Interobserver variation in organs at risk contouring in head and neck cancer according to the DAHANCA guidelines

Nielsen, Camilla Panduro; Lorenzen, Ebbe L.; Jensen, Kenneth; Eriksen, Jesper Grau; Johansen, Jørgen; Gyldenkerne, Niels; Zukauskaite, Ruta; Kjellgren, Martin; Maare, Christian; Lønkvist, Camilla Kjær; Nowicka-Matus, Kinga; Szejniuk, Weronika Maria; Farhadi, Mohammad; Ujmajuridze, Zaza; Marienhagen, Kirsten; Johansen, Tanja Stagaard; Friborg, Jeppe; Overgaard, Jens; Hansen, Christian Rønn

Published in:
Radiotherapy and Oncology

DOI:
10.1016/j.radonc.2024.110337

Publication date:
2024

Document version:
Final published version

Document license:
CC BY-NC-ND

Citation for pulished version (APA):
Nielsen, C. P., Lorenzen, E. L., Jensen, K., Eriksen, J. G., Johansen, J., Gyldenkerne, N., Zukauskaite, R., Kjellgren, M., Maare, C., Lønkvist, C. K., Nowicka-Matus, K., Szejniuk, W. M., Farhadi, M., Ujmajuridze, Z., Marienhagen, K., Johansen, T. S., Friborg, J., Overgaard, J., & Hansen, C. R. (2024). Interobserver variation in organs at risk contouring in head and neck cancer according to the DAHANCA guidelines. *Radiotherapy and Oncology*, *197*, Article 110337. https://doi.org/10.1016/j.radonc.2024.110337

Radiotherapy and Oncology 197 (2024) 110337

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Interobserver variation in organs at risk contouring in head and neck cancer according to the DAHANCA guidelines

Camilla Panduro Nielsen [a,b,*], Ebbe L. Lorenzen [a,b], Kenneth Jensen [c], Jesper Grau Eriksen [d,e], Jørgen Johansen [b,c,f], Niels Gyldenkerne [f], Ruta Zukauskaite [b,f], Martin Kjellgren [a], Christian Maare [g], Camilla Kjær Lønkvist [g], Kinga Nowicka-Matus [h], Weronika Maria Szejniuk [c,h,i], Mohammad Farhadi [j], Zaza Ujmajuridze [j], Kirsten Marienhagen [k], Tanja Stagaard Johansen [c,l], Jeppe Friborg [l], Jens Overgaard [e], Christian Rønn Hansen [a,b,c]

[a] *Laboratory of Radiation Physics, Odense University Hospital, Odense, Denmark*
[b] *Department of Clinical Research, University of Southern Denmark, Odense, Denmark*
[c] *Danish Centre for Particle Therapy, Aarhus University Hospital, Denmark*
[d] *Department of Oncology, Aarhus University Hospital, Denmark*
[e] *Department of Experimental Clinical Oncology, Aarhus University Hospital, Denmark*
[f] *Department of Oncology, Odense University Hospital, Denmark*
[g] *Department of Oncology, Copenhagen University Hospital Herlev, Denmark*
[h] *Department of Oncology & Clinical Cancer Research Center, Aalborg University Hospital, Denmark*
[i] *Department of Clinical Medicine, Aalborg University, Denmark*
[j] *Department of Oncology, Zealand University Hospital Næstved, Denmark*
[k] *Department of Oncology, University Hospital of North Norway, Tromsø, Norway*
[l] *Department of Oncology, Rigshospitalet, Denmark*

Introduction

Organs at risk (OARs) contouring is an important but time-consuming process in radiotherapy. OARs are continuously added to the national guidelines to minimise dose to OARs [1]. Furthermore, OAR doses are increasingly being used to predict toxicities [2,3], among others through Normal Tissue Complication Probability (NTCP) models [4,5], aiding oncologists and patients in selecting the most optimal radiation treatment.

There is a variation in the contouring of OARs between experts [6], potentially impacting both the generation of treatment plans and the reporting of OAR doses [7]. Contouring guidelines provide a consensus definition [1,8], but transferring that to individual medical images is no trivial task [9]. Anatomical variation among patients, the variation in guidelines interpretations among experts and treatment centres, together with limitations in image quality introduce complexity to the process. Additionally, the limited resources within clinical environments can further exacerbate these challenges.

The development of automated delineation of OARs, atlas-based, or recently by artificial intelligence (AI) segmentation algorithms, is rapidly increasing. AI has shown to be more consistent than oncologists when contouring OARs in head and neck cancer (H&N) [10] and, thus, might be a promising solution for improving the quality of dose planning in radiotherapy.

When training an AI algorithm, the data quality is important for the subsequent model performance [11,12]. The model performance is evaluated on its ability to produce an answer that matches a harmonised data set, i.e., what is believed to be the best possible reproduction of the true answer (also called gold standard data set). Such harmonised data sets are often based on experts' opinions and performances. However, experts may interpret delineation guidelines for radiation treatment differently. The harmonised data set may lead to biased results due to lacking knowledge of the variations in the experts' performances [13].

To analyse the interplay of factors affecting contouring in dose planning, this study investigated the interobserver variation among 15 experts across six treatment centres in Denmark. The interobserver variation was analysed geometrically to highlight areas of OARs where the guidelines might be less clear, or where OARs are difficult to contour. This knowledge may aid to clarify guidelines and can be incorporated into the dose planning by considering greater uncertainties in specific directions. Additionally, the subsequent harmonised data set can serve as a valuable tool for testing AI segmentation models, as the

[*] Corresponding author at: Odense University Hospital, Kløvervænget 19, 5000 Odense C, Denmark.
E-mail address: camilla.panduro.nielsen@rsyd.dk (C.P. Nielsen).

https://doi.org/10.1016/j.radonc.2024.110337
Received 9 February 2024; Received in revised form 24 April 2024; Accepted 14 May 2024
Available online 19 May 2024

segmentation can be directly compared with expert contours.

