Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring

van Dijk, Lisanne V; Van den Bosch, Lisa; Aljabar, Paul; Peressutti, Devis; Both, Stefan; J H M Steenbakkers, Roel; Langendijk, Johannes A; Gooding, Mark J; Brouwer, Charlotte L

Published in:
Radiotherapy and Oncology

DOI:
10.1016/j.radonc.2019.09.022

Radiotherapy and Oncology 142 (2020) 115–123

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring

Lisanne V. van Dijk [a,*], Lisa Van den Bosch [a], Paul Aljabar [b], Devis Peressutti [b], Stefan Both [a], Roel. J.H.M. Steenbakkers [a], Johannes A. Langendijk [a], Mark J. Gooding [b], Charlotte L. Brouwer [a]

[a] Department of Radiation Oncology, University of Groningen, University Medical Center Groningen, The Netherlands; [b] Mirada Medical Ltd, Oxford Centre for Innovation, UK

ARTICLE INFO

Article history:
Received 22 March 2019
Received in revised form 9 September 2019
Accepted 24 September 2019
Available online 22 October 2019

Keywords:
Head and neck
Organs at risks
Deep learning
Artificial intelligent
Auto segmentation
Contouring

ABSTRACT

Introduction: Adequate head and neck (HN) organ-at-risk (OAR) delineation is crucial for HN radiotherapy and for investigating the relationships between radiation dose to OARs and radiation-induced side effects. The automatic contouring algorithms that are currently in clinical use, such as atlas-based contouring (ABAS), leave room for improvement. The aim of this study was to use a comprehensive evaluation methodology to investigate the performance of HN OAR auto-contouring when using deep learning contouring (DLC), compared to ABAS.

Methods: The DLC neural network was trained on 589 HN cancer patients. DLC was compared to ABAS by providing each method with an independent validation cohort of 104 patients, which had also been manually contoured. For each of the 22 OAR contours – glandular, upper digestive tract and central nervous system (CNS)-related structures – the dice similarity coefficient (DICE), and absolute mean and max dose differences (|Δmean-dose| and |Δmax-dose|) performance measures were obtained. For a subset of 7 OARs, an evaluation of contouring time, inter-observer variation and subjective judgement was performed.

Results: DLC resulted in equal or significantly improved quantitative performance measures in 19 out of 22 OARs, compared to the ABAS (DICE/|Δmean dose|/|Δmax dose|: 0.59/4.2/4.1 Gy (ABAS); 0.74/1.1/0.8 Gy (DLC)). The improvements were mainly for the glandular and upper digestive tract OARs. DLC significantly reduced the delineation time for the inexperienced observer. The subjective evaluation showed that DLC contours were more often preferable to the ABAS contours overall, were considered to be more precise, and more often confused with manual contours. Manual contours still outperformed both DLC and ABAS; however, DLC results were within or bordering the inter-observer variability for the manual edited contours in this cohort.

Conclusion: The DLC, trained on a large HN cancer patient cohort, outperformed the ABAS for the majority of HN OARs.

Research on side effects of radiotherapy has been a steadily growing field of interest, as advances in treatment (e.g. multimodality imaging, proton therapy, targeted agents) have led to both increased life expectancy in cancer survivors and a greater degree of control in sparing organs-at-risk (OARs) [1]. Adequate delineation of OARs is crucial when investigating the association between radiation dose and side effects and when optimizing treatment planning. However, manual contouring of OARs is very time-consuming [2] and is prone to inter-observer variability [3,4]. This task requires significant expertise, especially for head and neck (HN) cancer patients, due to the complex anatomy. Meanwhile, the manual contouring burden on the clinic is rising as a consequence of the increasing number of OARs found to be associated with radiation-induced side effects [5–8].

Auto-contouring of OARs aims to reduce delineation time and effort, and to improve inter-observer consistency [9–11]. Atlas-based auto-contouring (ABAS) is a widely used method in which a set of representative patients with carefully delineated OARs serve as a reference set (i.e. atlas) for contouring new patients [12,13]. OAR contours of the atlas patient(s) are registered to new patients in order to transform (usually with a deformation) and combine their OAR contours in the new scan. Although ABAS has already reduced workload and improved consistency in many radiotherapy departments, there are a number of issues that leave room for improvement. First, ABAS generally underperforms for small and/or thin OARs, such as the swallowing muscles [14].

* Corresponding author at: Department of Radiation Oncology, University Medical Center Groningen, PO Box 30001, 9700 RB Groningen, The Netherlands.
E-mail address: l.v.van.dijk@umcg.nl (L.V. van Dijk).

https://doi.org/10.1016/j.radonc.2019.09.022
0167-8140/

Secondly, only a limited range of anatomical variation can be represented by typical sets of atlas patients (N = 10–30) as contouring performance generally plateaus with the inclusion of around 10–20 atlases [15,16], leading to poor delineation of structures in patients with anatomies differing from those in the transformed atlases [9]. Third, ABAS is also limited by the accuracy of the deformation between anatomies that, especially for CT, can be limited in areas with uniform intensity, such as soft tissues [17,18]. Finally, even for large databases, selecting the most appropriate atlases may be unreliable, potentially leading to sub-optimal performance [19].

Generating contours directly using deep learning techniques – derived from artificial intelligence research – has emerged as a promising method of addressing these challenges. Deep learning contouring (DLC) typically trains a convolutional neural network (CNN) model directly from the data without users needing to identify image features.

Improved computing power and training of neural networks have made deep learning methods more readily available for contouring purposes [20]. Several studies have already shown the potential of CNNs for HN contouring [21,22] and for other sites [23–25].

In this study, DLC was trained on a set of 589 HN cancer patients with complete sets of contours for 22 OARs (including glandular, upper digestive tract, central nervous system, bone and vessel related structures). OARs were carefully delineated according to the international OAR consensus guidelines [26]. The performance of DLC was comprehensively evaluated and compared to ABAS in an independent validation cohort of 104 HN cancer patients, using quantitative geometric and dosimetric measures, along with contouring time, inter-observer variation and subjective evaluations.

Methods

A schematic of the comparisons made in this study is shown in Fig. 1.

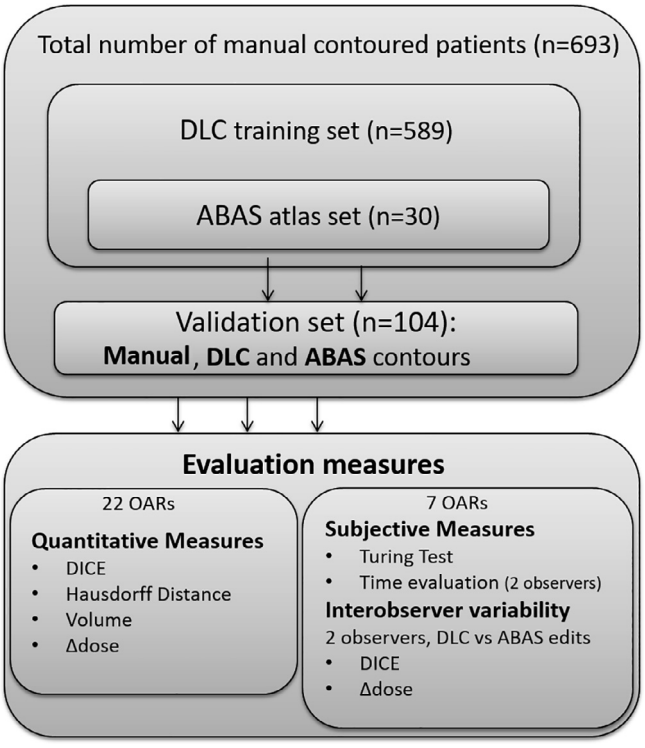

Fig. 1. Overview of evaluation methods.

