PAPER • **OPEN ACCESS**

Geometric evaluations of CT and MRI based deep learning segmentation for brain OARs in radiotherapy

To cite this article: Nouf Alzahrani *et al* 2023 *Phys. Med. Biol.* **68** 175035

IOP Publishing Phys. Med. Biol. 68 (2023) 175035 https://doi.org/10.1088/1361-6560/acf023

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
5 May 2023

REVISED
19 July 2023

ACCEPTED FOR PUBLICATION
14 August 2023

PUBLISHED
29 August 2023

PAPER

Geometric evaluations of CT and MRI based deep learning segmentation for brain OARs in radiotherapy

Nouf Alzahrani[1,2,3,*], **Ann Henry**[2,3], **Anna Clark**[3], **Louise Murray**[2,3], **Michael Nix**[3] and **Bashar Al-Qaisieh**[3]

[1] King Abdulaziz University, Department of Diagnostic Radiology, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia
[2] University of Leeds, School of Medicine, Leeds, United Kingdom
[3] St James's University Hospital, Department of Medical Physics and Engineering, Leeds Cancer Centre, Leeds, United Kingdom
[*] Author to whom any correspondence should be addressed.

Keywords: brain, organs at risk, autosegmentation, 3D U-net, deep learning, MRI scans, CT scans

Supplementary material for this article is available online

Abstract

Objective. Deep-learning auto-contouring (DL-AC) promises standardisation of organ-at-risk (OAR) contouring, enhancing quality and improving efficiency in radiotherapy. No commercial models exist for OAR contouring based on brain magnetic resonance imaging (MRI). We trained and evaluated computed tomography (CT) and MRI OAR autosegmentation models in RayStation. To ascertain clinical usability, we investigated the geometric impact of contour editing before training on model quality. *Approach.* Retrospective glioma cases were randomly selected for training ($n = 32, 47$) and validation ($n = 9, 10$) for MRI and CT, respectively. Clinical contours were edited using international consensus (gold standard) based on MRI and CT. MRI models were trained (i) using the original clinical contours based on planning CT and rigidly registered T1-weighted gadolinium-enhanced MRI (MRIu), (ii) as (i), further edited based on CT anatomy, to meet international consensus guidelines (MRIeCT), and (iii) as (i), further edited based on MRI anatomy (MRIeMRI). CT models were trained using: (iv) original clinical contours (CTu) and (v) clinical contours edited based on CT anatomy (CTeCT). Auto-contours were geometrically compared to gold standard validation contours (CTeCT or MRIeMRI) using Dice Similarity Coefficient, sensitivity, and mean distance to agreement. Models' performances were compared using paired Student's t-testing. *Main results.* The edited autosegmentation models successfully generated more segmentations than the unedited models. Paired t-testing showed editing pituitary, orbits, optic nerves, lenses, and optic chiasm on MRI before training significantly improved at least one geometry metric. MRI-based DL-AC performed worse than CT-based in delineating the lacrimal gland, whereas the CT-based performed worse in delineating the optic chiasm. No significant differences were found between the CTeCT and CTu except for optic chiasm. *Significance.* T1w-MRI DL-AC could segment all brain OARs except the lacrimal glands, which cannot be easily visualized on T1w-MRI. Editing contours on MRI before model training improved geometric performance. MRI DL-AC in RT may improve consistency, quality and efficiency but requires careful editing of training contours.

1. Introduction

The worldwide incidence of brain tumours is growing (Soomro *et al* 2023). In young adults, brain cancer is the third most common cause of death (Brunese *et al* 2020). Every year, over 5000 people die from brain cancer, and currently, in the UK, it is anticipated that 102 000 adults and children will have brain cancer (Brunese *et al* 2020).

Radiation therapy (RT) is commonly used to treat brain cancer, using ionizing radiation to destroy cancer cells. However, RT may cause damage to normal healthy tissues, called organs at risk (OARs). Damaging OARs

in the brain can lead to hearing and visual deficits and neurocognitive alteration (Scoccianti *et al* 2015). The side effects of treatment are minimized through the radiotherapy treatment planning process by targeting the dose to the tumour while reducing the dose to OARs. A radiation oncologist manually delineates the target volume of the tumour and surrounding OARs using computed tomography (CT) and/or magnetic resonance imaging (MRI) simulation scans. However, manual contouring is associated with several challenges. Firstly, contouring is time-consuming; previous studies have reported that each patient may take several hours of clinician's time to delineate all OARs (Cardenas *et al* 2019, Wang *et al* 2019). This could affect the treatment outcomes due to the delay in the start of the treatment. Secondly, manual contouring is subjective, as a radiation oncologist or dosimerist performs the delineation of OARs based on their previous experience and knowledge, which is a source of inconsistency (Cardenas *et al* 2019). Several studies have shown high inter-operator variability in contouring, which may lead to inappropriately treating normal areas (Scoccianti *et al* 2015, van Dijk *et al* 2020). Accordingly, there is great demand in the field of RT for autosegmentation to standardize and enhance the quality of contours and make the process more efficient by streamlining the clinical workflow and reducing staff workload.

