An investigation into the risk of population bias in deep learning autocontouring

McQuinlan, Yasmin; Brouwer, Charlotte L.; Lin, Zhixiong; Gan, Yong; Kim, Jin Sung; van Elmpt, Wouter; Gooding, Mark J.

Published in:
Radiotherapy and Oncology

DOI:
10.1016/j.radonc.2023.109747

Publication date:
2023

Radiotherapy and Oncology 186 (2023) 109747

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

An investigation into the risk of population bias in deep learning autocontouring

Yasmin McQuinlan [a], Charlotte L. Brouwer [b,*], Zhixiong Lin [c], Yong Gan [c], Jin Sung Kim [d], Wouter van Elmpt [e], Mark J. Gooding [f,g]

[a] Mirada Medical Ltd, Oxford, United Kingdom; [b] University of Groningen, University Medical Center Groningen, Department of Radiation Oncology, Groningen, The Netherlands; [c] Shantou University Medical Centre, Guangdong, China; [d] Yonsei University Health System, Seoul, Republic of Korea; [e] Department of Radiation Oncology (MAASTRO), GROW – School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands; [f] Mirada Medical Ltd; and [g] Inpictura Ltd, Oxford, United Kingdom

ARTICLE INFO

Article history:
Received 19 January 2023
Received in revised form 30 May 2023
Accepted 8 June 2023
Available online 16 June 2023

Keywords:
Radiotherapy
Autocontouring
Segmentation
Organ-at-Risk
Deep learning
Bias

ABSTRACT

Background and Purpose: To date, data used in the development of Deep Learning-based automatic contouring (DLC) algorithms have been largely sourced from single geographic populations. This study aimed to evaluate the risk of population-based bias by determining whether the performance of an autocontouring system is impacted by geographic population.

Materials and methods: 80 Head Neck CT deidentified scans were collected from four clinics in Europe (n = 2) and Asia (n = 2). A single observer manually delineated 16 organs-at-risk in each. Subsequently, the data was contoured using a DLC solution, and trained using single institution (European) data. Autocontours were compared to manual delineations using quantitative measures. A Kruskal-Wallis test was used to test for any difference between populations. Clinical acceptability of automatic and manual contours to observers from each participating institution was assessed using a blinded subjective evaluation.

Results: Seven organs showed a significant difference in volume between groups. Four organs showed statistical differences in quantitative similarity measures. The qualitative test showed greater variation in acceptance of contouring between observers than between data from different origins, with greater acceptance by the South Korean observers.

Conclusion: Much of the statistical difference in quantitative performance could be explained by the difference in organ volume impacting the contour similarity measures and the small sample size. However, the qualitative assessment suggests that observer perception bias has a greater impact on the apparent clinical acceptability than quantitatively observed differences. This investigation of potential geographic bias should extend to more patients, populations, and anatomical regions in the future.

Contouring is an integral part of the patient pathway for radiotherapy planning [2,12]. However, this task is affected by interobserver variation, even amongst the most experienced clinicians [3], which can lead to adverse patient outcomes [4]. Automation of contouring using deep learning contouring (DLC), can improve consistency and save time for clinicians [5–9]. DLC models are trained using large datasets, to learn the variations in the appearance of anatomy within the training population. Such datasets may be derived from hospitals or open-source collections [6,9–24]. Although artificial intelligence autocontouring models have been evaluated in a range of geographic regions [22,5–7,9–15,17–20,24–28], the clinical impact of applying these models across differing geographic populations is unclear.

Artificial intelligence (AI) models can be subject to several known biases including sample bias, temporal bias, and population bias [1,29]. Awareness of potential biases should be considered in the use of clinical deep learning solutions. To encourage the development of good AI-based clinical solutions, regulators from the United Kingdom, Canada and the United States have identified principles that could improve minimisation of biases. This paper explores the principle of the use of data sets that are representative of the intended patient population [30].

While the impact of patient diversity in training data has not been investigated in the context of DLC it has been investigated and found in cardiology [31–32]. However, radiotherapy patients already represent a cohort that stands apart from the general pop-

* Corresponding author at: Department of Radiation Oncology, University Medical Center Groningen, PO Box 30001, 9700 RB, Groningen, The Netherlands.
E-mail addresses: yasmin.mcquinlan@mirada-medical.com (Y. McQuinlan), c.l.brouwer@umcg.nl (C.L. Brouwer), zxlin5@qq.com (Z. Lin), y.gan@umcg.nl (Y. Gan), jinsung@yuhs.ac (J. Sung Kim), wouter.vanelmpt@maastro.nl (W. van Elmpt), mark.gooding@inpicturamedica.com (M.J. Gooding).

https://doi.org/10.1016/j.radonc.2023.109747
0167-8140/

Table 1
Publications showing deep learning autocontouring models and their derived data location.

Author	Total Datasets	Anatomy	Reported Geographic Data Source	Data reported disaggregated for race
Van Dijk et al [11]	589	Head Neck	Netherlands	N
Nikolov et al [12]	838	Head Neck	United Kingdom	N
Liu et al [13]	237	Pelvis	United Kingdom	N
Oktay et al [14]	761	Pelvis (519)Head Neck (242)	United Kingdom	N
Almberg et al [9]	200	Breast	Norway	N
Blanchard et al [15]	100	Head Neck	France	N
Iyer et al [16]	242	Head Neck	United States	N
Song et al [17]	199	Pelvis	China	N
Duan et al [18]	84	Pelvis	United States	N
Ma et al [19]	535	Pelvis	China	N
Byun et al [20]	111	Breast	South Korea	N
Kim et al [21]	100	Head Neck	South Korea	N
Fernandes et al [22]	127	Thorax	Netherlands	N
Kiljunen et al [6]	<900	Pelvis	Finland	N
Cardenas et al [23]	71	Head Neck	United States	N
Weston et al [24]	84	Abdomen	United States	N

Table 2
Model of CT Scanner used for data acquisition.

Clinic	CT Scanner Model(s)
University Medical Center Groningen, *Department of Radiation Oncology, Groningen, The Netherlands*	Siemens Somatom Definition AS
Yonsei Cancer Centre, *Yonsei University Health System, Seoul, South Korea*	Siemens Sensation OpenToshiba Acquilion
Maastro, *Department of Radiation Oncology, Maastricht, The Netherlands*	Phillips Gemini TF 64Siemens Biograph 40Siemens Sensation 10Siemens Sensation Open
Shantou University Medical Centre, *Guangdong, China*	Phillips Brilliance Big Bore

ulation. The images are acquired in a controlled manner, that distinguishes them from diagnostic imaging, for example the use of a flatbed. Typically, DLC models use data sourced from and reflecting a single geographic population (Table 1, Supplementary Data). Differing contouring protocols present one challenge to deploying models between different institutions and geographic regions, however the use of clinical contouring guidelines seeks to mitigate this risk [33]. There is also some evidence that differences in imaging protocols will impact performance [27]. Nevertheless, it is unclear whether a DLC model trained in one geographic population would perform as well in a difference geographic population with differing racial demographics, and whether any difference in performance represent a risk to equality of treatment of patients.

The goal of this study is to evaluate this risk by determining the impact differing geographic populations (as a proxy for racial demographics) have on performance of a deep learning autocontouring system trained using data from a single geographic region.

Materials and methods

Eighty primary head and neck cancer datasets were collected, 20 from each of four clinics. The participating clinics were from Europe (University Medical Center Groningen, Groningen, The Netherlands; Maastro, Maastricht, The Netherlands) and Asia

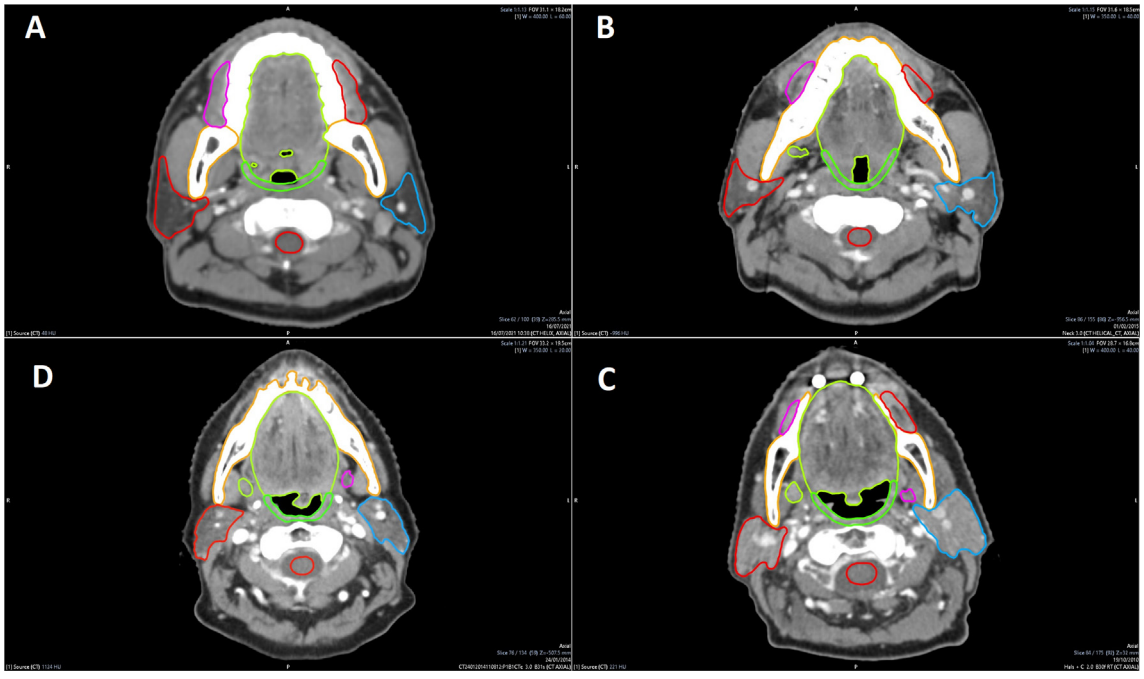

Fig. 1. An example of the single observer contours on randomly selected cases from each institution. Transverse slices taken at the caudal edge of the C1 spinous process are shown. A: Shantou University Medical College, China; B: Yonsei University Medical Centre, South Korea; C: Maastro, Netherlands; D: University Medical Center Groningen, Netherlands.

