Invited Review article

Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer

Abstract:

Advances in technical radiotherapy have resulted in significant sparing of organs at risk (OARs), reducing radiation-related toxicities for patients with cancer of the head and neck (HNC). Accurate delineation of target volumes (TVs) and OARs is critical for maximising tumour control and minimising radiation toxicities. When performed manually, variability in TV and OAR delineation has been shown to have significant dosimetric impacts for patients on treatment. Auto-segmentation (AS) techniques have shown promise in reducing both inter-practitioner variability and the time taken in TV and OAR delineation in HNC. Ultimately, this may reduce treatment planning and clinical waiting times for patients. Adaptation of radiation treatment for biological or anatomical changes during therapy will also require rapid re-planning; indeed, the time taken for manual delineation currently prevents adaptive radiotherapy from being implemented optimally. We are therefore standing on the doorstep of a transformation of routine radiotherapy planning via the use of artificial intelligence. In this article, we outline the current state-of-the-art for AS for HNC radiotherapy in order to predict how this will rapidly change with the introduction of artificial intelligence. We specifically focus on delineation accuracy and time saving. We argue that, if such technologies are implemented correctly, AS should result in better standardisation of treatment for patients and significantly reduce the time taken to plan radiotherapy.

Keywords: radiotherapy; artificial intelligence; deep learning; segmentation

Introduction:

To parallel the significant advances made in recent decades in technical radiotherapy delivery, we are currently standing on the threshold of a transformation in routine radiotherapy via the use of artificial intelligence. Recent advances in computing power, algorithms and in big data collection and analysis are already resulting in an explosion of new applications in other fields of health, including radiology and ophthalmology. In this article, we outline the current state-of-the-art for auto-segmentation for head and neck cancer radiotherapy in order to predict how this will imminently and rapidly change with the introduction of artificial intelligence.

Intensity-modulated radiotherapy (IMRT) enables the delivery of a conformal radiation dose distribution to target volumes (TVs) in head and neck cancer (HNC). The implementation of IMRT has resulted in significant sparing of organs at risk (OARs) compared to 2D- or 3D-conformal radiotherapy, reducing radiation-related toxicities. Radiation-induced xerostomia is the most commonly reported grade ≥2 late toxicity of HNC radiotherapy (1–3). It can result in difficulties in speech and swallowing, and the development of dental caries (2). Randomised clinical trial data have shown that sparing of the parotid gland with IMRT can reduce the rate of grade 2 xerostomia significantly one year after treatment (3). Swallowing dysfunction is the most common grade ≥3 late toxicity of HNC radiotherapy (4). Dose to pharyngeal constrictor muscles and the supraglottic larynx have been shown to be directly associated with late dysphagia (4–6).

Studies have shown that tumour control and radiotherapy toxicities are highly correlated with the accuracy of TV and OAR delineation (7,8). Steep dose gradients can occur outside the planning target volume (PTV), and structures that are not specifically delineated in IMRT as avoidance structures can receive significant absorbed radiation doses (9).

Accurate delineation of OARs is clearly important. However, delineation of TVs and OARs in HNC is a time-consuming and labour-intensive process that is typically undertaken by a clinical/radiation oncologist and/or radiographer/dosimetrist. According to published evidence, it takes an average of 2.7 to 3 hours to delineate a full set of TVs and OARs in a HNC patient (10–21). This human resource commitment is increasingly difficult to manage in health care systems on account of increasing demands for radiotherapy and shortages of adequately trained staff (22,23).

Moreover, the evolution of radiation therapy requires treatment to be adapted anatomically or biologically due to changes in the patient or the tumour during a course of therapy, termed adaptive radiotherapy. This requires rapid replanning and dosimetry, not only between fractions but ultimately during a fraction of radiotherapy in real-time. Repeated offline planning, which is the current standard, requires recontouring of the TVs and OARs followed by replanning of the optimal dosimetry. The time currently required to delineate OARs and TVs is a barrier to adaptive radiotherapy and therefore a barrier to the field of radiation oncology moving forward.

A major issue for quality assurance in radiotherapy is variation in outlining between practitioners and between centres. Manual delineation of TVs and OARs in HNC has also been shown to result in significant inter- and intra-practitioner variability; indeed, these differences may exceed planning and setup errors (10,24). This variation is not theoretical only: It has been shown that inter-practitioner variability in delineating OARs results in significant dosimetric impact for the patient (25).

Auto-segmentation (AS) software may aid the planning process and at least partially resolve some of these issues. AS has the potential to be time-saving, to reduce inter- and intra-practitioner variability and to permit online adaptive radiotherapy planning during a course of treatment. As shown in Figure 1, atlas-based auto-segmentation (ABAS) can be used, including hybrid auto-segmentation (HAS) techniques, or deep learning models (26). In this paper, we review the published literature on the use of ABAS, HAS and MBAS in the delineation of TVs and OARs in HNC, with a specific focus on delineation accuracy and time saving.

Methodology:

The search strategy for identification of studies was designed to ensure that the maximum number of studies in AS planning were included. Five databases were searched: EMBASE, PubMed, Science Direct, Google Scholar and arXiv. Original articles and reviews published in English between January 2005 and December 2018 were included. On account of the evolution of AS technology with reference to health, articles referring to AS prior to 2005 were not included. The following search terms were used:

1. Auto-segmentation AND (contouring OR organ-at-risk OR critical organs OR critical structure OR target OR head and neck patients);
2. Same search terms limited to search dates.

This initial search identified 569 abstracts. The title and abstract for each resulting citation were screened for relevance to this work and to avoid duplicated results. This resulted in the removal of 505 abstracts that did not present data on the evaluation of AS in HNC OAR or target volume delineation. Full articles were obtained on the remainder via University College London journal access. Sixty-two potentially eligible studies were assessed for relevance, quality and content by RS (radiation oncologist with 20 years experience) and MK (radiation oncologist with 7 years experience). Further searches and cross-checks of the reference lists of these articles were not performed. Data extraction and evaluation were conducted on the 33 articles assessed to be of direct relevance to this article.