Materials and methods

DAHANCA OAR definitions and guidelines

Defining OARs has always been an integral part of The Danish Head and Neck Cancer Group (DAHANCA) protocols and guidelines [1]. Initially, in the two-dimensional era, this was limited to describing the use of posterior electron fields over the spinal cord. With the introduction of Intensity Modulated Radiotherapy (IMRT) in the guidelines in 2004, an overview of OARs and constraints was introduced [14]. The list of OARs has since been expanded in the DAHANCA guidelines to include dysphagia-related OARs and pragmatically, included available published guidelines [1].

For the dysphagia-related OARs, the guidelines comprise a modified fusion of the guidelines of Christianen et al. [15] and Brouwer et al. [8]. In the DAHANCA guidelines, the upper esophageal inlet muscles and the cricopharyngeal muscle, defined by Christianen et al. [15], have been included in the esophagus, as there are no validated constraints for the three separate OARs, and the feasibility and reproducibility of the substructures were questionable. The separation of the pharyngeal constrictor muscle (PCM) into upper, middle and lower was maintained, as some of the early studies in dysphagia pointed to the upper constrictor as especially important for preserving swallowing function [4]. The intracranial OARs were discussed with the Danish Neuro-Oncology Group [16], and the Brouwer et al. [8] interpretation was primarily used, because the differences between different guidelines were minor. Brouwer et al. was used for most other OARs.

Prior to generation of a harmonised data set, the DAHANCA definitions of the H&N OARs were revised to quantify the expert interobserver variation, primarily through bi-annual DAHANCA quality assurance (QA) meetings, where a computed tomography (CT) scan of a single audit patient was sent to the seven departments treating H&N cancer in Denmark. All institutions provided OAR contours from the expert radiation oncologists. The variations and discrepancies were discussed at the QA meeting, and the written guidelines were clarified.

Subsequently, a CT scan of a second audit patient was distributed to the DAHANCA institutions. 20 expert oncologists and radiotherapy technologists (RTTs) contoured 17 OARs on the second audit patient, and these were discussed at the beginning of a two-day workshop. The guidelines were then further clarified. Subsequently, 15 experts attending the workshop were asked to contour OARs on CT scans of 26 patients with H&N cancer to generate multiple contours on the same patient. Intravenous contrast was used for all CT scans. The different treatment centres used different contrast agents, typically 300–350 mg/mL Iodine contrast. No clinical data was available apart from the CT scans; however, communication and exchange of information between experts was allowed during the contouring process to achieve optimal contours for later utilisation in a harmonised dataset. The experts chose how many and which OARs and patients to contour.

OAR definitions

Table 1 shows the definitions of the OAR delineations as discussed at the beginning of the workshop and as defined by DAHANCA [1]. The discussions and the definitions created the foundation for the subsequent contouring at the workshop.

All 15 experts at the workshop used the treatment planning system Eclipse v16.1 (Varian Medical Systems, Palo Alto, CA, USA) for delineation.

Patient cohort

The CT scans of the patients with H&N cancer used during the workshop were selected from a cohort of 600 patients with histologically proven squamous cell carcinoma of the pharynx or larynx planned for primary radiotherapy with curative intent, following the protocol for the DAHANCA 19 clinical trial [17,18]. Each institution contributed with around 5 patients, and the patients represented a range of body mass indices in order to select a heterogeneous cohort in terms of CT scans and body composition types. Patients' scans were anonymised for patient related data. Patients provided informed consent as part of a clinical trial, and the DAHANCA group has approved the project.

No patients had postoperative radiotherapy; therefore, tumour tissues could distort some of the delineation, especially the PCMs, and may explain some variation. In the discussion at the workshop, it was stressed that the PCMs should be delineated even though the organ delineation may be difficult and therefore, somewhat arbitrary; it was important due to the impact on the NTCP calculation of tumour embedded OARs.

Data curation

Post workshop, the segmentation data went through a sanity check, where small incorrect islands of segments were removed, and simple linear interpolation was performed in case of omitted contour interpolation. No further alterations of the contours were performed.

Comparison metrics

With several contours of the same OAR for each patient, the contour concordance was evaluated pairwise between a specific contour and each of the other contours for that OAR for the patient, until all contours of that OAR were compared for that patient. Metrics used were the dice similarity coefficient (Dice), Jaccard index (Jaccard), Mean Surface Distance (MSD), Hausdorff distance (HD), and Hausdorff 95 % distance (HD95). The metrics are introduced in Appendix A.

The metrics evaluate contour overlap based on different definitions, and metrics like Dice and Jaccard are volume dependent, meaning they should only be compared for the same OAR.

Surface mapping

To illustrate three-dimensional (3D) spatial areas of larger contouring variation, the mean standard deviation (SD) was investigated across all patients and observers, utilising surface mapping, following the procedure as described by Lorenzen et al. [16], see Fig. 1. Step 1: For each patient and each OAR, a reference contour was selected as the one with the lowest MSD in the pairwise comparison. The distance was then calculated from each point on the reference contour to the other contours for the patient, with a negative distance if the contour was inside the reference contour, and a positive distance, if it was outside. The SD was calculated of the distances, and the surface of the reference contour was coloured according to the SD in that surface point, giving an SD surface. Step 2: To visualise the variation in contouring for the specific OAR across all patients, affine transformation (linear transformation including translation, rotation, shear and scaling) was used on the reference contour from each patient to an arbitrarily chosen reference patient. This registration relationship was then used to register the SD surfaces to the coordinate system of the reference patient, and the mean SD surface was calculated by finding the nearest surface point [16]. The visualisation shows the interobserver variations when averaged across all patients and observers.

