Consistency in contouring of organs at risk by artificial intelligence vs oncologists in head and neck cancer patients

Nielsen, Camilla Panduro; Lorenzen, Ebbe Laugaard; Jensen, Kenneth; Sarup, Nis; Brink, Carsten; Smulders, Bob; Holm, Anne Ivalu Sander; Samsøe, Eva; Nielsen, Martin Skovmos; Sibolt, Patrik; Skyt, Peter Sandegaard; Elstrøm, Ulrik Vindelev; Johansen, Jørgen; Zukauskaite, Ruta; Eriksen, Jesper Grau; Farhadi, Mohammad; Andersen, Maria; Maare, Christian; Overgaard, Jens; Grau, Cai; Friborg, Jeppe; Hansen, Christian Rønn

Published in:
Acta Oncologica

DOI:
10.1080/0284186X.2023.2256958

Publication date:
2023

Document version:
Accepted manuscript

Citation for pulished version (APA):
Nielsen, C. P., Lorenzen, E. L., Jensen, K., Sarup, N., Brink, C., Smulders, B., Holm, A. I. S., Samsøe, E., Nielsen, M. S., Sibolt, P., Skyt, P. S., Elstrøm, U. V., Johansen, J., Zukauskaite, R., Eriksen, J. G., Farhadi, M., Andersen, M., Maare, C., Overgaard, J., ... Hansen, C. R. (2023). Consistency in contouring of organs at risk by artificial intelligence vs oncologists in head and neck cancer patients. *Acta Oncologica*, *62*(11), 1418-1425. https://doi.org/10.1080/0284186X.2023.2256958

Consistency in Contouring of Organs at Risk by Artificial Intelligence vs Oncologists in Head and Neck Cancer Patients

Camilla Panduro Nielsen[ab]*, Ebbe Laugaard Lorenzen[ab], Kenneth Jensen[d], Nis Sarup[a], Carsten Brink[ab], Bob Smulders[de], Anne Ivalu Sander Holm[f], Eva Samsøe[dg], Martin Skovmos Nielsen[h], Patrik Sibolt[i], Peter Sandegaard Skyt[d], Ulrik Vindelev Elstrøm[d], Jørgen Johansen[c], Ruta Zukauskaite[bc], Jesper Grau Eriksen[fj], Mohammad Farhadi[g], Maria Andersen[h], Christian Maare[i], Jens Overgaard[j], Cai Grau[d], Jeppe Friborg[de] and Christian Rønn Hansen[abd]

[a]*Laboratory of Radiation Physics, Odense University Hospital, Denmark;*
[b]*Institute of Clinical Research, University of Southern Denmark, Denmark;*
[c]*Department of Oncology, Odense University Hospital, Denmark;*
[d]*Danish Centre of Particle Therapy, Aarhus University Hospital, Denmark;*
[e]*Department of Oncology, Rigshospitalet, University Hospital of Copenhagen, Denmark;*
[f]*Department of Oncology, Aarhus University Hospital, Denmark;*
[g]*Department of Oncology, Zealand University Hospital, Naestved, Denmark;*
[h]*Department of Oncology, Aalborg University Hospital, Denmark;*
[i]*Department of Oncology, University Hospital Herlev, Denmark;*
[j]*Department of Experimental Clinical Oncology, Aarhus University Hospital, Denmark*

*Camilla Panduro Nielsen Camilla.Panduro.Nielsen@rsyd.dk

Words: 3106

Consistency in Contouring of Organs at Risk by Artificial Intelligence vs Oncologists in Head and Neck Cancer Patients

Background: In the Danish Head and Neck Cancer Group (DAHANCA) 35 trial, patients are selected for proton treatment based on simulated reductions of Normal Tissue Complication Probability (NTCP) for proton compared to photon treatment at the referring departments. After inclusion in the trial, immobilization, scanning, contouring and planning are repeated at the national proton centre. The new contours could result in reduced expected NTCP gain of the proton plan, resulting in a loss of validity in the selection process. The present study evaluates if contour consistency can be improved by having access to AI (Artificial Intelligence) based contours.

