Liu *et al. BioMedical Engineering OnLine* (2023) 22:104
https://doi.org/10.1186/s12938-023-01159-y

BioMedical Engineering
OnLine

REVIEW

Open Access

Deep learning algorithm performance in contouring head and neck organs at risk: a systematic review and single-arm meta-analysis

Peiru Liu[1,2†], Ying Sun[1†], Xinzhuo Zhao[3] and Ying Yan[1*]

†Peiru Liu, Ying Sun have contributed equally to this work.

*Correspondence:
yanyingdoctor@sina.com

[1] General Hospital of Northern Theater Command, Department of Radiation Oncology, Shenyang, China
[2] Beifang Hospital of China Medical University, Shenyang, China
[3] Shenyang University of Technology, School of Electrical Engineering,, Shenyang, China

Abstract

Purpose: The contouring of organs at risk (OARs) in head and neck cancer radiation treatment planning is a crucial, yet repetitive and time-consuming process. Recent studies have applied deep learning (DL) algorithms to automatically contour head and neck OARs. This study aims to conduct a systematic review and meta-analysis to summarize and analyze the performance of DL algorithms in contouring head and neck OARs. The objective is to assess the advantages and limitations of DL algorithms in contour planning of head and neck OARs.

Methods: This study conducted a literature search of Pubmed, Embase and Cochrane Library databases, to include studies related to DL contouring head and neck OARs, and the dice similarity coefficient (DSC) of four categories of OARs from the results of each study are selected as effect sizes for meta-analysis. Furthermore, this study conducted a subgroup analysis of OARs characterized by image modality and image type.

Results: 149 articles were retrieved, and 22 studies were included in the meta-analysis after excluding duplicate literature, primary screening, and re-screening. The combined effect sizes of DSC for brainstem, spinal cord, mandible, left eye, right eye, left optic nerve, right optic nerve, optic chiasm, left parotid, right parotid, left submandibular, and right submandibular are 0.87, 0.83, 0.92, 0.90, 0.90, 0.71, 0.74, 0.62, 0.85, 0.85, 0.82, and 0.82, respectively. For subgroup analysis, the combined effect sizes for segmentation of the brainstem, mandible, left optic nerve, and left parotid gland using CT and MRI images are 0.86/0.92, 0.92/0.90, 0.71/0.73, and 0.84/0.87, respectively. Pooled effect sizes using 2D and 3D images of the brainstem, mandible, left optic nerve, and left parotid gland for contouring are 0.88/0.87, 0.92/0.92, 0.75/0.71 and 0.87/0.85.

Conclusions: The use of automated contouring technology based on DL algorithms is an essential tool for contouring head and neck OARs, achieving high accuracy, reducing the workload of clinical radiation oncologists, and providing individualized, standardized, and refined treatment plans for implementing "precision radiotherapy". Improving DL performance requires the construction of high-quality data sets and enhancing algorithm optimization and innovation.

Keywords: Deep learning, Organs at risk, Head and neck cancer, Contouring, Systematic review, Meta-analysis

Introduction

Head and neck cancer is a highly malignant cancer with significant morbidity and mortality rates globally [1]. It comprises various types, such as nasopharyngeal, oropharyngeal, hypopharyngeal, and laryngeal cancers, all of which differ significantly in terms of clinical features, treatment, and prognosis [1]. The epidemiology of head and neck cancer differs based on ethnicity, nationality, gender, and age groups [2–4]. Tobacco and alcohol consumption, along with HPV infection, represent the primary risk factors for head and neck cancer. Specifically, HPV-16 seropositivity is associated with a nearly 30-fold higher risk of pharyngeal cancer [5–7]. Radiotherapy is a critical component of comprehensive treatment for head and neck cancer. Techniques such as 3D conformal radiotherapy, stereotactic radiotherapy, and intensity-modulated radiotherapy are commonly used for treating head and neck cancer. However, radiotherapy can also result in adverse effects, including xerostomia [8, 9], dysphagia [10, 11] and radiation osteonecrosis [12]. Accurate OAR contouring in the head and neck region can significantly reduce the incidence of adverse effects of radiotherapy, which will directly impact tumor control and long-term prognosis.

Accurate contouring of head and neck OARs has become a challenge for clinicians with the advent of precision radiotherapy. Presently, manual contouring of OARs is burdened with two challenges: reduced accuracy and increased time cost. Meanwhile, it has been demonstrated [13–16] that contouring of the target area varies among clinicians with different levels of experience, even for the same case. This could be due to the low pixel contrast on CT or MRI images and the clinicians' comprehension of the target area. Over-segmentation of OARs will make it difficult to optimize radiotherapy dose, while under-segmentation will subject OARs to an excessively high radiation dose, leading to irreversible side-effects on the patient's body (Table 1). The effectiveness of radiotherapy for cancer patients is seriously dependent on how accurately the OARs are contoured. The contouring of OARs is a labor-intensive task, and clinicians need to contour the cancerous foci and OARs layer by layer based on CT or MRI images (Figs. 1, 2), which will consume a lot of time [17–19]. With the development of deep learning technology, doctors have made great progress in contouring the target area, reducing radiological damage to patients, and even evaluating and improving patient prognosis [20–22].

Despite the fact that there is a relatively large body of research literature focusing on this subject, there is still a lack of comprehensive review and meta-analysis of this area. The purpose of this meta-analysis is to review, summaries and analyses the performance

Table 1 Dose limits and complication probability of head and neck radiotherapy OARs

OARs	Max dose	Possibility of toxicity	Most serious complication
Brain stem	54 Gy	< 5%	Neuropathy or necrosis
Spinal cord	50 Gy	0.2%	Radiological spinal cord injury
Mandible	60–65 Gy	< 5%	Radioactive osteonecrosis
Optic nerve/optic chiasm	< 55 Gy	< 3%	Nerve damage
Eyes (single)	< 35 Gy	–	Blindness
Parotid gland (single)	< 20 Gy	< 20%	Parotid function chronically less than 25%
Submandibular gland	< 35 Gy	–	Xerostomia

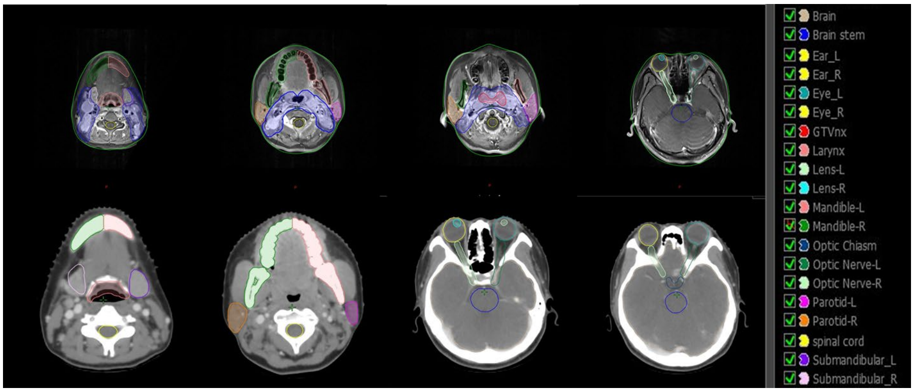
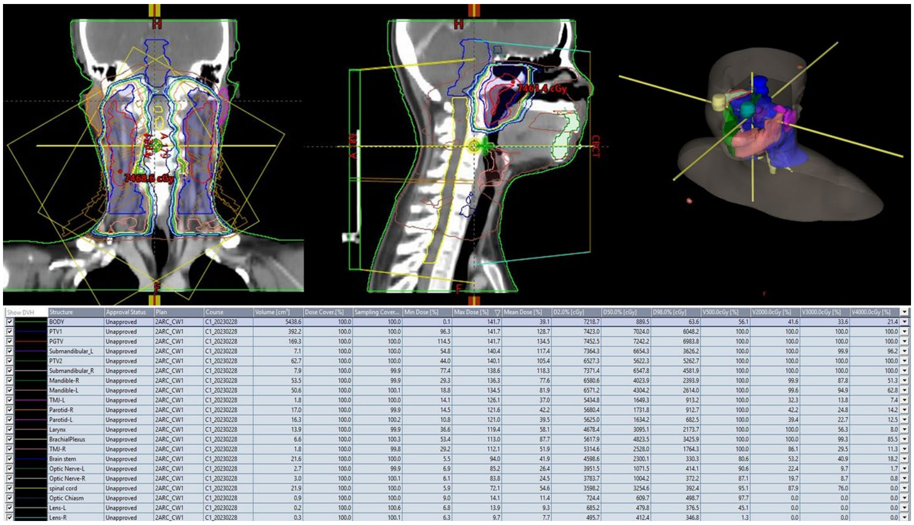