In the last decade, computing in RT has helped address manual contouring challenges through the development of autosegmentation algorithms. Deep learning based autosegmentation entered the field of RT after it was demonstrated that the convolutional neural networks (CNNs) could considerably improve image classification and recognition task predictions (Cardenas *et al* 2019, Brouwer *et al* 2020). Since then, there have been a considerable number of studies published on the performance of deep-learning autosegmentation for delineation of OARs, which demonstrate that it is outperforming traditional autosegmentation methods (Scoccianti *et al* 2015, Cardenas *et al* 2019, van Dijk *et al* 2020). The most popular method for medical images delineation is the U-net architecture, which was established by Ronneberger *et al* (Cardenas *et al* 2019). Typically, delineation of brain OARs is performed using a combination of CT and MRI images. CT is currently standard for treatment planning dose calculations, which are based on electron density. MRI is usually co-registered to CT and provides complimentary information for contouring, particularly for OARs that are very difficult to visualise on CT, such as the optic chiasm. Since CT is, however, used for dose calculations, some, more mobile, OARs may be contoured based on this, for example lenses and extra-cranial portions of the optic nerves. Recently, several efforts have been made to establish MRI-only treatment planning (Edmund and Nyholm 2017). Compared to CT, MRI offers better contrast for the soft tissue, consequently, it is a superior imaging modality for accurately detecting and localizing both the target volume and OARs (Schmidt and Payne 2015, Liu *et al* 2019). Additionally, MRI does not use ionizing radiation, which will reduce total radiation exposure to the patient. For MR-only RT treatment planning, instead of traditional CT, the needed electron density information is obtained through a synthetic-CT (sCT) produced from the MRI scan (Wiesinger *et al* 2018).

Compared to other treatment sites, few deep learning autosegmentation models currently identify brain OARs using MRI or CT scans. As far as we are aware, only one study has investigated commercial deep-learning autosegmentation software that uses a U-Net CNN to segment OARs in the brain using CT scans (Wong *et al* 2020). Three earlier research studies used 2D and 3D U-net with various modifications to develop MRI-based deep learning methods to delineate brain OARs (Chen *et al* 2019, Mlynarski *et al* 2020, Wiesinger *et al* 2021). Chen *et al* (2019) autosegmented six brain OARs (the orbits, optical nerves, brainstem, and chiasm) using T1-weighted MRI. Mlynarski *et al* (2020) used T1-weighted MRI to autosegment eleven OARs, including the orbits, brainstem, lenses, optic nerves, pituitary gland, optic chiasm, hippocampus, and brain. Wiesinger *et al* (2021) used T2-weighted MRI to autosegment fifteen OARs (the orbits, lenses, optical nerves, lacrimal glands, pituitary gland, chiasm, brainstem, brain, cochleas, and patient body contour). All these prior studies used deep learning to segment brain OARs on CT or MRI scans and produced acceptable segmentations, suitable for RT planning (Chen *et al* 2019, Mlynarski *et al* 2020, Wiesinger *et al* 2021). However, none of the proposed MRI deep-learning segmentation techniques are commercially available. Also, the previous studies focused on using only one imaging modality, CT or MRI. Clinically OARs must exist on the CT for RT planning, despite many being predominantly contoured on MRI by clinicians. The main objective of this study is to train and evaluate separate CT and MRI OAR deep learning segmentation models in RayStation (RaySearch AB, Stockholm) for brain radiotherapy, to ascertain clinical usability. Also, we aim to establish which modalities are required for the various OARs and whether standardising training data by editing clinical contours (on CT or MR) prior to training is beneficial (if the model's output improves the segmentation's quality) or necessary (if the model's output reduces the number of failed segmentations) for model performance.

2. Materials and methods

2.1. Dataset and clinical protocol

Sixty previously treated glioma cases with both CT and MRI available were randomly selected from a retrospective clinical cohort from the past 5 years using a computer generated simple-random list and used to build autosegmentation models for each modality. The total of 60 was chosen to enable careful quality assurance of the contours considering staff and time availability. The data was divided into 80% for training ($n = 48$) and 20% for testing ($n = 12$), which is the most popular and advised split ratio (Joseph 2022).

Brain CT scans was acquired using the following acquisition parameters: kVp: 120, FOV: 500 mm, 1 mm $\times$ 1 mm in-plane resolution, slice thickness: 2 mm, and scan type: helical scan on a Siemens Sensation. Moreover, the following acquisition parameters were used to acquire brain MRI scans: MRI sequence: T1w spin echo sequence, imaging plane: transverse, slice thickness: 2 mm, scanner: Siemens Magnetom Sola with 1 mm in-plane resolution and Gd contrast.

2.2. Brain OARs and gold standard atlas

The OAR selection was based on the four central nervous system clinical protocols at our institution: Meningioma, Pituitary, Glioma (Radical),and Glioma (Palliative). Thirteen OARs were selected for autosegmentation: brainstem, cochlea (left and right), orbits (left and right), lenses (left and right), optic chiasm, optic nerves (left and right), lacrimal glands (left and right), and pituitary gland.

A brain OAR atlas was developed as a gold standard example of contours with anatomical descriptions and contouring guidance, in line with international consensus guidelines (Scoccianti *et al* 2015, Eekers *et al* 2018, Ho *et al* 2018, Chen *et al* 2019, Mir *et al* 2020). All OARs were manually delineated using CT and MRI scans in combination, as per usual clinical practice. The atlas was reviewed and approved by the treating radiation neuro-oncology team.

2.3. Clinical contours and quality assurance (QA)

All the original clinical contours and image sets were reviewed in terms of image quality, contour accuracy and OAR labelling. OAR labelling was edited to be consistent with AAPM TG-263 guidelines Mayo *et al* (2018). The clinical contours were reviewed and edited where necessary to ensure alignment with the brain OAR atlas. The process was as follows: the original clinical contours (unedited contours-CT and MRI-based) were copied and then edited based on CT anatomy alone to create CT-edited contours. These were then copied onto the rigidly registered T1-weighted gadolinium-enhanced MRI and then reviewed and edited as necessary based on the MRI anatomy, again to align to clinical guidelines (MRI-edited contours).

2.4. Deep learning autosegmentation training

A commercially available 3D U-net (Çiçek *et al* 2016) was used to train all the autosegmentation models (RayStation 11 A, RaySearch Laboratories AB, Stockholm, Sweden).