Patients

A total of 693 HN cancer patients were included in this study (Table 1). All patients were treated with primary curative radiotherapy, with or without systemic treatment, between March 2007 and July 2016 at the University Medical Center Groningen (UMCG). The treatment modalities used were: 3D-CRT (12%), Simultaneous integrated boost (SIB) 7 Field IMRT (72%), or 2-ARC VMAT (16%). For most patients 70 Gy was prescribed to the primary tumour in 35 fractions. For each patient, a planning CT scan (Somatom Sensation Open, Somatom Definition AS or Biograph64, Siemens, Forchheim, Germany) was acquired approximately 2 weeks before treatment, with an average voxel size 0.98 × 0.98 × 2 mm (range: '0.62 × 0.62–1.37 × 1.37' × '2–4' mm); B30f or I40s\3; 80, 100–120 kV. An iterative metal artifact reduction reconstruction was used from 2015 on to limit the severity of artifacts in the scan on an minority of patients (~18%) and the majority of CT scans were contrast-enhanced (>90%). The clinical treatment plans were used when performing dose estimation. Patients were excluded if they had a salivary gland tumour, received prior HN surgery or radiotherapy. These exclusion criteria is related to an unpublished study cohort that is used for prediction of a comprehensive profile of side-effects of head and neck cancer. Additionally, when the CT scan had an in-plane voxel size larger than 1.5 mm scans were excluded.

Manual organ-at-risk (OAR) contouring

The OARs were manually delineated in the planning CT by a dedicated team of experts according to previously published

Table 1
Patient characteristics.

Characteristics	Train set		CV set		Test set		*p*-Value
	N = 549	%	N = 40	%	N = 104	%	
Sex							0.361
Female	139	25	13	33	21	20	
Male	410	75	27	68	83	80	
Age							0.293
18–65 years	368	67	20	50	64	62	
>65 years	181	33	20	50	40	38	
Tumour site							0.336
Oropharynx	194	35	15	38	45	43	
Nasopharynx	24	4	3	8	2	2	
Hypopharynx	53	10	2	5	10	10	
Larynx	255	46	18	45	38	37	
Oral cavity	23	4	2	5	9	9	
Other		0		0		0	
Tumour classification							0.895
Tis	3	1	0	0	0	0	
T1	87	16	7	18	14	13	
T2	180	33	12	30	29	28	
T3	133	24	10	25	30	29	
T4	146	27	11	28	31	30	
Node classification							0.702
N0	247	45	18	45	47	45	
N1	44	8	5	13	11	11	
N2	240	44	15	38	44	42	
N3	18	3	2	5	2	2	
Systemic treatment							0.410
Yes	235	43	13	33	45	43	
No	314	57	27	68	59	57	
Treatment technique							0.675
3D-CRT	65	12	6	15	9	9	
IMRT	394	72	27	68	81	78	
VMAT	90	16	7	18	14	13	
Neck irradiation							0.170
Bilateral	113	21	6	15	16	15	
Unilateral	12	2	3	8	2	2	
No	424	77	31	78	86	83	

Abbreviations: 3D-CRT: IMRT: Intensity-Modulated Radiation Therapy; VMAT: Volumetric Arc Therapy. CV: cross validation.

international consensus delineation guidelines [26]. Clinically available ABAS contours were often used as a basis for the contouring.

In this study, the following 22 OARs, divided in 3 sub-groups, were considered for auto-contouring:

1. *Glandular*: parotid glands (2×), submandibular glands (2x), and thyroid gland
2. *Upper digestive tract and airway-related*: arytenoids (2×), buccal mucosa (2×), extended oral cavity, pharyngeal constrictor muscle (PCM), cricopharyngeal inlet (cricoid), supraglottic area, glottic area, cervical esophagus
3. *Central nervous system (CNS), vessels and bone*: brainstem, cerebellum, cerebrum, spinal cord, mandible, and carotid arteries (2×)

Deep learning contouring (DLC)

The DLC implementation (DLCExpert™, Mirada Medical Ltd., UK) deploys multiple CNNs to predict a dense voxel-wise labelling for input CT images. A general 2D multiclass network with 14 layers predicts all OARs at a coarse resolution and its output, along with the CT image data, forms the input to a separately trained 10-layer OAR-specific network to predict the full resolution contours (loss function: cross entropy; batch size: 8 slices). Further details on this method can be found in the 2017 AAPM Challenge [27]. From the full set of 693 patients, 549 non-test cases were used for training and 40 were used for cross-validation to guide the training process. An independent validation and randomly selected cohort of 104 patients was excluded from training and cross-validation (Table 1) for the performance assessment.

Atlas-based auto-contouring (ABAS)

The ABAS implementation (WorkflowBox 1.4, Mirada Medical Ltd., UK) was designed using a representative set of 30 HN cancer patients taken from the training set. The atlas-patient image registration method was based on Lucas-Kanade Optic Flow [28], and a fixed set of atlases were used (i.e. no atlas selection) to generate a consensus contour using a sub-pixel precision form of majority voting [29,30]. The delineations of the atlas patients were carefully checked and patients with metal CT artifacts were excluded from the atlas patient set.

Quantitative evaluation: DICE, HD and dose

Both DLC and ABAS were applied to all 104 patients from the validation cohort. The performance of each method was evaluated by comparing the differences between the automatically generated and manual contours using the following metrics:

1. the Dice similarity coefficient (DICE) [31], which quantifies the overlap between contours A and B: $DICE = \frac{2(A \cap B)}{A+B}$
2. Hausdorff distance 95th-percentile (HD), i.e. the 95th percentile of the pairwise 3D point distances between two structures' contours [32]
3. Absolute dose difference between contours was determined using the clinical treatment plans:
   a. For the *glandular and upper digestive tract and airway related OARs,* the difference between mean dose was used (|Δmean-dose|)
   b. For the *CNS, vessels and bone OARs,* the difference between maximum dose was used (|Δmax-dose|)

Significance was assessed using the Wilcoxon signed rank test. DICE and HD was interpreted on a scale from "Very Poor" to "Good" based on previous inter-observer and contouring studies [3,33,34].

Time evaluation and observer-variability

Two observers, an expert (more than 10 years' experience) and a beginner (less than 2 years' experience), were randomly presented with either ABAS or DLC contours of 7 OARs which are among the most clinically relevant (left parotid and submandibular gland, thyroid gland, cricoid, glottic area, oral cavity, and PCM), and adjusted them as necessary to make them suitable for clinical use. Time was recorded from correcting first to last OAR and observers did not take breaks within this procedure. This was done for 14 patients taken from the test cohort and was carried out over two sessions, at least two months apart. The observers were blinded to the origin of the contours in each session.

