each considered segmentation was calculated for 50 000 random surface points spread on the baseline contours. The standard deviation of these distances was computed for each surface point and summarised with the root mean square across all surface points. In practice, a lower global 3D SD value suggests the considered segmentations are more consistent.

Despite our arguments against the use of the DSC above, we recognise that, as a widely reported metric, it is of some use to report alongside distance metrics. Additionally, following the recommendation of Peroni et al, we also computed the volume difference between the gold standard and predicted segmentations (Peroni et al 2013). These additional results are included in section SIII of the supplementary materials for completeness.
2.7. Experiments
To investigate the impact of dataset size and gold-standard consistency on the CNN performances, multiple auto-segmentation models were produced in the presented experiments. Each had an identical architecture and training procedure, but were trained with different sets of images or datasets. This results in different sets of model parameters which we label $\theta_X$, where $X$ denotes the training dataset (either A or B).

2.7.1. Internal 5-fold cross-validation
Firstly, we trained two sets of models on datasets A and B using a 5-fold cross-validation procedure (Borra and Ciaccio 2010). Both datasets A and B were divided into 5 folds to produce five unique parameter sets. For each fold, data was assigned to a ‘training’ subset to train the model parameters, a ‘validation’ subset to inform adjustments to the learning rate and training completion, and a ‘testing’ subset constituting of held-out data to evaluate the model performance.

We performed a 5-fold cross-validation for dataset A using the highly curated segmentations of the brainstem, parotid glands and the cervical section of the spinal cord for supervision. Five models were trained from scratch, $\theta_{A1}, \theta_{A2}, \ldots, \theta_{A5}$ ($\theta_A$), each using 24 training images, 3 validation images, and 7 testing images.

We similarly performed a 5-fold cross-validation for dataset B using the segmentations of the same OARs. Correspondingly, five models were trained from scratch, $\theta_{B1}, \theta_{B2}, \ldots, \theta_{B5}$ ($\theta_B$), each with 131 training images, 16 validation images, and 38 testing images.

We calculated the distance metrics for each $\theta_A$ and $\theta_B$ model using their corresponding testing subset. We additionally compared the predictions of the $\theta_A$ models to the second set of segmentations included in this open dataset to evaluate how well such models perform compared with manual inter-observer deviation. This extended analysis can be found in the supplementary materials.

2.7.2. Evaluation on an external dataset
To assess model generalisability we evaluated our CNN models in an external cohort of patients (dataset C), using the identical preprocessing scheme as detailed in section 2.2. For this stage, we separately followed two common practices (Raschka 2018, Bates et al. 2021).

We first selected the best of the five cross-validation models for datasets A and B, based on the performance on the held-out testing sets, which were then evaluated on dataset C. Each cross-validation model was ranked according to the mean of the HD95 and mDTA metrics across all OARs on the test set images. The top-performing model from $\theta_A$ and $\theta_B$ was selected by rank aggregation.

Second, we trained two further, independent segmentation models on datasets A and B, referred to as $\theta_A$ and $\theta_B$, which were evaluated on dataset C. The $\theta_A$ model was trained using the entirety dataset A split into 28 training and 6 validation images, and the $\theta_B$ model trained on dataset B split into 152 training and 33 validation images. Identical training processes were followed, including the same data augmentation strategies described in section SI.B of the supplementary materials.

We compared the OAR segmentations produced by each of the CNN models on the images belonging to dataset C with those of the five expert observers using the baseline contour as the gold-standard (the STAPLE estimate, see section 2.1). The global 3D SD was calculated for the five manual observers, and then for the combinations of the manual observers and each of the considered CNNs ($\theta_A$, $\theta_B$ and the top performing $\theta_A$ and $\theta_B$ models). This analysis quantifies the contouring variability each CNN model introduces when considered alongside the manual observers.

