Radiotherapy and Oncology 198 (2024) 110410

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

HaN-Seg: The head and neck organ-at-risk CT and MR segmentation challenge

Gašper Podobnik [a,*], Bulat Ibragimov [a,b], Elias Tappeiner [c], Chanwoong Lee [d,e], Jin Sung Kim [d,e,f], Zacharia Mesbah [g,h], Romain Modzelewski [g,i], Yihao Ma [j], Fan Yang [j], Mikołaj Rudecki [k], Marek Wodziński [k,l], Primož Peterlin [m], Primož Strojan [m], Tomaž Vrtovec [a]

[a] University of Ljubljana, Faculty Electrical Engineering, Tržaška cesta 25, Ljubljana 1000, Slovenia
[b] University of Copenhagen, Department of Computer Science, Universitetsparken 1, Copenhagen 2100, Denmark
[c] UMIT Tirol – Private University for Health Sciences and Health Technology, Eduard-Wallnöfer-Zentrum 1, Hall in Tirol 6060, Austria
[d] Yonsei University, College of Medicine, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea
[e] Yonsei Cancer Center, Department of RadiationOncology, 50-1 Yonsei-Ro, Seodaemun-gu, Seoul 03722, South Korea
[f] Oncosoft Inc, 37 Myeongmul-gil, Seodaemun-gu, Seoul 03722, South Korea
[g] Henri Becquerel Cancer Center, 1 Rue d'Amiens, Rouen 76000, France
[h] Siemens Healthineers, 6 Rue du Général Audran, CS20146, Courbevoie 92412, France
[i] Litis UR 4108, 684 Av. de l'Université, Saint- Étienne-du-Rouvray 76800, France
[j] Guizhou Medical University, School of Biology & Engineering, 9FW8+2P3, Ankang Avenue, Gui'an New Area, Guiyang, Guizhou Province 561113, China
[k] AGH University of Kraków, Department of Measurement and Electronicsal, Mickiewicza 30, Kraków 30-059, Poland
[l] University of Applied Sciences Western Switzerland, Information Systems Institute, Rue de la Plaine 2, Sierre 3960, Switzerland
[m] Institute of Oncology, Ljubljana, Zaloška cesta 2, Ljubljana 1000, Slovenia

ARTICLE INFO

Keywords:
Computational challenge
Segmentation
Deep learning
Organs-at-risk
Computed tomography
Magnetic resonance
Radiotherapy
Head and neck cancer

ABSTRACT

Background and purpose: To promote the development of auto-segmentation methods for head and neck (HaN) radiation treatment (RT) planning that exploit the information of computed tomography (CT) and magnetic resonance (MR) imaging modalities, we organized *HaN-Seg: The Head and Neck Organ-at-Risk CT and MR Segmentation Challenge*.

Materials and methods: The challenge task was to automatically segment 30 organs-at-risk (OARs) of the HaN region in 14 withheld test cases given the availability of 42 publicly available training cases. Each case consisted of one contrast-enhanced CT and one T1-weighted MR image of the HaN region of the same patient, with up to 30 corresponding reference OAR delineation masks. The performance was evaluated in terms of the Dice similarity coefficient (DSC) and 95-percentile Hausdorff distance (HD_{95}), and statistical ranking was applied for each metric by pairwise comparison of the submitted methods using the Wilcoxon signed-rank test.

Results: While 23 teams registered for the challenge, only seven submitted their methods for the final phase. The top-performing team achieved a DSC of 76.9 % and a HD_{95} of 3.5 mm. All participating teams utilized architectures based on U-Net, with the winning team leveraging rigid MR to CT registration combined with network entry-level concatenation of both modalities.

Conclusion: This challenge simulated a real-world clinical scenario by providing non-registered MR and CT images with varying fields-of-view and voxel sizes. Remarkably, the top-performing teams achieved segmentation performance surpassing the inter-observer agreement on the same dataset. These results set a benchmark for future research on this publicly available dataset and on paired multi-modal image segmentation in general.

Introduction

Radiation therapy (RT) is, in addition to surgery and systemic therapy, a cornerstone of head and neck (HaN) cancer treatment [1]. In parallel with improvements in organ-sparing surgery, introduction of new systemic agents and refinements of multi-modal treatment scenarios, advances in RT have over the past decades contributed significantly to the preservation of organ function and reduction of mortality

* Corresponding author at: University of Ljubljana, Faculty of Electrical Engineering, Tržaška cesta 25, Ljubljana SI-1000, Slovenia.
E-mail address: gasper.podobnik@fe.uni-lj.si (G. Podobnik).

https://doi.org/10.1016/j.radonc.2024.110410
Received 30 April 2024; Received in revised form 12 June 2024; Accepted 15 June 2024
Available online 23 June 2024

in HaN cancer patients [1]. With the introduction of artificial intelligence (AI), the RT workflow has witnessed a shift towards a more qualitative, standardized and fast implementation of a wide range of applications [2]. In comparison to manual slice-by-slice image delineation [3–5], AI-assisted methods for medical image analysis perform auto-segmentation of tumor volumes and organs-at-risk (OARs) from three-dimensional (3D) computed tomography (CT) [6–10], magnetic resonance (MR) [11–14] and/or positron emission tomography (PET) [15] images. In the process of creating optimal patient-specific radiation dose distribution plans, auto-segmentation provides a faster RT workflow that is less labor-intensive with reduced intra- and inter-observer variability [2].

With a variety of AI-assisted methods at our disposal [16,17], computational challenges [18] provide a systematic and objective method evaluation. In a competition-oriented setup, challenge organizers release images with known reference delineation masks used by participants for method development. These methods are then evaluated on images with reference delineation masks available only to challenge organizers. Since 2009, six different computational challenges focusing on HaN OAR segmentation have been organized [19–23]. Five teams segmented the mandible and brainstem from 25 CT images during the Medical Image Computing and Computer Assisted Interventions (MICCAI) 2009 conference [19], six teams segmented the parotid glands from the same database during MICCAI 2010 [20], and six teams segmented six OARs (i.e. brainstem, mandible, optic chiasm, optic nerves, parotid glands, submandibular glands) from 40 CT images during MICCAI 2015 [21] within the *Head and Neck Auto-Segmentation Challenge*. During the 2019 Annual Meeting of the American Association of Physicists in Medicine (AAPM), 10 teams segmented the parotid glands, submandibular glands and lymph nodes from 55 MR images within the *Auto-segmentation on MRI for Head-and-Neck Radiation Treatment Planning Challenge (RT-MAC)* [22], while during MICCAI 2019, 12 teams segmented 13 OARs (i.e. eyes, lens, optic nerves, optic chiasm, pituitary gland, brainstem, temporal lobes, spinal cord, parotid glands, inner ear, middle ear, temporo-mandibular joints and mandible) from 60 CT images within the *StructSeg2019: Automatic Structure Segmentation for Radiotherapy Planning Challenge*.[1] Finally, during MICCAI 2023, 12 teams segmented a complete set of 45 different OARs from 200 CT scans within the *SegRap2023: Segmentation of Organs-at-Risk and Gross Tumor Volume of Nasopharyngeal Carcinoma for Radiotherapy Planning Challenge* [23].

