A novel semi auto-segmentation method for accurate dose and NTCP evaluation in adaptive head and neck radiotherapy

Gan, Yong; Langendijk, Johannes A; Oldehinkel, Edwin; Scandurra, Daniel; Sijtsema, Nanna M; Lin, Zhixiong; Both, Stefan; Brouwer, Charlotte L

Published in:
Radiotherapy and Oncology

DOI:
10.1016/j.radonc.2021.09.019
10.1016/j.radonc.2021.09.019

Publication date:
2021

Radiotherapy and Oncology 164 (2021) 167–174

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

A novel semi auto-segmentation method for accurate dose and NTCP evaluation in adaptive head and neck radiotherapy

Yong Gan [a,*], Johannes A. Langendijk [a], Edwin Oldehinkel [a], Daniel Scandurra [a], Nanna M. Sijtsema [a], Zhixiong Lin [b], Stefan Both [a], Charlotte L. Brouwer [a]

[a] University of Groningen, University Medical Center Groningen, Department of Radiation Oncology, Groningen, The Netherlands; [b] Shantou University, Cancer Hospital of Shantou University Medical College, Department of Radiotherapy, China

ARTICLE INFO

Article history:
Received 2 May 2021
Received in revised form 15 August 2021
Accepted 17 September 2021
Available online 28 September 2021

Keywords:
Head and neck cancer
Organs at risk
Auto-segmentation
Deep learning contouring
Deformable image registration
Dosimetric changes

ABSTRACT

Background and purpose: Accurate segmentation of organs-at-risk (OARs) is crucial but tedious and time-consuming in adaptive radiotherapy (ART). The purpose of this work was to automate head and neck OAR-segmentation on repeat CT (rCT) by an optimal combination of human and auto-segmentation for accurate prediction of Normal Tissue Complication Probability (NTCP).

Materials and methods: Human segmentation (HS) of 3 observers, deformable image registration (DIR) based contour propagation and deep learning contouring (DLC) were carried out to segment 15 OARs on 15 rCTs. The original treatment plan was re-calculated on rCT to obtain mean dose (D_{mean}) and consequent NTCP-predictions. The average D_{mean} and NTCP-predictions of the three observers were referred to as the gold standard to calculate the absolute difference of D_{mean} and NTCP-predictions ($|\Delta D_{mean}|$ and $|\Delta NTCP|$).

Results: The average $|\Delta D_{mean}|$ of parotid glands in HS was 1.40 Gy, lower than that obtained with DIR and DLC (3.64 Gy, $p < 0.001$ and 3.72 Gy, $p < 0.001$, respectively). DLC showed the highest $|\Delta D_{mean}|$ in middle Pharyngeal Constrictor Muscle (PCM) (5.13 Gy, $p = 0.01$). DIR showed second highest $|\Delta D_{mean}|$ in the cricopharyngeal inlet (2.85 Gy, $p = 0.01$). The semi auto-segmentation (SAS) adopted HS, DIR and DLC for segmentation of parotid glands, PCM and all other OARs, respectively. The 90th percentile $|\Delta NTCP|$ was 2.19%, 2.24%, 1.10% and 1.50% for DIR, DLC, HS and SAS respectively.

Conclusions: Human segmentation of the parotid glands remains necessary for accurate interpretation of mean dose and NTCP during ART. Proposed semi auto-segmentation allows NTCP-predictions within 1.5% accuracy for 90% of the cases.

Previous studies have demonstrated that anatomical deformation and geometric changes of organs at risk (OARs) as well as target volume shrinkage during radiotherapy may incur changes in dose distributions of head and neck cancer patients [1–3]. Adaptive radiotherapy (ART) aims to address this issue but requires repeat CT-scans (rCT) and re-segmentation of OARs, which impedes clinical application [4]. To relieve the workload of re-segmentation, various auto-segmentation methods, varying from atlas-based to deep learning methods, have been developed trying to replace human segmentation (HS) [5].

Atlas-based auto-segmentation (ABAS) is available in several commercial implementations. Previous studies have shown the significant potential of ABAS as a timesaving and variability-decreasing tool in segmenting target volumes and OARs in head and neck, but it cannot be used independently without human intervention [6].

Deep learning contouring (DLC) applies deep learning, a subfield of machine learning, for image detection and segmentation [7]. Deep learning consists of multiple layers of data processing making it possible to learn by representing these data through several levels of abstraction [8]. Previous studies have demonstrated that DLC outperforms ABAS and performs sufficiently well for segmentation of most head and neck OARs, although it is still inferior to HS [9,10].

Deformable image registration (DIR) is commonly applied for contour propagation but controversial for dose propagation mainly due to the mass change of organs or tumours [11–14]. A comparison of re-segmentation on rCT between DIR and DLC has not yet been made. Previous papers have mostly focused on geometry or dose variability and rarely investigated the consequent deviations of normal tissue complication probability (NTCP) values compared to the nominal plans.

* Corresponding author at: Department of Radiation Oncology, University Medical Center Groningen, PO Box 30001, 9700 RB Groningen, The Netherlands.
E-mail address: y.gan@umcg.nl (Y. Gan).

https://doi.org/10.1016/j.radonc.2021.09.019
0167-8140/

Inaccuracies of auto-segmentation results are often related to the magnitude of the inter-observer variability seen in HS. In most studies on auto-segmentation, only one single HS was referred to as the gold standard [5,15,16]. During head and neck ART, clinically relevant changes need to be detected as quickly and accurately as possible. Actual anatomical changes can be small [17] while inter-observer variability of manual segmentation of lesions and OARs is sometimes large [16], so to select and identify patients who will benefit from ART requires an accurate segmentation method and clear thresholds.