Participants of eligible studies were aged 18 years or more with histologically proven HNC, with intent to treat with IMRT. All HNC sites and histological types were eligible if they used CT planning. Studies using conventional two-dimensional (2D) planning, 3D conformal radiotherapy (CRT) and paediatric HNC patients were excluded.

Due to the lack of published data, clinical outcomes were not studied. For this review, the outcomes of time saving and delineation accuracy were explored primarily. Additionally, if recorded, the influence of other external factors (observer variability, AS strategies adopted and stage of disease) was also noted.

For all studies, the gold standard (or “ground truth”) manual delineation (MD) volumes were compared with the AS volumes. The techniques used to perform these comparisons varied but included the Dice similarity coefficient (DSC), mean surface distance (MSD), Hausdorff surface distance (HSD), sensitivity and specificity analyses and receiver operator characteristics (ROC) curves. Further details on these techniques can be found in the Supplementary Material section.

Results

Atlas-Based Auto-Segmentation (ABAS)

ABAS is a technique that propagates volumes from an atlas onto a patient image dataset via deformable image registration. Single-ABAS techniques make use of a single dataset of pre-defined elective nodal TVs and OARs (gold standard segmentations). The combination of data from multiple atlases (multi-ABAS) can be performed to reduce the risk of significant variability in anatomy between the atlas and patient datasets. These multiple separate auto-segmentations can then be combined to form a population-based average atlas. Alternatively, a simultaneous truth and performance level estimation (STAPLE) or similarity and truth estimation for propagated segmentations (STEPS) algorithm can be used to fuse the data from multiple atlases simultaneously. These algorithms have been programmed to fuse the multiple atlases in different ways. Using the STAPLE algorithm, all atlases carry the same weight, whereas with STEPS, only the top-ranked atlases are used during the fusion process, thereby discarding the atlases with the least anatomical similarity to the patient (27).

OAR and TV delineation

A large number of the identified studies evaluated the geometric accuracy of ABAS techniques for OAR and TV delineation, as shown in Table 1. The metrics used for these evaluations and the range of performance reported in these studies are outlined in Table 2.

Hoang Duc et al undertook a comparison of the multi-ABAS techniques using the STAPLE and STEPS algorithm for the delineation of OARs in 100 HNC cases (16). STEPS outperformed STAPLE for large OAR structures such as the brainstem, spinal cord and parotid glands, but not for the smaller OARs of the visual pathway, including globes and optic chiasm. Several other studies have used the STAPLE algorithm to compare single- and multi-ABAS techniques. Teguh et al. showed that multi-ABAS consistently outperformed single-ABAS in the delineation of lymph node levels and OARs, with higher DSC and smaller MSD values in both N0 and N+ patients (18). Yang et al used the STAPLE algorithm to evaluate delineation of low-risk elective clinical target volumes (CTVs) in cases of unilateral tonsillar cancers (stage T1-4a, N0-2b). Overall, multi-ABAS performed at least comparably, if not superiorly, to the single-best atlas method, but without the need to identify the best atlas (15). A similar comparison of single-ABAS and multi-ABAS was undertaken by Hoogeman et al. Manual delineation of nodal levels I-V and OARs in ten HNC patients were used to construct the atlas. Multi-ABAS performed better than single-ABAS in nodal levels (DSC 0.65 vs 0.62) and OARs (0.50-0.78 vs 0.32-0.71) (28). Stapleford et al performed a study using STAPLE for the delineation of bilateral nodal CTVs in five HNC cases. Five clinicians performed manual CTV delineation, and the STAPLE algorithm was used to combine these volumes into a reference standard. Compared to the STAPLE-defined reference volumes, the DSC of AS volumes compared well with that of MD (DSC 0.76 versus 0.79 respectively) (29). A further study of the STAPLE algorithm in the delineation of the parotid glands alone showed similar results (30).

Levendag et al performed a study of multi-ABAS in ten HNC cases in which the atlas for each case was created from the other nine MD contours using a ‘leave-one-out’ method. DSC was 0.6 for unedited AS contours, which improved to 0.7 after manual editing (20). A similar study was performed by Han et al, in which DSC values ranged from 0.65 for the level III nodes, to 0.90 for the mandible for unedited AS contours (31).

Several studies have evaluated the performance of commercially available multi-ABAS software programmes. Gresswell et al performed a study using MIM-Maestro version 6.5.8 (PLACE). Patients were added sequentially into an in-house ABAS. They showed that the similarity of the in-house ABAS to the reference contours (defined by DSC and mean distance to agreement) increased as more patients were added to the atlas. This was offset by a slight increase in time to auto-populate the structures (an average of 23 seconds with each added patient) (32). Sims et al undertook a study of an ABAS system which uses an expectation maximisation algorithm to determine the mean OAR contour from 45-patient atlas library (ISOgray system by DOSIsoft, version 3.1). ABAS contours performed worse than edited ABAS contours, with systematic errors of this ABAS system identified including the over-estimation of the size of the parotid gland (19). La Macchia et al performed a study comparing the performance of three commercially available ABAS systems (ABAS 2.0, CMS-Elekta, Stockholm, Sweden; MIM 5.1.1, MIMVista Corp, Cleveland, Ohio; VelocityAI 2.6.2, Velocity Medical Systems, Atlanta, Georgia). Although differences in performance were seen between systems, the absolute values of such differences were modest (34).