Two CT autosegmentation models were trained using 47 cases (one case was excluded due to missing data). The first CT-based autosegmentation model was trained using the original clinical contours without editing, termed the CT unedited autosegmentation model (CTu). The second CT-based autosegmentation model was trained on the same dataset using the cases with CT-edited clinical contour termed the CT-edited autosegmentation model (CTeCT).

Three MRI autosegmentation models were trained on the same dataset using 32 cases (16 cases were excluded due to inconsistent MRI slice thickness). The first MRI-based model was trained using the original clinical contours copied from the CT scan without editing, termed the MRI unedited autosegmentation model (MRIu). The second MRI-based model was trained on the same dataset using the edited clinical contour (CTeCT), copied from the CT scan, termed the CT edited MRI autosegmentation model (MRIeCT). The third model was trained on the same dataset, using the CT-edited clinical contour, further edited based on the MRI scan, termed the MRI edited MRI autosegmentation model (MRIeMRI). After training, all the models were used to generate automatic contours on the validation cohort.

2.5. Deep learning autosegmentation validation

The performance of the models was geometrically evaluated on an independent dataset of 12 cases. Two cases were excluded from the CT validation cohort as no MRI scans were associated with them ($n = 10$ cases). Also, one MRI test case was excluded due to using different MRI sequences ($n = 9$ cases). The evaluation was done by comparing the generated contour to the gold standard contours in each modality, where clinical contours were edited based on each modality's anatomy in this validation cohort (i.e. CTeCT and MRIeMRI).

2.5.1. Geometric evaluation

The following test metrics were used for the geometric evaluation: the dice similarity coefficient (DSC) (Wong *et al* 2020), sensitivity (van Rooij *et al* 2019) and mean distance to agreement (MDA) (Jena *et al* 2010). Higher DSC and sensitivity scores indicate better agreement between the gold standard contour and autosegmentation, however lower MDA scores indicate that small distance errors exist between autosegmentation and gold standard contours.

To evaluate the statistical significance of these metrics and determine the impact of editing before training the model, each geometry test metric pair of the edited and unedited models was compared in each modality using the paired two-tailed Student's t-test. For the same patient, if the autosegmentation model failed to segment any OARs, and the comparable model was able to segment the missing OAR, this OAR was excluded from the pairwise comparison.

A Bonferroni correction was applied to factor in a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously. (3 metrics and 3 segmentation pairs for MRI, 3 metrics and one segmentation pair for CT). Bonferroni-corrected p-value thresholds for statistical significance were $\leqslant$0.005 (0.05/9) for MRI geometric evaluations, and $\leqslant$0.016 (0.05/3) for CT geometric evaluation.

3. Results

3.1. Comparison of CT versus MRI deep learning contours

CT and MRI- based deep-learning autocontouring (DL-AC) demonstrated excellent delineation quality for large structures such as brainstem, right and left orbits, with the exception of the CTu model which had poorer performance: average DSC and sensitivity scores ranged from 0.85 to 0.91 and from 0.85 to 0.94, respectively, across all three MRI-based models for these large structures (supplementary information tables S1 and S2) (figures 1 and 2). The CTeCT model average DSC and sensitivity scores ranged from 0.87 to 0.90 and 0.88 to 0.93, respectively across these OARs, while the CTu model average DSC and sensitivity scores ranged from 0.62 to 0.64 and from 0.62 to 0.63, respectively for the same set of structures (supplementary information tables S4 and S5) (figure 3).

The geometric assessments indicated that CT-based DL-AC performed worst in the delineation of the optic chiasm. The lowest DSC and sensitivity average scores were for the optic chiasm for both CT-based models. The average scores for DSC were 0.18 and 0.29, and the sensitivity was 0.15 and 0.28, for CTeCT and CTu, respectively (supplementary information tables S4 and S5). MDA evaluations showed that the CTeCT model had the highest average MDA score for the Optic chiasm (0.40 cm), whereas the CTu models had the highest score for the right lacrimal gland (0.43 cm) (supplementary information table S6).

In contrast, geometric evaluations showed that MRI-based DL-AC performed worst for delineation of the lacrimal gland: the lowest DSC and sensitivity average scores were obtained for the left and right lacrimal glands delineated by all MRI-based DL-AC models. For MRI-based DL-AC models, the average lacrimal gland DSC scores ranged from 0.02 to 0.15, and the sensitivity ranged from 0.02 to 0.10. Furthermore, the highest average MDA score for the MRIeMRI and MRIu models was that for the left lacrimal gland (0.23 cm and 0.42 cm, respectively), while for the MRIeCT model, the highest average MDA score was for the optic chiasm (0.33 cm) (supplementary information tables S1–S3).

3.2. The value of editing contours before training

The necessity of editing contours so that they align with an agreed atlas was established based on segmentation failure rates. Edited autosegmentation models generated more successful segmentations on OARs than unedited models in both modalities. The CTeCT model reduced the number of failed OAR segmentations compared to the CTu model (4 cf. 36) while the MRIeCT model resulted in a similar total number of failures compared to MRIu (21 cf. 22). MRIeMRI, however, reduced the failure rate to 13 and reduced failures to near zero for all organs except for the lacrimal glands, where more failures occurred with the edited MRI-based model (MRIeMRI) (10 of 13 failures were for the lacrimal glands). The MRIu model exhibited a high failure rate for the cochlea, which was almost entirely resolved when using the MRIeMRI model (supplementary information table S7).

A statistically significant quality difference between the CTeCT and CTu autosegmentation models was found only for the optic chiasm for all the geometry metrics (p = 0.009, 0.008, and 0.001 and effect size = 0.260, 0.160, and 0.150 cm for DSC, sensitivity, and the MDA, respectively) (supplementary information table S8).