Based on the comparison of mean dose (D_{mean}) and consequent NTCP-predictions, the current study aimed to define a fast and accurate semi auto-segmentation (SAS) method for OAR re-segmentation in head and neck cancer patients for evaluation of D_{mean} and NTCP-prediction during ART.

Materials and methods

Patients and selection of rCT

The population of this study was composed of 15 patients from our prospective data registration program that had a full set of weekly rCTs obtained according to the standard care of our department during the course of treatment. For each enrolled patient, the rCT with the largest geometric changes compared to the planning CT (visual evaluation) was selected for re-segmentation of OARs. Patients' age ranged from 53 to 72 years, primary tumour locations included oropharynx (n = 11), hypopharynx (n = 2), oral cavity (n = 1) and larynx (n = 1). All patients underwent radiotherapy with IMRT or VMAT, using 35 fractions of 2 Gy to a total dose of 70 Gy in 7 weeks. Segmentation of OARs on planning CT was done by the specialized segmentation team according to the international consensus guidelines [18]. Patients were excluded if they had previously received surgery or radiotherapy to the head and neck area and/or in case of metal artefacts on head and neck CT-scans.

CT scan

All patients received CT-scans (Somatom Sensation Open, Somatom Definition AS or Biograph64, Siemens, Forchheim, Germany) 2 weeks before radiotherapy and weekly during the course of radiotherapy 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.

Human segmentation (HS)

Three observers with more than 3 years' experience in OAR segmentation (designated as HS1, HS2 and HS3 respectively) carried out HS independently on rCT according to the international consensus guidelines [18]. Segmentation on planning CT was done according to clinical practice, while auto-segmentation (multi-atlas segmentation, Mirada Medical) was corrected by the specialized head and neck OAR segmentation team.

Deformable image registration (DIR)

The hybrid deformable registration algorithm (pre-setting: no controlling ROIs, default deformation strategy, resolution of 0.25 cm in three dimensions) was used for DIR in RayStation Research Version 8.99 (RaySearch Laboratories, Stockholm, Sweden). The produced deformation vector field was then used to map OAR contours from planning-CT to rCT.

Deep learning contouring (DLC)

All rCTs were exported to Mirada for DLC and post-processing (Workflow Box 2.0, DLCExpertTM, Mirada Medical Ltd., UK). Details of DLC can be found in the 2017 AAPM Challenge [19]. After completion of DLC in Mirada, the segmented structures were imported to RayStation Research.

Organs at risk (OARs)

A total of 15 OARs were segmented, including: Brain, Brainstem, Mandible, Oral Cavity,

Cricopharyngeal inlet (Crico), Glottic area, Supraglottic larynx, PCM_Superior (Superior pharyngeal constrictor muscles), PCM_Middle (Middle pharyngeal constrictor muscles), PCM_Inferior (Inferior pharyngeal constrictor muscles), PCM (combination of PCM_Superior, PCM_Middle and PCM_Inferior), Buccal mucosa (combination of bilateral Buccal mucosa), Submandibular glands (combination of bilateral submandibular glands), Arytenoids (combination of bilateral arytenoids), Parotid glands (combination of bilateral parotid glands).

Dosimetric parameters and NTCP models

- The most common dosimetric predictors adopted in NTCP models were D_{mean}. For all OARs, only D_{mean} was evaluated.
- 135 NTCP models describing multidimensional toxicity of head and neck radiotherapy were applied to translate the dosimetric parameters into potential clinical relevance. For more details of the NTCP models please refer to [20]

Absolute difference of D_{mean} and NTCP-predictions

The average D_{mean} and consequent NTCP-predictions of the three observers (HS1, HS2, HS3) was considered gold standard (HS*). The absolute difference of D_{mean} and consequent NTCP-prediction ($|\Delta D_{mean}|$ and $|\Delta NTCP|$) between HS* and each segmentation set were calculated. For HS, the maximum $|\Delta D_{mean}|$ and $|\Delta NTCP|$ of 3 observers was defined as absolute difference of human segmentation. After discussion with radiation oncologists, major $|\Delta NTCP|$ was defined as percentage point difference more than 3%.

Semi auto-segmentation (SAS)

For each OAR in SAS, the optimal segmentation from HS, DIR and DLC was manually selected based on the $|\Delta D_{mean}|$ of 15 patients. One previous study showed the $|\Delta D_{mean}|$ induced by inter-observer variability in segmenting OARs of head and neck could be up to 0.8 Gy on average [9]. To balance efficiency and accuracy of SAS, we preferentially selected segmentation from DIR or DLC for each OAR of which the $|\Delta D_{mean}|$ fulfilled at least one of the following criteria: 1, less than HS; 2, less than 0.8 Gy; 3, no statistically significant difference compared to HS, otherwise HS was selected.

In addition, we selected segmentation from the same method as many as possible to reduce the overlap of segmentation. Once the selection of segmentation for each OAR was determined, the combination of segmentation for all OARs was fixed and defined as SAS.

Statistical analysis and plotting

A paired sample t-test (normally distributed data) or related-samples Wilcoxon Signed Rank Test (non-normally distributed data) were utilized to test for statistically significant difference $|\Delta D_{mean}|$ between HS and each auto-segmentation. Because there are two comparisons within 3 groups of data, a Bonferroni correction was adopted with significant p-value of 0.05/2 (0.025). Graph-

Pad Prism 8.2.1 software (GraphPad Software Inc., San Diego, CA, USA) was used for statistical analysis and graphing.

Results

For each segmentation set, 15 OARs in 15 rCTs produced 225 D_{mean} parameters, 135 NTCP-models in 15 patients produced 2025 NTCP-predictions.