After the editing step, the 4 sets of contours, unmodified ABAS, unmodified DLC, and the modified versions of each were obtained. The DICE and |Δmean-dose| were calculated for each set against the initial manual contours.

Subjective evaluation: Turing test

A subjective evaluation of the contouring methods was carried out with a Turing test (also known as an imitation game), which assumes clinical usability of auto-contours if they are difficult to

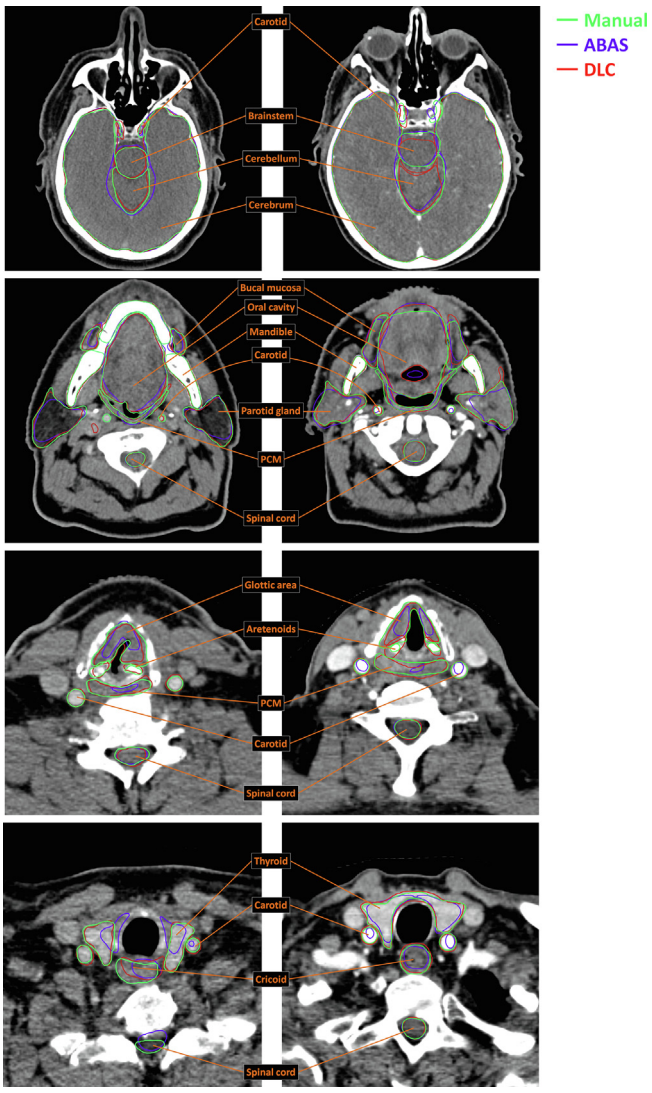

Fig. 2. Example of manual (green), ABAS (purple) and DLC (red) contours in two HN cancer patients for validation cohort for 4 different HN regions.

distinguish from human (i.e. manual) contours [35]. The expert observer was excluded from the test to prevent potential introduction of bias. Following the approach described by Gooding et al. [35], each observer was blindly presented with random slices that had an equal probability of featuring human, ABAS or DLC contours, for 7 OARs (see time evaluation) taken from the validation cohort patients. Using a web interface, the observers assessed the following questions for 100 scenarios:

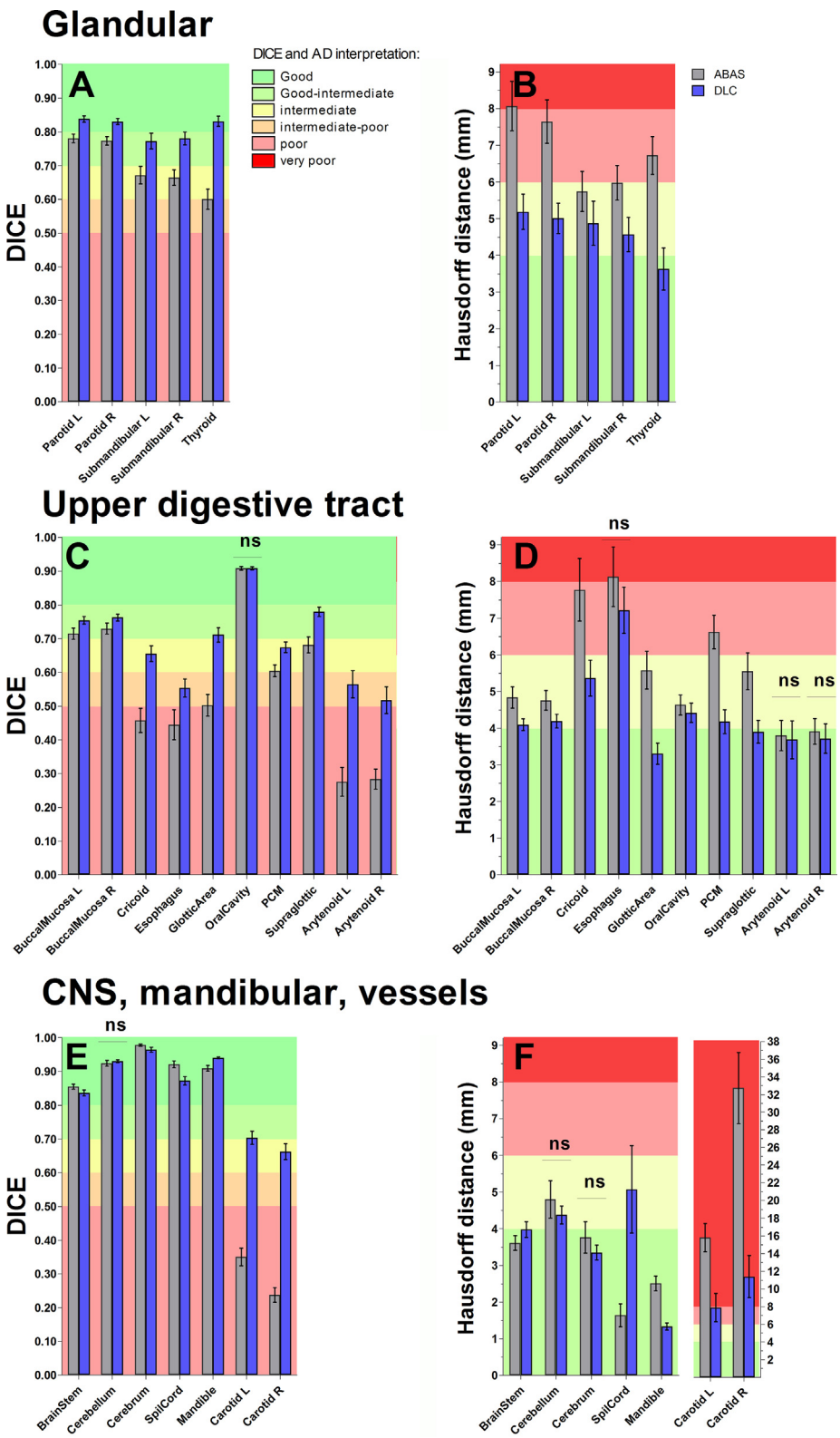

Fig. 3. Average and 95% Confidence Interval of DICE (Left) and HD (right) of ABAS (gray) and DLC (blue). Only non-significant difference between ABAS and DLC were indicated with ns. All others were significant with a p-value $\leq$0.001, except for the HD difference of the oral cavity (p-value = 0.02). OAR abbreviations: PCM: pharynx constrictor muscle; Cricoid: cricopharyngeal inlet; SpilCord: spinal cord; L: left; R: right.