(Shantou University Medical Centre, Guangdong, China; Yonsei University Health System, Seoul, Republic of Korea). Patients were immobilized supine in thermoplastic shell mask, and scanned from cranium, down to carina. The data were acquired using a range of CT scanners, detailed in Table 2 of the Supplementary Material. Patients who had an organ-at-risk removed, and or had disease burden to significantly disturb neighboring anatomy, were excluded to prevent more extreme outliers confounding results. Race is not recorded in the oncology electronic health record at any of the participating clinics. Therefore, it was assumed all patients submitted to the study were reasonably representative of racial demographics of each respective geographic location. Available demographic information for age and gender can be found in Table 8 (Supplementary Material). Datasets were anonymised using tools available on the local radiotherapy treatment planning system before analysis.

Table 3
Median volume(cm^3) of single observer reference contours. Bold italics indicate significance after Bonferroni correction for the multiple hypothesis testing across all organs with $p < 0.003125$.

Structure	Reference Contour Volume (cm^3): Median and IQR				Kruskall-Wallis Test (p-value)
	Groningen	Maastro	Shantou	Yonsei	
Brainstem	26.629 (28.469–24.038)	25.629 (28.584–24.000)	25.047 (26.976–23.698)	28.722 (29.895–26.312)	*0.0237*
Buccal Mucosa Left	8.545 (10.267–7.477)	9.312 (10.868–8.124)	5.362 (6.394–4.470)	6.243 (7.418–4.030)	***<0.001***
Buccal Mucosa Right	9.618 (10.728–8.811)	9.066 (10.626–7.569)	5.191 (6.373–4.743)	6.408 (7.911–3.876)	***<0.001***
Cricopharyngeal Inlet	5.450 (6.410–4.638)	5.001 (5.474–4.370)	3.893 (4.299–3.467)	3.579 (4.385–3.161)	***<0.001***
Glottic Area	3.0764 (4.1609–2.5219)	2.615 (3.648–1.737)	2.142 (2.807–1.614)	2.784 (3.455–2.360)	0.104
Mandible	66.585 (75.480–60.028)	61.574 (66.570–54.452)	88.420 (97.066–84.396)	90.890 (94.912–76.111)	***<0.001***
Oesophagus Cervical	2.638 (3.298–1.765)	2.732 (3.123–2.292)	1.9759 (2.4196–1.2697)	2.701 (3.318–1.718)	0.0714
Oral Cavity (Extended)	119.433 (132.743–111.174)	124.060 (135.737–108.696)	97.749 (110.795–88.645)	115.686 (130.416–109.002)	***<0.001***
Parotid Left	31.118 (38.055–26.199)	26.715 (31.509–23.091)	23.848 (29.468–21.322)	29.910 (36.219–22.778)	0.0528
Parotid Right	31.785 (37.667–28.842)	25.327 (32.875–21.799)	23.692 (27.879–21.397)	28.135 (40.020–22.274)	*0.0146*
Pharyngeal Constrictor Muscles	16.426 (19.759–14.656)	18.578 (22.400–15.493)	16.012 (18.565–12.717)	15.075 (17.751–13.125)	0.0727
Spinal Cord	16.693 (19.996–14.600)	17.262 (18.239–15.352)	12.571 (14.488–11.483)	15.979 (17.855–12.289)	***<0.001***
Submandibular Left	10.682 (12.047–8.558)	10.344 (12.922–8.685)	9.211 (10.402–7.828)	9.647 (12.151–8.101)	0.342
Submandibular Right	10.720 (12.918–9.464)	10.510 (13.571–9.055)	9.300 (10.598–8.005)	10.351 (12.484–8.899)	0.245
Supraglottic Larynx	16.461 (20.533–13.237)	15.941 (19.423–11.523)	11.279 (13.072–7.972)	13.507 (14.491–12.738)	***<0.001***
Thyroid	13.539 (18.684–9.044)	13.428 (21.535–9.939)	11.900 (16.946–8.796)	16.462 (18.396–14.515)	0.277

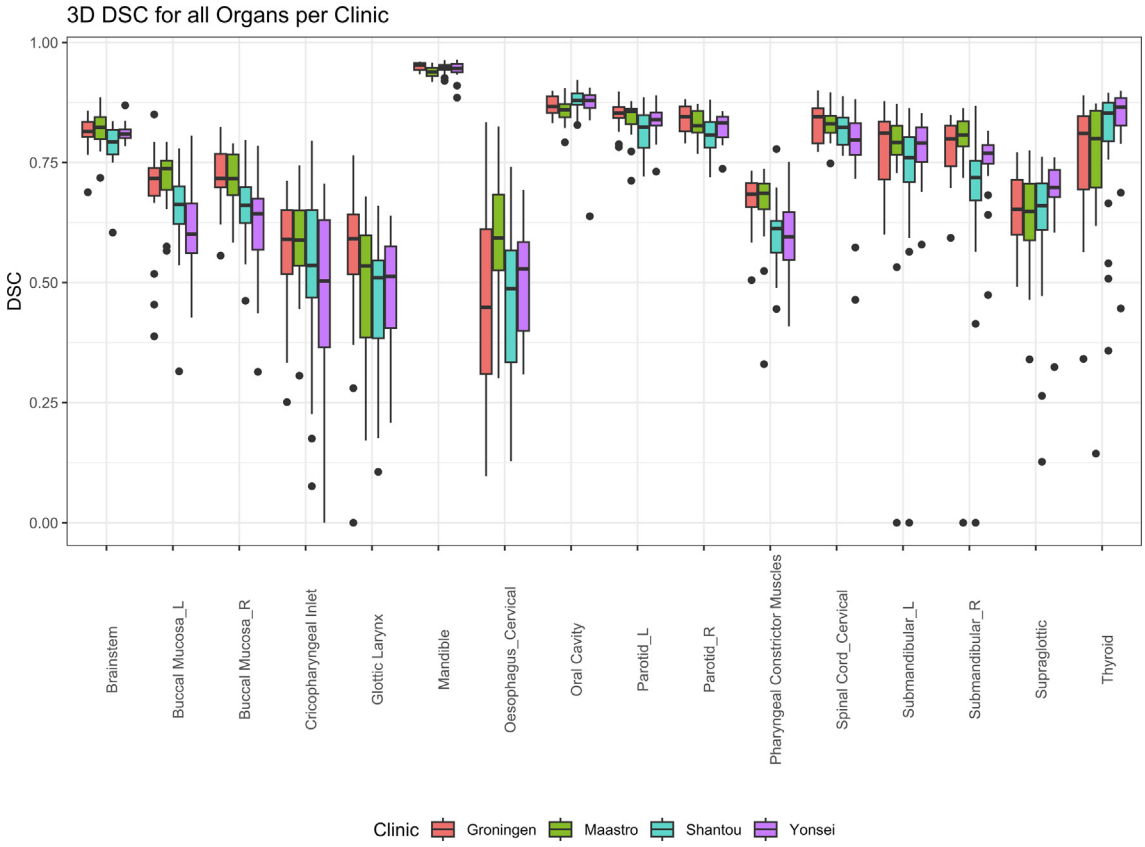

Fig. 2. DSC per organ for all patient origins. Small variations in scores can be observed for all organs. Some organs (e.g., PCM) have larger differences. Performance is not always better or worse for the same patient origin.

Table 4
Results of Kruskal-Wallis non-parametric ranking test. H(chi-squared) test statistic with 3 degrees of freedom; Italics indicate significance level of $p < 0.05$. Bold italics indicate significance after Bonferroni correction to $p < 0.00104$.

Structure	KWT DSC	KWT NAPL	KWT HD2D95%
Brainstem	H(3) = 8.78; *p = 0.0323*	H(3) = 9.44; *p = 0.024*	H(3) = 2.97; p = 0.396
Buccal Mucosa Left	H(3) = 17.1; ***p = 0.000659***	H(3) = 0.971; p = 0.808	H(3) = 4.90; p = 0.179
Buccal Mucosa Right	H(3) = 19.6; ***p = 0.000207***	H(3) = 1.04; p = 0.791	H(3) = 9.17; *p = 0.0271*
Cricopharyngeal Inlet	H(3) = 4.54; p = 0.209	H(3) = 0.771; p = 0.856	H(3) = 3.07; p = 0.381
Glottic Area	H(3) = 6.54; p = 0.0881	H(3) = 6.03; p = 0.11	H(3) = 12.5; *p = 0.00595*
Mandible	H(3) = 10.0; *p = 0.0183*	H(3) = 15.6; *p = 0.00138*	H(3) = 8.88; *p = 0.0309*
Oesophagus Cervical	H(3) = 9.14; *p = 0.0275*	H(3) = 8.16; *p = 0.0428*	H(3) = 4.65; p = 0.199
Oral Cavity (Extended)	H(3) = 8.47; *p = 0.0372*	H(3) = 11.1; *p = 0.0112*	H(3) = 3.58; p = 0.311
Parotid Left	H(3) = 7.86; *p = 0.0491*	H(3) = 9.63; *p = 0.022*	H(3) = 8.24; *p = 0.0413*
Parotid Right	H(3) = 16.7; *p = 0.00508*	H(3) = 17.8; *p = 0.00328*	H(3) = 6.3; p = 0.278
Pharyngeal Constrictor Muscles	H(3) = 19.1; ***p = 0.000256***	H(3) = 10.7; *p = 0.0132*	H(3) = 9.52; *p = 0.0231*
Spinal Cord	H(3) = 7.34; *p = 0.0618*	H(3) = 0.853; p = 0.837	H(3) = 4.97; p = 0.174
Submandibular Left	H(3) = 3.34; p = 0.342	H(3) = 9.36; *p = 0.0248*	H(3) = 2.14; p = 0.544
Submandibular Right	H(3) = 21.7; ***p = 7.7e-05***	H(3) = 20.3; ***p = 0.000149***	H(3) = 8.62; *p = 0.0347*
Supraglottic Larynx	H(3) = 4.10; *p = 0.0238*	H(3) = 2.38; p = 0.498	H(3) = 2.95; p = 0.4
Thyroid	H(3) = 9.4576; *p = 0.02379*	H(3) = 13.8; *p = 0.00313*	H(3) = 8.89; *p = 0.0307*

The study was submitted for ethical consideration despite being retrospective analysis and not including human subjects or special category data since the underlying assumption is that location relates to demographic ethnicity. The study was permitted by The Medical Ethics Review Board of the University Medical Center Groningen (METc UMCG) (METc 2022/315).