Except supraglottic larynx segmentations of DLC, all OAR segmentations by DIR and DLC showed larger average $|\Delta D_{mean}|$ than HS. Both DIR and DLC segmentations showed statistically significant larger values of $|\Delta D_{mean}|$ in parotid glands. DIR showed significantly larger $|\Delta D_{mean}|$ of brain and cricopharyngeal inlet than HS. Using DLC, PCM_Middle showed the largest and significantly larger $|\Delta D_{mean}|$ as compared to HS (Table 1).

For the parotid glands, both DIR and DLC showed significantly larger |ΔDmean| than HS (3.64 Gy, p-value < 0.001, and 3.72 Gy, p-value < 0.001 compared to 1.40 Gy, respectively). For PCM_Middle, the average |ΔDmean| was 5.13 Gy (DLC), 1.82 Gy (DIR) and 1.19 Gy (HS). The Crico showed the second highest |ΔDmean| (2.85 Gy, p-value = 0.01) in DIR, higher than DLC (1.70 Gy, p-value = 0.08) and HS (1.05 Gy). Brain by DIR, mandible by DLC showed significant differences of |ΔDmean| compared to by HS, but was only 0.17 Gy and 0.30 Gy on average respectively. For the other OARs, the |ΔDmean| obtained with DLC were more approximated to HS than with DIR except for the submandibular glands (DIR: 0.86 Gy, p-value = 0.17; DLC: 0.98 Gy, p-value = 0.17). The |ΔDmean| in 15 OARs of each patient were presented with a colour gradient in Fig. 1. For the difference of Dmean between the three human observers and gold standard please refer to supplementary data 1 and 2.

Based on the $|\Delta D_{mean}|$ and according to the pre-defined criteria of selecting segmentation, DLC was selected for segmentation of all OARs in SAS except for parotid glands and PCM, which were segmented by human and DIR respectively. For the D_{mean} of parotid glands, the one with the maximum $|\Delta D_{mean}|$ among 3 observers was used to calculate NTCP-prediction in SAS.

The 2025 NTCP predictions of each segmentation set were presented with median plus percentile value due to skewed distributions. The median |ΔNTCP| of HS, DIR and DLC were all less than 0.5%, the 90th percentile of DIR and DLC were 2.19% and 2.24% respectively, which is around two times of HS (1.10%). All the segmentation sets showed very large maximum |ΔNTCP|, with rates of 9.94%, 11.89% and 13.86% for HS, DIR and DLC respectively. The |ΔNTCP| for each segmentation set were presented with categorical colour code in Fig. 2. Histogram of |ΔNTCP| for HS, DIR, DLC and SAS are shown in Fig. 3.

In total, 131 (6.5%) and 122 (6.0%) of the 2025 NTCP-predictions showed major |ΔNTCP | with DIR and DLC respectively, the frequencies of OARs involved in these NTCP models were counted and shown in Fig. 4.

The parotid glands showed the highest frequency in models of major |ΔNTCP| in both DIR (53) and DLC (54). Frequency of major |ΔNTCP| for PCM was equivalent in DIR and DLC, but for its subregions, this frequency was higher with DLC than with DIR (PCM_Superior: 42 vs. 13; PCM_Middle: 34 vs. 13; PCM_Inferior: 26 vs. 13). Buccal mucosa, brain, mandible and brainstem were not involved in NTCP models of major |ΔNTCP|. Frequency of submandibular glands in models of major |ΔNTCP| in DLC was higher than in DIR (18 vs.14). The frequencies of other OARs were all less in DLC than in DIR. Fig. 5 showed larger |ΔNTCP| in models including only parotid glands than other models.

Compared to DIR and DLC, the median and 90% percentile of |ΔNTCP| in SAS decreased to 0.33% and 1.50%, which were very close to HS (0.30% and 1.10%), the maximum |ΔNTCP| in SAS was even smaller than HS (7.01% vs. 9.94%) (Fig. 2).

By using HS as the reference, DLC showed higher dice similarity coefficient and lower 95th percentile Hausdorff distance than DIR for Crico but in the contrary way for PCM_Middle. For more geometry results please refer to supplementary material (Supplementary Data 3–6). Figure Supplementary figure 2Figure Supplementary figure 3, Supplementary figure 4 and Supplementary figure 5

Discussion

This study showed discrepant $|\Delta D_{mean}|$ and |ΔNTCP| based on segmentation of humans, DIR and DLC on rCT in a comprehensive set of OARs in head and neck cancer patients. For the segmentation of parotid glands, human segmentation performed better than DIR and DLC ($|\Delta D_{mean}|$: 1.40, 3.64 and 3.72 Gy, respectively). For the segmentation of sub-regions of PCM, DIR performed much better than DLC. However, in most other OARs, segmentation of DLC was superior to DIR and showed very close $|\Delta D_{mean}|$ to that

Table 1
Absolute difference of D_{mean} compared to gold standard in 15 OARs of 15 patients (Mean ± SD).