Results

In total, 3545 OARs were delineated, 17 OARs per patient. The median number of experts per OAR was 9 (interquartile range (IQR) 7–9). Areas of low image contrast and transitions between structures resulted in higher contouring variation between 15 experts. The median and IQR volume, Dice, Jaccard, MSD, HD, and HD95 for the OARs contoured in the workshop are presented in Table 2. The Dice and MSD are further

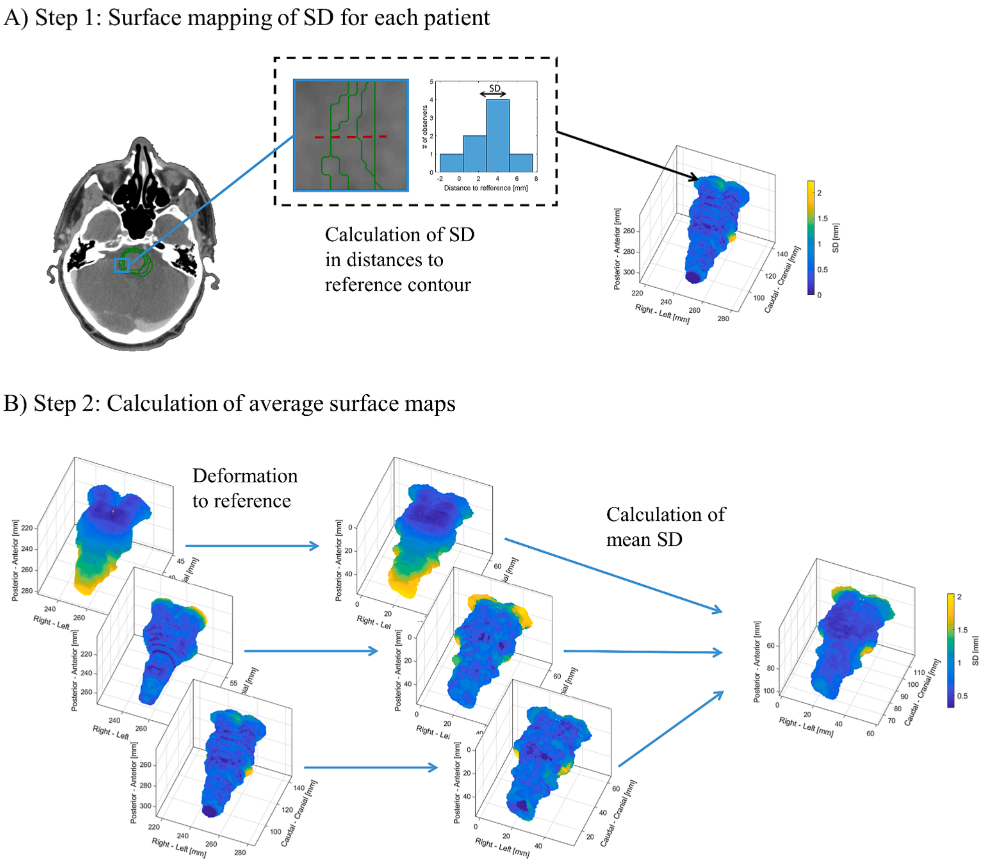
Table 1
Definitions of OARs used as the foundation for contouring of OARs at the workshop [1].

Organ	Cranial	Caudal	Anterior	Posterior	Lateral	Medial	Reference delineation*
Brainstem	Bottom of the 3rd ventricle	Tip of the dens of C2					Brouwer et al. [8]. Except cranial extended to the bottom of 3rd ventricle.
SpinalCord	Tip of the dens of C2						Brouwer et al. [8].
Esophagus (cervical esophagus + esophagus inlet muscle + cricopharyngeal muscle)	First slice caudal to the arytenoid cartilages	Sternal notch	Posterior edge of cricoid cartilage. tracheal lumen	Prevertebral muscle	Thyroid cartilage, fatty tissue, thyroid gland. Thyroid cartilage		Cervical esophagus + esophagus inlet muscle + cricopharyngeal muscle as in Christianen et al. [15].
LarynxG (glottic larynx)	Upper edge of the arythenoid cartilages	Lower edge of cricoid cartilage (if soft tissue is present)	Thyroid cartilage	Inferior PCM, pharyngeal lumen/ cricoid cartilage	Thyroid cartilage	Pharyngeal lumen (lumen excluded)	Christianen et al. [15].
LarynxSG (supraglottic larynx)	Tip of epiglottis	First slice cranial to the upper edge of the arytenoid cartilages	Hyoid bone, pre-epiglottic space, thyroid cartilage	Pharyngeal lumen, inferior PCM	Thyroid cartilage	Pharyngeal lumen (lumen excluded)	Christianen et al. [15].
OralCavity (=Brouwer extended oral cavity)	Hard palate mucosa and mucosal reflections near the maxilla	The base of tongue mucosa and hyoid posteriorly and the mylohyoid m. and ant. belly of the digastric m. anteriorly	Inner surface of the mandible and maxilla	Post. borders of soft palate, uvula, and more inferiorly the base of tongue	Inner surface of the mandible and maxilla		Brouwer et al. [8].
Parotid_L Parotid_R							No change to definition in Brouwer et al. [8].
PCM_Low (lower pharyngeal constrictor)	First slice caudal to the lower edge of hyoid bone	Lower edge of the arythenoid cartilages	Soft tissue of supraglottic/ glottic larynx	Prevertebral muscle	Superior horn of thyroid cartilage		Christianen et al. [15].
PCM_Mid (middle pharyngeal constrictor)	Upper edge of C3	Lower edge of hyoid bone	Base of tongue, hyoid	Prevertebral muscle	Greater horn of hyoid bone	Pharyngeal lumen	Christianen et al. [15].
PCM_Up (upper pharyngeal constrictor)	Caudal tip of the pterygoid plates (hamulus)	Lower edge of C2	Hamulus of pterygoid plate; mandibula; base of tongue; pharyngeal lumen	Prevertebral muscle	Medial pterygoid muscle	Pharyngeal lumen	Christianen et al. [15].
Submandibular_L Submandibular_R	Med. pterygoid m., mylohyoid m.	Fatty tissue	Lat. Surface mylohyoid m., hyoglossus m.	Parapharyngeal space, sternocleidomastoid m.	Med. surface med. pterygoid m., med. surface mandibular bone, platysma	Lat. surface mylohyoid m., hyoglossus m., superior and middle pharyngeal constrictor m., anterior belly of the digastric m.	Brouwer et al. [8].
Thyroid							No change to definition in Brouwer et al. [8].
Buccal mucosa	Bottom of maxillary sinus	Upper edge teeth sockets	Lips, teeth	Med. pterygoid m.	Buccal fat	Outer surface of the mandible and maxilla, oral cavity/base of tongue/soft palate	Brouwer et al. [8].