A DLC system (DLCExpert™, Mirada Medical Ltd., UK) was used to contour all cases. The head and neck model was trained from 589 datasets [11]. The training structures were delineated as per the international head and neck organ at risk delineation guidelines by the clinical institution providing the dataset [34], as previously reported [11]. This clinic also provided an independent evaluation set, acquired completely independently several years prior, thus the demographic of one evaluation group is intentionally matched to that of the training set.

A single observer recontoured all evaluation cases regardless of their geographic origin, to mitigate differences in contouring style between institutions being a confounding factor and to remove inter-observer variation in contouring. The structures contoured were Brainstem, Buccal Mucosa (BM) Left, BM Right, Cricopharyngeal Inlet, Glottic Area, Mandible, Oesophagus Cervical, Oral Cavity, Parotid Left and Parotid Right, Pharyngeal Constrictor Muscle (PCM), Spinal Cord, Submandibular Left, Submandibular Right, Supraglottic Larynx, Thyroid. All structures were delineated according to available consensus guidance [34] using Mirada RTx™ (Mirada Medical Ltd, Oxford, United Kingdom). Example contours from the single observer are shown in Fig. 1.

Quantitative evaluation

Open-source code [35] was used to compare the autocontours to the single observer reference contours, treating the single observer contours as "ground truth". Three similarity measures, 3D Dice similarity coefficient (DSC), 2D 95% Hausdorff Distance (HD2D95) and the Normalized Added Path Length (NAPL, used to give a broad understanding and to reduce the risk that any metric masks possible bias. An acceptance tolerance of 1 mm was used for the NAPL. The 2D 95th percentile Hausdorff Distance was selected as Radiotherapy Structure Sets are a 2D representation and this measure would give an indication of maximum in-plane error. Furthermore, the structure volume (cm^3) of the reference contour was compared between groups to evaluate any differences in the patient populations. A description of these measures, as defined [35], is given in the Supplementary Material.

A Kruskal-Wallis Test was used to determine if there was a statistically significant difference in quantitative measures between the four clinics. A significant level of 0.05 (prior to Bonferroni correction) was used to determine significance. Subsequently, Bonferroni correction was used to reduce the risk of Type I error. A Dunn's test was applied to organs that showed significance after correction, to identify differences between groups. Statistical tests and plots were performed, in RStudio (Posit Software PBC, Boston, MA, USA).

Qualitative assessment

Quantitative measures cannot easily indicate whether differences would impact clinical practice. Therefore, a subjective assessment of clinical acceptability was performed using a blinded web-based implementation [36] to understand better the clinical context. Each participating center had two clinicians answer a set of 200 questions each. The observers were experienced in the contouring of Head and Neck patients with an average 14 years' experience, a range of 2 years to 34 years. Observers were asked: "You have been asked to QA these contours for clinical use by a colleague. Would you…". The available choice of responses was a four-point Likert scale:

(1) "Accept them as they are; They contours are very precise."
(2) "Accept them as they are; There are minor errors, but these are clinically not significant."
(3) "Require them to be corrected; There are minor errors that need a small amount of editing."
(4) "Require them to be corrected; There are large, obvious, errors."

Each question showed a single randomised Head Neck image slice of an anonymised patient, from a collection of 142,976 images. Six organs included in the qualitative test were chosen prior to quantitative analysis and were selected to represent a wide range of functions and anatomical shapes. The image in each question was randomly selected to be uniformly distributed over the organs and then slices. Observers were shown only images that possessed a contour, as such false negatives from DLC were omitted. Randomisation of the images can result in duplicates. Observers were blinded to patient geographic origin and to the method of contour creation (manual or DLC). The single observer reference contours were used as the manual contours to remove guideline

interpretation as a confounding factor. The results were assessed when grouped by observer origin and by patient origin.

Results

The head & neck autocontouring model failed to predict some structures in the Test dataset; the Glottic Area for one patient and the Oesophagus in three patients, all from the Shantou dataset. Where a structure has failed to predict, the patient was excluded from the analysis of that structure.

In the quantitative assessment, the median volume (cm^3) and length (cm) of single observer reference contours were compared per clinic. Statistically significant volume differences ($p < 0.001$) between groups were observed for BM (L&R), Cricopharyngeal

Table 5
Reference Contour Volume (cm^3) of all organs.

Structure	Statistic	Reference Contour Volume (cm^3)			
		Groningen	Maastro	Shantou	Yonsei
Brainstem	Median	26.629	25.629	25.047	28.722
	Q1	24.038	24.000	23.698	26.312
	Q3	28.469	28.584	26.976	29.895
	IQR	4.331	4.584	3.278	3.582
Buccal Mucosa Left	Median	8.545	9.312	5.362	6.243
	Q1	7.477	8.124	4.470	4.030
	Q3	10.267	10.868	6.394	7.418
	IQR	2.790	2.744	1.924	3.388
Buccal Mucosa Right	Median	9.618	9.066	5.191	6.408
	Q1	8.811	7.569	4.743	3.876
	Q3	10.728	10.626	6.373	7.911
	IQR	1.917	3.057	1.630	4.036
Cricopharyngeal Inlet	Median	5.450	5.001	3.893	3.579
	Q1	4.638	4.370	3.467	3.161
	Q3	6.410	5.474	4.299	4.385
	IQR	1.772	11.04	8.32	1.225
Glottic Area	Median	3.0764	2.615	2.142	2.784
	Q1	2.5219	1.737	1.614	2.360
	Q3	4.1609	3.648	2.807	3.455
	IQR	1.639	1.911	1.193	1.095
Mandible	Median	66.585	61.574	88.420	90.890
	Q1	60.028	54.452	84.396	76.111
	Q3	75.480	66.570	97.066	94.912
	IQR	15.453	12.118	12.670	18.801
Oesophagus Cervical	Median	2.638	2.732	1.9759	2.701
	Q1	1.765	2.292	1.2697	1.718
	Q3	3.298	3.123	2.4196	3.318
	IQR	1.533	0.831	1.150	1.600
Oral Cavity (Extended)	Median	119.433	124.060	97.749	115.686
	Q1	111.174	108.696	88.645	109.002
	Q3	132.743	135.737	110.795	130.416
	IQR	21.569	27.041	22.150	21.414
Parotid Left	Median	31.118	26.715	23.848	29.910
	Q1	26.199	23.091	21.322	22.778
	Q3	38.055	31.509	29.468	36.219
	IQR	11.857	8.418	8.146	13.441
Parotid Right	Median	31.785	25.327	23.692	28.135
	Q1	28.842	21.799	21.397	22.274
	Q3	37.667	32.875	27.879	40.020
	IQR	8.825	11.076	6.482	17.746
Pharyngeal Constrictor Muscles	Median	16.426	18.578	16.012	15.075
	Q1	14.656	15.493	12.717	13.125
	Q3	19.759	22.400	18.565	17.751
	IQR	5.103	6.907	5.848	4.626
Spinal Cord	Median	16.693	17.262	12.571	15.979
	Q1	14.600	15.352	11.483	12.289
	Q3	19.996	18.239	14.488	17.855
	IQR	5.396	2.887	3.005	5.566
Submandibular Left	Median	10.682	10.344	9.211	9.647
	Q1	8.558	8.685	7.828	8.101
	Q3	12.047	12.922	10.402	12.151
	IQR	3.488	4.237	2.574	4.050
Submandibular Right	Median	10.720	10.510	9.300	10.351
	Q1	9.464	9.055	8.005	8.899
	Q3	12.918	13.571	10.598	12.484
	IQR	3.454	4.516	2.594	3.585
Supraglottic Larynx	Median	16.461	15.941	11.279	13.507
	Q1	13.237	11.523	7.972	12.738
	Q3	20.533	19.423	13.072	14.491
	IQR	7.296	7.900	5.100	1.753
Thyroid	Median	13.539	13.428	11.900	16.462
	Q1	9.044	9.939	8.796	14.515
	Q3	18.684	21.535	16.946	18.396
	IQR	9.640	11.596	8.150	3.881

Inlet, Mandible, Oral Cavity, Spinal Cord and Supraglottic Larynx (Table 3, Supplementary Material). All organs, except Cervical Oesophagus, showed length differences ($p < 0.001$) between groups (Table 7, Supplementary Material). Reference Contour Volume (cm^3) and 2D 95% HD to be moderately positively correlated ($r(75) = 0.613$, $p < 0.001$), for Oesophagus Cervical. Whilst Reference Contour Length (cm) and NAPL to have low negative correlation ($r(78) = -0.487$, $p < 0.001$) for Mandible.

The measures DSC (Fig. 2), NAPL and HD2D95 calculated for each organ, across all groups showed small differences between the clinics. Boxplots for NAPL (Fig. 7) and HD2D95 (Fig. 8) can be found in Supplementary Material. Visually, the PCM shows the

Table 6
Reference Contour Length (cm) for all organs.