On-gantry cone-beam CT (CBCT) is the most commonly used modality for image-guided RT (IGRT), but signal/noise ratio is significantly lower than with regular fan-beam CT images. This can limit the ability of ABAS techniques to delineate structures including target volumes (21). Zhang et al studied automated segmentation for online adaptive RT by using the patient’s own planning CT image and contours as the atlas for single-ABAS. They obtained DSC values of approximately 0.8 and mean surface distances of 1 mm (SD <2 mm) for most OARs (21). The concept of using the patient’s own CT dataset as an atlas for ABAS has also been investigated in patients receiving radiotherapy for lung cancers using 4DCT. Laub et al. used manual delineation of lung tumour target volumes and OARs on the 0% respiratory phase CT dataset as the reference atlas for propagation to the other nine phases of the 4DCT dataset (35).

Liu et al performed a study of a multi-ABAS technique in an adaptive radiotherapy pathway. They studied the performance of the Advanced Medical Imaging Registration Engine (ADMIRE) v1.05 (Elekta Software) on three separate CT datasets: pre-treatment planning CT, in-treatment planning CT, and a cone-beam CT (CBCT). Auto-contours generated by ADMIRE were compared with manual contours on the pre-treatment CT to evaluate their accuracy. Similar registration accuracy was achieved for intra-patient CT-to-CBCT deformable registration compared to intra-patient CT-to-CT deformable registration (36).

Irrespective of the method of fusion in multi-ABAS, its reliability remains dependent on the similarity between the underlying atlases and the patient. Large deformations of anatomy caused by HNC are difficult to correct for with registration algorithms.

Time saving

One of the main potential benefits of AS is in the reduction of time spent in the manual delineation of OARs and TVs. Several of the ABAS studies above also measured manual interaction time, with the aim to evaluate whether manual editing of AS contours was a more time-efficient process than full manual delineation.

In the study performed by Hoang Duc et al., there was a reduction in manual interaction time of 61% with manual editing of STEPS contours in comparison with fully manual OAR delineation (16). Teguh et al. produced ABAS contours in 7 minutes, but these required an additional 66 mins of editing (of which 31 minutes was for the nodal levels). Nevertheless, this compared favourably with MD contours, which took 180 minutes (18). Stapleford et al investigated single-ABAS for nodal CTV delineation. Physicians saved on average 11.5 minutes per patient by editing AS contours, equating to a 35% reduction in delineation time (29). Similar reductions were found by Daisne et al in both N0 and N+ cases (24). In the study performed by Levendag et al, manual delineation of TVs and OARs took 180 minutes versus 53 minutes for manual editing of ABAS contours (20).

The VelocityAI system studied by Sjöberg et al demonstrated a clear benefit in terms of time saving, with segmentation time reducing from 42.3 minutes for MD to 21.4 minutes for edited AS contours. However, not a single segmented structure was approved without editing (37). The comparison of three different commercially available ABAS systems performed by La Macchia et al demonstrated a range of time saving performance (69-113 minutes versus 163 minutes for MD) (34). In contrast, Ingle et al showed that the time saved amending AS OAR contours produced by the BrainLab iPlan® tool (25.2 minutes) was offset by the additional time taken to manually edit AS nodal volumes (24.9 minutes) (38).

Dosimetric impact

As detailed above, there is a range of performance observed with different ABAS systems in the geometric accuracy of delineation of different OARs and CTVs in HNC patients. However, it is important to understand the dosimetric impact of these geometric inaccuracies. In an international study of inter-clinician OAR delineation variability and its dosimetric impact, 32 different centres delineated OARs on a single CT dataset. Significant variations in dose were found for parotid glands, brainstem and spinal cord. This translated into differences in D_{mean} of parotids of up to 50% and into more than 20% in brainstem D_{max} (25). Voet et al performed a dosimetric evaluation comparing ABAS and edited ABAS contours performed by two independent observers. Clinically acceptable IMRT plans were constructed using the ABAS (unedited) plans. With regards to PTV coverage, the plans were evaluated for V_{95} – volume receiving 95% of prescribed dose; and D_{99} – near minimum dose in PTV. The edited volumes were larger by a mean of 8.7%. These edits led to a reduction in V_{95} of 7.2% ± 5.4% (1SD), and a mean reduction in D_{99} of 14.2 Gy. Even for DSC >0.8 and mean contour distances <1 mm, reductions in D_{99} of up to 11 Gy were observed (25). Similar findings have been reported by Tsuji et al, who evaluated the dosimetric impact of AS for adaptive IMRT. Evaluation of auto-contoured structures showed a significantly lower mean dose coverage of the manually delineated gross tumour volume (GTV) and CTV (26).

It is clear from these studies that, even when geometric differences between ABAS and edited contours are small, there is nevertheless a significant impact of adjusting the contours on the final clinically-relevant radiation dose distribution to target volumes. Indices of geometric similarity are of limited value in predicting TV coverage and plan quality.

Interobserver variability

The study by Sims et al included ground truth manual delineations performed by two clinicians. A comparison was made of the coefficient of variation (CV) between the two clinicians' MD contours and their edited AS contours. Lower CV values were obtained for all investigated OARs with edited AS contours, most notably in delineation of the brainstem

(CV 0.12 versus 0.46), illustrating the potential benefits of AS at reducing interobserver variability (19).

In Stapleford et al, five radiation oncologists manually delineated bilateral neck CTVs. An AS contour set was generated and then manually edited by the physicians to make them acceptable for planning. The STAPLE algorithm was used to combine the collections of contours, and the overlap of individual MD or edited AS contours with the STAPLE contours were analysed. Overall, the average DSC was higher for the edited-AS group than the MD group (0.89 versus 0.79), and MSD lower (1.8mm versus 2.8mm), indicating reduced interobserver variability with edited AS contours. The most dramatic differences were seen in delineated CTV volume, with a reduction in percentage false positive volume from 9% to 3% in the edited AS group (29).