Materials and Methods: The 63 patients in the DAHANCA 35 pilot trial had a CT from the local DAHANCA centre and one from the proton centre. A nationally validated convolutional neural network, based on nnU-Net, was used to contour OARs on both scans for each patient. Using deformable image registration, local AI and oncologist contours were transferred to the proton centre scans for comparison. Consistency was calculated with the Dice Similarity Coefficient (DSC) and Mean Surface Distance (MSD), comparing contours from AI to AI and oncologist to oncologist, respectively. Two NTCP models were applied to calculate NTCP for xerostomia and dysphagia.

Results: The AI contours showed significantly better consistency than the contours by oncologists. The median and interquartile range of DSC was 0.85 [0.78 – 0.90] and 0.68 [0.51 – 0.80] for AI and oncologist contours, respectively. The median and interquartile range of MSD was 0.9 mm [0.7 – 1.1] mm and 1.9 mm [1.5 – 2.6] mm for AI and oncologist contours, respectively. There was no significant difference in ΔNTCP.

Conclusions: The study showed that OAR contours made by the AI algorithm were more consistent than those made by oncologists. No significant impact on the ΔNTCP calculations could be discerned.

Keywords: AI; contouring; organs at risk; head and neck cancer; proton treatment

Background

Modern radiation treatment is complex and involves many manual and time-consuming procedures. Even with national and international guidelines, contouring of cancer targets and organs at risk (OARs) in head and neck (H&N) cancer patients varies between treatment centres and clinical experts [1-3]. Variability in the contouring of OARs has a significant dosimetric impact on patient treatment [2,4] and may influence the results of clinical trials [5].

In the Danish Head and Neck Cancer Group (DAHANCA) 35 trial (NCT04607694), H&N cancer patients are selected for proton treatment based on the simulated benefit in terms of Normal Tissue Complication Probability (NTCP) for xerostomia and dysphagia [6]. Contouring of OARs is part of the basis for treatment planning and, consequently, for the NTCP estimations. Thus, the accuracy and consistency of the OAR contours potentially affect the patient selection process for the DAHANCA 35 trial.

To test the feasibility and safety of proton treatment, the DAHANCA 35 pilot trial (NCT05423704) was conducted [7]. All patients selected for the DAHANCA 35 pilot trial received proton treatment, whereas the patients selected for the DAHANCA 35 trial will be randomised for proton or photon treatment [6]. Results from the DAHANCA 35 pilot trial showed that the NTCP estimates, based on contours and treatment plans from different treatment centres, had variations that could be related to target contouring. There were also indications that the variation in OAR contours could play an important role in the robustness of patient selection [7].

Artificial Intelligence (AI) has been investigated as a useful tool in radiotherapy [8,9], and previous studies implemented AI algorithms for auto-segmentation of OARs on computed tomography (CT) scans in H&N cancer treatment, showing improved efficiency and standardisation of treatment [10,11]. AI has also shown good

performance in contouring OARs compared to oncologist contours as ground truth [12,13]. It is therefore hypothesised that AI segmentation of OARs could improve the patient selection robustness for the DAHANCA 35 trial.

This study aimed to quantify the consistency and variation in the contouring by oncologists of relevant OARs on two different CT scans for the same patient. This was compared to the consistency in contours performed by an AI segmentation algorithm on the same CT scans. Secondly, the impact on estimated NTCP using AI contours of OARs compared to contours made by oncologists was assessed to evaluate the clinical relevance of the difference in contouring consistency between AI and oncologists.

Materials and Methods

From May 2019 to March 2021, 63 patients were included in the DAHANCA 35 pilot trial [6,7]. Each patient was diagnosed with squamous cell carcinoma of the pharynx or larynx at a local DAHANCA centre. Before potential inclusion in the trial, the following was performed at the local centre: a CT scan (local CT scan), contouring of the target volumes and OARs by radiation oncologists (local oncologist contours), as well as treatment planning for both photon treatment (local photon plan) and proton treatment (local proton plan).