Regarding the MRI autosegmentation models, there was no statistically significant difference between the MRIeCT and MRIu models for any geometric comparison, except for the right orbit as assessed by the sensitivity metric, where the effect size was small (p = 0.001, effect size = 0.080) (table 1).

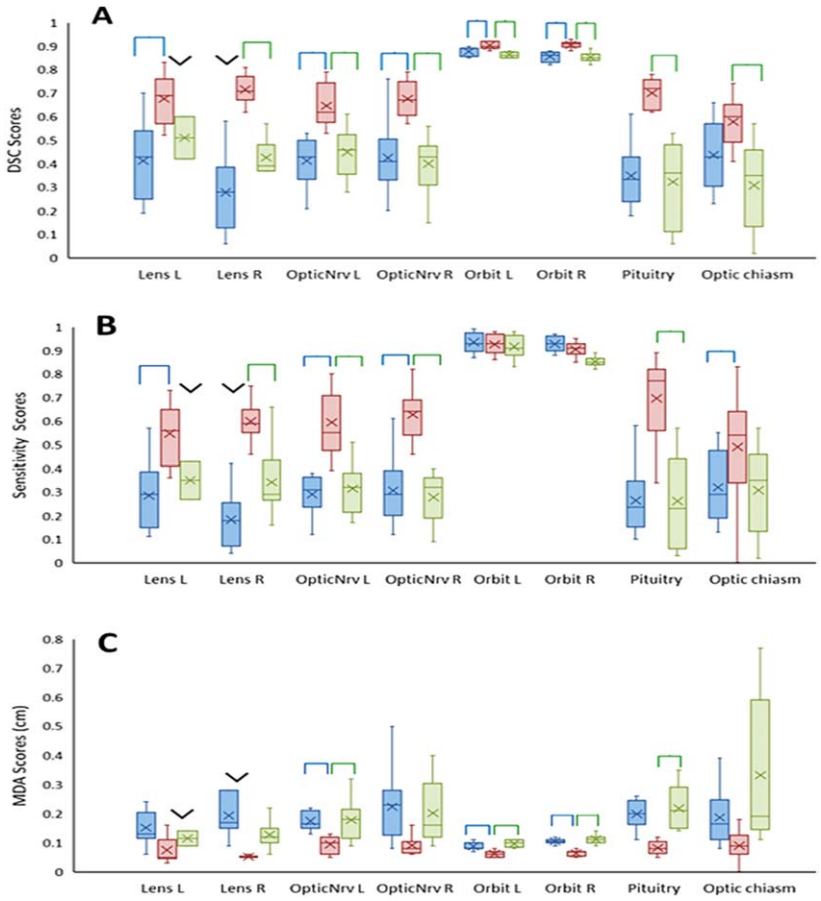

Figure 1. The distribution of the (a) DSC, (b) sensitivity, and (c) MDA (for MRI-based deep learning segmentations from three different MRI models): the MRIu segmentation (blue), the MRIeMRI segmentation (red), and the MRIeCT segmentation (green). The green square bracket denotes the statistically significant difference (paired T-test) between MRIeMRI and MRIeCT, the blue square bracket denotes the statistically significant difference between MRIeMRI and MRIu. Structures not segmented on one of the compared models were excluded from this analysis.The black chevron indicates that the statistical analysis was not performed in cases where less than 6 structures were segmented for any OAR or similar (outliers not shown for clarity).

A statistically significant difference was found between MRIeMRI versus both the MRIeCT and MRIu models in the delineation of the structures shown in figure 1. With the exception of the orbits, statistically significant differences (observed for lenses, optic nerves, pituitary, and optic chiasm) were associated with moderate to large effect sizes from 0.160 to 0.360 for DSC and from 0.230 to 0.540 for sensitivity and from 0.070 to 0.130 cm for MDA. The effect size for the orbits ranged from 0.010 to 0.240 in DSC and from 0.020 to 0.040 cm in MDA (table 1).

4. Discussion

This study examined the impact of editing clinical contours before training deep-learning autosegmentation models for brain OARs based on CT and MRI anatomy. Editing is a time-consuming process and should only be performed when there is evidence it will improve the model's performance.

The current study found that except for the lacrimal glands, MRI-based DL-AC is preferable for all brain OARs, particularly for delineating optic chiasm, which is known to be challenging for humans to delineate on CT due to lack of soft tissue contrast. CT based DL-AC was able to delineate optic chiasm (albeit with limited quality) given MRI derived clinical training contours. Conversely, lacrimal glands cannot be easily visualised on MRI without fat-saturation (Simon *et al* 1988), and even with CT derived clinical training contours, the performance of the MRI-based models for this OAR was not clinically acceptable. Accordingly, both modalities are needed for complete contouring of brain OARs, with lacrimal glands either segmented manually on CT or, potentially, via a separate CT-based DL-AC model. Alternately, a dual-modality autosegmentation model may overcome this issue, but may introduce inter-modality image registration issues (Mlynarski *et al* 2020). As there is a motivation to use MR-only RT for the brain, to allow improved target definition (Kazemifar *et al* 2019,