However not all studies have shown a clear reduction in interobserver variability with the use of AS techniques. The study performed by Ingle et al using the BrainLab iPlan® tool compared the performance of two clinicians in MD and edited AS delineation of nodal CTVs and OARs in five HNC patients (see Table 1 for details). There was no difference found in interobserver variability in the delineation of nodal CTV in the two groups, with mean percentage difference between the volumes delineated by the two clinicians being 4.26% with MD versus 4.22% with edited AS contours (37).

<u>Hybrid Auto-Segmentation (HAS)</u>

Model-based approaches can compensate somewhat for the lack of reliable image information (low soft tissue contrast, artefacts, insufficient image content) by imposing prior shape constraints in the segmentation process. This is typically done by statistical analysis of reference ground truth (gold standard) segmentations. Combining the advantages of local low-level features and global high level prior shape information in this way is a potentially beneficial technique for achieving a more reliable and robust AS.

For model-based approaches, deformable models of anatomical structures are often represented by flexible triangulated meshes, where the shape is designed to be close to the average shape of the structure in question. It can also possibly cover variations in shape by using techniques such as principle component analysis. In addition to shape, knowledge about the characteristic grey-value range, gradient direction, and strength of the region of interest can be incorporated into the model. In practice, this is often performed by manual drag-and-drop techniques which may require manual editing by using special mesh manipulation tools (27).

In order to work in a fully automated manner, model-based approaches can be combined with ABAS, where the patient dataset is registered with a reference image (single ABAS), or an averaged population containing some ground truth segmentations. Such hybrid approaches combine image registration and segmentation into a common framework where evolution of deformable models can act as a registration constraint, or be used to compensate for residual differences after the registration step.

OAR and TV delineation

The only prospective randomised double-blind in silico study of AS has been performed by Walker et al (8). Their study aimed to evaluate acceptability of individual OAR contours produced by HAS alone, versus HAS with manual editing. They studied a smart probabilistic image contouring engine (SPICE) HAS algorithm, which performs an initial registration, then a dense deformable registration, and then probabilistic (model-based) refinement. There was no statistically significant difference in the edited AS and MD contours, but the unedited HAS contours were significantly different for all OARs other than spinal cord and mandible. The authors concluded that HAS is a promising tool for workflow efficiency improvement, but human oversight remains critical for patient safety.

The SPICE algorithm was also evaluated by Thomson et al (39). OARs were manually delineated by five clinicians in ten cases. The STAPLE algorithm was used to create a reference standard from all five separate MD structure sets. The SPICE contours, modified SPICE contours and separate MD contours were then compared with the reference STAPLE contours. Without editing, DSC values were significantly lower for SPICE than MD in all OARs. Best performance of SPICE was seen in the parotid and submandibular glands with DSC 0.79 and 0.80 respectively. Manual modification of SPICE contours was still inferior to MD contours, with significantly lower mean DSC values (39). Comparable performance was found by Zhu et al in an evaluation of SPICE for OAR delineation in 32 HNC patients. Performance ranged from DSC 0.70 in submandibular glands to 0.96 for brain (40).

The advantages of adding intensity modelling to multi-ABAS was evaluated by Fortunati et al (41). They investigated this HAS technique in 18 patients receiving hyperthermia and radiotherapy treatment for HNC. The addition of intensity modelling to produce HAS contours outperformed multi-ABAS in all studied OARs other than brainstem and spinal cord. However, the addition of intensity modelling did not consistently improve maximum surface error (41). Therefore, larger errors made by AS are caused by inaccurate spatial prior models and this cannot be solved by adding intensity modelling.

An alternative HAS approach was investigated by Qazi et al (27). Atlas contours were used to guide a deformable registration algorithm. These models were then automatically fine-adjusted by a boundary refinement approach using a probabilistic mask. Segmentation was started at the global level (ABAS), and then refined down to voxel-level classification. Combining the advantages of local low-level features and global high level prior shape information is a potentially beneficial technique for achieving a more reliable and robust AS. DSC values for the OARs ranged from 0.93 (mandible) to 0.83 (parotid and submandibular glands) (27). The authors used a single, randomly selected atlas, so their technique could potentially be improved further by incorporating atlas selection, or combination of multiple atlases using the STAPLE algorithm.

The benefits of the addition of a model based approach to ABAS has also been investigated by Fritscher et al. They undertook a study of HAS in which they combined multi-ABAS with geodesic active contours (GAC) and statistical appearance models (SAM) for the delineation of the brainstem and parotid glands in HNC. The addition of the model approach showed statistically significant improvement when compared with the multi-ABAS technique alone.

Results for delineation of the brainstem with the HAS approach versus multi-ABAS alone were: DSC 0.87 vs 0.85; MSD 1.1 mm vs 1.3 mm; HSD 8 mm vs 10 mm (42).

Time saving

The potential benefits of HAS techniques for time saving in the delineation of HNC volumes have been evaluated by two studies with contrasting results. Along with evaluating the acceptability of contours created by SPICE, the second primary aim of the study by Walker et al was to determine the workflow feasibility and time saving of SPICE. A 30.9% time reduction was found comparing edited HAS contours with MD (19.7 vs 28.5 mins) (8). In contrast, the evaluation of SPICE performed by Thomson et al showed no observed time saving. The average MD time for delineation of OARs was 14.0 mins, compared with 16.2 mins for modification of SPICE contours (39).

Interobserver variability

The HAS study performed by Fortunati et al did not involve manual editing of AS contours, but instead compared the unedited performance of multi-ABAS, multi-ABAS with intensity modelling (HAS), and MD performed by three observers. Their HAS technique showed better robustness to variations in atlas labelling compared with multi-ABAS alone. Furthermore, the addition of intensity modelling improved the segmentation reproducibility compared with human observer’s segmentations (40).

Thomson et al compared SPICE contours, with and without manual editing, with MD delineation performed by five observers to assess interobserver and inter-technique variability. Editing of SPICE contours did reduce variability, but there was still a higher degree of variability seen between SPICE and MD contours than between the two most discordant manually delineated contours for all OARs studied.