Two NTCP models, one for xerostomia grade 2+ and one for dysphagia grade 2+, validated in a Danish cohort [14], were used to estimate the NTCP [15] for the local photon and proton treatment plans. If the difference in estimated NTCP (ΔNTCP) for the local photon plan compared to the local proton plan was larger than 5 %-point for either xerostomia, dysphagia or both, the patient was offered inclusion in the trial and, on informed consent, referred to the national proton centre for proton treatment.

At the proton centre, a new proton therapy compatible immobilisation mask and a new CT scan were made (clinical CT scan), as well as new contours (clinical oncologist contours) and a new proton treatment plan (clinical proton plan).

The present study is a retrospective analysis of contouring consistency and ΔNTCP, and thus does not influence the course of treatment of the patients included in the DAHANCA 35 pilot trial.

Artificial Intelligence Segmentation Model

The AI model used for segmentation was a Convolutional Neural Network (CNN) based on the nnU-Net model presented by Isensee et al. [16]. The AI model was trained on CT scans from a national trial [14] and relevant contours of OARs in H&N cancer recontoured by radiation oncologists following international standards [17]. The model performance was validated on OARs contoured by H&N oncology specialists from a Danish national workshop [18]. The model was used to retrospectively contour 12 OARs relevant for H&N cancer on both local and clinical CT scans for all patients from the DAHANCA 35 pilot trial (local AI contours and clinical AI contours, respectively). The OARs available from the model were: extended oral cavity, upper-, middle-, and lower constrictor muscles, glottic larynx, supraglottic larynx, left and right parotid, left and right submandibular, thyroid, and oesophagus.

Data Analyses

The data analyses were performed on data from the 63 patients in the pilot trial. Local AI contours on local CT scans were compared to clinical AI contours on clinical CT scans, and local oncologist contours on local CT scans were compared to clinical oncologist contours on clinical CT scans.

Data pre-processing and statistical analyses were performed in MATLAB R2022b.

Data Pre-Processing

One patient was missing the contour of the left submandibular on one of the CT scans, and another patient was missing contours of both the left and right submandibular. Thus, contours from 61 patients were used for the statistical comparison of the left submandibular and 62 for the right submandibular. Contours from all 63 patients were used for the remaining 10 OARs.

Using MIM software, Deformable Image Registration (DIR) was performed [19], transferring local AI and oncologist contours to the clinical CT scan, transferring both sets of contours using the same DIR for each patient. The DIR process in MIM first uses a rigid registration, then a coarse-to-fine multi-resolution approach, and finally, a custom-modified gradient descent for optimisation [20]. The result of the DIR was accepted based on visual inspection of the deform fusion alignment in MIM.

Contour Overlap

Contouring consistency was measured by contour overlap in terms of Dice Similarity Coefficient (DSC) [21,22] and Mean Surface Distance (MSD) [23]. The higher the DSC and the lower the MSD for a contour comparison, the better the consistency.

The oncologists contour oesophagus in the caudal direction until it is no longer deemed clinically relevant, concerning the radiotherapy treatment plan. However, as the dose tolerance is the mean dose, it can influence the plan. The oesophagus contours were analysed as directly contoured and after correcting the contours to have the same caudal length i.e., when comparing two contours of oesophagus, the contour with the caudal

part, most cranial determined the caudal length, and slices below that point were removed for the other contour.

Normal Tissue Complication Probability

NTCP is the foundation for patient selection in the DAHANCA 35 trial, and preliminary results from the pilot trial showed a disparity in ΔNTCP when comparing the local ΔNTCP ($NTCP_{local\ photon\ plan} - NTCP_{local\ proton\ plan}$) to the clinical ΔNTCP ($NTCP_{local\ photon\ plan} - NTCP_{clinical\ proton\ plan}$) [7]. NTCP for xerostomia grade 2+ and dysphagia grade 2+ was estimated using the Dutch models for selection [15]. The model for xerostomia included baseline xerostomia and mean dose to the contralateral parotid [24], the model for dysphagia included baseline dysphagia and mean dose to the upper pharyngeal constrictor muscle and extended oral cavity (regression coefficients in Supplementary Table 1).

In the present study, NTCP was calculated based on the original contours made by oncologists and the corresponding photon and proton plans, as described by Hansen et al. [7]. Additionally, NTCP was calculated based on the AI contours on the original photon and proton plans. The treatment plans were not optimised for the AI contours, as these were delineated retrospectively after patient treatment.