3. Results
3.1. Consistency of the datasets
3.1.1. Segmentation consistency in the training datasets
Figure 2 shows violin plots illustrating the distribution of the OAR segmentation volumes for these datasets. The spread of volumes for the brainstem and spinal cord are considerably smaller in dataset A than B, and slightly smaller in the parotid glands.

The summary statistics for the NCC analysis are shown in table 1. This analysis shows the images in dataset B to be more similar to each other than those in dataset A as the median NCC value is higher and the interquartile range (IQR) is smaller. Since the images within dataset B are more similar to each other, we can conclude that the differences observed in the volumes of the structures are not driven by systematic differences in the patient anatomy within this dataset. Instead, this is quantitative evidence that dataset B has lower observer segmentation consistency than dataset A.
3.1.2. Evaluating potential bias between the training and testing datasets

The median and IQR of the NCC values of the pairwise comparisons between datasets A and C, and datasets B and C are also shown in Table 1. The result of the KS test was non-significant \((p = 0.43)\) which suggests that the NCC values are drawn from the same underlying distribution. Therefore, the relationships between the CT images of datasets A and C and datasets B and C are not significantly different. This eliminates the possibility of the models trained with either dataset A or B being disadvantaged when tested on dataset C due to out-of-distribution imaging.

### Table 1. Summary statistics for the NCC analysis.

<table>
<thead>
<tr>
<th>Image pairs</th>
<th>NCC median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset A versus A</td>
<td>0.454 (0.170)</td>
</tr>
<tr>
<td>Dataset B versus B</td>
<td>0.516 (0.114)</td>
</tr>
<tr>
<td>Dataset A versus C</td>
<td>0.449 (0.165)</td>
</tr>
<tr>
<td>Dataset B versus C</td>
<td>0.457 (0.181)</td>
</tr>
</tbody>
</table>

3.2. Experimental results

Training the \(\theta_{A1}, \ldots, \theta_{A5}\) and \(\theta_{A}\) models took \(\sim 28\) hrs each and the \(\theta_{B1}, \ldots, \theta_{B5}\) and \(\theta_{B}\) models took \(\sim 66\) h each. Generating a set of OAR segmentation predictions took less than a second per 3D CT image.

3.2.1. Internal 5-fold cross-validation

The HD95 and mDTA cross-validation results are shown in Figures 3(a) and (b). Results for models \(\theta_{A1}, \ldots, \theta_{A5}\) and \(\theta_{B1}, \ldots, \theta_{B5}\) are shown with green and red boxes respectively. A direct comparison between the \(\theta_{Ai}\) and \(\theta_{Bi}\) models is not relevant here as the metrics are calculated on different images with the baseline segmentations produced by different observers. However, these results do give an impression of how the different models’ performance compares, with the \(\theta_{Ai}\) models consistently producing seemingly better segmentations than the \(\theta_{Bi}\) for all four OARs. This result is further supported by the DSC and volume difference metrics in figure S3 of the supplementary materials.

3.2.2. Evaluation on an external dataset

Figure 4 shows box-plots of the HD95 and mDTA results for each of the five manual observers (blue shades) and each of the CNN model variations when evaluated on dataset C. These include the best of the \(\theta_{Ai}\) and \(\theta_{Bi}\) cross-validation models (\(\theta_{A4}\) in green and \(\theta_{B5}\) in red) and the \(\theta_{A}\) and \(\theta_{B}\) retrained models (green and red hatched). The STAPLE contours, estimated from the five manual observers’ segmentations, were used as a common baseline for calculating the HD95 and mDTA metrics. The CNN variants trained on the small dataset, A, generalised very well to the external cohort of patients (dataset C), achieving similar accuracy as the 5 human observers (see figure 5).