1) A single contour: "How was this contour drawn?" Answer options: "By a human" or "By a computer".
2) Two contours: "Which contour do you prefer?" the preferred contour is selected by the observer.
3) A single contour: "You have been asked to Quality Assure this contour. Would you…"

Answer options:

a) "Require it to be corrected; there are large, obvious errors",
b) "Require it to be corrected; there are minor errors",
c) "Accept it as it is; There are minor errors that need a small amount of editing",
d) "Accept it as it is; the contour is very precise".

Results

For the 104 validation patients, ABAS failed to produce delineations for the esophagus in 6 cases, the glottic area in 2, and the left and right arytenoid in 66 and 43 patients, respectively. DLC failed to produce delineations for the glottic area in 5, the thyroid gland in 1, for the cricoid in 2, and the left and right arytenoid in 2 and 1 patients, respectively. If a contour was missing, the other auto-contoured OAR of the same patient was excluded from the evaluation to enable pair-wise comparison. Two example patients are shown in Fig. 2.

For all glandular OARs (Fig. 3A and B), DICE and HD values for the DLC significantly improved over ABAS ($p < 0.001$), with the largest difference for the thyroid gland, where DICE increased from 0.60 ± 0.15 (ABAS) to 0.83 ± 0.08 (DLC) and the HD decreased from 6.7 ± 2.6 (ABAS) to 3.6 ± 3.0 mm (DLC). DICE values for the parotid and submandibular glands increased on average from 0.72 ± 0.10 (ABAS) to 0.81 ± 0.08 (DLC). The mean dose differences (|Δmean-dose|) between the glandular manual and auto-contours were lower for DLC (0.9 ± 1.3 Gy) than for ABAS (1.9 ± 2.7 Gy) (Fig. 4A).

For all upper digestive tract and airway OARs (Fig. 3C and D), DICE values were significantly higher for DLC compared to ABAS for all OARs ($p < 0.001$), except for the oral cavity ($p = 0.84$). The largest differences were seen for the cricoid, supraglottic larynx, glottic area and PCM with average DICE values of 0.56 ± 0.14 (ABAS) and

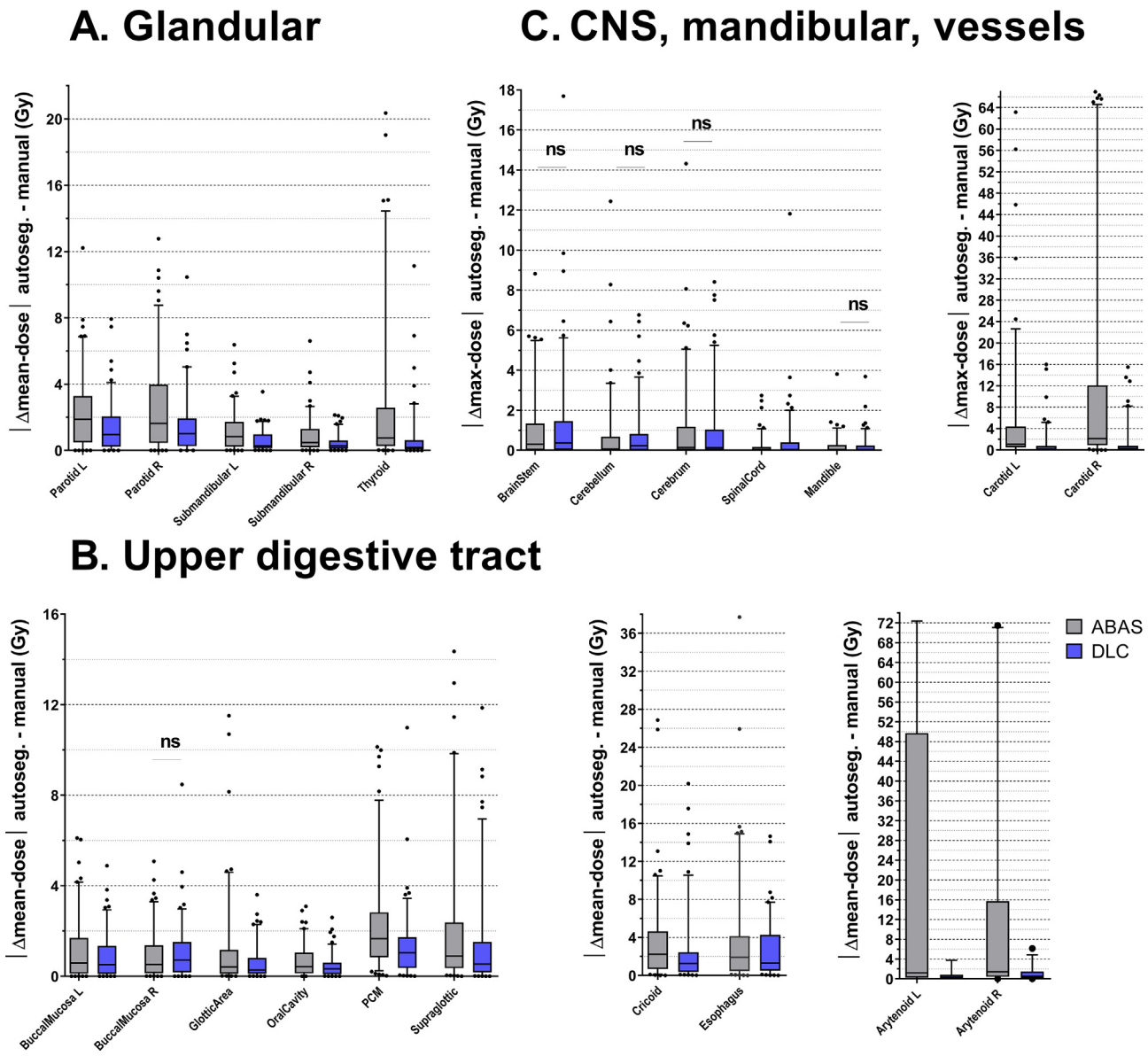

Fig. 4. Boxplots of absolute mean dose difference (|mean-dose|and for brainstem and spinal cord max dose difference (|max-dose|) between ABAS and manual contour (grey), and DLC and manual contour (blue) of all test patients (25th , 50% and 75%, 5th–95th percentile range, dot are outliers outside). For OAR abbreviations refer to Fig. 3.

0.71 ± 0.10 (DLC) and average HD values of 6.4 ± 3.1 mm (ABAS) and 4.1 ± 1.8 mm (DLC). The |Δmean-dose| values significantly decreased for all, except for the right buccal mucosa ($p = 0.72$) (Fig. 4B). For the arytenoids in particular, the |Δmean-dose| difference between ABAS (18.4 ± 27.8 Gy) and the DLC contours (0.8 ± 1.0 Gy) was large. The highest average |Δmean-dose| with DLC was found for the esophagus (2.6 ± 3.0 Gy) and cricoid (2.5 ± 3.6 Gy).