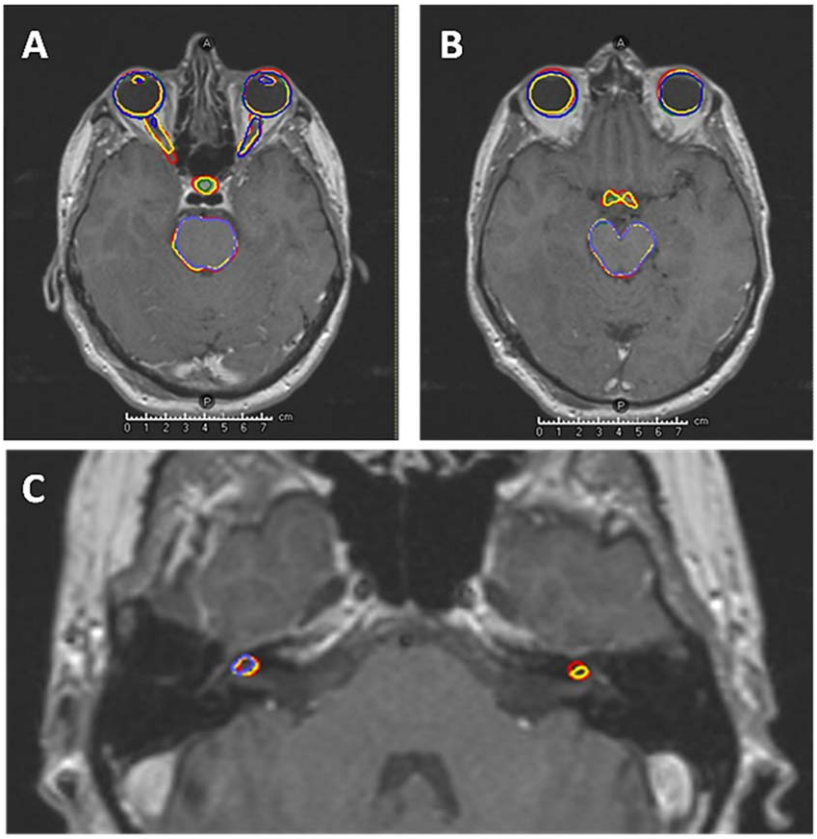

Figure 2. T1-weighted gadolinium-enhanced MRI showing examples of the predicted MRI deep learning segmentations compared to the gold standard segmentation of the orbits (a), (b), lenses (a), optic nerves (a), brainstem (a), (b), optic chiasm (b), cochlea (c), and pituitary (a). Red represents the gold standard segmentation. MRIeMRI is depicted in yellow, MRIeCT in green, and MRIu in blue. Lens L, cochlea L and R failed to be segmented by the MRIeCT model, while optic chiasm, pituitary, and cochlea L failed to be segmented by the MRIu model.

Lerner *et al* 2021, Ranta *et al* 2023), the T1-w MRI based DL-AC model would be sufficient to produce the segmentation , except for the lacrimal glands which would require manual contouring.

It has been recently demonstrated that T2-w MRI has the potential for direct DL-AC of lacrimal glands (Wiesinger *et al* 2021), creating the possibility for a multi-modality MRI model that could take advantage of the inherent registration of simultaneously acquired MR sequences.

The limitation of the current T1-w MRI model for lacrimal gland segmentation could also be related to training data quality. Lacrimal glands are typically segmented only on 2–3 slices, reducing the number of positive examples available to the model, exacerbating the lack of contrast available in non-fat saturated T1w imaging. The relatively small volume of the structures is also a factor, as it was previously reported that multi-organ DL-AC models can ignore small structures (Wang *et al* 2019), due to unbalanced losses. In our model, loss balancing across OARs was performed to minimise this effect.

Regarding other OARs, editing of clinical contours on MRI (MRIeMRI) reduced the number of failed segmentations to near zero for cochleae, lenses, optic chiasm, and pituitary and is therefore considered necessary (supplementary information table S7). The RayStation implementation of DL-AC uses an 'initialisation U-net' to find bounding boxes for each ROI and a set of 'refinement U-nets' to segment each ROI. If the initialisation network is unable to locate an organ; it will not be segmented at all. Hence, performance improvements in this network will affect the number of ROIs segmented, rather than the final segmentation quality. The number of ROIs that were segmented did increase after these structures were edited on MRI, suggesting that editing is crucial for the success of the initialization model.

Furthermore, significant differences ($p < 0.005$ after Bonferroni correction) between models were observed for at least one geometric measure for the following structures: optic nerves, orbits, lenses, optic chiasm, and pituitary (table 1). This indicates that editing these structures on MRI enhanced segmentation quality, even where the MRIu model successfully segmented the structure. For all structures showing statistically significant model-to-model performance variation, excluding orbits, effect sizes for DSC, sensitivity and MDA were often potentially clinically significant (Δ DSC > 0.2, Δ MDA > 0.1 cm and Δ sensitivity > 0.3). However, even though there was a significant difference between MRIeMRI versus MRIeCT and MRIu models in the

