Assessing artificial intelligence MRI autocontouring in Raystation and the AutoConfidence uncertainty model for brain radiotherapy

Nouf Mobarak Alzahrani

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Medicine

April 2024

Intellectual Property and Publication Statements

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and the other authors to this work has been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

My supervisory team included: Ann Henry, Louise Murray, Michael Nix and Bashar Al-Qaisieh. Anna Clark is a medical physicist who provided technical support.

Chapter 2 is based on work from the jointly authored publication: Nouf Alzahrani, Ann Henry, Anna Clark, Louise Murray, Michael Nix and Bashar Al-Qaisieh. Geometric evaluations of CT and MRI based deep learning segmentation for brain OARs in radiotherapy. Phys Med Biol. 2023 Aug 29;68(17). Available from: https://doi.org/10.1088/1361-6560/acf023

Nouf Alzahrani was responsible for collecting, preparing, and analysing the data, model training and testing, interpreting the results, and writing the manuscript. Michael Nix and Louise Murray provided essential guidance for the study design, reviewing the analysis and interpretation of the results. Ann Henry and Bashar Al-Qaisieh contributed to reviewing and approving the study design from the clinical and technical perspectives and providing the overall guidance of the project and data sourcing. Anna Clark supported in extracting the data from the treatment planning system. All authors contributed to the review of the manuscript, and all approved the final draft for submission.

Chapter 3 is based on work from the jointly authored publication: Nouf Alzahrani, Ann Henry, Anna Clark, Bashar Al-Qaisieh, Louise Murray, and Michael Nix. Dosimetric Impact of Contour Editing on CT and MRI Deep-Learning Autosegmentation for Brain OARs. J Appl Clin Med Phys. Accepted: March 5, 2024.

Nouf Alzahrani was responsible for collecting, preparing, and analysing the data, model training and testing, interpreting the results, and writing the manuscript. Michael Nix and Louise Murray provided essential guidance for the study design,

reviewing the analysis and interpretation of the results. Ann Henry and Bashar Al-Qaisieh contributed to reviewing and approving the study design from the clinical and technical perspectives and providing the overall guidance of the project and data sourcing. Anna Clark supported in extracting the data from the treatment planning system. All authors contributed to the review of the manuscript, and all approved the final draft for submission.

Chapter 4 is based on work from the jointly authored publication: Nouf Alzahrani, Ann Henry, Bashar Al-Qaisieh, Louise Murray and Michael Nix. Automated Confidence Estimation in Deep Learning Auto-Segmentation for Brain Organs at Risk on MRI for Radiotherapy. Submitted to J Appl Clin Med Phys on Jan 18th, 2024, and under revision March 21, 2024.
Nouf Alzahrani was responsible for collecting, preparing, and analysing the data, model training and testing, interpreting the results, and writing the manuscript. Michael Nix developed AutoConfidence (ACo) model and provided essential guidance for the study design. Louise Murray and Michael Nix provided essential guidance for reviewing the analysis and interpretation of the result. Ann Henry and Bashar Al-Qaisieh contributed to reviewing and approving the study design from the clinical and technical perspectives and providing the overall guidance of the project and data. All authors contributed to the review of the manuscript, and all approved the final draft for submission.

Acknowledgements

I would like to express my gratitude and appreciation to everyone who supported me in my PhD study.

I would like to say thank you from the bottom of my heart to all my supervisors (Dr. Ann Henry, Dr. Bashar Al-Qaisieh, Dr. Louise Murray, and Dr. Michael Nix) for their invaluable advice, guidance, support, patience during my PhD study. I am lucky to have worked with a very professional, knowledgeable, respectful, and supportive supervisors.

Special thanks to Dr. Mike Nix for his assistance at every stage of my PhD study, encouragement, understanding, and overall insight into the field.

I am deeply grateful to Dr. Ann Henry and Dr. Bashar Al-Qaisieh for their unwavering leadership, management, and support. I can reach them anytime for any support and advice.

Also, I would like to extend my sincere thanks to Dr. Louise Murray for her always clinical related advice, support, quick response, suggestions and comments.

Thanks to the medical physics and engineering department members at Leeds Cancer Centre for supporting and facilitating my studies.

I acknowledge the cooperation and support of RaySearch Laboratories AB.

I gratefully acknowledge the full support I got from my country (Saudi Arabia) and my workplace (King Abdulaziz University) throughout my PhD study.

I would like to express my deepest appreciation to my husband (Bader) for his constant support, kindness, love, tremendous understanding, and encouragement! I am truly blessed to have him by my side. My appreciation for him is countless!

I cannot begin to express my thanks to my mom and dad, who have believed in me since my school days! They have always supported me in my education and in everything I do. Your support has pushed me to keep moving forward, and I will be forever grateful for that! I must also thank my sisters and brothers for being by my side and spending countless hours listening and supporting me without any hesitation.

I am deeply indebted to my son (Anmar), whose presence has been the driving force behind my hard work and every success in my studies. Thanks for waking me up early every morning to drop you off at the school and then immediately starting my work. I know you will keep me busy when you come back. Love you! I could not imagine getting through this tough time without your beautiful smile and the loving sparkle in your eyes.

Lastly, special thanks to my friend Lamyaa Aljaafari for her personal support, and encouragement during this stage of my study. Her support has been invaluable to me; thanks for being with me on this journey.

Abstract

Background

In radiotherapy, deep learning autosegmentation (DL-AS) and automation of quality assurance (QA) have the potential to efficiently standardize and enhance the quality of contours.

Aim

To assess the performance of DL-AS in delineating organs-at-risk (OARs) in brain RT using the RayStation Treatment Planning System. Secondly, to build and test a novel artificial intelligence QA model called AutoConfidence (ACo).

Methods

Retrospective MRI and CT cases were randomly selected for training and testing. DL-AS models were evaluated from geometric and dosimetric perspectives, focusing on the impact of pre-training editing.

The ACo model was evaluated using two sources of autosegmentation: internal autosegmentations (IAS) produced from the ACo generator and two external DL-AS with different qualities (high and low quality) produced from RayStation models.

Results

The edited DL-AS models generated more segmentations than the unedited models. Editing pituitary, orbits, optic nerves, lenses, and optic chiasm on MRI before training significantly improved at least one geometry metric.

MRI-based DL-AS performed worse than CT-based in delineating the lacrimal gland, whereas the CT-based performed worse in delineating the optic chiasm.

Except for the right orbit, when delineated using MRI models, the dosimetric statistical analysis revealed no superior model in terms of the dosimetric accuracy between the MR and CT DL-AS models. The number of patients where the clinical significance threshold was exceeded was higher for the optic chiasm D1% than for other OARs, for all models.

ACo had excellent performance on both internal and external segmentations across all OARs (except lenses). Mathews Correlation Coefficient was higher on IAS and low-quality external segmentations than high-quality ones.

Conclusion

MRI DL-AS in RT may improve consistency, quality, and efficiency but requires careful editing of training contours. ACo was a reliable predictor of uncertainty and errors on DL-AS, demonstrating its potential as an independent, reference-free QA tool.

Table of Contents

List of Figures

List of Tables

Abbreviations

AAPM TG-263	The American Association of Physicists in Medicine task group 263
AC	Autocontouring
ACo	AutoConfidence
AI	Artificial intelligence
AN	Artificial neuron
AS	Autosegmentation
BN	Batch norm
CBTRUS	Central Brain Tumour Registry of the United States
CDC	Centres for Disease Control
cGAN	Conditional Generative Adversarial Networks
CNN	Convolutional neural networks
CNS	Central nervous system
CT	Computed Tomography
CTeCT	CT edited autosegmentation model
CTu	CT unedited autosegmentation model
3D CRT	Three-Dimensional Conformal Radiation Therapy
D	Discriminator
DL	Deep learning
DL-AC	Deep learning autocontouring
DL-AS	Deep learning autosegmentation
DSC	Dice similarity coefficient
DVH	Dose-volume histogram
d2GS	Difference to gold standard
D(tp)_max	The maximum dimension of the TP region
[D(x_tp)]	The minimum distance to the TP region from the central voxel
EDT	Euclidean Distance Transform
EM	Ensemble-model
EM-LQ	External model with low quality
EM-HQ	External model with high quality
FC	Fully connected network
FDA	Food and Drug Administration
FMEA	Failure mode and effects analysis
FN	False Negative
FNR	False Negative Rate
FOV	Field of view
FP	False Positive
FPR	False Positive Rate
G	Generator
GDC	Geometric Distance Correction
GBM	Glioblastoma multiform
GPU	Graphics processing unit
GS	Gold standard segmentations

HD	Hausdorff distance
IAS	Internal autosegmentations
IER	Intelligent Edge removal
IMRT	Intensity-modulated radiation therapy
IRAS	Integrated Research Application System
IT	Information technology professionals
kVp	kilovoltage peak
LCC	Leeds Cancer Centre
MCC	Matthew's correlation coefficient
MCD	Monte-Carlo Dropout
MDA	Mean distance to agreement
MDR	Medical Device Regulations
MDT	Multi-disciplinary team
MRI	Magnetic resonance imaging
MRIeMRI	MRI edited MRI autosegmentation model
MRIeCT	MRI edited CT autosegmentation model
MRIu	MRI unedited autosegmentation model
NCI	National Cancer Institute
NHS	National Health Service
NN	Neural network
OAR	Organ-at-risk
pGS	Confidence prediction
PTV	Planning target volume
QA	Quality assurance
RCR	Royal College of Radiologists
REC	Research Ethics Committee
ReLU	Rectified linear unit
ResNet	Residual network
RO	Radiation oncologists
ROI	Region of interest
RT	Radiotherapy
RTT	Radiation therapist
sCT	Synthetic CT
SNGP	Spectral-normalized Neural Gaussian Process
SPM	Spatial probability maps
TE	Time to Echo
TN	True Negative
TP	True Positive
TR	Repetition Time
T1-w, T2-w MRI	T1-weighted, T2-weighted MRI
T1w-Gd MRI	T1-weighted gadolinium-enhanced MRI

Chapter 1 Introduction

1.1 Overall Introduction

This PhD thesis focuses on approaches to developing robust automation in radiotherapy (RT) treatment planning, through deep learning autosegmentation (DL-AS). Specifically, the focus is on brain organ at risk (OAR) segmentation, using magnetic resonance (MR) imaging. Through robust validation and commissioning (Chapter 2 and 3), alongside advanced quality assurance (QA) methods (Chapter 4), a complete approach (Chapter 5) to efficient and safe use of DL-AS tools in clinical practice is laid out.

1.1.1 Brain cancer statistics

Over 90% of central nervous system (CNS) cancer occurs in the brain; it ranks among one of the most lethal forms of cancer (Miller et al., 2021, Uwishema et al., 2023, Aldape et al., 2019, Raghavapudi et al., 2021). It significantly impacts overall morbidity and mortality rates, affecting individuals of various ages, genders, and ethnicities (Lu et al., 2020, Miller et al., 2021). Annually, over 5000 people die from all types of brain cancer, and currently, at least 102,000 adults and children in the UK are expected to be living with brain cancer (Brunese et al., 2020). Moreover, the number of new cases continues to rise, with more than 11,000 reported annually and an estimated rise in incidence of 16% over the past decades (Lu et al., 2020). Different subtypes of brain cancer have very different prognosis.

1.1.2 Glioblastoma

Glioblastoma is the most common malignant primary brain tumour in adult according to the Central Brain Tumour Registry of the United States (CBTRUS), Centres for Disease Control (CDC), and the National Cancer Institute (NCI) (Ostrom et al., 2020). It tends to occur more frequently in males than females, and its incidence increases with age (Ostrom et al., 2020, Miller et al., 2021). It is a very aggressive tumour with an extremely poor prognosis, with a median overall survival rate from the diagnosis of between 12 and 16 months for those who receive treatment (Angom et al., 2023, Lu et al., 2020). Currently, the gold standard treatment for patients with glioblastoma is debulking surgery, then radiotherapy (RT) in combination with chemotherapy and then followed by 6

months of adjuvant chemotherapy (Lu et al., 2020, Angom et al., 2023, Ainslie et al., 2024).

1.1.3 Brain radiotherapy and side effects

RT for glioblastoma delivers ionizing radiation to try to destroy cancer cells. Different radiotherapy-delivery techniques can be used such as three-dimensional conformal RT (3D CRT), intensity modulated RT (IMRT), volumetric-modulated arc therapy (VMAT), and proton therapy (Angom et al., 2023).

While radiotherapy aims to destroy cancer cells, may also cause damage to the normal healthy tissues located near the tumour, called organs at risk (OARs). Damaging normal tissues in the brain can lead to several significant side effects during or after the treatment. Side effects are categorized as acute, or late effects. Acute symptoms may include headache, tiredness, hair loss and nausea (Raghavapudi et al., 2021). Late symptoms may appear after 3 months to years after the treatment such as impairments in neurocognition and radiation necrosis (Angom et al., 2023, Raghavapudi et al., 2021). Depending on severity, they can be life threatening, or may contribute to a reduction in the patient's quality of life. However, they can be minimized through the careful design of the dose distribution during the radiotherapy treatment planning process. More conformal radiotherapy techniques, such as IMRT or VMAT can help reduce the radiotherapy doses delivered to the normal brain (Raghavapudi et al., 2021). This may translate into reduced acute and late side effects.

1.1.4 Radiotherapy image segmentation

Radiotherapy is planned to meet the clinical goals by directing the radiation dose to the tumour, while minimizing the radiation dose to the nearby OARs to keep the toxicity levels low and acceptable (Hansen et al., 2022). Contouring, or image segmentation, is one of the most important steps of treatment planning to enable the calculation of dose delivered to the OARs and achieve this goal (Liesbeth et al., 2020). Image segmentation refers to the process where a radiation oncologist and/or technical staff use the RT simulation scan(s) to contour the target volume of the tumour and surrounding OARs such as the brainstem, optic nerves, optic chiasm, orbits, lenses, cochlea, and lacrimal glands. Depending on the target or OAR in question, image segmentation can be performed based on computed tomography (CT), magnetic resonance imaging (MRI), or CT co-registered with

the MRI simulation scans to improve the segmentation accuracy (Hansen et al., 2022).

Regarding MRI image segmentation, T1-weighted (T1-w) and T2-weighted (T2-w) are two common imaging sequences essential for segmentation (Soomro et al., 2023). Each provides crucial information about brain anatomy and pathology, which is essential for accurate radiotherapy treatment planning.

T1-weighted image demonstrates the differences in tissues' longitudinal relaxation time, referring to the time it takes for the tissues to return to their equilibrium state after being perturbed by the external radiofrequency pulse (Westbrook, Roth and Talbot, 2011). The contrast in the T1-w MRI image is influenced by the differences in T1 values, as each tissue has a unique recovery rate. The contrast of the T1-w image can be adjusted using the repetition time (TR) and echo time (TE). A short TR (300-700 ms) and TE (10-30 ms) are typically used for T1-w image (Westbrook, Roth and Talbot, 2011). Using these settings results in fat appearing bright (hyperintense), while fluids appear dark (hypointense) (Soomro et al., 2023). Tissues with a higher fat content have a shorter T1 value, which results in great magnetisation and thus appears brighter, while tissues filled with fluid appear dark.
T1-w pre-and post-contrast imaging is the recommended protocol for brain tumour (Kamepalli et al., 2023). T1-w MRI image is particularly useful for showing brain anatomy, which will aid the staff to delineate the OARs. Using the contrast media in a T1-w MRI image will demonstrate where the blood-brain barrier is compromised (Kamepalli et al., 2023). In the case of brain tumour such as gliomas, the blood brain barrier is disrupted, allowing the contrast media to leak into brain tissue, producing high signals that highlight the tumour’s growth area. This aids in accurate treatment planning and tumour delineation.

On the other hand, T2-weighted image demonstrates the differences in tissues' transverse relaxation time, which reflect the decrease of transverse magnetisation after the radiofrequency pulse (Westbrook, Roth and Talbot, 2011). The contrast of T2-w image can also be adjusted using TR and TE. A long TR (2000 ms) and TE (70 ms) are typically used to acquire T2-w image (Westbrook, Roth and Talbot, 2011). These settings result in fluid appearing

bright (hyperintense) while fat appearing dark (Soomro et al., 2023). This makes T2-w imaging valuable to delineate tumour as it will highlight tumour infiltration and edema. Moreover, the T2-w images are useful to precisely delineate OARs that filled with water. Since fluid appears bright, pathology is more easily detected.

Both T1-w and T2-w imaging can be used for delineating brain tumour and OARs. Clinicians select the appropriate imaging sequence based on local treatment protocols and individual clinical situations.

1.1.4.1 Manual contouring challenges

Contouring is typically done manually by the clinician or technical staff drawing around the OARs on each slice of a CT and/or MRI image set. However, it is associated with several challenges. First, it is a time-consuming task; prior studies have revealed that each patient may require a clinician to spend several hours to contour all targets and OARs (Cardenas et al., 2019, Wang et al., 2019b). Moreover, in brain RT, planning MR images are combined or co-registered with CT making more load for operators, as they need more time to refine the registration, and perform delineation based on information from different modalities (Rong et al., 2023).

Secondly, despite the availability of consensus contouring guidelines, manual contouring remains a subjective task. An oncologist or a member of technical staff performs the delineation of OARs based on their previous experience, skill, and anatomical knowledge (Rong et al., 2023, Gibbons et al., 2023). Many teams now hold regular multi-disciplinary contouring meetings but it remains a source of error and uncertainty that may not be recognised during review and plan assessment (Gibbons et al., 2023, Rong et al., 2023, Cardenas et al., 2019), leading to inconsistent treatment between patients and potentially suboptimal care. This can occur either by missing areas of pathological tissue that should have been treated or treating areas of OARs that are, in fact, normal (Rong et al., 2023, Liesbeth et al., 2020, van Dijk et al., 2020).

1.1.4.2 Autosegmentation

Over recent years, advances in RT computing have enabled automated approaches to image segmentation, potentially addressing the challenges of

manual contouring by introducing autosegmentation using autosegmentation algorithms. These algorithms use some form of prior knowledge to estimate the location of OARs on the CT or MRI simulation scans and then contour them automatically (Gibbons et al., 2023). Several approaches of autosegmentation have become commercially available and clinically implemented (Cardenas et al., 2019, Harrison et al., 2022). Until recently, uptake has been limited but, more recently, deep-learning approaches have become more popular. These technologies could potentially make a notable improvement in RT practice by improving contour quality, consistency, and workflow efficiency (Cardenas et al., 2019, Gibbons et al., 2023), but only with robust and consistent approaches to their evaluation and implementation. Here, I review some of the predominant approaches to autosegmentation.

1.1.4.2.1 Atlas-based segmentation

The most common autosegmentation approach prior to the last 5 years was atlas-based segmentation (Brouwer et al., 2020a). It uses deformable registration to transfer reference contours from (one or more) atlas patients to the target image set (Rong et al., 2023, Gibbons et al., 2023). However, the performance of the method depends on the quality of the atlas(es), in addition to the accuracy of the registration algorithms (Gibbons et al., 2023, Rong et al., 2023). Thus, single or multi-atlas-based segmentation approaches often don't meet the quality threshold of clinical acceptability without extensive manual editing on the generated segmentation, which still needs to be performed by the radiation oncologist or technical staff (Gibbons et al., 2023). Hence, it has proven challenging to achieve efficiency or consistency benefits compared to manual contouring with this technology. Further details about this approach can be found in sections 1.2.1.1 and 1.2.1.2.

1.1.4.2.2 Deep learning-based segmentation

In the last decade, there has been a transition in the research domain from atlas-based segmentation to deep learning-based segmentation approaches (Harrison et al., 2022). Deep learning-based autosegmentation (DL-AS) has emerged in radiotherapy as a promising solution, addressing many of the limitations related to atlas-based segmentation (Gibbons et al., 2023). DL-AS is trained on many more patient cases than it is practical to use in an atlas-based approach and is

able to synthesise information from all of them, rather than having to select one or a few cases for image registration.

DL-AS typically uses convolutional neural networks (CNN) to identify and recognize complex features of medical images and the relationships between these features and the pre-drawn contours (labels) provided with the images (Gibbons et al., 2023, Cardenas et al., 2019, Brouwer et al., 2020a). Training the network to extract these features and relationships, such that it can reproduce the human delineated structures in the training dataset enables the DL-AS model to predict the location of the OARs and contour them on a new image based, on the previous information analysed in the training phase (Gibbons et al., 2023, Rong et al., 2023).

Several researchers have compared the performance of the DL-AS model and the atlas-based segmentation model in the delineation of OARs in various treatment sites (Gibbons et al., 2023, Cardenas et al., 2019, Liesbeth et al., 2020, van Dijk et al., 2020, van der Heyden et al., 2019). These studies have shown that DL-AS generally outperforms atlas-based segmentation methods.(Gibbons et al., 2023, Liesbeth et al., 2020, Cardenas et al., 2019, van Dijk et al., 2020). DL-AS models, which are trained on data prepared from consensus data from a group of human experts, have also been shown to be comparable to or outperform an individual human expert (Rong et al., 2023, Tang et al., 2019). These studies highlight the power of using DL-AS to improve segmentation quality, minimize interobserver variability, and significantly reduce the time and effort required by operators to segment OARs. However, it is consistently noted that some structures may still require editing and all structures require clinician review before use for treatment planning (Rong et al., 2023, Tang et al., 2019). Further details about this approach can be found in section 1.2.1.3.

1.1.4.2.2.1 Brain OARs CT DL-AS

In comparison to other treatment sites, there are few commercial DL-AS models available to delineate brain OARs. Only two studies exist that investigate the performance of commercial DL-AS for brain OARs segmentation based on CT scans (Heilemann et al., 2023, Wong et al., 2020).

The first study investigated the performance of the Limbus Contour commercial model for the brain , head and neck, and prostate in comparison to manual reference contours from multiple radiation oncologists to find out if the accuracy

of the DL-AS model is comparable to radiation oncologist interobserver variability (Wong et al., 2020). For the brain, autosegmentation of the brainstem, optic chiasm, optic nerve, and globes was evaluated using Dice similarity coefficient (DSC) and 95% Hausdorff distance (HD) (Wong et al., 2020). The researchers found that contouring time was reduced compared to the manual contour. Moreover, the output of DL-AS was comparable to radiation oncologists (RO) contour (Wong et al., 2020).

The second study compared the performance of three different commercial software solutions; Annotate, Limbus Contour, and RayStation to choose one of them for clinical implementation (Heilemann et al., 2023). Segmentation of the following treatments sites was investigated: abdomen, pelvis, thorax, head and neck and brain. For brain radiotherapy, these structures were evaluated: brain, brainstem, eyes, lens, optic nerve, optic chiasm, and cochlea (Heilemann et al., 2023). The DSC, HD distance, a dose/volume assessment, and a blind rating by physicians were used for evaluation. The researchers found that all three DL-AS tools performed generally well compared to the manual reference contour for eye, lens, optic nerve, and optic chiasm with some variation in performance was noted. Accordingly, the researchers found that the decision to deploy the DL-AS tool is not direct and dependents on the focus of treatment sites that will be autosegmented.

1.1.4.2.2.2 Brain OARs MRI DL-AS

Regarding MRI based DL-AS, only one study exists that investigates the performance of a commercial DL-AS software (MVision) to segment the following brain OARs: brain, amygdala L and R, brainstem, cerebellum, corpus callosum, lacrimal gland, hippocampus, medulla oblongata, midbrain, optic chiasm, optic nerve, optic tract, pituitary, pons and thalamus using T1-weighted (w)MRI scans (Turcas et al., 2023). There are, however, other studies that investigate non-commercial models to segment brain OARs using MRI (Mlynarski et al., 2020, Wiesinger et al., 2021, Chen et al., 2019, Mekki et al., 2024). The overall findings from these previous studies are that DL-AS performance is a promising tool, despite variation in the selection of evaluated brain OARs and evaluation methods. These tools can aid operators in their clinical workflow and reduce interobserver variability.

The delineation of brain OARs for RT treatment planning currently relies on images acquired from both CT and MRI scans. However, in recent years, MRI-only radiotherapy planning has been the subject of extensive research, where the workflow depends on using MRI scans alone (Lerner et al., 2021, Ranta et al., 2023). This requires the generation of a synthetic CT (sCT) from the MRI data, which is outside the scope of this thesis, but has been an active and successful research programme in its own right (Lerner et al., 2021). MRI-only radiotherapy planning offers several advantages. First, it increases the treatment accuracy by minimizing the uncertainties that arise from co-registering CT and MRI scans acquired from different patient positions during the scans (An et al., 2022, Ranta et al., 2023, Lerner et al., 2021, Kazemifar et al., 2019). Moreover, utilizing MRI only enhances the precision of treatment planning, as images from MRI are known to have high soft tissue contrast compared to CT, especially in brain tissues. This enables accurate delineation of the brain tumour and OARs (Liu et al., 2019). This is an important feature as the accuracy in delineation will likely improve treatment outcomes, by reducing uncertainty and allowing more precise RT planning. Compared to the current practice of using both modalities, MRI-only radiotherapy will potentially reduce error, improve workflow, and be more convenient for patients and department.

Despite the potential advantages of using MRI in brain OAR autosegmentation, challenges persist (Soomro et al., 2023). MRI is not a well-defined imaging modality, with a plethora of contrasts, sequences, and parameters available and little consensus on the precise definition of a given scan type (e.g. T1w or T2w). The variability in imaging contrast is exacerbated by variations between manufacturers, models, and even individual scanners. The performance of DL-AS is improved when trained using a large consistent dataset, but given the inconsistency of data from different sources, this can be very challenging to achieve with MRI. Effectively utilizing multi-centre MRI data is challenging, as it needs techniques to address the diversity of the MRI scanner acquisition and resulting image contrast from different sources (Fatania et al., 2022). Overcoming this challenge is an active area of research, and the absence of a simple solution currently limits the development of large, high-quality MRI based DL-AS models. If centre specific DL-AS models are to be used, a large dataset needs significant effort and time from operators to label and contour OARs, which is another reason

for limited availability of MRI training datasets. To overcome these limitations and build an effective and transferable MRI DL-AS model, collaboration is needed between clinical and computer scientists, which needs time and scientific efforts to meet the practical needs of the clinical setting.

Meanwhile, centres interesting in DL-AS in the MRI only setting are limited to relatively small training datasets. Accordingly, this PhD thesis aims to address the challenges related to MRI DL-AS for OARs with limited data.

1.1.5 Gaps in knowledge and aims

The main aim of this project is to train and evaluate the performance of an MRI DL-AS model using a limited training dataset, focusing on the impact of pre-training editing of clinical segmentations on model quality. This investigation addresses two challenges identified above: the transferability of a reliable MRI DL-AS model between centres, and the difficulty of obtaining large single centre labelled datasets for training DL-AS models. This evaluation focused on the delineation of 13 clinically important brain OARs: orbits (right and left), lenses (right and left), optic nerves (right and left), optic chiasm, lacrimal glands (right and left), pituitary, brainstem, and cochlea (right and left).

The second aim of this work was to establish the utility of various forms of evaluation metric for the task of clinical commissioning of an MRI DL-AS model. Most of the previous literature has investigated the performance of MRI DL-AS based on geometric accuracy alone. However, based on the published recommendations (Liesbeth et al., 2020) on how to effectively assess the clinical applicability of DL-AS tools, a dosimetric assessment is also needed, to determine the clinical suitability of a DL-AS model for radiation treatment planning and delivery. Applying both types of evaluation will enable an investigation of the correlation between geometric and dosimetric performance and will indicate whether both evaluations are necessary or if geometric evaluation alone is sufficient. These comprehensive evaluations align with the published guidelines on how to effectively evaluate the performance of the DL-AS models and robustly evaluate MRI DL-AS for radiotherapy treatment planning in the brain.

Using DL-AS models clinically remains challenging, due to the uncertainty associated with predictions on unseen data, which are exacerbated in this case by the limited MRI data available for training and evaluating models. As previously mentioned, DL-AS models could produce errors when applied to a new dataset that is different from the training data (Claessens et al., 2022). Also, human interaction with the generated contour can be a cause for concern, due to human biases and subjectivity. Some operators may perform extensive editing on clinically acceptable generated contours, which is time-consuming, while others may not do enough editing, leaving clinically unacceptable errors in segmentations. The variability of human editing will also reintroduce interobserver variability, which DL-AS promised to eliminate.

Crucially, the errors produced from DL-AS models are different from those arising from manual contouring, potentially leading to mistakes during segmentation review. This tendency results from human biases as to the expected distribution of errors, which is based on clinical experience of other human derived contours or atlas-based AS. For example, a typical human segmentation failure mode is 'missing slices' in the middle of an OAR, which DL-AS will typically not do. Conversely, an experienced human operator is unlikely to include an anatomically incorrect structure, if it is superficially similar to the correct one, whereas DL-AS may well do so.

DL-AS models do not usually provide enough information about their uncertainty for each new case, and even if they do, this can often be poorly calibrated (i.e. high confidence is reported even for incorrect regions of OARs).
Consequently, this project also aims to build a novel artificial intelligence QA (AI-QA) model called AutoConfidence (ACo) to independently assess the segmentation uncertainty without ground-truth on a per-patient basis. Having a QA tool will improve clinical confidence in DL-AS, enhance the confidence and trust of the clinician to use the DL-AS tool, and aid users in deciding which area of the segmentation needs human review for potential editing. In turn this should reduce human editing variability and enhance efficiency savings, whilst simultaneously reducing the risk of clinically significant segmentation errors being missed in review.

1.1.6 Clinical implications

The outcomes of these investigations will have significant clinical implications. A robust DL-AS tool will potentially streamline the clinical workflow, reducing segmentation time and improving segmentation quality. By integrating an independent AI-QA tool (ACo) with DL-AS, clinicians can have more confidence in using the DL-AS tool, reduce review time for the generated segmentation, and avoid missing errors during the review process. Ultimately, having both DL-AS and ACo tools working together will enhance the safety, efficiency, and efficacy of the treatment planning process in the clinical setting, helping to realise the long-held promise of AI in RT.

1.2 Literature Review

1.2.1 Automatic segmentation

1.2.1.1 Single atlas-based autosegmentation

Single atlas-based segmentation is defined as using an expert segmented reference image, referred to as an atlas, to perform a new segmentation task through image registration (Schipaanboord et al., 2019) . To map the atlas contour to the input image, both the reference image and input image should be aligned to the same coordinate space, and deformations between the image sets computed. Then, the label information from the atlas image can be transferred to the input image (Schipaanboord et al., 2019).

The performance of single atlas-based segmentation is determined by the quality of deformable registration, which is affected by the differences in patient anatomy between the reference image and the new input image. As a result, single atlas-based models aren’t very accurate for most patients, becoming worse as the patient anatomy and positioning diverges from the atlas reference. It is worth noting that tumours and surgery can significantly change patient anatomy, making a single-atlas approach essentially unusable in oncology. However, segmentation accuracy may be improved somewhat by using a reference image that represents average patient anatomy (Harrison et al., 2022, Cardenas et al., 2019).

1.2.1.2 Multi-atlas-based autosegmentation

In response to the challenge of the single atlas-based segmentation approach, multi-atlas-based Autosegmentation was introduced, which uses several reference images to improve segmentation accuracy. Often, a similarity metric will be used to determine the ‘most similar’ patients in the multi-patient atlases, which can then be selected for registration to the current case.

The same process as the single atlas-based autosegmentation is applied (Schipaanboord et al., 2019). However, a further step, known as contour fusion, combines segmentations from several reference images to generate a segmentation that represents the best estimate of accurate contouring (Cardenas et al., 2019, Schipaanboord et al., 2019). The multi-atlas-based Autosegmentation approach is commercially available as a clinical tool (Cardenas et al., 2019, Brouwer et al., 2020a), and it has been validated in different clinical sites (Cardenas et al., 2019).

Despite the clinical success of multi-atlas-based autosegmentation, the main weakness is the limited ability to generalize to all patients. Patient variability can result from natural morphological differences, pathological reasons such as previous surgery, or healthy organs may be deformed because of growing tumours. These variations may not be represented in atlases, so generated contours may be inaccurate (Harrison et al., 2022, Mlynarski et al., 2020). Several studies show that multi-atlas-based autosegmentation has failed to meet the threshold of clinical acceptability, as radiation oncologists or technical staff must still perform extensive manual editing (Gibbons et al., 2023, Wang et al., 2019b), reducing the potential benefit and introducing the risk of incorrect OAR segmentations being used in RT planning.

1.2.1.3 Deep learning-based autosegmentation

Artificial intelligence (AI) is now extensively utilized in many fields, including medicine. It is defined as a computer system's ability to perform tasks that typically need human thinking using collections of complex computing and statistical algorithms (Liesbeth et al., 2020, Oh et al., 2019). AI research and applications have grown rapidly, especially in RT, aiming to improve treatment quality and save time. AI applications in RT include contouring, treatment planning, synthetic CT generation, and machine quality assurance (QA).

In 2020, a survey was conducted by (Brouwer et al., 2020b) among medical physicists in RT about the current clinical use of machine learning in the RT department. This revealed that 37% predominantly used AI for contouring (Brouwer et al., 2020b). This survey showed that most respondents (>90%) expected to introduce machine learning-based contouring and planning into their clinic over the next five years. Accordingly, further research is needed before clinically implementing DL-AS to establish the benefits and accuracy of autocontouring, address the challenges in implementation, and determine how to independently assess contour accuracy.

Deep learning is a subtype of machine learning technology, that employs deep networks of layers of 'artificial neurons', to efficiently process large amounts of data and extract the most important features to perform specific tasks (Du et al., 2020).

Modern deep learning methods entered common use in the 2010s when hardware improved, particularly following the use of graphical processing units (GPUs) for massive parallelisation of linear algebra (which is the core mathematical task for deep learning). The development of efficient gradient computation methods, especially 'backpropagation', enabled use of complex and deep architectures comprising many millions of parameters. These deep neural-networks began to demonstrate significant abilities in image processing tasks, especially for medical image classification and segmentation (Hesamian et al., 2019). The architecture of the deep neural network may change based on its purpose (Bibault and Giraud, 2024), with fully connected networks being common for structured (non-imaging) data problems and convolutional neural networks for imaging tasks.

1.2.1.3.1 Fully Connected networks

The simplest architecture for a neural network is that of fully connected (FC) layers (Bishop, 1995). In this architecture, the input to the layer is represented as a tensor of numerical values. Each of these is 'connected' to a value in an output tensor (which can have any shape and size) by an 'artificial neuron' (AN) (Bishop, 1995). These ANs consist of a linear function of the form:

$$y = wx + b$$

Where y is the output value, x is the input, w is the ‘weight’ and b is the ‘bias’. The output value y is then put through a second function, known as an ‘activation function’. This is a non-linear function that allows the network to learn non-linear relationships between input and output variables (Bishop, 1995). It is typically a simple form such as a sigmoid or tanh function with no parameters (figure 1.1). The network learns by optimising the weights and biases to produce the desired output from each input in the training dataset.