The disparity in local and clinical ΔNTCP was visualized using a scatterplot and the variation using a Bland-Altman plot.

Statistical Analyses

Wilcoxon Signed Rank test for non-parametric data was used to compare the consistency and ΔNTCP between AI and oncologist contours for each patient, using a significance level of 5 %.

Results

Contour Overlap

In terms of DSC and MSD, the AI contours showed significantly better consistency than the contours by the oncologists. The median and interquartile range of DSC across all 12 OARs was 0.85 [0.78 – 0.90] and 0.68 [0.51 – 0.80] for AI and oncologist contours, respectively. The median and interquartile range of MSD for all OARs was 0.9 mm [0.7 – 1.1] mm and 1.9 mm [1.5 – 2.6] mm for AI and oncologist contours, respectively. The DSC and MSD for the individual OARs are collected in Table 1. Comparing the DSC between AI and oncologists for each OAR, the DSC was significantly larger for AI contours. All p-values were significant ($p < 10^{-5}$). The MSD for all OARS was significantly lower for AI contours compared to oncologist contours. All p-values were significant ($p < 10^{-5}$). As seen in Table 1, the AI contours of oesophagus were still significantly more consistent than oncologist contours after correction to have the same caudal length ($p < 10^{-10}$).

	DSC AI		DSC oncologists		MSD AI		MSD oncologists	
	[index]	IQR	[index]	IQR	[mm]	IQR	[mm]	IQR
Extended oral cavity	0.95	0.94 - 0.95	0.84	0.81 - 0.89	0.8	0.7 - 1.0	2.2	1.7 - 2.9
PCM up	0.79	0.76 - 0.81	0.40	0.31 - 0.51	0.8	0.7 - 0.9	2.4	2.0 - 3.0
PCM mid	0.76	0.72 - 0.79	0.51	0.41 - 0.59	0.8	0.7 - 0.9	2.0	1.5 - 2.6
PCM low	0.75	0.70 - 0.79	0.50	0.39 - 0.60	0.9	0.7 - 1.1	2.3	1.7 - 3.5
Glottic larynx	0.79	0.76 - 0.82	0.45	0.35 - 0.57	0.8	0.6 - 0.9	2.2	1.7 - 3.0
Suppraglottic larynx	0.88	0.86 - 0.89	0.58	0.48 - 0.64	0.9	0.8 - 1.1	1.9	1.6 - 2.6
Parotid left	0.91	0.89 - 0.92	0.82	0.77 - 0.85	0.9	0.8 - 1.1	1.7	1.6 - 2.3
Parotid right	0.93	0.92 - 0.93	0.82	0.78 - 0.84	0.8	0.7 - 0.9	1.7	1.5 - 2.1
Submandibular left	0.86	0.83 - 0.88	0.78	0.74 - 0.82	1.0	0.9 - 1.2	1.4	1.1 - 1.8
Submandibular right	0.89	0.86 - 0.90	0.78	0.74 - 0.81	0.8	0.8 - 1.0	1.5	1.2 - 1.9
Thyroid	0.83	0.78 - 0.86	0.72	0.67 - 0.78	1.2	0.9 - 1.4	1.7	1.4 - 2.2
Oesophagus	0.81	0.78 - 0.85	0.63	0.55 - 0.71	1.1	0.9 - 1.6	3.3	2.1 - 6.2
Oesophagus*	0.84	0.81 - 0.86	0.72	0.67 - 0.77	0.8	0.7 - 1.0	1.9	1.3 - 2.6

Table 1: DSC and MSD for 12 OARs contoured by AI oncologists, respectively. PCM up: upper pharyngeal constrictor muscle, PCM mid: middle pharyngeal constrictor

muscle, PCM low: lower pharyngeal constrictor muscle. Interquartile range (IQR).
* Oesophagus after correction so the contours have the same caudal length.