Fig. 1 Sample CT/MRI image slices with OARs contours



Show DVH	Structure	Approval Status	Plan	Course	Volume [cm³]	Dose Cover.[%]	Sampling Cover...	Min Dose [%]	Max Dose [%]	Mean Dose [%]	D2.0% [cGy]	D50.0% [cGy]	D98.0% [cGy]	V500.0cGy [%]	V2000.0cGy [%]	V3000.0cGy [%]	V4000.0cGy [%]
☑	BODY	Unapproved	2ARC_CW1	C1_20230228	5438.6	100.0	100.0	0.1	141.7	39.1	7218.7	889.5	63.6	56.1	41.6	33.6	21.4
☑	PTV1	Unapproved	2ARC_CW1	C1_20230228	392.2	100.0	100.0	96.3	141.7	128.7	7423.0	7024.0	6048.2	100.0	100.0	100.0	100.0
☑	PGTV	Unapproved	2ARC_CW1	C1_20230228	169.3	100.0	100.0	114.5	141.7	134.5	7452.5	7242.2	6983.8	100.0	100.0	100.0	100.0
☑	Submandibular_L	Unapproved	2ARC_CW1	C1_20230228	7.1	100.0	100.0	54.8	140.4	117.4	7364.3	6654.3	3626.2	100.0	100.0	99.9	96.2
☑	PTV2	Unapproved	2ARC_CW1	C1_20230228	62.7	100.0	100.0	44.0	140.1	105.4	6527.3	5622.3	5262.7	100.0	100.0	100.0	100.0
☑	Submandibular_R	Unapproved	2ARC_CW1	C1_20230228	7.9	100.0	99.9	77.4	138.6	118.3	7371.4	6547.8	4581.9	100.0	100.0	100.0	100.0
☑	Mandible-R	Unapproved	2ARC_CW1	C1_20230228	53.5	100.0	99.9	29.3	136.3	77.6	6580.6	4023.9	2393.9	100.0	99.9	87.8	51.3
☑	Mandible-L	Unapproved	2ARC_CW1	C1_20230228	50.6	100.0	100.1	18.8	134.5	81.9	6571.2	4304.2	2614.0	100.0	99.6	94.9	62.8
☑	TMJ-L	Unapproved	2ARC_CW1	C1_20230228	1.8	100.0	100.0	14.1	126.1	37.0	5434.8	1649.3	913.2	100.0	32.3	13.8	7.4
☑	Parotid-R	Unapproved	2ARC_CW1	C1_20230228	17.0	100.0	99.9	14.5	121.6	42.2	5680.4	1731.8	912.7	100.0	42.2	24.8	14.2
☑	Parotid-L	Unapproved	2ARC_CW1	C1_20230228	16.3	100.0	100.2	10.8	121.0	39.5	5625.0	1634.2	682.5	100.0	39.4	22.7	12.5
☑	Larynx	Unapproved	2ARC_CW1	C1_20230228	13.9	100.0	99.9	36.6	119.4	58.1	4678.4	3095.1	2173.7	100.0	100.0	56.3	8.0
☑	BrachialPlexus	Unapproved	2ARC_CW1	C1_20230228	6.6	100.0	100.3	53.4	113.0	87.7	5617.9	4823.5	3425.9	100.0	100.0	99.3	85.5
☑	TMJ-R	Unapproved	2ARC_CW1	C1_20230228	1.8	100.0	99.8	29.2	112.1	51.9	5314.6	2528.0	1764.3	100.0	86.1	29.5	11.3
☑	Brain stem	Unapproved	2ARC_CW1	C1_20230228	21.6	100.0	100.0	5.5	94.0	41.9	4598.6	2300.1	330.3	80.6	53.2	40.9	18.2
☑	Optic Nerve-L	Unapproved	2ARC_CW1	C1_20230228	2.7	100.0	99.9	6.9	85.2	26.4	3951.5	1071.5	414.1	90.6	22.4	9.7	1.7
☑	Optic Nerve-R	Unapproved	2ARC_CW1	C1_20230228	3.0	100.0	100.1	6.1	83.8	24.5	3783.7	1004.2	372.2	87.1	19.7	8.7	0.8
☑	spinal cord	Unapproved	2ARC_CW1	C1_20230228	21.9	100.0	100.0	5.9	72.1	54.6	3598.2	3254.6	392.4	95.1	87.9	76.0	0.0
☑	Optic Chiasm	Unapproved	2ARC_CW1	C1_20230228	0.9	100.0	100.0	9.0	14.1	11.4	724.4	609.7	498.7	97.7	0.0	0.0	0.0
☑	Lens-L	Unapproved	2ARC_CW1	C1_20230228	0.2	100.0	100.6	6.8	13.9	9.3	685.2	479.8	376.5	45.1	0.0	0.0	0.0
☑	Lens-R	Unapproved	2ARC_CW1	C1_20230228	0.3	100.0	100.1	6.3	9.7	7.7	495.7	412.4	346.8	1.3	0.0	0.0	0.0

Fig. 2 Radiotherapy plans for head and neck cancer and OARs

of the DL technique for segmenting OARs in the head and neck region. The good image recognition and segmentation performance shown by the DL algorithm is promising. This study will focus on the following issues: the current status of DL algorithm segmentation in head and neck OARs, the influence of image modality and image type on the segmentation performance of the DL algorithm, and a systematic review of the key issues affecting the DL algorithm performance in the contouring of head and neck OARs and future development directions.

Methods

Search strategy

This single-arm meta-analysis is conducted based on the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [23]. We searched the literature in Pubmed, Embase, and Cochrane Library up to November 14, 2022, using the form of MeSH Terms + Entry Terms to search relevant literature. The search strategy is (Deep Learning OR Neural Networks) AND (Segmentation) AND

(Head and Neck Neoplasms) AND (Organs at Risk). The detailed search strategy can be found in Additional file 1: Table S1.

Selection criteria and data extraction

Studies with detailed OAR segmentation data, or studies able to calculate DSCs and their 95% confidence intervals (CIs) from other available data, are eligible. Studies with the following characteristics should be excluded: 1. studies on non-human species; 2. non-algorithmic studies, or contouring using mature segmentation software; 3. conference abstracts, reviews, book chapters, meta-analyses, editorials, duplicate literature; 4. non-English language studies; 5. lack of data; 6. unavailable literature; and 7. irrelevant studies.

Before data extraction, this study designed a data extraction form in conjunction with existing studies that will focus on the following data: 1. first author and year of publication; 2. country of first author attribution; 3. single-center or multicenter study; 4. prospective or retrospective study; 5. algorithm name; 6. image modality; 7. image type; 8. total number of patients; 9. test set sample size; and 10. head and neck OARs and corresponding DSC values and CI or standard deviations (SD).

Quality assessment and risk of bias

Accurately described detail of the development and validation of clinical prediction models is necessary to adequately assess the generalizability of specific studies. Therefore, the *Checklist for Artificial Intelligence in Medical Imaging* (CLAIM) for DL [24] is chosen as the standard for assessing the quality of the literature, see Additional file 1: Table S2 for materials related to the CLAIM criteria.

For risk of bias, the *Prediction Model Risk of Bias Assessment Tool* (PROBAST), which focuses on methodological evaluation, is selected [25], and PROBAST is a risk of bias assessment tool for predictive model studies published by the Cochrane Assist Group in 2019. Moreover, it has been revised to be more appropriate for DL studies and its related fields with reference to Frizzell et al. [26], see Additional file 1: Table S3 for materials related to the PROBAST criteria.

Quality assessment of the literature and risk of bias assessment is carried out by a single person, and in case of uncertainty about the results, the decision is discussed with a second person.

Statistical analysis

DSC is a quantitative analysis metric for evaluating graphic similarity in the field of computer vision. To calculate DSC, the computer first discrete the pixel points on the image and set the weight of each pixel point to 1. $AT \cup GT$ represents the sum of the weights of artificial intelligence target (AT) and ground truth (GT), $AT \cap GT$ represents the weight sum of the overlapping parts in AT and GT. The DSC takes values between [0, 1]. The closer the DSC is to 1, the better the fit between the AT and the GT area. In general, a DSC greater than 0.80 is considered to be a high similarity, a DSC greater than 0.70 is considered to be a moderate similarity, and a DSC less than 0.70 is considered to be a similarity that needs to be improved:

$$DSC = \frac{2(AT \cap GT)}{AT \cup GT}$$

The pooled effect size calculations, funnel plots and Egger's test for publication bias in this study were all done using Stata17. The calculation of the pooled effect size is based on the mean (mean) and 95% CI. For studies that did not report 95% CI data, reference is made to the methods in the *Cochrane handbook for systematic reviews of intervention*, using the test set sample size (*n*), the DSC mean (mean), the DSC SD were used to transform the data.

Higgins I^2 is used to test for heterogeneity between studies, with $I^2 < 25\%$ considered to have no heterogeneity, $25\% \le I^2 < 50\%$ considered to have low heterogeneity, $50\% \le I^2 < 75\%$ considered to have moderate heterogeneity, and $I^2 \ge 75\%$ considered to have high heterogeneity. The study selected either a fixed-effects model or a random-effects model based on the value of heterogeneity in the included literature.

Data analysis for this study were all performed using GraphPad Prism 9, and Student *t* test is used for comparison between groups. $p < 0.05$ (*) is considered a statistically significant difference, and vice versa (ns).

Results

Study selection and characteristics

With reference to the search strategy, a literature search is conducted in Pubmed, Embase and Cochrane Library for this study. 149 articles were retrieved and 106 articles were identified after excluding duplicates. After screening and detailed review and evaluation, a total of 22 articles were included in the meta-analysis (Fig. 3), involving 6,099 patients.

Among the 22 articles, 10 studies (45.45%) are from China, 5 studies (22.73%) are from the USA, 2 studies (9.09%) are from the Netherlands, 1 study (4.55%) is from Australia, 1 study (4.55%) is from the UK, 1 study (4.55%) is from Korea, 1 study (4.55%) is from Austria, and 1 study (4.55%) is from Austria. 2 studies (9.09%) are multicenter studies and 20 (90.91%) are single center studies. 18 studies (81.82%) perform contouring on CT images, 3 studies (13.64%) perform contouring on MRI images and 1 study (4.55%) on CT and MRI, respectively, and two DL models are trained. 5 studies (21.74%) use 2D images for contouring, 15 studies (65.22%) use 3D images for contouring, 1 study (4.55%) use 2.5D images for contouring and 2 studies (9.09%) do not specify the image type. 22 studies (100%) use internal validation sets to validate the algorithm performance and 8 studies (36.36%) use external validation sets. The detailed characteristics of the included literature can be found in Table 1 and the original data tables of the included literature can be found in Table 2.

Results of the meta-analysis

There are many OARs in head and neck region, and DSC is selected as an effect size for meta-analysis of the results of four categories (12 in total) of organs at risk from each study. Central nervous system (CNS): brainstem, spinal cord. Bony structures: mandible. Visual organs: right and left optic nerve, right and left eye, optic chiasm. Glandular structures: right and left parotid glands, right and left submandibular glands. Pooled

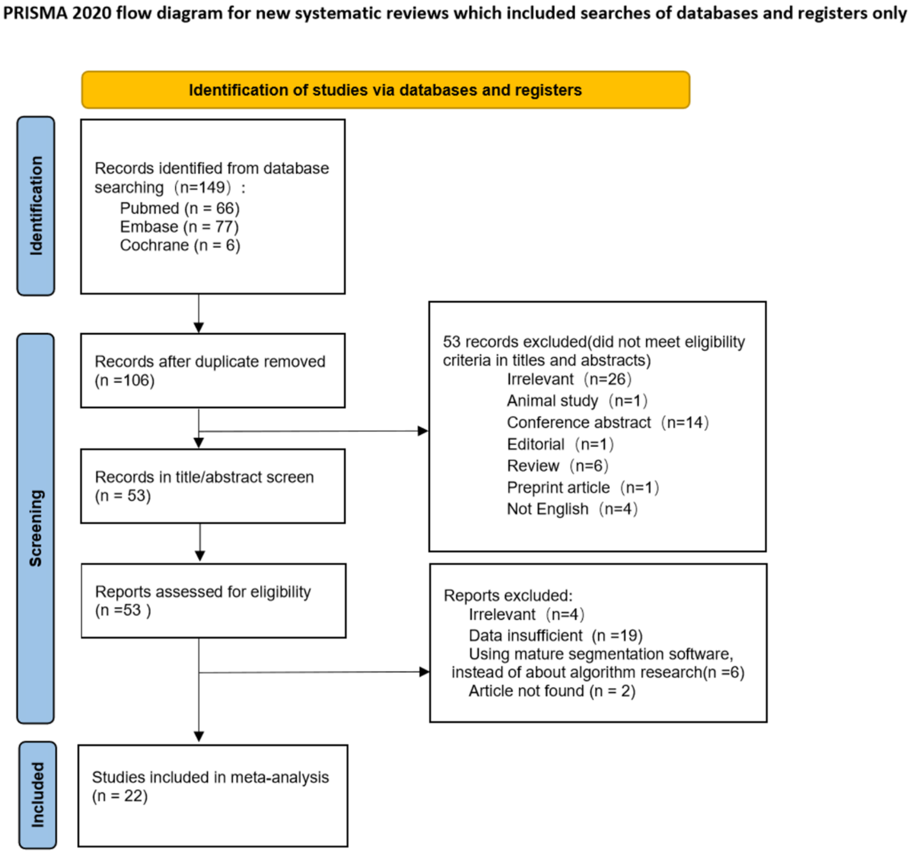

PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only
Identification of studies via databases and registers
Identification
Screening
Included
Records identified from database searching (n=149) :
Pubmed (n = 66)
Embase (n = 77)
Cochrane (n = 6)
Records after duplicate removed (n =106)
Records in title/abstract screen (n = 53)
Reports assessed for eligibility (n =53)
Studies included in meta-analysis (n = 22)
53 records excluded(did not meet eligibility criteria in titles and abstracts)
Irrelevant (n=26)
Animal study (n=1)
Conference abstract (n=14)
Editorial (n=1)
Review (n=6)
Preprint article (n=1)
Not English (n=4)
Reports excluded:
Irrelevant (n=4)
Data insufficient (n =19)
Using mature segmentation software, instead of about algorithm research(n =6)
Article not found (n = 2)

Fig. 3 PRISMA flowchart of eligible studies selection process

effect sizes, 95% CI can be found in Table 3. Forest plots for pooled effect size calculations can be found in Additional file 1: Fig. S1 (A–L).