While the number of images for method development and evaluation as well as the number of OARs to segment have increased with years, no computational challenge has yet targeted HaN OAR segmentation by combining multi-modal information from CT and MR images of the same patients. The information obtained from MR images has been recommended to complement CT images to improve the visualization of soft tissues [3,24,25], and proved to be particularly beneficial for delineating and segmenting both tumor volumes and OARs in the HaN region [3,26]. As a result, auto-segmentation methods integrating both modalities have been proposed [27–29] that may, by replacing CT image acquisition with synthetic CT image generation [30,31], further contribute to the paradigm of MR-only RT [32–35]. To promote the development of new and application of existing state-of-the-art auto-segmentation methods, we therefore organized *HaN-Seg: The Head and Neck Organ-at-Risk CT and MR Segmentation Challenge* that was held on the *Grand Challenge* online platform[2] between March 2023 and February 2024. In this report, we provide details about the HaN-Seg challenge organization, submitted methods and obtained results.

Materials and methods

Dataset

The HaN-Seg challenge data consisted of 56 cases [36], where each case is represented by one contrast-enhanced CT and one T1-weighted MR image of the HaN region of the same patient that was appointed for RT due to previous cancer diagnosis, with corresponding curated reference 3D delineation masks for CT images of up to 30 OARs (Fig. 1). Each case was delineated by an RT technologist or radiation oncologist, and the resulting delineations were curated by a medical imaging researcher. In each phase, the 3D OAR delineation masks were defined for CT images but were obtained by aid of co-registered MR images (please refer to [36] for dataset details). The cases were randomly split into a training set with 42 cases that is publicly available[3] [36] and was used by participants for method development, and a test set with 14 cases that is withheld and was used by the organizers for method evaluation.

Challenge setup

The task of the HaN-Seg challenge was to automatically segment 30 OARs in the HaN region from the devised test set given the availability of the training set, i.e. to provide 3D OAR segmentation masks in the coordinate system of each test CT image by considering a pair of CT and MR images is available for each training and test case. The challenge was organized by taking into account the current guidelines for biomedical image analysis competitions [18]. Participants were required to register as a team on the *Grand Challenge* online platform, and submit their methods in the form of Docker containers. In the Preliminary Test Phase, each participant was allowed to submit multiple methods (limited to one submission per week over a period of 31 weeks). Each submitted method was then executed on the platform with its performance estimated on four pre-selected cases from the test set, and the resulting team rankings were updated on the live public leaderboard. In the Final Test Phase, both existing and new participants were allowed to submit multiple methods (limited to one submission per day over a period of 15 weeks), which were executed on all 14 cases from the test set. The best-performing final methods were ranked according to their performance.

Evaluation metrics and ranking

The submitted methods were evaluated in terms of the Dice similarity coefficient (DSC) and 95-percentile Hausdorff distance (HD_{95}), implemented by Google DeepMind[4] [7], which are established metrics for assessing segmentation performance in RT [16,37,38]. Statistical ranking was applied for each metric by pairwise comparison of the methods using the Wilcoxon signed-rank test [39] with Bonferroni correction, resulting in a significance score and metric-specific rank. Specifically, separately for each metric, if a method resulted to be statistically significantly better in performance than the method it was compared to, its significance score was increased by 1 (from the initial zero value). According to the aggregated significance scores, ranks were assigned separately for each metric. Identical ranks were assigned to algorithms that showed only marginal performance differences so as to evaluate only statistically significant differences among methods. The final rank was obtained by aggregating the ranks over both metrics, and in the case it was equal for multiple methods, they were ordered according to the mean of both metrics.

[1] https://structseg2019.grand-challenge.org
[2] https://han-seg2023.grand-challenge.org
[3] https://doi.org/10.5281/zenodo.7442914
[4] https://github.com/google-deepmind/surface-distance

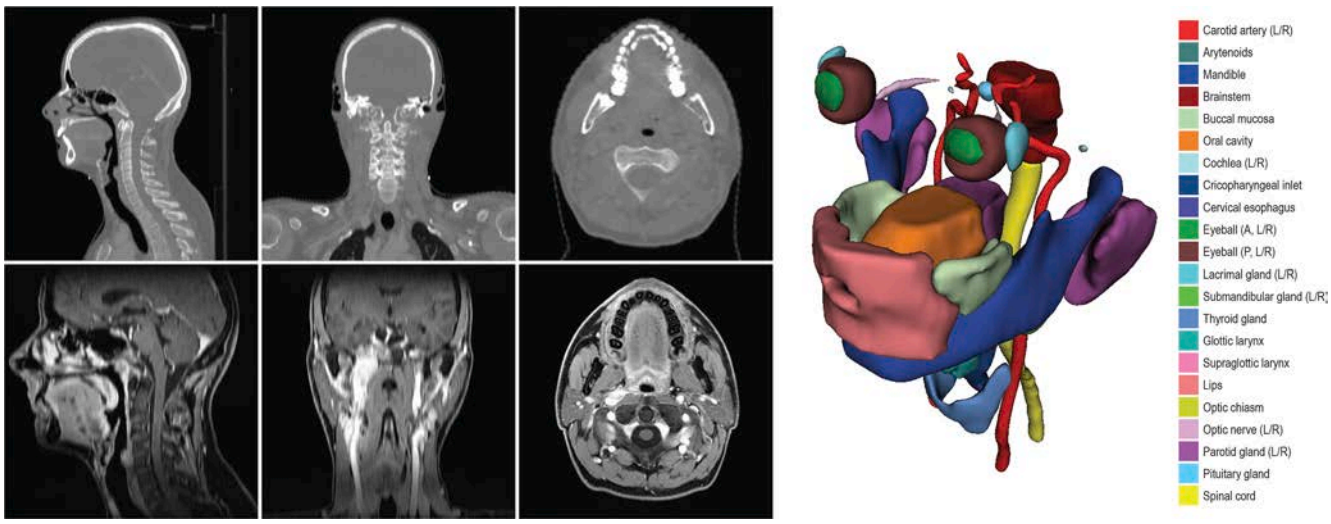

Fig. 1. An example of a case from the publicly available HaN-Seg dataset [36], consisting of one contrast-enhanced computed tomography (left, top) and one T1-weighted magnetic resonance (left, bottom) head and neck image of the same patient, with up to 30 corresponding reference organ-at-risk segmentation masks (right).

Table 1
The teams participating in the HaN-Seg challenge with corresponding short descriptions and main properties (i.e. method architecture, employed modalities and registration type).