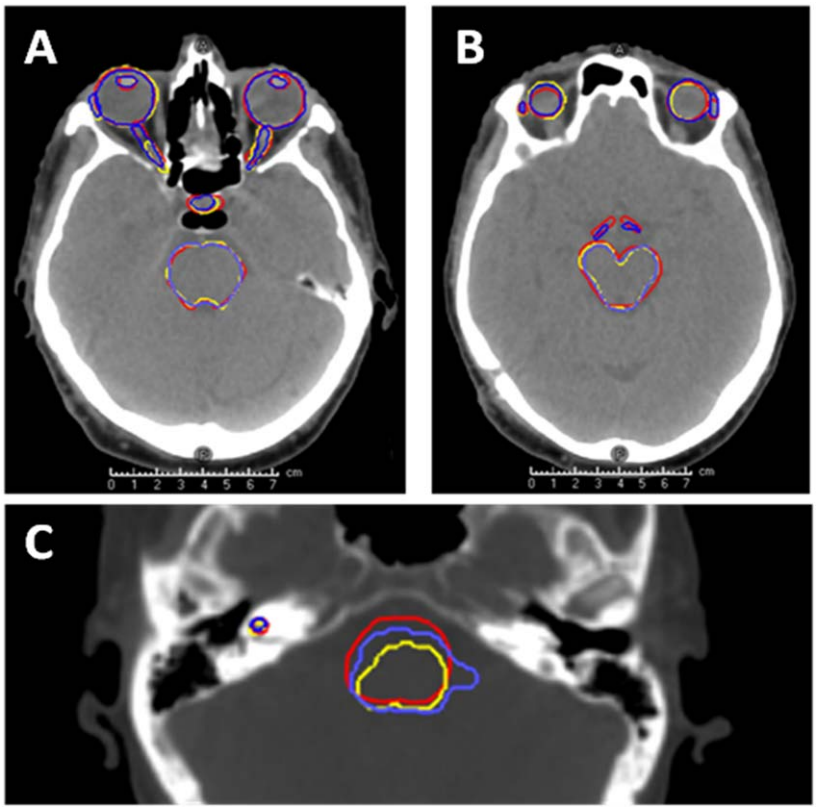

Figure 3. CT axial scans showing examples of the predicted CT deep learning segmentations compared to the gold standard segmentation of the orbits (a), (b), lenses (a), optic nerves (a), brainstem(a), (b), (c), optic chiasm (b), cochlea (c), lacrimal glands (a), (b), and pituitary (a). Red represents the gold standard segmentation. CTeCT is shown in yellow, while CTu in blue.

delineation of orbits ($p < 0.001$), the effect size was generally small (table 1). This was because the distribution of the DSC scores and the MDA for the orbits was narrow, due to their regular shape, so even a small effect was highly significant. The average DSC of the orbits was 0.91 (SD = 0.02) in the MRIeMRI, 0.86 (SD = 0.02) for MRIeCT and 0.87 (SD = 0.02) for MRIu model (supplementary information tables S1 and S3). These results imply that editing on MRI is beneficial for the above structures due to improved soft tissue contrast. The lack of soft tissue contrast and potential registration errors make editing on CT an inferior approach, where MR data are available.

For cochlea, insufficient cases were delineated by the MRIeCT and MRIu models to compare their performance with MRIeMRI. However, MRIeMRI was able to generate cochlea segmentations with high quality, average MDA = 0.84 mm (SD = 0.4 mm) (supplementary information table S3).

We have demonstrated a DL-AC model using a CE marked algorithm approved for clinical use, based on routine clinical T1-w MR imaging, for all clinically relevant brain OARs for RT. We demonstrated clinically acceptable geometric performance, following MRI based editing of training contours, comparable to previously published non-clinical algorithms for orbits brainstem and lenses, paving the way to the routine use of MR based DL-AC in brain RT (Chen *et al* 2019, Mlynarski *et al* 2020, Wiesinger *et al* 2021). Our model performed slightly worse for optic nerves and chiasm than the state-of-the-art non-clinical model (Wiesinger *et al* 2021) [DSC = 0.61 versus 0.66], but still achieved clinically useable performance despite a limited dataset. This is an important conclusion, given the need to train institution specific MRI-based models on small datasets due to sequence and scanner variability.

This work has important implications for developing a robust MRI autosegmentation model for brain OARs, by identifying how the training data should be defined and edited to enable segmentation for all brain OARs with acceptable quality, despite the lack of visibility of certain organs on specific image modalities. We found that editing directly on the T1w-MRI is necessary or beneficial in all cases, except lacrimal glands, which would require delineation on CT or the use of fat-saturated or T2-w MRI.

This study has certain limitations. The number of training cases was low due to the limited amount of available MRI data. However, editing the clinical contours before training the model enabled the DL-AC model to attain acceptable performance even with a small cohort. This model was also trained and tested using a single sequence, T1-w spin echo (SE) with gadolinium, as used locally. Thus, this model may not work well with similar

Table 1. Paired Student's t-test results comparing changes in DSC, MDA and sensitivity for all three pairs of MRI models. Bold values indicate statistically significant differences ($p \leqslant 0.005$). Insufficient successful segmentations were achieved by one of the models, this is noted ($$, **, or ##), indicating the superior model.

	Brainstem	Cochlea L	Cochlea R	Lacrimal L	Lacrimal R	Lens L	Lens R	Optic Chiasm	Optic Nrv L	Optic Nrv R	Orbit L	Orbit R	Pituitary
						DSC							
MRIeMRI versus MRIeCT (−means MRIeMRI performed better)													
p (*Threshold:* $\leqslant 0.005$)	0.074	**	**	$$	$$	**	**0.000**	**0.003**	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**
Effect size: Δ median	−0.010						−0.325	−0.200	−0.160	−0.240	−0.040	−0.060	−0.360
*N**	9	5	4	4	2	2	8	8	9	9	9	9	7
MRIeMRI versus MRIu (−means MRIeMRI performed better)													
p (*Threshold:* $\leqslant 0.005$)	0.397	**	**	##	##	**0.001**	**	0.068	**0.001**	**0.002**	**0.000**	**0.000**	0.009
Effect size: Δ median	0.010					−0.260		−0.185	−0.190	−0.260	−0.010	−0.050	−0.335
*N**	9	1	5	5	2	9	5	6	9	9	9	9	6
MRIeCT versus MRIu (−means MRIeCT performed better)													
p (*Threshold:* $\leqslant 0.005$)	0.430	$$	##	0.179	0.234	##	0.008	$$	0.321	0.638	0.035	0.622	$$
Effect size: Δ median	0.020			0.000	0.000		−0.115		−0.030	−0.020	0.030	0.010	
*N**	9	1	2	7	8	2	6	5	9	9	9	9	4
						MDA							
MRIeMRI versus MRIeCT (+means MRIeMRI performed better)													
p (*Threshold:* $\leqslant 0.005$)	0.042	**	**	$$	$$	**	0.173	0.031	**0.001**	0.006	**0.000**	**0.000**	**0.004**
Effect size: Δ median	0.020						0.080	0.100	0.080	0.080	0.040	0.040	0.130
*N**	9	5	4	4	2	2	8	8	9	9	9	9	7
MRIeMRI versus MRIu (+means MRIeMRI performed better)													
p (*Threshold:* $\leqslant 0.005$)	0.179	**	**	##	##	0.006	**	0.074	**0.002**	0.017	**0.000**	**0.000**	0.011
Effect size: Δ median	0.010					0.080		0.080	0.070	0.150	0.020	0.040	0.100
*N**	9	1	5	5	2	9	5	6	9	9	9	9	6
MRIeCT versus MRIu (+means MRIeCT performed better)													
p (*Threshold:* $\leqslant 0.005$)	0.325	$$	##	#	0.205	##	0.006	$$	0.222	0.686	0.086	0.282	$$
Effect size: Δ median	−0.010			0.140	0.060		0.055		−0.010	0.070	−0.020	0.000	
*N**	9	1	2	7	8	2	6	5	9	9	9	9	4
						Sensitivity							
MRIeMRI versus MRIeCT (−means MRIeMRI performed better)													
p (*Threshold:* $\leqslant 0.005$)	0.096	**	**	$$	$$	**	**0.001**	0.010	**0.000**	**0.000**	0.272	0.007	**0.001**
Effect size: Δ median	−0.010						−0.295	−0.145	−0.230	−0.320	−0.020	−0.060	−0.540
N*	9	5	4	4	2	2	8	8	9	9	9	9	7
MRIeMRI versus MRIu (−means MRIeMRI performed better)													
p (*Threshold:* $\leqslant 0.005$)	0.133	**	**	##	##	**0.001**	**	**0.005**	**0.000**	**0.000**	0.040	0.011	0.007
Effect size: Δ median	−0.020					−0.270		−0.340	−0.240	−0.350	0.000	0.020	−0.545
*N**	9	1	5	5	2	9	5	6	9	9	9	9	6