CNS organs

Brainstem: A total of 21 models from 20 studies presented results for the brainstem, with a pooled DSC effect size of 0.87 (95% CI 0.85–0.89) and a Higgins $I^2 = 98.1\%$.

Spinal cord: A total of 10 models from 10 studies presented results for the spinal cord, with a pooled DSC effect size of 0.83 (95% CI 0.81–0.85) and a Higgins $I^2 = 95.2\%$.

Bony structures

Mandible: A total of 19 models from 18 studies presented results for the spinal cord, with a pooled DSC effect size of 0.92 (95% CI 0.91–0.93) and a Higgins $I^2 = 98.5\%$.

Visual organs

Eye: A total of 9 models from 9 studies presented results for the left and right eyes, with pooled DSC effect sizes of 0.90 (95% CI 0.88–0.91) and Higgins $I^2 = 96.4\%$ for the left eye and 0.90 (95% CI 0.88–0.92) and Higgins $I^2 = 97.7\%$ for the right eye, respectively.

Optic nerve: A total of 17 models from 16 studies presented results for the left and right optic nerve, with pooled DSC effect sizes of 0.71 (95% CI 0.68–0.75), Higgins

Table 2 Characteristics of the included studies

Study ID	Country	Single center/ Multicenter	Retrospective/ Prospective research	Algorithm architecture name	Imaging modality	Imaging type	Internal validation	External validation	Amount of patient		
									Total	Dev	Test
Dai et al. (2021) [27]	America	Single center	Retrospective	MS–RCNN	MRI	3D	√	√	118	95	23
Tao et al. (2019) [28]	China	Single center	Retrospective	Boosting ResNets	CT	2D	√	–	140	120	20
Korte et al. (2021) [29]	Australia	Single center	Retrospective	–	MRI	–	√	–	41	31	10
Oktay et al. (2020) [17]	UK	Multicenter	Retrospective	Modified 3D U-Net	CT	3D	√	√	186	166	20
Ye et al. (2022) [18]	China	Multicenter	Retrospective	UaNet	CT	3D	√	√	502	176	326
Chan et al. (2019) [30]	America	Single center	Retrospective	LL-CNN	CT	3D	√	–	200	180	20
Chen et al. (2021) [23]	China	Single center	Retrospective	WBNet	CT	3D	√	√	180	150	30
Liang et al. (2019) [31]	China	Single center	Retrospective	ODS net	CT	2D	√	–	185	139	46
Kim et al. (2021) [32]	Korea	Single center	Retrospective	Modified FC-DenseNet	CT	2D	√	–	100	80	20
Gao et al. (2021) [33]	China	Single center	Retrospective	FocusNetv2	CT	3D	√	√	1164	1044	120
Nuo et al. (2018) [34]	China	Single center	Retrospective	FCNN + SRM	CT	3D	√	–	32	22	10
Fang et al. (2021) [35]	China	Single center	Retrospective	U-net	CT	2.5D	√	–	1000	800	200
V van Dijk et al. (2020) [36]	Netherlands	Single center	Retrospective	DLC	CT	3D	√	–	693	589	104
Dai et al. (2022) [37]	America	Single center	Retrospective	R-CNN	MRI	–	√	–	60	48	12
Tappeiner et al. (2019) [38]	Austria	Single center	Retrospective	HighRes3DNets	CT	3D	√	–	39	25	14
Gou et al. (2020) [39]	China	Single center	Retrospective	SCSA-Net	CT	3D	√	–	48	38	10
Nuo et al. (2019) [40]	China	Single center	Retrospective	SC-GAN-DenseNet	CT/MRI	3D	√	√	57	37	20
Zhang et al. (2021) [41]	China	Single center	Retrospective	–	CT	3D	√	–	170	150	20
Zhao et al. (2021) [42]	America	Single center	Retrospective	Weaving attention U-net	CT	3D	√	√	115	90	25
Liang et al. (2020) [43]	Not mentioned	Single center	Retrospective	–	CT	2D	√	√	48	33	15
Koo et al. (2022) [44]	America	Single center	Retrospective	–	CT	2D	√	–	864	–	–
Rooij et al. (2019) [20]	Netherlands	Single center	Retrospective	3D U-Net	CT	3D	√	–	157	142	15

Table 3 Raw data of the included studies

Study ID	Brain stem (±SD)	Spinal cord (±SD)	Mandible (±SD)	Optic nerve (Left) (±SD)	Optic nerve (Right) (±SD)	Eyes (Left) (±SD)	Eyes (Right) (±SD)	Optic chiasm (±SD)	Parotid gland (Left) (±SD)	Parotid gland (Right) (±SD)	Submandibular gland (Left) (±SD)	Submandibular gland (Right) (±SD)
Dai et al. (2021) [27]	0.93 ± 0.02	–	0.97 ± 0.01	0.8 ± 0.07	0.8 ± 0.08	–	–	0.73 ± 0.15	0.89 ± 0.05	0.88 ± 0.06	0.88 ± 0.05	0.85 ± 0.11
Tao et al. (2019) [28]	–	–	–	0.8994 ± 0.0431	0.8994 ± 0.0431	–	–	–	0.9188 ± 0.0351	0.9188 ± 0.0351	–	–
Korte et al. (2021) [29]	–	–	0.957 ± 0.023	–	–	–	–	–	0.86 ± 0.067	0.857 ± 0.063	0.83 ± 0.032	0.785 ± 0.123
Oktay et al. (2020) [17]	0.85 ± 0.037	0.84 ± 0.038	–	–	–	0.929 ± 0.016	0.931 ± 0.015	–	0.879 ± 0.038	0.878 ± 0.043	0.878 ± 0.043	0.875 ± 0.023
Ye et al. (2022) [18]	0.816 ± 0.053	0.815 ± 0.098	0.855 ± 0.135	0.676 ± 0.086	0.67 ± 0.097	0.0851 ± 0.132	0.862 ± 0.098	0.598 ± 0.158	0.832 ± 0.058	0.827 ± 0.062	0.792 ± 0.089	0.777 ± 0.092
Chan et al. (2019) [30]	0.89 ± 0.03	0.87 ± 0.03	0.91 ± 0.09	–	–	–	–	–	0.85 ± 0.03	0.86 ± 0.05	0.84 ± 0.01	0.85 ± 0.1
Chen et al. (2021) [23]	0.87 ± 0.03	–	0.94 ± 0.01	0.76 ± 0.06	0.75 ± 0.06	0.92 ± 0.03	0.93 ± 0.02	0.64 ± 0.15	0.85 ± 0.04	0.85 ± 0.04	0.82 ± 0.07	0.82 ± 0.08
Liang et al. (2019) [31]	0.896 ± 0.03	–	0.913 ± 0.035	0.661 ± 0.1	0.717 ± 0.1	0.932 ± 0.04	0.936 ± 0.03	–	0.852 ± 0.05	0.85 ± 0.05	–	–

Table 3 (continued)

Study ID	Brain stem (± SD)	Spinal cord (± SD)	Mandible (± SD)	Optic nerve (Left) (± SD)	Optic nerve (Right) (± SD)	Eyes (Left) (± SD)	Eyes (Right) (± SD)	Optic chiasm (± SD)	Parotid gland (Left) (± SD)	Parotid gland (Right) (± SD)	Submandibular gland (Left) (± SD)	Submandibular gland (Right) (± SD)
Kim et al. (2021) [32]	0.87 ± 0.02	0.82 ± 0.04	0.95 ± 0.01	0.7 ± 0.07	0.72 ± 0.07	0.91 ± 0.02	0.91 ± 0.02	0.53 ± 0.16	0.84 ± 0.04	0.85 ± 0.04	0.83 ± 0.06	0.81 ± 0.1
Gao et al. (2021) [33]	0.8926 ± 0.0317	0.826 ± 0.0532	0.9211 ± 0.0155	0.6876 ± 0.0998	0.7332 ± 0.0884	0.8928 ± 0.0195	0.8895 ± 0.0234	0.6115 ± 0.1244	0.8505 ± 0.0547	0.8687 ± 0.0382	–	–
Nuo et al. (2018) [34]	0.8697 ± 0.0295	–	0.936 ± 0.0121	0.6531 ± 0.0575	0.6889 ± 0.0471	–	–	0.5835 ± 0.1028	0.08387 ± 0.0287	0.8346 ± 0.0234	0.767 ± 0.0731	0.8131 ± 0.0645
Fang et al. (2021) [35]	0.82 ± 0.062	0.796 ± 0.089	–	0.63 ± 0.145	0.638 ± 0.14	0.863 ± 0.076	0.86 ± 0.082	–	0.782 ± 0.093	0.785 ± 0.092	–	–
V van Dijk et al. (2020) [36]	0.84 ± 0.04	0.87 ± 0.06	0.94 ± 0.01	–	–	–	–	–	0.84 ± 0.04	0.83 ± 0.05	0.77 ± 0.12	0.78 ± 0.1
Dai et al. (2022) [37]	0.89 ± 0.06	0.77 ± 0.15	0.82 ± 0.1	0.67 ± 0.11	0.68 ± 0.11	0.89 ± 0.07	0.89 ± 0.05	0.61 ± 0.14	0.85 ± 0.06	0.86 ± 0.05	–	–
Tap-peiner et al. (2019) [38]	0.82 ± 0.04	–	0.91 ± 0.02	0.64 ± 0.08	0.63 ± 0.06	–	–	0.42 ± 0.17	0.8 ± 0.1	0.81 ± 0.08	–	–
Gou et al. (2020) [39]	0.88 ± 0.02	–	0.94 ± 0.01	0.72 ± 0.05	0.71 ± 0.05	–	–	0.61 ± 0.06	0.87 ± 0.03	0.86 ± 0.05	0.78 ± 0.06	0.81 ± 0.07