Team	Description	Architecture	Modality	Registration
eli1 (E. T.)	A multi-modal nnU-Net [41] (i.e. a self-configuring U-Net [40]) was used with standard parameter settings for segmentation with the default Dice and cross-entropy loss [49,50]. To leverage the knowledge from the MR modality, they focused on an efficient and robust rigid MR-to-CT image registration that was based on SimpleElastix, an extension of the open-source toolbox elastix [51] within the SimpleITK framework [52]. Image preprocessing consisted of unifying image voxel size to CT resolution (i.e. voxel size was made consistent across images).	nnU-Net	CT & MR	Rigid
cwlg102 (C. L. & J. S. K.)	A localization stage was first used, where the OAR bounding boxes were detected in 2D axial CT image cross-sections using YOLOv7 [53], followed by a segmentation stage, where cropped OARs were segmented using the 3D patch-based DynUNet, which mimics the concept of nnU-Net [41] within the MONAI framework [54] with the Dice and binary cross-entropy loss [49,50]. Image preprocessing consisted of resizing all CT images to the axial resolution of 1024 × 1024 pixels and intensity histogram matching normalization. The final model was obtained by six-fold cross-validated ensemble voting on the training set. While they experimented with rigid image registration, their final submission relied exclusively on CT images.	DynUNet(nnU-Net)	CT	Not applied
CHB-QuantIF (Z. M. & R. M.)	Transfer learning was applied on the publicly available STU-Net-B model [55], which is a modified version of nnU-Net [41], pre-trained on 1204 TotalSegmentator CT images [56]. Additionally, the class-adaptive Dice loss function [57] was used to tackle the differences in OAR volumes. Image preprocessing consisted of unifying image voxel size, and rigid registration of MR to CT images using ANTsPy [58].	nnU-Net	CT & MR	Rigid
Mamaa (Y. M. & F.Y.)	A framework named UID-Net was used consisting of a U-Net segmentation backbone [40], where a basic convolution block was augmented with two Inception modules [59] and depthwise separable convolutions [60] to extract features at multiple scales, and enhance non-linearity and feature capabilities. Deep supervision [61] was used to enhance robustness, gradient propagation and feature representation, and Dice loss [49] to handle class imbalance. Image preprocessing consisted of unifying image voxel size, intensity normalization by segmental linear functions [62], and registration by translating the mid-coronal 2D cross-sections of MR images to align with those of CT images.	UID-Net(nnU-Net)	CT & MR	Translation
m.m.gs	The team chose not to participate in the report, and consequently the method description is omitted.			
TurboMiki (M. R. & M.W.)	The U-Net [40] with additional residual connections was used for segmentation that concatenated the CT and MR image into a single framework input, and applied a weighted sum of the Dice and focal loss [49,63]. Image preprocessing consisted of resizing all images to the axial resolution of 512 × 512 pixels.	U-Net	CT & MR	Not applied

CT: computed tomography; MR: magnetic resonance; OAR: organ-at-risk; 2D: two-dimensional; 3D: three-dimensional.

Results

In total, seven teams submitted their methods to the Final Test Phase of the HaN-Seg challenge, however, one team was excluded from further analysis because the submitted method produced trivial segmentation results (i.e. empty segmentation masks or masks without any overlap with or resemblance to the OARs in the HaN region), and one team chose not to participate in the challenge report. The remaining five teams and

their corresponding methods are introduced in Table 1.

The mean performance of each method was calculated from the results obtained on the 14 cases from the test set, and the results for each individual OAR are presented in Table 2 in terms of DSC and Table 3 in terms of HD_{95}. If a method produced an empty segmentation mask, it was assigned a DSC of zero and the maximum HD_{95} over all six methods for that particular organ. Results across all OARs and test cases are shown as box-plots in Fig. 2. In terms of DSC, the overall mean ± standard deviation performance was 76.9 ± 8.4 %, 76.8 ± 9.3 %, 75.1 ± 8.6 %, 75.2 ± 8.4 %, 73.1 ± 14.0 % and 60.9 ± 11.6 %, and the overall median (interquartile range) performance was 81.0 % (18.6 %), 80.2 % (17.2 %), 78.1 % (18.9 %), 78.0 % (18.5 %), 77.5 % (20.2 %) and 67.9 % (27.2 %) for teams eli1, cwlg102, CHB-QuantIF, Mamaa, m.m.gs and TurboMiki, respectively. In terms of HD_{95}, the overall mean ± standard deviation performance was 3.5 ± 2.4 mm, 3.8 ± 3.8 mm, 3.7 ± 2.3 mm, 3.9 ± 2.7 mm, 9.1 ± 18.4 mm and 14.0 ± 16.2 mm, and the overall median (interquartile range) performance was 2.7 mm (2.2 mm), 3.0 mm (2.6 mm), 3.0 mm (2.6 mm), 3.0 mm (2.5 mm), 3.2 mm (4.0 mm) and 5.5 mm (6.5 mm) for teams eli1, cwlg102, CHB-QuantIF, Mamaa, m.m.gs and TurboMiki, respectively. According to the obtained results, the most difficult OARs to segment were the optic chiasm, arytenoids and lacrimal glands, while the best segmentation performance was observed for the mandible, posterior part of the eyeballs, brainstem and thyroid gland.

The statistical ranking results, obtained from computing significance scores and metric-specific ranks, are reported in Table 4. The overall differences in the performance between the first- and second-ranked teams (i.e. eli1 and cwlg102, respectively) were relatively small. However, the differences become more evident when observing the results for specific OARs (Table 2 and Table 3). In fact, when the first-ranked team outperformed the second-ranked team, the differences were larger than for the reverse scenario, especially in the case of HD_{95}. The third-ranked teams (i.e. CHB-QuantIF and Mamaa) closely follow the first two teams, with no statistically significant differences between them. Finally, the two bottom-ranked teams (i.e. m.m.gs and TurboMiki) exhibit a considerable drop in performance, both in terms of DSC (Table 2) and HD_{95} (Table 3).

Discussion

In this study, we present the outcomes of the HaN-Seg computational challenge, where participants were tasked with segmenting 30 OARs from paired CT and MR images. With 42 segmented image pairs for training, methods in the form of Docker containers were evaluated on 14 unseen pairs. The evaluation employed DSC and HD_{95} metrics, with final rankings determined through statistical testing. The winning approach utilized rigid MR to CT registration, achieving 76.9 % and 3.5 mm in terms of mean DSC and mean HD_{95}, respectively, across all OARs. Advancements in various aspects of AI, specifically deep learning methodologies, including training efficiency, data augmentation, feature fusion techniques, and inference methods, have been driving the progress in medical image segmentation. Computational challenges provide a unique opportunity to objectively compare different methods on held-out test datasets, utilizing uniform evaluation metrics and minimizing bias. We structure our discussion of the HaN-Seg results into two primary components: methodological considerations of the submitted approaches and their clinical implications. We conclude by reflecting on the limitations of our study and suggesting potential avenues for future research.

Table 2
The HaN-Seg challenge segmentation results in terms of the mean Dice similarity coefficient (DSC) for each participating team and each organ-at-risk (OAR). The best mean values are in bold.