(continued on next page)

Table 1 (*continued*)

Organ	Cranial	Caudal	Anterior	Posterior	Lateral	Medial	Reference delineation*
Lips	Hard palate (lateral), anterior nasal spine (at the midline)	Lower edge teeth sockets, cranial edge mandibular body	Outer surface of the skin	Mandibular body, teeth, tongue, air (if present)	Depressor anguli oris m., buccinator m. levator anguli oris m./risorius m. (the mentioned mucles are all lateral to the m. orbicularis oris)	Hard palate (lateral), anterior nasal spine (at the midline)	Brouwer et al. [8].


A) Step 1: Surface mapping of SD for each patient

Calculation of SD in distances to reference contour

B) Step 2: Calculation of average surface maps

Deformation to reference

Calculation of mean SD

Fig. 1. Demonstration of surface mapping depicting interobserver variation in segmentation. Initially, as depicted in (A), the SD in distance between various segmentations and a designated reference segmentation was computed for each patient, with the results graphically represented on the surface corresponding to the reference contour. Subsequently, as outlined in (B), these patient-specific surfaces reflecting the SD were normalised to a reference patient, facilitating the plotting of individual surfaces onto a unified plane. This approach enables the calculation of SD across patients while ensuring comparability. Figure adapted with permission from "A national study on the inter-observer variability in the delineation of organs at risk in the brain" by Lorenzen et al. [16].

visualised in Fig. 2.

Traditionally delineated OARs like the brainstem, oral cavity, parotids, submandibular glands, esophagus and thyroid, showed high median Dice, above 0.8. The OARs that were added more recently to the guidelines, and are not well-defined on CT, like PCMs, lips, and buccal mucosa, had the lowest median Dice, below 0.6.

The structures with the highest tissue contrast, like submandibular and thyroid glands, showed a median MSD of around 1 mm, while other OARs had a median MSD below 3 mm. The spinal cord showed a large variation linked to observers' inconsistency in following the guidelines of contouring the full length in the caudal direction, but only minor variation was observed in the axial plane.

Fig. 3 shows the areas of the OAR surfaces where the mean SD is high (yellow) for the individual OARs i.e., where interobserver variation is large across all patients and observers. Both buccal mucosa show a mean SD above 1.5 mm. The transition between PCM low and esophagus showed a large variation. Likewise, there is roughly one CT slice of variation for the transition between the glottic larynx and supraglottic larynx. Both submandibular glands have the largest variation in the cranial direction. For the spinal cord, there was substantial variation in the caudal direction; however, the variation results from the spinal cord not being contoured on the full scan length, as instructed, and not due to interobserver variation.

Discussion

After thorough discussions and clarification of DAHANCA contouring guidelines [1], a new national harmonised data set from CT scans of 26 patients with H&N cancer was created.

Several publications investigated the interobserver variation in the delineation of clinical target volume and OARs and reported substantial variations [19,20]. Van der Veen et al., analysed the interobserver OAR

Table 2
Median volume and contour comparison metrics with corresponding IQR in parentheses.