Table 3 (continued)

Study ID	Brain stem (±SD)	Spinal cord (±SD)	Mandible (±SD)	Optic nerve (Left) (±SD)	Optic nerve (Right) (±SD)	Eyes (Left) (±SD)	Eyes (Right) (±SD)	Optic chiasm (±SD)	Parotid gland (Left) (±SD)	Parotid gland (Right) (±SD)	Submandibular gland (Left) (±SD)	Submandibular gland (Right) (±SD)
Nuo et al. (2019) (model 1) [40]	0.8672 ± 0.029	–	0.9391 ± 0.013	0.6638 ± 0.048	0.6991 ± 0.043	–	–	0.5916 ± 0.097	0.8549 ± 0.017	0.8577 ± 0.024	0.8065 ± 0.05	0.8186 ± 0.049
Nuo et al. (2019) (model 2) [40]	0.9157 ± 0.0285	–	0.8164 ± 0.0444	0.7165 ± 0.0446	0.9631 ± 0.0658	–	–	0.5892 ± 0.0722	0.8648 ± 0.0501	0.8248 ± 0.0534	–	–
Zhang et al. (2021) [41]	0.87 ± 0.04	–	0.89 ± 0.02	0.73 ± 0.09	0.72 ± 0.09	0.89 ± 0.03	0.88 ± 0.03		0.71 ± 0.12	0.77 ± 0.07	–	–
Zhao et al. (2021) [42]	0.92 ± 0.03	0.89 ± 0.04	0.95 ± 0.04	0.76 ± 0.07	0.78 ± 0.09	–	–	0.74 ± 0.18	0.86 ± 0.06	0.88 ± 0.05	0.87 ± 0.07	0.85 ± 0.09
Liang et al. (2020) [43]	0.923 ± 0.1	–	0.941 ± 0.07	0.738 ± 0.046	0.734 ± 0.051	–	–	0.713 ± 0.083	0.882 ± 0.013	0.87 ± 0.015	0.815 ± 0.029	0.8 ± 0.034
Koo et al. (2022) [44]	0.88 ± 0.02	0.81 ± 0.07	0.87 ± 0.04	–	–	–	–	–	0.83 ± 0.11	0.83 ± 0.07	0.81 ± 0.1	0.83 ± 0.06
Rooij et al. (2019) [20]	0.64 ± 0.16	–	–	–	–	–	–	–	0.83 ± 0.02	0.82 ± 0.02	0.82 ± 0.07	0.81 ± 0.13

$I^2 = 97.4\%$ for the left optic nerve and 0.74 (95% CI 0.70–0.78), Higgins $I^2 = 97.8\%$ for the right optic nerve, respectively.

Optic chiasm: A total of 13 models from 12 studies presented results for optic chiasm, with a pooled DSC effect size of 0.62 (95% CI 0.59–0.65) and Higgins $I^2 = 84.7\%$.

Glandular structures

Parotid glands: A total of 23 models from 22 studies presented results for the left and right parotid glands, with pooled DSC effect sizes of 0.85 (95% CI 0.84–0.86) and Higgins $I^2 = 94.5\%$ for the left parotid gland and 0.85 (95% CI 0.83–0.86) and Higgins $I^2 = 94.4\%$ for the right parotid gland, respectively.

Submandibular glands: A total of 15 models from 15 studies presented results for the left and right submandibular glands with combined DSC effect sizes of 0.82 (95% CI 0.81–0.84), Higgins $I^2 = 92.4\%$ for the left submandibular gland and 0.82 (95% CI 0.80–0.94) for the right submandibular gland with Higgins $I^2 = 93.5\%$.

Publication bias

Publication bias is evaluated qualitatively using funnel plots and quantitatively using the Egger test. The funnel plot for the bias analysis can be found in Additional file 1: Fig. S2 (A–L). No publication bias is detected in the Egger test for the four categories of organs ($p > 0.05$), see Table 3 for the results.

Subgroup analysis: comparison of contours on CT and MRI images

In this study, four representative organs (brainstem, mandible, left optic nerve, left parotid gland) were selected among the four types of OARs for study [Additional file 1: Fig. S3 (A–H)]. For the DL segmentation performance of DL on CT and MRI, the pooled effect sizes for the four types of organs in the studies using CT images for segmentation were 0.86 (95% CI 0.85–0.88), 0.92 (95% CI 0.91–0.93), 0.71 (95% CI 0.67–0.75), 0.84 (95% CI 0.83–0.86). The pooled effect sizes for the four types of organs in studies using MRI images for segmentation were 0.92 (95% CI 0.90–0.94), 0.90 (95% CI 0.84–0.95), 0.73 (95% CI 0.66–0.80), and 0.87 (95% CI 0.86–0.89), respectively. Among the organs' contours in the two types of image modalities, the difference in brainstem is statistically significant ($p = 0.0139$), suggesting that DL is able to better contour the brainstem on MRI images. The segmentation result of the mandible, left optic nerve and left parotid gland is somewhat different (Fig. 4A) but did not show a statistically significant difference between the two modalities ($p > 0.05$).

Subgroup analysis: comparison of contours on 2D and 3D modalities

This study also investigated the performance of DL in contouring the four organs mentioned above in different image types [Additional file 1: Fig. S4 (A–H)]. For DL segmentation performance on 2D and 3D modalities, the pooled effect sizes for the four types of organs in the study using 2D modalities for segmentation were 0.88 (95% CI 0.87–0.90), 0.92 (95% CI 0.87–0.96), 0.75 (95% CI 0.64–0.86), 0.87 (95% CI 0.84–0.89). The pooled effect sizes for the four types of organs in studies using 3D modalities for segmentation were: 0.87 (95% CI 0.84–0.89), 0.92 (95% CI 0.91–0.93), 0.71 (95% CI 0.68–0.74), 0.85 (95% CI 0.84–0.86). DL contours the brainstem, left optic nerve and left parotid gland

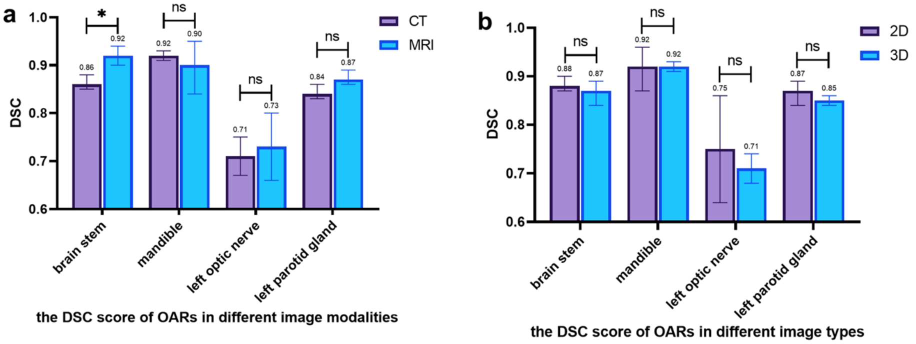


Fig. 4 Bar chart of OARs DSC score in head and neck cancer patients of different image modalities and different image types

Table 4 Pooled dice similarity coefficient and Egger test of the publication bias of DL segmentation model

	References	Pooled estimate	95% CI	I^2	Egger's bias	Egger's 95% CI	Egger's *P* value
Brain stem	[17, 18, 20, 23, 27, 30–44]	0.87	0.855 to 0.885	98.10%	− 0.363099	− 6.434718 to 5.70852	0.902
Spinal cord	[17, 18, 30, 32, 33, 35–37]	0.83	0.815 to 0.855	95.20%	1.172798	− 9.669338 to 12.01493	0.809
Mandible	[18, 23, 27, 29–34, 36, 37, 39–44]	0.92	0.907 to 0.930	98.50%	− 4.291492	− 10.1736 to 1.590612	0.142
Eye (Left)	[17, 18, 23, 31–33, 35, 37, 41]	0.90	0.882 to 0.914	96.40%	0.3947261	− 7.959603 to 8.749055	0.914
Eye (Right)	[17, 18, 23, 31–33, 35, 37, 41]	0.90	0.881 to 0.917	97.70%	− 1.326596	− 14.50229 to 11.8491	0.819
Optic nerve (Left)	[18, 23, 27, 28, 31–35, 37–43]	0.71	0.677 to 0.748	97.40%	1.666572	− 5.986573 to 9.319716	0.649
Optic nerve (Right)	[18, 23, 27, 28, 31–35, 37–43]	0.74	0.698 to 0.776	97.80%	4.376849	− 4.080193 to 12.83389	0.287
Optic chiasm	[18, 23, 27, 32–34, 37–40, 42, 43]	0.62	0.586 to 0.648	84.70%	0.4959904	− 2.725629 to 3.717609	0.741
Parotid gland (Left)	[17, 18, 20, 23, 27–44]	0.85	0.836 to 0.860	94.50%	− 0.3678269	− 4.289085 to 3.553431	0.847
Parotid gland (Right)	[17, 18, 20, 23, 27–44]	0.85	0.835 to 0.859	94.40%	0.1090494	− 3.900165 to 4.118264	0.955
Submandibular gland (Left)	[17, 18, 20, 23, 27, 29, 30, 32, 34, 36, 39, 40, 42–44]	0.82	0.806 to 0.839	92.40%	− 1.411071	− 4.497547 to 1.675404	0.341
Submandibular gland (Right)	[17, 18, 20, 23, 27, 29, 30, 32, 34, 36, 39, 40, 42–44]	0.82	0.796 to 0.843	93.50%	− 0.2339747	− 4.336639 to 3.868689	0.904

better on 2D images than on 3D images (Fig. 4B). The mandible had the same results on both types of images. All four types of organs did not show a statistical difference ($p > 0.05$) between the two types of images, Table 4.