No studies of the dosimetric impact of HAS were identified.

Deep learning-based algorithms

The development of deep learning-based algorithms for radiotherapy planning in HNC has focused primarily on geometric accuracy of target volume and OAR delineation. No studies were identified in our literature search that evaluated these techniques for time saving, dosimetric impact or interobserver variability in HNC. Our search did reveal one study (43) that evaluated the impact of deep learning-based algorithms on TV and OAR dosimetry, but this study was performed in brain tumour patients, and is discussed below.

OAR and TV delineation

Ibragimov et al (44) performed the first evaluation of a deep learning-based algorithm for segmentation of OARs in HNC CT images. They used convolutional neural networks (CNN) that were trained using ground truth contours in reference CT images. Deep learning techniques such as CNNs have demonstrated impressive performance in computer vision and medical image analysis applications (45). CNNs first study the appearance of regions of interest in a training set of segmented images. CNNs take input images and pass them through a sequence of learning functions or layers that extract and recognise consistent intensity patterns and make a pixel-wise prediction according to these patterns. CNNs can take spatial information and relationships into account by analysing neighbouring pixels together. In this study, a Markov random fields algorithm was then used to smooth the obtained classification results. Cavities in the volume were removed using dilate-erode operations. The resulting MBAS OAR volumes in 50 cases were evaluated for accuracy of delineation of a number of OARs (Table 1). A wide variation in the geometric accuracy of the technique was found, ranging from 0.90 for the mandible and 0.87 for the spinal cord, to 0.34 for the optic chiasm. More challenging is the presence of metallic dental restorations, which can hamper identification of the borders of the mandible, but also corrupt the appearances of surrounding structures such as parotids, tongue, submandibular glands. CNNs were able to correctly model the composition of low and high intensity voxel groups that characterise dental artefacts, an ability that is beyond ABAS techniques. CNNs performed less well than existing algorithms in delineating the submandibular glands (DSC 0.71). This is due to lack of distinguishable intensity features, and the fact that ABAS deformation is relatively restricted and will roughly identify the position, whereas CNN relies purely on image intensities around glands (44). As with ABAS techniques, CNNs rely considerably on the quality and representativeness of the training dataset.

Fritscher et al. (46) presented an approach in three steps. First, they pre-align the images by using an affine transformation that register them with respect to an available reference atlas. Second, they extract 2D orthogonal patches in the sagittal, coronal, and axial plane, for each structure of interest. Finally, they apply their CNN. This CNN contains three pathways, one for each plane, which learn their specific low-level features, followed by a common sequence of fully connected layers. Segmentation is carried out by applying this network at different locations, using input patch sizes of 31 x 31 pixel planes. Experiments have been made to segment parotid gland, submandibular gland, and optic chiasm, obtaining a DSC of 0.81, 0.65, and 0.52 respectively.

Močnik et al (47) proposed a bimodal method that segments parotid glands using both CT and MR images, based on the fact that these glands have better visibility in MR scans. In order to combine these two modalities, image registration is done as a first step, transforming the MR image according to the reference CT image. The resultant two aligned 3D images are the input of a CNN. The CNN used here presents several resemblances to that of Fritscher et al (46). The results showed that the proposed approach combining both CT and MR modalities obtained a DSC of 0.79, whereas using CT alone obtained a DSC of 0.77.

Ren et al (48) proposed another CNN-based approach for the segmentation of the chiasm and the left and right optic nerves from CT images. Similar to previous studies, they performed image registration as a pre-processing, and used a CNN-based network to classify the central voxel of the patch provided as an input. However, their proposed approach contains some novelty. Firstly, the network takes as input patches of different scales centred at the same voxel. Secondly, the overall method is composed of a sequence of interconnected CNNs that keep refining the predictions made by previous CNNs. The authors use two different networks, one for segmenting chiasm, and another for segmenting both left and right optic nerves. The results using four iterations of this scheme obtain the best results, achieving a DSC of 0.56 for chiasm, and 0.72 and 0.70 for the left and right optic nerves.

Men et al performed a study of the performance of an end-to-end deep deconvolutional neural network (DDNN) for segmentation of gross tumour volumes and clinical target volumes in cases of nasopharyngeal carcinoma. 184 patients were chosen as a training set to adjust the parameters of DDNN, and the technique was tested on 46 patients. DDNN achieved mean DSC values ranging from 0.62 for metastatic nodal GTV to 0.83 for CTV. The authors attributed the relatively poor performance of DDNN in the metastatic nodes to the considerable difference in shape, volume and location between patients, and they expect DDNN performance to improve with further expansion of the training dataset and combination of MR images (49).

It should be noted that investigators of brain tumour segmentation have included some OARs of relevance to HNC. Agn et al (43) propose a generative model which leverages prior information about the anatomy and the imaging process. The approach is composed of two stages. The first one models the likelihood of the input data given the segmentation labels by means of Gaussian mixture models. The second part models the segmentation prior, which includes prior knowledge constraints. The authors explore two ways of modelling this. They test this approach on 3 datasets, including 2015 BRATS (50), obtaining comparable results to CNN-based models. The authors also performed a dosimetric evaluation that showed some significant differences in dosimetry between the AS and MD volumes for some but not all of the OARs and TVs evaluated. Overall, significant differences to treatments could be expected if automatic segmentation were to be used when optimising the radiotherapy dose plan. Due to the limited size of the datasets available, it is currently unclear how well their approach would transfer to a larger training set in comparison to CNNs.

Two recent studies are worthy of particular mention. In the AnatomyNet, an end-to-end, atlas-free, three dimensional squeeze-and-excitation U-Net (3D SE U-Net), fast and fully automated whole-volume HNC anatomical segmentation for 9 structures was achieved from a training set of 261 HNC CT images. AnatomyNet takes only 0.12 seconds to segment all 9 organs. Compared to previous AS methods, this approach improved DSC by 3.3% on average (51).