Both the \(\theta_{B}\) and \(\theta_{B5}\) models show poorer mDTA performance in the parotid glands compared to the manual observers and the dataset A model variants, and considerably poorer performance for the HD95 metric. The \(\theta_{B}\) also performed poorly in the spinal cord, however, the HD95 and mDTA results for the \(\theta_{B5}\) model were considerably better for this OAR and comparable to the manual observers. Models trained on dataset B,
especially $\theta_{B}$, failed to generalise well to the external cohort of patients and produced poor quality OAR segmentations. For example, see figure 5 where we show axial and sagittal CT slices of a patient from dataset C with the segmentations of the manual experts, the CNN models and the STAPLE prediction overlaid. 3D visualisations of the segmentations shown in figure 5 can be found in section IV of the supplementary materials.

The results of the global 3D SD analysis are presented in table 2. The global 3D SD result for the manual observers alone provides an indicative baseline measure of the inter-observer variation for each OAR. When each CNN ($\theta_{A}, \theta_{B}, \theta_{A}, \theta_{B}$) was added to the consideration the global 3D SD measure rose or fell, suggesting the addition of that CNNs segmentations increased or decreased the amount of variability present.

In general, for both parotid glands the addition of models trained on dataset A kept the global 3D SD stable. However, adding models trained on dataset B resulted in considerably higher global 3D SD measures, suggesting the predictions of these CNNs had a low level of agreement with the manual observers. For the spinal cord, the...
addition of each CNN to the manual observers had little effect on the global 3D SD. Figure 6 shows three examples of recurring errors made by the \( \theta_B \) model. The first failure mode shows the insensitivity of the \( \theta_B \) model towards the accessory parotid glands (APG). The APG are an anterior extension of salivary tissues that are present in >30% of the population (Rosa et al 2020) and are recommended for inclusion in the parotid gland OAR delineation (van de Water et al 2009, Brouwer et al 2015). The \( \theta_A \) model and several of the manual observers identified the presence of the APGs (figure 6(a)). Furthermore, the \( \theta_B \) model produces significant segmentation errors for a patient with a bolus covering the nose, incorrectly labelling part of a bolus as parotid gland. Another recurring error includes \( \theta_B \) recognising portions of the back of the head as spinal cord. Overall, the \( \theta_B \) model produces OAR segmentation predictions in reasonable locations but lacking critical detail or with significant errors.

The corresponding DSC and volume difference results for the evaluation on an external dataset are shown in figure S4 in the supplementary materials. However, both these metrics fail to identify the gross errors made by the \( \theta_B \) model shown in figure 6. The Hausdorff distance (100th percentile maximum distance or HD100) was additionally calculated in each of experiments 1, 2 and 3. The resulting trends of each were identical as observed for the exhibited HD95 metric (results not included).

4. Discussion

In this paper we took an established 3D HN auto-segmentation architecture, which is capable of producing high-quality OAR contours, and evaluated how the consistency of the gold-standard annotations in the training data effects model performance. We have showed that models, trained with a small, well-curated dataset (dataset
A), greatly outperform models of identical architecture which are trained on a far larger dataset containing less consistent gold-standard annotations (dataset B).

We validated all models with an external cohort of patients containing OAR segmentations provided by five expert observers (dataset C) and further assessed the quality of model-predicted segmentations using an analysis of observer variation approach. The results of this approach closely aligned with the narrative conveyed from the distance metrics in figure 4, that models trained with dataset A produced segmentations in greater agreement with the five expert observers. The one exception being the spinal cord global 3D SD measure for the combination of the manual observers and model $θ_B$. Both the HD95 and mDTA measures show large outliers for spinal cord segmentations of $θ_B$. However these large errors were not reflected by a rise in variation shown by an elevated global 3D SD. This case reveals a weakness in the global 3D SD measure in that the distances are calculated in a single direction, from the baseline contour to the closest point on each of the observers contours. This implementation is unlikely to be sensitive to large outliers in the observers contours or disconnected erroneous segmentation volumes. To investigate this potential we visually assessed the predicted segmentations and found evidence of such failures (figure 6(c)), accounting for the discrepancy between the distance metrics and global 3D SD measure for $θ_B$.