OAR	DSC: mean ± standard deviation (%)					
	eli1	cwlg102	CHB-QuantIF	Mamaa	m.m.gs	TurboMiki
Carotid artery (L)	82.8 ± 5.1	**85.2 ± 4.5**	79.6 ± 6.0	80.1 ± 5.8	71.2 ± 19.4	63.8 ± 9.5
Carotid artery (R)	85.2 ± 3.3	**86.8 ± 3.5**	82.4 ± 3.8	81.7 ± 5.2	69.8 ± 22.1	63.8 ± 7.3
Arytenoids	59.9 ± 14.3	52.3 ± 23.9	55.4 ± 13.1	**62.1 ± 12.5**	47.7 ± 26.5	39.6 ± 14.2
Mandible	94.3 ± 1.7	**95.0 ± 1.6**	94.2 ± 1.3	93.2 ± 2.0	94.6 ± 2.1	88.1 ± 2.7
Brainstem	**88.5 ± 4.7**	84.9 ± 3.1	86.9 ± 3.4	84.6 ± 3.6	85.4 ± 4.9	79.9 ± 3.4
Buccal mucosa	69.1 ± 8.9	**71.1 ± 7.3**	67.2 ± 9.5	68.0 ± 9.9	70.9 ± 8.4	59.7 ± 11.1
Oral cavity	89.4 ± 4.6	89.4 ± 4.5	**89.6 ± 4.5**	88.9 ± 4.0	87.6 ± 6.3	85.6 ± 4.4
Cochlea (L)	73.4 ± 10.1	**78.8 ± 7.8**	69.1 ± 8.7	72.5 ± 8.6	69.9 ± 14.7	61.6 ± 12.4
Cochlea (R)	74.1 ± 10.6	**78.2 ± 8.1**	68.6 ± 12.2	66.2 ± 12.5	67.4 ± 18.9	58.9 ± 19.1
Cricopharyngeal inlet	**64.5 ± 11.4**	62.8 ± 9.4	62.8 ± 11.2	62.4 ± 11.0	59.9 ± 18.0	55.5 ± 11.0
Cervical esophagus	**63.1 ± 13.2**	62.1 ± 12.3	61.8 ± 11.9	58.3 ± 14.6	54.8 ± 17.5	52.6 ± 15.2
Eyeball (A, L)	76.6 ± 7.5	**79.1 ± 7.0**	77.6 ± 4.8	76.3 ± 8.0	75.0 ± 14.0	62.9 ± 22.4
Eyeball (A, R)	78.5 ± 7.2	**80.6 ± 5.7**	78.6 ± 5.0	78.8 ± 5.1	72.3 ± 20.6	64.9 ± 13.9
Eyeball (P, L)	92.0 ± 2.0	**93.0 ± 1.4**	91.5 ± 1.9	91.5 ± 2.3	92.5 ± 2.2	86.0 ± 6.3
Eyeball (P, R)	91.2 ± 1.6	**92.6 ± 1.4**	91.1 ± 1.4	91.5 ± 1.9	91.7 ± 2.1	84.9 ± 6.3
Lacrimal gland (L)	60.8 ± 14.8	65.3 ± 10.4	61.4 ± 13.6	61.8 ± 9.6	**66.3 ± 10.3**	45.0 ± 16.2
Lacrimal gland (R)	61.9 ± 15.2	**64.2 ± 13.7**	61.1 ± 11.9	59.4 ± 13.2	61.2 ± 16.3	39.6 ± 24.3
Submandibular gland (L)	**85.7 ± 8.2**	82.2 ± 10.8	83.1 ± 10.8	82.7 ± 10.3	82.5 ± 10.5	74.7 ± 10.4
Submandibular gland (R)	84.5 ± 6.5	82.9 ± 7.0	83.1 ± 6.6	83.2 ± 9.0	**85.1 ± 5.3**	73.8 ± 6.6
Thyroid gland	**88.3 ± 4.7**	85.4 ± 15.3	86.9 ± 5.5	86.5 ± 7.0	86.9 ± 6.3	79.3 ± 8.9
Glottic larynx	**75.0 ± 6.9**	67.5 ± 11.2	73.3 ± 6.4	73.1 ± 6.8	70.6 ± 14.1	64.1 ± 11.0
Supraglottic larynx	**80.9 ± 5.1**	79.3 ± 7.0	79.7 ± 6.5	79.4 ± 4.9	79.0 ± 6.1	73.8 ± 8.6
Lips	**74.0 ± 8.0**	73.8 ± 10.2	73.5 ± 8.7	72.1 ± 8.8	71.0 ± 11.9	65.5 ± 8.5
Optic chiasm	44.6 ± 8.6	**45.6 ± 13.2**	41.1 ± 16.2	44.4 ± 10.1	35.3 ± 15.4	31.9 ± 16.1
Optic nerve (L)	67.7 ± 11.8	**71.7 ± 8.7**	64.3 ± 10.0	65.8 ± 11.2	65.8 ± 16.9	47.1 ± 11.0
Optic nerve (R)	72.5 ± 7.5	**75.5 ± 5.3**	67.2 ± 8.3	71.8 ± 5.6	68.5 ± 13.6	0.0 ± 0.0
Parotid gland (L)	**86.7 ± 2.2**	85.1 ± 3.8	86.1 ± 2.2	83.9 ± 3.1	83.0 ± 11.5	76.3 ± 7.8
Parotid gland (R)	**85.7 ± 3.7**	82.1 ± 6.6	84.0 ± 4.1	82.8 ± 4.6	82.1 ± 6.2	72.6 ± 8.3
Pituitary gland	**73.6 ± 9.2**	68.7 ± 12.4	68.0 ± 14.4	71.5 ± 12.7	63.5 ± 22.4	0.0 ± 0.0
Spinal cord	**83.3 ± 2.4**	82.0 ± 3.9	82.4 ± 2.9	80.2 ± 5.0	80.2 ± 7.5	74.6 ± 4.5
Mean performance	**76.9 ± 8.4**	76.8 ± 9.3	75.1 ± 8.6	75.2 ± 8.4	73.1 ± 14.0	60.9 ± 11.6

L: left; R: right; A: anterior; P: posterior.

Table 3
The HaN-Seg challenge segmentation results in terms of the mean 95-percentile Hausdorff distance (HD_{95}) for each participating team and each organ-at-risk (OAR). The best mean values are in bold.