OAR	HS $\lvert\Delta D_{mean}\rvert$ (Gy)	DIR $\lvert\Delta D_{mean}\rvert$(Gy)	DLC $\lvert\Delta D_{mean}\rvert$(Gy)	SAS $\lvert\Delta D_{mean}\rvert$ (Gy)
Brain	0.10 ± 0.06	0.17 ± 0.09*	0.12 ± 0.18	0.12 ± 0.18
Mandible	0.16 ± 0.10	0.36 ± 0.29	0.30 ± 0.23*	0.30 ± 0.23
Oral cavity	0.29 ± 0.16	0.92 ± 0.98	0.61 ± 0.51	0.61 ± 0.51
Submandibular glands	0.64 ± 0.48	0.86 ± 0.71	0.98 ± 0.85	0.98 ± 0.85
PCM_Superior	0.97 ± 0.91	1.45 ± 1.73	2.22 ± 2.43	1.45 ± 1.73
Arytenoids	1.04 ± 0.97	2.66 ± 3.20	1.67 ± 1.51	1.67 ± 1.51
Cricopharyngeal inlet	1.05 ± 1.37	2.85 ± 2.51*	1.70 ± 1.09	1.70 ± 1.09
PCM_Middle	1.19 ± 1.45	1.82 ± 1.88	5.13 ± 4.61*	1.82 ± 1.88
Glottic area	1.23 ± 1.98	2.18 ± 2.28	1.61 ± 1.12	1.61 ± 1.12
Buccal mucosa	1.27 ± 0.98	1.79 ± 1.46	1.57 ± 2.05	1.57 ± 2.05
Supraglottic larynx	1.31 ± 2.05	2.03 ± 2.28	1.10 ± 1.18	1.10 ± 1.18
PCM_Inferior	1.31 ± 1.69	2.04 ± 2.08	1.92 ± 1.17	2.04 ± 2.08
Parotid glands	1.40 ± 0.75	3.64 ± 2.55*	3.72 ± 2.74*	1.40 ± 0.75
PCM	1.48 ± 0.71	1.47 ± 1.12	1.59 ± 0.83	1.47 ± 1.12
Brainstem	1.54 ± 0.85	1.33 ± 1.88	1.01 ± 0.93	1.01 ± 0.93
Overall	1.00 ± 1.20	1.71 ± 2.03*	1.68 ± 2.15*	1.17 ± 1.38

HS = human segmentation; DIR = deformable image registration; DLC = deep learning contouring.
SAS = semi auto-segmentation; PCM = pharyngeal constrictor muscle.
$|\Delta D_{mean}|$ = absolute difference of D_{mean} compared to HS.
*p-value < 0.025 with Related-Samples Wilcoxon Signed Rank Test compared to HS.

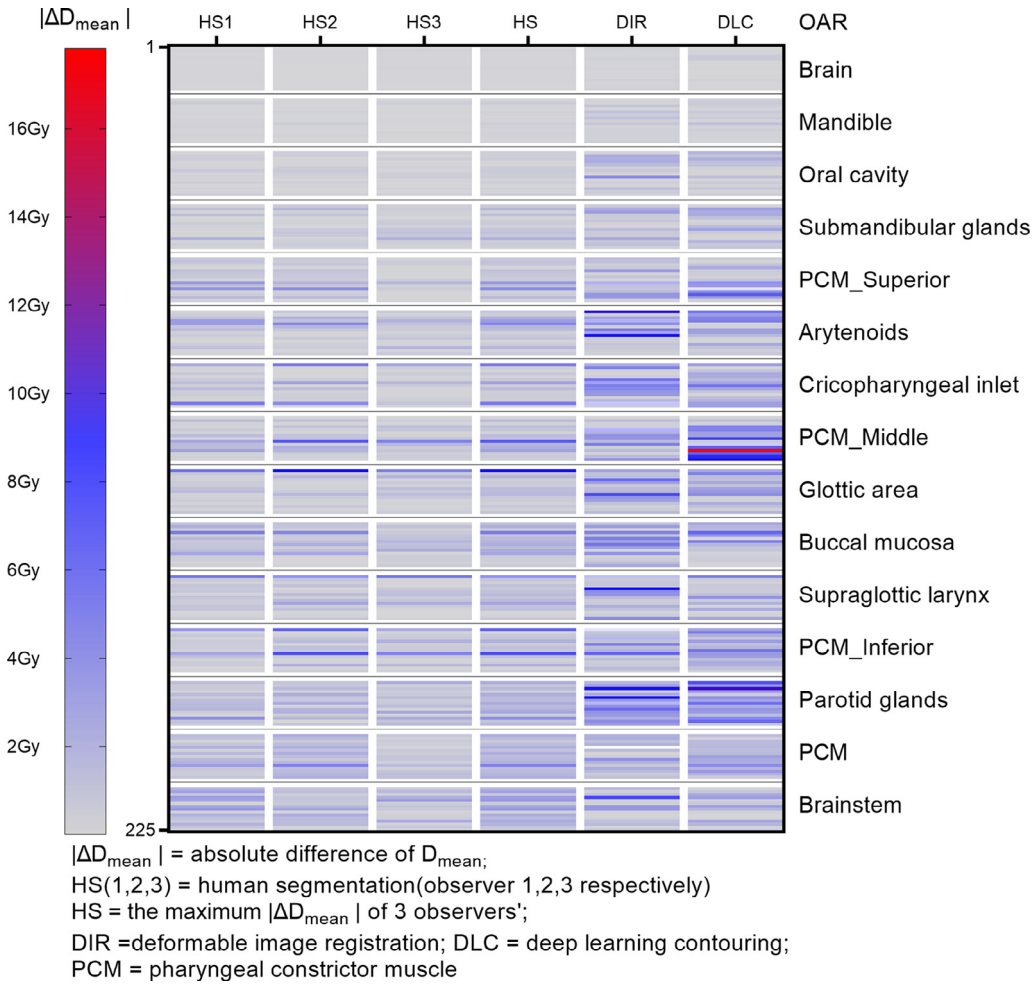


Fig. 1. Heatmap of absolute difference of D_{mean} in 15 OARs of 15 patients for HS, DIR, DLC.

observed with HS. For NTCP-predictions, parotid glands are the most common OARs involved in the NTCP-models of major |ΔNTCP|.