Structure	Statistic	Reference Contour Length (cm)			
		Groningen	Maastro	Shantou	Yonsei
Brainstem	Median	230.1	153.1	152.0	160.6
	Q1	218.8	144.0	144.4	154.9
	Q3	245.7	182.7	159.2	168.8
	IQR	27.0	38.7	14.8	13.9
Buccal Mucosa Left	Median	132.99	100.90	717.8	68.57
	Q1	121.74	85.86	57.06	43.50
	Q3	153.79	115.36	83.13	87.48
	IQR	32.05	29.5	26.07	42.98
Buccal Mucosa Right	Median	143.32	96.19	68.14	73.56
	Q1	133.99	88.45	62.83	48.21
	Q3	159.59	124.47	82.61	90.43
	IQR	25.6	36.02	19.78	42.22
Cricopharyngeal Inlet	Median	113.48	69.69	56.09	54.78
	Q1	90.08	62.12	53.33	50.12
	Q3	130.06	93.19	60.65	62.42
	IQR	39.98	31.07	7.32	12.3
Glottic Area	Median	71.72	47.85	34.76	43.73
	Q1	62.51	33.76	31.17	39.18
	Q3	80.22	63.40	46.12	56.31
	IQR	17.71	29.64	14.95	17.13
Mandible	Median	902.8	554.4	625.6	645.9
	Q1	794.7	504.1	602.3	601.1
	Q3	944.3	702.6	671.0	676.6
	IQR	149.54	198.5	68.7	75.5
Oesophagus Cervical	Median	37.15	33.97	25.295	37.36
	Q1	28.04	25.32	12.970	22.53
	Q3	48.16	40.87	30.567	43.33
	IQR	20.12	15.55	17.597	20.8
Oral Cavity (Extended)	Median	641.5	487.3	354.7	375.6
	Q1	545.0	393.9	317.5	354.4
	Q3	765.6	602.0	418.5	421.8
	IQR	220.6	208.1	101.0	67.4
Parotid Left	Median	316.3	193.2	192.4	211.0
	Q1	293.3	174.6	151.3	186.8
	Q3	388.8	267.5	218.8	238.1
	IQR	95.5	92.9	67.5	51.3
Parotid Right	Median	336.8	196.4	183.3	215.1
	Q1	299.3	169.3	159.5	168.6
	Q3	393.7	251.1	196.8	240.8
	IQR	94.4	81.8	37.3	72.2
Pharyngeal Constrictor Muscles	Median	469.9	340.6	293.8	314.1
	Q1	427.6	290.6	238.7	262.7
	Q3	512.5	521.5	315.5	326.3
	IQR	84.9	230.9	76.8	63.6
Spinal Cord	Median	299.1	205.8	173.1	200.5
	Q1	270.3	188.0	167.3	173.5
	Q3	346.3	237.9	181.8	212.3
	IQR	76.0	49.9	14.5	38.8
Submandibular Left	Median	135.24	84.91	79.09	79.21
	Q1	113.25	74.39	66.99	72.46
	Q3	148.79	99.73	82.81	89.47
	IQR	35.54	25.34	15.11	17.01
Submandibular Right	Median	131.51	90.04	81.19	79.12
	Q1	115.29	77.65	72.92	73.05
	Q3	149.42	111.44	86.56	91.42
	IQR	34.13	33.79	13.64	18.37
Supraglottic Larynx	Median	257.9	207.9	134.07	158.98
	Q1	231.2	161.0	97.80	128.82
	Q3	300.2	285.9	167.39	175.80
	IQR	69.0	124.9	69.59	46.98
Thyroid	Median	236.10	203.41	161.36	189.9
	Q1	184.63	149.96	132.85	167.7
	Q3	328.87	260.67	196.91	218.4
	IQR	144.24	110.71	64.06	50.7

Table 7
Median Reference Contour Length (cm) of single observer reference contours. Bold italics indicate significance after Bonferroni correction for the multiple hypothesis testing across all organs with p < 0.003125.

Structure	Reference Contour Length (cm): Median (IQR)				Kruskall-Wallis Test (p-value)
	Groningen	Maastro	Shantou	Yonsei	
Brainstem	230.1 (245.7–218.8)	153.1 (182.7–144.0)	152.0 (159.2–144.4)	160.6 (168.8–154.9)	***<0.001***
Buccal Mucosa Left	132.99 (153.79–121.74)	100.90 (115.36–85.86)	717.8 (83.13–57.06)	68.57 (87.48–43.50)	***<0.001***
Buccal Mucosa Right	143.32 (159.59–133.99)	96.19 (124.47–88.45)	68.14 (82.61–62.83)	73.56 (90.43–48.21)	***<0.001***
Cricopharyngeal Inlet	113.48 (130.06–90.08)	69.69 (93.19–62.12)	56.09 (60.65–53.33)	54.78 (62.42–50.12)	***<0.001***
Glottic Area	71.72 (80.22–62.51)	47.85 (63.40–33.76)	34.76 (46.12–31.17)	43.73 (56.31–39.18)	***<0.001***
Mandible	902.8 (944.3–794.7)	554.4 (702.6–504.1)	625.6 (671.0–602.3)	645.9 (676.6–601.1)	***<0.001***
Oesophagus Cervical	37.15 (48.16–28.04)	33.97 (40.87–25.32)	25.295 (30.567–12.970)	37.36 (43.33–22.53)	0.0186
Oral Cavity (Extended)	641.5 (765.6–545.0)	487.3 (602.0–393.9)	354.7 (418.5–317.5)	375.6 (421.8–354.4)	***<0.001***
Parotid Left	316.3 (388.8–293.3)	193.2 (267.5–174.6)	192.4 (218.8–151.3)	211.0 (238.1–186.8)	***<0.001***
Parotid Right	336.8 (393.7–299.3)	196.4 (251.1–169.3)	183.3 (196.8–159.5)	215.1 (240.8–168.6)	***<0.001***
Pharyngeal Constrictor Muscles	469.9 (512.5–427.6)	340.6 (521.5–290.6)	293.8 (315.5–238.7)	314.1 (326.3–262.7)	***<0.001***
Spinal Cord	299.1 (346.3–270.3)	205.8 (237.9–188.0)	173.1 (181.8–167.3)	200.5 (212.3–173.5)	***<0.001***
Submandibular Left	135.24 (148.79–113.25)	84.91 (99.73 -	79.09 (82.81–66.99)	79.21 (89.47–72.46)	***<0.001***
Submandibular Right	131.51 (149.42–115.29)	90.04 (111.44–77.65)	81.19 (86.56–72.92)	79.12 (91.42–73.05)	***<0.001***
Supraglottic Larynx	257.9 (300.2–231.2)	207.9 (285.9–161.0)	134.07 (167.39–97.80)	158.98 (175.80–128.82)	***<0.001***
Thyroid	236.10 (328.87–184.63)	203.41 (260.67–149.96)	161.36 (196.91–132.85)	189.9 (218.4–167.7)	0.00516

Table 8
Patient Demographic Data per Site.

Site	Age	Sex	Contrast
Yonsei	56	M	Y
	67	M	Y
	68	M	Y
	56	M	Y
	33	M	Y
	48	M	Y
	35	F	Y
	74	M	Y
	62	M	Y
	54	M	Y
	58	M	Y
	48	M	Y
	43	M	Y
	47	M	Y
	49	M	Y
	65	M	Y
	53	M	Y
	51	M	Y
	70	F	Y
	42	M	Y
Shantou	60	M	N
	46	M	Y
	40	F	Y
	45	F	Y
	38	M	Y
	70	M	Y
	56	M	N
	67	M	Y
	66	M	Y
	67	F	Y
	23	M	Y
	30	M	Y
	63	M	Y
	51	F	Y
	81	M	Y
	58	M	Y
	60	F	Y
	63	M	Y
	74	M	Y
	50	M	Y
Groningen	72	F	Y
	80	M	N
	61	M	N
	76	M	N
	78	F	N
	54	M	N
	67	M	N
	94	F	N
	62	M	N
	74	F	N
	56	M	N
	77	M	N
	89	F	N
	84	M	N
	74	M	Y
	66	M	Y
	67	M	Y
	70	M	Y
	71	F	Y
	57	M	Y
Maastro	66	M	Y
	66	M	Y
	59	F	Y
	74	M	N
	57	M	Y
	73	F	Y
	86	M	Y
	80	F	Y
	76	M	Y
	71	M	Y
	73	F	N
	66	F	Y
	73	M	Y
	62	M	Y
	72	F	Y
	64	M	Y
	69	F	Y
	69	M	Y
	91	M	N
	72	M	N
Summary	Mean (63)Median (66)	M (61)F (20)n/a (2)	Contrast (61)Non-Contrast (24)

NB 1: Data was anonymized according to local patient data privacy regulations.

most noticeable difference in the box plots of DSC. Prior to Bonferroni correction, statistically significant differences between the groups were found for at least on measure for all organs except the Cricopharyngeal Inlet. Following Bonferroni correction, statistically significant differences continued to be observed for only; BM L&R for DSC, PCM for DSC, and Submandibular R for DSC and NAPL. Results for these organs will be expanded below. All quantitative results are given in Tables 4–6 in the Supplementary Material.