OAR	HD_{95}: mean ± standard deviation (mm)					
	eli1	cwlg102	CHB-QuantIF	Mamaa	m.m.gs	TurboMiki
Carotid artery (L)	3.5 ± 4.6	**3.0 ± 4.2**	4.7 ± 5.2	5.0 ± 5.3	44.7 ± 21.1	17.8 ± 3.6
Carotid artery (R)	**2.7 ± 4.6**	4.3 ± 9.5	3.2 ± 3.1	3.0 ± 4.1	55.7 ± 42.3	27.3 ± 16.4
Arytenoids	3.0 ± 1.4	4.7 ± 4.8	3.6 ± 1.4	**2.8 ± 1.3**	6.2 ± 5.4	5.0 ± 2.5
Mandible	1.3 ± 0.8	**1.2 ± 0.8**	1.4 ± 0.7	1.6 ± 0.8	1.7 ± 1.1	5.2 ± 2.3
Brainstem	**3.9 ± 2.1**	4.7 ± 1.6	4.6 ± 1.8	4.5 ± 1.6	4.8 ± 2.7	5.5 ± 1.3
Buccal mucosa	5.3 ± 2.4	**5.1 ± 2.5**	5.5 ± 2.6	5.8 ± 2.7	5.5 ± 2.4	6.5 ± 2.2
Oral cavity	5.3 ± 2.6	5.4 ± 2.4	**5.0 ± 2.1**	5.8 ± 2.7	15.8 ± 35.3	7.2 ± 2.2
Cochlea (L)	1.4 ± 0.7	**1.3 ± 0.8**	2.0 ± 0.6	1.7 ± 0.8	9.9 ± 30.3	9.6 ± 27.8
Cochlea (R)	1.9 ± 0.9	**1.6 ± 0.9**	2.3 ± 0.7	2.4 ± 1.3	1.9 ± 1.0	2.3 ± 0.9
Cricopharyngeal inlet	6.4 ± 3.5	6.5 ± 3.5	**6.2 ± 3.7**	7.2 ± 6.7	6.4 ± 4.0	7.5 ± 3.7
Cervical esophagus	8.0 ± 4.9	**7.5 ± 3.7**	7.6 ± 3.8	8.1 ± 3.3	7.9 ± 3.9	8.4 ± 3.9
Eyeball (A, L)	**2.3 ± 0.6**	2.4 ± 0.7	2.5 ± 0.5	**2.3 ± 0.7**	2.6 ± 0.9	3.2 ± 1.4
Eyeball (A, R)	2.3 ± 1.2	**2.1 ± 0.7**	2.3 ± 0.7	**2.1 ± 0.6**	2.5 ± 1.4	3.5 ± 1.5
Eyeball (P, L)	1.9 ± 0.5	**1.5 ± 0.3**	1.8 ± 0.6	1.8 ± 0.5	1.6 ± 0.7	2.9 ± 1.4
Eyeball (P, R)	1.9 ± 0.6	**1.6 ± 0.5**	2.1 ± 0.5	1.8 ± 0.5	1.9 ± 0.6	2.8 ± 1.0
Lacrimal gland (L)	3.9 ± 1.7	**3.1 ± 1.2**	**3.1 ± 1.6**	3.4 ± 1.7	3.3 ± 1.5	4.1 ± 1.7
Lacrimal gland (R)	3.7 ± 1.9	**3.6 ± 1.6**	3.8 ± 1.8	3.9 ± 1.7	4.1 ± 2.1	6.1 ± 2.9
Submandibular gland (L)	**3.1 ± 2.4**	6.5 ± 11.3	4.3 ± 4.4	5.1 ± 5.9	7.7 ± 12.4	5.0 ± 2.8
Submandibular gland (R)	3.9 ± 2.8	4.1 ± 2.8	4.1 ± 2.4	4.5 ± 3.2	**3.8 ± 2.1**	6.2 ± 2.3
Thyroid gland	**2.6 ± 2.0**	4.0 ± 7.0	2.9 ± 2.1	3.1 ± 3.2	3.3 ± 2.3	4.7 ± 2.6
Glottic larynx	**2.8 ± 1.2**	3.7 ± 1.9	3.0 ± 1.2	3.4 ± 1.3	14.4 ± 39.9	4.3 ± 2.1
Supraglottic larynx	3.5 ± 1.2	4.2 ± 1.7	**3.4 ± 1.3**	3.6 ± 1.3	4.4 ± 2.0	4.2 ± 1.9
Lips	6.5 ± 3.2	**6.3 ± 3.0**	6.5 ± 2.9	6.9 ± 3.0	8.1 ± 4.1	8.3 ± 4.3
Optic chiasm	4.2 ± 1.7	4.4 ± 2.1	4.4 ± 1.8	**4.1 ± 1.6**	5.4 ± 2.1	6.3 ± 2.5
Optic nerve (L)	2.6 ± 1.5	**2.2 ± 0.9**	2.4 ± 0.8	2.4 ± 0.9	6.4 ± 8.9	5.4 ± 3.3
Optic nerve (R)	2.7 ± 1.9	**2.5 ± 1.6**	2.8 ± 0.9	2.7 ± 0.9	5.5 ± 7.4	28.5 ± 0.0
Parotid gland (L)	**5.1 ± 2.6**	**5.1 ± 2.5**	5.3 ± 3.5	5.9 ± 3.0	8.3 ± 8.4	8.8 ± 3.5
Parotid gland (R)	**5.3 ± 3.1**	6.6 ± 5.1	6.2 ± 3.1	6.4 ± 3.5	6.0 ± 2.4	51.1 ± 36.1
Pituitary gland	**2.1 ± 0.7**	2.4 ± 1.0	2.6 ± 1.1	2.3 ± 1.0	18.6 ± 60.4	159.9 ± 73.2
Spinal cord	2.6 ± 2.4	2.1 ± 1.2	**2.0 ± 0.6**	2.3 ± 1.1	5.9 ± 6.5	3.7 ± 2.6
Mean performance	**3.5 ± 2.4**	3.8 ± 3.8	3.7 ± 2.3	3.9 ± 2.7	9.1 ± 18.4	14.0 ± 16.2

L: left; R: right; A: anterior; P: posterior.

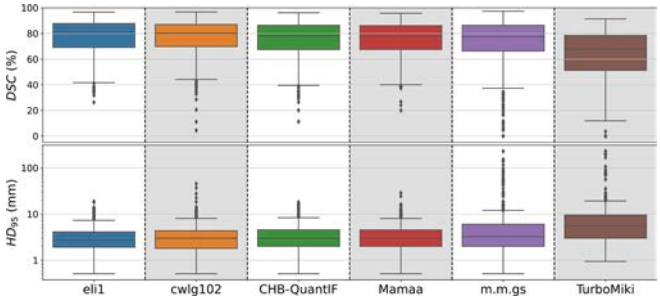

Fig. 2. Box plots comparing all six teams in terms of the Dice similarity coefficient (DSC) and 95th percentile Hausdorff distance (HD_{95}), shown in the top and bottom rows, respectively.

Table 4
The results of the statistical ranking of submitted methods in the HaN-Seg challenge according to the Dice similarity coefficient (DSC) and 95-percentile Hausdorff distance (HD_{95}) significance score (*S*) and corresponding rank (*R*).

Method	DSC		HD_{95}		Ranking	
	S	*R*	*S*	*R*	Aggregate	Final
eli1	4	1	4	1	2	1
cwlg102	4	1	3	2	3	2
CHB-QuantIF	1	3	2	3	6	3*
Mamaa	1	3	2	3	6	3*
m.m.gs	1	3	1	5	8	5
TurboMiki	0	6	0	6	12	6

* Team Mamaa had a higher mean DSC and team CHB-QuantIF had a higher mean HD_{95}, therefore resulting in a shared final rank.