Figure 1 shows the DSC represented as box plots with raw data points overlaid for all 12 OARs, and Figure 2 shows the MSD. Supplementary Figure 1 and 2 shows the DSC and MSD, respectively, for the oesophagus with and without correction to have the same caudal length. When correcting for the same length, the median DSC and MSD match the other OARs.

Normal Tissue Complication Probability

Considering the local ΔNTCP compared to the clinical ΔNTCP, no significant difference was found between the ΔNTCP calculated based on AI contours and ΔNTCP calculated on oncologist contours.

Figure 3**Fejl! Henvisningskilde ikke fundet.** shows a scatter plot with the local ΔNTCP on the x-axis and the clinical ΔNTCP on the y-axis. The ΔNTCP is shown in %-point. The dotted black line shows the identity line. If there was no difference between the local and clinical ΔNTCP, all samples would be on the dashed identity line.

Figure 4 shows a Bland-Altman plot of the mean between the local and clinical ΔNTCP on the x-axis and the difference between the local and clinical ΔNTCP on the y-axis. The ΔNTCP is shown in %-point.

For the model used in the DAHANCA 35 pilot trial, the median difference and interquartile range in ΔNTCP for xerostomia was 0 %-point [-2 – 3] %-points for AI and 0 %-point [-1 – 4] %-points for oncologists. The p-value was not significant ($p = 0.45$). For dysphagia, the median difference in ΔNTCP was 1%-point [-1 – 4] for AI and 1%-point [-1 – 4] for oncologists. The p-value was not significant ($p = 0.72$).

Discussion

The results of the present study showed that contours generated by the AI segmentation algorithm were significantly more consistent than contours made by oncologists. The AI contours investigated in the present study are not adjusted by oncologists, which they would be if used for patient treatment. Furthermore, different centres might have slightly different procedures when working with AI contours. This could alter the consistency between contours, but it would presumably still be more harmonised, as all oncologists would use a more consistent starting point. Implementing an AI segmentation algorithm, with consistency as shown in this study, would therefore introduce less inter-observer variability in a clinical trial, assuming that the post-correction done by oncologists would be limited.

The contouring consistency was lower for oncologists than for AI across all OARs; however, the consistency for oncologists was highest in the OARs that have the longest history of guidelines and where the oncologist interpretation has been discussed over the years (i.e. extended oral cavity, left and right parotid, left and right submandibular, and thyroid) [25-27]. This is especially evident in Figure 1, showing the DSC. The consistency for contours by oncologists was lower for OARs implemented in the guidelines most recently, like the glottic larynx, supraglottic larynx, and constrictor muscles.

The variation in consistency in the oesophagus contouring mainly depends on the length of the contoured organ as determined by the oncologist. The AI contours were still significantly more consistent after correcting the oesophagus contours to have the same caudal length. The corrected contours might give a more fair comparison, as the length of oesophagus outside of what is clinically relevant is less important. The AI contours of oesophagus without correction were also significantly more consistent than

the oncologist contours after correction; thus, using these would give a more representative mean dose.

The contrast on CT scans differs for different OARs, requiring an individual oncologist's specific interpretation of anatomy. This could explain some of the lack of consistency for oncologists, as some OARs might be more difficult to distinguish. Here the AI algorithm places a typical segmentation that matches the patient in shape and size. The algorithm works in 3D, whereas the oncologist works in 2D in the three different planes, which again could explain why the AI contours are more consistent.

The lower consistency in contours made by oncologists supports the statement that even with national and international guidelines [17,27], there is a gap between what has been generally accepted and what is practically performed at different treatment centres [28]. In Denmark, every treatment centre adheres to the same guidelines [27], but even then, this study indicates that the interpretation and execution differ across the country. Implementing AI for contouring could reduce the gap between guidelines, interpretation, and execution. Contouring at the national proton centre is always conducted using an MR scan performed in the treatment position in addition to the planning CT scan. Although it is recommended to use MR [29], it is not always acquired at the local centres, and the use of MR for contouring OARs is not always used. This difference may explain some of the inconsistencies in contours between oncologists.