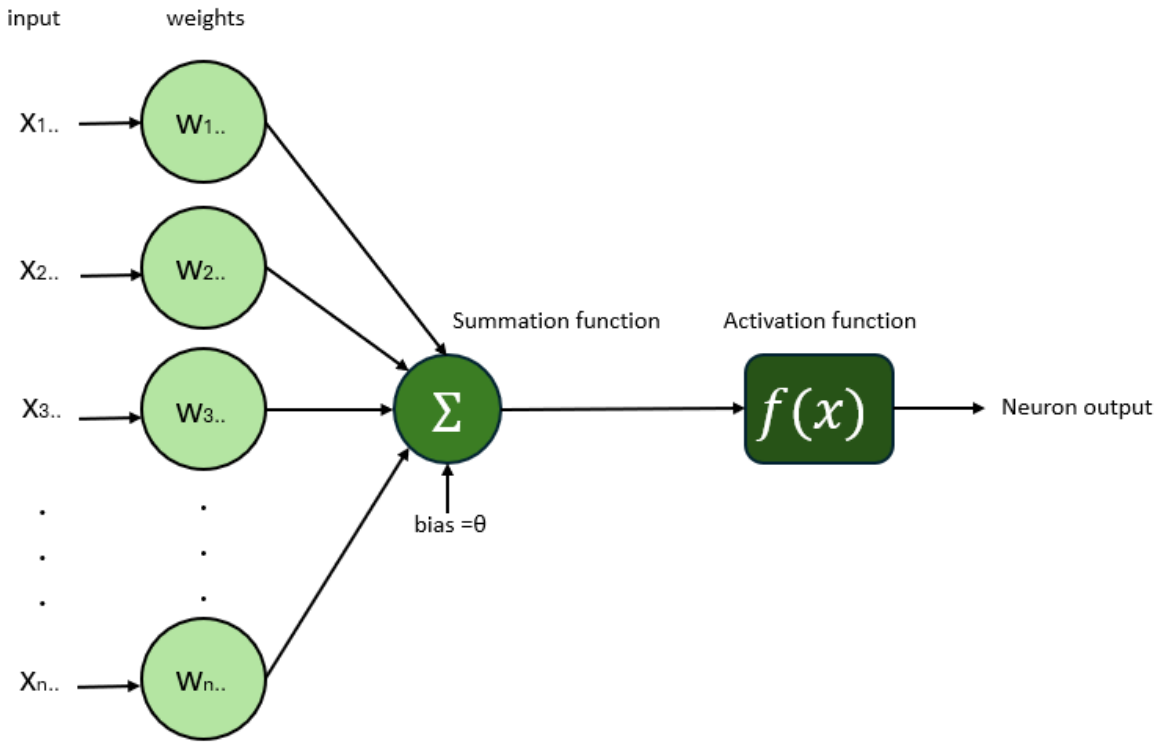


Figure 1.1:Artificial neuron with activation function. The input tensor of numerical values (X1-Xn) and their corresponding weights (W1-Wn) are summed, and a bias (b) is added. Then, the activation function f is applied to the outputs.

In an FC network, every input value is connected by an AN to every output value (figure 1.2). The weight and biases are adjusted iteratively, as the model is trained against the known ‘ground truth’ predictions, or labels (Bishop, 1995). In order to do this, a loss-function is used to compute the difference between the expected and actual prediction for each example in the training dataset. By differentiating this ‘loss’ with respect to the weights and biases, using an algorithm called back-projection, the gradient, or sensitivity of each parameter can be computed

(Bishop, 1995). This allows an optimiser to adjust the parameters in such a way as to minimise the loss and improve the network predictions.

The limitation of FC networks is one of scaling. For an input image of size 256*256 pixels (values) and an output size of 128*128 pixels, a single FC layer would correspond to approximately $1x10^9$ parameters, which is clearly prohibitive.

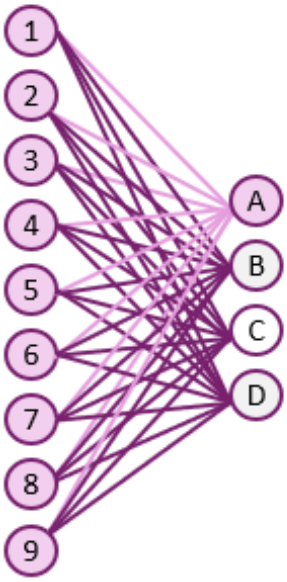


Figure 1.2: Architecture of fully connected (FC) layers. Each value in the input tensor of numerical values is 'connected' to a value in the output tensor by an artificial neuron

1.2.1.3.2 Convolutional neural networks

Due to the constraints of FC layers, convolutional layers were developed. Here, each output value is connected, with a small array (kernel) of ANs, to the input layer. This kernel is typically a square array of 3*3 or 4*4 ANs, resulting in 18 or 32 parameters. The kernel is moved over the input array in 2D, but the parameters remain the same, massively reducing the amount of computation needed to optimise their values, and making image based deep-learning feasible (figure 1.3).

The Convolutional neural network (CNN) algorithm is the most common deep learning algorithm for imaging tasks that has abilities to perform imaging tasks, such as image recognition, classification, segmentation and synthesis (Bibault and Giraud, 2024, Yang and Yu, 2021).

Each layer takes the output of the last layer as its input (Hesamian et al., 2019). The input for the first layer of the CNN is the (normalised) pixel values of a medical image.

By stacking many convolutional layers, each with an internal set of kernels (also known as filters), producing a multi-channel output, deep-CNNs can encode features of images at many scales (resolutions). As well as convolving the kernel across the image, and applying a non-linear activation function, each layer also down-samples the image size, typically by a factor of 2, using a pooling layer. (Hesamian et al., 2019). The dimension of the activation map is typically reduced by a pooling layer.

The output for each convolutional layer will create an activation map, or encoded representation of the image information. The first layers encode local features, such as edges and shapes, whereas the deeper layers are able to synthesis these simple features into more complex and contextual information. Based on each pixel or voxel value, and the surrounding voxels, the CNN can be trained to predict some information about the image. This may be a simple binary classification (e.g. contains tumour vs. does not contain tumour) or a complex set of information.

For image classification, the final layer is an FC layer, which extracts high-level abstractions as it has full connections to every activation unit in the previous layers (Hesamian et al., 2019).

By taking the output of the deepest layer of the CNN and using it as the input to an ‘inverted CNN’ it is possible to generate information in the form of a new image, predicted on the basis of the input image. This architecture is known as an ‘encoder-decoder’ and is the core of most modern approaches to image based deep-learning. The output of the encoder is known as the ‘latent space’ and is the input to the decoder.

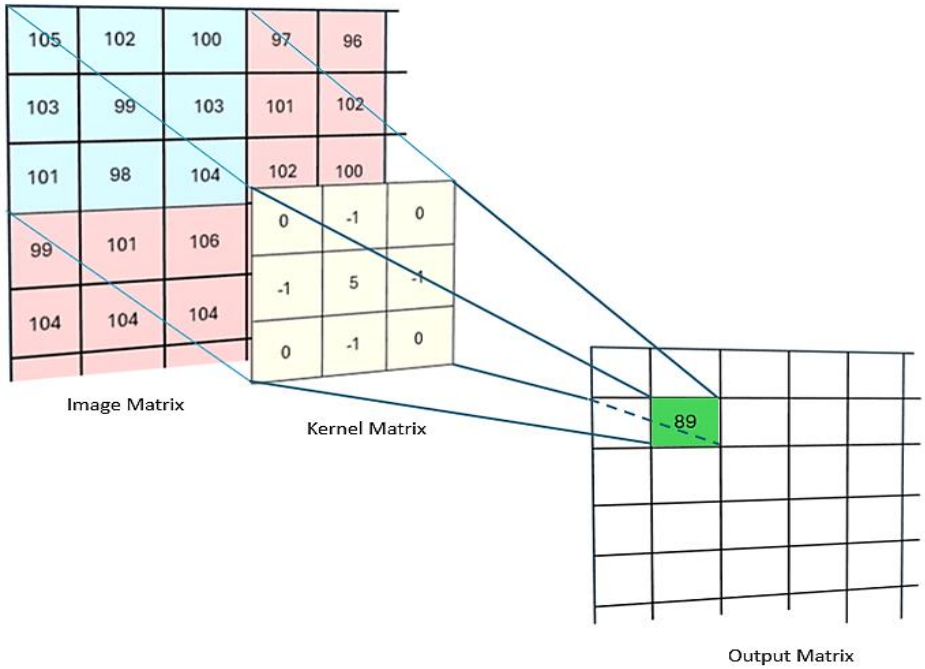


Figure 1.3: Illustration of the Convolutional Neural Network (CNN). The input image is processed by different layers of CNN using filters of artificial neuron to detect image features. This is followed by pooling layers to reduce the dimension of feature map and activation function to learn complex pattern in the image.

An encoder-decoder CNN, trained for image segmentation, can classify each voxel as belonging to either an OARs, a target, or the background, leading to a mask of the segmentation for each region of interest (ROI), at the input resolution of the images (Liesbeth et al., 2020).

Fully-convolutional encoder-decoder CNNs (Long et al., 2015) have two paths, firstly the encoding path, which is as the same as regular CNNs, and secondly, the decoding path, which removes the last fully connected layer to perform a learned upsampling via transposed convolutions to generate accurate segmentations (Cardenas et al., 2019). There has been a growing use of Encoder-decoder CNNs in radiotherapy since 2017 (Bibault and Giraud, 2024).

1.2.1.3.2.1 U-Net

There are different types of CNN architecture, and U-net is the most widely used CNN architecture for medical image delineation (Cardenas et al., 2019). The U-net architecture was developed in 2015 (figure 1.4) (Ronneberger et al., 2015). It consists of two paths: the encoding path (the same as the structure of CNN) and the decoding path, which is called the expansion path, involving an up-sampling

and deconvolution layers (Hesamian et al., 2019). One key limitation of CNNs for image segmentation is the fact that the latent representation is compressed - it contains less information than the original image. This means that reconstructed images and segmentations from the decoder are typically quite blurry, or uncertain at the edges of the masks, in the case of segmentation.

The most powerful and innovative feature of the U-net approach is using skip-connections, which combine features from the encoding contraction path directly to the decoding expansion path, bypassing the latent space. This approach helps to improve localization, sharpness and accuracy, when learning representations from input images (Cardenas et al., 2019).The U-net paper also demonstrated that with fewer labelled training data, which is a common situation in medical imaging, the networks may still be trained to generate acceptable segmentations.

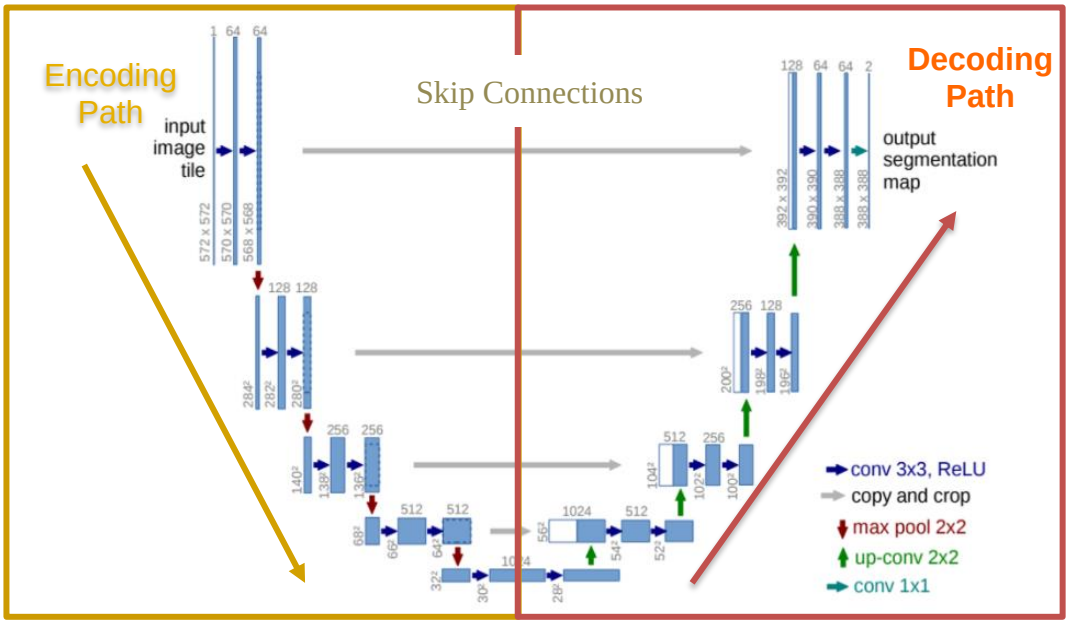


Figure 1.4: A modified diagram for the U-net architecture from figure 1 (Ronneberger et al., 2015). It consists of the encoding path, the decoding path and skip-connections.

In 2016, there was an update to the original 2D U-net method to use 3D images to train the network (Çiçek et al., 2016). The impact of using the 2D U-net architecture to segment 3D medical images, was that images slices were processed individually, losing information stored in the stack direction. This could lead to data from the original 3D medical image not being used in the segmentations, leading to slice-to-slice variation in the predicted ROIs. The 3D U-net allowed for direct analysis of the 3D images (Du et al., 2020), using 3D

convolutional kernels (e.g. 3*3*3). However, computational limitations on handling full 3D datasets in memory have limited the depth of 3D CNNs, leading to a performance trade-off. Other researchers have created modifications of the U-net structures. They developed a three-dimensional version of the U-net, called the V-net (Milletari et al., 2016). The resulting segmentation of this approach was fast and accurate (Milletari et al., 2016).

To summarize, autosegmentation based on deep learning (DL-AS) has been successfully applied in the medical research field using different strategies in which a neural network can automatically learn features from the medical image, which predict the segmentation. The U-net network has demonstrated outstanding performance across many different segmentation tasks for medical images (Bibault and Giraud, 2024, Cardenas et al., 2019, Hesamian et al., 2019). DL-AS demonstrated much improved performance compared to the atlas-based segmentation methods (Cardenas et al., 2019, Liesbeth et al., 2020, van Dijk et al., 2020, Gibbons et al., 2023).

1.2.1.3.3 Deep learning-based autosegmentation challenges

Despite the potential of deep learning automatic segmentation, it has some limitations. Deep learning autosegmentation typically generates accurate segmentations for large structures, but accuracy for small structures is typically sacrificed (Wang et al., 2019b). This can be due to the loss function, which is scaled by the number of voxels in the structure. However, some techniques can improve the accuracy of the generated contour for small structures by separating segmentation between the small and large structures using a two-stage framework to localize and segment OARs (Wang et al., 2019b). Alternatively, more heavily weighting the importance of the small structures can help with this problem.

Deep learning in general has low interpretability. There are problems understanding how image features such as image intensity or anatomical structures can affect the trained network during the process of segmentation prediction. This makes it difficult to understand and determine the source of incorrect segmentations (Cardenas et al., 2019).

Differences in patient anatomy or positioning between training and clinical cases can affect the model output, as can using different immobilization devices (Claessens et al., 2022, Brouwer et al., 2020a). Using various image acquisition

protocols can also negatively affect the performance of a deep learning autosegmentation (Claessens et al., 2022). These differences between the individual patient and training cases could cause geometric segmentation errors varying from minor issues, such as misplaced boundaries, to significant issues such as missing segmenting of the entire slice or structure (Claessens et al., 2022).

Moreover, there are different challenges during the training of the DL-AS model. First, the consistency and accuracy of the manual segmentations used to train the model significantly affect the model output (Cardenas et al., 2019). However, standardizing manual contouring through the implementation of international consensus guidelines can overcome this limitation to some extent (Cardenas et al., 2019). Training deep learning models needs large datasets to expose the model to sufficient different examples, which increases the accuracy and reliability of the model (Thwaites et al., 2021). This is still challenging for many reasons, such as a shortage of high-quality labelled data, and privacy concerns from sharing data among different institutions.

Having limited graphics processing unit (GPU) memory can be an obstacle to using 3D deep neural networks, as training the network needs intensive computational resources (Bibault and Giraud, 2024).

Lastly, clinicians and technical staff find it hard to have appropriate confidence in using deep learning tools such as autosegmentation clinically (Harrison et al., 2022). There is a risk of both under- and over-confidence, which can vary from patient to patient. This risks both allowing errors to go undetected, and editing contours unnecessarily, which its time consuming and can re-introduce operator inconsistency. Accordingly, independent AI-QA tools are needed to detect uncertainty in the segmentations. This would support clinicians in gaining confidence to use DL-AS tools and accelerate their review process of the generated segmentations.

1.2.2 Commissioning, clinical implementation, and quality assurance

Any new medical software or device intended for use in RT clinical practice should be quality assured with an initial commissioning phase followed by clinical implementation and ongoing routine QA. It is a critical QA task to verify the performance of the deep learning autosegmentation techniques that will be clinically used in the RT department. Any inaccuracy could lead to incidents

where unintended radiation dose is delivered to healthy tissues or inadequate target volume coverage occurs. Previous studies developed general recommendations on commissioning, clinical implementation, and quality assurance of DL-AS models (Liesbeth et al., 2020, Claessens et al., 2022, Bibault and Giraud, 2024). They recommend good practices at the development, implementation, and clinical use stages. These recommendations should be followed to ensure the model's output and predictions are and remain accurate.

1.2.2.1 The development phase

The commissioning phase includes the training phase followed by the testing phase (Bibault and Giraud, 2024, Liesbeth et al., 2020).

1.2.2.1.1 Training phase

During the training phase, the neural network learns to perform an accurate segmentation by finding the correlation between the image features and the relevant segmentation labels. The loss function will calculate the differences between the generated and gold standard segmentation (Bibault and Giraud, 2024). Then, the networks weights will be adjusted during the training based on the outputs of the loss function to reduce the discrepancy between the generated and gold standard segmentation (Bibault and Giraud, 2024).

Model training can start in different ways, such as initializing all weights to zero or using pre-trained weights from a previous training episode. When all weights are initialized to zero, the training of neural networks begins from scratch. Conversely, using pre-trained weights from previous models may help the model to work faster.

1.2.2.1.1.1.1 Training data criteria

A considerable amount of high-quality, labelled data needs to be used to train DL-AS models (El Naqa et al., 2018). The quality of the data includes the quality of the medical images themselves (resolution, noise level, artefact) and manual segmentations, which are key factors that can affect the model’s output. Thus, all the clinical data should be reviewed before being used for training (Cardenas et al., 2019). Moreover, it is preferable to use locally clinical acquired data to follow the imaging methods, equipment and clinical guidelines of the department (Liesbeth et al., 2020). This is particularly important for MRI, which is much less standardised than CT. Using clinical data from other institutions with different

imaging or contouring guidelines could significantly affect the model output when tested on a different dataset (Cardenas et al., 2019). The training data should also include many variations in patient anatomy to reflect the population variability of the clinical data (Liesbeth et al., 2020, Thwaites et al., 2021).

Based on previous studies, the number of scans required to train the model still has to be determined (Cardenas et al., 2019), and likely is dependent on the clinical site and amount of anatomical variation present. Currently, state-of-the-art CNN-based segmentation models typically have over 100 patients (van Dijk et al., 2020). On the other hand, models of 50–100 patients have been proven to segment OARs with sufficient precision (Van der Veen et al., 2019) in some cases. A survey was done among medical physicists in radiotherapy to investigate the current utilization and requirements to support deep learning implementation (Brouwer et al., 2020b). 114 of 147 participants responded to questions about the training and preparing data to use AI tools (Brouwer et al., 2020b). Most of them noted that they use their own data instead of vendor data with a sample size of < 100 patients. Except for 10% of cases, where they used > 200 cases. Most of the participants stated that they reviewed the training cases before training the model. Some respondents mentioned that they excluded unusual patients or setup positions from the training data. So, most physicists and centres have their own training data practices.

1.2.2.1.2 Validation phase

Validation aims to develop an independent assessment of the model's final performance and determine which types of patients the model may be used for (Liesbeth et al., 2020, Thwaites et al., 2021). An independent dataset is needed to assess the model quantitatively and qualitatively (Liesbeth et al., 2020).This dataset should reflect the variation of the clinical data that the model will be used for (Liesbeth et al., 2020). The average number of patients in a test phase is about 20 patients (Willems et al., 2018), or 20% of the training dataset size (Joseph, 2022) . A minimum of ten cases is recommended. However, this number should be increased if the results are highly variable (Cardenas et al., 2019). The autosegmentation for the test cases should be compared with the reference contours (gold standard contour) (Harrison et al., 2022), to ensure clinically acceptable performance.

Various metrics are used to evaluate the reproducibility and accuracy of the output of the model compared to the gold standard contours (Harrison et al., 2022).

1.2.2.1.2.1 Evaluation metrics

Quantitative geometric agreement between generated segmentation and the reference contour (Vaassen et al., 2020, Liesbeth et al., 2020, Harrison et al., 2022) can be measured using volume-, surface-, and moment-based metrics.

Volume-based metrics includes Dice Similarity Coefficient (DSC) and sensitivity. DSC measures the overlap volume between the generated and gold standard contour.

$$DSC\ (A, B) = 2(A \cap B)/(A+B),$$

where ∩ is the area of overlap between two contours. A is the generated segmentation, and B is the gold standard segmentation (Sherer et al., 2021). The range of DSC is from 0 to 1. Zero denotes no overlap between the two contours, while 1 denotes complete overlap (Bibault and Giraud, 2024). Sensitivity measures the ability of the autosegmentation model to correctly predict the pixels located within the OAR gold standard contour (van Rooij et al., 2019).

$$Sensitivity = TP / (TP + FN)$$

where TP is the true positive (presenting pixels in both gold standard contour and autosegmentation) and FN is the false negative (presenting pixels inside the gold standard contour but outside the autosegmentation). If the sensitivity is near 1, this indicates the model correctly predicted most of the pixels within the OAR gold standard contour, while a result near 0 indicates the model is missing most of the pixels within the OAR gold standard contour.

Surface-based metrics include the median, 95% and maximal Hausdorff Distances (HD) between surfaces, and mean distance to agreement (MDA) (Vaassen et al., 2020, Liesbeth et al., 2020, Harrison et al., 2022). HD is based on the histogram of distance differences between two contour surfaces (Mackay et al., 2023). Mean distance to agreement is the mean distance required to make the outline points of the automatic segmentation and outline points of the gold standard contour match perfectly. Small values of HD and MDA indicate both contours are close to each other, while greater values indicate dissimilarity between the two contours (Mackay et al., 2023).

Moment-based metrics compare the centre of the generated contour to the centre of the gold standard contour of a specific structure in x, y, and z dimensions (Mackay et al., 2023).

However, these methods (volume-, surface- and moment-based metrics) do not necessarily correlate with clinical (dosimetric) applicability or the time needed to edit DL-AS contours to an acceptable standard.

Thus, there are other clinically significant metrics to consider, such as the time required for editing (Liesbeth et al., 2020, Vaassen et al., 2020, Harrison et al., 2022, Sherer et al., 2021) and dose variation between the generated and gold standard contours of the OAR (van Rooij et al., 2019, van Dijk et al., 2020, Harrison et al., 2022). However, dosimetric analysis can be extremely challenging, as it depends on both the underlying geometric contour error and the position of that error relative to the high dose gradients of the patient plan. Hence, dosimetric results can be very individual and hard to generalise to future cases.

Furthermore, there are qualitative analysis methods using subjective judgment (qualitative scoring systems) to assess the quality of the generated segmentation in each patient, compared to the gold standard contour by one or more radiation oncologists (RO) (Sherer et al., 2021, Liesbeth et al., 2020). In cases of inaccurate segmentation, specific limitations should be identified and documented. These can then be used to generate restrictions or advice for the use of DL-AS in the clinic.

Once all the generated segmentations are evaluated, and any error recognised and documented, the commissioning phase is completed. The autosegmentation model can be then used in routine clinical practice if the evaluation scores confirmed its applicability for clinical use (Cardenas et al., 2019). If not, the commissioning phase should be repeated until the performance of the model improved by providing clinically acceptable segmentation.

1.2.2.2 DL-AS Implementation phase

Even following validation of the DL-AS model, which is a critical step of the commissioning and implementation process, it is still unclear how to most efficiently implement AI-based technology into clinical practice (Brouwer et al., 2020b, Rong et al., 2023). However, recent recommendations exist for using AI technology in the clinical setting. First, it is highly recommended to create a multi-

disciplinary team (MDT) of the clinicians, medical physicists, RTT, and information technology professionals (IT) involved in the clinical implementation (Liesbeth et al., 2020). This MDT should meet regularly to assess the strengths and limitations of the models, the way in which they will be integrated into the workflows, the responsibilities of each user group, how to educate users model usage, and how to check and edit the output. Meeting regularly is key to detecting any potential problems at an early stage. Moreover, it is recommended to conduct a feedback system for maintaining clinical practice quality and safety over time.

If any changes are made to models, either to address issues uncovered or as part of an external provider's update programme, the validation phase of the commissioning process should be repeated. Retraining the model with a new training dataset may be required in cases of a change in the clinical workflow or a systematic drop in model performance (Kalet et al., 2020). In these cases, the model must be recommissioned with a new test dataset, which matches the new clinical scenario (Liesbeth et al., 2020).

The current UK Medical Device Regulations (MDR) emphasise the importance of assessing and control risks, and if followed appropriately, should ensure the safety of using the model in clinical practice (Brouwer et al., 2020b, Thwaites et al., 2021). Importantly, it is the entire system which should be risk assessed, including workflow, checking processes and human-system integrated performance. It is insufficient to depend solely on technical model performance assessment.

Accordingly, conducting a prospective risk analysis assessment is essential to find and assess factors that may affect the model output, such as using failure mode and effects analysis (FMEA), the most commonly employed risk assessment for machine learning applications (Brouwer et al., 2020b).

1.2.2.2.1 DL-AS implementation challenges

Integrating the DL-AS into clinical practice is still challenging, as highlighted by previous literature and surveys addressing the barriers of using these tools in radiotherapy.

1.2.2.2.1.1 A lack of information and resources

The first challenge is related to a lack of information and resources on how best to deploy DL-AS tools into clinical practice (Mackay et al., 2023). Clear guidelines

should be developed in collaboration with vendors, healthcare institutions, and researchers to support the effective integration of DL-AS into clinical practice and ongoing quality assurance. There are ongoing projects to deliver these guidelines through the Royal College of Radiologists (RCR) and other professional bodies.

1.2.2.2.1.2 A need for training and education

There is a need for training and educating the operator to know the purpose of implementing this tool, how to use it, the clinically relevant outcomes generated by the model, possible limitations, and proper documentation practice (Brouwer et al., 2020b, Thwaites et al., 2021). This training is crucial to ensure the operators are competent to use these tools and to increase their acceptance and confidence to use them (Karalis, 2024). Misunderstanding how to use this tool and what is expected could lead to clinically significant errors or misalignment with the clinical or efficiency saving goals and workflow (Karalis, 2024).

1.2.2.2.1.3 Ethical and legal considerations

An additional challenge is related to ethical and legal considerations (McCague et al., 2023, Karalis, 2024). The implementation and use of this tool should follow the regulations such as the Food and Drug Administration (FDA) approval process or MDR (McCague et al., 2023). Also, it's essential to determine who is responsible in the event of errors caused by using DL-AS tool and what the consequences are from ethical and legal aspects. Accordingly, the integration of the DL-AS tool in clinical practice is not straightforward and required ethical, legal, technical, and educational considerations that necessary to be addressed to ensure safe use of such tool, especially with the expectation to use this tool more within the following year (Hindocha et al., 2023). In current practice, DL-AS is exclusively used within a human-in-the-loop workflow, where human operators check edit and verify all DL-AS outputs, mitigating these legal challenges and risks. To move beyond that to fully automated operation will be extremely challenging from a liability perspective.

1.2.2.3 Clinical use phase – quality assurance

Currently there is no clear consensus on the required quality assurance schedule for DL-AS tools. However, based on the previous literature's recommendation were established for developing two forms of QA: routine and case-specific QA (Liesbeth et al., 2020). Routine QA is performing general regular tests to ensure

this model is still working properly (Liesbeth et al., 2020), and has not been affected by external changes to data or workflows. Case-specific QA focuses on the performance of the model in each case (Liesbeth et al., 2020).

1.2.2.3.1 Routine QA

It is essential to have a quality management program to regularly check the performance of the model after any update in the software or any changes in the clinical workflow (Claessens et al., 2022, Liesbeth et al., 2020). Changes include imaging protocol, demographics, patient setup, and immobilization devices. When the output of the routine QA does not meet the expected performance, recommissioning the model may be required (Liesbeth et al., 2020, Claessens et al., 2022).

1.2.2.3.2 Manual verification and monitoring human interaction

The generated segmentation in each case should be reviewed, edited if required, and approved by clinical staff (Liesbeth et al., 2020, Cardenas et al., 2019). It is recommended to document any corrections to keep track of the cases where the autosegmentation model is underperforming and monitor how the user interacts with the model (Liesbeth et al., 2020). Several studies prove that supervision of the output is one of the most valuable tool to track the quality of the model (Liesbeth et al., 2020, Brouwer et al., 2020b). Moreover, a survey of medical physicists from 202 RT centres about their developed methods of QA for deep learning applications during clinical use reported that monitoring the performed manual interactions/edits is the most commonly used method (Brouwer et al., 2020b). However, some respondents stated there is no regular QA of their AI applications. This could be due to the lack of information about how to conduct the QA for the AI application in the clinical workflow (Brouwer et al., 2020b).

To our knowledge, only two studies evaluated how users interacted with a commercial deep learning autosegmentation CT model in delineating OARs after a year of routine clinical use (Brouwer et al., 2020a, Wong et al., 2021). The first study aimed to assess the performance of the deep learning autosegmentation model in the clinical workflow based on user experience and objective comparison metrics (Wong et al., 2021). A subjective assessment was through a survey conducted among dosimetrists, RTTs and radiation oncologists after they completed reviewing and editing of each generated segmentation (Wong et

al., 2021). The objective assessment metrics compare the generated segmentation without editing and after editing the automatic segmentation using 95% HD, and DSC (Wong et al., 2021). Using a survey to report users' experience and comparing unedited with edited autosegmentations will help highlight scenarios where the model is underperforming and address the issue with further training (Wong et al., 2021). These methods could be used to give feedback about the impact of the autosegmentation model on the clinical workflow and identify how it can be improved.

Moreover, another previous study conducted after a year of routine clinical use aimed to identify what manual adjustments were made during the review of the clinical cases of the autosegmentation to identify what modifications could be applied to improve the model output (Brouwer et al., 2020a). The findings showed that evaluating the manual adjustments of the autosegmentation in a typical clinical setting was a valuable tool to improve the quality of practice in autosegmentation by identifying which technical improvements are needed and highlighting the necessity for continued training of RTTs, as there was variation in their understanding of the guidelines (Brouwer et al., 2020a).

1.2.2.3.3 Statistical QA models and an independent AI-QA model

The concerns about routine manual checking and editing are that it depends on the individual user, which could introduce bias and could reduce the benefit of segmentation automation by spending time reviewing and editing the cases. Also, it is prone to missing clinically significant errors and failing to edit incorrect segmentation.

Accordingly, it is recommended to use data-driven QA methods to assist the users in making decisions regarding the accuracy of the generated segmentation and what modifications are needed. However, this is made very challenging by the absence of ground truth (manual gold standard) contours for routine clinical cases. There are potential solutions to this problem. For instance, a statistical model could be implemented to highlight outliers through the evaluation of the OARs characteristics (e.g. shape, volume, and centroid) to check if there is any significant deviation from expectation in the segmented volume (Claessens et al., 2022, Liesbeth et al., 2020, Cardenas et al., 2019). However, these methods can be insensitive and do not account well for unusual patient anatomy.

Moreover, another suggested method is using a secondary independent DL-AS algorithm as a verification method depending on AI. This is still emerging technology (Claessens et al., 2022). The aim is to detect the uncertainty in the segmentation, as disagreement between algorithms and identify in which areas the performance of the segmentation algorithm is suboptimal (Claessens et al., 2022). This will increase users' confidence and trust in the use of the DL-AS tool. Moreover, it can potentially expedite the verification process and increase safety. However, all these approaches support, but cannot replace the necessity for carefully reviewing the generated segmentation by the clinical team (Liesbeth et al., 2020, Cardenas et al., 2019).
Undoubtedly, patient specific QA for DL-AS is challenging, specifically in instances where the interpretability of the model is lacking such as trying to identify the reasons of certain segmentation prediction.

Three previous studies investigated different QA approaches to assess the quality of autosegmentation in RT. Spatial probability maps (SPMs) based on Monte-Carlo Dropout (MCD) were investigated on the salivary glands as a paradigm (van Rooij et al., 2021). The second study explored using a secondary neural network (NN) to predict Dice similarity Coefficient from breast CT-AS pairs, through internal class probability as an input to the QA network (Chen et al., 2020). The third study investigated the QA of DL-AS on head and neck CT using radiomics features located near contour boundaries to predict DSC scores (Luan et al., 2023).

However, these methods depend on the prediction-generating model and the distribution of training data to generate uncertainty estimates, and hence fail the test of independence. Moreover, they are also susceptible to internal probability calibration issues as well as creating practical challenges for use with existing or commercial AS models, which may not provide probabilistic predictions. Moreover, predicting DSC is not particularly valuable as it will not provide information about the location of the error, and it depends on the structure size and shape. Accordingly, it is important to develop an independent QA approach that can assess the quality of the segmentation without relying on DSC prediction or internal AS model probability.

1.3 Study Overview

1.3.1 Identifying research problems

Based on the previous literature regarding the performance of the DL-AS model on OARs and general recommendations for model training and validation, three gaps in knowledge need to be addressed.

First, the main issue with the performance of the DL-AS model is not related to the optimal model architecture, which has now been well established for some time, but it is more related to the quality and availability of training examples. It is well known that training the DL-AS model on a large dataset will tend to perform better on unseen data. However, it is challenging to acquire a large training dataset for MRI based DL-AS. This challenge is related to the difficulty in using MRI data from different institutions due to variations in the MRI scanner acquisition process. Moreover, obtaining large, high-quality annotated data for training is challenging as it needs significant effort and time to consistently contour and label all the structures and review them. We will investigate the use of small datasets, and the importance of the way editing is done on the gold standard label contours, by examining the effect of editing on the RayStation model performance.

The second area is related to DL-AS evaluation methods. Most researchers use geometric metrics alone to evaluate the model. However, based on recommendations on how to effectively assess the clinical applicability of DL-AS model, dosimetric evaluation is also needed. There is no agreed method for identifying the acceptable clinical threshold for dosimetric variation between the gold standard and autosegmentation. Moreover, the correlation between the geometric and dosimetric outputs needs to be investigated to determine if geometric evaluation alone is enough.

The third gap in the knowledge is related to the QA of generated segmentations in routine clinical settings. It is widely understood that regardless of the pre-clinical evaluation, the DL-AS model could generate incorrect segmentations for several reasons, such as the individual patient position, anatomy or disease pathology being different from the training data. Thus, currently, the generated segmentation should always be reviewed and edited if needed by the clinician. To avoid reducing the benefit of segmentation automation, an independent AI-

QA tool needs to be developed to aid the operators during the review process and enhance the safety and confidence of using the DL-AS tool clinically.
This PhD thesis aims to address these knowledge gaps related to training data for MRI DL-AS for OARs, necessity of geometric and dosimetric evaluations and enhancing the confidence and trust of using DL-AS tools clinically.

1.3.2 Aims

- To assess the accuracy and practicality of utilizing MRI DL-AS models for brain OARs, when trained using limited training dataset, focusing on the impact of pre-training editing on model quality from geometric and dosimetric perspectives.
- To build and test a novel AI-QA model called AutoConfidence (ACo) to assess the autosegmentation uncertainty without ground-truth or gold standard labels, on a per-patient basis.

1.3.3 Overall hypothesis and study focus

- We will test the hypothesis that there is no geometric difference between the DL-AS model trained with edited clinical contour versus the DL-AS model trained without editing of clinical contours.
- We will test the hypothesis that there is no dosimetric difference in OAR doses between the DL-AS model trained with edited clinical contour versus the DL-AS model trained without editing of clinical contours.
- We will test the hypothesis that it is possible to build the AutoConfidence QA model to detect segmentation uncertainty without ground-truth label data in routine clinical use.