Methodological details

A high unification regarding the deep learning architectures is observed across all teams, all employing a modification of the U-Net architecture [40], predominantly based on the nnU-Net framework [41]. We hypothesize that this is due to the strong inductive bias of the convolutional neural networks (CNNs), which has been identified as particularly beneficial when the training dataset is of moderate size [42]. Moreover, CNNs can rival and outperform the performance of novel attention-based networks such as Transformers or Mamba networks [43]. Last but not least, the U-Net architecture and its variants, such as the nnU-Net framework, are established in the medical imaging community, are easy to access through publicly available implementations, and proved to generate state-of-the-art results on several publicly available datasets used in different biomedical segmentation challenges [41]. Perhaps surprising is that all multi-modal methods used the early fusion approach, i.e. input level channel concatenation of CT and MR modalities [44], whereas hybrid or late fusion might be advantageous to extract high-level features before the fusion [45]. Although not related to CT and MR image fusion, the method proposed by team cwlg102 is the only one using majority voting based on an ensemble of networks trained on different folds of the training set.

The challenge was designed to resemble the challenging real-world clinical scenario, by providing non-registered planning CT and MR images, which included metal and motion-related artifacts. As seen in clinical practice, CT and MR images of the same subject exhibit varying fields-of-view (FoV), with MR modalities commonly featuring a smaller FoV compared to CT due to acquisition intricacies such as MR acquisition time. Consistent with the clinical workflow, participants were tasked with providing OAR segmentation masks in the CT image space,

where voxel sizes were not standardized across all CT images. Before feeding the images into their networks, teams tackled this challenge by either resizing the images to a uniform size or resampling them to achieve consistent voxel spacing, after which they applied cropping. Following inference, segmentation masks underwent resampling to match the physical space of the source CT images.

Clinical implications

The clinical utility of the proposed auto-segmentation tools can be assessed by comparing their performance with the intra- and inter-observer agreement [46,47]. Despite the demanding nature of the problem, the two top-performing methods by teams eli1 and cwlg102 performed better than the inter-observer agreement for the majority of organs, as observed between a junior and a senior expert on CT images from the same dataset as used in the challenge [46]. This indicates that auto-segmentation tools are increasingly applicable in clinical practice, which is reinforced by the recent FDA approval of several auto-contouring tools for OARs [48].

The incorporation of the MR image modality in the auto-segmentation framework remains a challenging task. Besides the fact that the multi-modal setting inherently poses a more computationally demanding training due to the input dimensionality growth, the method needs to robustly perform multi-modal (MR-CT) registration and be able to handle missing information in order to achieve complete end-to-end automation. For example, in case the MR image has a smaller FoV than the CT image, one trivial solution is to simply pad the image, however, the segmentation model still needs to be trained to produce high-quality segmentation masks, e.g. by relying solely on information provided by the CT modality. Indeed, this is not trivial – team cwlg102 tested their models with or without the inclusion of rigidly registered MR images and opted to completely disregard the MR image modality and use an ensemble that relies solely on CT images.

While oncologists find MR images an invaluable source of information for OAR delineation, there are several potential reasons why CNNs did not benefit from MR images, which can be observed from the solid performance of team cwlg102 that relied on CT images only, in comparison to other teams that relied on both CT and MR images (Table 4). Firstly, CNN performance is expected to significantly depend on the registration accuracy, as modality misalignments may hinder CNNs in accurately extracting meaningful intensity patterns from the images. Rigid registration selected by some teams due to its computation efficiency cannot result in perfect CT and MR alignment. While being more accurate, non-rigid registration can be sometimes too time-consuming and therefore violate the 15-minute time frame allocated for each segmentation by the challenge setting. Secondly, the moderate size of the dataset could be a limiting factor. Given the complexities of registration and the presence of various artifacts affecting CT and MR image quality, it is plausible that a larger dataset is required to effectively train CNNs on MR-CT pairs compared to CT images alone. Thirdly, the early fusion approach employed by the CNNs may not be optimal for extracting meaningful features from MR images. Layer- or late-fusion techniques could potentially be more effective in mitigating registration-related misalignments and extracting valuable features from both modalities [44]. This suggests that alternative fusion strategies may be worth exploring to improve the utilization of the MR modality to aid model performance. Lastly, although obtained by aid of co-registered MR images, the reference 3D OAR delineation masks were defined in the coordinate system of CT images [36], which may inherently cause to favor CT over MR images.

Limitations and future perspectives

One limitation of this challenge is the moderate dataset size, consisting of 42 training and 14 test cases. However, to ensure objectivity and rigorous comparison between the competing teams, rankings were determined through statistical testing, with 420 measurements per method (i.e. 14 test cases multiplied by 30 OARs) for each metric. This is illustrated by a third place tie between the team CHB-QuantIF and team Mamaa, a scenario that would not occur if rankings were based solely on mean values. Another potential limitation was the restricted execution time, capped at 15 min per test case, and the available inference VRAM limited to 16 GB. Feedback from the participants suggested that time constraints led to the prevalence of rigid over deformable registration. Additionally, limited GPU VRAM necessitated the adjustment of patch sizes and overlap ratios during sliding window inference. Nonetheless, while these resource limitations required participants to adapt their methods, they reflect practical constraints often encountered in clinical settings, where time and computational resources may be limited as well.

Despite the limitations, the challenge unequivocally demonstrated that auto-segmentation tools can achieve performance comparable to or even surpassing inter-observer agreement, thus underscoring their clinical utility. However, while quantitative analyses provide valuable insights, multi-center qualitative studies are imperative to validate the practical applicability of these methods in clinical practice. We hope that this challenge will inspire future endeavors on larger datasets of paired CT and MR multi-modal images, facilitating extensive experiments on how deep learning methods can leverage both modalities amidst various artifacts and image quality variations.

Conclusion

The HaN-Seg challenge, alongside the publicly available HaN-Seg dataset [36], establishes a benchmark for objectively comparing new methods in OAR segmentation within the HaN region from CT, MR, or combined CT and MR images. The obtained performance results, coupled with the prevalence of U-Net architectures among the participants, suggest that innovations in method architecture may not be paramount for method performance, and that more emphasis is instead put on method robustness and improving worst-case performance. We hope that our findings will encourage and facilitate the development of novel general-purpose multi-modal methods for semantic segmentation.

Sources of support

This study was approved by the Ethics Committee of the Institute of Oncology Ljubljana, Slovenia, under ERID-EK/139, and supported by the Slovenian Research and Innovation Agency (ARIS) under grants J2-1732, P2-0232 and P3-0307, and in part by the Novo Nordisk Foundation under grant NFF20OC0062056.

CRediT authorship contribution statement

Gašper Podobnik: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Bulat Ibragimov:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Elias Tappeiner:** Software. **Chanwoong Lee:** Software. **Jin Sung Kim:** Software. **Zacharia Mesbah:** Software. **Romain Modzelewski:** Software. **Yihao Ma:** Software. **Fan Yang:** Software. **Mikołaj Rudecki:** Software. **Marek Wodziński:** Software. **Primož Peterlin:** Project administration, Data curation, Conceptualization. **Primož Strojan:** Project administration, Data curation, Conceptualization. **Tomaž Vrtovec:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, 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.

Acknowledgments

This study was approved by the Ethics Committee of the Institute of Oncology Ljubljana, Slovenia, under ERID-EK/139, and supported by the Slovenian Research and Innovation Agency (ARIS) under grants J2-1732, P2-0232 and P3-0307, and in part by the Novo Nordisk Foundation under grant NFF20OC0062056.