Quality assessment and risk of bias

The six sections of the CLAIM criteria are presented as percentages in Fig. 5A. In the title/abstract section, 97.7% of the studies clearly and accurately described the type of artificial intelligence (AI), study design protocol, etc. 2.3% of the studies do not clearly specify these elements. In the "Introduction" section, all studies (100%) have described the disciplinary background, research objectives and research hypotheses. In the "Methods" section, 59.7% of the studies accurately provide detailed descriptions of the AI architecture, data sources, and training process, while 40.3% of the

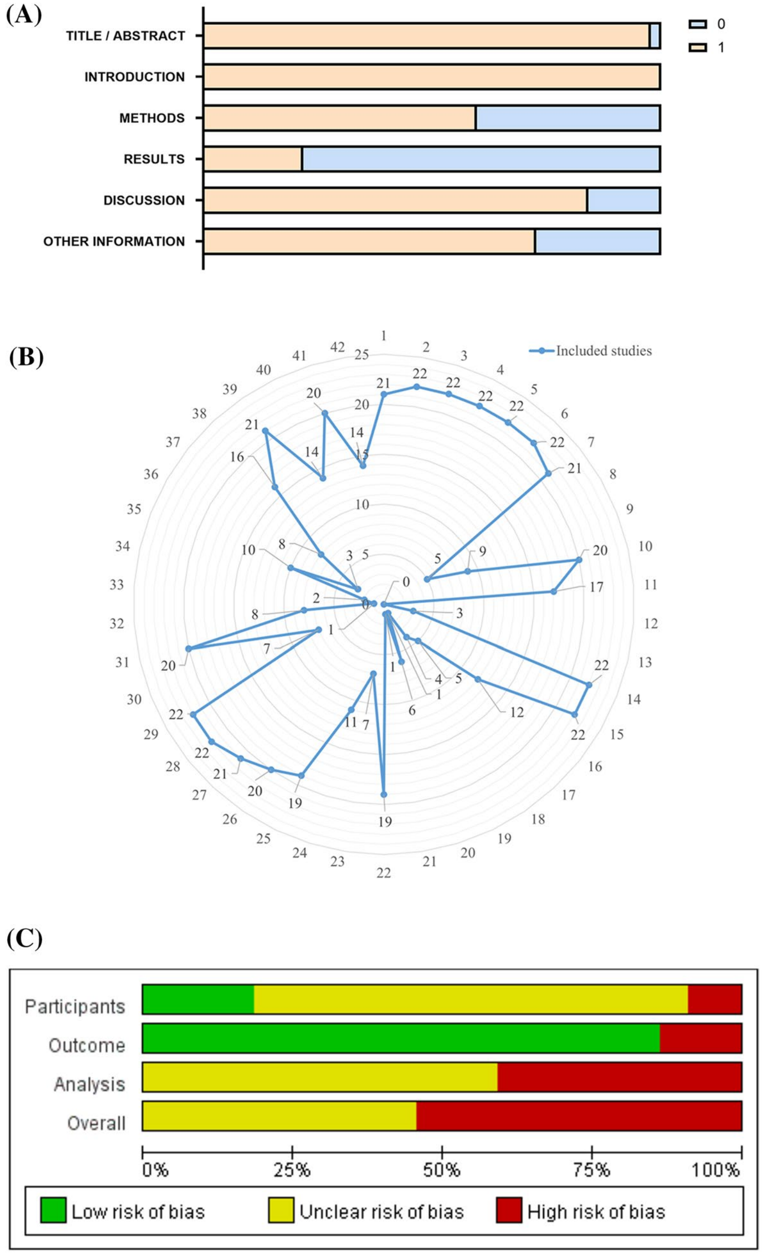
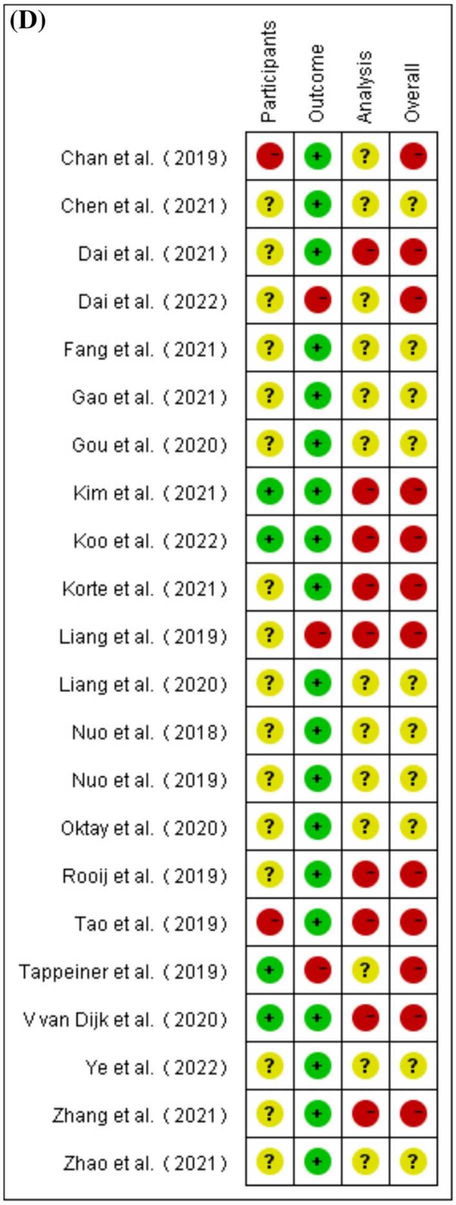


Fig. 5 **A** Summary of CLAIM assessments of included studies. **B** Number of included studies meeting each CLAIM criterion. **C** Risk of bias graph according to PROBAST. **D** Risk of bias summary according to PROBAST



	Participants	Outcome	Analysis	Overall
Chan et al. (2019)	−	+	?	−
Chen et al. (2021)	?	+	?	?
Dai et al. (2021)	?	+	−	−
Dai et al. (2022)	?	−	?	−
Fang et al. (2021)	?	+	?	?
Gao et al. (2021)	?	+	?	?
Gou et al. (2020)	?	+	?	?
Kim et al. (2021)	+	+	−	−
Koo et al. (2022)	+	+	−	−
Korte et al. (2021)	?	+	−	−
Liang et al. (2019)	?	−	−	−
Liang et al. (2020)	?	+	?	?
Nuo et al. (2018)	?	+	?	?
Nuo et al. (2019)	?	+	?	?
Oktay et al. (2020)	?	+	?	?
Rooij et al. (2019)	?	+	−	−
Tao et al. (2019)	−	+	−	−
Tappeiner et al. (2019)	+	−	?	−
V van Dijk et al. (2020)	+	+	−	−
Ye et al. (2022)	?	+	?	?
Zhang et al. (2021)	?	+	−	−
Zhao et al. (2021)	?	+	?	?

Fig. 5 continued

studies do not provide detailed descriptions of the data sources, pre-processing steps, or how to handle missing data. In the "Results" section, 78.2% of the studies are unclear about the inclusion/exclusion criteria for researchers, simply state the source of the CT/MRI images of the included patients, or lack an accurate assessment of the performance of the model, do not analyze cases that are incorrectly contoured. In the "Discussion" section, 84.1% of the studies comment on the limitations of this study, while 15.9% omit this element. For other information, 72.7% of the studies indicate information, such as the location, where the full study protocol could be accessed. Compliance with the CLAIM criteria for the 22 studies include in the meta-analysis ranged from 50% to 71.4%, with a mean of 61.0%. The number of studies meeting the

42 criteria in the CLAIM criteria can be found in Fig. 5B. Detailed results for the CLAIM criteria can be found in Additional file 1: Table S4.

About half (54.5%) of the studies show a high risk of bias according to the PROBAST assessment, Fig. 5C, D. The main source of high risk of bias is that the analysis section did not provide an accurate and comprehensive assessment of DL, including failure to assess metrics, such as specificity and sensitivity or failure to report model over-fitting, under-fitting and solutions. The risk of bias is unclear in less than half (45.5%) of the studies, mainly because the inclusion/exclusion criteria for the cases included in the study are not detailed. Detailed results of PROBAST can be found in Additional file 1: Table S5.

Discussion

DL has the ability to produce high-precision contours of head and neck OARs automatically.

In this study, it is found that DL has the capacity to generate highly precise contours in the automatic contouring of head and neck OARs. Overall, DL can attain a high level of similarity (DSC > 0.8) for CNS organs, bony structures, visual organs (eyes) and glandular structures, and a moderate level of similarity (DSC > 0.7) for the optic nerve in visual organs, while the ability to contour the optic chiasm needs to be improved (DSC < 0.7) (Fig. 6).

Radiation therapy for head and neck cancer is often associated with various radiotoxic reactions; these include optic nerve damage [27], cognitive deficits [45], and central nervous demyelinating lesions [46]. This requires the clinicians to strike a balance between maximizing the extent of tumor control and minimizing toxic effects, where even small differences in contouring may result in a difference in dose [47]. As the radiotherapy process progresses, the anatomy of head and neck region will change dramatically [48–50]. The location and shape of the tumor and surrounding OARs will change

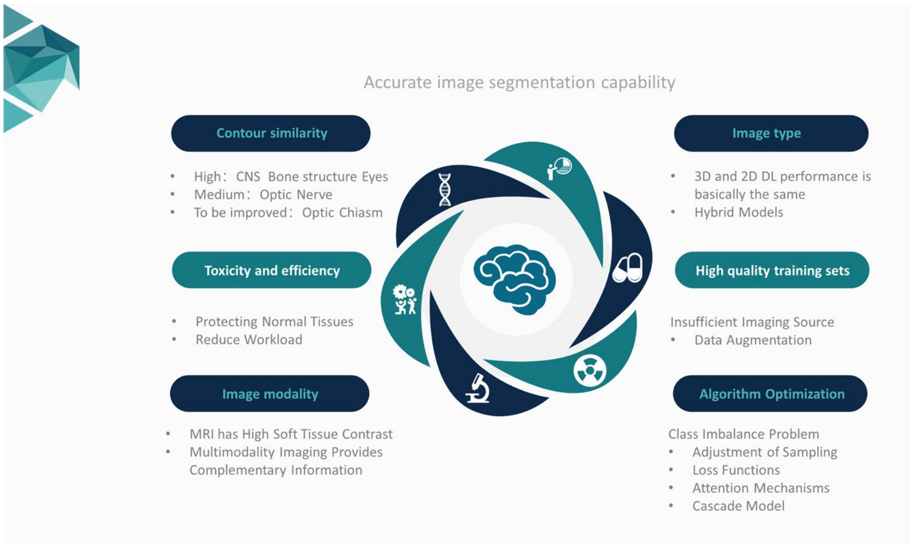

Accurate image segmentation capability
Contour similarity
High：CNS Bone structure Eyes
Medium：Optic Nerve
To be improved：Optic Chiasm
Toxicity and efficiency
Protecting Normal Tissues
Reduce Workload
Image modality
MRI has High Soft Tissue Contrast
Multimodality Imaging Provides Complementary Information
Image type
3D and 2D DL performance is basically the same
Hybrid Models
High quality training sets
Insufficient Imaging Source
Data Augmentation
Algorithm Optimization
Class Imbalance Problem
Adjustment of Sampling
Loss Functions
Attention Mechanisms
Cascade Model

Fig. 6 Contouring similarity and optimization directions for accurate image segmentation algorithms

as a result of the exposure to the radiation. Moreover, the gradient of radiation dose distribution in image-guided radiotherapy changes drastically, and if no corresponding adjustments are made to the dose distribution according to the changes in the lesion and surrounding organs, damage to normal tissues will be exacerbated, while the tumor is not well-controlled [51].