	Volume [cm^3]	Dice [index]	Jaccard [index]	MSD [mm]	HD [mm]	HD95 [mm]
Brainstem	24.5 (22.6–27.5)	0.82 (0.78–0.84)	0.70 (0.65–0.73)	1.7 (1.4–2.0)	8.1 (7.1–9.6)	4.6 (4.0–5.6)
Spinal cord	20.6 (15.4–25.8)	0.74 (0.70–0.77)	0.61 (0.55–0.64)	2.9 (2.0–4.0)	31.2 (20.2–43.0)	13.9 (9.3–17.1)
Esophagus	10.6 (8.4–12.0)	0.73 (0.69–0.77)	0.59 (0.53–0.64)	1.8 (1.6–2.4)	13.1 (10.6–17.5)	7.7 (5.7–10.1)
Glottic larynx	4.2 (2.9–7.0)	0.57 (0.49–0.64)	0.41 (0.35–0.49)	2.0 (1.6–3.1)	10.2 (8.0–13.4)	5.4 (4.3–7.3)
Suppraglottic larynx	11.7 (9.5–15.6)	0.67 (0.60–0.72)	0.52 (0.43–0.58)	1.8 (1.4–2.3)	9.7 (8.3–12.9)	5.1 (4.5–6.5)
Oral cavity	103.0 (89.2–122.3)	0.83 (0.80–0.86)	0.72 (0.68–0.76)	2.6 (2.1–3.0)	15.2 (12.2–17.6)	9.0 (7.3–10.3)
Left parotid	27.1 (20.8–30.2)	0.81 (0.77–0.84)	0.68 (0.64–0.73)	1.8 (1.5–2.1)	11.4 (10.0–14.7)	5.5 (4.5–7.2)
Right parotid	27.5 (21.7–31.8)	0.82 (0.78–0.86)	0.70 (0.65–0.75)	1.6 (1.4–2.0)	11.9 (10.5–14.7)	5.5 (4.3–6.4)
Lower PCM	5.4 (3.5–6.9)	0.56 (0.47–0.62)	0.41 (0.35–0.48)	2.1 (1.7–2.7)	10.4 (8.4–15.4)	6.3 (5.3–8.8)
Middle PCM	4.6 (3.3–6.2)	0.56 (0.48–0.63)	0.41 (0.34–0.47)	2.0 (1.5–2.9)	11.8 (9.2–15.5)	5.8 (4.4–8.2)
Upper PCM	10.3 (7.4–13.3)	0.54 (0.49–0.56)	0.38 (0.34–0.42)	2.1 (1.8–2.6)	13.4 (11.5–16.2)	6.9 (5.8–8.3)
Left submandibular	8.8 (6.9–10.6)	0.84 (0.76–0.86)	0.72 (0.63–0.76)	1.1 (0.9–1.6)	6.9 (5.8–9.1)	4.0 (3.2–5.1)
Right submandibular	9.2 (7.5–10.7)	0.84 (0.81–0.87)	0.73 (0.69–0.77)	1.0 (0.8–1.3)	6.2 (4.2–7.7)	3.2 (2.5–4.3)
Thyroid	20.2 (13.4–27.5)	0.83 (0.79–0.87)	0.71 (0.66–0.78)	0.9 (0.8–1.1)	7.2 (5.7–8.6)	2.7 (2.4–3.6)
Left buccal mucosa	5.9 (4.6–8.6)	0.48 (0.39–0.52)	0.33 (0.27–0.38)	3.1 (2.5–4.1)	16.0 (12.8–18.4)	8.8 (7.2–11.1)
Right buccal mucosa	6.0 (4.8–7.9)	0.48 (0.38–0.55)	0.34 (0.26–0.40)	2.9 (2.5–3.9)	15.7 (12.3–18.3)	8.4 (7.1–10.7)
Lips	15.9 (11.7–23.7)	0.49 (0.46–0.57)	0.35 (0.30–0.42)	3.0 (2.5–3.6)	15.1 (13.5–16.9)	8.0 (6.7–9.1)

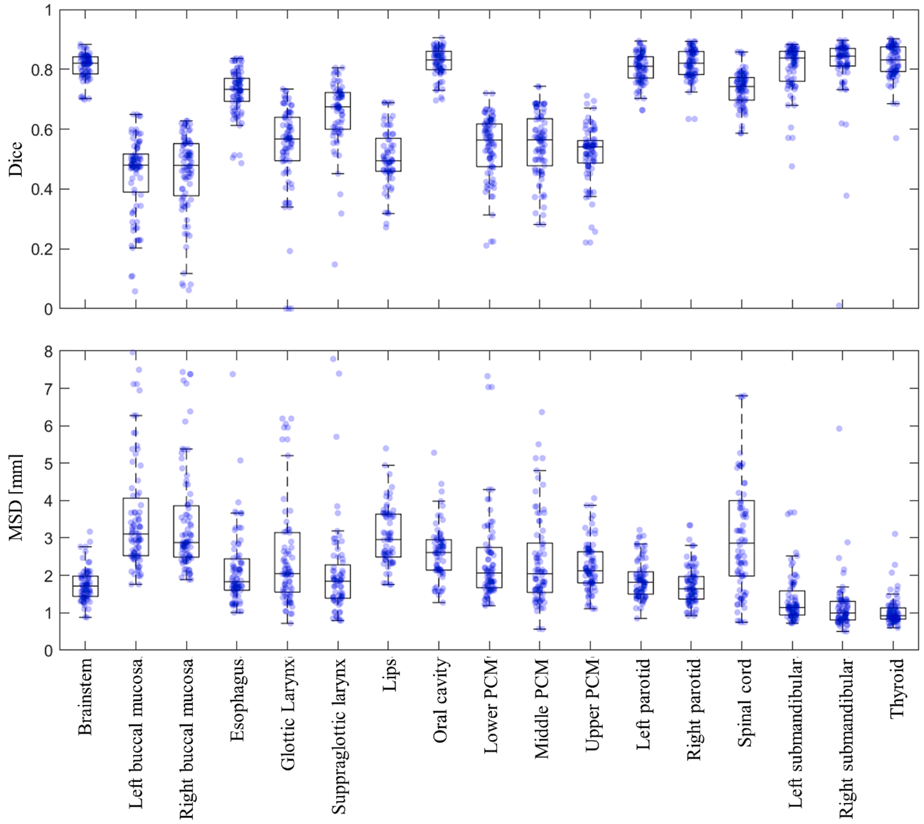


Fig. 2. Box plots with samples overlaid showing Dice and MSD for the 17 OARs investigated.

variation across the Belgian radiotherapy centres for delineations of five scans of H&N cancer patients through a survey [9]. The results are similar to the current study, apart from higher Dice and lower MSD in the current study. This was anticipated, since the current results were obtained after a dedicated workshop, including detailed discussions and interpretations of the guidelines. The results of Van der Veen et al. are presumably much closer to everyday clinical practice, while the current study probably defines the best achievable interobserver variation, as they are linked to the workshop.

Nelms et al., showed that the interobserver variation in contouring of six OARs in the H&N (brainstem, brain, left and right parotids, mandible, and spinal cord) led to substantial differences in mean and maximum doses in treatment plans [21]. Feng et al. found that the mean dose differences were around 1 Gy SD across 10 oropharynx patients contoured by three experts [22]. Additionally, it's important to note that while maximum and minimum doses are substantially affected by contouring variation, most NTCP models rely on the reported mean dose, which is less influenced by these variations [23,24]. The current study does not investigate the dosimetric consequences; however, the 3D interobserver variation indicates which areas would result in the dose differences due to contouring variation.