References

[1] Chow L. Head and neck cancer. N Engl J Med 2020;382:50–72. https://doi.org/10.1056/NEJMra1715715.
[2] 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;153:55–66. https://doi.org/10.1016/j.radonc.2020.09.008.
[3] Brouwer C, Steenbakkers R, 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;117:83–90. https://doi.org/10.1016/j.radonc.2015.07.041.
[4] Eekers D, 't Ven L, Roelofs E, et al. The EPTN consensus-based atlas for CT- and MR-based contouring in neuro-oncology. Radiother Oncol 2018;128:37–43. https://doi.org/10.1016/j.radonc.2017.12.013.
[5] Eekers D, Di Perri D, Roelofs E, et al. Update of the EPTN atlas for CT- and MR-based contouring in neuro-oncology. Radiother Oncol 2021;160:259–65. https://doi.org/10.1016/j.radonc.2021.05.013.
[6] Gao Y, Huang R, Yang Y, et al. FocusNetv2: imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images. Med Image Anal 2021;67:101831. https://doi.org/10.1016/j.media.2020.101831.
[7] Nikolov S, Blackwell S, Zverovitch A, 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. https://doi.org/10.2196/26151.
[8] Kawahara D, Tsuneda M, Ozawa S, et al. Stepwise deep neural network (stepwise-net) for head and neck auto-segmentation on CT images. Comput Biol Med 2022;143:105295. https://doi.org/10.1016/j.compbiomed.2022.105295.
[9] Henderson E, Vasquez Osorio E, van Herk M, Brouwer C, Steenbakkers R, Green A. Accurate segmentation of head and neck radiotherapy CT scans with 3D CNNs: consistency is key. Phys Med Biol 2023;68:085003. https://doi.org/10.1088/1361-6560/acc309.
[10] Clark B, Hardcastle N, Johnston L, Korte J. Transfer learning for auto-segmentation of 17 organs-at-risk in the head and neck: bridging the gap between institutional and public datasets. Med Phys 2024;51:4767–77. https://doi.org/10.1002/mp.16997.
[11] Korte J, Hardcastle N, Ng S, Clark B, Kron T, Jackson P. Cascaded deep learning-based auto-segmentation for head and neck cancer patients: organs at risk on T2-weighted magnetic resonance imaging. Med Phys 2021;48:7757–72. https://doi.org/10.1002/mp.15290.
[12] Hague C, McPartlin A, Lee L, et al. An evaluation of MR based deep learning auto-contouring for planning head and neck radiotherapy. Radiother Oncol 2021;158:112–7. https://doi.org/10.1016/j.radonc.2021.02.018.
[13] Paczona V, Capala B, Deák-Karancsi ME, et al. Magnetic resonance imaging-based delineation of organs at risk in the head and neck region. Adv Radiat Oncol 2022;8:101042. https://doi.org/10.1016/j.adro.2022.101042.
[14] Bologna M, Corino V, Cavalieri S, et al. Prognostic radiomic signature for head and neck cancer: development and validation on a multi-centric MRI dataset. Radiother Oncol 2023;183:109638. https://doi.org/10.1016/j.radonc.2023.109638.
[15] Wang Y, Lombardo E, Huang L, et al. Comparison of deep learning networks for fully automated head and neck tumor delineation on multi-centric PET/CT images. Radiat Oncol 2024;19:3. https://doi.org/10.1186/s13014-023-02388-0.
[16] 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.
[17] Lin D, Wahid K, Nelms B, et al. E pluribus unum: prospective acceptability benchmarking from the contouring collaborative for consensus in radiation oncology crowdsourced initiative for multiobserver segmentation. J Med Imaging 2023;10:S11903. https://doi.org/10.1117/1.JMI.10.S1.S11903.
[18] Maier-Hein L, Reinke A, Kozubek M, et al. BIAS: transparent reporting of biomedical image analysis challenges. Med Image Anal 2020;66:101796. https://doi.org/10.1016/j.media.2020.101796.
[19] Pekar V, Allaire S, Kim J, Jaffray D. Head and neck auto-segmentation challenge. MIDAS J 2009;5:5. https://doi.org/10.54294/263mqy.
[20] Pekar V, Allaire S, Kim J, Jaffray D. Head and neck auto-segmentation challenge: segmentation of the parotid glands. In: Medical Image Analysis for the Clinic: A Grand Challenge 2010, MICCAI, Beijing, China, 2010, pp. 273–280.
[21] Raudaschl P, Zaffino P, Sharp G, et al. Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. Med Phys 2017;44:2020–36. https://doi.org/10.1002/mp.12197.
[22] Armato S, Tahir B, Sharp G. AAPM grand challenges symposium: Rtmac. Med Phys 2019;46:e485–6. https://doi.org/10.1002/mp.13589.
[23] Luo X, Fu J, Zhong Y, et al. SegRap2023: a benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma, arXiv (2023) 2312.09576. https://doi.org/10.48550/arXiv.2312.09576.
[24] McGee K, Tyagi N, Bayouth J, et al. Findings of the AAPM Ad Hoc committee on magnetic resonance imaging in radiation therapy: unmet needs, opportunities, and recommendations. Med Phys 2021;48:4523–31. https://doi.org/10.1002/mp.14996.
[25] Goodburn R, Philippens M, Lefebvre T, et al. The future of MRI in radiation therapy: challenges and opportunities for the MR community. Magn Reson Med 2022;88:2592–608. https://doi.org/10.1002/mrm.29450.
[26] Lekshmi R, Manoj G, Sweety G, et al. Comparison of magnetic resonance imaging and CT scan-based delineation of target volumes and organs at risk in the radiation treatment planning of head and neck malignancies. J Med Imaging Radiat Sci 2023;54:503–10. https://doi.org/10.1016/j.jmir.2023.03.034.
[27] Bollen H, Willems S, Wegge M, Maes F, Nuyts S. Benefits of automated gross tumor volume segmentation in head and neck cancer using multi-modality information. Radiother Oncol 2023;182:109574. https://doi.org/10.1016/j.radonc.2023.109574.
[28] Wei Z, Ren J, Korreman S, Nijkamp J. Towards interactive deep-learning for tumour segmentation in head and neck cancer radiotherapy. Phys Imaging Radiat Oncol 2023;25:100408. https://doi.org/10.1016/j.phro.2022.12.005.
[29] Zhong Z, He L, Chen C, et al. Full-scale attention network for automated organ segmentation on head and neck CT and MR images. IET Image Proc 2023;17:660–73. https://doi.org/10.1049/ipr2.12663.
[30] Qi M, Li Y, Wu A, Lu X, Zhou L, Song T. Multi-sequence MR generated sCT is promising for HNC MR-only RT: a comprehensive evaluation of previously developed sCT generation networks. Med Phys 2022;49:2150–8. https://doi.org/10.1002/mp.15572.
[31] Bird D, Speight R, Andersson S, Wingqvist J, Al-Qaisieh B. Deep learning MRI-only synthetic-CT generation for pelvis, brain and head and neck cancers. Radiother Oncol 2024;191:110052. https://doi.org/10.1016/j.radonc.2023.110052.
[32] Boeke S, Mönnich D, van Timmeren J, Balermpas P. MR-guided radiotherapy for head and neck cancer: current developments, perspectives, and challenges. Front Oncol 2021;11:616156. https://doi.org/10.3389/fonc.2021.616156.
[33] Habrich J, Boeke S, Nachbar M, et al. Repeatability of diffusion-weighted magnetic resonance imaging in head and neck cancer at a 1.5 T MR-Linac. Radiother Oncol 2022;174:141–8. https://doi.org/10.1016/j.radonc.2022.07.020.
[34] Huynh E, Boyle S, Campbell J, et al. Toward implementation of MR-guided radiation therapy for laryngeal cancer with healthy volunteer imaging and a custom MR-CT larynx phantom. Med Phys 2022;49:1814–21. https://doi.org/10.1002/mp.15472.
[35] Lombardo E, Rabe M, Xiong Y, et al. Evaluation of real-time tumor contour prediction using LSTM networks for MR-guided radiotherapy. Radiother Oncol 2023;182:109555. https://doi.org/10.1016/j.radonc.2023.109555.
[36] Podobnik G, Strojan P, Peterlin P, Ibragimov B, Vrtovec T. HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset. Med Phys 2023;50:1917–27. https://doi.org/10.1002/mp.16197.
[37] Mackay K, Bernstein D, Glocker B, Kamnitsas K, Taylor A. A review of the metrics used to assess auto-contouring systems in radiotherapy. Clin Oncol 2023;35:354–69. https://doi.org/10.1016/j.clon.2023.01.016.
[38] Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods 2024;21:195–212. https://doi.org/10.1038/s41592-023-02151-z.
[39] Wiesenfarth M, Reinke A, Landman B, et al. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci Rep 2021;11:2369. https://doi.org/10.1038/s41598-021-82017-6.
[40] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, Vol. 9351 of LNCS, Springer, Munich, Germany 2015;234-41. https://doi.org/10.1007/978-3-319-24574-4_28.
[41] Isensee F, Jaeger P, Kohl S, Petersen J, Maier-Hein K. nnUNet: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11. https://doi.org/10.1038/s41592-020-01008-z.
[42] Liu Z, Mao H, Wu C, et al. A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022;11976-86. https://doi.org/10.1109/CVPR52688.2022.01167.
[43] Isensee F, Wald T, Ulrich C, et al. nnU-Net Revisited: A call for rigorous validation in 3D medical image segmentation, arXiv (2024) 2404.09556. https://doi.org/10.48550/arXiv.2404.09556.
[44] Podobnik G, Strojan P, Peterlin P, Ibragimov B, Vrtovec T. Multimodal CT and MR segmentation of head and neck organs-at-risk, in: 26th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, Vol. 14223 of LNCS, Springer, Vancouver, Canada, 2023;745-55. https://doi.org/10.1007/978-3-031-43901-8_71.
[45] Zhang Y, Sidibé D, Morel O, Mériaudeau F. Deep multimodal fusion for semantic image segmentation: a survey. Image Vis Comput 2021;105:104042. https://doi.org/10.1016/j.imavis.2020.104042.
[46] Podobnik G, Ibragimov B, Peterlin P, Strojan P, Vrtovec T. vOARiability: interobserver and intermodality variability analysis in OAR contouring from head and neck CT and MR images. Med Phys 2024;51:2175–86. https://doi.org/10.1002/mp.16924.
[47] Nielsen C, Lorenzen E, Jensen K, et al. Interobserver variation in organs at risk contouring in head and neck cancer according to the DAHANCA guidelines.