Table 1. (Continued.)

	Brainstem	Cochlea L	Cochlea R	Lacrimal L	Lacrimal R	Lens L	Lens R	Optic Chiasm	Optic Nrv L	Optic Nrv R	Orbit L	Orbit R	Pituitary
MRIeCT versus MRIu (−means MRIeCT performed better)													
p (*Threshold:* $\leqslant 0.005$)	0.609	$$	##	0.190	0.288	##	0.006	$$	0.462	0.520	0.010	**0.001**	$$
Effect size: Δ median	−0.010			0.000	0.000		−0.115		−0.010	−0.030	0.020	0.080	
N[*]	9	1	2	7	8	2	6	5	9	9	9	9	4

[*]Number of compared segmentations (successfully segmented by both models considered)

[**] MRIeMRI is better based on producing the segmentation for more cases.

$$ MRIeCT is better based on producing the segmentation for more cases.

MRIu is better based on producing the segmentation for more cases.

MDA unreliable due to insufficient overlap of OARs.

data from other institutions, due to lack of harmonisation between scanners. This study, on the other hand, is focussed on assessing the impact of standardising the clinical contours before training the model on its performance rather than in developing a general DL-AC model that can work with data from different institutions. We have shown the feasibility of training and using a CE-marked MR-based model clinically, with the limitations of deep-learning architecture and training dataset this implies.

Further research is needed to identify the impact of training data editing on radiotherapy dosimetry. The correlation between the geometric and dosimetric evaluation of contour quality is known to be complex and we intend to investigate this in future, to establish which geometric and dosimetric tests are necessary to determine the clinical usability of DL-AC models in brain OAR contouring.

5. Conclusion

The clinical delineation of brain OARs is typically performed manually and requires both CT and MRI scans. However, manual delineation is time-consuming and variable between operators. Developing a robust deep learning-based segmentation model is therefore essential. In this work, separate deep learning-based segmentation models for CT and MRI were developed and assessed. The T1-weighted gadolinium-enhanced MRI deep learning segmentation model was able to segment all brain OARs except for the lacrimal glands, which are difficult to see on T1w-MRI. CT scans are needed for the complete contouring of brain OARs if it is necessary to delineate lacrimal glands. These could be manually segmented on the CT scan or via a separate CT-based DL-AC model. A dual-modality autosegmentation model could also be developed to solve this problem. Editing MRI contours to be consistent with gold standard, before training models enhanced the geometric performance and reduced the number of failed segmentations, except for lacrimal glands. MRI-based deep-learning autosegmentation in RT may improve consistency, quality, and efficiency but requires careful editing of training contours on MRI.

Acknowledgments

We acknowledge the cooperation and support of RaySearch Laboratories AB. Also, we acknowledge N Alzahrani's sponsor, King Abdulaziz University, Jeddah, Saudi Arabia.

Dr L Murray is an Associate Professor funded by Yorkshire Cancer Research (award number L389LM).

Dr M Nix is funded by Cancer Research UK for the Leeds Radiotherapy Research Centre of Excellence (RadNet; C19942/A28832).

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary information files).

Ethical statement

Ethical approval for retrospective use of de-identified patient data was given by Leeds East REC, reference: 19/YH/0300, IRAS project ID: 255 585.