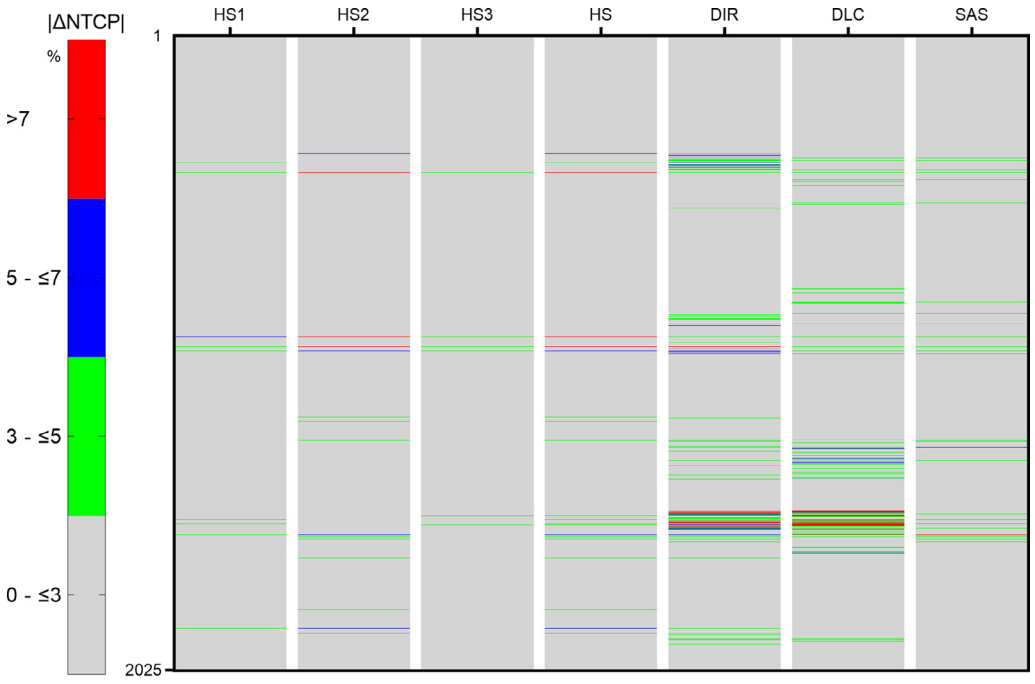
Previous studies indicated the limited impact of geometric similarity on dose in comparisons between different segmentation methods [21]. From this point of view, it is important to evaluate both changes in dose and NTCP-predictions when comparing auto-segmentation and HS. This study evaluated the absolute difference of HS, DIR and DLC regarding D_{mean} and consequent NTCP-predictions. In addition, we developed a semi auto-segmentation method by combining HS (parotid glands), DIR (PCM) and DLC (all other OARs) for fast and accurate re-segmentation of OARs and NTCP-prediction in ART.

Many studies have shown favourable geometric performance of DLC for the parotid glands [15,22,23]. Van Dijk et al. [15] observed a $|\Delta D_{mean}|$ of 0.9 ± 1.3 Gy between DLC and HS for the parotid glands. However, the $|\Delta D_{mean}|$ of DLC in the current study was 3.72 ± 2.74 Gy. This could be explained in two ways: First, previous studies revealed that the dosimetry of the parotid glands was susceptible to the variability in the segmentation of the deep lobe and geometric overlap between parotid glands and target volume [24,25], which means that even small geometric difference in the parotid glands could result in significant dosimetric difference. Second, the inter-observer variability between different HS which was referred to as the gold standard has been defined in various ways and this could also have caused different results. In current study, parotid glands showed large $|\Delta D_{mean}|$ in DIR and DLC (Table1) and most often involved in NTCP models of major |ΔNTCP| (Fig. 4), which also induced large |ΔNTCP|(Fig. 5), suggested that parotid glands should be segmented by human regarding accurate D_{mean} evaluation and NTCP-prediction.

In a previous study [15], the dice similarity coefficient of PCM and Crico between DLC and HS was 0.68 and 0.66 respectively, while the corresponding $|\Delta D_{mean}|$ was 1.3 Gy and 2.5 Gy. The median dice similarity coefficient in current study were both less than 0.6 (Supplementary data3) Figure Supplementary figure 2, but the corresponding $|\Delta D_{mean}|$ were 1.59 Gy for the PCM and 1.70 Gy for Crico, which were similar to that study.

DLC showed poor geometric performance in subregions of PCM (Supplementary data 3, 4) Figure Supplementary figure 2 and Supplementary figure 3and the largest $|\Delta D_{mean}|$ in PCM_Middle. These results were not completely consistent with others. Van Rooij et al. also observed poorer geometric performance for the esophagus, brainstem, PCM and Crico, but only found a significantly higher $|\Delta D_{mean}|$ compared to HS in the PCM_Inferior (1.4 Gy) and esophagus (2.2 Gy) [10]. This was different with the current study and could be explained partly by the variability of HS and DLC.