In general, the models trained with dataset A produced superior OAR segmentations than those trained with dataset B, both in terms of the measured distance metrics (HD95 and mDTA) and in the global 3D SD analysis. Combinations of observers including the $θ_A$ and $θ_B$ predictions were shown to introduce a higher geometrical variation when compared to the five manual observers. Whereas, the global 3D SD measures for combinations containing predictions from models trained on dataset A were very similar to that for the combination containing the five manual observers alone.

We believe the key factor for the improved segmentation performance was the consistency of the gold-standard annotations. Dataset A has gold-standard OAR segmentations performed by a single radiographer and further audited by four highly-trained clinicians each following the same delineation guidelines. Meanwhile, dataset B—over five times the size of A—contains OAR gold-standard segmentations created by different observers in the course of routine treatment.

We assessed consistency of the OAR segmentations in the training datasets A and B and found quantitative evidence that the the observer segmentation consistency of dataset A was higher than that of dataset B. This evidence is based on the segmentation volumes which also include anatomical differences. To evaluate the role of anatomical differences we performed an analysis of the image similarity of pairs of images in the respective datasets using the NCC. NCC is a measure of both image similarity but also alignment. This opens the possibility that the higher NCC values for dataset B are due to the images being better globally aligned compared to those in dataset A. However, the NCC measure was computed after the CT scans were cropped to an anatomically consistent sub-region, section 2.2. Therefore, the NCC was computed for globally aligned images in all cases, ensuring the robustness and relevance of our consistency evaluation. The impact of the lower observer segmentation consistency of dataset B can be seen directly in the decreased performance of models trained on this dataset in the 5-fold cross-validations in figure 3 compared models trained with dataset A.
Additionally we have evaluated for potential bias in the test dataset against the training datasets. From a further analysis of the imaging with NCC, both datasets A and B vary similarly from dataset C in terms of the CT images. Therefore there is no significant advantage gained by models trained with either dataset A or B due to image similarity with the testing dataset C. We conclude that although the consistency of the gold-standard OAR segmentations in A and B are different, our results on the external testing set are not impacted by an imaging bias.

For 2D tasks, transfer learning from large, pre-trained backbone models, such as ResNet, is often used to improve performance when training data is limited. Unfortunately, analogous backbone models are not yet readily available in 3D. Despite this shortcoming of 3D deep learning, CNN models for auto-segmentation of CT images can be trained with a dataset several orders of magnitude smaller than those used in traditional computer vision tasks, (hundreds (Nikolov et al. 2018, Brunenbeneg et al. 2020, van Dijk et al. 2020) rather than millions (Deng et al. 2009) of images) (Sun et al. 2017). We believe this is possible because radiotherapy planning CT images are highly standardised, both in voxel intensity (brightness) and patient setup. The intensity of voxels are calibrated to Hounsfield units (HU), a scale of tissue density with fixed reference values (air = −1024 HU and water = 0 HU). Consequently, the CNN will ‘see’ different tissue types in highly consistent intensity ranges. In planning CTs, patients are consistently positioned laid on their back using a neck rest with minor rotational differences caused by patient-specific neck flexion. As a result, a CNN model does not need to develop high rotational (or shift) invariance to be able to generalise well to planning CT images.

In this work, we were able to train an accurate, well-generalised CT auto-segmentation model \( \theta_A \) with fewer than 10% of the training samples compared to the 3D HN auto-segmentation models presented in the studies by Nikolov et al. (Nikolov et al. 2018) and van Dijk et al. (van Dijk et al. 2020). We believe that this was because for this specific task, the emphasis for developing an accurate, well-generalised model is shifted from sheer quantity of training samples to the quality and consistency of the gold-standard segmentations. In supervised learning, this involves ensuring the training data domain is complementary to the desired outcome.