For the CNS OARs (Fig. 3E and F), the DICE values were good for ABAS (DICE >0.86) but were slightly lower for DLC (DICE >0.84) for the brainstem, cerebrum, and spinal cord ($p < 0.001$). The HD values were significant higher with DLC compared to ABAS for the brainstem and spinal cord ($p < 0.001$). The |Δmax-dose| values were comparable between the ABAS and DLC, except for a significant increase for the spinal cord ($p = 0.04$). *For the mandible*, the DICE and HD values were good for ABAS, but significantly improved with DLC ($p < 0.001$). *For the carotid arteries*, DICE values were substantially higher for DLC (0.68 ± 0.11) than for ABAS (0.29 ± 0.12) and HD substantially lower (DLC: 9.6 ± 10.4 mm ABAS: 24.3 ± 15.8 mm). The |Δmax-dose| improved markedly, reducing the average |Δmax-dose| < 1.3 Gy). The complete list of average DICE, HD and |Δmean-dose| values for all OARs is in Supplementary data 1.

For the time evaluation with the blinded approach, the 2 observers adjusted ABAS and DLC contours in two alternating sessions separated by 2 months. The average adjustment delineation time for the expert and beginner observer for the ABAS contours were 36 ± 7 and 59 ± 14 minutes, reducing slightly to 34 ± 6 and 54 ± 8 minutes for the DLC contours, respectively. The delineation time reduction was only significant for the beginner (Fig. 5).

Assessing the inter-observer variation, the DICE values between the adjusted auto-contours and the initial manual contours were for ABAS 0.79 ± 0.08 (beginner) and 0.81 ± 0.05 (expert) and for DLC 0.80 ± 0.06 (beginner) and 0.82 ± 0.05 (expert). Additionally, manual adjustments of the DLC contours showed little improvement in DICE compared to the unedited DLC contours, tested in the validation cohort, in 4 of the 7 OARs (refer to Supplementary data 2 for individual DICE values). The |Δmean-dose| did not change significantly in 5 of the 7 OARs. The average |Δmean-dose| for all 7 OARs and patients was 2.5 ± 2.3 Gy (ABAS) and 1.3 ± 1.6 Gy (DLC) without manual adjustment and 0.9 ± 0.7 Gy (ABAS) and 0.8 ± 0.7 Gy (DLC) with manual adjustment (Supplementary data 3). This suggests that DLC is approaching the level of inter-observer variability for these OARs.

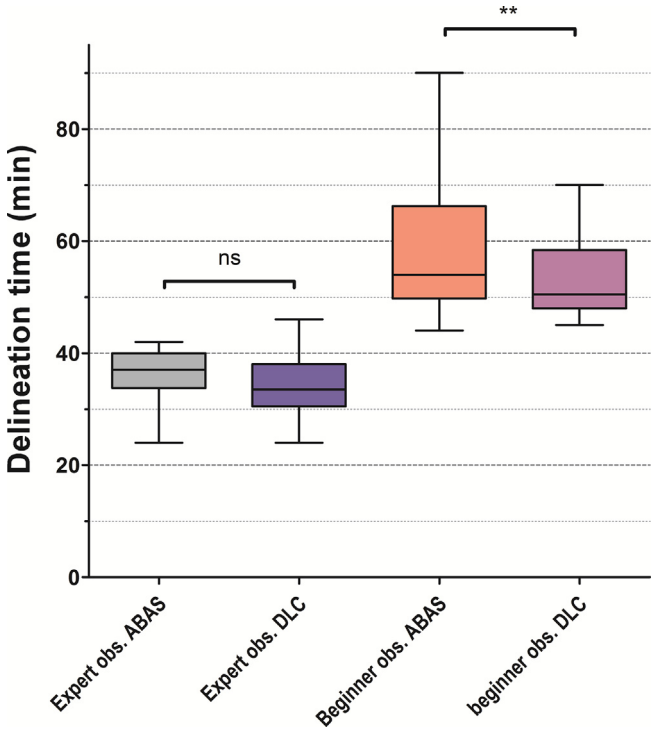

Fig. 5. Time evaluation for adjusting 7 OARs with ABAS vs. DLC for the expert and beginner observer (noted as obs.).

The subjective evaluation (Turing test) was performed by 12 observers: 10 physicians and 2 radiation oncology technicians (involved in clinical practice in OAR contouring and treatment planning). The Turing test was fully completed by 9 observers, generating evaluations for 965 scenarios. For the question "whether contours were human or computer-created", 40% of the human-created contours were misclassified as computer-generated (Fig. 6A). The misclassification rate for ABAS contours was 26% and for DLC contours it was 35%. For 5 of the 7 OARs, DLC contours were misclassified being human-created (34%) more often than ABAS (15%) (Fig. 6B). The difference was greatest for the thyroid and parotid gland. For the question "which of 2 contours was preferred", human contours were substantially more often preferred (81–82%) than either ABAS (18%) or DLC (19%) (Fig. 6C). For observers choosing between DLC and ABAS, DLC was selected substantially more often (66%). For the individual OARs, DLC was preferred more often over ABAS for all OARs except for the glottic area (Fig. 6D). The responses to the question "Would you correct the contour?" suggest low rates of obvious errors in both human and DLC contours, 7% and 9% respectively (Fig. 6E). In contrast, 30% of ABAS contours were considered to have obvious errors. Nevertheless, DLC had higher minor error rates than human contours, and consequently lower rates of being considered "precise" by the observers. For all individual OARs (Fig. 6F), human and DLC contours were 'required to be corrected' less often than ABAS except for the oral cavity. Similarly, human-drawn contours were required to be corrected less often than DLC contours, except for the thyroid, cricoid and PCM.

An overview of the improvement of all organs is given in Fig. 7. For additional volume analysis refer to Supplementary data 4.

Discussion

DLC was trained on a large cohort of 589 HNC patients with 22 high-quality manual HN OAR contours. This work showed that DLC significantly improved the auto-contours of the majority of 22 OARs (Fig. 7), compared to ABAS. The comprehensive evaluation that was carried out on an independent test set ($N = 104$), using quantitative measures of overlap, distance and dose, as well as subjective assessments and editing time, demonstrated that the greatest improvements were observed in the glandular structures, the upper digestive tract and airway OARs (e.g. the cricoid, PCM, supraglottic larynx and glottic area), arytenoids, and the carotid arteries. These results also translated to a decreased absolute difference between the mean dose/max dose of the auto-contoured and manual OARs (|Δmean-dose|max-dose|in addition to reduced variability between patients. The small dose differences indicate that DLC can be a valuable tool for large retrospective dose analyses such as development and validation of normal tissue complication probability (NTCP) models.