The results on consistency in this study are a combination of contouring consistency, DIR, and differences in procedures between local and clinical centres. The OARs were investigated in terms of volume change before and after DIR; for most OARs, the volume changed by approximately 10%. This could be because of the DIR process, but also differences in scanning procedures between local centres and the

proton centre, where the local CT scans are always performed with contrast and the proton centre CT scans are without. However, each set of local oncologist and AI contours was transferred using the same DIR process, which means that the potential change due to the DIR process was applied to both. Therefore it was assumed not to alter the overall conclusion of the study.

A change in OAR contouring will affect the treatment planning, which in turn can affect estimations of NTCP [2]. Brouwer et al. investigated the effect of differences in delineation on resulting NTCP estimations. They found little NTCP differences in the majority of patients and large NTCP differences of $> 10\%$ in a few patients [2]. For a clinical trial like DAHANCA 35, utilising patient selection based on NTCP estimates, consistency in NTCP is important. Results from the DAHANCA 35 pilot trial, showed that for patients selected for a specific toxicity, the mean ΔNTCP for xerostomia and dysphagia was significantly reduced from the local centre to the national proton centre [7]. The mean local ΔNTCP for xerostomia was 7.3 %-point, and the mean clinical ΔNTCP was 4.9 %-point. For dysphagia, the mean local ΔNTCP was 6.9 %-point, and the mean clinical ΔNTCP was 5.3 %-point [7]. The present study did not show significant differences in ΔNTCP between contours by oncologists and AI, potentially because the treatment plans were not optimised according to the AI contours. It would be expected that improving the contour consistency of OARs and target as well as optimising treatment plan quality, would result in more consistent NTCP estimates. The consistency of OAR contours could be improved by using AI, as suggested in this study. Even though DAHANCA guidelines have already improved the contouring consistency of clinical target volumes [3], it could be further enhanced by implementing AI for segmentation of target volumes as a starting point for oncologists [30]. Furthermore, optimising treatment plan quality to spare OARs could be done using

automated and knowledge-based treatment planning tools [31-33]. The field of AI continues to develop, and better segmentation models will likely be developed for contouring for both OARs and cancer targets, thus improving consistency between treatment centres. Similarly, dose prediction AI algorithms [34-36] will potentially help improve NTCP estimates. The dose distribution can be predicted without simulating the full complex photon or proton plan, presumably increasing the consistency.

The results of this study do not indicate whether contours made by AI or oncologists are more correct, only that AI contours are more consistent for the same patient. Before implementing an AI model for the segmentation of OARs, it should be investigated if the AI model performs to a clinically acceptable standard. This was not investigated in this study; however, the current AI model performance was investigated in a study by Lorenzen et al. [37], who found that it performed as well as, or better than, the expert oncologists for almost all OARs investigated here. An exception was the upper pharyngeal constrictor muscle, where the model was trained on segmentation from a vague definition of the upper pharyngeal constrictor muscle. For this reason, the AI model is being updated.

Higher consistency in contouring would contribute to increasing the chance of more consistent treatment planning across treatment centres, influencing NTCP estimates. In combination with improved target contouring, it would thus potentially result in improved patient selection for the trial, potentially improving the overall outcome of the trial. In general, AI OAR segmentation could provide a common starting point which, in the long run, could lead to harmonised treatment procedures and improve the local selection of patients for appropriate treatment, independent of local expertise and workload, and hence improve equality in health care.

This study investigated the consistency in contouring when using AI, but AI may also be useful for Quality Assurance (QA) of clinical trials and clinical practice [36]. QA could be implemented like in this study, where AI constitutes a second opinion to consider, or as a tool for decision support to form the basis for oncologist contouring. It could also be used directly for QA of the AI-generated contours [38].

In conclusion, AI is more consistent for segmentation of OARs in H&N cancer patients compared to oncologist contours. However, the more consistent contours did not translate into more consistent ΔNTCP estimates.

Funding

Supported by the Novo Nordisk Foundation (NNF18OC0034612), DCCC Radiotherapy - The Danish National Research Center for Radiotherapy, Danish Cancer Society (grant no. R191-A11526), Danish Comprehensive Cancer Center, University of Southern Denmark Faculty of Health Sciences Scholarship, and Odense University Hospital.

Data Availability Statement

The data used in this study is part of a clinical trial and is not available.