1.3.4 Chapter overview

- Chapter 1: Overall introduction
- Chapter 2: Geometric Evaluations of CT and MRI based Deep Learning Segmentation for Brain OARs in Radiotherapy
- Chapter 3: Dosimetric Impact of Contour Editing on CT and MRI Deep Learning Autosegmentation for Brain OARs
- Chapter 4: Automated Confidence Estimation in MRI Deep Learning Segmentation (MRI DL-AS) for Brain OARs in Radiotherapy
- Chapter 5: Discussion, future work, and conclusions

1.4 References

AINSLIE, A. P., KLAVER, M., VOSHART, D. C., GERRITS, E., DEN DUNNEN, W. F. A., EGGEN, B. J. L., BERGINK, S. & BARAZZUOL, L. 2024. Glioblastoma and its treatment are associated with extensive accelerated brain aging. *Aging Cell,* 23**,** e14066.

ALDAPE, K., BRINDLE, K. M., CHESLER, L., CHOPRA, R., GAJJAR, A., GILBERT, M. R., GOTTARDO, N., GUTMANN, D. H., HARGRAVE, D., HOLLAND, E. C., JONES, D. T. W., JOYCE, J. A., KEARNS, P., KIERAN, M. W., MELLINGHOFF, I. K., MERCHANT, M., PFISTER, S. M., POLLARD, S. M., RAMASWAMY, V., RICH, J. N., ROBINSON, G. W., ROWITCH, D. H., SAMPSON, J. H., TAYLOR, M. D., WORKMAN, P. & GILBERTSON, R. J. 2019. Challenges to curing primary brain tumours. *Nat Rev Clin Oncol,* 16**,** 509-520.

AN, L., CHEN, J., CHEN, P., ZHANG, C., HE, T., CHEN, C., ZHOU, J. H., YEO, B. T. T., ALZHEIMER'S DISEASE NEUROIMAGING, I., AUSTRALIAN IMAGING, B. & LIFESTYLE STUDY OF, A. 2022. Goal-specific brain MRI harmonization. *Neuroimage,* 263**,** 119570.

ANGOM, R. S., NAKKA, N. M. R. & BHATTACHARYA, S. 2023. Advances in Glioblastoma Therapy: An Update on Current Approaches. *Brain Sci,* 13.

BIBAULT, J. E. & GIRAUD, P. 2024. Deep learning for automated segmentation in radiotherapy: a narrative review. *Br J Radiol,* 97**,** 13-20.

BROUWER, C. L., BOUKERROUI, D., OLIVEIRA, J., LOONEY, P., STEENBAKKERS, R., LANGENDIJK, J. A., BOTH, S. & GOODING, M. J. 2020a. 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.

BROUWER, C. L., DINKLA, A. M., VANDEWINCKELE, L., CRIJNS, W., CLAESSENS, M., VERELLEN, D. & VAN ELMPT, W. 2020b. Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. *Physics and Imaging in Radiation Oncology,* 16**,** 144-148.

BRUNESE, L., MERCALDO, F., REGINELLI, A. & 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. & BROCK, K. B. 2019. Advances in Auto-Segmentation. *Semin Radiat Oncol,* 29**,** 185-197.

CHEN, H., LU, W., CHEN, M., ZHOU, L., TIMMERMAN, R., TU, D., NEDZI, L., WARDAK, Z., JIANG, S., ZHEN, X. & GU, X. 2019. A recursive ensemble organ segmentation (REOS) framework: application in brain radiotherapy. *Phys Med Biol,* 64**,** 025015.

CHEN, X., MEN, K., CHEN, B., TANG, Y., ZHANG, T., WANG, S., LI, Y. & DAI, J. 2020. CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy. *Front Oncol,* 10**,** 524.

ÇIÇEK, Ö., ABDULKADIR, A., LIENKAMP, S. S., BROX, T. & RONNEBERGER, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International conference on medical image computing and computer-assisted intervention, 2016. Springer, 424-432.

CLAESSENS, M., ORIA, C. S., BROUWER, C. L., ZIEMER, B. P., SCHOLEY, J. E., LIN, H., WITZTUM, A., MORIN, O., NAQA, I. E., VAN ELMPT, W. & VERELLEN, D. 2022. Quality Assurance for AI-Based Applications in Radiation Therapy. *Semin Radiat Oncol,* 32**,** 421-431.

DU, G., CAO, X., LIANG, J., CHEN, X. & ZHAN, Y. 2020. Medical Image Segmentation based on U-Net: A Review. *Journal of Imaging Science and Technology,* 64**,** 20508-1-20508-12.

EL NAQA, I., RUAN, D., VALDES, G., DEKKER, A., MCNUTT, T., GE, Y., WU, Q. J., OH, J. H., THOR, M. & SMITH, W. 2018. Machine learning and modeling: Data, validation, communication challenges. *Medical physics,* 45**,** e834-e840.

FATANIA, K., CLARK, A., FROOD, R., SCARSBROOK, A., AL-QAISIEH, B., CURRIE, S. & NIX, M. 2022. Harmonisation of scanner-dependent contrast variations in magnetic resonance imaging for radiation oncology, using style-blind auto-encoders. *Phys Imaging Radiat Oncol,* 22**,** 115-122.

GIBBONS, E., HOFFMANN, M., WESTHUYZEN, J., HODGSON, A., CHICK, B. & LAST, A. 2023. Clinical evaluation of deep learning and atlas-based auto-segmentation for critical organs at risk in radiation therapy. *J Med Radiat Sci,* 70 Suppl 2**,** 15-25.

HANSEN, C. R., HUSSEIN, M., BERNCHOU, U., ZUKAUSKAITE, R. & THWAITES, D. 2022. Plan quality in radiotherapy treatment planning -

Review of the factors and challenges. *J Med Imaging Radiat Oncol,* 66**,** 267-278.

HARRISON, K., PULLEN, H., WELSH, C., OKTAY, O., ALVAREZ-VALLE, J. & JENA, R. 2022. Machine Learning for Auto-Segmentation in Radiotherapy Planning. *Clin Oncol (R Coll Radiol),* 34**,** 74-88.

HEILEMANN, G., BUSCHMANN, M., LECHNER, W., DICK, V., ECKERT, F., HEILMANN, M., HERRMANN, H., MOLL, M., KNOTH, J., KONRAD, S., SIMEK, I. M., THIELE, C., ZAHARIE, A., GEORG, D., WIDDER, J. & TRNKOVA, P. 2023. Clinical Implementation and Evaluation of Auto-Segmentation Tools for Multi-Site Contouring in Radiotherapy. *Phys Imaging Radiat Oncol,* 28**,** 100515.

HESAMIAN, M. H., JIA, W., HE, X. & KENNEDY, P. 2019. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *J Digit Imaging,* 32**,** 582-596.

HINDOCHA, S., ZUCKER, K., JENA, R., BANFILL, K., MACKAY, K., PRICE, G., PUDNEY, D., WANG, J. & TAYLOR, A. 2023. Artificial Intelligence for Radiotherapy Auto-Contouring: Current Use, Perceptions of and Barriers to Implementation. *Clin Oncol (R Coll Radiol),* 35**,** 219-226.

JOSEPH, V. R. 2022. Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal,* 15**,** 531-538.

KALET, A. M., LUK, S. M. & PHILLIPS, M. H. 2020. Radiation therapy quality assurance tasks and tools: the many roles of machine learning. *Medical physics,* 47**,** e168-e177.

KARALIS, V. D. 2024. The Integration of Artificial Intelligence into Clinical Practice. *Applied Biosciences,* 3**,** 14-44.

KAZEMIFAR, S., MCGUIRE, S., TIMMERMAN, R., WARDAK, Z., NGUYEN, D., PARK, Y., JIANG, S. & 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. & OLSSON, L. E. 2021. Prospective Clinical Feasibility Study for MRI-Only Brain Radiotherapy. *Front Oncol,* 11**,** 812643.

LIESBETH, V., MICHAËL, C., ANNA, M. D., CHARLOTTE, L. B., WOUTER, C. & DIRK, V. 2020. Overview of artificial intelligence-based applications in

radiotherapy: recommendations for implementation and quality assurance. *Radiotherapy and Oncology*.

LIU, F., YADAV, P., BASCHNAGEL, A. M. & MCMILLAN, A. B. 2019. MR-based treatment planning in radiation therapy using a deep learning approach. *Journal of applied clinical medical physics,* 20**,** 105-114.

LONG, J., SHELHAMER, E. & DARRELL, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 3431-3440.

LU, Y., PATEL, M., NATARAJAN, K., UGHRATDAR, I., SANGHERA, P., JENA, R., WATTS, C. & SAWLANI, V. 2020. Machine learning-based radiomic, clinical and semantic feature analysis for predicting overall survival and MGMT promoter methylation status in patients with glioblastoma. *Magn Reson Imaging,* 74**,** 161-170.

LUAN, S., XUE, X., WEI, C., DING, Y., ZHU, B. & WEI, W. 2023. Machine Learning-Based Quality Assurance for Automatic Segmentation of Head-and-Neck Organs-at-Risk in Radiotherapy. *Technol Cancer Res Treat,* 22**,** 15330338231157936.

MACKAY, K., BERNSTEIN, D., GLOCKER, B., KAMNITSAS, K. & TAYLOR, A. 2023. A Review of the Metrics Used to Assess Auto-Contouring Systems in Radiotherapy. *Clin Oncol (R Coll Radiol),* 35**,** 354-369.

MCCAGUE, C., MACKAY, K., WELSH, C., CONSTANTINOU, A., JENA, R., CRISPIN-ORTUZAR, M. & IMAGING, A. I. E. C. G. 2023. Position statement on clinical evaluation of imaging AI. *Lancet Digit Health,* 5**,** e400-e402.

MEKKI, L., ACHARYA, S., LADRA, M. & LEE, J. 2024. Deep learning segmentation of organs-at-risk with integration into clinical workflow for pediatric brain radiotherapy. *J Appl Clin Med Phys,* 25**,** e14310.

MILLER, K. D., OSTROM, Q. T., KRUCHKO, C., PATIL, N., TIHAN, T., CIOFFI, G., FUCHS, H. E., WAITE, K. A., JEMAL, A., SIEGEL, R. L. & BARNHOLTZ-SLOAN, J. S. 2021. Brain and other central nervous system tumor statistics, 2021. *CA Cancer J Clin,* 71**,** 381-406.

MILLETARI, F., NAVAB, N. & AHMADI, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision (3DV), 2016. IEEE, 565-571.

MLYNARSKI, P., DELINGETTE, H., ALGHAMDI, H., BONDIAU, P. Y. & AYACHE, N. 2020. Anatomically consistent CNN-based segmentation of organs-at-risk in cranial radiotherapy. *J Med Imaging (Bellingham),* 7**,** 014502.

OH, S., KIM, J. H., CHOI, S.-W., LEE, H. J., HONG, J. & KWON, S. H. 2019. Physician confidence in artificial intelligence: an online mobile survey. *Journal of medical Internet research,* 21**,** e12422.

OSTROM, Q. T., PATIL, N., CIOFFI, G., WAITE, K., KRUCHKO, C. & BARNHOLTZ-SLOAN, J. S. 2020. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013-2017. *Neuro Oncol,* 22**,** iv1-iv96.

RAGHAVAPUDI, H., SINGROUL, P. & KOHILA, V. 2021. Brain Tumor Causes, Symptoms, Diagnosis and Radiotherapy Treatment. *Curr Med Imaging,* 17**,** 931-942.

RANTA, I., WRIGHT, P., SUILAMO, S., KEMPPAINEN, R., SCHUBERT, G., KAPANEN, M. & 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.

RONG, Y., CHEN, Q., FU, Y., YANG, X., AL-HALLAQ, H. A., WU, Q. J., YUAN, L., XIAO, Y., CAI, B., LATIFI, K., BENEDICT, S. H., BUCHSBAUM, J. C. & QI, X. S. 2023. NRG Oncology Assessment of Artificial Intelligence Deep Learning-Based Auto-segmentation for Radiation Therapy: Current Developments, Clinical Considerations, and Future Directions. *Int J Radiat Oncol Biol Phys*.

RONNEBERGER, O., FISCHER, P. & BROX, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, 2015. Springer, 234-241.

SCHIPAANBOORD, B., BOUKERROUI, D., PERESSUTTI, D., VAN SOEST, J., LUSTBERG, T., DEKKER, A., ELMPT, W. V. & GOODING, M. J. 2019. An Evaluation of Atlas Selection Methods for Atlas-Based Automatic Segmentation in Radiotherapy Treatment Planning. *IEEE Trans Med Imaging,* 38**,** 2654-2664.

SHERER, M. V., LIN, D., ELGUINDI, S., DUKE, S., TAN, L. T., CACICEDO, J., DAHELE, M. & GILLESPIE, E. F. 2021. Metrics to evaluate the

performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother Oncol,* 160**,** 185-191.

SOOMRO, T. A., ZHENG, L., AFIFI, A. J., ALI, A., SOOMRO, S., YIN, M. & GAO, J. 2023. Image Segmentation for MR Brain Tumor Detection Using Machine Learning: A Review. *IEEE Rev Biomed Eng,* PP.

TANG, H., CHEN, X., LIU, Y., LU, Z., YOU, J., YANG, M., YAO, S., ZHAO, G., XU, Y. & CHEN, T. 2019. Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nature Machine Intelligence,* 1**,** 480-491.

THWAITES, D., MOSES, D., HAWORTH, A., BARTON, M. & HOLLOWAY, L. 2021. Artificial intelligence in medical imaging and radiation oncology: Opportunities and challenges. *J Med Imaging Radiat Oncol,* 65**,** 481-485.

TURCAS, A., LEUCUTA, D., BALAN, C., CLEMENTEL, E., GHEARA, C., KACSO, A., KELLY, S. M., TANASA, D., CERNEA, D. & ACHIMAS-CADARIU, P. 2023. Deep-learning magnetic resonance imaging-based automatic segmentation for organs-at-risk in the brain: Accuracy and impact on dose distribution. *Phys Imaging Radiat Oncol,* 27**,** 100454.

UWISHEMA, O., FREDERIKSEN, K. S., BADRI, R., PRADHAN, A. U., SHARIFF, S., ADANUR, I., DOST, B., ESENE, I. & ROSSEAU, G. 2023. Epidemiology and etiology of brain cancer in Africa: A systematic review. *Brain Behav,* 13**,** e3112.

VAASSEN, F., HAZELAAR, C., VANIQUI, A., GOODING, M., VAN DER HEYDEN, B., CANTERS, R. & VAN ELMPT, W. 2020. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology,* 13**,** 1-6.

VAN DER HEYDEN, B., WOHLFAHRT, P., EEKERS, D. B. P., RICHTER, C., TERHAAG, K., TROOST, E. G. C. & VERHAEGEN, F. 2019. Dual-energy CT for automatic organs-at-risk segmentation in brain-tumor patients using a multi-atlas and deep-learning approach. *Sci Rep,* 9**,** 4126.

VAN DER VEEN, J., WILLEMS, S., DESCHUYMER, S., ROBBEN, D., CRIJNS, W., MAES, F. & NUYTS, S. 2019. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiotherapy and Oncology,* 138**,** 68-74.

VAN DIJK, L. V., VAN DEN BOSCH, L., ALJABAR, P., PERESSUTTI, D., BOTH, S., R, J. H. M. S., LANGENDIJK, J. A., GOODING, M. J. & BROUWER, C. L. 2020. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiother Oncol,* 142**,** 115-123.

VAN ROOIJ, W., DAHELE, M., BRANDAO, H. R., DELANEY, A. R., SLOTMAN, B. J. & VERBAKEL, W. F. 2019. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. *International Journal of Radiation Oncology* Biology* Physics,* 104**,** 677-684.

VAN ROOIJ, W., VERBAKEL, W. F., SLOTMAN, B. J. & DAHELE, M. 2021. Using Spatial Probability Maps to Highlight Potential Inaccuracies in Deep Learning-Based Contours: Facilitating Online Adaptive Radiation Therapy. *Adv Radiat Oncol,* 6**,** 100658.

WANG, Y., ZHAO, L., WANG, M. & 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**,** 144591-144602.

Westbrook, C., Roth,C.K. and Talbot, J. (2011). MRI in Practice. 4th ed. Chichester: Wiley-Blackwell.

WIESINGER, F., PETIT, S., HIDEGHÉTY, K., HERNANDEZ TAMAMES, J., MCCALLUM, H., MAXWELL, R., PEARSON, R., VERDUIJN, G., DARÁZS, B., KAUSHIK, S., COZZINI, C., BOBB, C., FODOR, E., PACZONA, V., KÓSZÓ, R., EGYÜD, Z., BORZASI, E., VÉGVÁRY, Z., TAN, T., GYALAI, B., CZABÁNY, R., DEÁK-KARANCSI, B., KOLOZSVÁRI, B., CZIPCZER, V., CAPALA, M. & RUSKÓ, L. 2021. Deep-Learning-based Segmentation of Organs-at-Risk in the Head for MR-assisted Radiation Therapy Planning. *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies.*

WILLEMS, S., CRIJNS, W., SAINT-ESTEVEN, A. L. G., VAN DER VEEN, J., ROBBEN, D., DEPUYDT, T., NUYTS, S., HAUSTERMANS, K. & MAES, F. 2018. Clinical implementation of DeepVoxNet for auto-delineation of organs at risk in head and neck cancer patients in radiotherapy. *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis.* Springer.

WONG, J., FONG, A., MCVICAR, N., SMITH, S., GIAMBATTISTA, J., WELLS, D., KOLBECK, C., GIAMBATTISTA, J., GONDARA, L. & 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**,** 152-158.

WONG, J., HUANG, V., WELLS, D., GIAMBATTISTA, J., GIAMBATTISTA, J., KOLBECK, C., OTTO, K., SAIBISHKUMAR, E. P. & ALEXANDER, A. 2021. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. *Radiation Oncology,* 16**,** 1-10.

YANG, R. & YU, Y. 2021. Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis. *Frontiers in Oncology,* 11**,** 573.

Chapter 2 Geometric Evaluations of CT and MRI based Deep Learning Segmentation for Brain OARs in Radiotherapy

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 MRI. We trained and evaluated 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 DSC, sensitivity, and MDA. 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.

Keywords: Brain, Organs at risk, Autosegmentation, 3D U-net, Deep learning, MRI scans, CT scans.

2.1 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., 2019b). 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 dosimetrist 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., 2020a). 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 (Mlynarski et al., 2020, Wiesinger et al., 2021, Chen et al., 2019). 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 (Mlynarski et al., 2020, Wiesinger et al., 2021, Chen et al., 2019). 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.

2.2 Materials and Methods

2.2.1 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 cohort contains both biopsy and resection. 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: kilovoltage peak (kVp): 120, Field of view (FOV): 500mm, 1mm*1mm in-plane resolution, slice thickness: 2mm, 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: 2mm, scanner: Siemens Magnetom Sola with 1mm in-plane resolution and Gd contrast.

2.2.2 Brain OARs and gold standard atlas

The OAR selection was based on the four central nervous system (CNS) 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 (Eekers et al., 2018, Chen et al., 2019, Scoccianti et al., 2015, Mir et al., 2020, Ho et al., 2018). 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.

Note: the delineation of OARs was done by an expert radiation oncologist. The anatomical descriptions and contouring guidance in the atlas were written by me after reviewing several international consensus guidelines (Eekers et al., 2018, Chen et al., 2019, Scoccianti et al., 2015, Mir et al., 2020, Ho et al., 2018). Then, the atlas was reviewed and approved by the treating radiation neuro-oncology team.

2.2.3 Clinical contours and quality assurance (QA)

All image sets and original clinical contours, which were manually delineated by the clinician for radiotherapy treatment planning, were reviewed for image quality, contour accuracy and OAR labelling. The 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).

Note: The entire process of review and editing was done by me after being trained by an expert radiation oncologist and was then reviewed and approved by an expert radiation oncologist.

2.2.4 Deep learning autosegmentation training

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

2.2.5 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).

The gold standard contour represents the most accurate and widely accepted delineation, aligned with international consensus guidelines, and serves as the reference for comparisons. This contour was drawn by the same PhD student, following guidance from the brain OARs atlas, and was then reviewed and approved by an expert radiation oncologist

2.2.5.1 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 ≤ 0.005 (0.05/9) for MRI geometric evaluations, and ≤ 0.016 (0.05/3) for CT geometric evaluation.

2.3 Results

2.3.1 Comparison of CT vs 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 (suppl. Info. Table S1 and S2) (figure 2.1 and 2.2). 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 (suppl. Info. Table S4 and S5) (figure 2.3).
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 (suppl. Info. Table 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) (suppl. Info. 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) (suppl. Info. Table S1, S2 and S3).

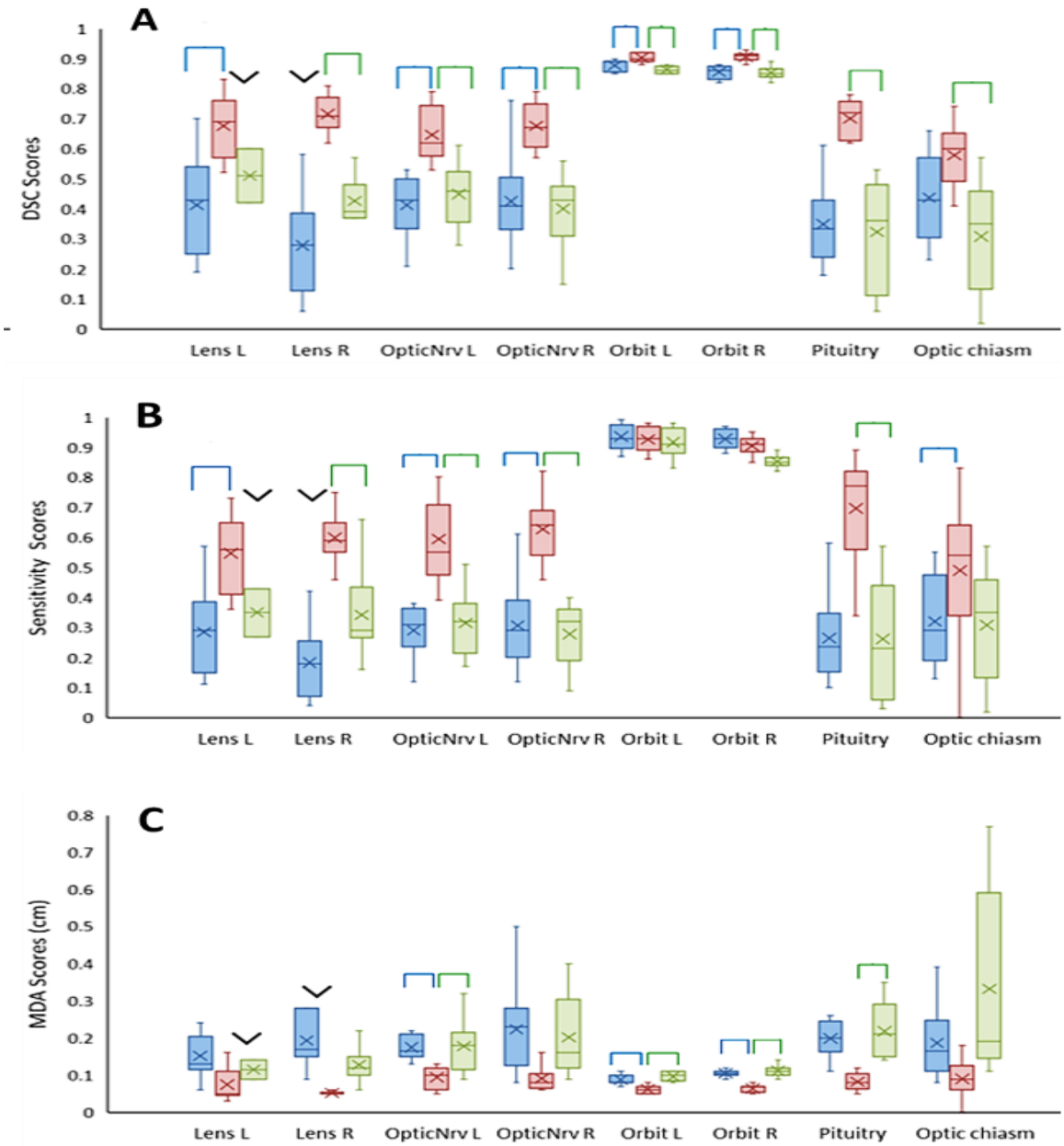


Figure 2.1: The distribution of the a) DSC, b) sensitivity, and c) MDA (for MRI-based deep learning segmentations from three different MRI models): the MRIu segmentation (blue), the MRIeMRI segmentation (red), and the MRIeCT segmentation (green). The green square bracket denotes the statistically significant 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).

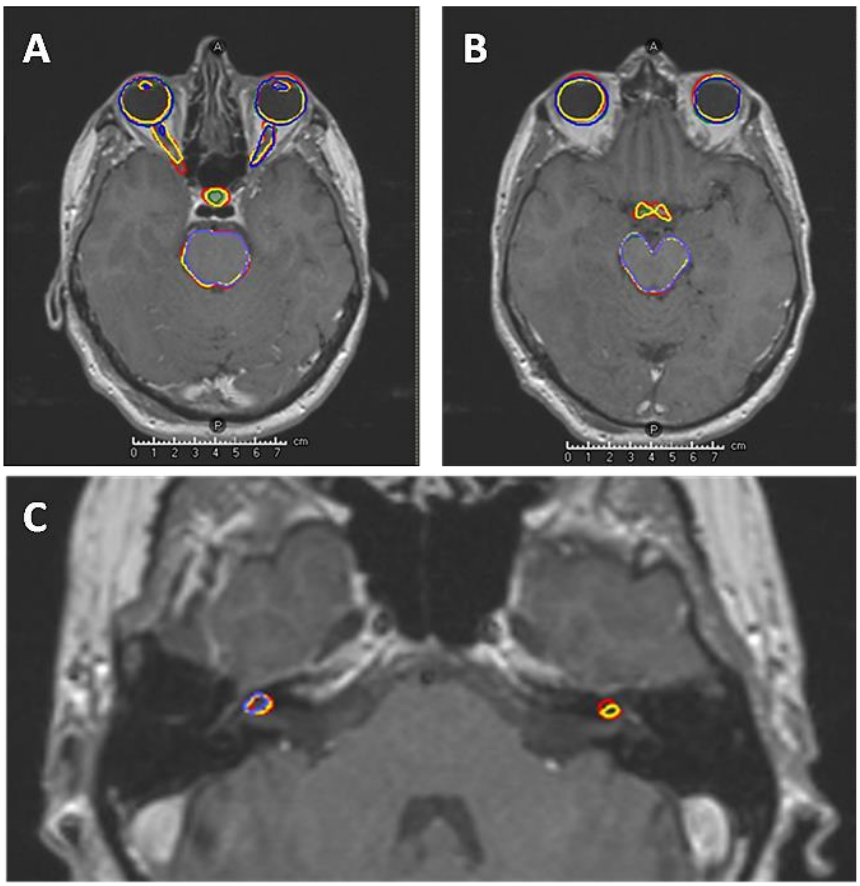


Figure 2.2: T1-weighted gadolinium-enhanced MRI showing examples of the predicted MRI deep learning segmentations compared to the gold standard segmentation of the orbits (a, b), lenses (a), optic nerves (a), brainstem (a, b), optic chiasm (b), cochlea (c), and pituitary(a). Red represents the gold standard 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.

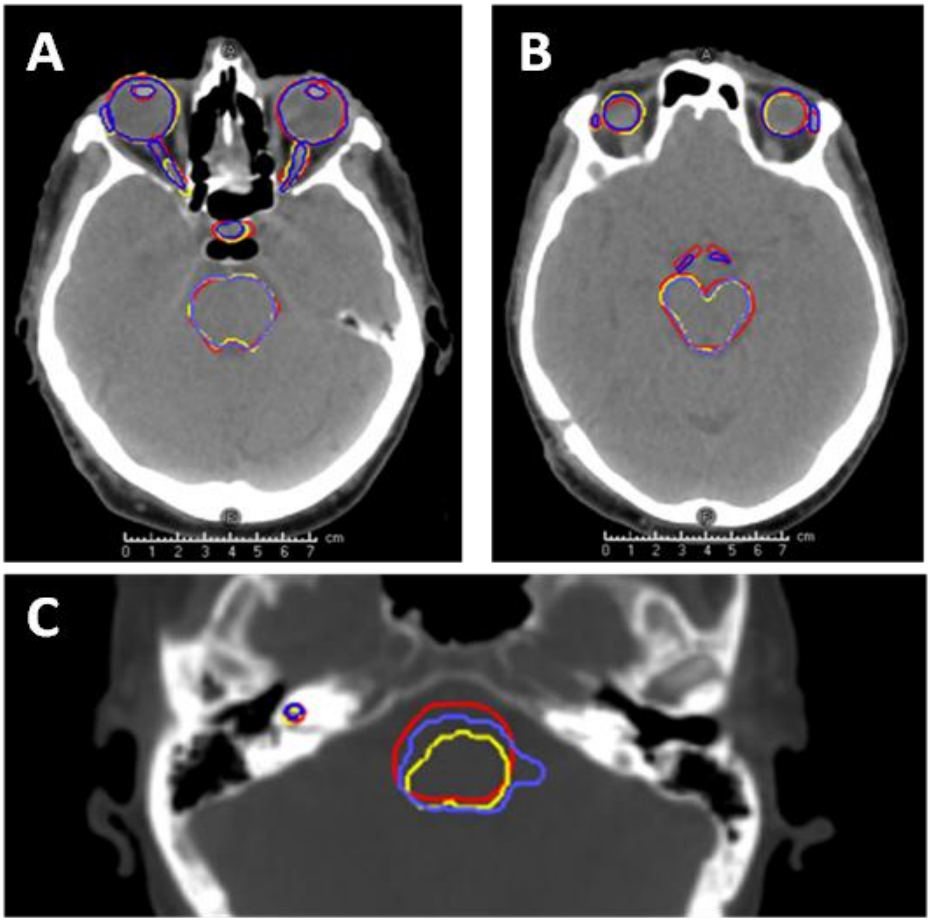


Figure 2.3: CT axial scans showing examples of the predicted CT deep learning segmentations compared to the gold standard segmentation of the orbits (a,b), lenses (a), optic nerves (a), brainstem(a,b,c), optic chiasm (b), cochlea (c), lacrimal glands (a,b), and pituitary (a). Red represents the gold standard segmentation. CTeCT is shown in yellow, while CTu in blue.

2.3.2 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 numbers, where the model failed to produce the segmentation. 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 number of failures 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 number of failures for the cochlea, which was almost entirely resolved when using the MRIeMRI model (suppl. Info. 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 (suppl. Info. 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 2.1).
A statistically significant difference was found between MRIeMRI vs both the MRIeCT and MRIu models in the delineation of the structures shown in fig 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 2.1).

	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 vs MRIeCT (- means MRIeMRI performed better)													
p* (*Threshold: ≤ 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 vs MRIu (- means MRIeMRI performed better)													
p* (*Threshold: ≤ 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 vs MRIu (- means MRIeCT performed better)													
p* (*Threshold: ≤ 0.005)	0.430	$$	##	0.179	0.234	##	0.008	$$	0.321	0.638	0.035	0.622	$$
Effect size: Δ median	0.020			0.000	0.000		-0.115		-0.030	-0.020	0.030	0.010	
N*	9	1	2	7	8	2	6	5	9	9	9	9	4
MDA													
MRIeMRI vs MRIeCT (+means MRIeMRI performed better)													
p* (*Threshold: ≤ 0.005)	0.042	**	**	$$	$$	**	0.173	**0.031**	**0.001**	0.006	**0.000**	**0.000**	**0.004**
Effect size: Δ median	0.020						0.080	0.100	0.080	0.080	0.040	0.040	0.130
N*	9	5	4	4	2	2	8	8	9	9	9	9	7
MRIeMRI vs MRIu (+means MRIeMRI performed better)													
p* (*Threshold: ≤ 0.005)	0.179	**	**	##	##	**0.006**	**	0.074	**0.002**	0.017	**0.000**	**0.000**	0.011
Effect size: Δ median	0.010					0.080		0.080	0.070	0.150	0.020	0.040	0.100
N*	9	1	5	5	2	9	5	6	9	9	9	9	6
MRIeCT vs MRIu (+means MRIeCT performed better)													
p* (*Threshold: ≤ 0.005)	0.325	$$	##	#	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 vs MRIeCT (- means MRIeMRI performed better)													
p* (*Threshold: ≤ 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 vs MRIu (- means MRIeMRI performed better)													
p* (*Threshold: ≤ 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
MRIeCT vs MRIu (- means MRIeCT performed better)													
p* (*Threshold: ≤ 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.
MRI Unedited is better based on producing the segmentation for more cases.
 # MDA unreliable due to insufficient overlap of OARs.

Table 2.1: 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 ≤ 0.005). Insufficient successful segmentations were achieved by one of the models, this is noted ($$, **, or ##), indicating the superior model.

2.4 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 (Lerner et al., 2021, Kazemifar et al., 2019, 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. However, since T1-w and T2-w images are acquired separately, there is a potential for patient movement and misregistration of the sequences, which complicates the use of the multi-modality MRI model.

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., 2019b), due to unbalanced losses. In our model, loss balancing across OARs was performed to minimise this effect. Loss balancing was achieved using a weighting factor proportional to the inverse of the volume of each OAR.

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 (suppl. Info. 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 2.1). 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 vs MRIeCT and MRIu models in the delineation of orbits (p<0.001), the effect size was generally small (Table 2.1). 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 (suppl. Info. Table 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) (suppl. Info. 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(Mlynarski et al., 2020, Wiesinger et al., 2021, Chen et al., 2019). 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 vs. 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 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.

2.5 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: 255585.

2.6 References

BROUWER, C. L., BOUKERROUI, D., OLIVEIRA, J., LOONEY, P., STEENBAKKERS, R., LANGENDIJK, J. A., BOTH, S. & 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. & 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. & BROCK, K. B. 2019. Advances in Auto-Segmentation. *Semin Radiat Oncol,* 29**,** 185-197.

CHEN, H., LU, W., CHEN, M., ZHOU, L., TIMMERMAN, R., TU, D., NEDZI, L., WARDAK, Z., JIANG, S., ZHEN, X. & GU, X. 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. & ZOU, C. 2019. Development of a contouring guide for three different types of optic chiasm: A practical approach. *J Med Imaging Radiat Oncol,* 63**,** 657-664.

ÇIÇEK, Ö., ABDULKADIR, A., LIENKAMP, S. S., BROX, T. & RONNEBERGER, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International conference on medical image computing and computer-assisted intervention, 2016. Springer, 424-432.

EEKERS, D. B., IN 'T VEN, L., ROELOFS, E., POSTMA, A., ALAPETITE, C., BURNET, N. G., CALUGARU, V., COMPTER, I., COREMANS, I. E. M., HOYER, M., LAMBRECHT, M., NYSTROM, P. W., MENDEZ ROMERO, A., PAULSEN, F., PERPAR, A., DE RUYSSCHER, D., RENARD, L., TIMMERMANN, B., VITEK, P., WEBER, D. C., VAN DER WEIDE, H. L., WHITFIELD, G. A., WIGGENRAAD, R., TROOST, E. G. C. & EUROPEAN PARTICLE THERAPY NETWORK" OF, E. 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. & 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. & 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. *Statistical Analysis and Data Mining: The ASA Data Science Journal,* 15**,** 531-538.

KAZEMIFAR, S., MCGUIRE, S., TIMMERMAN, R., WARDAK, Z., NGUYEN, D., PARK, Y., JIANG, S. & 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. & OLSSON, L. E. 2021. Prospective Clinical Feasibility Study for MRI-Only Brain Radiotherapy. *Front Oncol,* 11**,** 812643.