The Kruskal-Wallis test of the Pharyngeal Constrictor Muscle showed statistical differences for the PCM ($p < 0.05$) for all quanti-

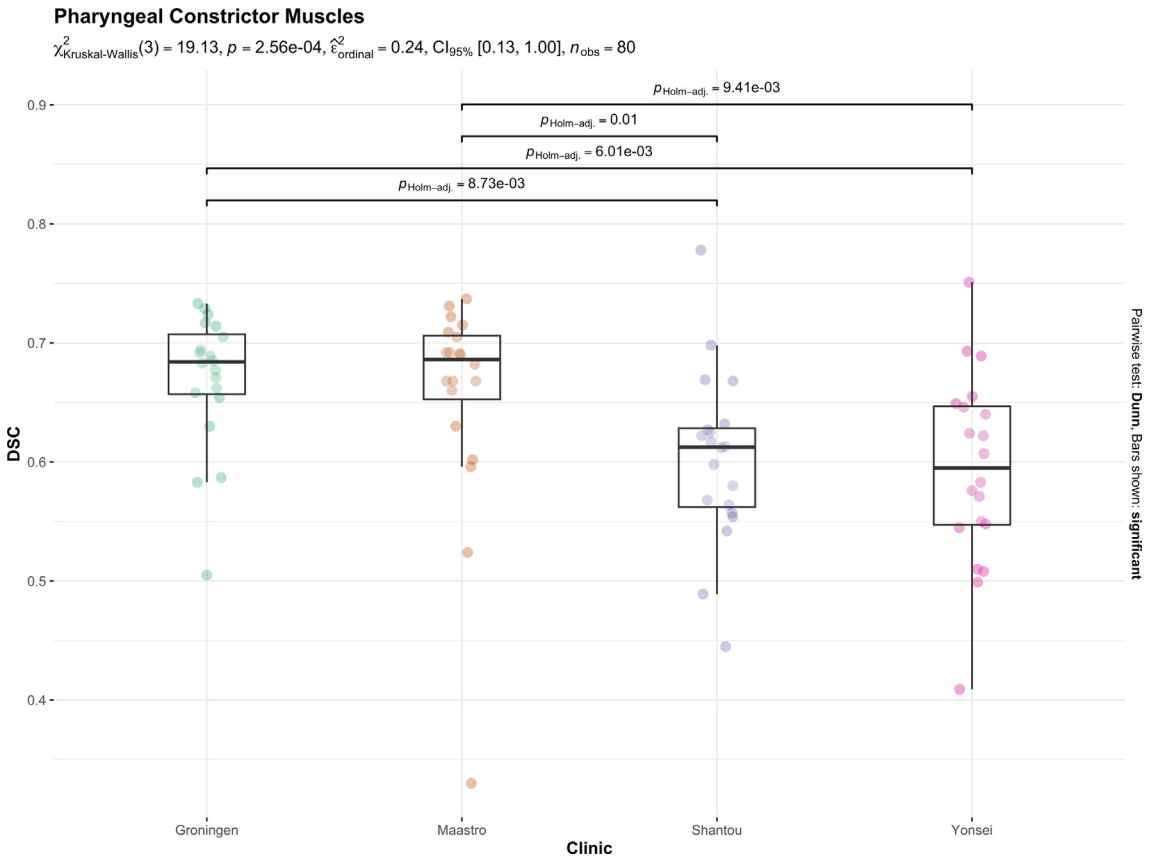

Fig. 3. Boxplot showing significant difference between clinics for DSC scores of Pharyngeal Constrictor Muscles.

tative measures prior to Bonferroni correction. After correction ($p < 0.00104$), the statistical significance remained for DSC. A post-hoc Dunn test showed a statistical difference between the European clinics (Groningen and Maastro) and the Asian clinics (Shantou and Yonsei), visualized in Fig. 3, indicating the differences between the European and Asian grouped clinics.

The Buccal Mucosae DSC scores for both left and right BM were found have a significant difference between clinics. A Dunn test showed differences between both European clinics (Groningen and Maastro) and Yonsei, as seen in Fig. 4. HD2D95 and NAPL showed no observed differences.

The quantitative measures showed significant differences between the groups for Submandibular Right for all measures prior to Bonferroni correction, and for DSC and NAPL post-correction. However, no significant differences were found for the left side. The Dunn test on the DSC scores indicated the results from both Groningen and Maastro differed to those from Shantou. An additional difference was found between Maastro and Yonsei for NAPL. There is a significant difference between the European grouped clinics and Shantou, as seen in Fig. 5.

Quantitative evaluation of organs reference volume and length showed anatomical variations between the Asian and European population (Figs. 9 and 10).

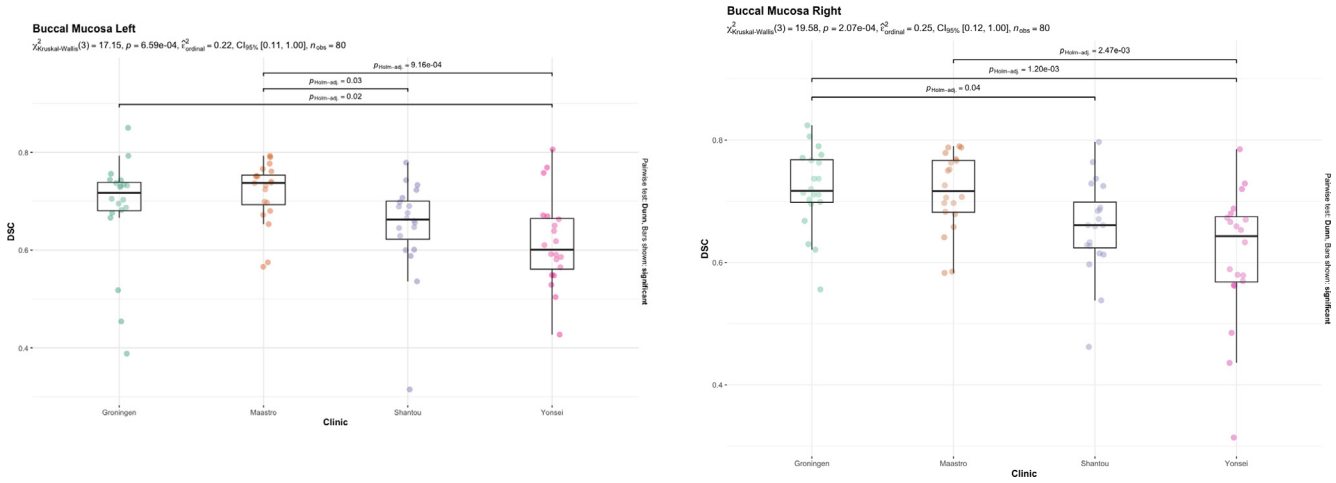

Fig. 4. Boxplots showing significance and relationship between clinics for DSC of both Buccal Mucosae.

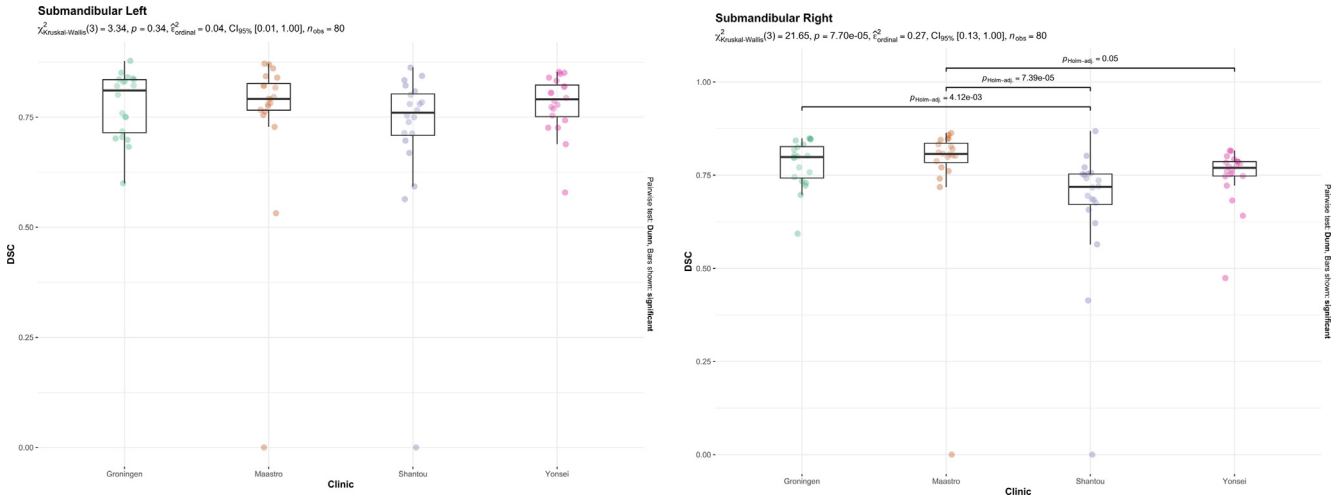

Fig. 5. Boxplots showing significance and relationship between clinics for DSC for the Submandibular Left and Submandibular Right.

In the qualitative assessment, no clear trend was seen when observing results for the blinded acceptance test by Patient Origin, shown in the plots in the panel 'A' of Fig. 6 . The rates of clinical acceptance were comparable across all structures for the manual contours, with minor variation in acceptance for BM Right, as seen in lower panel 'A'. Rates of clinical acceptance of the DLC contours did not mirror to the rates seen for the GT. However, there was no clear trend, as seen in the top row of panel 'A'. For example, acceptance was highest for the left parotid in the Korean patient group and lowest in the Chinese patient group, but this trend is reversed for the PCM.

Similarly, there was no clear trend amongst observers across the ground truth data grouped according to Observer Origin, in panel 'B' of Fig. 6. Largely, all observers have similar rates of clinical acceptance of the manual contours, regardless of observer origin. This is shown in lower panel 'B' of Fig. 6. Generally, the level of acceptance was lower for the DLC contours, except for the Mandible. However, there is a marked difference in how much the contours are considered clinically acceptable according to observer origin. Observers from South Korea accepted the DLC contours at similar rates to the manual contours. Except for the glottic larynx, which the Chinese and Dutch observers were more critical of the DLC contours, other than the Mandible. This can be seen in the acceptance of DL plots in lower panel 'B' of Fig. 6.