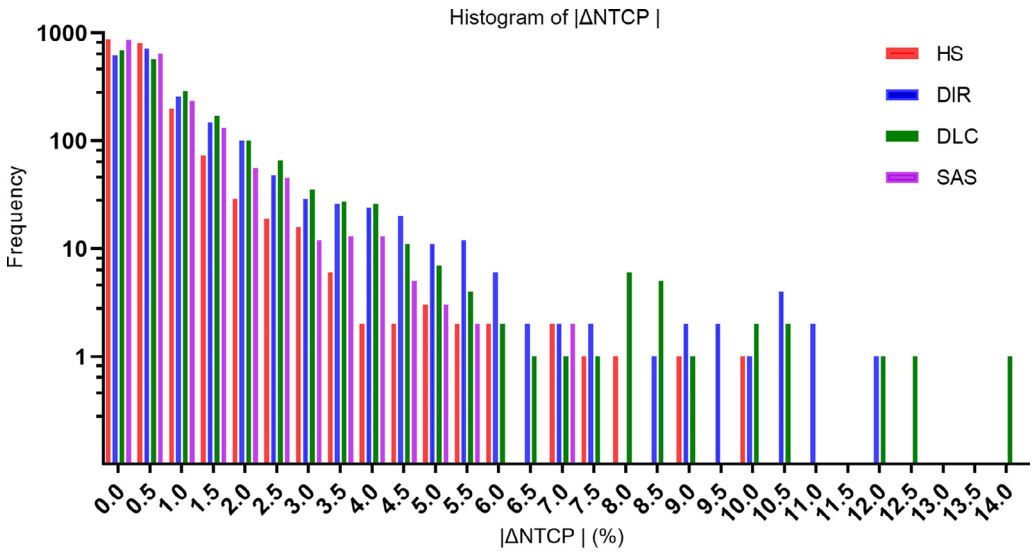
PCM and Crico are morphologically similar organs but show disparate performance between DIR and DLC. The mechanism of DIR and DLC may be accountable for this disparity. The indispensable rigid image registration before DIR makes structures propagation more adherent to the bony borders but more vulnerable to organ motion. In addition, the DLC in the current study firstly segment PCM and then divided it evenly into three sections: the



	HS1	HS2	HS3	HS	DIR	DLC	SAS
Minimum (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10th Percentile (%)	0.03	0.02	0.02	0.06	0.07	0.05	0.04
Median (%)	0.17	0.19	0.16	0.30	0.49	0.47	0.33
90th Percentile (%)	0.77	0.89	0.63	1.10	2.19	2.24	1.50
Maximum (%)	5.23	9.94	4.71	9.94	11.89	13.86	7.01

| ΔNTCP| = absolute difference of NTCP-prediction;
HS(1,2,3) = human segmentation(observer1,2,3 respectively)
HS = the maximum | ΔNTCP| of 3 observers'; DIR = deformable image registration;
DLC = deep learning contouring; SAS = semi auto-segmentation

Fig. 2. Heatmap of absolute difference of 2025 NTCP-predictions for HS, DIR, DLC and SAS.



|ΔNTCP | = absolute difference of NTCP-prediction;
HS = human segmentation; DIR = deformable image registration;
DLC = deep learning contouring; SAS = semi auto-segmentation

Fig. 3. Histogram of absolute difference of NTCP-predictions for HS, DIR, DLC and SAS.

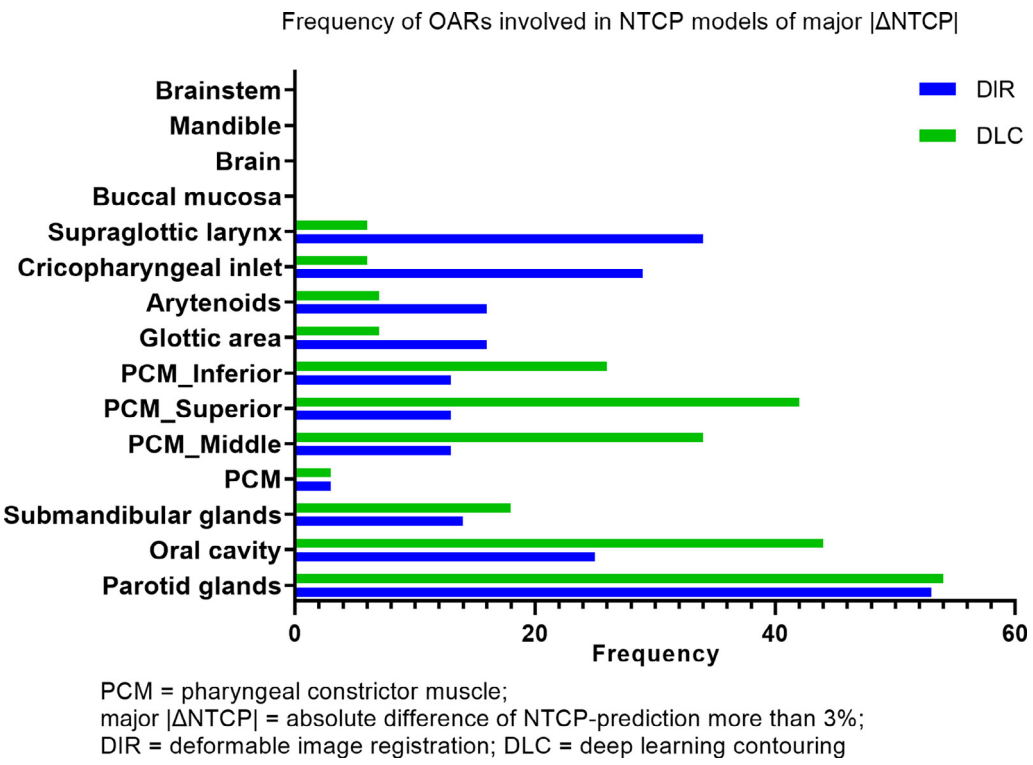


Fig. 4. Frequency of OARs involved in NTCP models showing absolute difference of NTCP-prediction more than 3%.