The most exciting landmark in the field is the recent demonstration that 3D U-Net architecture deep learning can achieve performance metrics similar to experts in delineating a wide range of HNC OARs. The model was trained on a dataset of 689 CT scans acquired in routine clinical practice. Its generalisability has been suggested by applying the model to 24 CT scans available from the Cancer Imaging Archive collected at multiple international sites previously unseen to the model (52). This application of deep learning appears to hold significant potential as a major step forward compared to the limitations of ABAS and HAS methods. It is a high priority for application to other datasets, including clinical evaluation and dosimetry as validation for clinical scenarios.

Conclusions:

There is no doubt that automatic image segmentation will play a critical role in the future of clinical radiotherapy planning, particularly to facilitate the implementation of adaptive radiotherapy techniques into routine clinical practice. Ultimately, for intra-fraction adaptive radiotherapy to be optimal, auto-segmentation of critical organs for any part of the body should only take seconds and should not require significant editing by experts. This has to be the goal for the investigators making progress in this field of study.

At least eleven auto-contouring software solutions are currently available commercially (53), with varying claims on their potential for lowering the segmentation time and reducing inter-observer variability. The studies reviewed here report varying degrees of success with some structures being clinically acceptable while others require considerable manual editing. In general, AS performs well for the delineation of the brainstem, spinal cord and parotid glands. It shows moderate levels of performance for the optic apparatus, cochleae and elective nodal groups. It performs poorly for volumetric delineation of tumour gross target volumes as well as in situations of abnormal anatomy, such as in the post-operative setting. Deep learning approaches appear to hold the greatest potential for addressing these limitations.

There are a number of limitations to the ABAS approach (54). The deformable image registration is never perfect, the contrast in the images between boundaries of organs may be indistinct leading to errors in the deformation, and the image registration algorithm might not sufficiently account for anatomical deformation. Multiple atlases can be employed, but these algorithms make their decision based on the set of contours alone, without reference to the underlying image data. If there is a random error between the deformed atlases at a point on the surface of an organ to be segmented, then the greater the likelihood of a systematic error. Unfortunately, having more atlases does not always mean better segmentation. Overall, ABAS accuracy is highly dependent on the similarity of the atlas and the underlying patient and inaccurate delineation may result in time-consuming manual post-processing or treatment error.

As new technologies are tested, it is important to use the best metrics. A description of metrics for comparing contours has been reviewed previously (55-57). The Dice similarity coefficient (DSC) is used widely but is very difficult to interpret because its value is dependent on the volume being compared and it gives no indication of the distance between the two contours. The normalised dice similarity coefficient (nDSC) simultaneously removes the dependence on volume and also attaches a clinical significance to the discrepancy. The mean distance to agreement gives a better indication of the magnitude of the error in terms of distance (52), but it can also be misleading in terms of clinical impact. Valentini and colleagues (58) recommend the development of a combination of conformation scores, metric elements and clinical risk assessment into a new class of indices that would provide a more robust tool for the evaluation of a test contour against a reference contour. They also emphasise the necessity of building up a reliable ‘‘gold structure set’’ which will represent the unique benchmark of the study and will be the referral contour to which all other contours should be compared with. Gold standard structure sets in HNC have been published by Brouwer et al (59) with regard to OARs, and Gregoire et al (17) with regard to the delineation of lymph node levels and related CTVs in the node-negative neck.

In addition to time saving and improved workflow, application of the new technologies should also perform comparisons between multiple operators, or between manual delineations and auto-contouring, to show improved standardisation and reduced error between operators and between centres. Discrepant contouring practices between radiation oncologists within an institution and between institutions is a current limitation to the field of radiation oncology.

In conclusion, although we are currently standing on a very exciting threshold of a transformation in routine radiotherapy via the use of artificial intelligence, implementation of time-saving AS needs to be optimal. AS should facilitate the adoption of international consensus guidelines across centres to result in more favourable and more standardised routine clinical practice. These advances represent a real opportunity to improve outcomes for patients, akin to the application of quality assurance in clinical trials that has improved clinical outcomes for patients taking part in radiotherapy trials (60).

Acknowledgements: CN is supported by the RMH-ICR Biomedical Research Centre. RAS is supported by the NIHR University College London Hospitals Biomedical Research Centre, the Cancer Research UK University College London Experimental Cancer Medicine Centre, CRUK-UCL Centre Award [C416/A25145], the Cancer Research UK Centres Network Accelerator Award Grant (A21993) to the ART-NET Consortium,.

References

1. Wijers O, Levendag P, Braaksma M, Boonzaaijer M, Visch L, Schmitz P. Patients with head and neck cancer cured by radiation therapy : a survey of the dry mouth syndrome in long-term survivors. Head Neck. 2002;24(8):737–47.
2. Guchelaar H, Vermes A, Meerwaldt J. Radiation- - induced xerostomia : pathophysiology , clinical course and supportive treatment . Support Care Cancer. 1997;5(4):281–8.
3. Nutting CM, Morden JP, Harrington KJ, Urbano TG, Bhide SA, Clark C, 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;12:127–36.
4. Jensen K, Lambertsen K, Grau C. Late swallowing dysfunction and dysphagia after radiotherapy for pharynx cancer : Frequency, intensity and correlation with dose and volume parameters. Radiother Oncol. 2007;85:74–82.
5. Dirix P, Abbeel S, Vanstraelen B, Hermans R, Nuyts S. Dsyphagia after chemoradiotherapy for head-and-neck squamous cell carcinoma: dose-effect relationships for the swallowing structures. Int J Radiat Oncol Biol Phys. 2009;75(2):385–92.
6. Caudell J, Schaner P, Desmond R, Meredith R, Spencer S, Bonner J. Dosimetric factors associated with long-term dysphagia after definitive radiotherapy for squamous cell carcinoma of the head and neck. Int J Radiat Oncol Biol Phys. 2010;76(2):403–9.
7. Mukesh M, Benson R, Jena R, Hoole A, Roques T, Scrase C, et al. Interobserver variation in clinical target volume and organs at risk segmentation in post-parotidectomy radiotherapy: can segmentation protocols help? Br J Radiol. 2012;85(August):e530-6.
8. Walker G, Awan M, Tao R, Koay E, Boehling N, Grant J, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. Radiother Oncol. 2015;112(3):321–5.