Figure Legends

Figure 1: Box plot with individual samples overlaid showing the DSC for the 12 OARs. Green boxes and samples show the DSC for the AI contours, and blue boxes and samples are results comparing oncologist contours. The raw data points are shown to visualise the distribution.

Figure 2: Box plot with individual samples overlaid showing the MSD for the 12 OARs. Green boxes and samples show the MSD for the AI contours, and blue boxes and samples are results comparing oncologist contours. The raw data points are shown to

visualise the distribution. For visualisation, the plot has been scaled, omitting two outliers from oesophagus and one from glottic larynx for oncologist contours.

Figure 3: Scatter plot of the local ΔNTCP ($NTCP_{local\ photon\ plan} - NTCP_{local\ proton\ plan}$) and clinical ΔNTCP ($NTCP_{local\ photon\ plan} - NTCP_{clinical\ proton\ plan}$) based on AI (green data points) and oncologist (blue data points) contours, respectively, for xerostomia and dysphagia.

Figure 4: Bland-Altman plot showing the mean and difference between the local and clinical ΔNTCP for xerostomia and dysphagia. The green data points represent the ΔNTCP based on AI contours, and the blue data points represent the ΔNTCP calculated based on oncologist contours.

References

1. Brouwer CL, Steenbakkers RJHM, van den Heuvel E, et al. 3D Variation in delineation of head and neck organs at risk. Radiation Oncology. 2012 2012/03/13;7(1):32.
2. Brouwer CL, Steenbakkers RJ, Gort E, et al. Differences in delineation guidelines for head and neck cancer result in inconsistent reported dose and corresponding NTCP. Radiother Oncol. 2014 Apr;111(1):148-52.
3. Hansen CR, Johansen J, Samsøe E, et al. Consequences of introducing geometric GTV to CTV margin expansion in DAHANCA contouring guidelines for head and neck radiotherapy. Radiother Oncol. 2018 Jan;126(1):43-47.
4. Voet PW, Dirkx ML, Teguh DN, et al. 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 Mar;98(3):373-7.
5. Peters LJ, O'Sullivan B, Giralt J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. J Clin Oncol. 2010 Jun 20;28(18):2996-3001.
6. Friborg J, Jensen K, Eriksen JG, et al. Considerations for study design in the DAHANCA 35 trial of protons versus photons for head and neck cancer. Radiotherapy and Oncology. 2023;Under review.
7. Hansen CR, Jensen K, Smulders B, et al. Evaluation of decentralised model-based selection of head and neck cancer patients for a proton treatment study. DAHANCA 35. Radiother Oncol. 2023 Jul 20:109812.
8. Chufal KS, Ahmad I, Chowdhary RL. Artificial intelligence in radiation oncology: How far have we reached? International Journal of Molecular and Immuno Oncology. 2023;8.
9. Lim JY, Leech M. Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. Acta Oncologica. 2016 2016/07/02;55(7):799-806.