Brouwer et al. visualised the 3D interobserver variation for the spinal cord, parotids and submandibular glands [6], providing 3D information on the highest interobservers' disagreement and potential need for acceptance of the highest difference from an automated contouring approach. In the visualisation of the parotids, the caudal and cranial

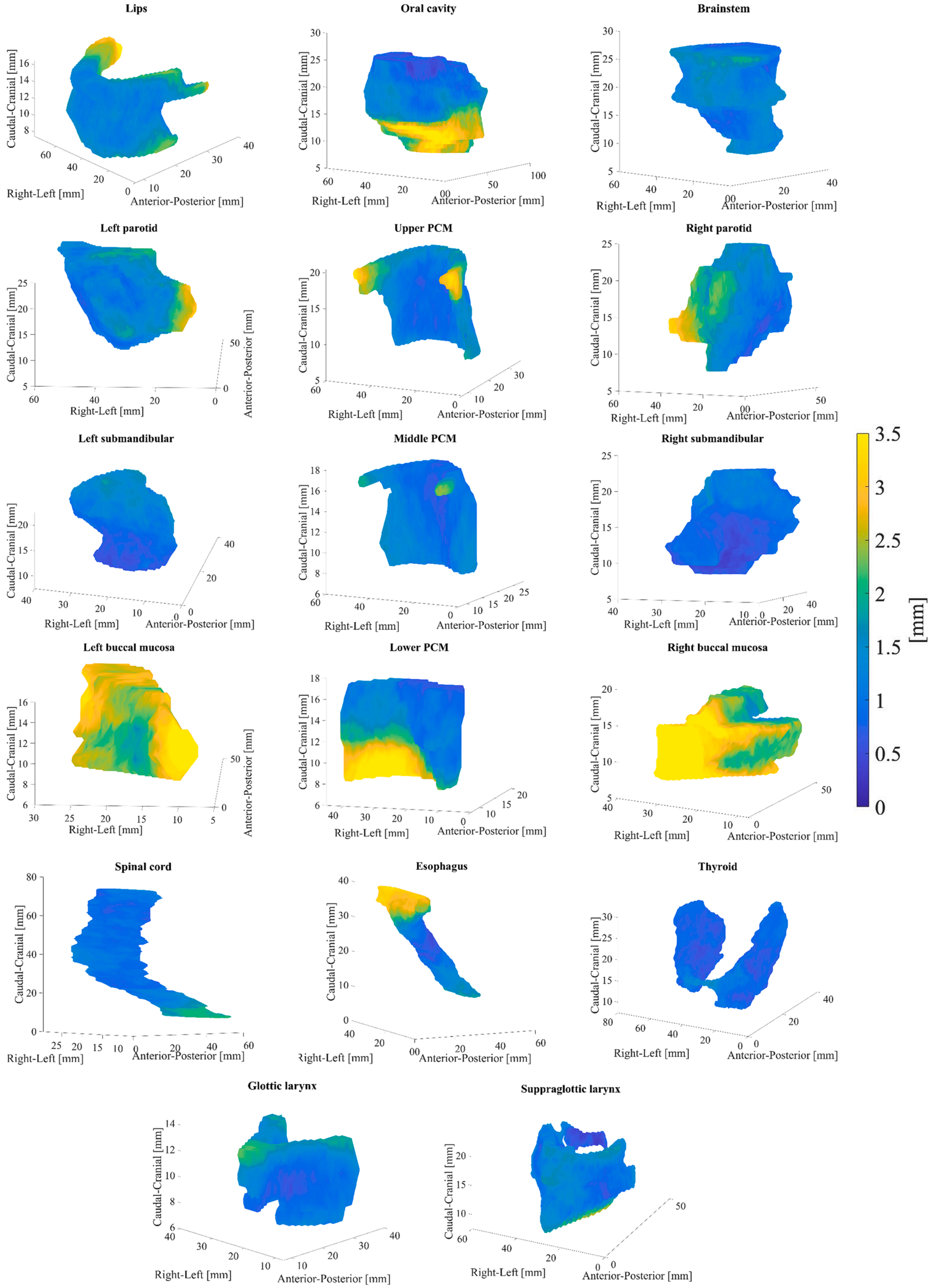


Fig. 3. Mean SD surfaces averaged across all patients and observers for each OAR. The colour of the SD surfaces are according to SD, so a yellow surface point shows a high SD, and a blue surface point shows a low SD. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

parts had an interobserver variation SD of 3 mm or more [6]. In contrast, the current study does not confirm this finding, but shows a mean SD of 3 mm in the medial part of the parotid. The mean SD of the submandibular glands is below 2 mm in both studies.

Defining accurate boundaries for OARs is a critical stage in automated contouring. While some OARs align with anatomical atlases, others, such as the PCMs have a more complex shape. Overlapping and not easily distinguishable from each other on a CT scan, these contours rely more on radiotherapy-oriented and arbitrary definitions than purely anatomical elements. In the DAHANCA guidelines [1], the lower PCM encompasses the cricopharyngeal inlet and the esophagus delineation includes the cervical esophagus, which differs from the globally accepted OAR contouring guidelines proposed by Brouwer et al. [6].

The included patients had cancer at different sites and stages, and across the 26 patients they represent a general cohort of head and neck cancer patients. The tumour could potentially add a small variation to the contours when embedded in OARs, but as the tumour size and stage varies across the patients, the contribution to the variation by this should be small. Furthermore, experts were instructed to delineate the OAR as normal even though it might be difficult with distortions.

The present study showed that the variation in contouring of OARs in H&N patients is organ-specific. The 3D visualisation method used here, as demonstrated in Fig. 3, provides a relevant tool for oncologists to visualise the areas of specific OARs that are less defined. This could draw attention to which organs and directions that pose major difficulties in defining the precise boundaries of relevant normal tissue structures. Furthermore, such visualisation could be used for educational purposes.