Image modality is an important factor affecting the performance of the DL algorithm, and multi-modality images can provide more accurate automatic contouring

DL's ability to segment CNS organs, visual organs and glandular structures on MRI images superior to that of similar algorithms on CT images, although this does not produce a statistical difference. This suggests that MRI is better equipped to segment soft tissue.

It is found that DL's ability to contour bony structures on MRI images is lessened in comparison with CT, which aligns with Tong et al.'s findings [40]. The bone cortex's low water content results in a low signal on MRI sequences, whereas bone tissue can strongly influence the ray beam attenuation on CT images, leading to a high-density signal. Due to the imaging limitations of single-modality images, DL faces challenges in extracting numerous imaging histological features from CT or MRI [27]. This severely restricts the accuracy of contouring head and neck OARs, which in turn, might have implicational effects on radiotherapy planning. Ibragimov et al. [52] also found that the convolutional neural network (CNN)-based DL algorithm is highly capable of identifying organs with clear boundaries on CT images, and for organs, such as optic chiasm, which are not well-defined, may require additional information to aid in contouring. Kieselmann et al. [53] are exploring the creation of sMRI image synthesized from CT by an algorithm based on generative adversarial networks. sMRI has the advantage of providing good complementary information on soft and bone tissues, and compared to image segmentation on CT alone, sMRI has a significant improvement in predicting optic chiasm, the cochlea and other organs [27, 54]. Multimodality images can provide additional imaging information for the accurate contouring of OARs.

Image type is not a key factor in algorithm performance

The creation of 3D deep learning models necessitates numerous training parameters, leading to considerable computational overhead and potential overfitting hazards. [55, 56]. Due to hardware limitations, 3D DL model neural network depth is typically shallower than that of 2D DL models. This results in a reduced ability of 3D DL models to extract features and contour individual CT/MRI images, which explains why there is no significant difference in algorithm performance across image types. Furthermore, 2D DL models are fast, computationally efficient, and independent of layer thickness. [57, 58]. Medical images are often stored in 3D format in computers, and 3D DL models can efficiently utilize the correlation information from several contiguous images to provide more precise anatomical details and lesion features, therefore, overcoming the deficiency of information amongst body layers that is present in 2D DL models [27]. In general, 3D DL models produce more uniform, intricate, and lifelike contour of OARs. These models are capable of accurately modeling organs that have relatively stable anatomical positions [43]. 3D DL models have many advantages that are currently the focus of attention in the field of image segmentation. However, it is clear that 3D DL models do not currently demonstrate superior performance to 2D DL models.

To enhance the accuracy of segmentation for OARs with respect to image type, Fang et al. [35] applied a 2.5D U-Net model for OARs segmentation. For central slice information prediction, 2.5D images also entail the use of adjacent slices as input, even though the convolution kernel remains in 2D. 2.5D DL models enable the extraction of surrounding 3D information, while also reducing computational complexity, making them more efficient than traditional 3D CNNs [59]. Nuo et al. integrated shape representation models into 3D DL networks to predict images, along with a priori OAR shape features. [40]. In addition to single image types, there is ongoing research on hybrid 2D–3D CNN models. Various studies have implemented 2D–3D hybrid neural networks for organ segmentation [60–62], combining the semantic information of single slices extracted by 2D methods and the contextual semantic information extracted by 3D methods to achieve better segmentation results. Lee et al. [63] incorporated migration learning into organ segmentation. All of these schemes offer potential research ideas for accurately segmenting head and neck OARs. It is essential to emphasize that achieving accurate segmentation requires adequate pre-processing of medical images, irrespective of whether the algorithm segmentation performance is enhanced by the image modality or type. Operations such as removing artifacts, normalizing data, and aligning images can reduce the likelihood of inaccurate segmentation and facilitate image analysis. [64].

Building a high-quality training set and enhancing innovation in the optimization of the algorithm are developing directions to further improve the performance of the algorithm

High-quality training data are a prerequisite for DL algorithms to achieve accurate predictions [40, 44]. A high-quality training data set is often a simpler and more effective means of enhancing of DL algorithms than a low-quality yet high-volume training data set [18]. Although DL models are robust to the noise of image data labels, Rolnick D's study showed a significant negative correlation between the amount of noise and the performance of automatic segmentation algorithms [65]. High-quality data sets are expensive to create, requiring a clinician's medical background, a significant amount of time and effort, among other factors. To overcome the challenge of limited access to such data sets, data augmentation has emerged as a potential solution. This involves generating variations of the original image by rotating, panning, cropping, and applying other techniques such as grayscale perturbation, scaling, and stretching to enhance diversity in the training data set. [34, 53]. Edward [66] performed data augmentation using limited data and evaluated the segmentation effect of a custom model (3D CNN) on a small data set, and the algorithm yielded an average surface distance of only 0.81 mm for the brainstem. Zhao et al. [67] used a principal component analysis model to randomly deform the original CT image to produce new data, and data augmentation provided small-sample-high-quality variants of the contours for DL. Asma Amjad et al. [68] used the adaptive spatial resolution method to improve the problem of low default spatial resolution ($2 \times 2 \times 2$ mm^3) for identifying small organs, and a higher resolution of ($1 \times 1 \times 2$ mm^3) for tissues, such as the optic nerve. Results on the test set showed that the DL algorithm contoured improved DSC values for all nine OARs, including the brainstem, inner ear and optic nerve.

The class imbalance problem is a major obstacle to computer image segmentation. It leads to a bias toward larger objects at the expense of smaller ones, resulting in higher rates of false positives and increased computational demands [41]. Currently, there are four main solutions for addressing the class imbalance problem: adjusting sampling methods, developing a new type of loss function, utilizing attention mechanisms, and implementing cascade models. Sampling method adjustments typically involve undersampling and oversampling techniques [69]. Undersampling adjusts the imbalance of categories by reducing the majority class samples, but may lead to information loss. On the other hand, oversampling methods can be used to expand imbalanced data, such as random oversampling, SMOTE oversampling [70], adaptive integrated oversampling [71], and random undersampling [72] which are popular sampling methods. The design of appropriate loss function is also one of the effective strategies to mitigate the impact of class imbalance, the advantage is that it will not destroy the original data distribution, the loss function mainly includes the loss function based on Dice, the loss function based on cross-entropy, or a combination of both. For example, researchers optimise the patch size of the segmentation architecture nnU-Net and use the class-adaptive Dice loss function to reduce the possibility of false positives brought by the image class imbalance problem [73], and Yeung et al. [74] combined dice- and cross-entropy-based loss to deal with the class imbalance, which reduces the loss to the class imbalance while converting voxel measurements to semantically labelled overlap measurements sensitivity to imbalance effects. The attention mechanism can selectively assign different weights to the input variables according to the importance differences, which can highlight useful information in image features while suppressing irrelevant information without the need for a large number of parameters and computational overheads. Ke Sheng et al. [39] designed a network architecture based on a spatial attention learning mechanism and a channel attention learning mechanism, which is able to priorities the invocation of neurons in regions potentially related to OARs, thus identifying meaningful features, which reduces the requirement of computer arithmetic and decreases the segmentation time. The fourth category of methods is cascade models, using cascade models can effectively take advantage of multiple models for image segmentation, e.g., James C. Korte et al. [29] used cascade CNNs to segment organs, such as submandibular gland and parotid gland of head and neck tumor patients, and still performed the image segmentation task using the original image resolution on a low dimensional image. In conclusion, the class imbalance problem is a key issue in DL-based segmentation of head and neck OARs, and future work on optimization at the level of data sources, algorithms, and hybrid models will help to improve global accuracy and reduce misclassification.

In conclusion, with the rapid development of computer vision and image processing technology, DL has immense potential for application in various fields, including healthcare, as well as image recognition and classification. This research paper assesses the performance of DL contouring OARs in head and neck region. Its excellent performance confirms the value of DL for clinical applications. However, there are also some urgent problems that need to be solved. For the future development of DL, it is necessary to strengthen theoretical research and innovation of algorithms while simultaneously building large medical image data sets. In addition, it is important to explore more intelligent, automated, and precise radiotherapy techniques.

This systematic review and meta-analysis analyzed the contouring performance of DL in contouring head and neck OARs for radiotherapy. There is some heterogeneity in the literature included in the study, which is an inherent limitation of single-arm meta-analysis. In addition, the low level of publication bias ensured the stability of the analysis results. The field of literature quality assessment and bias analysis of AI is highly controversial [75–79], the development of clinical prediction models necessitates comprehensive information to serve as a foundation to aid researchers in evaluating the models' performance and generalizability. The absence of adequate information to reiterate the model will heighten the risk of bias in articles, to a certain extent. In this paper, the assessment of article quality and analysis of bias did not yield very satisfactory results. This is due to the specific details reported in each research literature, resulting in a common lack of information among low quality/high bias studies. Therefore, this paper does not utilize study quality or risk of bias as a criterion for literature exclusion, but as an informative reference to aid researchers in carefully and objectively assessing high-level clinical evidence, rather than blindly utilizing it for clinical decision-making.

The limitations of this paper are as follows: 1. only the DSC metric is used to measure segmentation performance. Other parameters used in the field of computer vision to evaluate algorithm performance include mean surface distance, Hausdorff distance, Jaccard distance, and contouring time. The incorporation of additional, objective evaluation metrics would enhance the comprehensiveness of algorithm performance assessment. In addition, evaluating segmentation performance solely based on DSC does not fully indicate the effectiveness of response treatment [80], and in some studies it has been found that even large differences between DL contour and true contour do not necessarily affect the dosimetry or clinical feasibility of OARs [81]. Dose accuracy [22], normal tissue complication probability values [82] and the applicability of the target area to the clinic [36] are all subject to critical review by clinicians. 2. Diverse image sources: algorithm performance is closely tied to factors, such as imaging modality, device parameters, and characteristics of the patient population, each of which may directly affect the performance of the DL algorithm. Non-homogeneous parameter metrics may present a potential risk of bias.3. Assessing the interobserver contour variability of head and neck OARs and the impact of variability on the performance in DL algorithms has an important value in furthering the understanding and application of DL contours [44, 83], which will be one of the key elements of future research.

Conclusion

The potential of DL is enormous, and it should be optimized and innovated in the future to coordinate with multiple institutions to create large-scale, multi-modality, high-quality medical data sets that integrate multiple information. DL is expected to become a powerful tool to promote the implementation of "precision radiotherapy" and provide individualized, standardized and refined treatment plans for patients.

Abbreviations

OARs	Organs at risk
DL	Deep learning
DSC	Dice similarity coefficient
CLAIM	Checklist for Artificial Intelligence in Medical Imaging
PROBAST	Prediction Model Risk of Bias Assessment Tool
AI	Artificial intelligence
CNN	Convolutional neural network

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12938-023-01159-y.