Dataset A was contoured according to consensus guidelines defined by Brouwer et al. (Brouwer et al. 2015), whereas the segmentations of dataset B were created created in the course of routine treatment practice and the guidelines followed to delineate the OARs were unreported (Grossberg et al. 2018). Dataset C was contoured according to delineation guidelines published by van de Water et al. (van de Water et al. 2009) which is a predecessor to that used by dataset A but has identical guidelines for the parotid glands. Therefore, dataset A has a higher guideline consistency with dataset C than dataset B. While this is a distinct concept from the observer segmentation consistency, poorly defined (or poor adherence to) contouring guidelines for dataset B may have induced high variability in the manual segmentations, leading to inferior performance in both the internal cross validation and evaluation in an external dataset. We believe that a combination of the lack of guideline consistency and observer segmentation consistency contributed to the considerable improvement in performance of models trained on dataset A over those trained on dataset B. Therefore, we recommend selecting training data with segmentations that closely follow the desired delineation guidelines.

Despite the ‘macroscopic’ standardisation of CT imaging discussed previously, patient scans will differ between treatment centres with different types of immobilisation devices and the use of intravenous contrast which will enhance the imaging contrast of certain structures. If a model is not trained on examples showing these characteristics, it may struggle to generalise to new examples. In this work, we used extensive data augmentation in an attempt to enrich our small dataset and improve the generalisation performance, however, it is difficult to include all possible scenarios with augmentation strategies alone. It would be of interest to measure the impact of these types of variation on the performance of auto-segmentation CNN models, but this is beyond the scope of this work.

In this work, all of the three datasets were preprocessed using a localisation model to crop the CT scans to anatomically consistent sub-volumes. Whilst this mitigates some of the variation between patients and allows the models to be trained on full-resolution scans on a single consumer-grade GPU, it necessitates our proposed models to be used in combination with this localisation model. Fortunately, this is openly-available and adds little time or computational effort to the overall process (Henderson et al. 2022).

In section 2.7.2 we independently followed two common approaches when assessing the generalisation performance of models trained on datasets A and B on the external cohort, dataset C. We first evaluated the best-performing of the five cross-validation models produced in section 2.7.1 for datasets A and B \( \theta_{A4} \) and \( \theta_{B5} \). Additionally, we trained two further models on the entirety of each of datasets A and B \( \theta_A \) and \( \theta_B \). Retraining a model in this way after completing a k-fold cross-validation is not recommended by Bates et al. since the performance estimates generated in the cross-validation are not valid for the newly trained model (Bates et al. 2021). However, we felt it was of interest to compare the performance of the resulting models. For models trained with dataset A it appears that the retrained model slightly outperformed the cross-validation model. This intuitively makes sense for such a small dataset, where any additional data at training time constitutes a large proportional increase. For models trained on dataset B, it was the best cross-validation model that produced
superior segmentations, especially for the spinal cord. This phenomena would be interesting to investigate further.

When testing the generalisation performance of our CNN models trained on datasets A and B, we compared the predicted segmentations in dataset C to the 5 manual observers using the STAPLE-estimated contour as the baseline (or gold-standard) for distance metric calculation. The STAPLE contour was estimated based on the segmentations of the manual observers alone, so there will be a bias when comparing results for the segmentations of the CNN models. The external observers did not delineate the brainstem which shares an adjacent boundary with the spinal cord. The definition of the anatomical boundary between these two OARs is vague (Brouwer et al. 2015). As a result of this, we observed a small bias for the spinal cord predictions of the \( \theta_A \) model to be shorter at the superior end by \( \sim 1-2 \) slices than the STAPLE prediction in favour of the inferior end of the brainstem. We do not adjust our calculations to adjust for this bias, however, in practice, both OARs have a very similar dose constraint and often an identical expanded margin is included around both structures. The spinal cord segmentations frequently extended to the inferior extent of the automatically cropped CT sub-volume, and so we did not observe any large vertical differences between the segmentations at the inferior end.