When comparing the interobserver variation studies, there is often a lack of geometric description. Vinod et al. reviewed the literature regarding interobserver variation in both OAR and targets [25] and recommended how to improve the reporting of these studies [26]. Several metrics have been reported to facilitate comparison between many studies. Some, like the Dice, are very volume-dependent, and care should be taken when comparing this metric across different OARs. The MSD metric can be compared across OARs; however, between studies, one should ensure that MSD is calculated in the same way.

Contouring is time-consuming, and an automated process could have substantial clinical benefits. Several publications investigated auto contouring, and quantified the contour quality by measuring the amount of manual editing needed for the automated contour to be clinically acceptable [27]. Minor corrections will almost always be performed with some interobserver variation due to interpretation differences. However, if the contouring differences are solely influenced by subjective opinions, one could argue that corrections may be unnecessary. The 3D visualisations in Fig. 3 visualises and quantifies these differences and a scale. The next rational step is to investigate the dosimetric implications of the variation in order to define the dosimetric consequence of a given uncertainty in delineation and whether correction is meaningful.

OARs that have been introduced more recently to H&N treatment planning, like the PCMs and buccal mucosa, are rarely included in the automated segmentation algorithms [28–30]; however, these automated OAR contours could potentially give the less experienced oncologists a solid starting point, making clinical implementation more feasible [31]. Inclusion of more OARs should improve treatment plans by sculpting dose away from specific OARs. Therefore, the more OARs that are segmented, the more options will be available to improve treatment plan quality [32].

The present data set can be used to validate external OAR segmentations. If the segmentation looks reasonable on a few local test cases (face validity), performing a systematic quantitative validation on the atlas is informative. It is important to perform a local validation first, as a model can perform well in one environment but fail in a different setting.

A harmonised data set can be used prospectively as a gold standard for validating and implementing segmentation tools. Through the Danish national treatment plan database, DcmCollab [33], it will soon be possible to validate automated OAR segmentations for some of the gold standard data patients. The analyses of metrics will be performed automatically.

In conclusion, the current study presents the interobserver variation in contouring H&N OARs during a national DAHANCA workshop, producing a harmonised data set visualising the interobserver variation in 3D and by several recommended contouring metrics. The data set can be used for educational purposes and for automated segmentation validation.

CRediT authorship contribution statement

Camilla Panduro Nielsen: Writing – original draft, Visualization, Formal analysis. **Ebbe L. Lorenzen:** Writing – review & editing, Validation, Methodology, Formal analysis, Data curation. **Kenneth Jensen:** Writing – original draft, Methodology, Conceptualization. **Jesper Grau Eriksen:** Writing – review & editing, Methodology, Data curation. **Jørgen Johansen:** Writing – review & editing, Validation, Data curation, Conceptualization. **Niels Gyldenkerne:** Writing – review & editing, Conceptualization. **Ruta Zukauskaite:** Writing – review & editing, Formal analysis, Data curation. **Martin Kjellgren:** Writing – review & editing, Data curation. **Christian Maare:** Writing – review & editing, Data curation. **Camilla Kjær Lønkvist:** Writing – review & editing, Data curation. **Kinga Nowicka-Matus:** Methodology, Data curation. **Weronika Maria Szejniuk:** Writing – review & editing, Data curation. **Mohammad Farhadi:** Writing – review & editing, Data curation. **Zaza Ujmajuridze:** Writing – review & editing, Data curation. **Kirsten Marienhagen:** Writing – review & editing, Data curation. **Tanja Stagaard Johansen:** Writing – review & editing, Data curation. **Jeppe Friborg:** Writing – review & editing, Formal analysis, Data curation. **Jens Overgaard:** Writing – review & editing, Methodology, Conceptualization. **Christian Rønn Hansen:** Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization.

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

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

Appendix A. Comparison metrics

For comparison of contour overlap, the dice similarity coefficient (Dice), Jaccard index (Jaccard), Mean Surface Distance (MSD), Hausdorff distance (HD), and Hausdorff 95 % distance (HD95) were used. The metrics are introduced below for two volumes A and B.

The Dice metric calculates the overlap of two volumes (A and B), where 100 % overlap is 1, and 0 % overlap is 0:

$$Dice = \frac{2(A \cap B)}{A + B}$$

Like Dice, Jaccard measures overlap between 0 and 1, like this:

$$Jaccard = \frac{A \cap B}{A \cup B}$$

The MSD calculates the average distance between the surface of two volumes A and B, by first calculating the mean squared distance between each point *a* on the surface of volume *A* and its nearest neighbour on *B*, normalised by the number of points *N*:

$$meansquareddistance(A, B) = \frac{1}{N}\sum_{a \in A}\min_{b \in B}\|a - b\|$$

$\|a - b\|$ is the Euclidian distance between points *a* and *b*. The MSD is then the mean squared distance between sets *A* and *B* calculated in both directions:

$$MSD = \frac{1}{2}\left(meansquareddistance(A, B) + meansquareddistance(B, A)\right)$$

The HD measure returns the maximum distance between a surface and the closest point on the other surface:

$$hd(A, B) = \max_{a \in A}\min_{b \in B}\|a - b\|$$

$$HD = \max(hd(A, b) + hd(B, A))$$

Instead of calculating the maximum distance, HD95 calculates the 95th percentile of the distances between a surface and the closest point on the other surface.

References

[1] Jensen K, Friborg J, Hansen CR, Samsøe E, Johansen J, Andersen M, et al. The danish head and neck cancer group (DAHANCA) 2020 radiotherapy guidelines. Radiother Oncol 2020;151:149–51.