Additional file 1: Table S1. Search strategie**s. Table S2.** Checklist for Artificial Intelligence in Medical Imaging (CLAIM). **Table S3.** PROBAST (Prediction model Risk of Bias Assessment Tool) Review Items. Table **S4.** Result of CLAIM. **Table S5.** Result of PROBAST. **Figure S1 (A–L)** Forest plot of the pooled DSC of 12 OARs. **Figure S2 (A–L)** Funnel plots for meta-analysis of 12 OARs. **Figure S3 (A–H)** Forest plot of the DSC of segmentation of 4 OARs in CT or MRI images. **Figure S4 (A–H)** Forest plot of the DSC of segmentation of 4 OARs in 2D or 3D images.

Acknowledgements

Not applicable.

Author contributions

The review is designed and revised by LP, SY and YY. LP and SY conducted the literature collection. LP draw the figures. SY, ZZ, YY provided guidance and revised this manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study is supported by the grants from Liaoning Provincial Education Department Surface Project (No. LJKZ0127) and National Natural Science Foundation of China (No.62101357).

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

The review did not require approval by an Ethical Committee.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 14 July 2023 Accepted: 21 September 2023
Published online: 01 November 2023

References

1. Cohen N, Fedewa S, Chen AY. Epidemiology and demographics of the head and neck cancer population. Oral Maxillofac Surg Clin North Am. 2018;30(4):381–95.
2. Mazul AL, Chidambaram S, Zevallos JP, Massa ST. Disparities in head and neck cancer incidence and trends by race/ethnicity and sex. Head Neck. 2023;45(1):75–84. https://doi.org/10.1002/hed.27209. (**Epub 20221006**).
3. Daraei P, Moore CE. Racial disparity among the head and neck cancer population. J Cancer Educ. 2015;30(3):546–51. https://doi.org/10.1007/s13187-014-0753-4.
4. Cole L, Polfus L, Peters ES. Examining the incidence of human papillomavirus-associated head and neck cancers by race and ethnicity in the U.S,, 1995–2005. PLoS ONE. 2012;7(3):32657. https://doi.org/10.1371/journal.pone.0032657. (**Epub 20120320**).
5. Larsson SC, Burgess S. Appraising the causal role of smoking in multiple diseases: a systematic review and meta-analysis of mendelian randomization studies. EBioMedicine. 2022;82: 104154. https://doi.org/10.1016/j.ebiom.2022.104154. (**Epub 20220708**).
6. Di Credico G, Polesel J, Dal Maso L, Pauli F, Torelli N, Luce D, et al. Alcohol drinking and head and neck cancer risk: the joint effect of intensity and duration. Br J Cancer. 2020;123(9):1456–63. https://doi.org/10.1038/s41416-020-01031-z. (**Epub 20200824**).

7. Applebaum KM, Furniss CS, Zeka A, Posner MR, Smith JF, Bryan J, et al. Lack of association of alcohol and tobacco with Hpv16-associated head and neck cancer. J Natl Cancer Inst. 2007;99(23):1801–10. https://doi.org/10.1093/jnci/djm233. (**Epub 20071127**).
8. Merlano M. alternating chemotherapy and radiotherapy in locally advanced head and neck cancer: an alternative? Oncologist. 2006;11(2):146–51. https://doi.org/10.1634/theoncologist.11-2-146.
9. Gujral DM, Nutting CM. Patterns of failure, treatment outcomes and late toxicities of head and neck cancer in the current era of imrt. Oral Oncol. 2018;86:225–33. https://doi.org/10.1016/j.oraloncology.2018.09.011. (**Epub 20181004**).
10. Baudelet M, Van den Steen L, Tomassen P, Bonte K, Deron P, Huvenne W, et al. Very late xerostomia, dysphagia, and neck fibrosis after head and neck radiotherapy. Head Neck. 2019;41(10):3594–603. https://doi.org/10.1002/hed.25880. (**Epub 20190722**).
11. Crowder SL, Douglas KG, Yanina Pepino M, Sarma KP, Arthur AE. Nutrition impact symptoms and associated outcomes in post-chemoradiotherapy head and neck cancer survivors: a systematic review. J Cancer Surviv. 2018;12(4):479–94. https://doi.org/10.1007/s11764-018-0687-7. (**Epub 20180320**).
12. Jham BC, da Silva Freire AR. Oral Complications of Radiotherapy in the Head and Neck. Braz J Otorhinolaryngol. 2006;72(5):704–8. https://doi.org/10.1016/s1808-8694.
13. 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(1):120. https://doi.org/10.1186/s13014-020-01677-2.
14. Geets X, Daisne JF, Arcangeli S, Coche E, De Poel M, Duprez T, Nardella G, Grégoire V. 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(1):25–31. https://doi.org/10.1016/j.radonc.2005.04.010.
15. Peng YL, Chen L, Shen GZ, Li YN, Yao JJ, Xiao WW, Yang L, Zhou S, Li JX, Cheng WQ, et al. Interobserver variations in the delineation of target volumes and organs at risk and their impact on dose distribution in intensity-modulated radiation therapy for nasopharyngeal carcinoma. Oral Oncol. 2018;82:1–7.
16. Zukauskaite R, Rumley CN, Hansen CR, Jameson MG, Trada Y, Johansen J, Gyldenkerne N, Eriksen JG, Aly F, Christensen RL, et al. Delineation uncertainties of tumour volumes on MRI of head and neck cancer patients. Clin Transl Radiat Oncol. 2022;36:121–6.
17. Oktay O, Nanavati J, Schwaighofer A, Carter D, Bristow M, Tanno R, Jena R, Barnett G, Noble D, Rimmer Y, et al. Evaluation of deep learning to augment image-guided radiotherapy for head and neck and prostate cancers. JAMA Netw Open. 2020;3(11): e2027426.
18. 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(1):6137. https://doi.org/10.1038/s41467-022-33178-z. (**Epub 20221017**).
19. Vorwerk H, Zink K, Schiller R, Budach V, Böhmer D, Kampfer S, Popp W, Sack H, Engenhart-Cabillic R. Protection of quality and innovation in radiation oncology: the prospective multicenter trial the German Society of Radiation Oncology (DEGRO-QUIRO study). Evaluation of time, attendance of medical staff, and resources during radiotherapy with IMRT. Strahlenther Onkol. 2014;190(5):433–43.
20. La Macchia M, Fellin F, Amichetti M, Cianchetti M, Gianolini S, Paola V, Lomax AJ, Widesott L. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. Radiat Oncol. 2012;7:160.
21. Douglass M, Gorayski P, Patel S, Santos A. Synthetic cranial MRI from 3D optical surface scans using deep learning for radiation therapy treatment planning. Phys Eng Sci Med. 2023;46(1):367–75.
22. Chen X, Sun S, Bai N, Han K, Liu Q, Yao S, Tang H, Zhang C, Lu Z, Huang Q, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiother Oncol. 2021;160:175–84.
23. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339: b2700.
24. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers. Radiol Artif Intell. 2020;2(2): 200029. https://doi.org/10.1148/ryai.2020200029. (**Epub 20200325**).
25. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. Probast: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8. https://doi.org/10.7326/m18-1376.
26. Frizzell TO, Glashutter M, Liu CC, Zeng A, Pan D, Hajra SG, et al. Artificial intelligence in brain mri analysis of alzheimer's disease over the past 12 years: a systematic review. Ageing Res Rev. 2022;77: 101614. https://doi.org/10.1016/j.arr.2022.101614. (**Epub 20220328**).
27. Dai X, Lei Y, Wang T, Zhou J, Roper J, McDonald M, et al. Automated delineation of head and neck organs at risk using synthetic mri-aided mask scoring regional convolutional neural network. Med Phys. 2021;48(10):5862–73. https://doi.org/10.1002/mp.15146. (**Epub 20210818**).
28. Zhong T, Huang X, Tang F, Liang S, Deng X, Zhang Y. Boosting-based cascaded convolutional neural networks for the segmentation of ct organs-at-risk in nasopharyngeal carcinoma. Med Phys. 2019. https://doi.org/10.1002/mp.13825. (**Epub 20190916**).
29. Korte JC, Hardcastle N, Ng SP, 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(12):7757–72. https://doi.org/10.1002/mp.15290. (**Epub 20211101**).
30. Chan JW, Kearney V, Haaf S, Wu S, Bogdanov M, Reddick M, et al. A convolutional neural network algorithm for automatic segmentation of head and neck organs at risk using deep lifelong learning. Med Phys. 2019;46(5):2204–13. https://doi.org/10.1002/mp.13495. (**Epub 20190404**).