10. van der Veen J, Willems S, Deschuymer S, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. Radiother Oncol. 2019 Sep;138:68-74.
11. Kosmin M, Ledsam J, Romera-Paredes B, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiother Oncol. 2019 Jun;135:130-140.
12. Sartor H, Minarik D, Enqvist O, et al. Auto-segmentations by convolutional neural network in cervical and anorectal cancer with clinical structure sets as the ground truth. Clin Transl Radiat Oncol. 2020 Nov;25:37-45.
13. Wong J, Fong A, McVicar N, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiother Oncol. 2020 Mar;144:152-158.
14. Hansen CR, Friborg J, Jensen K, et al. NTCP model validation method for DAHANCA patient selection of protons versus photons in head and neck cancer radiotherapy. Acta Oncol. 2019 Oct;58(10):1410-1415.
15. Langendijk JA, Lambin P, De Ruysscher D, et al. Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. Radiotherapy and oncology. 2013;107(3):267-273.
16. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021 Feb;18(2):203-211.
17. Brouwer CL, Steenbakkers RJ, Bourhis J, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. Radiother Oncol. 2015 Oct;117(1):83-90.
18. Jensen K, Lorentzen E, Eriksen JG, et al. MO-0713 Inter-expert observer variance of organs at risk according to the DAHANCA guidelines. ESTRO 2023 - Abstract Book. 2023:596-597.
19. Kristensen MH, Hansen CR, Zukauskaite R, et al. Co-registration of radiotherapy planning and recurrence scans with different imaging modalities in head and neck cancer. Phys Imaging Radiat Oncol. 2022 Jul;23:80-84.
20. Piper J, Nelson A, Harper J. Deformable Image Registration in MIM Maestro® Evaluation and Description. MIM Software Inc. 2018 (White Paper).
21. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945;26(3):297-302.
22. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Det Kongelige Danske Videnskabernes Selskab. 1948;5(4):1-34.
23. Lorenzen EL, Kallehauge JF, Byskov CS, et al. A national study on the inter-observer variability in the delineation of organs at risk in the brain. Acta Oncol. 2021 Nov;60(11):1548-1554.
24. Beetz I, Schilstra C, van der Schaaf A, 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. Radiotherapy and oncology. 2012;105(1):101-106.
25. Hansen CR, Johansen J, Kristensen CA, et al. Quality assurance of radiation therapy for head and neck cancer patients treated in DAHANCA 10 randomized trial. Acta Oncol. 2015;54(9):1669-73.

26. Overgaard J, Hoff CM, Hansen HS, et al. DAHANCA 10 - Effect of darbepoetin alfa and radiotherapy in the treatment of squamous cell carcinoma of the head and neck. A multicenter, open-label, randomized, phase 3 trial by the Danish head and neck cancer group. Radiother Oncol. 2018 Apr;127(1):12-19.
27. Jensen K, Friborg J, Hansen CR, et al. The Danish Head and Neck Cancer Group (DAHANCA) 2020 radiotherapy guidelines. Radiother Oncol. 2020 Oct;151:149-151.
28. van der Veen J, Gulyban A, Willems S, et al. Interobserver variability in organ at risk delineation in head and neck cancer. Radiation Oncology. 2021 2021/06/28;16(1):120.
29. Jensen K, Al-Farra G, Dejanovic D, et al. Imaging for Target Delineation in Head and Neck Cancer Radiotherapy. Semin Nucl Med. 2021 Jan;51(1):59-67.
30. Wei Z, Ren J, Korreman SS, et al. Towards interactive deep-learning for tumour segmentation in head and neck cancer radiotherapy. Physics and Imaging in Radiation Oncology. 2023 2023/01/01/;25:100408.
31. Tol JP, Delaney AR, Dahele M, et al. Evaluation of a knowledge-based planning solution for head and neck cancer. Int J Radiat Oncol Biol Phys. 2015 Mar 1;91(3):612-20.
32. Hansen CR, Bertelsen A, Hazell I, et al. Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans. Clin Transl Radiat Oncol. 2016 Dec;1:2-8.
33. Hussein M, Heijmen BJM, Verellen D, et al. Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations. Br J Radiol. 2018 Dec;91(1092):20180270.
34. Gronberg MP, Beadle BM, Garden AS, et al. Deep Learning-Based Dose Prediction for Automated, Individualized Quality Assurance of Head and Neck Radiation Therapy Plans. Pract Radiat Oncol. 2023 May-Jun;13(3):e282-e291.
35. Baroudi H, Brock KK, Cao W, et al. Automated Contouring and Planning in Radiation Therapy: What Is 'Clinically Acceptable'? Diagnostics (Basel). 2023 Feb 10;13(4).
36. Vandewinckele L, Claessens M, Dinkla A, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol. 2020 Dec;153:55-66.
37. Lorenzen EL, Zukauskaite R, Kyndt M, et al. OC-0118 First results on DAHANCA automatic segmentation algorithms of organs at risk. ESTRO 2023 - Abstract Book. 2023:92-93.
38. Luan S, Xue X, Wei C, et al. Machine Learning-Based Quality Assurance for Automatic Segmentation of Head-and-Neck Organs-at-Risk in Radiotherapy. Technology in cancer research & treatment. 2023;22:15330338231157936-15330338231157936.