Our results show similar DICE scores to those by Liang et al. [22] (e.g. parotid gland: 0.85 ± 0.05, oral cavity: 0.91 ± 0.04), who introduced the innovative approach of training two CNNs on 186 nasopharynx patients, the first detecting a bounding box surrounding the OAR, the second contouring the OAR within it. The higher patient number in our study would theoretically make the system more robust to outlier cases. For the parotid gland, this resulted in

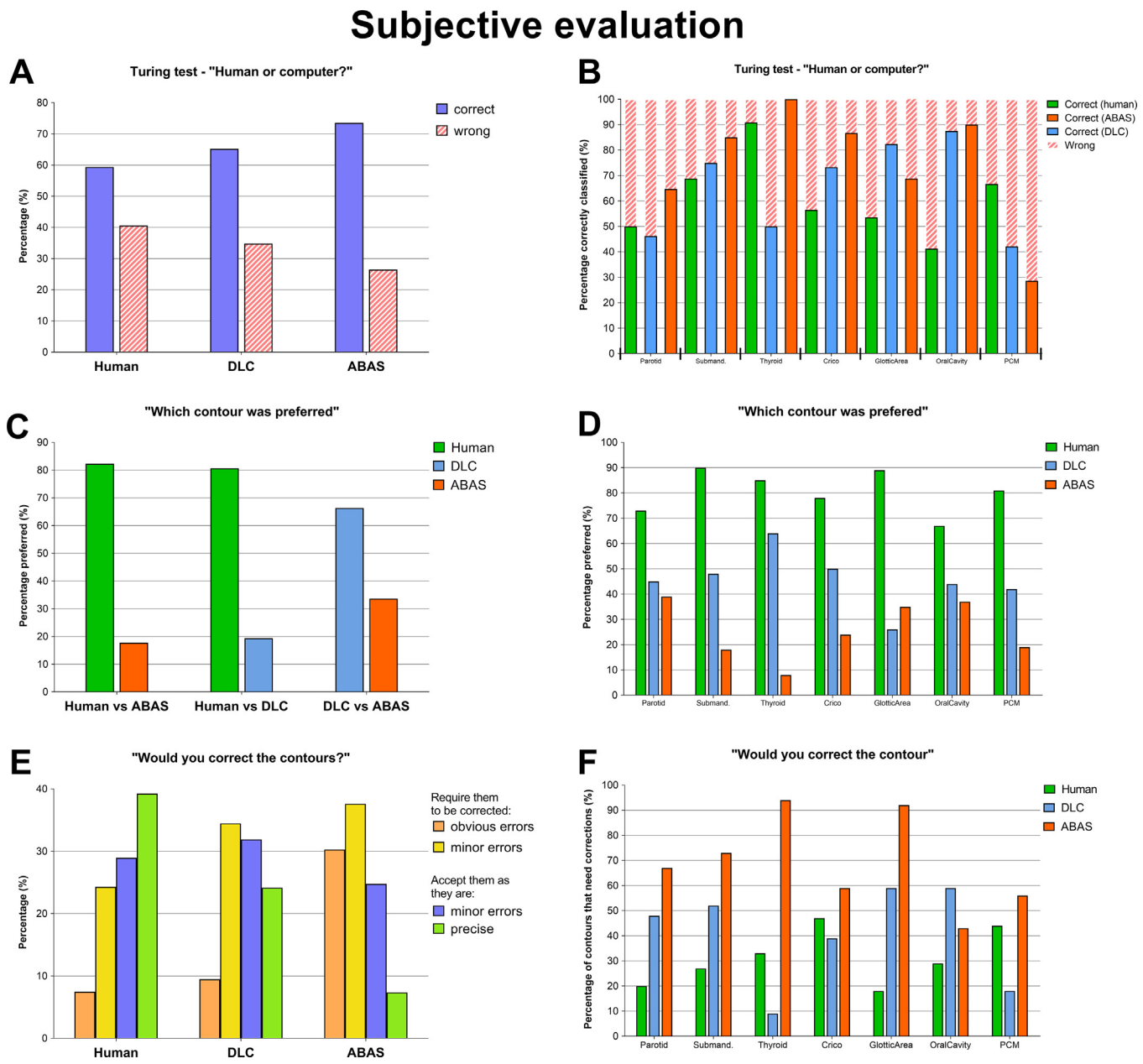

Fig. 6. Subjective evaluation. Response to questions: "How was this contour drawn: by a human or computer?" (A and B); "Which contour do you prefer?" choice between 2 blinded contour (C and D); "You have been asked to Quality Assure this contour. Would you... require them to be corrected or accept them as they are?" (E and F). For OAR abbreviations refer to Fig. 3.

only marginal improvement. Ibragimov et al. [21] showed slightly lower DICE values (e.g. parotid gland: 0.78 ± 0.05, submandibular gland: 0.73 ± 0.09), with a CNN trained on fewer than 50 HN cases. In contrast to these studies, in the current study, two consecutive CNNs were trained on a much larger and diverse patient cohort, in addition to using an independent validation cohort, providing better estimates of the generalizability and performance. The models also cover a greater number of OARs, including all upper digestive tract and airway OARs in accordance with the most recent delineation guidelines [26].

The time evaluation suggests that DLC may reduce the overall delineation time compared to ABAS, especially for inexperienced delineators. ABAS has been employed for a number of years in our clinic now, manual contouring of the limited OARs studied for the time evaluation study can take up to 90 minutes per patient (internal evaluation). This is in line with previous reported manual contouring time of 108 min for a set of six OARs in the head and neck area [36].

The comparison between the contours of both observers and the DLC, suggests that DLC is approaching the level of inter-observer variability for these OARs in terms of DICE and |mean-dose|The DICE variation reported between observers is in line with previous studies reporting inter-observer variability [3,37]. Since it is challenging to assess clinical usability from geometric measures alone (e.g. DICE showed little improvement after manual editing in the time evaluation), a subjective evaluation was also performed. This confirmed that manual contours remain preferable over both DLC and ABAS contours by the observers (Fig. 6C). However, the Turing test also showed that it remained relatively difficult to identify contours as being human- or computer created (Fig. 6B). The overall misclassification of human contours was 41% (Fig. 6A), and 32% of the human contours were marked as requiring correc-