Discussion

This study sought to investigate whether racial/demographic bias in auto-contouring exists for a model training in a single institution, and what the impact of any bias might be.

Quantitative evaluation of organ volume revealed statistical differences between the Asian and European population for several organs. These anatomical variations between populations may explain the observed differences in the quantitative measures of autocontouring performance for some organs. For the BM, differences were observed, with the European patient population having larger volumes, and higher DSC. It is known that DSC correlates with organ volume [37] and this difference may not reflect a difference in performance but rather a difference in the organ volume. As for the Mandible, there was a significant volume difference between the Asian and European population, with the former having larger volumes. However, statistical differences were not observed for any measure of contouring performance. This may be attributed to the Mandibles high contrast, which consistently performs well with DLC. Differences found between the Asian and European population for volume, were shared for organ length. Measures concerned with length, NAPL and HD2595, revealed population differences for many organs prior to Bonferroni correction.

There were two organs for which quantitative performance measure differences were observed but population (organ volume) differences were not. The Submandibular R is perhaps the most interesting of these, since the results were not mirrored in the left side. In this case, outliers with poor performance appears to skew the Shantou and Maastro results, respectively. This highlights the challenge of performing such analysis with small sample sizes. Therefore, it appears that there is limited bias overall in the quantitative performance of the DLC for the various populations.

While helpful, quantitative metrics alone are not sufficient to determine clinical acceptability, and a qualitative test was required to evaluate the contours in a manner that closely simulates clinical practice [36]. Overall, the qualitative data supports the assumption that DLC is not biased to patient origin, since there is no trend for accepting contours from one population more than another. However, there appears to be a bias in the perception of the clinical acceptance of DL contours between observers from different origins. While the acceptance of manual contours was consistent between observers, there was a strong difference in acceptance with the European clinics being more critical and the Korean clinic being more accepting of the DL contours. Though this study is insufficient to determine the cause of this bias, we can speculate that clinical workload, familiarity with advanced technologies, training and experience as possible factors that may contribute to the bias seen. Nevertheless, this study shows that focus is required to understand why there are differences in perception and how it may be better aligned amongst observers. Much like the differences seen in interobserver variation studies for delineation, the same phenomenon may be shared when assessing contours [11,27].

There is growing awareness that AI solutions could amplify racial bias and healthcare disparities [38–39]. A report commissioned by NHS Race and Health Observatory [40] found that research studies using clinical data, often did not include race data. Within diagnostic and therapeutic medical imaging, this may not routinely be recorded as it is assumed that medical imaging tasks are race agnostic [41]. As such, it is unclear if the patients in the population are equitably served.

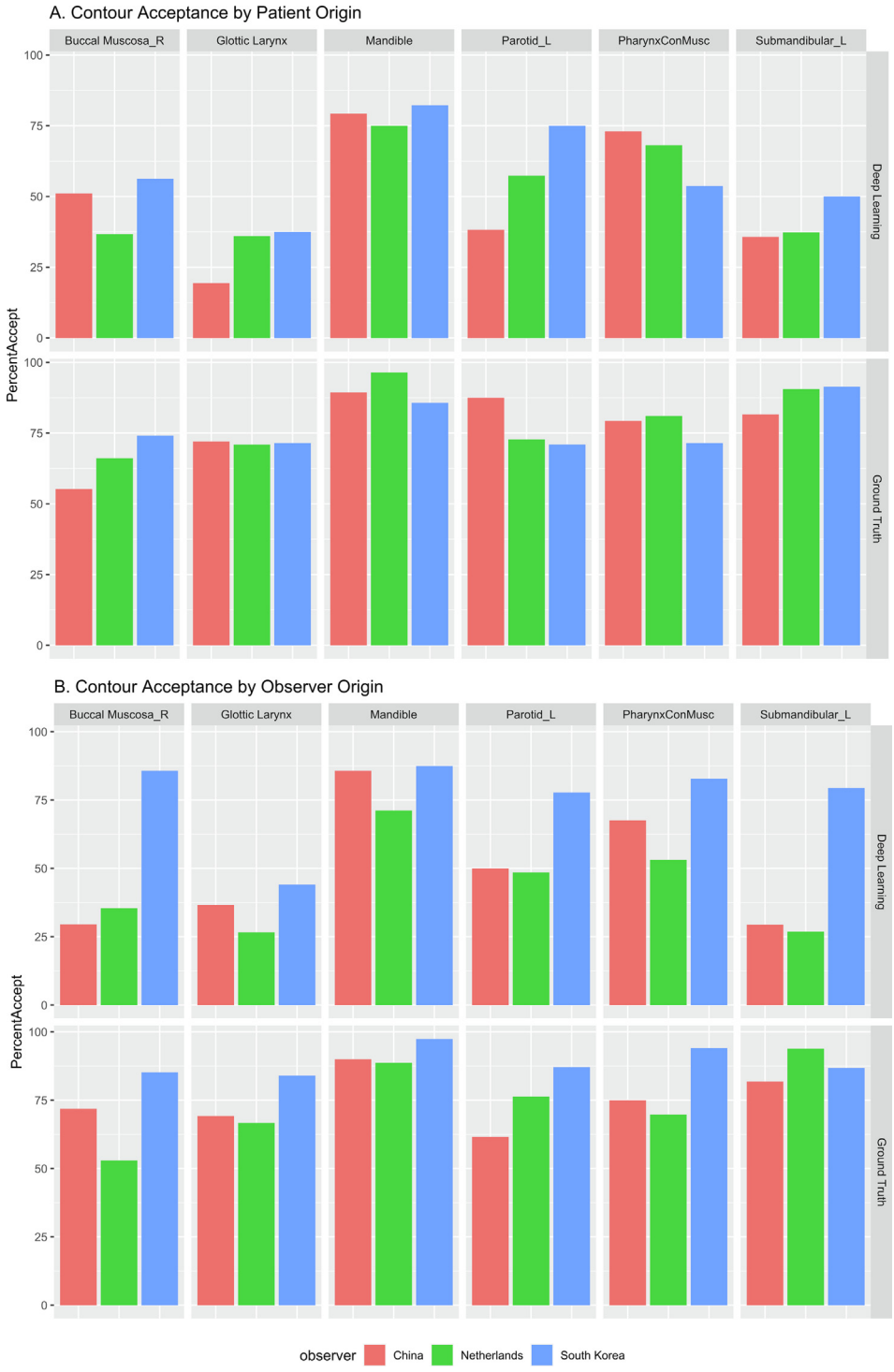

Fig. 6. Results from blinded acceptance test reviewing manual and DLC contours. Bar graphs show clinical acceptance, grouped per Organ and Clinic and by either Patient or Observer Origin.

Race is considered a protected attribute and is classed as special category data [42]protected under the Equality Act. The first step in representing relevant population data to build these models, is to first document patient race actively. However, protected attribute data is unlikely to be available to an AI model developer – a challenge that must be overcome if the risk of bias in AI is to be mitigated.

The generalisation of the results is limited as only one model with specific parameters was studied and other models may exhibit greater or different biases. While further investigation is required, this study provides some encouragement that autocontouring appears robust to demographic. Independent testing with a coordinated effort to collect data that represents many geographic populations, is needed. It is not enough to report the diversity of the training data used to develop a deep learning model, without evaluating the resulting generalisation of the model to a diverse population.

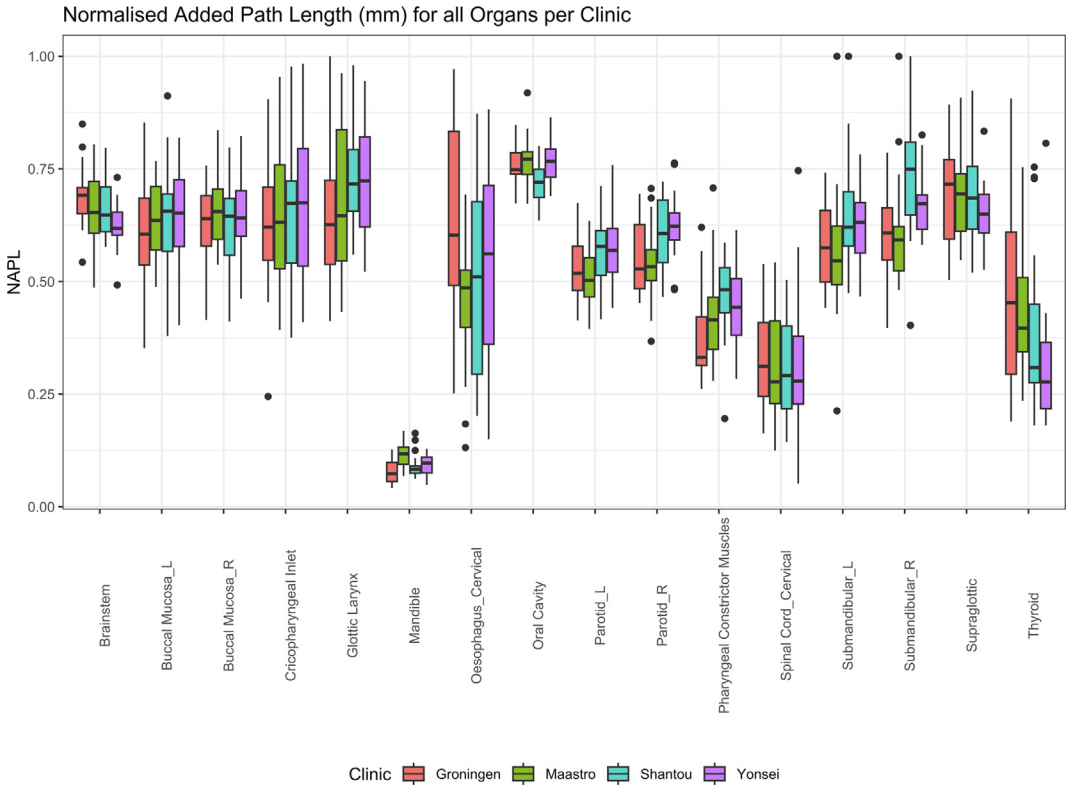

Fig. 7. Normalized Added Path Length (with 1 mm acceptable tolerance) for each organ and patient origin.