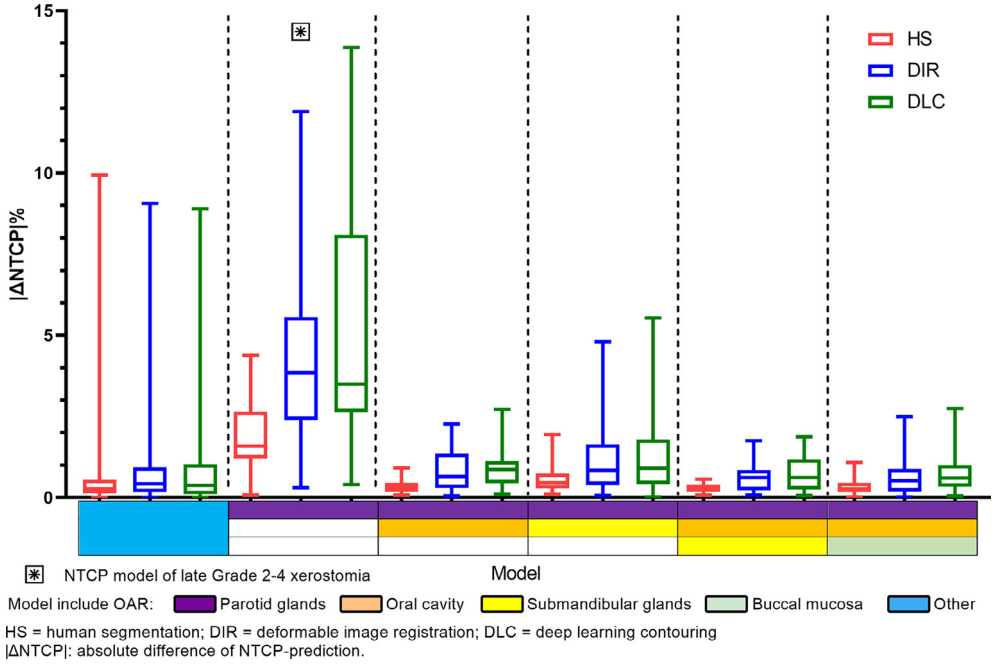


Fig. 5. Boxplot of absolute difference of NTCP-prediction for models including OARs of only parotid glands, parotid glands and others.

PCM_Superior, PCM_Middle and PCM_Inferior, regardless of the border definition within subregions of PCM. This also explained why PCM of DLC, as an aggregation, did not show such a larger $|\Delta D_{mean}|$. Results suggested this post-processing method for subregions of PCM in DLC is not fit for accurate D_{mean} evaluation and consequent NTCP-prediction. The figure in supplementary data 7 Figure Supplementary figure 6illustrates the difference between DIR and DLC in segmentation of the PCM and Crico.

Previous studies demonstrated better geometric performance of DIR in large OARs than in small OARs [11,26]. In the current study, the small volume OARs, such as the arytenoids, glottic area and supraglottic larynx showed higher $|\Delta D_{mean}|$ compared to other OARs in DIR. It has to be noted that small OARs are more susceptible to geometry change than large OARs due to the nature of D_{mean} which is also related to volume. For DLC, the $|\Delta D_{mean}|$ of other OARs was all less than 2 Gy, similar to another study [15]. Regard-

ing dosimetry, DLC performed better than DIR in all the other OARs except for the submandibular glands. Although DLC showed statistically significant $|\Delta D_{mean}|$ in the mandible, the values were very low (average $|\Delta D_{mean}|$ of DLC, DIR and HS: 0.30 Gy, 0.36 Gy and 0.16 Gy), which were only larger than the Brain.

In the current study, SAS produced comparable NTCP-predictions as HS. Fig. 2 showed the majority of major |ΔNTCP| in SAS were less than 5%, which was even better than HS.

The higher $|\Delta D_{mean}|$ values in Fig. 1 were not always present in the same horizontal position, indicating the differentiated performance of DIR and DLC on different OAR segmentations or different patients. In Fig. 2, major |ΔNTCP| were likely to present in the same horizontal position (NTCP model), indicating that the |ΔNTCP| was model- or patient-specific, or OAR- or disease-specific. Therefore, further studies are needed to develop individualized SASs for patients of different features.

Several shortcomings of the current study should be mentioned. First, the proposed SAS was evaluated based on the NTCP-models including only D_{mean} from the planning technique of swallowing sparing IMRT and VMAT, which means it may be inapplicable for other treatment planning techniques and NTCP models with non-D_{mean} parameters. Second, the small sample size might reduce the representativeness of conclusions, however this is a trade off against having multiple human segmentations. Third, the selected rCTs of large geometry change could weaken the performance of DIR while having no impact on DLC, but the result could be more beneficial for ART. It should be noted the gold standard applied in this study was defined as the average value of 3 observers', therefore the calculated $|\Delta D_{mean}|$ of HS was actually smaller than the real inter-observer difference (more information on inter-observer difference in Supplementary data 8, 9) Figure Supplementary figure 7. Figure Supplementary figure 1

For application of SAS in ART in clinical practice, it remains necessary to identify outliers in SAS by human or automated anomaly detection for a robust and high-quality result.

Conclusions

On repeat CT-scans of head and neck cancer patients presenting large geometry changes compared to the planning CT-scan, DLC outperformed DIR segmentation for evaluation of mean dose for the majority of OARs. Human segmentation of parotid glands remains necessary for accurate interpretation of mean dose and NTCP-prediction during ART. The proposed semi auto-segmentation allows NTCP-prediction within 1.5% accuracy for at least 90% of the cases.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Leonardo van der Wijk and Wouter Rutgers to carry out human segmentations that we used for comparison. We also want to thank Lisa Van den Bosch for guidance and assistance on NTCP calculation, thank H.R. van de Glind and Sanne van Dijk for assistance in geometry evaluation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2021.09.019.

References

[1] Barker JL, Garden AS, Ang KK, O'Daniel JC, Wang He, Court LE, et al. Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system. Int J Radiat Oncol Biol Phys 2004;59:960–70. https://doi.org/10.1016/j.ijrobp.2003.12.024.