LIU, F., YADAV, P., BASCHNAGEL, A. M. & MCMILLAN, A. B. 2019. MR-based treatment planning in radiation therapy using a deep learning approach. *Journal of applied clinical medical physics,* 20**,** 105-114.

MAYO, C. S., MORAN, J. M., BOSCH, W., XIAO, Y., MCNUTT, T., POPPLE, R., MICHALSKI, J., FENG, M., MARKS, L. B., FULLER, C. D., YORKE, E., PALTA, J., GABRIEL, P. E., MOLINEU, A., MATUSZAK, M. M., COVINGTON, E., MASI, K., RICHARDSON, S. L., RITTER, T., MORGAS, T., FLAMPOURI, S., SANTANAM, L., MOORE, J. A., PURDIE, T. G., MILLER, R. C., HURKMANS, C., ADAMS, J., JACKIE WU, Q. R., FOX, C. J., SIOCHI, R. A., BROWN, N. L., VERBAKEL, W., ARCHAMBAULT, Y., CHMURA, S. J., DEKKER, A. L., EAGLE, D. G., FITZGERALD, T. J., HONG, T., KAPOOR, R., LANSING, B., JOLLY, S., NAPOLITANO, M. E., PERCY, J., ROSE, M. S., SIDDIQUI, S., SCHADT, C., SIMON, W. E., STRAUBE, W. L., ST JAMES, S. T., ULIN, K., YOM, S. S. & YOCK, T. I. 2018. American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. *Int J Radiat Oncol Biol Phys,* 100**,** 1057-1066.

MIR, R., KELLY, S. M., XIAO, Y., MOORE, A., CLARK, C. H., CLEMENTEL, E., CORNING, C., EBERT, M., HOSKIN, P. & HURKMANS, C. W. 2020. Organ at risk delineation for radiation therapy clinical trials: Global harmonization group consensus guidelines. *Radiotherapy and Oncology,* 150**,** 30-39.

MLYNARSKI, P., DELINGETTE, H., ALGHAMDI, H., BONDIAU, P. Y. & 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. & 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.

SCHMIDT, M. A. & PAYNE, G. S. 2015. Radiotherapy planning using MRI. *Physics in Medicine & Biology,* 60**,** R323.

SCOCCIANTI, S., DETTI, B., GADDA, D., GRETO, D., FURFARO, I., MEACCI, F., SIMONTACCHI, G., DI BRINA, L., BONOMO, P., GIACOMELLI, I., MEATTINI, I., MANGONI, M., CAPPELLI, S., CASSANI, S., TALAMONTI, C., BORDI, L. & LIVI, L. 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. & 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. & GAO, J. 2022. Image Segmentation for MR Brain Tumor Detection Using Machine Learning: A Review. *IEEE Rev Biomed Eng,* PP.

VAN DIJK, L. V., VAN DEN BOSCH, L., ALJABAR, P., PERESSUTTI, D., BOTH, S., R, J. H. M. S., LANGENDIJK, J. A., GOODING, M. J. & BROUWER, C. L. 2020. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiother Oncol,* 142**,** 115-123.

VAN ROOIJ, W., DAHELE, M., BRANDAO, H. R., DELANEY, A. R., SLOTMAN, B. J. & VERBAKEL, W. F. 2019. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. *International Journal of Radiation Oncology* Biology* Physics,* 104**,** 677-684.

WANG, Y., ZHAO, L., WANG, M. & 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**,** 144591-144602.

WIESINGER, F., BYLUND, M., YANG, J., KAUSHIK, S., SHANBHAG, D., AHN, S., JONSSON, J. H., LUNDMAN, J. A., HOPE, T., NYHOLM, T., LARSON, P. & COZZINI, C. 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**,** 1440-1451.

WIESINGER, F., PETIT, S., HIDEGHÉTY, K., HERNANDEZ TAMAMES, J., MCCALLUM, H., MAXWELL, R., PEARSON, R., VERDUIJN, G., DARÁZS, B., KAUSHIK, S., COZZINI, C., BOBB, C., FODOR, E., PACZONA, V., KÓSZÓ, R., EGYÜD, Z., BORZASI, E., VÉGVÁRY, Z., TAN, T., GYALAI, B., CZABÁNY, R., DEÁK-KARANCSI, B., KOLOZSVÁRI, B., CZIPCZER, V., CAPALA, M. & RUSKÓ, L. 2021. Deep-Learning-based Segmentation of Organs-at-Risk in the Head for MR-assisted Radiation Therapy Planning. *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies.*

WONG, J., FONG, A., MCVICAR, N., SMITH, S., GIAMBATTISTA, J., WELLS, D., KOLBECK, C., GIAMBATTISTA, J., GONDARA, L. & 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**,** 152-158.

2.7 Supplementary Material

	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 scores for the MRIeCT													
Average	0.89	0.51	0.55	0.01	0.04	0.51	0.44	0.31	0.45	0.40	0.86	0.85	0.32
SD	0.02	0.29	0.30	0.03	0.09	0.23	0.16	0.21	0.11	0.13	0.01	0.02	0.21
DSC score for the MRIeMRI													
Average	0.90	0.57	0.49	0.10	0.15	0.68	0.67	0.51	0.65	0.68	0.90	0.91	0.67
SD	0.02	0.15	0.25	0.12	0.10	0.10	0.26	0.22	0.09	0.08	0.02	0.02	0.12
DSC score for the MRIu													
Average	0.89	0.73	0.52	0.04	0.02	0.41	0.28	0.44	0.41	0.43	0.88	0.86	0.35
SD	0.02	0.24	0.29	0.06	0.04	0.17	0.20	0.25	0.10	0.16	0.02	0.02	0.21

Table S1: Average and SD of the DSC scores for the MRI deep learning-based segmentation models

	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
The sensitivity score of the MRIeCT													
Average	0.86	0.37	0.45	0.02	0.05	0.35	0.34	0.31	0.31	0.28	0.92	0.85	0.26
SD	0.041	0.223	0.247	0.037	0.128	0.159	0.150	0.206	0.108	0.104	0.050	0.021	0.208
The sensitivity score of the MRIeMRI													
Average	0.88	0.50	0.42	0.06	0.10	0.55	0.55	0.49	0.59	0.63	0.93	0.90	0.70
SD	0.024	0.173	0.234	0.077	0.063	0.129	0.240	0.244	0.139	0.108	0.042	0.030	0.177
The sensitivity score of the MRI unedited													
Average	0.85	0.62	0.44	0.05	0.02	0.29	0.18	0.32	0.29	0.31	0.94	0.93	0.27
SD	0.037	0.207	0.271	0.067	0.033	0.151	0.139	0.202	0.086	0.148	0.043	0.032	0.187

Table S2: Average and SD of the sensitivity scores for the MRI deep learning-based segmentation models

	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
The MDA scores of the MRIeCT													
Average	0.10	0.10	0.06	0.32	0.29	0.12	0.13	0.33	0.18	0.20	0.10	0.11	0.22
SD	0.1	0.07	0.06	0.32	0.27	0.12	0.12	0.19	0.18	0.16	0.1	0.11	0.21
The MDA scores of the MRIeMRI													
Average	0.08	0.07	0.10	0.23	0.18	0.07	0.09	0.10	0.09	0.09	0.06	0.06	0.09
SD	0.02	0.03	0.05	0.12	0.06	0.04	0.11	0.05	0.03	0.03	0.01	0.01	0.04
The MDA scores of the MRIu													
Average	0.09	0.08	0.08	0.42	0.31	0.15	0.19	0.19	0.20	0.22	0.09	0.11	0.20
SD	0.02	0.03	0.05	0.23	0.14	0.06	0.11	0.13	0.08	0.13	0.01	0.01	0.11

Table S3: Average and SD of the MDA scores for the MRI deep learning-based segmentation models

	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 scores for the CTeCT													
Average	0.87	0.51	0.63	0.40	0.44	0.71	0.72	0.18	0.50	0.49	0.90	0.90	0.44
SD	0.04	0.21	0.22	0.15	0.16	0.24	0.31	0.14	0.12	0.16	0.02	0.02	0.31
DSC scores for the CTu													
Average	0.62	0.58	0.48	0.30	0.36	0.72	0.67	0.29	0.47	0.57	0.64	0.64	0.47
SD	0.43	0.29	0.29	0.14	0.22	0.35	0.37	0.16	0.24	0.29	0.44	0.44	0.29

Table S4: Average and SD of the DSC scores for the CT deep learning-based segmentation models

	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
The sensitivity score of the CTeCT													
Average	0.86	0.41	0.62	0.38	0.40	0.66	0.69	0.15	0.39	0.39	0.88	0.93	0.35
SD	0.065	0.241	0.261	0.185	0.202	0.238	0.310	0.130	0.110	0.147	0.064	0.052	0.277
The sensitivity score of the CTu													
Average	0.62	0.57	0.48	0.41	0.35	0.70	0.67	0.28	0.33	0.48	0.63	0.62	0.36
SD	0.432	0.307	0.298	0.217	0.216	0.360	0.386	0.181	0.169	0.258	0.440	0.431	0.236

Table S5: Average and SD of the sensitivity scores for the CT deep learning-based segmentation models

	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
The MDA scores of the CTeCT (cm)													
Average	0.11	0.08	0.06	0.18	0.29	0.05	0.05	0.40	0.15	0.15	0.07	0.08	0.17
SD	0.04	0.05	0.03	0.10	0.41	0.02	0.03	0.38	0.05	0.08	0.02	0.01	0.14
The MDA scores of the CTu (cm)													
Average	0.10	0.07	0.11	0.28	0.43	0.05	0.04	0.36	0.16	0.12	0.06	0.06	0.15
SD	0.05	0.04	0.09	0.12	0.40	0.03	0.02	0.38	0.09	0.09	0.03	0.03	0.09

Table S6: Average and SD of the MDA scores for the CT deep learning-based segmentation models

	Deep Learning Segmentation Models					
OARs	CTeCT	CTu		MRIeCT	MRIeMRI	MRIu
Optic Nerve L		3				
Optic Nerve R		3				
Cochlea L		3		4		8
Cochlea R	1	3		5	2	3
Lacrimal Gland L		8		2	3	1
Lacrimal Gland R		5			7	1
Lens L	1	3		7		
Lens R	2	2			1	3
Optic Chiasm		3		1		3
Pituitary		3		2		3
TOTAL	4	36		21	13	22

Table S7: the number of failed segmentations based on each deep learning 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													
P* (*Threshold:* ≤ *0.016)	0.083	0.176	0.205	**	**	0.644	0.574	**0.009**	0.407	0.047	0.093	0.102	0.213
Effect size: Δ median	-0.015	0.020	-0.060			-0.020	0.050	0.260	-0.070	0.050	-0.005	-0.005	0.180
N*	10	7	7	2	5	6	8	7	7	7	10	10	7
Sensitivity													
P* (*Threshold:* ≤ *0.016)	0.088	0.105	0.392	**	**	0.354	0.796	**0.008**	0.231	0.065	0.109	0.048	0.317
Effect size: Δ median	0.025	0.220	-0.020			0.020	0.050	0.160	-0.100	0.080	-0.035	-0.100	0.150
N*	10	7	7	2	5	6	8	7	7	7	10	10	7
MDA (cm)													
P* (*Threshold:* ≤ *0.016)	0.466	0.321	0.222	**	**	0.456	0.084	**0.001**	0.796	0.084	0.084	0.135	0.213
Effect size: Δ median	0.000	0.000	0.000			0.000	0.000	0.150	0.000	-0.010	0.000	0.020	0.030
N*	10	7	7	2	5	6	8	7	7	7	10	10	7

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

**** CT edited segmented more cases**

(-) means CTeCT is better)

Table S8: Paired T-test results comparing changes in DSC, MDA and sensitivity for both pairs of CT models (CTeCT and CTu). Bold values indicate statistically significant differences ($p \leq 0.016$). Insufficient successful segmentations were achieved by one of the models, this is noted (**), indicating the superior model.

Chapter 3 Dosimetric Impact of Contour Editing on CT and MRI Deep-Learning Autosegmentation for Brain OARs

Abstract

Purpose

To establish the clinical applicability of deep-learning organ-at-risk autocontouring models (DL-AC) for brain radiotherapy. The dosimetric impact of contour editing, prior to model training, on performance was evaluated for both CT and MRI-based models. The correlation between geometric and dosimetric measures was also investigated to establish whether dosimetric assessment is required for clinical validation.

Method

CT and MRI-based deep learning autosegmentation models were trained using edited and unedited clinical contours. Autosegmentations were dosimetrically compared to gold standard contours for a test cohort. D1%, D5%, D50%, and maximum dose were used as clinically relevant dosimetric measures. The statistical significance of dosimetric differences between the gold standard and autocontours was established using paired Student's t-tests. Clinically significant cases were identified via dosimetric headroom to the OAR tolerance. Pearson's Correlations were used to investigate the relationship between geometric measures and absolute percentage dose changes for each autosegmentation model.

Results

Except for the right orbit, when delineated using MRI models, the dosimetric statistical analysis revealed no superior model in terms of the dosimetric accuracy between the CT DL-AC models or between the MRI DL-AC for any investigated brain OARs. The number of patients where the clinical significance threshold was exceeded was higher for the optic chiasm D1% than other OARs, for all autosegmentation models.

A weak correlation was consistently observed between the outcomes of dosimetric and geometric evaluations.

Conclusions

Editing contours before training the DL-AC model had no significant impact on dosimetry. The geometric test metrics were inadequate to estimate the impact of contour inaccuracies on dose. Accordingly, dosimetric analysis is needed to evaluate the clinical applicability of DL-AC models in the brain.

***Keywords*:** Brain cancer, Organs at risk, Autosegmentation, Deep learning, MRI scans, CT scans, dosimetric evaluation

3.1 Introduction

With the advancement of technology and the increasing number of brain cancer patients(Soomro et al., 2023), clinical use of brain OARs deep learning autocontouring (DL-AC) models in the radiotherapy department has become attractive. It promises to improve the standardization and efficiency of organ-at-risk (OAR) contouring (Sherer et al., 2021). However, appropriate evaluation of contour quality and clinical acceptability is a challenge. Whilst geometric evaluation is straightforward, generalisable, and quantitative, its connection to clinical impact is difficult to establish (Baroudi et al., 2023, Harrison et al., 2022). The most popular methods for evaluating autosegmentation geometric quality are the Dice similarity coefficient (DSC) and distance-to-agreement metrics (DTA) (Baroudi et al., 2023). Overlap metrics can be sensitive to structure size and are frequently poor predictors of impact on clinically relevant dosimetric parameters (Harrison et al., 2022, Baroudi et al., 2023). Conversely, dosimetric analysis depends on local treatment protocols and clinical criteria, as well as individual patient anatomy and dose distributions, making it harder to draw general conclusions about model performance.

Researchers have reported that the optimal evaluation method depends on the aim of autosegmentation (Sherer et al., 2021). Where autocontours will be checked and edited by human operators, geometric or editing-time based analysis may be sufficient, although dosimetric analysis can inform operators about the clinical significance of editing and hence maximise time savings (Sherer et al., 2021). If contours will be used directly, e.g., in online adaptive therapy, with minimal or zero human intervention, a higher bar of both geometric and dosimetric testing is needed to ensure patient safety.

Therefore, to determine the clinical feasibility of autosegmentation for radiation treatment planning and delivery, several evaluation strategies, including geometric, dosimetric, and physician assessment, are ideally required (Sherer et al., 2021).

Dosimetric evaluation is most directly linked to clinical relevance (Harrison et al., 2022). However, this analysis requires treatment planning data (Harrison et al., 2022). Also, there is no standard method or agreed threshold of acceptability for dosimetric variation (Vinod et al., 2016). Accordingly, there is little research on

the dosimetric effects of contour variations between manual and autosegmentation, and even less on the dosimetric consequences of editing contours either before model training (as here) or post autosegmentation (Johnston et al., 2022).

Recent research (Sherer et al., 2021, Baroudi et al., 2023) raises questions about the correlation between common geometric measures, dose planning statistics, and clinical acceptability of OAR contours. Hence, it is difficult to establish whether a segmentation model is clinically usable in a specific clinical scenario, sufficiently limiting the risk of overexposing normal tissue and allowing the precise delivery of RT dose to targets.

This study investigates the dosimetric impact of autocontouring OARs in the brain, in the context of RT for common brain cancers. This work is built upon a geometric evaluation which was previously published and hence focusses on the clinically relevant dosimetric aspects (Alzahrani et al., 2023)**.** The correlation of dosimetry with the geometric accuracy of MRI and CT-based DL-AC models, established previously (Johnston et al., 2022, van Rooij et al., 2019, Zhu et al., 2020), is also addressed. Further, we determine the dosimetric impact of editing clinical contours to gold standard quality before training CT and MRI DL-AC models. Previous geometric analysis showed that DL-AC models trained with edited clinical contours successfully generated more segmentations than the models trained with unedited clinical contours. Also, editing contours on MRI before model training improved the geometric performance (Alzahrani et al., 2023)**.** However, generating gold standard contours is a time-consuming process that may require several clinicians, it severely limits the amount of high-quality labelled data available for model training. Also, there are no specific guidelines on the level of editing required, and the trade-off between training data quantity and quality. Whilst DL-AC delineations are usually checked or edited before use, poorer quality results from model involving limited unedited data may cause loss of efficiency and increase risk. However, if found editing contours to be unnecessary before training the DL-AC model, larger amounts of un-curated data could be a more efficient route to high-quality autosegmentation models for OARs in RT, particularly for MRI models, where limited data with equivalent sequences is available.

Understanding the impact of autosegmentation on RT dosimetry could also improve guidance for the critical assessment and editing of autocontours in clinical practice, maximising time-efficiency gains whilst avoiding an increased risk of toxicity from overexposing OARs.

Overdosing brain OARs can lead to, for example, visual and hearing deficits, making understanding of OAR segmentation accuracy a critical requirement in delivering high-quality RT (Scoccianti et al., 2015).

Previous studies of autocontouring for brain OARs using deep learning relied only on geometric assessment (Mlynarski et al., 2020, Wiesinger et al., 2021, Chen et al., 2019, Alzahrani et al., 2023). By evaluating the correlation between geometric and dosimetric measures, we aim to establish whether geometric assessment alone is sufficient to evaluate brain OAR autosegmentation tools or whether an additional dosimetric evaluation is also needed.

Regarding other treatment sites (thoracic, oesophageal, and head and neck), several studies have assessed the dosimetric impact of deep learning segmentation (Johnston et al., 2022, Zhu et al., 2020, van Rooij et al., 2019). Correlations between the geometric and dosimetric measures in thoracic and head and neck OARs have not been identified (Johnston et al., 2022, van Rooij et al., 2019). In contrast, a study investigating oesophageal OARs revealed that DSC and OAR dose had a statistically significant overall correlation, although this correlation was not always present at the level of individual patients or OARs (Zhu et al., 2020).

Finally, it is essential to consider the clinical significance of a dosimetric error. Whilst for a given test case, it is possible to say whether the dosimetric change caused a dose constraint to be exceeded, this is highly dependent on the details of the individual dose distribution and may not generalise to other cases. Here, we detail a pragmatic approach for determining the likely clinical significance of dose differences across a patient cohort, with a view to prospective clinical use of the model.

3.2 Materials and Methods

3.2.1 Dataset and clinical protocol

As this study was built based on previously published geometry study, you can find a summary of essential details information such as data preparation, OAR selection, gold standard contours, and image acquisitions in that earlier publication (Alzahrani et al., 2023).

A computer-generated simple-random list was used to select randomly 60 Brain cases from a retrospective clinical cohort treated in our institution over the past five years. 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. This UK ethics committee approval indicates that our study is conformant with the Declaration of Helsinki, the UK Policy Framework for Health and Social Care Research and the EMA guidelines on Good Clinical Practice. The data for training and testing was randomly chosen: 80% for training (n=48) and 20% for testing (n=12), which is the most popular split ratio (80/20) (Alzahrani et al., 2023). As the model used was a commercially approved model, on which we did not perform hyperparameter tuning, there was no need for in-training validation. More information about the available training parameter can be found in the supplementary information.

Using the same dataset, two CT autosegmentation models were trained with a total of 47/48 cases (one case was excluded due to missing data), and three MRI autosegmentation models were trained using 32 cases (16/48 cases were excluded due to inconsistent MRI slice thickness) (Alzahrani et al., 2023). For testing, three test cases were excluded. Two CT test cases were excluded because no MRI images were associated with them (n=10 cases) and one additional MRI test case was excluded (n=9 cases) due to the use of different MRI sequence (Alzahrani et al., 2023). In addition, All test cases were treated for either high-grade or low-grade glioma using volumetric modulated arc therapy (VMAT). Total RT dose was 60 Gy in 30 fractions, for glioblastoma multiforme (GBM) and grade III glioma (protocol A), or total RT dose was 54 Gy in 30 fractions for low-grade glioma (protocol B). The clinical OAR dose constraints are shown in Table 3.1. D1, 5, 50% denotes a minimum dose to the most exposed 1, 5, or 50% of the OAR volume, respectively.

Table 3.1: Dose constraints for Glioma Radical-Primary VMAT (60 Gy in 30# and 54 Gy in 30#)

OARs	Dose constrains	Dosimetric metrics
Brainstem	54 Gy	D5%
Lenses	6 Gy	D1%
Optic Chiasm	54 Gy	D1%
Optic Nerves	54 Gy	D1%
Orbit	45 Gy	D1%
Lacrimal Glands	30 Gy	D1%
Pituitary	45 Gy	Max Dose
Cochlea	45 Gy	D50%

3.2.2 Deep learning autosegmentation training

The OAR contours used for clinical treatment were based on a combination of the anatomy as seen on co-registered MRI (specifically brainstem, optic chiasm, and intra-cranial component of the optic nerves) and radiotherapy planning CT (specifically extra-cranial portions of the optic nerves, lenses, globes, cochlea, and lacrimal glands). From these, the contours used in this project were derived:

- Unedited clinical contours as above (used for both CT and MRI -based autosegmentation models, termed the CT unedited and MRI unedited models, CTu and MRIu, respectively- please see next paragraph)
- Clinical contours edited to correspond with a departmental contouring guide (the ‘gold standard’) and edited to be entirely based on CT anatomy (used for CT and MRI -based autosegmentation models, termed the edited models CTeCT and MRIeCT- please see next paragraph)
- Clinical contours edited to correspond with a departmental contouring guide and edited to be entirely based on MRI anatomy alone (used for the MRI-based autosegmentation model termed the MRI edited model, MRIeMRI- please see next paragraph)

The same MRI and CT DL-AC models that were built for geometric evaluation were used for dosimetric evaluation as follows:(Alzahrani et al., 2023)

DL-AC models were trained using a 3D U-net (Çiçek et al., 2016) architecture (RayStation 11A, RaySearch Laboratories AB, Stockholm, Sweden). Five separate autosegmentation models (two CT- and three MRI-based) were trained: i) CT-based, using the unedited clinical contours (CTu), and ii) CT-based using contours edited to gold standard based on CT anatomy (CTeCT). Both contour sets were rigidly registered to T1-weighted gadolinium-enhanced MRI (T1w-Gd MRI) to train the iii) MRI-based model using the unedited clinical contours (MRIu), and the iv) MRI-based model using the CT edited contours (MRIeCT). Finally, an MRI-based model that used these contours edited based on MRI anatomy (MRIeMRI). After training, all the autosegmentation models were used to generate automatic contours on the test cohort.

3.2.3 Dosimetric Evaluation

Dose statistics were computed (Raystation 11A) to compare the CT and MRI autosegmentation models with gold standard contours in each modality, where clinical contours were edited based on each modality's anatomy in this test cohort (i.e., CTeCT and MRIeMRI). Dose evaluation for MRI autosegmentation was performed by copying the CT dose distribution to T1w-Gd MRI via rigid image registration.

The statistical significance of differences in dose metrics due to autosegmentation models was evaluated using a paired two-tailed Student's t-test. The three MRI-based models were compared statistically, as were the two CT-based models. The Bonferroni corrected statistical significance threshold was $p \leq 0.01$ (0.05/3) and ≤ 0.05 for the MRI and CT dosimetric evaluations, respectively. *More information is available in the supplementary materials about dosimetric evaluation and statistical analysis.*

3.2.4 Clinical evaluation

The question of 'what is a clinically significant dose difference?' is challenging. If an OAR dose is close to or at tolerance, any change could be significant, but we would normally accept a 2-3% tolerance due to other uncertainties in dose calculation and setup (for example). However, that arbitrary 2-3% tolerance would be overly restrictive if the OAR dose were said 30% below tolerance. Thus, the clinical significance of dosimetric differences for each OAR was determined

using a pragmatic approach I developed under the guidance of an experienced radiation oncologist.

For first-order OARs (where the dosimetric tolerance is a hard limit for RT dose planning) with near maximal dose statistics (e.g., D1% or D5%), the average dosimetric headroom between the gold standard contour dose and the tolerance dose in Table 3.1 was computed.

50% of the average dosimetric headroom was used as the clinical significance threshold for these first-order OARs: brainstem, orbits, optic chiasm, and optic nerves. A case was considered clinically significant if the dose changes between the gold standard contour and autosegmentation was more than half the average dose headroom in either direction (Figure 3.1a).

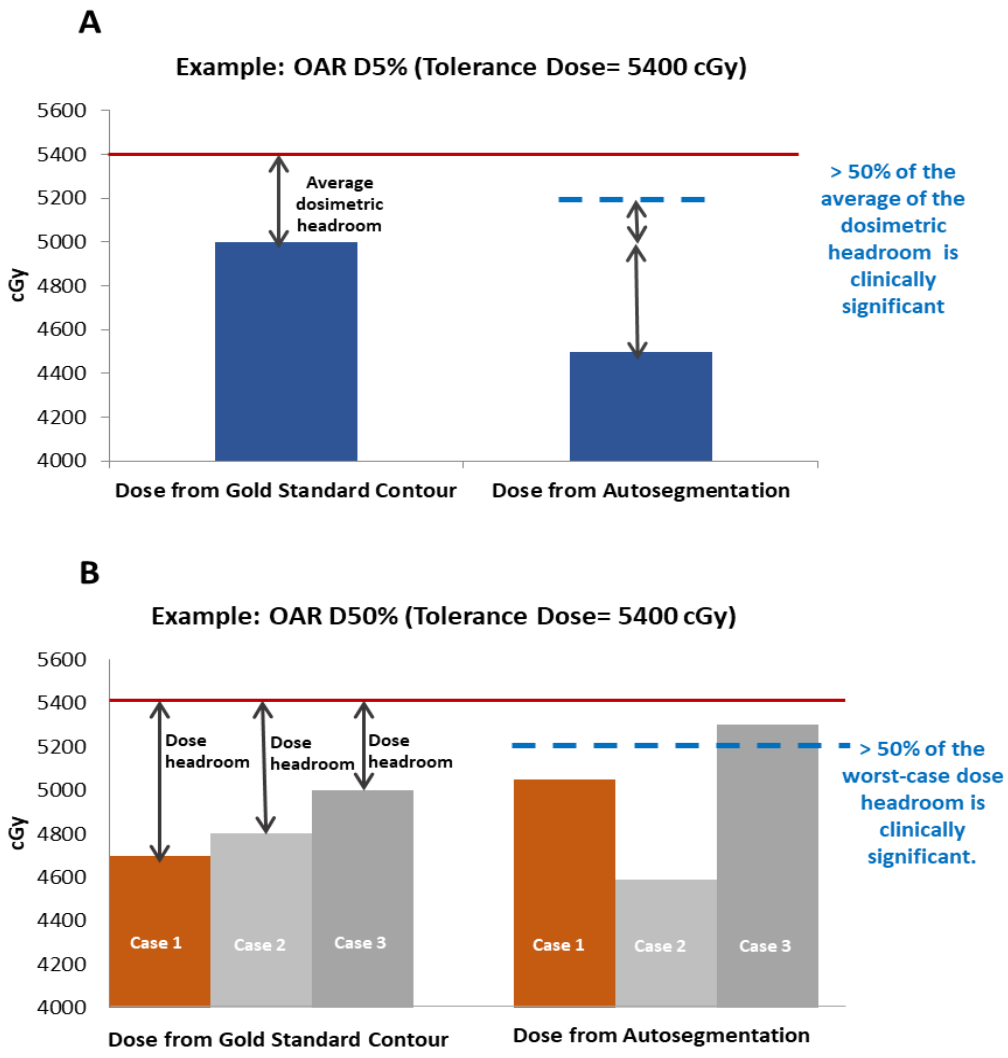


Figure 3.1: Clinical dose evaluation: a) the average metric approach which relates to the average dose change, b) the worst-case scenario approach.

For second-order OARs (where dose tolerances are optimal, rather than mandatory) where the dosimetric statistics are mean-dose-like (e.g., cochlea D50%), the approach was based on the worst-case scenario in the test cohort. The worst-case scenario was defined as the case with the least headroom to the tolerance dose, using the gold standard OAR contours. 50% of the worst-case scenario headroom was used as the clinical significance threshold. A case was considered clinically significant if the dose changes between gold standard contours and autosegmentation was more than half of this threshold in either direction (Figure 3.1b).

For other second-order OARs, the clinical significance of the dosimetric change was more challenging to define. The evaluation was therefore based on a comparison of relative model dosimetric performance as above, rather than any clinical significance threshold. This approach was applied for lenses, lacrimal glands, and pituitary gland as they were treated in some cases to more than the optimal tolerance dose, which would result in a negative clinical significance threshold by the methods described above.

All cases identified as having clinically significant dosimetric changes were visually reviewed in the treatment planning system with an experienced clinical oncologist to identify the cause (e.g., proximity of an OAR to a dose gradient).

By aligning our approach with the perception of an experienced radiation oncologist, we enhanced the reliability of this clinically significant metric in identifying the potential clinically significant cases. As we mentioned in the introduction, there is no standard method or agreed threshold of acceptability for dosimetric variation.

3.2.5 Correlation between the geometric and dosimetric output:

Pearson's Correlation Coefficient (r) (Mukaka, 2012) was applied to measure correlations between geometric test metrics (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) and absolute percentage dose change for each autosegmentation model.

3.3 Results

3.3.1 Overall effect of using autosegmentation vs. gold standard human contouring on dosimetry

Figures 3.2, 3.3, and 3.4 represent the overall patterns of dosimetric change for CT and MRI DL-AC models relative to the gold standard contours. The lacrimal glands are presented separately due to the relatively larger dose changes. The dosimetric change for the MRI autosegmentations vs. gold standard contour was greatest in the lacrimal glands D1%, followed by the optic nerves D1% (Table 3.2) (Figures 3.2 and 3.4). The average absolute dosimetric change for the lacrimal glands D1% and optic nerves D1% varied from 25% (MRIeMRI) to 143% (MRIeCT) and 9% (MRIeMRI) to 20% (MRIu and MRIeCT), respectively (Table 3.2). The remaining OARs had less dosimetric change relative to the gold standard contour, ranging from 1% to 12%. (Table 3.2)

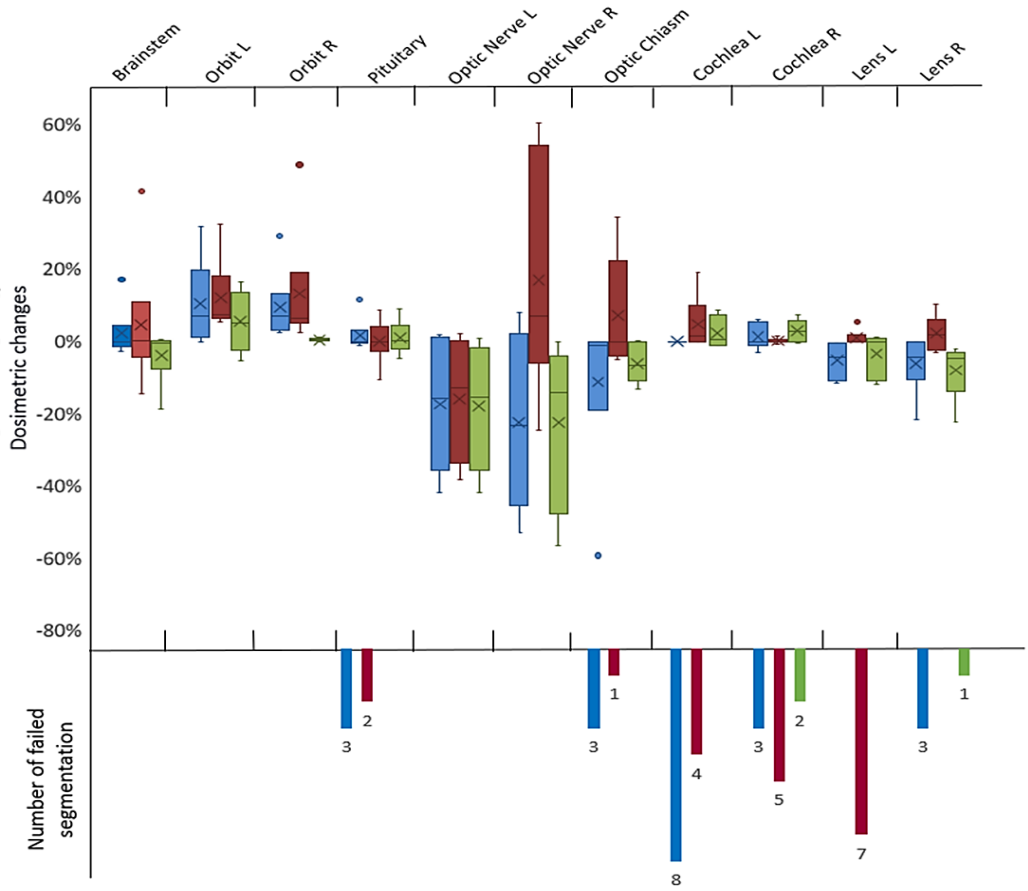


Figure 3.2: Distribution of the dosimetric change of all OARs delineated by MRI DL-AC models (excluding lacrimal glands). The number of failed segmentations is when autosegmentation model failed to produce structures. In some cases, the

small dosimetric change is affected by the number of failed cases such as cochlea, pituitary, and lens L. MRIu is shown in blue, MRIeCT in red, and MRIeMRI in green.

The greatest dosimetric change for the CT DL-AC vs. gold standard contour was observed for the right lens D1% and optic chiasm D1% for the CTu and CTeCT models, respectively (Table 3.3 and Figure 3.3). The average absolute dosimetric change was 57% (CTu) for the right lens D1% and 18% (CTeCT) for the optic chiasm D1%. The marked dosimetric change was also reported for L and R orbits D1% delineated by the CTu model (21%, 25%) and L and R optic nerves D1% delineated by the CTeCT model (14%, 15%) (Table 3.3). The remaining OARs had less dosimetric change relative to the gold standard contour, ranging from 1% to 17% (Table 3.3).