9. Prescribing, recording, and reporting photon-beam intensity modulated radiotherapy (IMRT): International Commission on Radiation Units and Measurements. ICRU Report 83. 2010.

10. Hong T, Tome W, Chappell R, Harari P. Variations in Target Delineation for Head and Neck IMRT: A International Multi-institutional Study. Int J Radiat Oncol Biol Phys. 2004;60(1):S157-158.

11. Harari P, Song S, Tome W. Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer. Int J Radiat Oncol Biol Phys. 2010;77(3):950–8.

12. Harari P, Jaradat H, Connor N, Tome W. Refining target coverage and normal tissue avoidance with helical tomotherapy vs linac-based IMRT for oropharyngeal cancer. IInt J Radiat Oncol Biol Phys. 2004;60(1):S160.

13. Nelms B, Tome W, Robinson G, Wheeler J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. Int J Radiat Oncol Biol Phys. 2012;82(1):368–78.

14. Lim JY, Leech M. Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. Acta Oncol (Madr). 2016;55(7):799–806.

15. Yang J, Beadle BM, Garden AS, Gunn B, Rosenthal D, Ang K, et al. Auto-segmentation of low-risk clinical target volume for head and neck radiation therapy. Pract Radiat Oncol. 2014;4(1):e31–7.

16. Hoang Duc AK, Eminowicz G, Mendes R, Wong S-L, Mcclelland J, Modat M, et al. Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. Med Phys. 2015;42(9):5027–34.

17. Gregoire V, Levendag P, Ang KK, Bernier J, Braaksma M, Budach V, et al. CT-based delineation of lymph node levels and related CTVs in the node-negative neck : DAHANCA , EORTC , GORTEC , NCIC , RTOG consensus guidelines. Radiother Oncol. 2003;69:227–36.

18. Teguh D, Levendag P, Voet P, Al-Mamgani A, Han X, Wolf T, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and

normal tissue (swallowing/mastication) structures in the head and neck. Int J Radiat Oncol Biol Phys. 2011;81(4):950–7.

19. Sims R, Isambert A, Grégoire V, Bidault F, Fresco L, Sage J, et al. Automatic segmentation A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. Radiother Oncol. 2009;93(3):474–8.

20. Levendag P, Hoogeman M, Teguh D, Wolf T, Hibbard L, Wijers O, et al. Atlas Based Auto-segmentation of CT Images: Clinical Evaluation of using Auto-contouring in High-dose, High-precision Radiotherapy of Cancer in the Head and Neck. Int J Radiat Oncol Biol Phys. 2017;40:S401.

21. Zhang T, Chi Y, Meldolesi E, Yan D. Automatic delineation of on-line head-and-neck computed tomography images: towards on-line adapative radiotherapy. Int J Radiat Oncol Biol Phys. 2007;68(2):522–30.

22. Round CE, Williams M V, Mee T, Kirkby NF, Cooper T, Hoskin P, et al. Radiotherapy Demand and Activity in England 2006 - 2020. Clin Oncol [Internet]. Elsevier Ltd; 2013;25(9):522–30. Available from: http://dx.doi.org/10.1016/j.clon.2013.05.005

23. Rosenblatt E, Zubizarreta E. International Atomic Energy Agency: Radiotherapy in Cancer Care: Facing the Global Challenge [Internet]. 2017. Available from: https://www-pub.iaea.org/MTCD/Publications/PDF/P1638_web.pdf

24. Daisne J, Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes : a clinical validation. Radiat Oncol. 2013;8:154.

25. Voet PWJ, Dirkx MLP, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJM. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage ? A dosimetric analysis. Radiother Oncol. 2011;98(3):373–7.

26. Tsuji S, Hwang A, Weinberg V, Yom S, Quivey J, Xia P. Dosimetric evaluation of automatic segmentation for adapative IMRT for head-and-neck cancer. Int J Radiat Oncol Biol Phys. 2010;77(3):707–14.

27. Qazi AA, Pekar V, Kim J, Xie J, Breen SL, Jaffray DA. Auto-segmentation of normal and target structures in head and neck CT images : A feature-driven model-based approach. Med Phys. 2011;38(11):6160–70.

28. Hoogeman MS, Han X, Teguh DN, Voet P, Nowak P, Wolf T, et al. Atlas-based Auto-segmentation of CT Images in Head and Neck Cancer: What is the Best Approach? Int J Radiat Oncol Biol Phys. 2008;72(1):S591.

29. Stapleford LJ, Lawson JD, Perkins C, et al. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. Int J Radiat Oncol Biol Phys. 2010;77(3): 959-966.

30. Han X, Hibbard LS, Connell NPO, Willcut V. Automatic Segmentation of Parotids in Head and Neck CT Images using Multi-atlas Fusion. Medical Image Analysis for the Clinic: A Grand Challenge. 2011. p. 297–304.

31. Han X, Hoogeman MS, Levendag PC, Hibbard LS, Teguh DN, Voet P, et al. Atlas-Based Auto-segmentation of Head and Neck CT Images. International Conference on Medical Image Computing and Computer-assisted Intervention. 2008. p. 434–41.

32. Gresswell S, Renz P, Werts D, and Arshoun Y. Impact of increasing atlas size on accuracy of an atlas-basde auto-segmentation program (ABAS) for organs-at-risk (OARS) in head and neck (H&N) cancer patients. Int J Radiat Oncol Biol Phys. 2017;98(2):SE31.