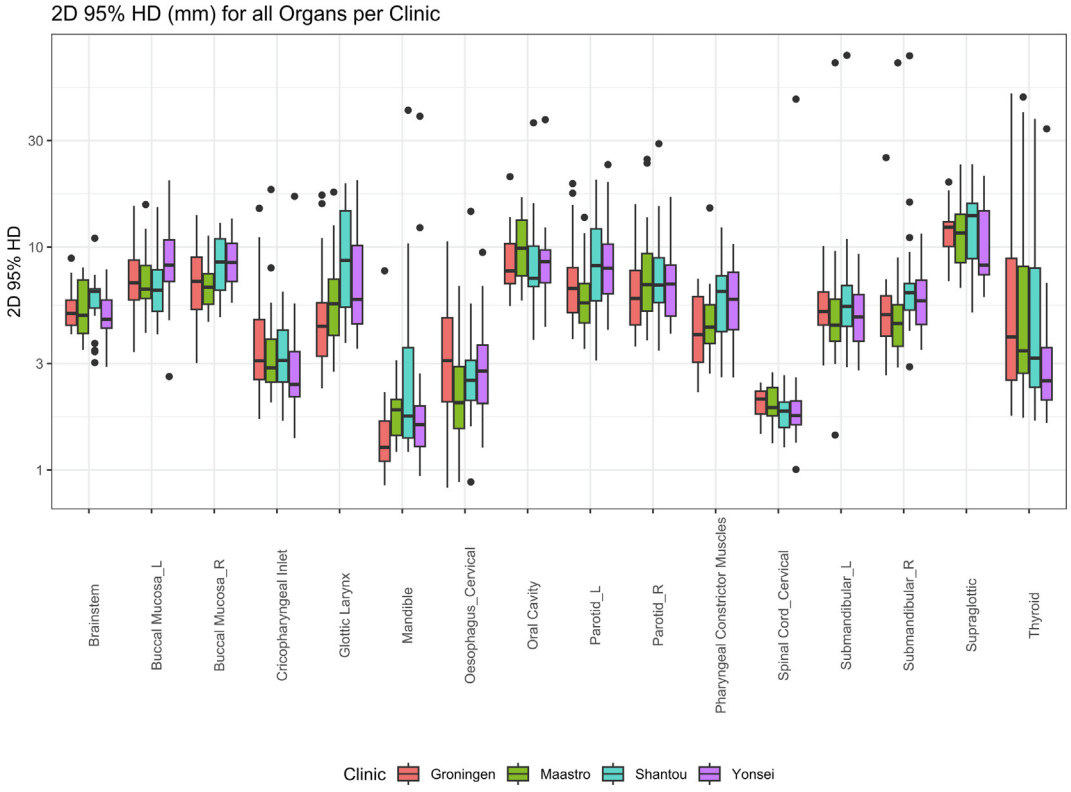

Fig. 8. 2D 95% Hausdorff Distance per organ and patient origin.

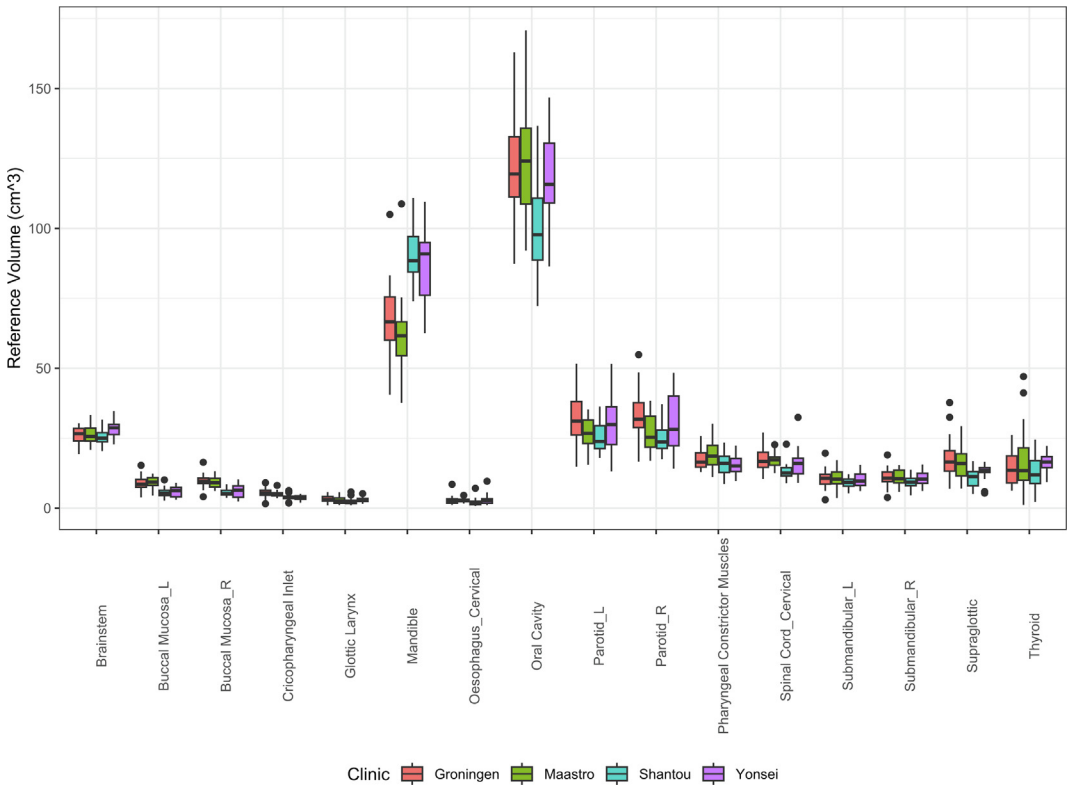

Fig. 9. All Organs Reference Volume (cm^3).

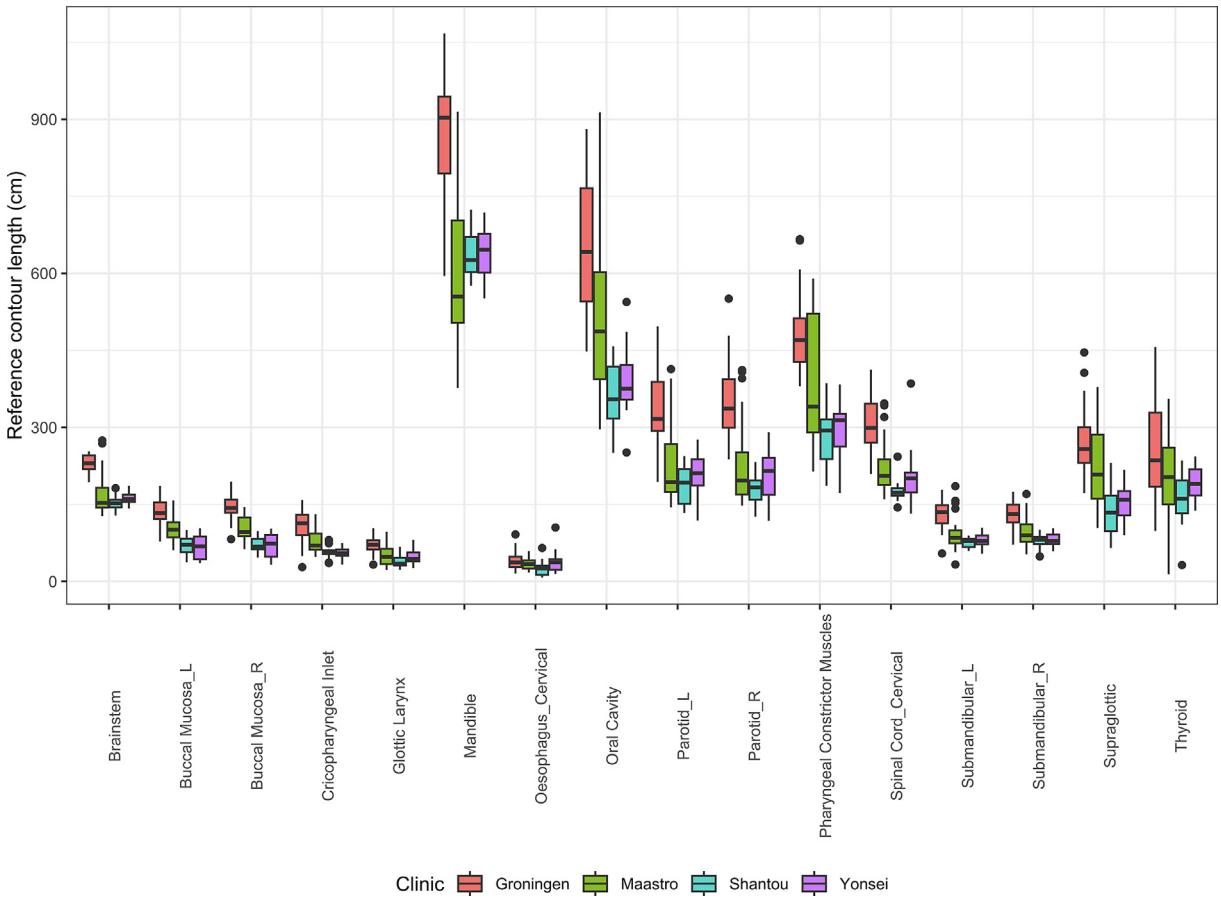

Fig. 10. All Organs Reference Length (cm).