[2] Hansen CR, Bertelsen A, Zukauskaite R, Johnsen L, Bernchou U, Thwaites DI, et al. Prediction of radiation-induced mucositis of H&N cancer patients based on a large patient cohort. Radiother Oncol 2020;147:15–21.

[3] 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.

[4] Langendijk JA, Lambin P, De Ruysscher D, Widder J, Bos M, Verheij M. Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach. Radiother Oncol 2013;107:267–73.

[5] Hansen CR, Friborg J, Jensen K, Samsøe E, Johnsen L, Zukauskaite R, et al. NTCP model validation method for DAHANCA patient selection of protons versus photons in head and neck cancer radiotherapy. Acta Oncol 2019;58:1410–5.

[6] Brouwer CL, Steenbakkers RJHM, 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.

[7] Brouwer CL, Steenbakkers RJ, Gort E, Kamphuis ME, van der Laan HP, Van't Veld AA, et al. Differences in delineation guidelines for head and neck cancer result in inconsistent reported dose and corresponding NTCP. Radiother Oncol 2014;111:148–52.

[8] Brouwer CL, Steenbakkers RJ, 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.

[9] van der Veen J, Gulyban A, Willems S, Maes F, Nuyts S. Interobserver variability in organ at risk delineation in head and neck cancer. Radiat Oncol 2021;16:120.

[10] Nielsen CP, Lorenzen EL, Jensen K, Sarup N, Brink C, Smulders B, et al. Consistency in contouring of organs at risk by artificial intelligence vs oncologists in head and neck cancer patients. Acta Oncol 2023;1–8.

[11] Rangineni S. An analysis of data quality requirements for machine learning development pipelines frameworks. Int J Computer Trends and Tech 2023;71:16–27.

[12] Barragán-Montero AM, Thomas M, Defraene G, Michiels S, Haustermans K, Lee JA, et al. Deep learning dose prediction for IMRT of esophageal cancer: the effect of data quality and quantity on model performance. Phys Med 2021;83:52–63.

[13] Lebovitz S, Levina N, Lifshitz-Assaf H. Is AI ground truth really true? the dangers of training and evaluating AI tools based on experts' know-what. MIS Q 2021;45:1501–26.

[14] Hansen CR, Johansen J, Kristensen CA, Smulders B, Andersen LJ, Samsøe E, et al. Quality assurance of radiation therapy for head and neck cancer patients treated in DAHANCA 10 randomized trial. Acta Oncol 2015;54:1669–73.

[15] Christianen ME, Langendijk JA, Westerlaan HE, van de Water TA, Bijl HP. Delineation of organs at risk involved in swallowing for radiotherapy treatment planning. Radiother Oncol 2011;101:394–402.

[16] Lorenzen EL, Kallehauge JF, Byskov CS, Dahlrot RH, Haslund CA, Guldberg TL, et al. A national study on the inter-observer variability in the delineation of organs at risk in the brain. Acta Oncol 2021;60:1548–54.

[17] Kristensen MH, Sørensen MK, Tramm T, Alsner J, Sørensen BS, Maare C, et al. Tumor volume and cancer stem cell expression as prognostic markers for high-dose loco-regional failure in head and neck squamous cell carcinoma - a DAHANCA 19 study. Radiother Oncol 2024;193:110149.

[18] Zukauskaite R, Horsholt Kristensen M, Grau Eriksen J, Johansen J, Samsøe E, Johnsen L, et al. Comparison of 3-year local control using DAHANCA radiotherapy guidelines before and after implementation of five millimetres geometrical GTV to high-dose CTV margin. Radiother Oncol 2024;196:110284.

[19] 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.

[20] Mukesh M, Benson R, Jena R, Hoole A, Roques T, Scrase C, et al. Interobserver variation in clinical target volume and organs at risk segmentation in post-parotidectomy radiotherapy: can segmentation protocols help? Br J Radiol 2012;85 (e530-6).

[21] Nelms BE, Tome WA, Robinson G, Wheeler J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. Int J Radiat Oncol Biol Phys 2012;82:368–78.

[22] 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).

[23] Hansen CR, Jensen K, Smulders B, Holm AIS, Samsøe E, Nielsen MS, et al. Evaluation of decentralised model-based selection of head and neck cancer patients for a proton treatment study. DAHANCA 35. Radiother Oncol 2023;109812.

[24] Gan Y, Langendijk JA, Oldehinkel E, Scandurra D, Sijtsema NM, Lin Z, et al. A novel semi auto-segmentation method for accurate dose and NTCP evaluation in adaptive head and neck radiotherapy. Radiother Oncol 2021;164:167–74.

[25] Vinod SK, Min M, Jameson MG, Holloway LC. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. J Med Imaging Radiat Oncol 2016;60:393–406.

[26] Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. Radiother Oncol 2016;121:169–79.

[27] Ye X, Guo D, Ge J, Yan S, Xin Y, Song Y, et al. Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. Nat Commun 2022;13.

[28] 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.

[29] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. J Med Internet Res 2021;23:e26151.

[30] Wong J, Fong A, McVicar N, Smith S, Giambattista J, Wells D, 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;144:152–8.

[31] Lorenzen EL, Zukauskaite R, Kyndt M, Eriksen JG, Sarup N, Johansen J, et al. OC-0118 First results on DAHANCA automatic segmentation algorithms of organs at risk. Abstract Book; 2023. p. 92–3.

[32] Hansen CR, Hussein M, Bernchou U, Zukauskaite R, Thwaites D. Plan quality in radiotherapy treatment planning - review of the factors and challenges. J Med Imaging Radiat Oncol 2022;66:267–78.

[33] Krogh SL, Brink C, Lorenzen EL, Samsøe E, Vogelius IR, Zukauskaite R, et al. A national repository of complete radiotherapy plans: design, Results, and experiences. Acta Oncol 2023;62:1161–8.