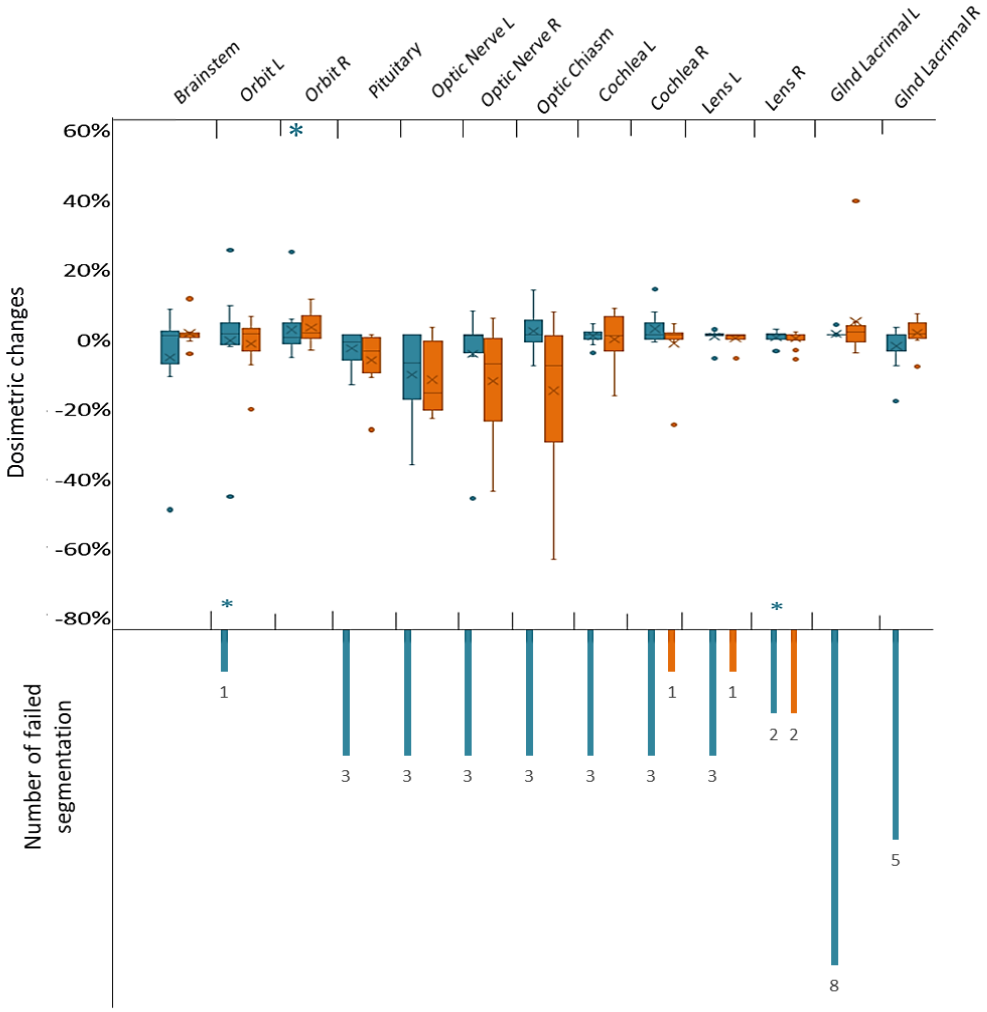


Figure 3.3: Distribution of the dosimetric changes of all OARs delineated by the CT DL-AC relative to the gold standard contour. The number of failed

segmentations is when autosegmentation model failed to produce structures. In some cases, the small dosimetric change is affected by the number of failed cases. CTu is shown in turquoise, while CTeCT is in orange. (*) indicates that outliers have been removed from the plot for clarity.

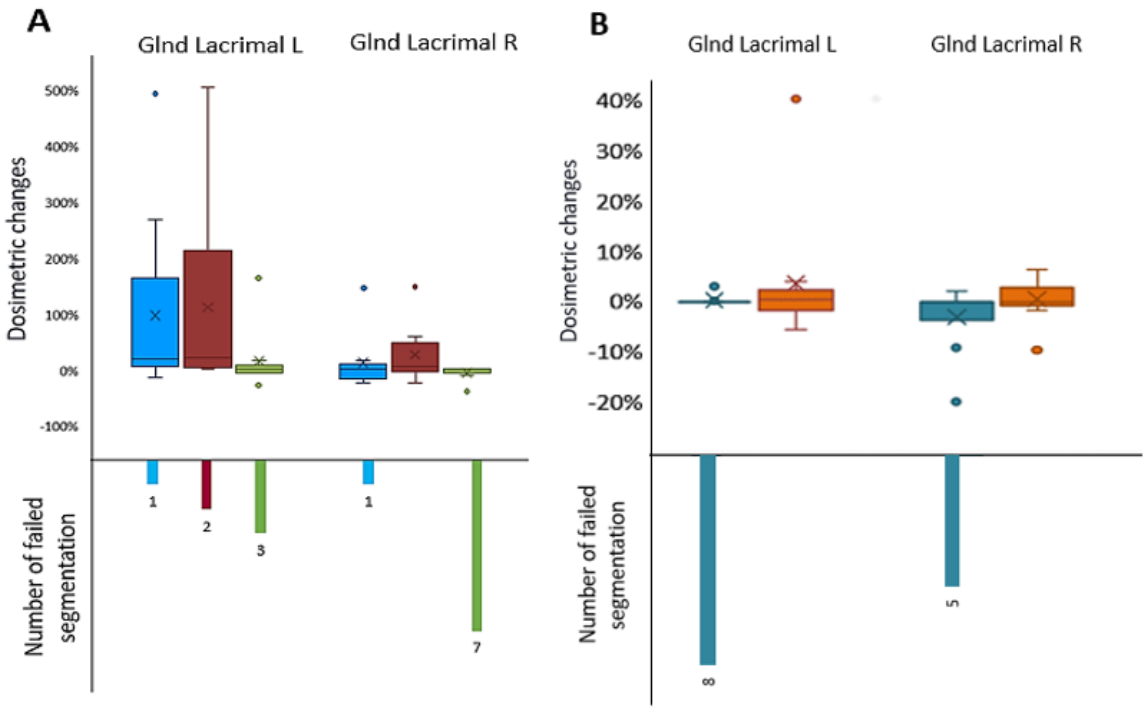


Figure 3.4: Distribution of the dosimetric change of the lacrimal glands segmented by a) MRI DL-AC models and b) CT DL-AC models relative to the gold standard contour. MRIu is shown in blue, MRIeCT in red, while MRIeMRI in green, CTu is shown in turquoise, while CTeCT in orange.

3.3.2 Impact of editing

For orbits, optic nerves, and optic chiasm, the MRIeMRI model showed less average dosimetric changes than other MRI models (Table 3.2 and Figure 3.2). However, differences between MRI DL-AC model dosimetry were not statistically significant, except in the right orbit, where a statistically significant effect was found comparing the MRIu and MRIeMRI models (P= 0.012, effect size (Δ median dosimetric change) = 7%). However, it was clinically insignificant (See Supp.info Table S1).

The CTeCT model demonstrated smaller average dosimetric changes, relative to the gold standard, than the CTu model for the following structures: orbits, lenses, and brainstem (Table 3.3 and Figure 3.3). Again, however, dosimetric differences

between the CT DL-AC models were not statistically significant (See Supp.info Table S2). In 3 cases, the CTu model generated incorrectly located segmentations (DSC = 0) for several of these OARs. These cases were visually qualitatively assessed in the treatment planning system.

Table3.2: The absolute average dosimetric change between the MRI autosegmentations and gold standard contour.

	Brainstem D5%	Orbit L D1%	Orbit R D1%	Optic Nrv L D1%	Optic Nrv R D1%	Optic Chiasm D1%	Cochlea L D50%	Cochlea R D50%	Pituitary Max dose	Lacrimal L D1%	Lacrimal R D1%	Lens L D1%	Lens R D1%
MRIu													
Δ absolute Average dosimetric change	3%	9%	10%	20%	20%	12%	1%	3%	3%	113%	29%	4%	7%
N*	9	9	9	9	9	6	1	6	6	8	8	9	6
MRIeCT													
Δ absolute Average dosimetric change	7%	10%	11%	15%	20%	9%	6%	1%	4%	143%	33%	3%	4%
N*	9	9	9	9	9	8	5	4	7	7	9	2	9
MRIeMRI													
Δ absolute Average dosimetric change	3%	6%	1%	9%	18%	4%	3%	10%	3%	36%	25%	3%	7%
N*	9	9	9	9	9	9	9	7	9	6	2	9	8

N represents the number of the successful segmentation by each model.*

Table 3.3: The absolute average dosimetric change between the CT autosegmentations and gold standard contour.

	Brainstem D5%	Orbit L D1%	Orbit R D1%	Optic Nrv L D1%	Optic Nrv R D1%	Optic Chiasm D1%	Cochlea L D50%	Cochlea R D50%	Pituitary Max dose	Lacrimal L D1%	Lacrimal R D1%	Lens L D1%	Lens R D1%
CTu													
Δ absolute Average dosimetric change	9%	21%	25%	17%	10%	6%	2%	4%	6%	2%	8%	2%	57%
N*	10	9	10	7	7	7	7	7	7	8	5	7	8
CTeCT													
Δ absolute Average dosimetric change	2%	5%	4%	14%	15%	18%	5%	4%	8%	6%	3%	1%	2%
N*	10	10	10	10	10	10	10	9	10	10	10	9	8

N* represents the number of the successful segmentation by each model.

3.3.3 MRI vs CT DL-AC - effect on dosimetry

Differences in dosimetric changes relative to the gold standard contour between the CT and MRI DL-AC models, were most noticeable in the lacrimal glands D1% (Figure 3.4). The dosimetric change for lacrimal glands D1% delineated by the CT DL-AC models vs gold standard contour was considerably smaller than that of the MRI DL-AC models. Additionally, the MRIeMRI model failed to segment the lacrimal glands in 9 cases.

3.3.4 Correlation between the geometric and dosimetric evaluations

All models showed a weak correlation between absolute dosimetric change and geometric evaluation metrics. Negative correlations were observed between DSC and absolute dosimetric change and between sensitivity and absolute dosimetric change ($r \leq -0.40$ and $r \leq -0.38$, respectively). A positive correlation was observed between mean DTA and absolute dosimetric change ($r \leq 0.54$) (Table 3.4). None of the observed correlations reached statistical significance at $p=0.05$. *All results related to the geometric output used for this evaluation can be found in the previous published work.(Alzahrani et al., 2023)*

Table 3.4: Correlation between geometric and dosimetric outputs.

Autosegmentation Models	Absolute dosimetric change and DSC	Absolute dosimetric change and sensitivity	Absolute dosimetric change and mean DTA
MRIeCT	r = -0.299	r = -0.256	r = 0.262
MRIeMRI	r = -0.402	r = -0.381	r = 0.328
MRIu	r = -0.304	r = -0.255	r = 0.543
CTeCT	r = -0.343	r = -0.378	r = 0.288
CTu	r = -0.386	r = -0.359	r = 0.106

3.3.5 Clinical significance of autosegmentation models on dosimetry

3.3.5.1 First-order OARs

Tables 3.5 and 3.6 demonstrate the number of clinically significant cases according to the definitions outlined above, and the average dosimetric change relative to the gold standard contour. In both CT DL-AC and MRI DL-AC, the number of cases that exceeded the clinical significance threshold for optic chiasm D1% was higher than for other first-order OARs (n≥ 4 cases). In both modalities, models trained with edited contours based on CT scans (MRIeCT and CTeCT) demonstrated the largest frequency of clinically significant errors (n=7 with Δ average dose= 590.0 and 1376.1 cGy, respectively) (Table 3.6).

Only one clinically significant case was observed for the brainstem D5% in each MRI DL-AC model (n= 3 cases, Δ average dose= 203.5 cGy) (Table 3.5). However, the MRIeCT exhibited greater dosimetric change relative to the gold standard contour than the MRIeMRI and MRIu models (Table 3.5).

Training the CT DL-AC model with edited contours, on the other hand, reduced the frequency of clinically significant dosimetric errors for the brainstem D5% and demonstrated smaller dosimetric changes relative to the gold standard contour compared to the CTu model (n= 3 cases, Δ average= 246.6 cGy) (Table 3.6).

3.3.5.2 Second-order OARs

Amongst the second-order OARs (waterfall plots - supplementary information Figures S1-S2), the lacrimal glands demonstrated the largest dosimetric change in the MRI DL-AC models. In the worst case, the dose was changed in the left lacrimal gland by 505% for the MRIeCT model (Figure1S:C), relative to the gold standard. On the other hand, the right lens had the largest dosimetric change in the CT DL-AC (446% worst-case for the CTu model). (Figure 2S: b). Otherwise, the dosimetric changes associated with DL-AC compared to gold standard were generally lower for second-order OARs. For CT DL-AC, these ranged from 0% to 40%, whereas they ranged from 0% to 22% for the MRI DL-AC (Figure 1S (a-e) and 2S(a-e)).

Table 3.5: Significant clinical cases and their average of the dosimetric change compared to gold standard.

	Threshold (cGy)		MRIeCT		MRIeMRI		MRIu	
OARs	**protocol A**	**protocol B**	**A (n=5 cases)**	**B (n=4 cases)**	**A (n=5 cases)**	**B (n=4cases)**	**A (n=5 cases)**	**B (n=4 cases)**
Brainstem D5%	54.222	660.410	(n=1) 411.3 cGy	(n=0) -	(n=1) 58.9 cGy	(n=0) -	(n=1) 140.3cGy	(n=0) -
Optic Nrv L D1%	521.292	1031.981	(n=1) 869.8 cGy	(n=0) -	(n=2) 611.3 cGy	(n=0) -	(n=2) 1269.393 cGy	(n=1) 1192 cGy
OpticNrv R D1%	1194.213	1063.127	(n=0) -	(n=1) 1313.7 cGy	(n=0) -	(n=1) 1123.8 cGy	(n=0) -	(n=1) 1073.1 cGy
Optic Chiasm D1%	50.231	286.879	(n=5) 164.8 cGy	(n=2) 1015.2 cGy	(n=3) 298.1 cGy	(n=1) 722.0 cGy	(n=3) 160 cGy	(n=1) 1745.9 cGy
Cochlea L D50%	548.633	1001.294	(n=1) 636.4 cGy	(n=0) -	(n=0) -	(n=0) -	(n=0) -	(n=0) -
Total clinically significant cases			**11**		**9**		**10**	

Table 3.6: Significant clinical cases and their average of the dosimetric change compared to the gold standard.

	Threshold (cGy)		CTu		CTeCT	
OARs	**protocol A**	**protocol B**	**A (n=7 cases)**	**B (n=3 cases)**	**A (n=7 cases)**	**B (n=3 cases)**
Brainstem D5%	171.161	43.310	(n=3) 447.9 cGy	(n=2) 279.4 cGy	(n=2) 395.8 cGy	(n=1) 97.3 cGy
Orbit L D1%	1806.533	2038.483	(n=1) 2572.2 cGy	(n=0) -	(n=0) -	(n=0) -
Optic Nrv L D1%	769.175	821.592	(n=1) 768 cGy	(n=1) 1605.9 cGy	(n=2) 996.2 cGy	(n=1) 1025.7cGy
Optic Nrv R D1%	1092.199	561.628	(n=1) 1226.1 cGy	(n=0) -	(n=1) 1173.0 cGy	(n=0) -
Optic Chiasm D1%	186.969	161.208	(n=3) 402.527 cGy	(n=1) 503.8 cGy	(n=4) 825.6 cGy	(n=3) 1926.5cGy
Cochlea L D50%	287.110	79.972	(n=0) -	(n=0) -	(n=1) 747.9 cGy	(n=0) -
Total clinically significant cases			**13**		**16**	

3.4 Discussion

This study investigated the dosimetric impact of clinical contour editing before training MRI and CT DL-AC models for brain OARs to establish clinical applicability. This study also examined the correlation between geometric and dosimetric outcomes, in order to guide centres as to whether the geometric assessment alone is sufficient to evaluate and commission DL-AC models in radiotherapy or whether a dosimetric evaluation is also necessary.

Except for the right orbit, when delineated by the MRI models, the dosimetric statistical analysis revealed no superior model between the CT DL-AC models or between the MRI DL-AC in terms of the dosimetric accuracy for any investigated brain OARs (Table S1 and S2). The significant finding for the right orbit likely results from a slight registration inaccuracy in mapping CT-derived evaluation contours to MRI, rather than any feature of the DL-AC model. As a result, editing contours for brain OAR structures on the CT or MRI scans before training the model had no significant effect on OAR dosimetry. The lack of superiority indicates that both models perform well dosimetrically. This occurs for two main reasons. Firstly, doses in brain RT for GBM are relatively homogeneous, meaning that most differences between these complex OAR contours lie in either uniformly high or low dose regions. Only occasionally will a contouring difference occur on a high dose gradient, leading to a significant dosimetric impact. Secondly, the metrics used clinically tend to be of the 'near-maximal dose' type, which are insensitive to contouring changes which occur in regions of lower dose. This is in contrast to metrics such as mean doses or V20 Gy, which might be used in the thorax for example.

Clinical dosimetric evaluation was performed as a secondary assessment of potential clinical impact, using the average metric approach and the worst-case scenario approach.

The number of patients that exceeded the derived clinical significance threshold for optic chiasm D1% was higher compared to other OARs (brainstem D5%, orbits D1%, optic nerves D1%, and cochlea D50%) in both modalities.

The absolute dosimetric changes of the optic chiasm D1% relative to the gold standard of the clinically significant cases were ≤ 67% and ≤ 59% across all the CT and MRI models, respectively. The DSC, sensitivity and mean DTA scores were ≤ 0.37, 0.54, 1.44 cm and ≤ 0.74, 0.83, 0.77 cm for the CT and MRI models,

respectively (Alzahrani et al., 2023). In comparison with the CTeCT model, the CTu model shows a smaller number of significant cases with more acceptable percentage change, which is surprising at first sight since the edited contours should be more closely correlated with the underlying CT anatomy. However, as the optic chiasm is very poorly visualised on CT, the segmentation model relies not on the correlation with imaging features, but more on the consistency of the shape and location of the optic chiasm, to predict its segmentation. In the unedited data (used for the CTu model), this consistency is high, due to the original clinical contours being based on MRI rather than CT anatomy (see section 2.2), enabling the model to learn. By editing the optic chiasm on CT anatomy alone (used for the CTeCT model), this consistency is degraded, and the correlation with image features is not improved, as there are none present on CT. Hence, CTeCT performs worse, as it struggled to learn a consistent shape and location for the optic chiasm and hence produced a high dosimetric change with more significant cases compared to CTu.

On the other hand, MRIeMRI showed a more acceptable dosimetric change than other MRI models, showing the benefits of editing optic chiasm on MRI prior to model training. Based on visual inspection of the optic chiasm in the treatment planning system, the level of dose discrepancy was independent of dose gradient location and appeared well correlated to the geometric error. This is expected because optic chiasm is a small structure and has a complicated shape. The model failed to delineate all the optic chiasm on each slice accurately. This indicates that post-segmentation editing may be required for optic chiasm.

Regarding the Brainstem D5%, in a comparison with the CTu model, the CTeCT model demonstrated a lower number of significant changes for brainstem D5% (3 significant cases) (Table 3.6) with less dosimetric change relative to the gold standard (≤ 11% in either direction). Notably, the CTu model segmented several OARs in completely the wrong location, leading to the significant increase in mean dosimetric errors for the brainstem (and orbits). Editing prior to model training resolved these failures.

On the other hand, the number of clinically significant cases for the brainstem D5% for the MRI was just one for each model (Table 3.5). The dosimetric differences compared to the gold standard contour were ≤ 19 % in either direction, but the geometric error was low (DSC and sensitivity scores ≥ 0.89 and 0.85, while mean DTA score ≤ 0.10 cm) (Alzahrani et al., 2023). This dosimetric

error appears clinically significant because in most clinical cases, D5% Brainstem is at or near to the PTV, so even a slight difference is significant.

It was noticeable that the geometric error of the clinically significant cases for D5% brainstem in both modalities was generally low (DSC score ≥ 0.8) (Alzahrani et al., 2023), except for CTu failure cases mentioned above. On visual assessment, the superior part of the brainstem was found to overlap PTV or at a distance, resulting in significant dose gradients (Figure 3.5: a). These results show that clinical dosimetric evaluation is essential in some cases, and the geometric evaluation alone is insufficient to demonstrate the clinical utility of autosegmentation, due to the extreme inhomogeneity of dose distributions. Geometric errors only translate to dosimetric errors where they overlap steep dose gradients. Similarly, a recent study evaluating the dose for the thoracic OARs delineated by CNN-based autosegmentation found that significant dose-volume variations were more strongly correlated with areas of high-dose gradient than geometric segmentation errors (Johnston et al., 2022). Moreover, previous studies have identified significant dosimetric differences between test and standard segmentation observed for OARs with high-dose gradients, even when geometric measures show good overlap (Harrison et al., 2022). On the other hand, OARs within homogeneous dose regions may reveal poor volumetric agreement but minimal dosimetric differences (Harrison et al., 2022). Accordingly, the superior part of the brainstem autosegmentation must be corrected when needed due to the poor performance of the contouring models in this portion and due to its proximity to PTV in many of the GBM cases (Figure 3.5: a). For modern, highly conformal arc therapies, the PTV is often a good surrogate for the location of high-dose gradients, but care should be taken with fixed beam angle treatments, where steep gradients may exist far from the PTV.

It was noted that clinically significant dosimetric changes for optic nerves were mostly reductions in dose relative to the gold standard for two main reasons. First, after reviewing the treatment planning system, the CT models were found to have failed to correctly identify all the boundaries on each slice, while the MRI models failed to identify the posterior limit of the optic nerves (Figure 3.5: b). In either case, the segmentation was incomplete, resulting in reduced dose statistics. Second, a considerable reduction in dose was noticed in some other cases, even though there was relatively good visual agreement between the generated

contours and the gold standard. In these cases, part of the gold standard contour was near PTV, whereas the DL-AC contour was not.

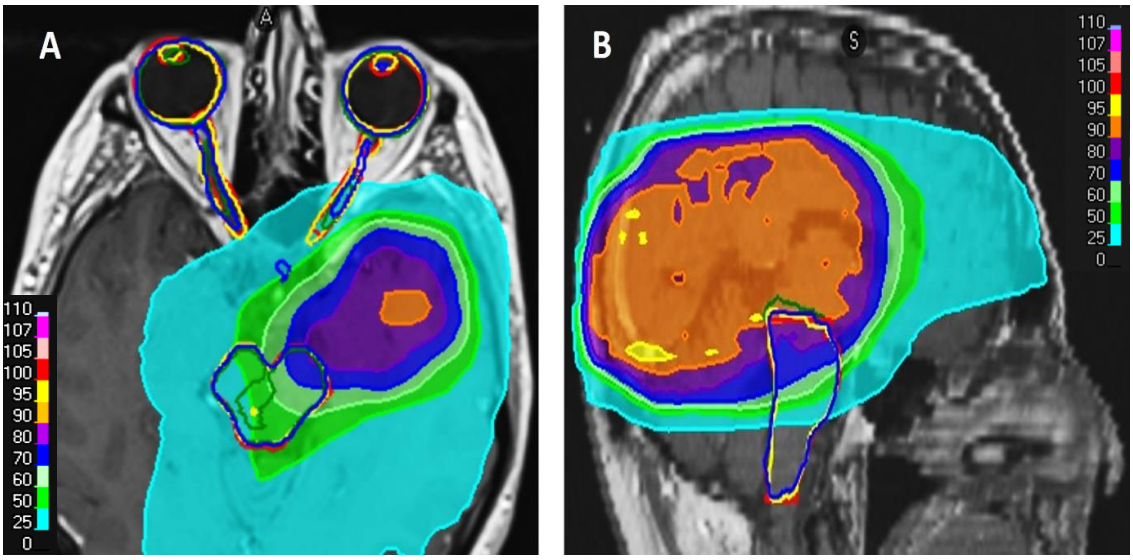


Figure 3.5: a) axial and b) sagittal T1w-Gd MRI with overlying dose distribution, showing examples of different geometrical changes of predicted MRI autosegmentations compared to the gold standard. Red outline represents the gold standard contour. The MRIeMRI contours are depicted as yellow outlines, the CTeMRI contours as green outlines, and MRIu contours as blue outlines. The colourwash represents the percentage dose distribution, relative the prescription dose, according to the inset colorbar. The dosimetric impact for a given geometric error is large only in high-dose gradients (e.g., as seen on sagittal image, the dosimetric impact of the yellow contour, relative to the gold standard (red) is 7% (411 cGy), as there is a steep dose gradient, whereas the dosimetric difference of the green contour relative to gold standard (red) is only 1% (38 cGy), as it lies in a more homogenous region of dose.) Overall, this dependence on dose gradient leads to the observed weak overall correlation of dosimetric impact and geometric error.

Regarding the small structures (lenses, lacrimal glands, and pituitary), lacrimal glands delineated by MRI models demonstrated a remarkably high dosimetric change relative to the gold standard. This correlated with geometric inaccuracy due to the difficulty visualising this organ using T1-w MRI scans (Alzahrani et al., 2023). Uniquely, on CT, these glands are more visible than on MRI. However, CTu model showed that the right lens had the largest dosimetric change (446% worst-case). This was a failed segmentation, which falsely identified a region of

the brain far from the right lens, instead of simply producing no contour. The dose in that region was approximately 4.5x higher than in the lens, as it was by chance on the PTV boundary, leading to this extreme result.

This study found a weak correlation between the geometric and dosimetric outcomes in both modalities. The correlation direction of the geometric and dosimetric results followed our expectations. The absolute percentage dose difference was negatively correlated with the sensitivity and DSC scores, and positively correlated with the mean DTA. Higher DSC and sensitivity scores indicate improved geometric performance, whereas higher mean DTA scores indicate larger geometric (and hence dosimetric) errors.

This suggests geometric test metrics were insufficient to predict the effect of contour inaccuracies on dose, due primarily to variability in the location of dose gradients. Also, geometry test metrics such as DSC can be impacted by the structure size and are often a poor indicator of clinically significant dosimetric impact (Baroudi et al., 2023, Harrison et al., 2022).

A recent study examined the correlation between geometrical measures and dose-volume variations for thoracic OARs (Johnston et al., 2022). Researchers found no significant correlation between them (Johnston et al., 2022). The weak correlation identified in this current work may indicate that dose distributions exhibit more variance in the thorax than the brain; hence, geometric performance was found to be an insufficient metric for clinical utility. Consequently, it is crucial also to perform dosimetric tests to demonstrate the clinical applicability and accuracy of autosegmentation models.

The fact that specific organs are prone to exhibiting large geometric errors, and the likelihood that these are in high-dose gradient regions, potentially allows human operators to prioritise their contour editing to the critical organs that are likely to be in the vicinity of high-dose gradients, further improving efficiency in checking contours, and avoiding spending time editing geometric errors which will not translate to dosimetric errors.

This study has certain limitations. The relatively small number of cases analysed makes it possible that outlier cases have not been captured (e.g., where a small OAR lies very close to a high-dose gradient). Additionally, the clinician's time editing contours needs to be investigated to measure efficiency savings from DL-AC.

3.5 Conclusion

As technology advances and the number of brain cancer patients increases, clinical use of brain OARs DL-AC models in the radiotherapy department becomes attractive. However, adequate assessment of contour accuracy and clinical applicability are essential. In this study, the dosimetric impact of autocontouring OARs in the brain was investigated. Specifically, the dosimetric impact of editing clinical contours to gold standard quality before training CT and MRI DL-AC models was assessed. Moreover, the correlation of dosimetry with geometric accuracy of MRI and CT-based DL-AC models was determined. Generally, we found that editing the clinical contour before training the model had no statistically significant impact on the dosimetry, despite clear geometric effects. However, by assessing the clinical significance of dosimetric changes as a secondary assessment of potential clinical impact, some geometric errors resulted in clinically significant dosimetry changes, despite the small underlying geometrical errors.

Our results suggest that an MRIeMRI model could be used clinically for treatment planning despite some structures requiring manual contour editing. This is due to its generated segmentation generally showing less dosimetric change relative to the gold standard contours for most of the OARs. It also produced the fewest clinically significant dosimetric errors, indicating that the improvements in geometric performance can lead to dosimetric improvements in specific cases.

Generally, a weak, and statistically insignificant correlation between the geometric and dosimetric outcomes for brain OARs in both modalities was found. Accordingly, geometric test metrics are insufficient to establish the impact of autocontouring inaccuracies on RT dose, mainly due to the variability in the location of dose gradients relative to OARs and geometric errors. For robust evaluation and commissioning of autocontouring, both geometric and dosimetric evaluation is recommended.

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).

Conflict of Interest Statement

No conflict of interest.

3.6 References

ALZAHRANI, N., HENRY, A., CLARK, A., MURRAY, L., NIX, M. & AL-QAISIEH, B. 2023. Geometric evaluations of CT and MRI based deep learning segmentation for brain OARs in radiotherapy. *Phys Med Biol,* 68.

BAROUDI, H., BROCK, K. K., CAO, W., CHEN, X., CHUNG, C., COURT, L. E., EL BASHA, M. D., FARHAT, M., GAY, S., GRONBERG, M. P., GUPTA, A. C., HERNANDEZ, S., HUANG, K., JAFFRAY, D. A., LIM, R., MARQUEZ, B., NEALON, K., NETHERTON, T. J., NGUYEN, C. M., REBER, B., RHEE, D. J., SALAZAR, R. M., SHANKER, M. D., SJOGREEN, C., WOODLAND, M., YANG, J., YU, C. & ZHAO, Y. 2023. Automated Contouring and Planning in Radiation Therapy: What Is 'Clinically Acceptable'? *Diagnostics (Basel),* 13.

CHEN, H., LU, W., CHEN, M., ZHOU, L., TIMMERMAN, R., TU, D., NEDZI, L., WARDAK, Z., JIANG, S., ZHEN, X. & GU, X. 2019. A recursive ensemble organ segmentation (REOS) framework: application in brain radiotherapy. *Phys Med Biol,* 64**,** 025015.

ÇIÇEK, Ö., ABDULKADIR, A., LIENKAMP, S. S., BROX, T. & RONNEBERGER, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International conference on medical image computing and computer-assisted intervention, 2016. Springer, 424-432.

HARRISON, K., PULLEN, H., WELSH, C., OKTAY, O., ALVAREZ-VALLE, J. & JENA, R. 2022. Machine Learning for Auto-Segmentation in Radiotherapy Planning. *Clin Oncol (R Coll Radiol),* 34**,** 74-88.

JENA, R., KIRKBY, N. F., BURTON, K. E., HOOLE, A. C., TAN, L. T. & BURNET, N. G. 2010. A novel algorithm for the morphometric assessment of radiotherapy treatment planning volumes. *Br J Radiol,* 83**,** 44-51.

JOHNSTON, N., DE RYCKE, J., LIEVENS, Y., VAN EIJKEREN, M., AELTERMAN, J., VANDERSMISSEN, E., PONTE, S. & VANDERSTRAETEN, B. 2022. Dose-volume-based evaluation of convolutional neural network-based auto-segmentation of thoracic organs at risk. *Phys Imaging Radiat Oncol,* 23**,** 109-117.

MLYNARSKI, P., DELINGETTE, H., ALGHAMDI, H., BONDIAU, P. Y. & AYACHE, N. 2020. Anatomically consistent CNN-based segmentation of organs-at-risk in cranial radiotherapy. *J Med Imaging (Bellingham),* 7**,** 014502.

MUKAKA, M. 2012. Statistics corner: a guide to appropriate use of correlation in medical research. *Malawi Med J,* 24**,** 69-71.

SCOCCIANTI, S., DETTI, B., GADDA, D., GRETO, D., FURFARO, I., MEACCI, F., SIMONTACCHI, G., DI BRINA, L., BONOMO, P., GIACOMELLI, I., MEATTINI, I., MANGONI, M., CAPPELLI, S., CASSANI, S., TALAMONTI, C., BORDI, L. & LIVI, L. 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.

SHERER, M. V., LIN, D., ELGUINDI, S., DUKE, S., TAN, L. T., CACICEDO, J., DAHELE, M. & GILLESPIE, E. F. 2021. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother Oncol,* 160**,** 185-191.

SOOMRO, T. A., ZHENG, L., AFIFI, A. J., ALI, A., SOOMRO, S., YIN, M. & GAO, J. 2023. Image Segmentation for MR Brain Tumor Detection Using Machine Learning: A Review. *IEEE Rev Biomed Eng,* PP.

VAN ROOIJ, W., DAHELE, M., BRANDAO, H. R., DELANEY, A. R., SLOTMAN, B. J. & VERBAKEL, W. F. 2019. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. *International Journal of Radiation Oncology* Biology* Physics,* 104**,** 677-684.

VINOD, S. K., JAMESON, M. G., MIN, M. & HOLLOWAY, L. C. 2016. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiother Oncol,* 121**,** 169-179.

WIESINGER, F., PETIT, S., HIDEGHÉTY, K., HERNANDEZ TAMAMES, J., MCCALLUM, H., MAXWELL, R., PEARSON, R., VERDUIJN, G., DARÁZS, B., KAUSHIK, S., COZZINI, C., BOBB, C., FODOR, E., PACZONA, V., KÓSZÓ, R., EGYÜD, Z., BORZASI, E., VÉGVÁRY, Z., TAN, T., GYALAI, B., CZABÁNY, R., DEÁK-KARANCSI, B., KOLOZSVÁRI, B., CZIPCZER, V., CAPALA, M. & RUSKÓ, L. 2021. Deep-Learning-based Segmentation of Organs-at-Risk in the Head for MR-assisted Radiation Therapy Planning. *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies.*

WONG, J., FONG, A., MCVICAR, N., SMITH, S., GIAMBATTISTA, J., WELLS, D., KOLBECK, C., GIAMBATTISTA, J., GONDARA, L. & 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**,** 152-158.

ZHU, J., CHEN, X., YANG, B., BI, N., ZHANG, T., MEN, K. & DAI, J. 2020. Evaluation of Automatic Segmentation Model With Dosimetric Metrics for Radiotherapy of Esophageal Cancer. *Front Oncol,* 10**,** 564737.

3.7 Supplementary Information

- ***Training parameter information***:

  The network consisted of a two-stage 3D-U-net design, where the first (localisation network) was used to extract smaller regions of the 3D image data for the second level models to refine the per-OAR segmentation. The initial network used 4 max pooling convolutional layers, with reLU activation, a 3*3 convolutional kernel and 24 filters at the initial layer. Dropout (p=0.5) and instance normalisation was applied to each layer. Transposed convolution was used for up sampling.

  Optimisation was performed using the Adam optimiser with learning rate 1*10-4 and a batch size of 1. Losses were computed using categorical cross entropy with Softmax activation. Losses in the localisation model were weighted by inverse average OAR volume, to encourage localisation of small OARs. Data augmentation including translation, rotation and elastic deformation was applied during training.

- ***Dosimetric Evaluation: Statistical analysis:***

  To evaluate the statistical significance of these metrics and determine the impact of editing before training the model, each 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. (1 metric and 3 segmentation pairs for MRI, 1 metrics and one segmentation pair for CT). The Bonferroni corrected statistical significance threshold was $p \leq 0.01$ (0.05/3) and ≤ 0.05 for the MRI and CT dosimetric evaluations, respectively.