Limitations

While this study sought to investigate whether racial/demographic bias in auto-contouring exists, to do so is challenging for several reasons; racial information was not recorded and clinic location had to be used as a surrogate, variations in clinical contouring guidelines, and quantitative metrics do not necessarily reflect clinical impact. Acquisition parameters were similar, but CT scanners differed (Table 2, Supplementary Material) making it a potential confounder. Sample size of this study is also a significant limitation of this study. However, it was decided to use a single observer recontouring all cases to mitigate contouring style as a confounder, making a larger sample size difficult to achieve. Further research of this nature should be conducted with larger sample sizes, other anatomical regions and possibly exploring further territories.

Conclusion

The deep learning autocontouring model for radiotherapy showed some organs with statistically significant differences in quantitative scores across geographic populations. However, some of these apparent quantitative differences in performance may be attributed to the choice of metric. The results of the qualitative evaluation showed that no bias was found regarding patient origin, rather that there was an observed difference in the perceived acceptance of deep learning autocontours amongst the observers. Further research should be undertaken to understand geographic biases extending into other anatomies and geographies. In addition, those developing autocontouring models should be mindful that training populations should reflect the treatment population. The implementation of independent testing on a diverse dataset would be a starting point towards improving generalisability and data diversity.

Conflict of Interest

Yasmin McQuinlan and Mark Gooding are employees of Mirada Medical Ltd.

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.

References

[1] Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. Br J Radiol 2019;92:20190001. https://doi.org/10.1259/bjr.20190001.

[2] Van den Bosch L, van der Laan HP, van der Schaaf A, et al. Patient-Reported toxicity and quality-of-life profiles in patients with head and neck cancer treated with definitive radiation therapy or chemoradiation. Int J Radiat Oncol 2021;111:456–67. https://doi.org/10.1016/j.ijrobp.2021.05.114.

[3] Brouwer CL, Steenbakkers RJ, van den Heuvel E, et al. 3D Variation in delineation of head and neck organs at risk. Radiat Oncol 2012;7:32. https://doi.org/10.1186/1748-717X-7-32.

[4] Peters LJ, O'Sullivan B, Giralt J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. J Clin Oncol 2010;28:2996–3001. https://doi.org/10.1200/JCO.2009.27.4498.

[5] Lustberg T, van Soest J, Gooding M, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother Oncol 2018;126:312–7. https://doi.org/10.1016/j.radonc.2017.11.012.

[6] Kiljunen T, Akram S, Niemelä J, et al. A deep learning-based automated CT segmentation of prostate cancer anatomy for radiation therapy planning-a retrospective multicenter study. Diagnostics 2020;10:959. https://doi.org/10.3390/diagnostics10110959.

[7] Zabel WJ, Conway JL, Gladwish A, et al. Clinical evaluation of deep learning and atlas-based auto-contouring of bladder and rectum for prostate radiation therapy. Pract Radiat Oncol 2021;11:e80–9. https://doi.org/10.1016/j.prro.2020.05.013.

[8] Mattiucci GC, Boldrini L, Chiloiro G, et al. Automatic delineation for replanning in nasopharynx radiotherapy: what is the agreement among experts to be considered as benchmark? Acta Oncol 2013;52:1417–22. https://doi.org/10.3109/0284186X.2013.813069.

[9] Almberg SS, Lervåg C, Frengen J, et al. Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. Radiother Oncol J Eur Soc Ther Radiol Oncol 2022;173:62–8. https://doi.org/10.1016/j.radonc.2022.05.018.

[10] Wong J, Huang V, Giambattista JA, et al. Training and validation of deep learning-based auto-segmentation models for lung stereotactic ablative radiotherapy using retrospective radiotherapy planning contours. Front Oncol 2021;11:. https://doi.org/10.3389/fonc.2021.626499626499.

[11] van Dijk LV, Van den Bosch L, Aljabar P, 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.

[12] 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.

[13] Liu Z, Liu X, Guan H, et al. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. Radiother Oncol 2020;153:172–9. https://doi.org/10.1016/j.radonc.2020.09.060.

[14] Oktay O, Nanavati J, Schwaighofer A, et al. Evaluation of deep learning to augment image-guided radiotherapy for head and neck and prostate cancers. JAMA Netw Open 2020;3:e2027426. https://doi.org/10.1001/jamanetworkopen.2020.27426.

[15] Blanchard P, Gregoire V, Petit C, et al. A blinded prospective evaluation of clinical applicability of deep learning-based auto contouring of OAR for head and neck radiotherapy. Int J Radiat Oncol Biol Phys 2020;108:e780–1. https://doi.org/10.1016/j.ijrobp.2020.07.239.

[16] Iyer A, Thor M, Onochie I, et al. Prospectively-validated deep learning model for segmenting swallowing and chewing structures in CT. Phys Med Biol 2022;67:. https://doi.org/10.1088/1361-6560/ac4000024001.

[17] Song Y, Hu J, Wu Q, et al. Automatic delineation of the clinical target volume and organs at risk by deep learning for rectal cancer postoperative radiotherapy. Radiother Oncol 2020;145:186–92. https://doi.org/10.1016/j.radonc.2020.01.020.

[18] Duan J, Bernard M, Downes L, et al. Evaluating the clinical acceptability of deep learning contours of prostate and organs-at-risk in an automated prostate treatment planning process. Med Phys 2022;49:2570–81. https://doi.org/10.1002/mp.15525.

[19] Ma C, Zhou J, Xu X, et al. Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer. J Appl Clin Med Phys 2022;23. https://doi.org/10.1002/acm2.13470.

[20] Byun HK, Chang JS, Choi MS, et al. Evaluation of deep learning-based autosegmentation in breast cancer radiotherapy. Radiat Oncol 2021;16:203. https://doi.org/10.1186/s13014-021-01923-1.

[21] 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;13:702. https://doi.org/10.3390/cancers13040702.

[22] Garrett Fernandes M, Bussink J, Stam B, et al. Deep learning model for automatic contouring of cardiovascular substructures on radiotherapy planning CT images: dosimetric validation and reader study based clinical acceptability testing. Radiother Oncol 2021;165:52–9. https://doi.org/10.1016/j.radonc.2021.10.008.

[23] Cardenas CE, Beadle BM, Garden AS, et al. Generating high-quality lymph node clinical target volumes for head and neck cancer radiation therapy using a fully automated deep learning-based approach. Int J Radiat Oncol 2021;109:801–12. https://doi.org/10.1016/j.ijrobp.2020.10.005.

[24] Weston AD, Korfiatis P, Philbrick KA, et al. Complete abdomen and pelvis segmentation using U-net variant architecture. Med Phys 2020;47:5609–18. https://doi.org/10.1002/mp.14422.

[25] Robert C, Munoz A, Moreau D, et al. Clinical implementation of deep-learning based auto-contouring tools–Experience of three French radiotherapy centers. Cancer/Radiothérapie 2021;25:607–16. https://doi.org/10.1016/j.canrad.2021.06.023.

[26] Christianen MEMC, Langendijk JA, Westerlaan HE, van de Water TA, Bijl HP. Delineation of organs at risk involved in swallowing for radiotherapy treatment planning. Radiother Oncol 2011;101:394–402. https://doi.org/10.1016/j.radonc.2011.05.015.

[27] Brunenberg EJL, Steinseifer IK, van den Bosch S, et al. External validation of deep learning-based contouring of head and neck organs at risk. Phys Imaging Radiat Oncol 2020;15:8–15. https://doi.org/10.1016/j.phro.2020.06.006.

[28] Wong J, Huang V, Wells D, et al. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. Radiat Oncol 2021;16:101. https://doi.org/10.1186/s13014-021-01831-4.

[29] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. Published online January 25, 2022. Accessed July 14, 2022.

[30] Medicines and Healthcare products Regulatory Agency. Good Machine Learning Practice for Medical Device Development: Guiding Principles -

GOV.UK. Published October 27, 2021. Accessed July 19, 2022. https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles.

[31] Noseworthy PA, Attia ZI, Brewer LC, et al. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. Circ Arrhythm Electrophysiol 2020;13:e007988. https://doi.org/10.1161/CIRCEP.119.007988.

[32] Puyol Anton E, Ruijsink B, Piechnik SK, et al. Fairness in AI: are deep learning-based CMR segmentation algorithms biased? Eur Heart J 2021;42. https://doi.org/10.1093/eurheartj/ehab724.3055. ehab724.3055.

[33] Vinod SK, Min M, Jameson MG, Holloway LC. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. J Med Imaging Radiat Oncol 2016;60:393–406. https://doi.org/10.1111/1754-9485.12462.

[34] Brouwer CL, Steenbakkers RJHM, 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.

[35] Yang J, Sharp GC, Gooding MJ. Auto-Segmentation for Radiation Oncology: State of the Art. CRC Press; 2021.

[36] Gooding MJ, Smith AJ, Tariq M, et al. Comparative evaluation of autocontouring in clinical practice: a practical method using the Turing test. Med Phys 2018;45:5105–15. https://doi.org/10.1002/mp.13200.

[37] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29. https://doi.org/10.1186/s12880-015-0068-x.

[38] Brault N, Saxena M. For a critical appraisal of artificial intelligence in healthcare: the problem of bias in MHEALTH. J Eval Clin Pract 2021;27:513–9. https://doi.org/10.1111/jep.13528.

[39] DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. J Am Med Inform Assoc 2020;27:2020–3. https://doi.org/10.1093/jamia/ocaa094.

[40] Kapadia D, Zhang J, Salway S, et al. Ethnic Inequalities in Healthcare: A Rapid Evidence Review. NHS Race & Health Observatory; 2022:17. Accessed September 14, 2022. https://www.nhsrho.org/wp-content/uploads/2022/02/RHO-Rapid-Review-Final-Report_v.7.pdf.

[41] Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health 2022;4:e406–14. https://doi.org/10.1016/S2589-7500(22)00063-2.

[42] Information Commissioner's Office. Special category data. Published February 15, 2021. Accessed July 19, 2022. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/.