31. Liang S, Tang F, Huang X, Yang K, Zhong T, Hu R, et al. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. Eur Radiol. 2019;29(4):1961–7. https://doi.org/10.1007/s00330-018-5748-9. (**Epub 20181009**).
32. Kim N, Chun J, Chang JS, Lee CG, Keum KC, Kim JS. Feasibility of continual deep learning-based segmentation for personalized adaptive radiation therapy in head and neck area. Cancers. 2021. https://doi.org/10.3390/cancers13040702. (**Epub 20210209**).
33. Gao Y, Huang R, Yang Y, Zhang J, Shao K, Tao C, et al. Focusnetv 2: 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. (**Epub 20201010**).
34. Tong N, Gou S, Yang S, Ruan D, Sheng K. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. Med Phys. 2018;45(10):4558–67. https://doi.org/10.1002/mp.13147. (**Epub 20180919**).
35. Fang Y, Wang J, Ou X, Ying H, Hu C, Zhang Z, et al. the impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients. Phys Med Biol. 2021. https://doi.org/10.1088/1361-6560/ac2206. (**Epub 20210914**).
36. Van Dijk LV, Van den Bosch L, Aljabar P, Peressutti D, Both S, Steenbakkers RJ, et al. Improving automatic delineation for head and neck organs at risk by deep learning contouring. Radiother Oncol. 2020;142:115–23. https://doi.org/10.1016/j.radonc.2019.09.022. (**Epub 20191022**).
37. Dai X, Lei Y, Wang T, Zhou J, Rudra S, McDonald M, et al. Multi-organ auto-delineation in head-and-neck mri for radiation therapy using regional convolutional neural network. Phys Med Biol. 2022. https://doi.org/10.1088/1361-6560/ac3b34. (**Epub 20220121**).
38. Tappeiner E, Pröll S, Hönig M, Raudaschl PF, Zaffino P, Spadea MF, et al. Multi-organ segmentation of the head and neck area: an efficient hierarchical neural networks approach. Int J Comput Assist Radiol Surg. 2019;14(5):745–54. https://doi.org/10.1007/s11548-019-01922-4. (**Epub 20190307**).
39. Gou S, Tong N, Qi S, Yang S, Chin R, Sheng K. Self-channel-and-spatial-attention neural network for automated multi-organ segmentation on head and neck Ct images. Phys Med Biol. 2020;65(24):245034. https://doi.org/10.1088/1361-6560/ab79c3. (**Epub 20201211**).
40. Tong N, Gou S, Yang S, Cao M, Sheng K. Shape constrained fully convolutional densenet with adversarial training for multiorgan segmentation on head and neck ct and low-field mr images. Med Phys. 2019;46(6):2669–82. https://doi.org/10.1002/mp.13553. (**Epub 20190506**).
41. Zhang S, Wang H, Tian S, Zhang X, Li J, Lei R, et al. A slice classification model-facilitated 3d encoder-decoder network for segmenting organs at risk in head and neck cancer. J Radiat Res. 2021;62(1):94–103. https://doi.org/10.1093/jrr/rraa094.
42. Zhang Z, Zhao T, Gay H, Zhang W, Sun B. Weaving attention U-Net: a novel hybrid cnn and attention-based method for organs-at-risk segmentation in head and neck ct images. Med Phys. 2021;48(11):7052–62. https://doi.org/10.1002/mp.15287.
43. Liang S, Thung KH, Nie D, Zhang Y, Shen D. Multi-view spatial aggregation framework for joint localization and segmentation of organs at risk in head and neck Ct images. IEEE Trans Med Imaging. 2020;39(9):2794–805. https://doi.org/10.1109/tmi.2020.2975853.
44. Koo J, Caudell JJ, Latifi K, Jordan P, Shen S, Adamson PM, et al. Comparative evaluation of a prototype deep learning algorithm for autosegmentation of normal tissues in head and neck radiotherapy. Radiother Oncol. 2022;174:52–8. https://doi.org/10.1016/j.radonc.2022.06.024.
45. DeAngelis LM, Delattre JY, Posner JB. Radiation-induced dementia in patients cured of brain metastases. Neurology. 1989;39(6):789–96. https://doi.org/10.1212/wnl.39.6.789.
46. Wolfson AH, Bae K, Komaki R, et al. Primary analysis of a phase II randomized trial Radiation Therapy Oncology Group (RTOG) 0212: impact of different total doses and schedules of prophylactic cranial irradiation on chronic neurotoxicity and quality of life for patients with limited-disease small-cell lung cancer. Int J Radiat Oncol Biol Phys. 2011;81(1):77–84. https://doi.org/10.1016/j.ijrobp.2010.05.013.
47. Goyal H, Singh N, Gurjar OP, Tanwar RK. Radiation induced demyelination in cervical spinal cord of the head and neck cancer patients after receiving radiotherapy. J Biomed Phys Eng. 2020;10(1):1–6. https://doi.org/10.31661/jbpe.v0i0.1055.
48. Nelms BE, Tomé 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(1):368–78. https://doi.org/10.1016/j.ijrobp.2010.10.019.
49. Caudell JJ, Torres-Roca JF, Gillies RJ, et al. The future of personalised radiotherapy for head and neck cancer[J]. Lancet Oncol. 2017;18(5):266–73.
50. Grkgoire V, Jeraj R, Lee JA, et al. Radiotherapy for head and neck tumours in 2012 and beyond:Conformal, tailored, and adaptive[J]. Lancet Oncol. 2012;13(7):292–300.
51. Castelli J, Simon A, Lafond C, et al. Adaptive radiotherapyfor head andneck cancer[J]. Acta Oncologica Taylor Francis. 2018;57(10):1284–92.
52. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys. 2017;44:547–57. https://doi.org/10.1002/mp.12045.
53. Kieselmann JP, Fuller CD, Gurney-Champion OJ, Oelfke U. Cross-modality deep learning: contouring of MRI data from annotated CT data only. Med Phys. 2021;48:1673–84. https://doi.org/10.1002/mp.14619.
54. Liu Y, Lei Y, Fu Y, Wang T, Zhou J, Jiang X, et al. Head and neck multi-organ auto-segmentation on CT images aided by synthetic MRI. Med Phys. 2020;47:4294–302. https://doi.org/10.1002/mp.14378.

55. Yee E, Ma D, Popuri K, Chen S, Lee H, Chow V, et al. 3D hemisphere-based convolutional neural network for whole-brain MRI segmentation. Comput Med Imaging Graph. 2022;95: 102000. https://doi.org/10.1016/j.compmedimag.2021.102000.
56. Huo Y, Xu Z, Xiong Y, Aboud K, Parvathaneni P, Bao S, et al. 3D whole brain segmentation using spatially localized atlas network tiles. Neuroimage. 2019;194:105–19. https://doi.org/10.1016/j.neuroimage.2019.03.041.
57. Yu J, Yang B, Wang J, Leader J, Wilson D, Pu J. 2D CNN versus 3D CNN for false-positive reduction in lung cancer screening. J Med Imag. 2020;7: 051202. https://doi.org/10.1117/1.Jmi.7.5.051202.
58. Gaikar R, Zabihollahy F, Elfaal MW, Azad A, Schieda N, Ukwatta E. Transfer learning-based approach for automated kidney segmentation on multiparametric MRI sequences. J Med Imaging. 2022;9: 036001. https://doi.org/10.1117/1.Jmi.9.3.036001.
59. Vu MH, Grimbergen G, Nyholm T, Löfstedt T. Evaluation of multislice inputs to convolutional neural networks for medical image segmentation. Med Phys. 2020;47:6216–31. https://doi.org/10.1002/mp.14391.
60. Zhang R, Zhuo L, Chen M, Yin H, Li X, Wang Z. Hybrid Deep Feature Fusion of 2D CNN and 3D CNN for Vestibule Segmentation from CT Images. Comput Math Methods Med. 2022;2022:6557593. https://doi.org/10.1155/2022/6557593.
61. Valdez-Rodríguez JE, Calvo H, Felipe-Riverón E, Moreno-Armendáriz MA. Improving depth estimation by embedding semantic segmentation: a hybrid CNN model. Sensors. 2022. https://doi.org/10.3390/s22041669.
62. Gu L, Cai XC. Fusing 2D and 3D convolutional neural networks for the segmentation of aorta and coronary arteries from CT images. Artif Intell Med. 2021;121: 102189. https://doi.org/10.1016/j.artmed.2021.102189.
63. Lee J, Nishikawa RM. Cross-organ, cross-modality transfer learning: feasibility study for segmentation and classification. IEEE Access. 2020;8:210194–205. https://doi.org/10.1109/access.2020.3038909.
64. Mzoughi H, Njeh I, Wali A, Slima MB, BenHamida A, Mhiri C, et al. Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification. J Digit Imaging. 2020;33:903–15. https://doi.org/10.1007/s10278-020-00347-9.
65. Yu S, Chen M, Zhang E, Wu J, Yu H, Yang Z, et al. Robustness study of noisy annotation in deep learning based medical image segmentation. Phys Med Biol. 2020;65: 175007. https://doi.org/10.1088/1361-6560/ab99e5.
66. Henderson EGA, Vasquez Osorio EM, van Herk M, Green AF. Optimising a 3D convolutional neural network for head and neck computed tomography segmentation with limited training data. Phys Imag Radiat Oncol. 2022;22:44–50. https://doi.org/10.1016/j.phro.2022.04.003.
67. Zhao Y, Rhee DJ, Cardenas C, Court LE, Yang J. Training deep-learning segmentation models from severely limited data. Med Phys. 2021;48:1697–706. https://doi.org/10.1002/mp.14728.
68. Amjad A, Xu J, Thill D, Lawton C, Hall W, Awan MJ, et al. General and custom deep learning autosegmentation models for organs in head and neck, abdomen, and male pelvis. Med Phys. 2022;49:1686–700. https://doi.org/10.1002/mp.15507.
69. Gnip P, Vokorokos L, Drotár P. Selective oversampling approach for strongly imbalanced data. PeerJ Comput Sci. 2021;7: e604.
70. Welvaars K, Oosterhoff JHF, van den Bekerom MPJ, Doornberg JN, van Haarst EP. Implications of resampling data to address the class imbalance problem (IRCIP): an evaluation of impact on performance between classification algorithms in medical data. JAMIA Open. 2023;6(2):ooad033.
71. Priyadharshini M, Banu AF, Sharma B, Chowdhury S, Rabie K, Shongwe T. Hybrid multi-label classification model for medical applications based on adaptive synthetic data and ensemble learning. Sensors. 2023;23(15):6836.
72. Kishore A, Venkataramana L, Prasad DVV, Mohan A, Jha B. Enhancing the prediction of IDC breast cancer staging from gene expression profiles using hybrid feature selection methods and deep learning architecture. Med Biol Eng Comput. 2023. https://doi.org/10.1007/s11517-023-02892-1.
73. Tappeiner E, Welk M, Schubert R. Tackling the class imbalance problem of deep learning-based head and neck organ segmentation. Int J Comput Assist Radiol Surg. 2022;17:2103–11. https://doi.org/10.1007/s11548-022-02649-5.
74. Yeung M, Sala E, Schönlieb CB, Rundo L. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput Med Imaging Graph. 2022;95: 102026.
75. Gou S, Tong N, Qi S, Yang S, Chin R, Sheng K. Self-channel-and-spatial-attention neural network for automated multi- organ segmentation on head and neck CT images. Phys Med Biol. 2020;65: 245034. https://doi.org/10.1088/1361-6560/ab79c3.
76. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1:e271–97. https://doi.org/10.1016/s2589-7500(19)30123-2.
77. Venema E, Wessler BS, Paulus JK, Salah R, Raman G, Leung LY, et al. Large-scale validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. J Clin Epidemiol. 2021;138:32–9. https://doi.org/10.1016/j.jclinepi.2021.06.017.
78. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350: g7594. https://doi.org/10.1136/bmj.g7594.
79. de Jong Y, Ramspek CL, Zoccali C, Jager KJ, Dekker FW, van Diepen M. Appraising prediction research: a guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). Nephrology. 2021;26:939–47. https://doi.org/10.1111/nep.13913.
80. Belue MJ, Harmon SA, Lay NS, Daryanani A, Phelps TE, Choyke PL, et al. The low rate of adherence to checklist for artificial intelligence in medical imaging criteria among published prostate MRI artificial intelligence algorithms. J Am Coll Radiol. 2022. https://doi.org/10.1016/j.jacr.2022.05.022.

81. Kieselmann JP, Kamerling CP, Burgos N, Menten MJ, Fuller CD, Nill S, et al. Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region. Phys Med Biol. 2018;63: 145007. https://doi.org/10.1088/1361-6560/aacb65.
82. Delaney AR, Dahele M, Slotman BJ, Verbakel W. Is accurate contouring of salivary and swallowing structures necessary to spare them in head and neck VMAT plans? Radiother Oncol. 2018;127:190–6. https://doi.org/10.1016/j.radonc.2018.03.012.
83. 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. https://doi.org/10.1016/j.radonc.2021.09.019.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.