33. Tao C, Yi J, Chen N, Ren W, Cheng J, Tung S, et al. Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma : A multi-institution clinical study. Radiother Oncol. 2015;115(3):407–11.

34. La Macchia M, Fellin F, Amichetti M, Cianchetti M, Gianolini S, Paola V, et al. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck , prostate and pleural cancer. Radiat Oncol. 2012;7:160.

35. Laub W, Kalpathy-Cramer J, Fuss M. A Comparison between Atlas Based Auto-Segmentation and Manual Contouring in 4D-Image Based Treatment Planning. Radiother Oncol. 2009;92:S162.

36. Liu Q, Qin A, Liang J, Yan D. Evaluation of Atlas-Based Auto-Segmentation and Deformable Propagation of Evaluation of Atlas-Based Auto-Segmentation and Deformable Propagation of Organs-at-Risk for Head-and-Neck Adaptive Radiotherapy. Recent Patents Top Imaging. 2015;5(2):1–9.

37. Sjöberg C, Lundmark M, Granberg C, Johansson S, Ahnesjö A, Montelius A. Clinical evaluation of multi-atlas based segmentation of lymph node regions in head and neck and prostate cancer patients. Radiat Oncol. 2013;8(1):1–7.

38. Ingle C, Parker R, James H, Scrase CD. Evaluation of a commercial auto-segmentation tool in the outlining of lymph nodes & organs at risk in head and neck cancer. Radiother Oncol. 2011;99:S578.

39. Thomson D, Boylan C, Liptrot T, Aitkenhead A, Lee L, Sykes A, et al. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. Radiat Oncol. 2014;9:173.

40. Zhu M, Bzdusek K, Brink C, Eriksen JG, Hansen O, Jensen HA, et al. Multi-institutional quantitative evaluation and clinical validation of Smart Probabilistic Image Contouring Engine (SPICE) autosegmentation of target structures and normal tissues on computer tomography images in the head and neck, thorax, liver, and male. Int J Radiat Oncol Biol Phys. 2013;87(4):809–16.

41. Fortunati V, Lijn F Van Der, Niessen WJ, Veenland JF, Paulides MM, Walsum T Van. Tissue segmentation of head and neck CT images for treatment planning : A multiatlas approach combined with intensity modeling. Med Phys. 2013;40(7):71905-1–14.

42. Fritscher KD, Peroni M, Zaffino P, Spadea MF, Schubert R, Sharp G. Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. Med Phys. 2015;41(5):051910-1–11.

43. Agn M, Rosenschold PM, Puonti O, et al. A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiotherapy planning. Med Image Analysis, 2018, in press.

44. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys. 2017;44(2):547–57.

45. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.

46. Fritscher K, Raudaschl P, Zaffino P, et al. Deep Neural Networks for Fast Segmentation of 3D Medical Images. Medical Image Computing and Computer-Assisted Intervention –MICCAI 2016; 158-165.

47. Močnik D, Ibragimov B, Lei Xing L, et al. Segmentation of parotid glands from registered CT and MR images. Physica Medica, 2018; 52: 33-41.

48. Ren X, Xiang L, Nie D, et al. Interleaved 3D-CNNs for Joint Segmentation of Small-Volume Structures in Head and Neck CT Images. Medical physics. 2018;45(5):2063-2075.

49. Men K, Chen X, Zhang Y, Zhang T, Dai J, Yi J, et al. Deep Deconvolutional neural network for Target segmentation of nasopharyngeal cancer in Planning computed Tomography images. Front Oncol. 2017;7(December):1–9.

50. Menze BH, Jakab A, Bauer S et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging, 2015; 34(10): 1993-2024.

51. Zhu W, Huang Y, Xie X. AnatomyNet: Deep 3D squeeze-and-excitation U-Nets for fast and fully automated whole-volume anatomical segmentation. 2018. https://github.com/wentaozhu/AnatomyNet-for-anatomical-segmentation.git

52. Nikolov S, Blackwell S, Mendes R, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. 2018. https://arxiv.org/abs/1809.04430

53. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. Medical physics. 2014;41(5):050902.

54. Sykes J. Reflections on the current status of commercial automated segmentation systems in clinical practice. Journal of Medical Radiation Sciences. 2014;61(3):131-4.

55. Jameson MG, Holloway LC, Vial PJ, Vinod SK, Metcalfe PE. A review of methods of analysis in contouring studies for radiation oncology. Journal of medical imaging and radiation oncology. 2010;54(5):401-10.

56. Ezhil M, Choi B, Starkschall G, Bucci MK, Vedam S, Balter P. Comparison of rigid and adaptive methods of propagating gross tumor volume through respiratory phases of four-dimensional computed tomography image data set. Int J Radiat Oncol Biol Phys. 2008;71(1):290-6.

57. Speight R, Sykes J, Lindsay R, Franks K, Thwaites D. The evaluation of a deformable image registration segmentation technique for semi-automating

internal target volume (ITV) production from 4DCT images of lung stereotactic body radiotherapy (SBRT) patients. Radiotherapy and Oncology. 2011;98(2):277-83.

58. Valentini V, Boldrini L, Damiani A, Muren LP. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2014;112(3):317-20.

59. Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, 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.

60. Wuthrick EJ, Zhang Q, Machtay M, Rosenthal DI, Nguyen-tan PF, Fortin A, et al. Institutional Clinical Trial Accrual Volume and Survival of Patients With Head and Neck Cancer. J Clin Oncol. 2015;33(2):156–64.

Figure Legends

Figure 1: Overview of techniques used for auto-segmentation.
Autosegmentation can be atlas-based or deep learning. Within atlas-based techniques, model performance is expected to improve from single-atlas to multi-atlas to hybrid approaches.