Radiother Oncol 2024;197:110337. https://doi.org/10.1016/j.radonc.2024.110337.
[48] Strolin S, Santoro M, Paolani G, et al. How smart is artificial intelligence in organs delineation? Testing a CE and FDA-approved Deep-Learning tool using multiple expert contours delineated on planning CT images, Front Oncol 2023;13;1089807. https://doi.org/10.3389/fonc.2023.1089807.
[49] Sudre C, Li W, Vercauteren T, Ourselin S, Cardoso M. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support - DLMIA 2017 and MLCDS 2017, Vol. 10553 of LNCS, Springer, Québec City, QC, Canada, 2017;240-48. https://doi.org/10.1007/978-3-319-67558-9_28.
[50] Mao A, Mohri M, Zhong Y. Cross-entropy loss functions: theoretical analysis and applications. In: 40th International Conference on Machine Learning - ICML 2023, Honolulu, HI, USA: PMLR, 2023;202;23803-28. https://doi.org/10.48550/arXiv.2304.07288.
[51] Klein S, Staring M, Murphy K, Viergever M, Pluim J. elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2010;29:196–205. https://doi.org/10.1109/TMI.2009.2035616.
[52] Lowekamp B, Chen D, Ibáñez L, Blezek D. The design of SimpleITK. Front Neuroinf 2013;7:45. https://doi.org/10.3389/fninf.2013.00045.
[53] Wang C-Y, Bochkovskiy A, Liao H-Y. YOLOv7: trainable bag-of freebies sets new state-of-the-art for real-time object detectors, arXiv (2023) 2207.02696. https://doi.org/10.48550/arXiv.2207.02696.
[54] Cardoso M, Li W, Brown R, et al. MONAI: an open-source framework for deep learning in healthcare, arXiv (2022) 2211.02701. https://doi.org/10.48550/arXiv.2211.02701.
[55] Huang Z, Wang H, Deng Z, et al. STU-Net: scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training, arXiv 2023;2304.06716. https://doi.org/10.48550/arXiv.2304.06716.
[56] Wasserthal J, Breit H-C, Meyer M, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol Artif Intell 2023;5:e230024. https://doi.org/10.1148/ryai.230024.
[57] Tappeiner E, Welk M, Schubert R. Tackling the class imbalance problem of deep learning-based head and neck organ segmentation. Int J Comput Assisted Radiol Surg 2022;17:2103–11. https://doi.org/10.1007/s11548-022-02649-5.
[58] Tustison N, Cook P, Holbrook A, et al. The ANTsX ecosystem for quantitative biological and medical imaging. Sci Rep 2021;11:9068. https://doi.org/10.1038/s41598-021-87564-6.
[59] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition – CVPR 2015. Boston, MA, USA: IEEE; 2015. p. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.
[60] Chollet F. Xception: deep learning with depthwise separable convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition - CVPR 2017. Honolulu, HI, USA: IEEE. 2017;1800-7. https://doi.org/10.1109/CVPR.2017.195.
[61] Lee C-Y, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. In: 18th International Conference on Artificial Intelligence and Statistics - AISTATS 2015. San Diego, CA, USA: PMLR. 2015;38;562-70. https://doi.org/10.48550/arXiv.1409.5185.
[62] Lei W, Mei H, Sun Z, et al. Automatic segmentation of organs-at-risk from head-and-neck CT using separable convolutional neural network with hard-region-weighted loss. Neurocomputing 2021;442:184–99. https://doi.org/10.1016/j.neucom.2021.01.135.
[63] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;43:318–27. https://doi.org/10.1109/TPAMI.2018.2858826.