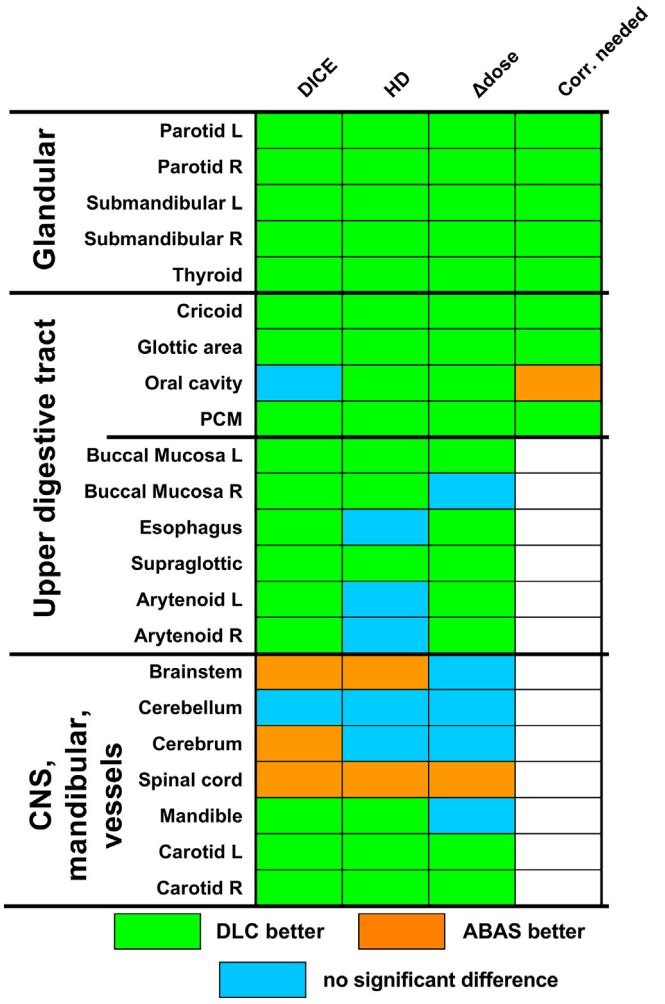

Fig. 7. Overview of the results of all HN OARs. Green indicates that DLC is significantly better than ABAS, orange that ABAS is significantly better than DLC and blue indicates that there is no significant difference. Corr. Needed = Objective evaluation indicated that manual corrections are advised.

tion (Fig. 6E). The latter indicates the inter-observer variation. Moreover, DLC performed substantially better than the ABAS, since the DLC contours for almost all OARs were more often confused with human contours (Fig. 6B). They were also more often preferred (Fig. 6D) and considered directly clinically usable (Fig. 6E and F) than the ABAS contours for nearly all OARs. Overall, DLC appeared to perform exceedingly well for the parotid and thyroid gland, cricoid inlet, and PCM, in line with the quantitative and |mean-dose|results (Fig. 4).

For the CNS OARs, performance measures with ABAS were already good and the DLC (DICE > 0.84 and HD < 3.3 mm) did not out-perform ABAS, approaching the inter-observer variation [38]. This study may be somewhat biased towards the ABAS, since the basis of the manual delineations used in this study were often ABAS (yet with different atlas patients) contours, especially for the CNS and mandible contours. This bias is minimized by the thorough checks and edits of all manual delineations to conform them to the new consensus guidelines, which were performed by a small team of experts.

Another limitation concerning the subjective evaluation is the presentation of a single slice per contour to the observers, having access to three-dimensional information might change the results. In addition, ABAS of a single system was compared to the DLC. Other studies showed similar DICE values between ABAS and manual contours with other systems [13,14], but the ABAS contours were subjectively better scored [14]. Moreover, the time and inter-observer evaluation was limited by the number of observers and patients. Furthermore, the DLC did not perform well on CT scans with very different voxel sizes (in-plane >1.5 mm), but did perform well for both non- and contrast-enhanced CT scans, and for patients with the different tumour sites. In cases where the auto-contouring failed to produce output, no contour was returned. For DLC, this is because no activation threshold was reached to consider any voxel inside the contour. For ABAS, this is because there were no voxels where the majority of the multi-atlas registrations contours overlapped. Additionally, while ABAS failed to produce arytenoid contours for many patients (69), DLC failed for only 2, illustrating the difficulty for the ABAS to contour small OARs and the relative robustness of DLC. Nevertheless, DLC occasionally omitted to produce contours and this might be helped by an even larger cohort. Finally, our results may be improved by incorporating other image modalities when training DLC models.

A DLC model can be more robust than ABAS since it can be trained using as much data as is available, including patients with metal artifacts and diverse anatomy. Additionally, it incorporates appearance and patterns of structures with convolution filters, rather than the anatomic location-and intensity-only approach of the ABAS. For these two reasons, DLC likely outperforms ABAS especially in areas with more anatomic variability [27,39].

As the Turing Test showed a higher acceptance of DLC contours than ABAS contours (Fig. 6E and F) it was expected to find a much larger time saving from the time evaluation. The reason for the substantial editing times might be a difference between the observer groups. Medical doctors that participated in the Turing Test might be willing to accept minor errors, while the delineators in the time evaluation were more focused on editing all minor errors. This is partly confirmed by the findings presented in Supplementary Data 2, where it can be seen that, for some OARs, the inter-observer variability results in similar DICE scores for human–human comparisons as for human-DLC comparisons. Future work should therefore focus on identifying the reasons and significance of the large editing time of DLC that was found in this study.

In conclusion, DLC, trained on a large cohort, outperformed ABAS for the majority of HN OARs. The improved quantitative performance translated to smaller dosimetric differences compared to the manual contours. Although manual contours were clearly preferred and less often required to be corrected than both DLC and ABAS contours, DLC outperformed the ABAS in nearly all subjective assessments. Additionally, the time evaluation showed significant benefit for the inexperienced delineator. DLC has the potential to reduce clinical burden by reducing the delineation time required to produce acceptable contours and currently replaced ABAS in routine use in our clinic.

Conflict of interest

University Medical Center Groningen has research collaboration with Mirada Medical, Oxford, UK. Authors MJG and PA are employees of Mirada Medical Ltd. DP was an employee of Mirada Medical Ltd.

Acknowledgements

We want to thank Harriëtte van der Laan, Roelof Pot and Erik Bakker for their efforts contouring for the time evaluation section of the study and for producing manual contours that were used to train and validate the DLC.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2019.09.022.

References

[1] Grégoire V, Langendijk JA, Nuyts S. Advances in radiotherapy for head and neck cancer. J Clin Oncol 2015;33:3277–84. https://doi.org/10.1200/JCO.2015.61.2994.

[2] Vorwerk H, Zink K, Schiller R, Budach V, Kampfer S, Popp W, et al. Protection of quality and innovation in radiation oncology: the prospective multicenter trial the German Society of Radiation Oncology (DEGRO-QUIRO study). Evaluation of time, attendance of medical staff, and resources during radiotherapy with IMRT. Strahlenther Onkol 2014:433–43. https://doi.org/10.1007/s00066-014-0634-0.

[3] Brouwer CL, Steenbakkers RJ, van den Heuvel E, Duppen JC, Navran A, Bijl HP, et al. 3D Variation in delineation of head and neck organs at risk. Radiat Oncol 2012;7:32. https://doi.org/10.1186/1748-717X-7-32.

[4] Geets X, Daisne JF, Arcangeli S, Coche E, De Poel M, Duprez T, et al. Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: comparison between CT-scan and MRI. Radiother Oncol 2005;77:25–31. https://doi.org/10.1016/j.radonc.2005.04.010.

[5] Beetz I, Schilstra C, Van Der Schaaf A, Van Den Heuvel ER, Doornaert P, Van Luijk P, et al. NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: the role of dosimetric and clinical factors. Radiother Oncol 2012;105:101–6. https://doi.org/10.1016/j.radonc.2012.03.004.

[6] Wopken K, Bijl HP, Van Der Schaaf A, Christianen ME, Chouvalova O, Oosting SF, et al. Development and validation of a prediction model for tube feeding dependence after curative (Chemo-) radiation in head and neck cancer. PLoS ONE 2014;9:1–8. https://doi.org/10.1371/journal.pone.0094879.