[2] Noble DJ, Yeap P-L, Seah SYK, Harrison K, Shelley LEA, Romanchikova M, et al. Anatomical change during radiotherapy for head and neck cancer, and its effect on delivered dose to the spinal cord. Radiother Oncol 2019;130:32–8. https://doi.org/10.1016/j.radonc.2018.07.009.

[3] Marzi S, Pinnarò P, D'Alessio D, Strigari L, Bruzzaniti V, Giordano C, et al. Anatomical and dose changes of gross tumour volume and parotid glands for head and neck cancer patients during intensity-modulated radiotherapy: effect on the probability of xerostomia incidence. Clin Oncol 2012;24:e54–62. https://doi.org/10.1016/j.clon.2011.11.006.

[4] Bertholet J, Anastasi G, Noble D, Bel A, van Leeuwen R, Roggen T, et al. Patterns of practice for adaptive and real-time radiation therapy (POP-ART RT) part II: Offline and online plan adaption for interfractional changes. Radiother Oncol 2020;153:88–96. https://doi.org/10.1016/j.radonc.2020.06.017.

[5] Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. Med Phys 2020;47:e929–50. https://doi.org/10.1002/mp.14320.

[6] Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. Sem Radiat Oncol 2019;29:185–97. https://doi.org/10.1016/j.semradonc.2019.02.001.

[7] L. Boldrini, J. E. Bibault, C. Masciocchi, Y. Shen, M.I. Bittner, Deep Learning: A Review for the Radiation Oncologist, Front Oncol 2019:9, doi: 10.3389/fonc.2019.00977.

[8] 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.

[9] 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 2019;29:1961–7. https://doi.org/10.1007/s00330-018-5748-9.

[10] van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, Verbakel WF. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. Int J Radiat Oncol Biol Phys 2019;104:677–84. https://doi.org/10.1016/j.ijrobp.2019.02.040.

[11] Kumarasiri A, et al., Deformable image registration based automatic CT-to-CT contour propagation for head and neck adaptive radiotherapy in the routine clinical setting, Med Phys 2014;41, doi: 10.1118/1.4901409.

[12] Hardcastle N, Tomé WA, Cannon DM, Brouwer CL, Wittendorp PWH, Dogan N, et al. A multi-institution evaluation of deformable image registration algorithms for automatic organ delineation in adaptive head and neck radiotherapy. Radiat Oncol 2012;7. https://doi.org/10.1186/1748-717X-7-90.

[13] Chetty IJ, Rosu-Bubulac M. Deformable registration for dose accumulation. Semin Radiat Oncol 2019;29(3):198–208. https://doi.org/10.1016/j.semradonc.2019.02.002.

[14] Dowling JA, O'Connor LM. Deformable image registration in radiation therapy. J Med Radiat Sci 2020;67:257–9. https://doi.org/10.1002/jmrs.v67.410.1002/jmrs.446.

[15] van Dijk LV, Van den Bosch L, Aljabar P, Peressutti D, Both S, J.H.M. Steenbakkers R, et al. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. Radiother Oncol 2020;142:115–23. https://doi.org/10.1016/j.radonc.2019.09.022.

[16] Joskowicz L, Cohen D, Caplan N, Sosna J. Inter-observer variability of manual contour delineation of structures in CT. Eur Radiol 2019;29:1391–9. https://doi.org/10.1007/s00330-018-5695-5.

[17] Brouwer CL, Steenbakkers RJHM, Langendijk JA, Sijtsema NM. Identifying patients who may benefit from adaptive radiotherapy: Does the literature on anatomic and dosimetric changes in head and neck organs at risk during radiotherapy provide information to help? Radiother Oncol 2015;115:285–94. https://doi.org/10.1016/j.radonc.2015.05.018.

[18] 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.

[19] Ahn H et al. Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. Physiol Behav 2017;176:139–48. https://doi.org/10.1002/mp.13141.Auto-segmentation.

[20] Van den Bosch L, van der Schaaf A, van der Laan HP, Hoebers FJP, Wijers OB, van den Hoek JGM, et al. Comprehensive toxicity risk profiling in radiation therapy for head and neck cancer: a new concept for individually optimised treatment. Radiother Oncol. 2021;157:147–54. https://doi.org/10.1016/j.radonc.2021.01.024.

[21] Kosmin M, Ledsam J, Romera-Paredes B, Mendes R, Moinuddin S, de Souza D, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiother Oncol 2019;135:130–40. https://doi.org/10.1016/j.radonc.2019.03.004.

[22] W. Zhu et al., AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy, arXiv. 2018.

[23] Tappeiner E, Pröll S, Hönig M, Raudaschl PF, Zaffino P, Spadea MF, et al. Multi-organ segmentation of the head and neck area: an efficient hierarchical neural networks approach. Int J Comput Assist Radiol Surg 2019;14:745–54. https://doi.org/10.1007/s11548-019-01922-4.

[24] Millunchick CH, Zhen H, Redler G, Liao Y, J., v., Turian,. A model for predicting the dose to the parotid glands based on their relative overlapping with planning target volumes during helical radiotherapy. J Appl Clin Med Phys 2018;19:48–53. https://doi.org/10.1002/acm2.12203.

[25] Loo SW, Martin WMC, Smith P, Cherian S, Roques TW. Interobserver variation in parotid gland delineation: A study of its impact on intensity-modulated radiotherapy solutions with a systematic review of the literature. Br J Radiol 2012;85:1070–7. https://doi.org/10.1259/bjr/32038456.

[26] L. Zhang, Z. Wang, C. Shi, T. Long, X.G. Xu, The impact of robustness of deformable image registration on contour propagation and dose accumulation for head and neck adaptive radiotherapy, 2018;185–194. doi: 10.1002/acm2.12361.