	ΔBrainstem D5%	ΔOrbit L D1%	ΔOrbit R D1%	ΔOptic Nrv L D1%	ΔOptic Nrv R D1%	ΔOptic Chiasm D1%	ΔCochlea L D50%	ΔCochlea R D50%	ΔPituitary Max dose	ΔLacrimal L D1%	ΔLacrimal R D1%	Δlens L D1%	ΔLens R D1%
MRIeMRI* vs MRIeCT**													
P. Value Threshold: $p \leq 0.01$	0.318	0.083	0.074	0.093	0.115	0.323	**	**	0.614	$$	$$	**	0.070
Effect size: Δ median	0%	4%	6%	0%	4%	1%			1%				-1%
N*	9	9	9	9	9	8	5	4	7	4	2	2	8
MRIeMRI* vs MRIu**													
P. Value Threshold: $p \leq 0.01$	0.965	0.208	0.012	0.033	0.298	0.316	**	**	0.243	##	##	0.435	**
Effect size: Δ median	0%	4%	7%	10%	6%	1%			0%			1%	
N*	9	9	9	9	9	6	1	5	6	5	2	9	5
MRIeCT* vs MRIu**													
P. Value Threshold: $p \leq 0.01$	0.139	0.178	0.687	0.146	0.818	$$	$$	##	$$	0.318	0.263	##	0.371
Effect size: Δ median	0%	0%	2%	10%	2%					0%	-3%		2%
N*	9	9	9	9	9	5	1	2	4	7	8	2	6

*** Number of compared segmentations (successfully segmented by both models considered)**
**** MRIeMRI segmented more cases**
$$ MRIeCT segmented more cases
MRI Unedited segmented more cases

*Table S1: Paired T-test results comparing dosimetric difference between the MRI models. Bold values indicate statistically significant differences ($p \leq 0.016$). Insufficient successful segmentations were achieved by one of the models, this is noted ($$, **, or ##), indicating the superior model. A positive effect size indicates the tested model (***) was closer to the gold standard than the comparison model.*

CTeCT*** vs CTu													
	ΔBrainstem D5%	**ΔOrbit L D1%**	**ΔOrbit R D1%**	**ΔOptic Nrv L D1%**	**ΔOptic Nrv R D1%**	**ΔOptic Chiasm D1%**	**ΔCochlea L D50%**	**ΔCochlea R D50%**	**ΔPituitary Max dose**	**ΔLacrimal L D1%**	**ΔLacrimal R D1%**	**Δlens L D1%**	**ΔLens R D1%**
P. Value *Threshold: ≤0.05*	0.139	0.461	0.147	0.695	0.269	0.097	0.250	0.872	0.163	**	**	0.692	0.359
Effect size: Δ median	3%	0%	3%	-8%	0%	-7%	-2%	1%	-4%			0%	0%
N*	10	10	10	7	7	7	7	7	7	2	5	6	8

*** Number of compared segmentations (successfully segmented by both models considered)**
**** CTeCT segmented more cases**

*Table S2: Paired T-test results comparing dosimetric difference between the CT models. Insufficient successful segmentations were achieved by one of the models, this is noted ($$, **, or ##), indicating the superior model. A positive effect size indicates the tested model (***) was closer to the gold standard than the comparison model.*

Figure 1S: a: Perecentage Dose change between the Gold Standard contour and MRI autosegmentation models for the Lens_L D1%

Note:Stars (*)for the cases were not segmented by the model in all the following graphs.

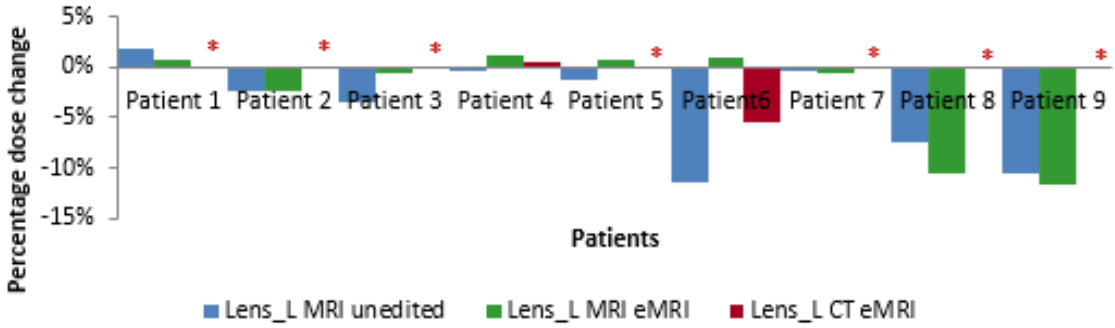


Figure 1S: b) Perecentage Dose change between the Gold Standard contour and MRI autosegmentation models for the Lens_R D1%

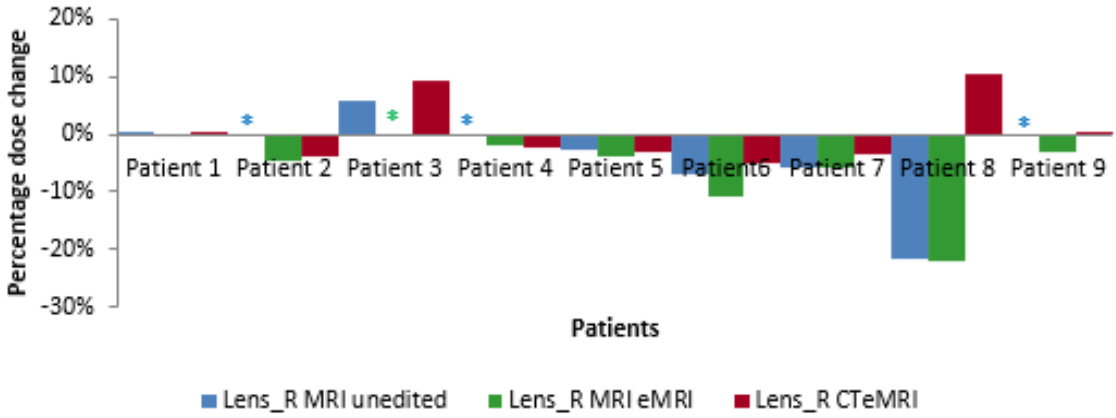
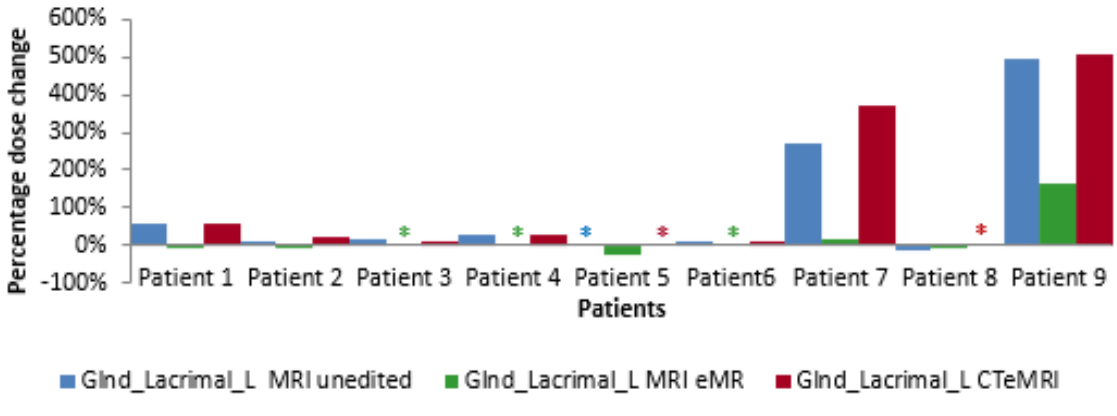


Figure 1S: c) Perecentage Dose change between the Gold Standard contour and MRI autosegmentation models for the Glnd_Lacrimal_L D1%



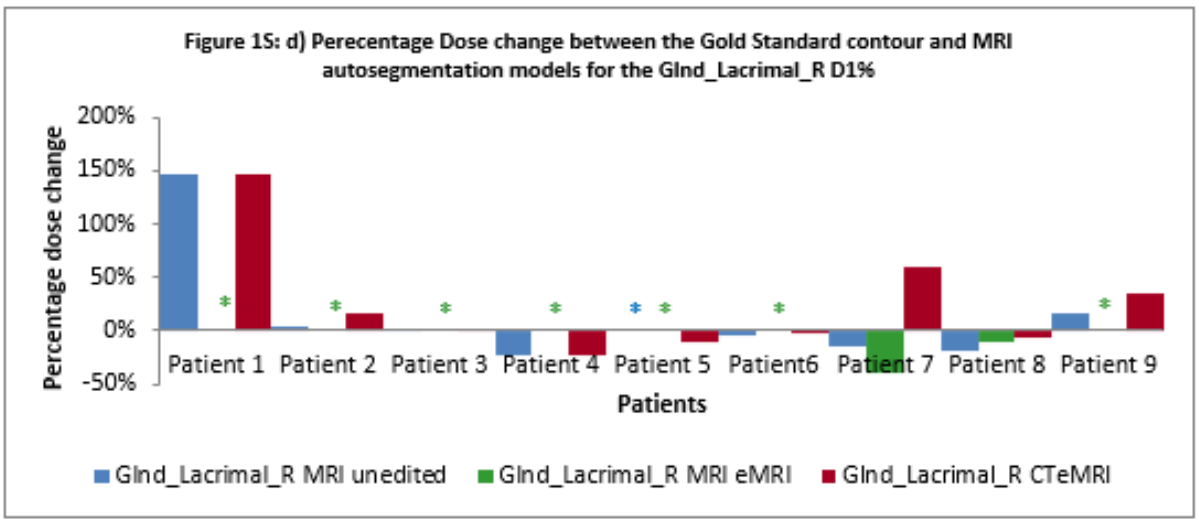


Figure 1S: d) Perecentage Dose change between the Gold Standard contour and MRI autosegmentation models for the Glnd_Lacrimal_R D1%
Percentage dose change
200%
150%
100%
50%
0%
-50%
Patient 1
Patient 2
Patient 3
Patient 4
Patient 5
Patient6
Patient 7
Patient 8
Patient 9
Patients
Glnd_Lacrimal_R MRI unedited
Glnd_Lacrimal_R MRI eMRI
Glnd_Lacrimal_R CTeMRI

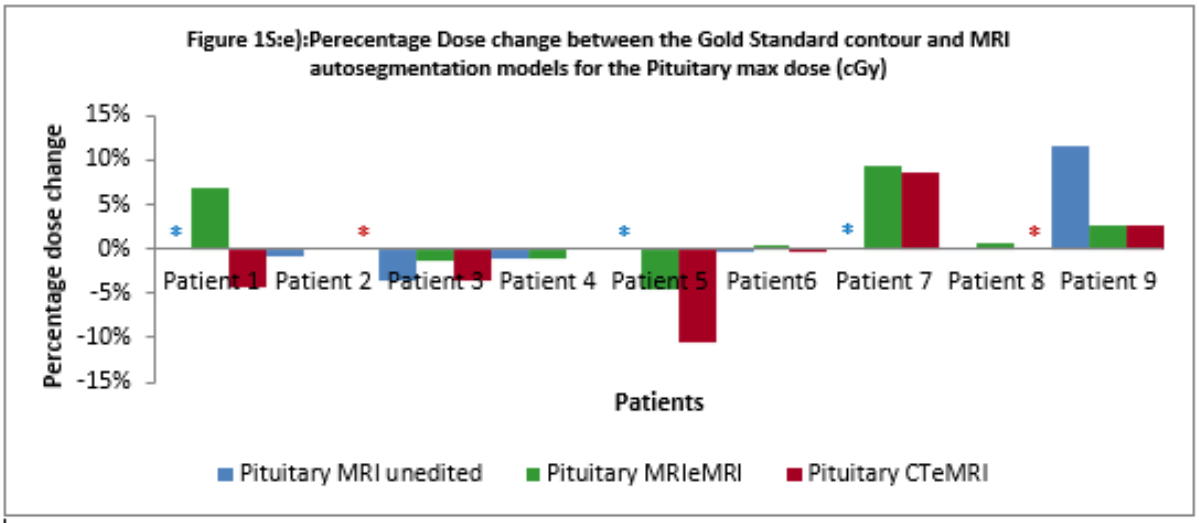


Figure 1S:e):Perecentage Dose change between the Gold Standard contour and MRI autosegmentation models for the Pituitary max dose (cGy)
Percentage dose change
15%
10%
5%
0%
-5%
-10%
-15%
Patient 1
Patient 2
Patient 3
Patient 4
Patient 5
Patient6
Patient 7
Patient 8
Patient 9
Patients
Pituitary MRI unedited
Pituitary MRIeMRI
Pituitary CTeMRI

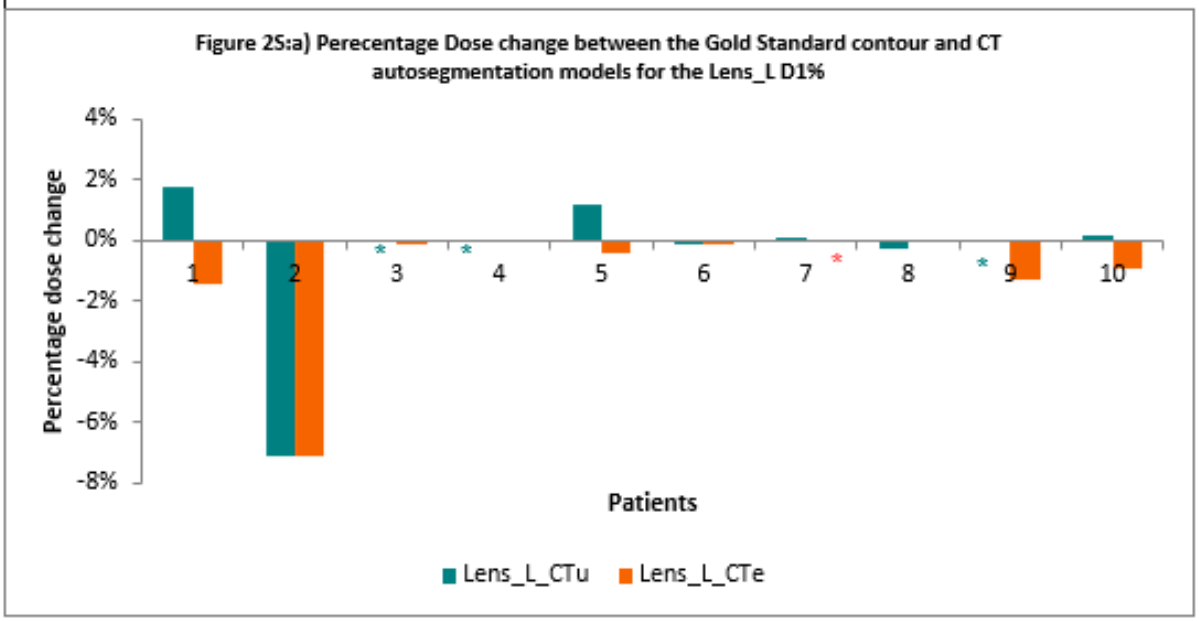


Figure 2S:a) Perecentage Dose change between the Gold Standard contour and CT autosegmentation models for the Lens_L D1%
Percentage dose change
4%
2%
0%
-2%
-4%
-6%
-8%
1
2
3
4
5
6
7
8
9
10
Patients
Lens_L_CTu
Lens_L_CTe

Figure 2S:b) Perecentage Dose change between the Gold Standard contour and CT autosegmentation models for the Lens_R D1%

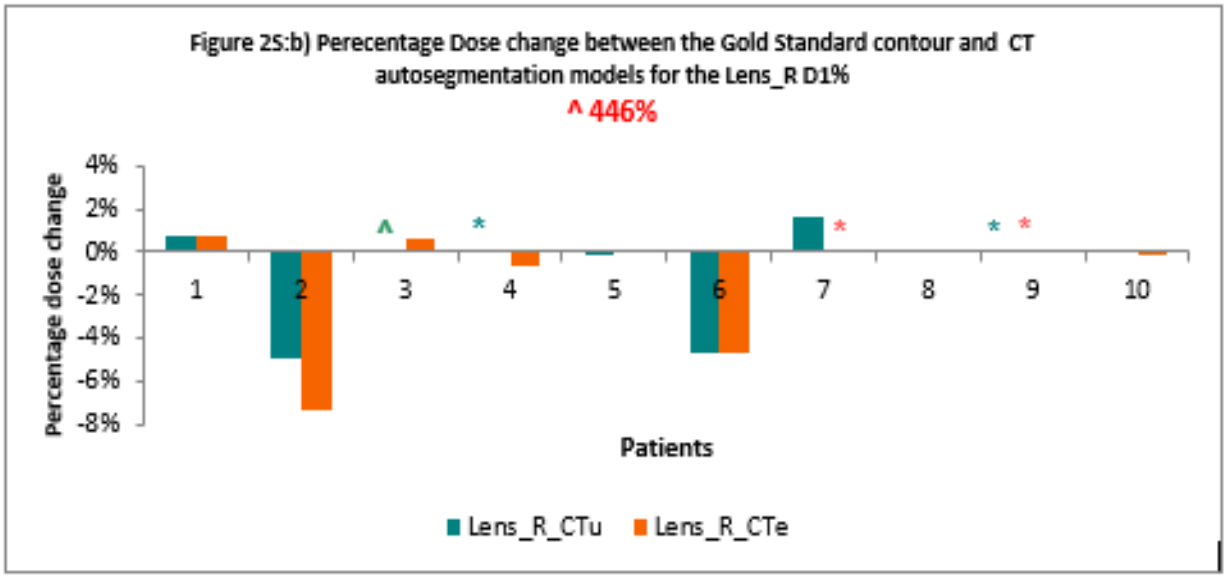


Figure 2S:c) Perecentage Dose change between the Gold Standard contour and CT autosegmentation models for the Glnd_Lacrimal_L D1%

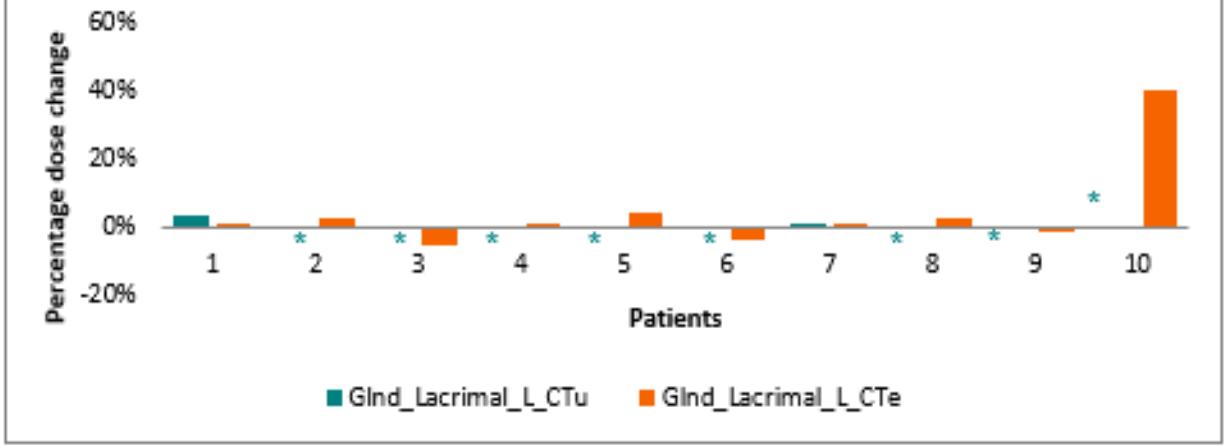
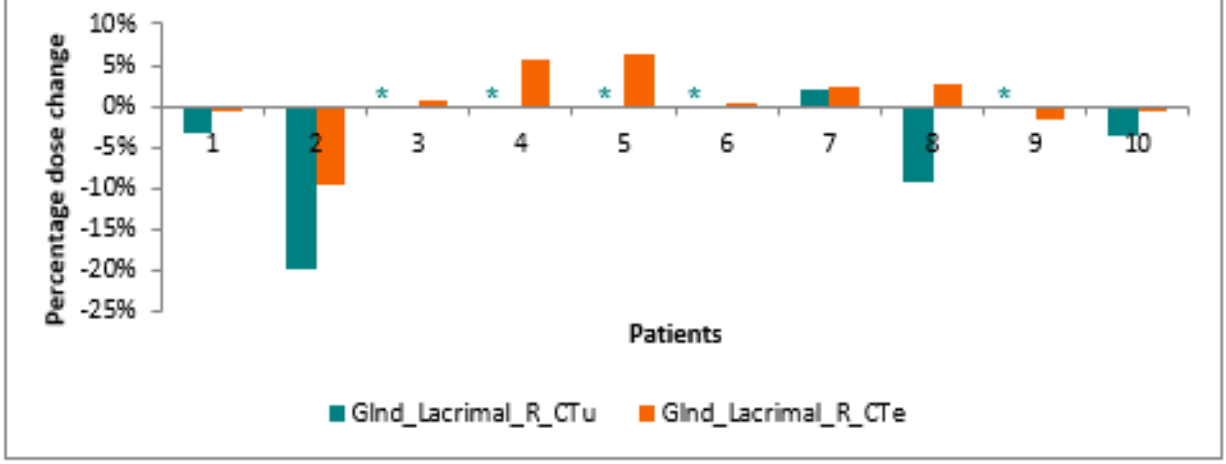


Figure 2S: d) Perecentage Dose change between the Gold Standard contour and CT autosegmentation models for the Glnd_Lacrimal_R D1%



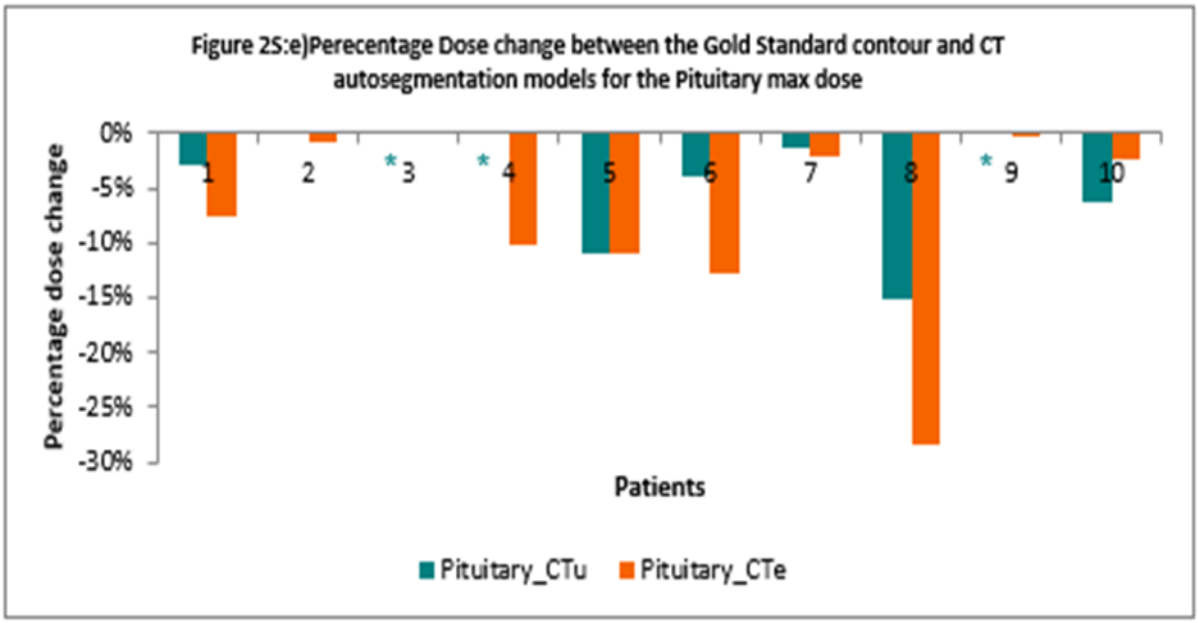


Figure 25:e)Perecentage Dose change between the Gold Standard contour and CT autosegmentation models for the Pituitary max dose
Percentage dose change
0%
-5%
-10%
-15%
-20%
-25%
-30%
1
2
*3
*4
5
6
7
8
*9
10
Patients
Pituitary_CTu
Pituitary_CTe

Chapter 4 Automated Confidence Estimation in Deep Learning Auto-Segmentation for Brain Organs at Risk on MRI for Radiotherapy

Abstract:

Purpose:

We have built a novel AI-driven QA method called AutoConfidence (ACo), to estimate segmentation confidence on a per-voxel basis without gold standard segmentations, enabling robust, efficient review of automated segmentation (AS). We have demonstrated this method in brain OAR AS on MRI, using internal and external (third-party) AS models.

Methods:

32 retrospectives, MRI planned, glioma cases were randomly selected from a local clinical cohort for ACo training. A generator was trained adversarialy to produce internal autosegmentations (IAS) with a discriminator to estimate voxel-wise IAS uncertainty, given the input MRI. Confidence maps for each proposed segmentation were produced for operator use in AS editing and were compared with 'difference to gold-standard' error maps. Nine cases were used for testing ACo performance on IAS and validation with two external deep-learning segmentation model predictions (external model with low quality AS (EM-LQ) and external model with high quality AS (EM-HQ)). Matthew's correlation coefficient (MCC), False Positive Rate (FPR), False Negative Rate (FNR), and visual assessment were used for evaluation. Edge removal and geometric distance corrections were applied to achieve more useful and clinically relevant confidence maps and performance metrics.

Results:

ACo showed generally excellent performance on both internal and external segmentations, across all OARs (except lenses). MCC was higher on IAS and low-quality external segmentations (EM-LQ) than high-quality ones (EM-HQ). On IAS and EM-LQ, average MCC (excluding lenses) varied from 0.6 to 0.9, while average FPR and FNR were ≤0.13 and ≤0.21, respectively. For EM-HQ, average

MCC varied from 0.4 to 0.8, while average FPR and FNR were ≤0.37 and ≤0.22, respectively.

Conclusion:

ACo was a reliable predictor of uncertainty and errors on AS generated both internally and externally, demonstrating its potential as an independent, reference-free QA tool which could help operators deliver robust, efficient autosegmentation in the radiotherapy clinic.

Key words: Deep-learning, AI, Confidence, Autosegmentation, uncertainty, MRI scans, Brain, organs at risk, Radiotherapy

4.1 Introduction

Deep-learning (DL) promises efficiency and quality improvements for radiotherapy (RT) (van den Berg and Meliado, 2022) but also raises concerns around potential adverse safety consequences (van den Berg and Meliado, 2022, Faghani et al., 2023). The most developed and commonly deployed DL application in radiotherapy is autosegmentation (AS) (Cardenas et al., 2019, van Dijk et al., 2020). Despite efficiency gains, errors and uncertainty resulting from data and model limitations necessitate time-consuming human review and editing, which can reduce or eliminate efficiency gains (Claessens et al., 2022). Editing is also prone to inter-operator variability and bias, reintroducing inconsistencies, and potential segmentation errors (Claessens et al., 2022, Apolle et al., 2019). Uncertainty quantification is vital to allow operators to evaluate appropriate localised confidence in AS (Faghani et al., 2023, Abdar et al., 2021). It has been attempted previously, using internal probability estimates from the segmentation model itself (Claessens et al., 2022).

However, research into uncertainty estimation for AS remains nascent (van den Berg and Meliado, 2022). Without any ground-truth, predicting and evaluating uncertainty remains challenging (Abdar et al., 2021).

Clinical quality assurance (QA) for AS is still developing, and remains a manual process in general; however, recommendations have been made that any QA tool should be independent of the underlying AS model, rather than relying solely on internal model probabilities (Claessens et al., 2022, Liesbeth et al., 2020).

Whilst most DL models can provide internal probabilities alongside class predictions, these are typically poorly calibrated (Asgharnezhad et al., 2022, Yeung et al., 2023), with high-probability predictions for most voxels, except those with predictions extremely close to decision boundaries. This behaviour raises the spectre of ‘confident-but-wrong’ predictions, which are a major concern in safety critical applications like RT.

A more useful concept than probability is uncertainty. Uncertainty can be broken down into epistemic (driven by model limitations) and aleatoric (driven by data variability) uncertainty(van den Berg and Meliado, 2022, Wang et al., 2019a, Abdar et al., 2021). Several approaches to estimating these uncertainties have been developed, including Monte-Carlo Dropout (MCD), ensemble-model (EM) and Spectral-normalized Neural Gaussian Process (SNGP) (Abdar et al., 2021,

Liu et al., 2020). MCD and EM predict a distribution of possible results, while SNGP integrates variance into a classifier to estimate prediction uncertainty, based on distance-awareness (Abdar et al., 2021, Liu et al., 2020).

Spatial probability maps based on MCD have been previously explored as a QA method for AS in RT, albeit with limited correlation between predicted uncertainty and observed error (van Rooij et al., 2021). Similarly, a secondary neural network (NN) was used to predict Dice similarity Coefficient from CT-AS pairs (Chen et al., 2020), with internal class probability as an input to the QA network. However, these techniques all rely on the prediction-generating model and training data distribution to produce uncertainty estimates, and hence fail the test of independence. They are also susceptible to internal probability calibration issues as well as creating practical challenges for use with existing AS models, which may not provide probabilistic predictions.

Herein, we propose a model-agnostic, independent uncertainty estimator, based on correlation of features in the underlying image and proposed segmentation, rather than internal AS model probability. In principle, a secondary neural network (NN) can estimate errors in AS predictions, if these errors are known, relative to data labels. However, clinical confidence depends on identifying both detectable errors, and regions of high uncertainty, even if the predicted AS is 'correct'. We aim to combine these objectives, producing a 'confidence map' that highlights both i) regions of likely error and ii) regions of low confidence due to either aleatoric uncertainty (e.g., low image contrast, and dissimilar image acquisition) or epistemic uncertainty whether they correlate with actual errors to labels or not. To achieve the latter, we require a distribution of AS predictions on each image, to train a secondary confidence model.

By leveraging adversarial learning, in which one NN learns to critique the predictions of another, we can address both objectives in a unified framework. Our discriminative network is trained to estimate the voxel-wise probability that a test segmentation is derived from the gold standard population. This probability is consistently low if the AS network prediction is consistently incorrect, and highly variable if the AS network predictions varies. The network therefore learns not only to predict where discrepancies with gold standard labels occur at the sample level but also regions of uncertainty in which such discrepancies are more likely at the population level.

We present AutoConfidence (ACo), a novel AI-driven QA method to estimate autosegmentation confidence on a per-patient basis, without a gold standard reference. We demonstrate it for AS of brain OARs on MRI, against internal segmentations from the generator network (IAS) and external (third party) AS models of varying quality. ACo produces a voxel-wise confidence map enabling efficient and robust manual verification and editing. ACo was designed to focus users' attention on regions of low confidence which require attention to avoid significant segmentation errors, potentially improving safety, confidence, and efficiency in clinical practice.

4.2 Materials and Methods:

4.2.1 Network concept and architecture

ACo is an adversarial architecture based on the concept of a segmentation generator (G) discriminator (D) pair, working against each other to produce convincing segmentations and estimating the probability (pGS) that the prediction is from the distribution of correct segmentations, from which the (noisy) gold standard examples are drawn. D therefore estimates the probability that a segmentation is 'plausible' for a given image. (Figure 4.1)

In contrast to conditional Generative Adversarial Networks (cGAN) (Isola et al., 2017), where D exists only to train G, here G exists solely to provide a training example distribution for D, which is used independently at imputation to provide confidence estimates on segmentations from any source.

G is a typical 2D-U-net segmentation network with 8 convolutional layers, featuring batch norm (BN), dropout and 'relu' activation. G takes 2D MRI slices as input and predicts segmentations per voxel, with Softmax activation at the final layer. Focal log loss (g = 2) against gold standard segmentation labels is used as the non-adversarial loss.

In further contrast to cGANs, D followed a U-RESnet architecture, extending the concept of a patch-wise discriminator to a shallow encoder-decoder, providing voxel-wise discrimination, with residual blocks to improve performance in the shallow architecture. D takes image-segmentation pairs as input and learns to predict voxel-wise difference to gold-standard (d2GS) as a surrogate for pGS.

D discriminates predicted segmentations from gold standard segmentations (GS), on a per-voxel basis, with focal log loss against 'difference-to-gold-

standard' (d2GS), which is provided as a binary voxel-wise label. D was trained alternately with gold standard contours and AS predictions, for each input image. G and D were updated sequentially for each batch during training.

The output of D therefore represented a confidence map, across the segmentation, which was used as an adversarial loss to augment training of G. This adversarial loss encouraged G to produce different AS predictions where D predicted low confidence. Hence, errors were reduced, making the task more difficult and consequently improving the performance of D on hard examples. Detailed model architecture and hyperparameters are provided in supplementary information (research method).

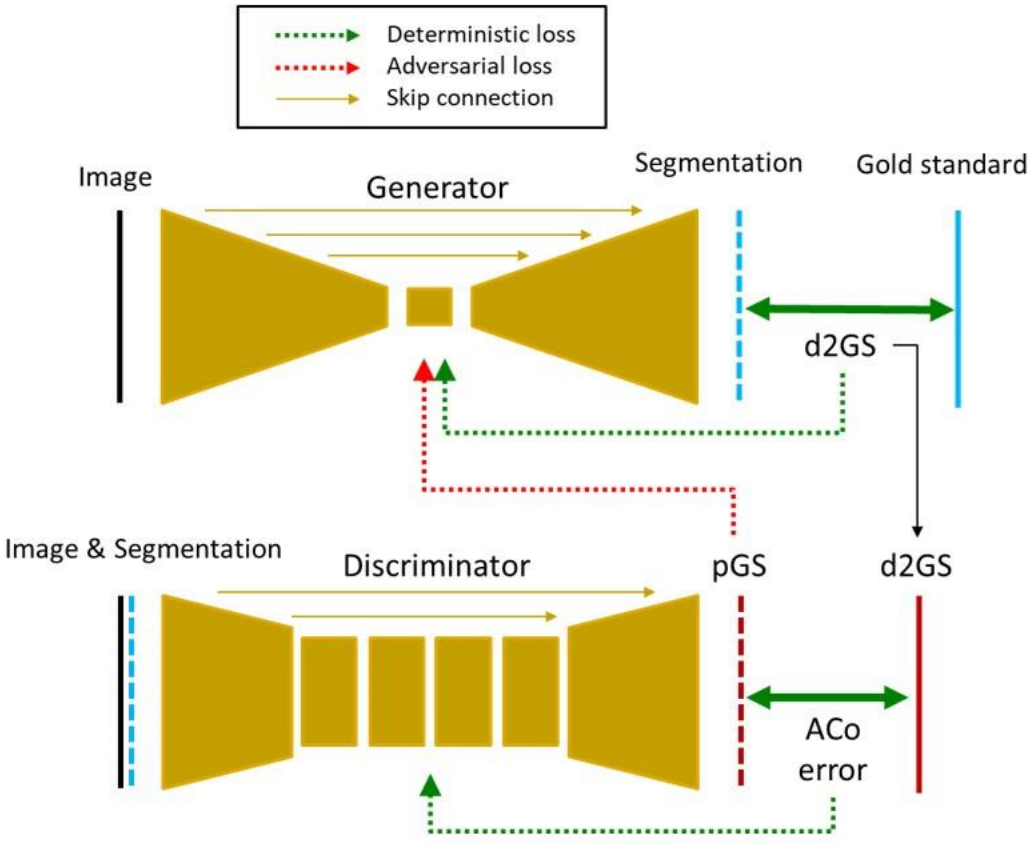


Figure 4.1: Model architecture overview for ACo, showing the segmentation generator and confidence estimator (discriminator). Segmentation difference to gold standard (d2GS) is used as a loss to the generator and a reference for the discriminator. Confidence prediction (pGS) is compared to d2GS, resulting in ACo error, which is used as a loss to the discriminator.

4.2.2 Data and training

ACo models were trained using 32 retrospective, clinical glioma cases, each containing ~100 clinical T1-weighted gadolinium-enhanced MRI (T1w-Gd MRI) slices, with gold-standard contours, manually quality assured by editing clinical contours. 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.

Thirteen commonly delineated brain OARs, which were previously included in AS models developed in-house, were selected. Details concerning image acquisition, OARs selection and gold standard contour delineation are found in our previous published work (Alzahrani et al., 2023).

ACo was initially trained as described above. A second, otherwise identical, model was trained by adding synthetic errors to the IAS produced by G, before D estimated pGS. These included random geometric deformations, random OAR class perturbations, and random removal of an entire OAR, constrained to remain within plausible physiological limits. These synthetic errors were intended to mimic potential real-world errors from AS algorithms.