[7] Schwartz DL, Hutcheson K, Barringer D, Tucker SL, Kies M, Holsinger FC, et al. Candidate dosimetric predictors of long-term swallowing dysfunction after oropharyngeal intensity-modulated radiotherapy. Int J Radiat Oncol Biol Phys 2010;78:1356–65. https://doi.org/10.1016/j.ijrobp.2009.10.002.

[8] Deasy JO, Moiseenko V, Marks L, Chao KSC, Nam J, Eisbruch A. Radiotherapy dose-volume effects on salivary gland function. Int J Radiat Oncol Biol Phys 2010;76:S58–63. https://doi.org/10.1016/j.ijrobp.2009.06.090.

[9] Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. Med Phys 2014;41:50902. https://doi.org/10.1118/1.4871620.

[10] 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;55:799–806. https://doi.org/10.3109/0284186X.2016.1173723.

[11] Valentini V, Boldrini L, Damiani A, Muren LP. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. Radiother Oncol 2014;112:317–20. https://doi.org/10.1016/j.radonc.2014.09.014.

[12] Brandt R, Menzel R. Quo Vadis, Atlas-based segmentation? 2007. doi: 10.1007/0-306-48608-3.

[13] Han X, Hoogeman MS, Levendag PC, Hibbard LS, Teguh DN, Voet P, et al. Atlas-based auto-segmentation of head and neck CT images. Med Image Comput Comput Assit Interv 2008;5242:434–41. https://doi.org/10.1007/978-3-540-85990-1-52.

[14] Teguh DN, Levendag PC, Voet PWJ, Al-Mamgani A, Han X, Wolf TK, 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;81:950–7. https://doi.org/10.1016/j.ijrobp.2010.07.009.

[15] Van De Velde J, Wouters J, Vercauteren T, De Gersem W, Achten E. Optimal number of atlases and label fusion for automatic multi-atlas-based brachial plexus contouring in radiotherapy treatment planning. Radiat Oncol 2016:1–9. https://doi.org/10.1186/s13014-015-0579-1.

[16] Larrue A, Gujral D, Nutting C, Gooding M. The impact of the number of atlases on the performance of automatic multi-atlas contouring. Phys Med 2015;31:. https://doi.org/10.1016/j.ejmp.2015.10.020e30.

[17] Yeo UJ, Supple JR, Taylor ML, Smith R, Kron T, Franich RD. Perfomance of 12 DIR algorithms in low-contrast regions for mass and density conserving deformation. Med Phys 2013;40:1–12. https://doi.org/10.1118/1.4819945.

[18] Zhong H, Kim J, Chetty IJ. Analysis of deformable image registration accuracy using computational modeling. Med Phys 2010;37:970–9. https://doi.org/10.1118/1.3302141.

[19] Peressutti D, Schipaanboord B, van Soest J, Lustberg T, van Elmpt W, Kadir T, et al. How effective are current atlas selection methods for atlas-based auto-contouring in radiotherapy planning?. Med Phys 2016;43:3738–9.

[20] Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. Comput Biol Med 2018;98:126–46. https://doi.org/10.1016/j.compbiomed.2018.05.018.

[21] Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys 2017;44:547–57. https://doi.org/10.1002/mp.12045.

[22] Liang S, Tang F, Huang X, Yang K, Zhong T, Hu R, et al. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. Eur Radiol 2018. https://doi.org/10.1007/s00330-018-5748-9.

[23] Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother Oncol 2018;126:312–7. https://doi.org/10.1016/j.radonc.2017.11.012.

[24] Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. Med Phys 2017;44:6377–89. https://doi.org/10.1002/mp.12602.

[25] Ibragimov B, Toesca D, Chang D, Koong A, Xing L. Combining deep learning with anatomical analysis for segmentation of the portal vein for liver SBRT planning. Phys Med Biol 2017;62:8943–58. https://doi.org/10.1088/1361-6560/aa9262.

[26] 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;117:83–90. https://doi.org/10.1016/j.radonc.2015.07.041.

[27] Jinzhong Y, Iii SGA, Kirby JS, Oliveira B, Zamdborg L, Gooding M, et al. Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017. Med Phys 2017;45:4568–82.

[28] Schipaanboord B, Boukerroui D, Peressutti D, Van Soest J, Lustberg T, Dekker A, et al. An evaluation of atlas selection methods for atlas-based automatic segmentation in radiotherapy treatment planning. IEEE Trans Med Imaging 2019:1. https://doi.org/10.1109/TMI.2019.2907072.

[29] Wardman K, Gooding M, Preswich R, Speight R. The feasibility of atlas-based automatic segmentation of MRI for H&N radiotherapy planning. Radiother Oncol 2016;119:S891–2. https://doi.org/10.1016/s0167-8140(16)33137-1.

[30] Schipaanboord B, Boukerroui D, Peressutti D, Van Soest J, Lustberg T, Kadir T, et al. Can atlas-based auto-segmentation ever be perfect? Insights from extreme value theory. IEEE Trans Med Imaging 2019;38:99–106. https://doi.org/10.1109/TMI.2018.2856464.

[31] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302. https://doi.org/10.2307/1932409.

[32] Heimann T, Van Ginneken B, Styner MA, Arzhaeva Y, Aurich V, Bauer C, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging 2003;28:280. https://doi.org/10.1109/TMI.2009.2013851.

[33] Thomson D, Boylan C, Liptrot T, Aitkenhead A, Lee L, Yap B, et al. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. Radiat Oncol 2014;9:1–12. https://doi.org/10.1186/1748-717X-9-173.

[34] Kieselmann JP, Kamerling CP, Burgos N, Menten MJ, Fuller CD, Nill S, et al. Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region. Phys Med Biol 2018:63. https://doi.org/10.1088/1361-6560/aacb65.

[35] Gooding MJ, Smith AJ, Tariq M, Aljabar P, Peressutti D, van der Stoep J, et al. Comparative evaluation of auto-contouring in clinical practice: a practical method using the Turing Test. Med Phys 2018:1–11. https://doi.org/10.1002/mp.13200.

[36] Fung NTC, Hung WM, Sze CK, Lee MCH, Ng WT. Automatic segmentation for adaptive planning in nasopharyngeal carcinoma IMRT: time, geometrical, and dosimetric analysis. Med Dosim 2019:3–8. https://doi.org/10.1016/j.meddos.2019.06.002.

[37] Feng M, Demiroz C, Vineberg KA, Eisbruch A, Balter JM. Normal tissue anatomy for oropharyngeal cancer: contouring variability and its impact on optimization. Int J Radiat Oncol Biol Phys 2012;84:e245–9. https://doi.org/10.1016/j.ijrobp.2012.03.031.

[38] Beddok A, Faivre JC, Coutte A, Le Guévelou J, Welmant J, Clavier JB, et al. Practical contouring guidelines with an MR-based atlas of brainstem structures involved in radiation-induced nausea and vomiting. Radiother Oncol 2019;130:113–20. https://doi.org/10.1016/j.radonc.2018.08.003.

[39] Aljabar P, Peressutti D, Brunenberg E, Smeenk R, Van Leeuwen R, Gooding M. OC-0419: Comparison of auto-contouring methods for regions of interest in prostate CT. Radiother Oncol 2018;127:S218–9. https://doi.org/10.1016/s0167-8140(18)30729-1.