ORCID iDs

Ann Henry https://orcid.org/0000-0002-5379-6618
Anna Clark https://orcid.org/0000-0003-4359-3697
Louise Murray https://orcid.org/0000-0003-0658-6455
Michael Nix https://orcid.org/0000-0001-7228-7344

References

Brouwer C L, Boukerroui D, Oliveira J, Looney P, Steenbakkers R, Langendijk J A, Both S and Gooding M J 2020 Assessment of manual adjustment performed in clinical practice following deep learning contouring for head and neck organs at risk in radiotherapy *Phys. Imaging Radiat. Oncol.* **16** 54–60

Brunese l, Mercaldo F, Reginelli A and Santone A 2020 An ensemble learning approach for brain cancer detection exploiting radiomic features *Comput. Methods Programs Biomed.* **185** 105134

Cardenas C E, Yang J, Anderson B M, Court L E and Brock K B 2019 Advances in auto-segmentation *Semin. Radiat. Oncol.* **29** 185–97
Chen H *et al* 2019 A recursive ensemble organ segmentation (REOS) framework: application in brain radiotherapy *Phys. Med. Biol.* **64** 025015
Chen W, Zhang H, Zhang W, Su M, Xie R, Li K, Xia X and Zou C 2019 Development of a contouring guide for three different types of optic chiasm: a practical approach *J. Med. Imaging Radiat. Oncol.* **63** 657–64
Çiçek Ö, Abdulkadir A, Lienkamp S S, Brox T and Ronneberger O 2016 3D U-Net: learning dense volumetric segmentation from sparse annotation *Lecture Notes in Computer Science* **9901** pp 424–32 Int. Conf. on Medical Image Computing and Computer-assisted Intervention - Medical Image Computing and Computer-Assisted Intervention (MICCAI 2016) Springer
Eekers D B *et al* 2018 The EPTN consensus-based atlas for CT and MR-based contouring in neuro-oncology. *Radiother. Oncol.* **128** 37–43
Ho F, Tey J, Chia D, Soon Y Y, Tan C W, Bahiah S, Cheo T and Tham I W K 2018 Implementation of temporal lobe contouring protocol in head and neck cancer radiotherapy planning: a quality improvement project *Medicine (Baltimore)* **97** e12381
Jena R, Kirkby N F, Burton K E, Hoole A C, Tan L T and Burnet N G 2010 A novel algorithm for the morphometric assessment of radiotherapy treatment planning volumes *Br. J. Radiol.* **83** 44–51
Joseph V R 2022 Optimal ratio for data splitting *Stat. Analy. Data Min.: ASA Data Sci. J.* **15** 531–8
Kazemifar s, Mcguire S, Timmerman R, Wardak Z, Nguyen D, Park Y, Jiang S and Owrangi A 2019 MRI-only brain radiotherapy: assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach *Radiother. Oncol.* **136** 56–63
Lerner M, Medin J, Jamtheim Gustafsson C, Alkner S and Olsson L E 2021 Prospective clinical feasibility study for MRI-only brain radiotherapy *Front Oncol.* **11** 812643
Liu F, Yadav P, baschnagel A M and Mcmillan A B 2019 MR-based treatment planning in radiation therapy using a deep learning approach *J. Appl. Clin. Med. Phys.* **20** 105–14
Mayo C S *et al* 2018 American association of physicists in medicine task group 263: standardizing nomenclatures in radiation oncology *Int. J. Radiat. Oncol. Biol. Phys.* **100** 1057–66
Mir R, Kelly S M, Xiao Y, Moore A, Clark C H, Clementel E, Corning C, Ebert M, Hoskin P and Hurkmans C W 2020 Organ at risk delineation for radiation therapy clinical trials: global harmonization group consensus guidelines *Radiother. Oncol.* **150** 30–9
Mlynarski P, Delingette H, Alghamdi h, boNDIAU P Y and Ayache N 2020 Anatomically consistent CNN-based segmentation of organs-at-risk in cranial radiotherapy *J. Med. Imaging (Bellingham)* **7** 014502
Ranta I, Wright P, Suilamo S, Kemppainen R, Schubert G, Kapanen M and Keyriläinen J 2023 Clinical feasibility of a commercially available MRI-only method for radiotherapy treatment planning of the brain *J. Appl. Clin. Med. Phys.* e14044
van Rooij W, Dahele M, Brandao H R, Delaney A R, Slotman B J and Verbakel W F 2019 Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation *Int. J. Radiat. Oncol. Biol. Phys.* **104** 677–84
Schmidt M A and Payne G S 2015 Radiotherapy planning using MRI *Phys. Med. Biol.* **60** R323–61
Scoccianti S *et al* 2015 Organs at risk in the brain and their dose-constraints in adults and in children: a radiation oncologist's guide for delineation in everyday practice *Radiother. Oncol.* **114** 230–8
Simon J, Szumowski J, Totterman S, Kido d, Ekholm S, Wicks A and Plewes D 1988 Fat-suppression MR imaging of the orbit *AJNR Am. J. Neuroradiol.* **9** 961–8
Soomro T A, Zheng L, Afifi A J, Ali A, Soomro S, Yin M and Gao J 2023 Image segmentation for MR brain tumor detection using machine learning: a review *IEEE Rev. Biomed. Eng.* **16** 70–90
van Dijk L V, Van Den Bosch L, Aljabar P, Peressutti D, Both S, Steenbakkers R J H M, Langendijk J A, Gooding M J and Brouwer C L 2020 Improving automatic delineation for head and neck organs at risk by deep learning contouring *Radiother Oncol,* **142** 115–23
Wang Y, Zhao L, Wang M and Song Z 2019 Organ at risk segmentation in head and neck ct images using a two-stage segmentation framework based on 3D U-Net *IEEE Access* **7** 144591–602
Wiesinger F *et al* 2018 Zero TE-based pseudo-CT image conversion in the head and its application in PET/MR attenuation correction and MR-guided radiation therapy planning *Magn. Reson. Med.* **80** 1440–51
Rusko L *et al* 2021 Deep-learning-based segmentation of organs-at-risk in the head for MR-assisted radiation therapy planning *Proc. of the 14th Int. Joint Conf. on Biomedical Engineering Systems and Technologies* (*BIOIMAGING*) 2 31–43
Wong J, Fong A, Mcvicar N, Smith S, Giambattista J, Wells D, Kolbeck C, Giambattista J, GONDara l and alexander A 2020 Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning *Radiother. Oncol.* **144** 152–8