4.2.3 Evaluation of AutoConfidence performance

Nine independent glioma test cases were used to evaluate the performance of both models on the IAS. The following metrics were used to compare the outputs of the ACo map that was produced by each model relative to the d2GS: using: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), Matthew's correlation coefficient (MCC), False Positive Rate (FPR), False Negative Rate (FNR), FP/FN ratio.

Two corrections (Intelligent Edge removal and Geometric Distance Correction) were applied to the model output, and their impact was assessed using the same methodology. The corrections were justified, applied, and optimised as follows:

1) **Intelligent Edge removal (IER)**

IER was applied to remove 'thin' (clinically insignificant) error regions around OAR borders, which resulted from partial volume effects at the edges of binary segmentation masks. Sobel edges of the segmentation were computed and dilated to a width of n voxels. This edge mask was applied to the confidence map and d2GS map, removing errors close to OARs boundaries. Errors remaining after this process were regrown by k+2 voxels back into this region, to avoid

deleting significant errors adjacent to OAR boundaries (Figure 4.2). Visual assessment was used to determine the optimal mask width as k=3, as the minimum value which resulted in removal of thin error predictions along all OAR boundaries.

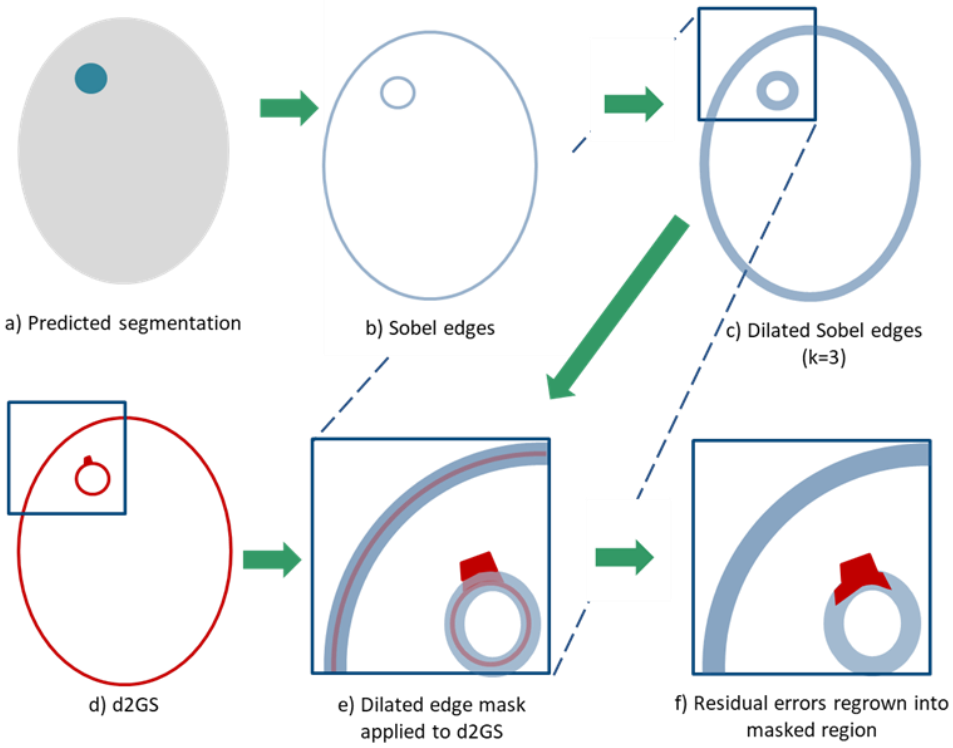


Figure 4.2: IER algorithm. Sobel edges (b) were generated from predicted segmentations (a) and dilated (c) with a kernel size of 3 voxels. d2GS (d) was masked with the dilated edges (e) to remove errors at the OAR boundaries. Finally, remaining errors were regrown into the boundary region (f).

2) Geometric Distance Correction (GDC)

As ACo was designed to detect both errors to gold standard and regions of correct prediction but low confidence, there was a difference in the clinical significance of both FP and FN prediction errors, which depends on their proximity to TP predictions. FP voxels may be valid depending on their location. Low confidence regions (e.g., where image contrast is low) are typically geometrically close to actual error regions but extend beyond them.

Hence, when a FP voxel was ‘close’ to a sufficiently large true positive region, it was considered clinically useful as an indicator of potential error (i.e., low confidence), and counted as TP for Aco performance evaluation. Equally, if a FN voxel was close to a large TP region, it could be disregarded clinically as the TP region would draw operator attention.

Our definition of ‘closeness’ depended on the size of the true positive region within a local image patch, which should be larger than the distance to it from the voxel in question.

To make quantitative performance analysis more clinically relevant, prediction classes were updated as follows: Within a local patch of the image, the maximum dimension of the TP region [D(tp)_max] was computed via Euclidean Distance Transform (EDT). Secondly, the minimum distance to the TP region from the central voxel [D(x_tp)] was computed, also via EDT. If D(x_tp) < D(tp_max) and the central voxel was either FP or FN, it was reclassified as TP or TN respectively (see supp info Figure S1).

4.2.4 Validation with external autosegmentation

The same nine test cases were used to evaluate the performance of ACo model on external MRI AS that generated using custom models based on a commercial 3D medical image segmentation U-net (RayStation 11A, RaySearch AB). Two external models were used; external model - low quality (EM-LQ) and external model - high quality (EM-HQ), which were previously trained, using clinical and gold standard contours respectively, allowing evaluation of AutoConfidence in both scenarios. (*More information about the difference between these models can be found in our previous work)(Alzahrani et al., 2023).*

Evaluation was done pre- and post-application of IER and GDC, as described above.

‘Four-colour maps‘ were produced for visualisation of FP, FP, FN regions and the regions modified by GDC. Visual assessment was performed by 3 expert observers, to evaluate the clinical utility of the ACo model predictions, based on the location, type and appropriateness of regions highlighted as low confidence, especially where no explicit errors existed in the d2GS reference.

4.3 Results

4.3.1 Internal Autosegmentation Quality

Relative to gold standard, G produced acceptable IAS, with average test DSC from 0.47 to 0.85 for all structures except lacrimal glands (DSC ≤ 0.30), (Table 1S).

4.3.1.1 Internal versus External Autosegmentation Quality

Excluding lacrimal glands, IAS performance (DSC 0.47 to 0.85) was comparable to EM-HQ (DSC 0.49 to 0.91) (Table 1S). IAS outperformed EM-LQ (DSC 0.28 to 0.89). All models performed equally poorly for lacrimal glands (DSC ≤ 0.26)

4.3.2 AutoConfidence results

4.3.2.1 AutoConfidence outputs on the internal AS:

ACo trained with synthetic errors was superior to ACo without synthetic errors. FNR, using d2GS as a reference, showed a mean reduction of 7% across 11/13 OARs (Table 2S and 3S), using synthetic errors. FPR was not strongly affected (mean ΔFPR = 0.4%). Mean MCC improved by 0.075. All subsequent ACo results were obtained with synthetic errors during model training.
MCC on IAS, following application of IER and GDC, was > 0.69 (Figure 4.3a) for all OARs except lenses (MCC ≤ 0.40). FPR and FNR on IAS were ≤ 0.13 and ≤ 0.17, respectively (≤0.53 and ≤0.16 for lenses, resulting in a high FP/FN ratio of 9.0 (Table 4S).

ACo performance was shown to rely on both MRI and AS inputs via ablation testing. Removing either images or segmentations during training and testing severely affected confidence estimation (mean MCC 0.05 and 0.08 respectively). Furthermore, training IAS generator and ACo networks sequentially, as opposed to adversarially, severely impacted ACo test performance (mean MCC < 0.23).

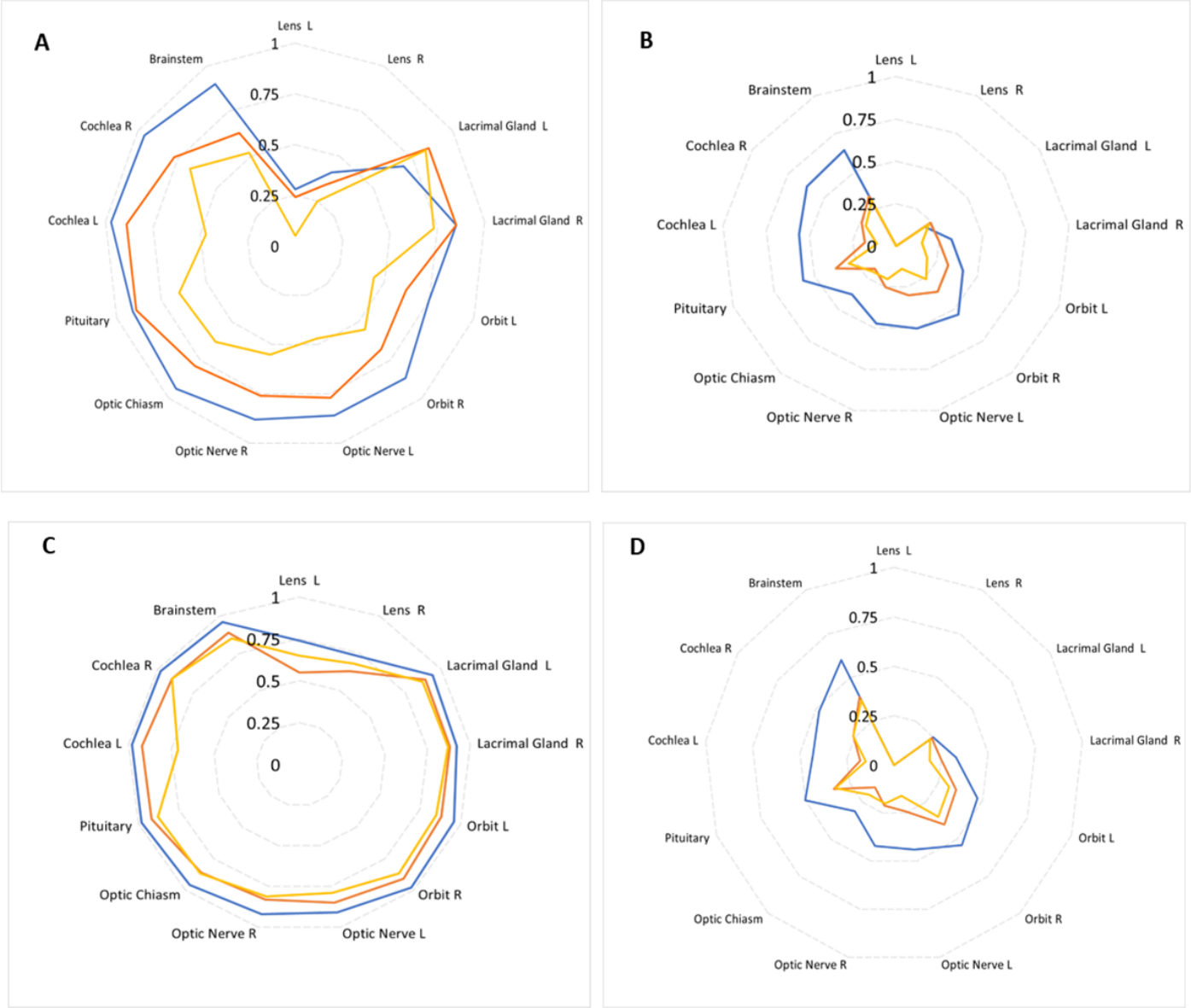


Figure 4.3: Mean MCC for AutoConfidence per OAR and per segmentation model. Performance is shown on IAS (blue), MRIu (orange) and MRIeMRI (yellow) segmentations. a) MCC of ACo with IER and GDC combined, b) IER only c) GDC only, and d) baseline ACo output without corrections.

4.3.2.2 AutoConfidence outputs on the external AS

An example of the ACo output confidence map and a colour-coded comparison to d2GS is shown in figure 4.4. Following IER and GDC, and excluding lenses, mean MCC for ACo on EM-LQ autosegmentations ranged from 0.62 to 0.89, with left orbit, right orbit and brainstem having the lowest scores (0.62, 0.68, and 0.63, respectively), and a high error ratio (FP/FN ≥2.13). (Figure 4.3a, table 5S). Mean

MCC for ACo on the EM-HQ autosegmentations ranged from 0.44 to 0.83, with left orbit, left cochlea, and left optic nerve having the lowest scores (0.44, 0.47, and 0.47, respectively), and FP/FN ≥ 2.67 (Figure 4.3a, table 6S).

The lowest mean MCC scores for ACo across all the Raystation MRI AS models were for lenses (average MCC ≤ 0.34) (Figure 4.3a). FP/FN for lenses on EM-LQ was ≤ 0.67, but for EM-HQ was 3.0 (Tables 5S and 6S).

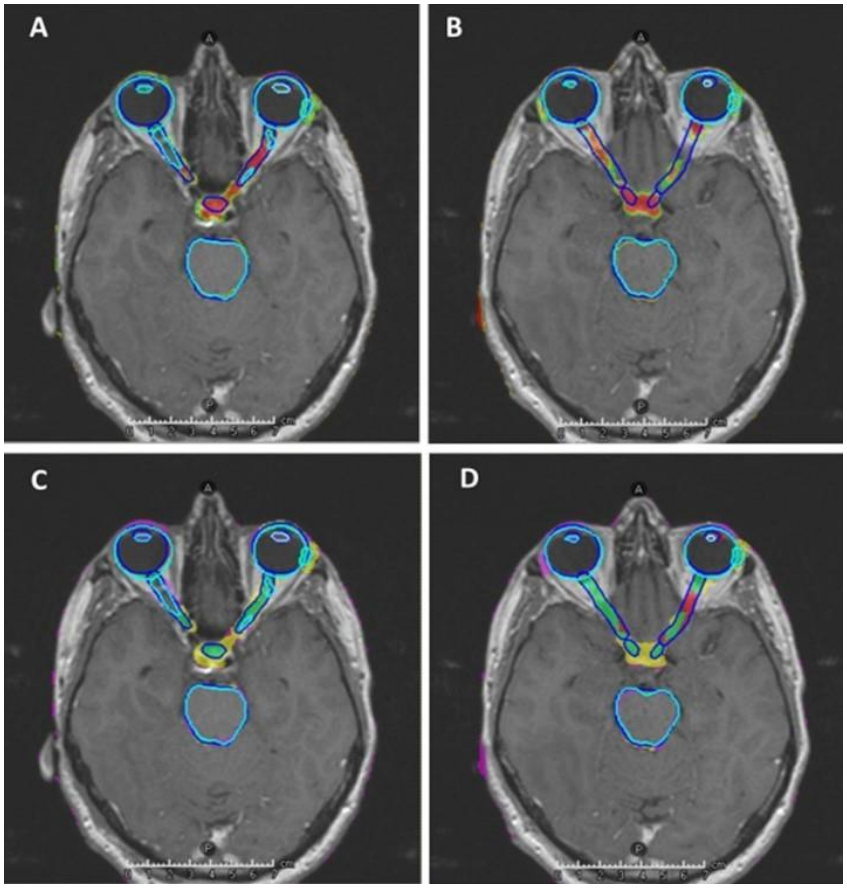


Figure 4.4: Axial T1w-Gd MRI with dark-blue contours representing gold standard and light blue representing EM-LQ autosegmentation. a) example showing a high uncertainty level for missing pituitary segmentation and errors in optic nerves. b) demonstrating high uncertainty for missing optic chiasm and apparent false-negative predictions for optic-nerves, which are in fact due to non-anatomical GS contours (see main text). FP in the region near lacrimal glands is typical of ACo on MRI, where lacrimal glands are very hard to visualise and therefore highly uncertain in location. c and d) four-colour-map showing regions of TP (green), FP (pink) and GDC modified FP (yellow) relative to the differences-to-gold-standard, for the ACo prediction.

4.3.3 Impact of postprocessing on ACo

4.3.3.1 Intelligent Edge removal (IER)

IER alone had limited impact on quantitative ACo performance. Mean MCC, improved by 0.03 and 0.01 compared to baseline for IAS and EM-LQ autosegmentations respectively, and reduced by 0.04 for EM-HQ autosegmentations (Figure 4.3b).

Mean FPR showed a reduction of 0.09, 0.08, and 0.11 relative to baseline, on IAS, EM-LQ, and EM-HQ autosegmentations, respectively. Absolute FP count was also reduced for all OARs, driving the change in FPR.

Mean FNR increased by 0.06, 0.06 and 0.14 on IAS, EM-LQ and EM-HQ autosegmentations, respectively (Tables 7S- 9S). Absolute FN count did not increase significantly for any OARs. TP count decreased significantly for Orbits and Brainstem on internal and EM-LQ autosegmentations. For EM-HQ, all OARs exhibited significant decreases in TP count. These changes in TP count drove the observed changes in FNR (Table 9S).

Mean FPR, FNR, and MCC for the ACo outputs using the IER can be found in Supplementary Tables 7S-8S.

4.3.3.2 Geometric Distance Correction (GDC)

GDC improved mean MCC across all OARs by 0.50, 0.63 and 0.62 for IAS, EM-LQ and EM-HQ AS respectively, relative to the baseline (Figure 4.3c).

GDC reduced both FPR and FNR relative to baseline across all OARs and AS models. Performance on IAS, EM-LQ, and EM-HQ showed reductions of 0.35, 0.38 and 0.32 respectively. This change in FPR was directly driven by corresponding changes in FP count (see supp info Tables 10S-12S).

Mean FNR showed a reduction with GDC of 0.19, 0.22, and 0.24 on the IAS, EM-LQ, and EM-HQ AS respectively. These changes were driven in part by small reductions in FN count but were predominantly a result of increased TP count.

Mean FPR, FNR and MCC for ACo using GDC can be found in Supplementary Tables 10S-12S.

4.3.3.3 Combined Corrections (IER and GDC)

Combining the two postprocessing corrections improved mean MCC across all OARs by 0.43, 0.49, and 0.33 relative to baseline for IAS, EM-LQ and EM-HQ AS, respectively (Figure 4.3a).

IER and GDC combined decreased mean FPR relative to baseline for all OARs by 0.27, 0.33,0.23 for IAS, EM-LQ, and EM-HQ AS, respectively. Mean FNR on IAS, EM-LQ and EM-HQ AS was reduced by 0.15, 0.14, and 0.11, respectively.

The mean FPR and FNR for baseline and for combined corrections can be found in Supplementary Tables 4S-6S.

4.4 Discussion

There is strong demand in radiotherapy for deep-learning OAR and target autosegmentation to improve efficiency. However, efficiency gains can be eroded by the need for time-consuming manual checking and editing. Robust automation is desirable but challenging due to lack of automated segmentation QA tools. Our novel AI-driven QA tool (AutoConfidence) can estimate segmentation confidence from the underlying image and proposed segmentation, on a localised, independent, per-patient basis, without gold standard reference segmentation. Our approach is based on adversarial generative learning, utilising the errors and variability of an internal segmentation neural network to train an independent discriminative network to predict a map of segmentation confidence, which is then used to optimise the segmentation model and adversarially train both models. ACo can estimate confidence for segmentations from any source (including manual contours). ACo highlights regions of likely error and low confidence, which may help focus users' attention on regions for review, enabling safe, robust, and efficient implementation of autosegmentation in the clinic.

Importantly, ACo was designed to highlight regions of low confidence (high uncertainty) to the user. These may correspond to delineation errors, relative to gold standard or represent regions of correct segmentation relative to gold standard, but low confidence due to low contrast, high variability in gold standard contouring or imaging artefact. In either situation, ACo will attract operator attention to QA the contour in the suspect region.

Due to lack of ground-truth for segmentation confidence as defined above, quantitative analysis of ACo performance was challenging. As a surrogate, we used 'difference-to-gold-standard' (d2GS), the difference between autosegmented contours and gold standard human contours. However, this surrogate is imperfect for two reasons. Firstly, gold-standard contours can be erroneous, such that apparent false-negative or false-positive regions can derive from label inaccuracy rather than prediction error. Secondly, d2GS only accounts for regions of actual difference between prediction and gold standard, whereas ACo also identifies low-confidence regions, even when the prediction and gold standard align perfectly.

For these reasons, the quantitative analysis against d2GS presented here represents a conservative, partial surrogate for clinical utility and safety of ACo. Nevertheless, ACo showed generally excellent performance on both internal and external segmentations, across all OARs except lenses (Figure 4.3) (Table 4S, 5S, and 6S), once IER and GDC were applied.

IER and GDC were necessary to address two issues with the raw ACo predictions. There are separate rationales for these corrections, which therefore have been designed with different objectives. IER removed thin regions of low-confidence predictions within 2 pixels of an OAR boundary, improving visual perception and enabling the operator to focus on more significant low-confidence regions. GDC did not affect the predicted confidence maps, but rather aimed to modify the statistical performance metrics on the surrogate labels of d2GS, to better correlate with clinical significance when validating ACo.

IER (w=3) reduced average FPR by ~0.1 across all MRI AS models and OARs (Tables 7S-9S), making the ACo map more useful and clinically relevant, as positives very near OAR boundaries, resulting mainly from partial volume effects are not clinically relevant to RT planning. The IER algorithm ‘re-grows’ larger low-confidence signals into this boundary region, ensuring clinically significant predictions are not affected.

The observed increase in FNR, of between 0.06 and 0.11, appeared superficially concerning, as it implied an increased propensity to miss genuine segmentation errors relative to gold standard (increased FN count). If a predicted uncertainty region was fully eroded by edge-removal, but the d2GS map was not, re-growing of eroded d2GS into the boundary region could lead to new FN voxels near the boundary. However, this was found to occur rarely, as evidenced by the very limited increase of FN count with IER (Tables 7S-9S). Furthermore, these artefacts were clinically unimportant, as they were very close to the OAR boundary.

Rather, the observed increase in FNR with IER was primarily driven by removal of TP voxels near the OAR boundaries. This effect was especially pronounced for orbits and brainstem, which were generally well segmented, leaving the majority of residual error voxels at OAR boundaries. Hence, removal of predicted uncertainties along OAR boundaries via IER was safe and should aid visual interpretation of ACo maps.

GDC accounted for the fact that not all incorrectly predicted (FP and FN) voxels were equally important. Statistical performance measures derived from the naive confusion matrix did not account for these effects and hence gave an overly conservative assessment of clinically relevant performance (Figure 4.3d). As ACo was designed to highlight low confidence (potential error) regions to human operators for review, it was considered acceptable or even desirable to predict larger regions of uncertainty than d2GS, provided they were in appropriate locations and not grossly oversized. Likewise, it was acceptable to have small regions of FN, if they were proximal to larger TP regions, as the human operator's attention would be drawn to this area anyway.

GDC strongly reduced FPR and FNR across all OARs and segmentation models. The reduction in FPR showed that most FP voxels were near significant TP regions. The FNR reduction was partly due to GDC acting on FN directly, but mainly due to increased TP, driven by the reclassification of geometrically close FP voxels.

Combining GDC with IER reduced the effect, relative to GDC alone (Table 4S-6S and 10S-12S). This was due to GDC acting on overprediction of errors in the clinically unimportant OAR boundary regions, which were removed by IER before the GDC algorithm was applied.

The model performance and the need for IER and GDC may be altered by a 3D model because the current ACo model is running in 2D, potentially leading to over- or under-prediction of errors at superior and inferior OAR limits. A 3D version of ACo will potentially resolve these issues and improve the model performance.

During the training of ACo, synthetic errors were applied. Without this, the model trained only on good segmentations and became insensitive to errors which it had not seen later in training. Synthetic errors reduced mean FPR and FNR and improved mean MCC across all OARs and test segmentations.

For the optimised ACo model (with IER, GDC and synthetic training errors), MCC was higher on low-quality external segmentations (EM-LQ), vs. high quality ones (EM-HQ) (Figure 4.3a). This was partially due to the dependence of MCC on class-imbalance (fewer errors to detect) and also due to the higher quality model making more subtle errors, which were harder to identify. Furthermore, imperfect labels at the voxel level represent a theoretical limit to apparent ACo performance, which will be approached at segmentation quality improves. The

low mean FPR, FNR and a high FP/FN ratio on EM-HQ AS, indicated the method was safe, as it tended to be over-cautious (FP/FN > 1), and could be used clinically (except for lenses) (Tables 6S). ACo performed slightly better on internally generated test segmentations, as the errors present were generated by the model used during adversarial ACo discriminator training.

Visual assessment by expert oncologists indicated ACo was able to successfully identify inaccurate segmentations and missing structures that generally correlated well with errors to gold standard, as well as regions of low confidence associated with poor image contrast or artefact, highlighting them for clinical review. Figure 4.4a-c show examples of successful error detection in the optic nerves and the missing pituitary gland. The ACo confidence map (Figure 4.4a) shows regions of low confidence, while figure 4.4c shows the graphical confusion matric map, highlighting TP, FP, and FN voxels relative to gold standard.

Where ACo predictions did not correlate with d2GS, visual inspection often revealed a systematic difference due to gold standard contouring definitions. For example, local protocol for manual optic nerve segmentation indicates that nerves should be continuous on each slice, which does not necessarily align with anatomical reality. Hence, the gold standard was not a good comparator in these regions, leading to apparent underperformance of ACo, which appropriately identified errors based on the 3D anatomical image features (Figure 4.4b-d). Additionally, low-contrast regions (e.g., boundary of brainstem - figure 4.5) or image artefacts often led to low-confidence predictions by ACo, in the absence of error to gold-standard. This low confidence was deemed clinically appropriate, as there was genuine uncertainty in such regions.

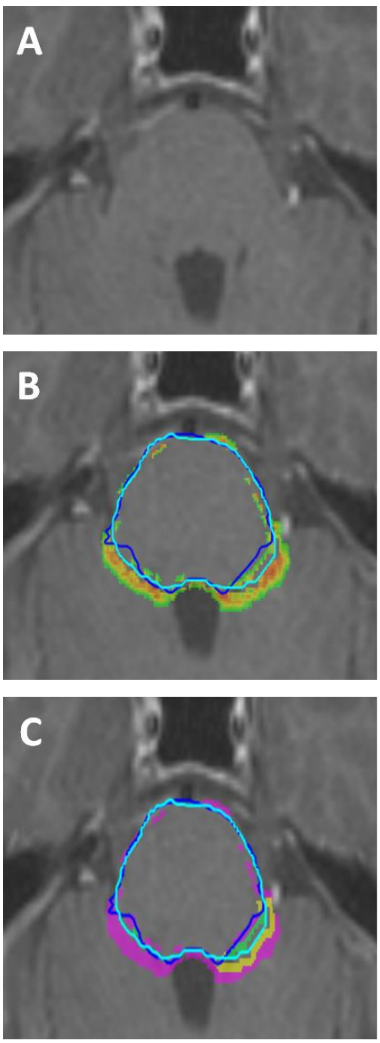


Figure 4.5: Axial T1w-Gd MRI illustrating a) a low contrast in the brainstem region b) ACo prediction (heat map), EM-HQ segmentation (light-blue) and gold standard (dark-blue), showing uncertainty due to the low contrast region around brainstem. c) Four-colour-map showing regions of TP (green), FP (pink) and GDC modified FP (yellow) relative to the differences to gold-standard, for the ACo prediction.

Visual inspection also confirmed that IER made the final confidence maps more useful by removing unimportant but visually distracting low confidence predictions around OAR boundaries. Whilst partial volume effects in the axial plane were removed via IER, both the ACo model itself and the IER algorithm operated in 2D. Hence, the superior and inferior limit of orbits and the cranio-caudal extent of the optic nerves exhibited some residual errors from these partial volume effects, which were not removed by IER, leading to under- or over-prediction of

error in these regions. This could in principle be addressed by a 3D ACo and IER algorithm.

Interestingly, ACo detected uncertainty in erroneous regions of the external segmentation, even though it was not exposed to such errors during training. The generator did not predict external segmentations, and gold standard ones were used as input to the ACo discriminator. Nevertheless, ACo learned how external contours should look relative to images, based on many gold standard external segmentations/image pairs. Any test external-segmentation region which departed from this expectation was labelled as uncertain. Thus, in the unseen error scenario, ACo operated as a 'zero-shot outlier detector', flagging previously unseen error types based on dissimilarity to the gold-standard distribution (low pGS), potentially improving robustness in clinical use.

Despite generally excellent performance of ACo, there were limitations. ACo used a 2D recurrent-Unet, due to GPU memory constraints, whilst segmentations are inherently 3D, leading to over- or under-prediction of errors at superior and inferior OAR limits. Furthermore, as IER was also a 2D algorithm, these issues were not mitigated by post-processing. Lastly, the ACo model trained with only 32 cases (~4500 slices), due to limited data availability for brain MRI. Relatively consistent cranial anatomy made this acceptable, but for a clinical site with greater anatomical variation, more data would be required. This effect was demonstrated by the poor performance of both IAS and ACo for lenses (Figure 4.6) which exhibited greater motion artefact and variability than any other OAR in the brain.

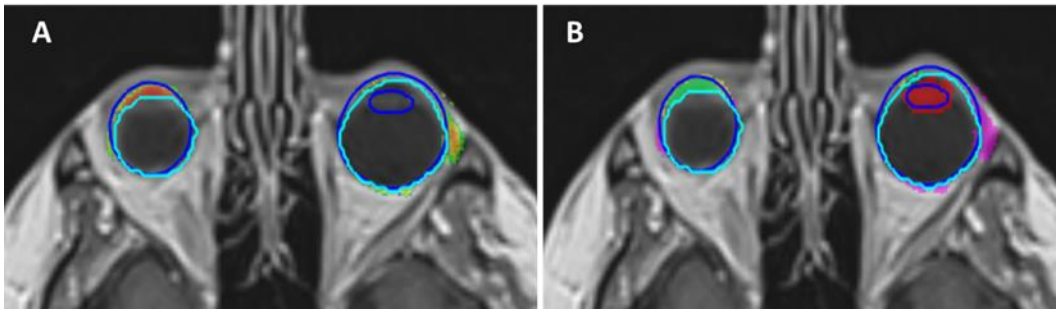


Figure 4.6: Axial T1w-Gd MRI showing the failure of AutoConfidence prediction to detect the missing autosegmentation for the left lens, resulting in false-negative. Blue represents the gold standard, while light blue represents EM-HQ autosegmentation. b) four-colour-map showing regions of TP (green), FP (pink) and GDC modified FP (yellow) relative to the differences-to-gold-standard, for the ACo prediction.

Also, ACo struggled to determine the inferior limit of brainstem (see supp info Figure S2). This limit was landmark driven, defined by the tip of the dens of C2, which was not directly adjacent to brainstem, making it difficult to learn from 2D image data alone. The performance of ACo could potentially be improved in this regard by segmenting the dens, or by a 3D ACo method. A 3D version of ACo is being investigated and will potentially resolve many of these limitations.

ACo is a model-agnostic, fully independent estimator of potential errors and combined (aleatoric and epistemic) uncertainty for medical image segmentation, at a per-voxel, per-prediction level. This approach is different to previously published DL autosegmentation QA methods based on internal DL-model probability, or Monte Carlo dropout (Claessens et al., 2022). The advantage of our approach, which does not require any additional information from, or access to, the segmentation model, is both its generalisability (deep-learning, atlas-based, or even manual contours can be assessed by ACo) and its independence. Furthermore, internal probability estimates from deep-learning classifiers are notoriously poorly calibrated, typically resulting in high confidence predictions for all voxels, except very near decision boundaries, leading to the model being ‘confident but wrong’, a scenario one would like to avoid in safety critical medical applications.

4.5 Conclusion

ACo was able to successfully predict regions of low confidence, including errors to gold standard, in both internal and external autosegmentations, from a commercially available segmentation algorithm, without reference to the gold-standard segmentations themselves. These confidence estimates did not depend on the internal confidence of the segmentation model. Indeed, they require only the proposed segmentation and underlying image, making them fully independent and applicable to segmentations from any source, including manual delineations. ACo confidence maps could serve as a per-patient, reference-free segmentation QA tool, increasing clinical confidence in autosegmentation and potentially reducing editing time, whist improving patient safety. The additional ability to detect error types of unseen during training (zero-shot detection) enhances the robustness of ACo for clinical use.

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).

Author contribution statement

Michael Nix developed AutoConfidence (ACo) model and provided essential guidance for the study design. Nouf Alzahrani was responsible for collecting, preparing, and analysing the data, model training and testing, interpreting the results, and writing the manuscript. Louise Murray and Michael Nix provided essential guidance for reviewing the analysis and interpretation of the result. Ann Henry and Bashar Al-Qaisieh contributed to reviewing and approving the study design from the clinical and technical perspectives and providing the overall guidance of the project and data.

Conflict of Interest Statement

No conflict of interest.

4.6 References

ABDAR, M., POURPANAH, F., HUSSAIN, S., REZAZADEGAN, D., LIU, L., GHAVAMZADEH, M., FIEGUTH, P., CAO, X., KHOSRAVI, A., ACHARYA, U. R., MAKARENKOV, V. & NAHAVANDI, S. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion,* 76**,** 243-297.

ALZAHRANI, N., HENRY, A., CLARK, A., MURRAY, L., NIX, M. & AL-QAISIEH, B. 2023. Geometric evaluations of CT and MRI based deep learning segmentation for brain OARs in radiotherapy. *Phys Med Biol,* 68.

APOLLE, R., APPOLD, S., BIJL, H. P., BLANCHARD, P., BUSSINK, J., FAIVRE-FINN, C., KHALIFA, J., LAPRIE, A., LIEVENS, Y., MADANI, I., RUFFIER, A., DE RUYSSCHER, D., VAN ELMPT, W. & TROOST, E. G. C. 2019. Inter-observer variability in target delineation increases during adaptive treatment of head-and-neck and lung cancer. *Acta Oncol,* 58**,** 1378-1385.

ASGHARNEZHAD, H., SHAMSI, A., ALIZADEHSANI, R., KHOSRAVI, A., NAHAVANDI, S., SANI, Z. A., SRINIVASAN, D. & ISLAM, S. M. S. 2022. Objective evaluation of deep uncertainty predictions for COVID-19 detection. *Sci Rep,* 12**,** 815.

CARDENAS, C. E., YANG, J., ANDERSON, B. M., COURT, L. E. & BROCK, K. B. 2019. Advances in Auto-Segmentation. *Semin Radiat Oncol,* 29**,** 185-197.

CHEN, X., MEN, K., CHEN, B., TANG, Y., ZHANG, T., WANG, S., LI, Y. & DAI, J. 2020. CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy. *Front Oncol,* 10**,** 524.

CLAESSENS, M., ORIA, C. S., BROUWER, C. L., ZIEMER, B. P., SCHOLEY, J. E., LIN, H., WITZTUM, A., MORIN, O., NAQA, I. E., VAN ELMPT, W. & VERELLEN, D. 2022. Quality Assurance for AI-Based Applications in Radiation Therapy. *Semin Radiat Oncol,* 32**,** 421-431.

FAGHANI, S., MOASSEFI, M., ROUZROKH, P., KHOSRAVI, B., BAFFOUR, F. I., RINGLER, M. D. & ERICKSON, B. J. 2023. Quantifying Uncertainty in Deep Learning of Radiologic Images. *Radiology,* 308**,** e222217.

ISOLA, P., ZHU, J.-Y., ZHOU, T. & EFROS, A. A. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 1125-1134.

LIESBETH, V., MICHAËL, C., ANNA, M. D., CHARLOTTE, L. B., WOUTER, C. & DIRK, V. 2020. Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. *Radiotherapy and Oncology*.

LIU, J., LIN, Z., PADHY, S., TRAN, D., BEDRAX WEISS, T. & LAKSHMINARAYANAN, B. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems,* 33**,** 7498-7512.