In computer vision tasks, training data quantity is often considered king, with performance scaling logarithmically with dataset size (Sun et al. 2017). However, it is dangerous to assume that purely increasing the size of a dataset will improve performance and generalisation of models for 3D segmentation of medical images. In our study, we highlight the potential pitfalls of using large datasets with lower quality annotations. Analogous to the process of clinicians following detailed delineation consensus guidelines in an effort to minimise delineation variations (Brouwer et al. 2012, 2015, Vinod et al. 2016), we believe all training gold-standard annotations used in supervised learning of a CNN model should reflect the required delineation standard. Manual observers follow consensus guidelines in an attempt to reduce inter-observer variability. It naturally follows that when training auto-segmentation models, the gold-standard annotations guiding the learning process ought to conform to the consensus guidelines.

We have only experimented with a limited scope of data quality and quantity. This speaks to the nature of the problem in radiotherapy, where large datasets of highly consistent data are scarce. A more thorough follow up study is required with a larger variety of training dataset sizes and quantified levels of segmentation quality. Automatic methods to determine segmentation quality, such as the work of Henderson et al. (Henderson et al. 2022), could be applied and used to gain further insight on the interaction between segmentation quality and training dataset size. In a similar vein, further work is required to confirm the clinical suitability of contours produced by our models.

While this study cannot prescribe guidelines on exactly how much data is enough to train a high performance CNN for HN auto-segmentation, we show that such a question is multi-faceted. Future work could identify how far the training dataset size could be reduced using highly consistent training annotations, while still capturing the necessary clinical variability of patients. However, based on this result, our current practical recommendation to train a high performing model is to focus on acquiring a limited number of segmentations reviewed by multiple expert observers to make them highly conformal to the appropriate guidelines rather than gathering a larger cohort.

5. Conclusion

In this study, we have empirically shown that the consistency of segmentations used for supervision has a greater impact than the size of the dataset used to train 3D segmentation models for radiotherapy. We trained multiple models, using an identical CNN architecture and training procedure, and found models trained using a small dataset with a well-curated set of gold-standard OAR segmentations outperformed those trained with a larger dataset but less consistent supervision annotations, and produces segmentations with a similar accuracy as expert human observers.

Our findings reflect the common notion, ‘garbage in—garbage out’: in a supervised learning regime, the quality of the input data annotations will significantly influence the resulting models performance. Expanding a training dataset with poorly annotated data can severely impact the performance of the model. We wish to further advocate the importance of fully understanding the data used to train auto-segmentation models for medical imaging, particularly for applications in radiotherapy.

Acknowledgments

Marcel van Herk was supported by NIHR Manchester Biomedical Research Centre. This work was also supported by Cancer Research UK via funding to the Cancer Research Manchester Centre [C147/A25254] and by Cancer Research UK RadNet Manchester [C1994/A28701].
Data availability statement

The data cannot be made publicly available upon publication due to legal restrictions preventing unrestricted public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Edward G A Henderson  @  https://orcid.org/0000-0003-3752-4054
Eliana M Vasquez Osorio  @  https://orcid.org/0000-0003-0741-994X
Andrew F Green  @  https://orcid.org/0000-0002-8297-0953

References

Brouwer C L et al 2012 3D variation in delineation of head and neck organs at risk Radiat. Oncol. 7 32
Brouwer C L et al 2015 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. 117 83–90
Clark K et al 2013 The cancer imaging archive (TCIA): maintaining and operating a public information repository J. Digit. Imaging 26 1045–57
Elhalawani H et al 2017 Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges Sci. Data 4 170077
Grossberg A L et al 2018 Data descriptor: imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy Sci. Data 5 180173
Strojan P et al 2017 Treatment of late sequelae after radiotherapy for head and neck cancer Cancer Treat. Rev. 59 79–92


Wu Y and He K 2018 Group Normalization Proc. of the European Conf. on Computer Vision (ECCV) (Cham: Springer) (https://doi.org/10.1007/978-3-030-01261-8_1)