VAN DEN BERG, C. A. T. & MELIADO, E. F. 2022. Uncertainty Assessment for Deep Learning Radiotherapy Applications. *Semin Radiat Oncol,* 32**,** 304-318.

VAN DIJK, L. V., VAN DEN BOSCH, L., ALJABAR, P., PERESSUTTI, D., BOTH, S., R, J. H. M. S., LANGENDIJK, J. A., GOODING, M. J. & BROUWER, C. L. 2020. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiother Oncol,* 142**,** 115-123.

VAN ROOIJ, W., VERBAKEL, W. F., SLOTMAN, B. J. & DAHELE, M. 2021. Using Spatial Probability Maps to Highlight Potential Inaccuracies in Deep Learning-Based Contours: Facilitating Online Adaptive Radiation Therapy. *Adv Radiat Oncol,* 6**,** 100658.

WANG, G., LI, W., AERTSEN, M., DEPREST, J., OURSELIN, S. & VERCAUTEREN, T. 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing (Amst),* 335**,** 34-45.

YEUNG, M., RUNDO, L., NAN, Y., SALA, E., SCHÖNLIEB, C. B. & YANG, G. 2023. Calibrating the Dice Loss to Handle Neural Network Overconfidence for Biomedical Image Segmentation. *J Digit Imaging,* 36**,** 739-752.

4.7 Supplementary information

- **Research method:**

The generator is a U-net neural network that contains downsampling and upsampling paths and a Skip connection. The downsampling path includes 8 layers each has 2D Convolution layers with a rectified linear unit (*ReLU*) activation and batch normalization (batch norm). The upsampling path includes 8 layers, which 7 of have transposed 2D convolutional layer, drop out layer, and batch norm. The final layer contains transposed 2D convolution. Skip connection enhances the output of the downsampling to upsampling layers.

The discriminator is a shallow U-net neural network that contains downsampling, upsampling paths and a series of residual blocks skip connection. The downsampling path includes five layers with each having 2D Convolution layer with *ReLU* activation and batch norm. The upsampling path includes 4 layers with each having transposed 2D convolution with *ReLU* activation, drop-out layer, and batch norm.

Five blocks residual skip connection in each layer with block contains two 2D convolutional neural layers, *ReLU* activation, drop out layer, and batch norm. The output of the residual blocks it’s the output of its convolutional neural layers and the output from downsampling.

The combination of residual network architecture with U-net architecture in the discriminator is to have the ability to classify in a high voxel resolution.

Tables:

	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
						IAS							
Average	0.85	0.52	0.47	0.07	0.26	0.77	0.51	0.60	0.68	0.63	0.81	0.78	0.73
						EM-HQ							
Average	0.9	0.57	0.49	0.1	0.15	0.68	0.67	0.51	0.65	0.68	0.9	0.91	0.67
						EM-LQ							
Average	0.89	0.73	0.52	0.04	0.02	0.41	0.28	0.44	0.41	0.43	0.88	0.86	0.35

Table S1: Average DSC across OARs and segmentations.

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	4.30	4.28	10.34	10.19	55.83	63.24	14.21	15.79	9.21	4.94	1.57	1.63	110.22
FP	7.92	8.59	6.38	5.64	51.56	48.98	12.54	11.40	10.87	6.79	1.29	1.47	194.75
TN	3.59	3.03	11.54	8.25	140.71	140.91	24.37	23.18	14.80	11.66	3.24	3.25	504.71
FN	2.77	2.20	4.84	3.36	17.84	12.05	10.14	8.79	6.44	1.69	0.53	0.52	45.54
MCC	-0.08	-0.08	0.32	0.35	0.45	0.53	0.24	0.31	0.16	0.33	0.43	0.42	0.35
FPR	0.69	0.74	0.36	0.41	0.27	0.26	0.34	0.33	0.42	0.37	0.28	0.31	0.28
FNR	0.39	0.34	0.32	0.25	0.24	0.16	0.42	0.36	0.41	0.25	0.25	0.24	0.29
(FP/FN)	-	7.21	2.86	3.90	1.32	1.68	2.89	4.06	1.24	1.30	1.69	4.02	2.43

Table S2: AutoConfidence outputs for the IAS trained without synthetic errors relative to the gold standard (baseline).

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	3.75	4.02	3.39	3.18	46.98	53.72	20.28	19.64	13.27	7.24	2.82	2.66	152.65
FP	10.21	9.16	2.8	1.79	49.13	41.95	9.7	12.37	12.58	5.4	1.3	1.48	98.66
TN	3.36	3.15	3.46	4.34	156.13	157.01	23.86	21.76	12.4	11.21	1.9	2.32	568.03
FN	1.27	1.77	1.5	1.96	13.7	12.5	7.42	5.4	3.07	1.22	0.62	0.41	35.89
MCC	-0.01	-0.05	0.25	0.33	0.47	0.54	0.44	0.42	0.31	0.5	0.43	0.48	0.6
FPR	0.75	0.74	0.45	0.29	0.24	0.21	0.29	0.36	0.50	0.33	0.41	0.39	0.15
FNR	0.25	0.31	0.31	0.38	0.23	0.19	0.27	0.22	0.19	0.14	0.18	0.13	0.19
(FP/FN)	8.04	5.18	1.87	0.91	3.59	3.36	1.31	2.29	4.10	4.43	2.10	3.61	2.75

Table S3: AutoConfidence outputs for the IAS trained with synthetic errors relative to the gold standard (baseline).

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	4.2	5.77	4.4	3.98	50.68	62.96	26.62	26.93	23.15	9.94	3.73	3.49	201.97
FP	7.23	5.69	0.76	0.08	18	10.21	1.16	1.34	0.5	0.89	0.03	0.1	22.19
TN	6.32	6.01	5.04	6.45	191.13	188.49	30.24	28.52	16.92	14.07	2.83	3.23	619.51
FN	0.83	0.64	0.95	0.75	6.13	3.52	3.24	2.37	0.75	0.18	0.06	0.05	11.57
MCC	0.28	0.41	0.69	0.85	0.75	0.87	0.86	0.88	0.94	0.91	0.97	0.96	0.9
FPR	0.53	0.49	0.13	0.01	0.09	0.05	0.04	0.04	0.03	0.06	0.01	0.03	0.03
FNR	0.17	0.10	0.18	0.16	0.11	0.05	0.11	0.08	0.03	0.02	0.02	0.01	0.05
(FP/FN)	8.71	8.89	0.80	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table S4: AutoConfidence outputs for the IAS relative to the gold standard utilizing the IER and GDC.

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	3.38	3.63	5.81	4.06	34.88	40.39	23.93	25.53	22.36	14.38	3.95	1.82	95.45
FP	2.65	1.33	0.1	0.04	23.16	18.89	2.93	2.62	1.64	0.41	0.15	0.55	73.38
TN	8.59	8.52	4.48	6.35	198.59	197.06	30.24	26.42	14.66	9.36	2.36	4.33	668.32
FN	3.95	4.62	0.76	0.82	9.31	8.85	4.16	4.59	2.67	0.93	0.18	0.16	18.08
MCC	0.24	0.34	0.85	0.85	0.62	0.68	0.77	0.76	0.79	0.89	0.89	0.77	0.63
FPR	0.24	0.14	0.02	0.01	0.10	0.09	0.09	0.09	0.10	0.04	0.06	0.11	0.10
FNR	0.54	0.56	0.12	0.17	0.21	0.18	0.15	0.15	0.11	0.06	0.04	0.08	0.16
(FP/FN)	0.67	0.29	N/A	N/A	2.49	2.13	N/A	N/A	N/A	N/A	N/A	N/A	4.06

Table S5: AutoConfidence outputs for the EM-LQ AS relative to the gold standard utilizing the IER and GDC.

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	1.33	2.73	5.3	3.6	19.6	27.34	10.18	10.53	13.06	5.07	1.46	1.32	65.54
FP	5.49	4.67	0.03	0.23	37.22	34.95	9.78	7.72	5.53	3.09	1.86	0.69	83.43
TN	9.91	9.1	4.83	6.15	203.32	199.17	37.64	37.7	20.63	16.29	3.19	4.63	686.77
FN	1.85	1.6	0.98	1.28	5.8	3.72	3.65	3.21	2.1	0.63	0.12	0.22	19.5
MCC	0.05	0.25	0.83	0.73	0.44	0.55	0.47	0.55	0.63	0.65	0.47	0.67	0.52
FPR	0.36	0.34	0.01	0.04	0.15	0.15	0.21	0.17	0.21	0.16	0.37	0.13	0.11
FNR	0.58	0.37	0.16	0.26	0.23	0.12	0.26	0.23	0.14	0.11	0.08	0.14	0.23
(FP/FN)	2.97	2.92	N/A	N/A	6.42	9.40	2.68	2.40	2.63	4.90	15.50	3.14	4.28

Table S6: AutoConfidence outputs for the EM-HQ AS relative to the gold standard utilizing the IER and GDC.

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	1.51	2.57	2.69	2.66	32.05	42.8	20.16	18.06	13.02	6.73	2.96	2.67	146.27
FP	9.92	8.89	2.46	1.39	36.62	30.38	7.62	10.21	10.63	4.1	0.8	0.91	77.89
TN	5.87	4.98	4.08	4.83	178.34	176.13	25.84	25.4	14.51	12.98	2.23	2.87	593.82
FN	1.28	1.66	1.92	2.38	18.92	15.88	7.64	5.48	3.16	1.27	0.66	0.41	37.26
MCC	-0.06	-0.03	0.2	0.32	0.41	0.54	0.5	0.47	0.38	0.57	0.56	0.62	0.64
FPR	0.63	0.64	0.38	0.22	0.17	0.15	0.23	0.29	0.42	0.24	0.26	0.24	0.12
FNR	0.46	0.39	0.42	0.47	0.37	0.27	0.27	0.23	0.20	0.16	0.18	0.13	0.20
(FP/FN)	7.75	5.36	1.28	0.58	1.94	1.91	1.00	1.86	3.36	3.23	1.21	2.22	2.09

Table S7: AutoConfidence outputs for the IAS relative to the gold standard utilizing the IER alone.

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	1.6	2.56	3.05	2.26	19.45	25.02	11.2	12.11	8.93	7.13	0.97	0.51	45.44
FP	4.44	2.4	2.85	1.84	38.59	34.25	15.66	16.04	15.07	7.66	3.13	1.85	123.39
TN	7.21	5.47	3.76	5	192.27	185.76	29.14	24.86	13.75	8.96	2.31	4.24	660.84
FN	5.34	7.67	1.49	2.16	15.63	20.15	5.25	6.15	3.58	1.33	0.23	0.25	25.57
MCC	-0.16	-0.06	0.24	0.25	0.32	0.36	0.3	0.25	0.18	0.37	0.18	0.24	0.33
FPR	0.38	0.30	0.43	0.27	0.17	0.16	0.35	0.39	0.52	0.46	0.58	0.30	0.16
FNR	0.77	0.75	0.33	0.49	0.45	0.45	0.32	0.34	0.29	0.16	0.19	0.33	0.36
(FP/FN)	0.83	0.31	1.91	0.85	2.47	1.70	2.98	2.61	4.21	5.76	13.61	7.40	4.83

Table S8: AutoConfidence outputs for the EM-LQ AS relative to the gold standard utilizing the IER alone.

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	0.62	1.4	2.64	1.77	9.53	12.34	4.34	4.41	5.49	2.12	0.34	0.4	37.23
FP	6.2	6.01	2.69	2.06	47.29	49.96	15.62	13.84	13.09	6.03	2.99	1.61	111.74
TN	9.37	7.79	4.22	5.13	199.15	195.45	36.66	37.17	19.7	15.99	3.17	4.55	677.48
FN	2.39	2.91	1.59	2.3	9.97	7.44	4.63	3.74	3.03	0.94	0.15	0.3	28.78
MCC	-0.15	-0.1	0.23	0.15	0.19	0.26	0.14	0.2	0.2	0.29	0.11	0.21	0.3
FPR	0.40	0.44	0.39	0.29	0.19	0.20	0.30	0.27	0.40	0.27	0.49	0.26	0.14
FNR	0.79	0.68	0.38	0.57	0.51	0.38	0.52	0.46	0.36	0.31	0.31	0.43	0.44
(FP/FN)	2.59	2.07	1.69	0.90	4.74	6.72	3.37	3.70	4.32	6.41	19.93	5.37	3.88

Table S9: AutoConfidence outputs for the EM-HQ AS relative to the gold standard utilizing the IER alone.

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	12.52	11.81	6.14	4.95	93.22	95	29.95	31.88	25.84	12.54	4.12	4.13	249
FP	1.44	1.37	0.05	0.01	2.89	0.67	0.03	0.13	0.02	0.11	0	0	2.32
TN	4.08	4.26	4.66	5.86	168.07	168.22	28.49	25.01	14.79	12.3	2.47	2.68	593.24
FN	0.54	0.66	0.3	0.44	1.75	1.29	2.79	2.15	0.67	0.14	0.06	0.05	10.69
MCC	0.74	0.73	0.94	0.92	0.96	0.98	0.91	0.92	0.96	0.98	0.98	0.98	0.96
FPR	0.26	0.24	0.01	0.00	0.02	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00
FNR	0.04	0.05	0.05	0.08	0.02	0.01	0.09	0.06	0.03	0.01	0.01	0.01	0.04
(FP/FN)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table S10: AutoConfidence outputs for the IAS relative to the gold standard utilizing the GDC alone.

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	6.56	6.29	6.26	4.64	78.97	78.87	32.56	32.95	26.21	15.83	4.16	2.92	222.11
FP	1.4	0.74	0.03	0.01	8.06	5.29	0.24	0.48	0.23	0.15	0.06	0.15	25.33
TN	7.81	8.32	4.25	5.93	172.83	176.22	24.07	21.08	12.32	8.3	2.23	3.6	593.16
FN	2.82	2.74	0.61	0.68	6.08	4.8	4.4	4.65	2.56	0.81	0.18	0.19	14.64
MCC	0.55	0.63	0.89	0.88	0.88	0.91	0.85	0.83	0.86	0.92	0.92	0.9	0.89
FPR	0.15	0.08	0.01	0.00	0.04	0.03	0.01	0.02	0.02	0.02	0.03	0.04	0.04
FNR	0.30	0.30	0.09	0.13	0.07	0.06	0.12	0.12	0.09	0.05	0.04	0.06	0.06
(FP/FN)	0.50	0.27	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table S11: AutoConfidence outputs for the EM-LQ AS relative to the gold standard utilizing the GDC alone.

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	8.55	8.55	6.08	4.62	76.45	79.14	27.32	25.64	21.91	10.37	3.16	2.81	196.69
FP	2.1	1.86	0.04	0.01	12.93	13.02	2.67	2.15	0.72	1.01	0.87	0.11	32.65
TN	6.78	6.7	4.33	5.84	172.2	170.11	27.43	27.98	16.75	13.21	2.46	3.71	609.03
FN	1.16	1	0.7	0.79	4.35	2.91	3.84	3.39	1.94	0.48	0.15	0.23	16.87
MCC	0.65	0.68	0.87	0.87	0.85	0.87	0.79	0.81	0.87	0.88	0.71	0.9	0.85
FPR	0.24	0.22	0.01	0.00	0.07	0.07	0.09	0.07	0.04	0.07	0.26	0.03	0.05
FNR	0.12	0.10	0.10	0.15	0.05	0.04	0.12	0.12	0.08	0.04	0.05	0.08	0.08
(FP/FN)	1.81	1.86	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table S12: AutoConfidence outputs for the EM-HQ AS relative to the gold standard utilizing the GDC alone.

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	2.41	3.65	3.21	2.58	33.08	38.27	13.96	14.71	10.21	7.59	1.02	0.83	82.76
FP	5.55	3.39	3.08	2.08	53.94	45.89	18.84	18.72	16.23	8.39	3.21	2.24	164.67
TN	5.23	4.23	3.53	4.62	162.99	163.06	22.76	19.54	11.28	7.88	2.18	3.49	580.44
FN	5.39	6.83	1.33	1.99	15.92	17.96	5.70	6.19	3.60	1.23	0.23	0.30	27.36
MCC	-0.21	-0.10	0.24	0.25	0.35	0.40	0.24	0.21	0.15	0.34	0.18	0.26	0.39
FPR	0.51	0.44	0.47	0.31	0.25	0.22	0.45	0.49	0.59	0.52	0.60	0.39	0.22
FNR	0.69	0.65	0.29	0.44	0.32	0.32	0.29	0.30	0.26	0.14	0.18	0.27	0.25
(FP/FN)	1.03	0.50	2.32	1.05	3.39	2.56	3.31	3.02	4.51	6.82	13.96	7.47	6.02

Table S13: AutoConfidence outputs for the EM-LQ AS relative to the gold standard (baseline).

	Lens L	Lens R	Lacrimal Gland L	Lacrimal Gland R	Orbit L	Orbit R	Optic Nerve L	Optic Nerve R	Optic Chiasm	Pituitary Gland	Cochlea L	Cochlea R	Brainstem
TP	2.88	3.15	3.04	2.32	26.18	27.69	8.94	8.68	7.76	3.77	0.71	0.84	74.00
FP	7.76	7.26	3.08	2.32	63.21	64.46	21.05	19.12	14.87	7.62	3.32	2.08	155.35
TN	5.13	4.74	3.66	4.58	165.33	164.47	26.14	26.83	15.61	12.71	2.42	3.60	594.24
FN	2.80	2.96	1.37	2.05	11.22	8.56	5.12	4.54	3.08	0.98	0.19	0.34	31.65
MCC	-0.09	-0.09	0.23	0.19	0.31	0.35	0.16	0.20	0.20	0.33	0.15	0.26	0.37
FPR	0.60	0.61	0.46	0.34	0.28	0.28	0.45	0.42	0.49	0.37	0.58	0.37	0.21
FNR	0.49	0.48	0.31	0.47	0.30	0.24	0.36	0.34	0.28	0.21	0.21	0.29	0.30
(FP/FN)	2.77	2.45	2.25	1.13	5.63	7.53	4.11	4.21	4.83	7.78	17.47	6.12	4.91

Table S14: AutoConfidence outputs for the EM-HQ AS relative to the gold standard (baseline).

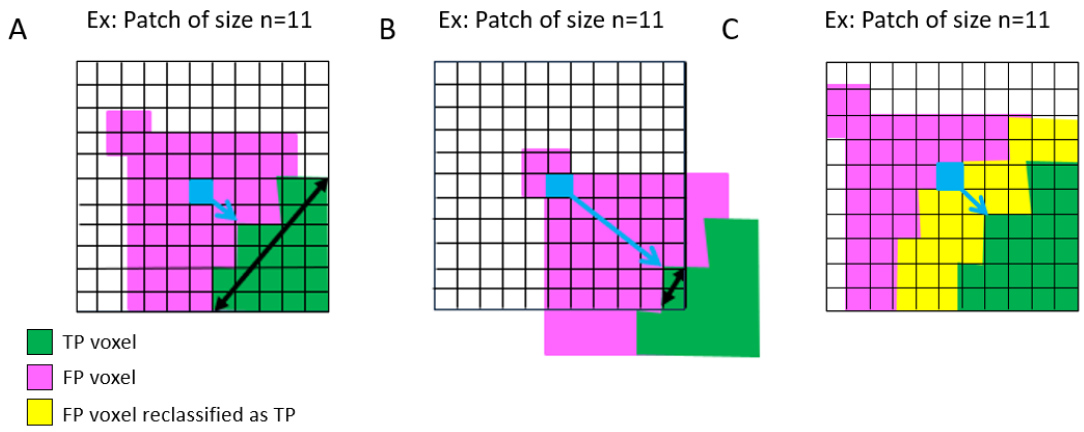


Figure S1: a and b) Illustration of GDC in different scenarios. The voxel under consideration is the central blue voxel. FP voxels are represented in pink, TP voxels in green. A) Example conversion of a FP voxel to TP. Blue arrow (D(fp_tp)) is shorter than the black arrow (D(tp)_max). B) A voxel which does not quality for conversion, as it lies too far from the TP region (blue arrow longer than black arrow). C) The final classification, showing converted voxels in yellow.

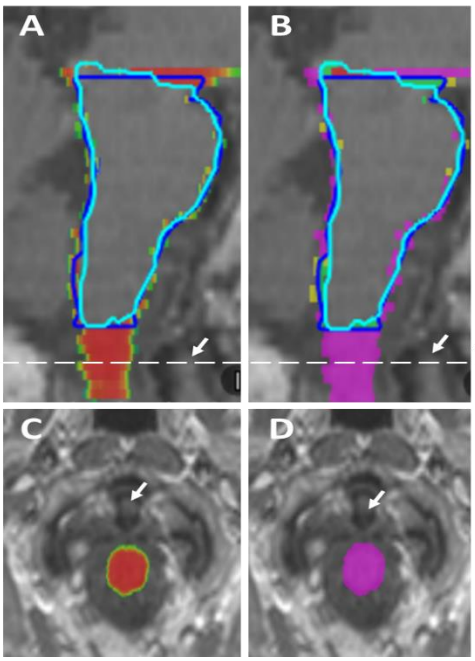


Figure S2: a) Sagittal (dotted line represents axial slice), and c) axial T1w-Gd MRI showing ACo overprediction of uncertainty beyond the inferior limit of brainstem, resulting in false-positives. Blue represents the gold standard contours; light blue represents EM-HQ AS. The white arrows show the tip of the dens of C2, the anatomical definition of the inferior limit of brainstem. b) Sagittal and d) axial four-colour-map showing regions of TP (green), FP (pink) and GDC modified FP (yellow) relative to the differences to gold-standard, for the ACo prediction.

Chapter 5 Discussion, future works, and conclusion

5.1 Summary of research aim and studies

AI research and applications have grown rapidly in radiation therapy, aiming either to improve treatment quality or save time, or both. The foremost example of this technology is auto-segmentation for treatment planning. This PhD thesis focuses on the practical use of auto-segmentation technology in brain radiotherapy. It explores the performance of deep learning auto-segmentation (DL-AS) models for delineating OARs in brain radiotherapy, using the deep-learning framework within the RayStation Treatment Planning System. Also, it investigates auto-contouring accuracy on a per-patient basis using a novel AI-driven QA method called AutoConfidence (ACo).

Three studies were designed and performed, to achieve a comprehensive evaluation of accuracy and usability of DL-AS models in delineating brain OARs. Several models were investigated, comprising CT and MRI based models, with varying degrees of pre-training editing of the labelled data. The objective of this was to determine the optimal balance of model performance against (costly) editing of training data.

The first study concerned geometric evaluation of DL-AS for brain OARs in radiotherapy (Chapter 2). The second study explored the dosimetric impact of contour editing on CT and MRI Deep-Learning autosegmentation for Brain OARs (Chapter 3). The third study concerned confidence estimation in MRI deep learning segmentation (MRI DL-AS) for Brain OARs in Radiotherapy (Chapter 4)

5.2 Objectives

5.2.1 Investigating the impact of editing the clinical contours before training DL-AS models on geometric and dosimetric accuracy

Accurate OAR delineation is an essential step for effective treatment planning, as every following step of treatment planning and delivery depends on the quality of the delineation. Typically, the delineation of brain OARs in treatment planning requires the use of both CT and MR imaging together.

Recently, the clinical workflow for many sites, especially the brain, is moving towards MRI only radiotherapy planning (Edmund and Nyholm, 2017). MRI images have greater soft tissue contrast than CT, making it more effective in

delineating exact boundaries for both target and OARs to reduce geometric uncertainty and limit the dose to critical structures (Liu et al., 2019). MRI only planning may also reduce the uncertainty produced in the final dose plan, which derives from inaccuracy in registration between CT and MRI (Lerner et al., 2021). Hence, I was motivated to investigate how to utilize a commercial MRI DL-AS model efficiently, effectively, and safely for brain OARs delineations.

Due to scanner-dependent intensity and contrast, where scanner contrast can be affected by variations in the scan parameters such as Repetition Time (TR), Time to Echo (TE), and differences in the manufacturing of the scanners, the transferability and generalizability of MRI DL-AS between centres can be severely limited (Fatania et al., 2022). As a result, centre or scanner specific models may be required, severely limiting the available data for training and validation. Hence, the trade-off between training data volume and accuracy may be different for MRI DL-AS as compared to CT, due to the limited dataset setting. Feature normalisation is a standard approach in Radiomic analysis (Demircioglu, 2024) but as DL-AS works directly on images, not radiomic features, image normalisation is required. Hence, methods such as z-score or white-stripe normalisation may improve the generalisability of DL-AS models for MRI between centres.

Accordingly, in this work, the impact of editing clinical contours to be consistent prior to training the DL-AS model was investigated due to the limited availability of MRI data for training in addition to limited transferability of MRI DL-AS models to other centres. The focus of this part of the thesis is MRI DL-AS models, while investigations of the impact of editing on CT DL-AS model performance was used as a reference for comparison with the MRI model results.

Chapters 2 and 3 of this thesis involve a comprehensive evaluation of the geometric and dosimetric impact of CT and MRI-based DL-AS for brain OARs. The primary goal was to establish their clinical safety and utility, with a further focus on investigating the impact of editing the clinical contours before model training to be consistent with contouring guidelines. Models based on edited and unedited clinical contours were compared to investigate the impact of editing on CT and MRI segmentation accuracy. Thirdly, I investigated the correlation between geometric and dosimetric measures, to determine if the geometric evaluation is sufficient to establish segmentation safety and utility in clinical use.

5.2.2 Utilising AutoConfidence as a method to estimate DL-AS autosegmentation uncertainty.

Despite careful contour editing, training and pre-clinical validation, DL-AS models often produce errors in the clinical setting. These errors may be systematic (epistemic – deriving from limitations of the model), or random (aleatoric – deriving from variability in the input data) (van den Berg and Meliado, 2022, Abdar et al., 2021, Wang et al., 2019a).

Factors such as individual patient anatomical variations, low image quality, and artifacts can produce these errors (Claessens et al., 2022, Mackay et al., 2023).Such data issues can occur during training or in clinical use. Whilst careful data selection, pre-processing and editing can mitigate some errors at a model level, variability of new data in clinical use is harder to deal with.

Accordingly, the output of DL-AS should always be reviewed, edited, and approved by the clinician or dosimetrist before treatment planning. This review and editing process is time-consuming and may introduce interobserver variability as it is user-dependent, and it may reduce the productivity benefits of segmentation automation (Claessens et al., 2022, Mackay et al., 2023).

To address this issue, an AI QA method, called AutoConfidence (ACo) was built for DL-AS, aiming to estimate segmentation confidence on a per-patient basis, without a gold standard, enabling robust, efficient review of MRI DL-AS brain OARs. This ACo model is described and evaluated in Chapter 4.

ACo performance was evaluated on estimation of uncertainty for two different contouring sources. It was tested on segmentations from its own internal U-net, against which it was adversarialy trained, and separately validated on the segmentations generated from the independent (external to ACo) MRI DL-AS models described in Chapters 2 and 3 (MRIu and MRIeMRI). They have different qualities (external model with low-quality AS (EM-LQ, MRIu) and external model with high-quality AS (EM-HQ, MRIeMRI)).

The objective of this study was to determine the potential utility of ACo in identifying DL-AS contouring errors, and regions of low confidence, such that human operators' attention could be focussed on likely problematic regions of the segmentation. The intention is that this could make review of DL-AS both more efficient and more robust to inter-observer variability.

5.3 Summary of main findings

5.3.1 Geometric assessments (Chapter 2)

5.3.1.1 CT MRI DL-AS and MR DL-AS overall segmentation quality

MRI and CT DL-AS models provided excellent contouring quality for larger structures such as orbits and brainstem, apart from the unedited CT model which performed poorly on these structures. Average DSC and sensitivity scores ranged from 0.85 to 0.90 and from 0.85 to 0.99, respectively, across all three MRI DL-AS models and CTeCT model for these large structures (Alzahrani et al., 2023). The CTu DL-AS 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 (tables S4-S5) (Alzahrani et al., 2023). This is expected from the DL-AS model as it is known to perform well with large structures. However, the CTu model performed poorly in delineating the orbits, which is related to the quality of the training data. For orbits, some operators precisely contouring the exact shape of the orbits on the CT scan, while others simplified this step by using a standard spherical shape in the treatment planning system. This variability could explain the inferior performance of CTu DL-AS in the delineation of the orbits. Regarding the brainstem, there is variability between the operator in the superior and inferior limits in addition to difficulty in visualising the exact boundaries of the brainstem on all the CT slices. This is not the case with other models (CTeCT, MRIeCT and MRIeMRI), as the clinical contour was edited based on guidelines from the atlas by one person. Moreover, it was reviewed by one experienced neuro-radiation oncologist. Notably, the CTu model produced incorrectly located segmentations in 3 cases, (DSC = 0). Editing clinical contour prior training the model helped to resolve these failures (result section 3.3.2, Chapter 3).

The same issues did not occur in the case of the MRIu model due to the high soft tissue contrast of the images used in the training, which helped the model to perform better to find exact boundaries and location of the structure.

Regarding small structures, accuracy of the MRI and CT DL-AS model delineation is more challenging compared to the larger structures (tables S1, S2, S3, S4, S5, and S6) (Alzahrani et al., 2023). The geometric assessment indicated that MRI DL-AS models performed worse than CT for contouring the lacrimal gland (Chapter 2, tables S1-S2), which is unsurprising given the difficulty of

visualising this structure on MRI. In contrast, the geometric assessment indicated that CT DL-AS performed particularly poorly in the segmentation of the optic chiasm, (tables S4 and S5)(Alzahrani et al., 2023), which is known to be essentially invisible on CT imaging.

5.3.1.2 Impact of editing

Our investigation showed the value of editing clinical contours before training the DL-AS models. I found that depending on the model and structure, editing to gold standard is either essential or beneficial. I determined that editing is essential based on the segmentation failure numbers, while assessment of benefit was based on the statistical significance of comparisons between geometry test metrics for paired edited and unedited models in each modality.

5.3.1.2.1 Pre-training editing of clinical contours is necessary for successful segmentation

Editing was found to be necessary for both CT and MRI, as the models trained with edited clinical contours produced a greater number of successful segmentations on OARs than unedited segmentation (Alzahrani et al., 2023).

From the supplementary information: Chapter 2, table S7, it can be seen the number of failed segmentations reduced when using MRIeMRI DL-AS model, compared to MRIeCT and MRIu for the following structures; cochlea, lenses, optic chiasm and pituitary (Alzahrani et al., 2023).

For CT DL-AS, the number of failed segmentations reduced when using CTeCT DL-AS model, compared to CTu DL-AS model for the following structures; optic nerves, lacrimal glands, cochlea, lenses, optic chiasm and pituitary (Alzahrani et al., 2023).

5.3.1.2.2 Pre-training editing of clinical contours is beneficial for segmentation quality

Except for the lacrimal glands, table 2.1 in Chapter 2 revealed that editing the clinical contours on MRI before training the model is desirable. For orbits, lenses, optic nerves, pituitary, and optic chiasm, autosegmentation geometric performance improved on at least one metric (Alzahrani et al., 2023).

Regarding CT DL-AS model, no geometric improvement was observed from pre-training editing of clinical contours (Chapter 2, tables S8). Interestingly, optic chiasm delineated by the DL-AS model without editing (CTu) showed better

geometric performance than the edited version CTeCT DL-AS model (tables S8) (Alzahrani et al., 2023). This is likely due to the fact that the original optic chiasms were segmented clinically on MRI, and subsequent editing on CT in fact degraded the quality, due to the extremely poor CT contrast for this organ.

5.3.2 Dosimetric assessments (Chapter 3)

5.3.2.1 CT DL-AS and MR DL-AS

For MRI DL-AS models, the greatest dosimetric change between the DL-AS and gold standard contours was found for the lacrimal glands D1% (Δ absolute Average dosimetric change ≤ 143% across all MRI models) (Chapter 3, table 3.2, figures 3.4a), which also showed the worst performance on geometric analysis.

For CT DL-AS models, the greatest dosimetric change between the DL-AS and gold standard contour was revealed for the right lens D1% (Δ absolute Average dosimetric change =57%) and optic chiasm D1% (Δ absolute Average dosimetric change =18%) delineated by the CTu and CTeCT models, respectively (Chapter 3, table 3.3 and figure 3.3). The optic chiasm was worst performing geometrically as discussed. The Δ absolute average dosimetric change of the right lens was affected by one outlier (446% worst-case for the CTu model) (result section 3.5.2, Chapter 3) (Chapter 3, figure 2S: b). The failure of this case was discussed in the discussion section, Chapter 3). However, the dosimetric change relative to the gold standard in the remaining cases ranged from 0% to 5% in either direction, and two cases were not segmented by the model.

The remaining OARs showed less dosimetric change relative to the gold standard on MRI and CT (Δ absolute Average dosimetric change ≤ 20% and ≤ 25% across all MRI DL-AS and CT DL-AS models respectively) (Chapter 3, tables 3.2 and 3.3 and figures 3.2 and 3.3)

The largest differences in dosimetric change were found between the MRI and CT DL-AS models in the lacrimal glands D1%. The CT DL-AS models showed less dosimetric change vs gold standard contour compared to the MRI DL-AS models. (Chapter 3, figure 3.4a and b).

5.3.2.2 Impact of editing

Except for the right orbit, all the dosimetric assessments were not statistically significant between the CT DL-AS models or the MRI DL-AS models, for any investigated brain OARs. A statistically significant effect (*P*= 0.012, effect size (Δ

median dosimetric change) = 7%) was found between the MRIeMRI DL-AS and MRIu DL-AS models in the delineation of right orbit, although this was considered clinically insignificant, and likely resulted from slight misalignment of clinical gold standard unedited contours with the MRI images, resulting from imperfect rigid registration (result section 3.3.2, discussion, Chapter 3).

However, editing the clinical contours before training reduced the dosimetric differences between doses computed for contours from the DL-AS models and gold standard contours. This was consistently observed in both modalities for certain structures, even though it was not statistically significant.

Compared to MRIeCT and MRIu DL-AS models, the MRIeMRI DL-AS model demonstrated less average dosimetric change, relative to the gold standard contour for the following structures: optic nerves, orbits, and optic chiasm (Chapter 3, table 3.2 and figure 3.2).

Similarly, compared to the CTu DL-AS model, the CTeCT model demonstrated smaller average dosimetric changes for the following structures: lenses, orbits, and brainstem (Chapter 3, table 3.3 and figure 3.3). Notably, the CTu model produced incorrectly located contouring in 3 cases (DSC = 0) for several of these OARs, leading to gross changes in dose.

5.3.2.3 Clinical significance of dosimetric evaluation

The analysis showed that the number of cases exceeding the derived clinical significance threshold for optic chiasm (D1%) was higher compared to other OARs (brainstem D5%, orbits D1%, optic nerves D1%, and cochlea D50%) in both modalities (Chapter 3, tables 3.5 and 3.6). The number of cases that showed clinically significant error was greater in the models that were based on clinical contours edited on CT (7 cases for both MRIeCT and CTeCT DL-AS models) (Chapter 3, tables 3.5 and 3.6). The inferior performance resulting from editing the clinical contour of optic chiasm on CT was previously discussed as being due to the extremely poor CT contrast for the optic chiasm, leading to worse dosimetric performance and clinically significant dosimetric changes (Discussion section, Chapter3).

5.3.2.4 Correlation between the dosimetric and different geometric outputs

A weak, correlation was demonstrated between the dosimetric and geometric outputs for all OARs. This correlation indicates that whilst the dosimetry is dependent on the contouring accuracy, it is also strongly dependent on other factors, such as the location of high dose gradients, which is patient specific. (Chapter 3, table 3.4).

5.3.3 Automated Confidence estimation (Chapter 4)

This section describes the results of the AutoConfidence (ACo) model trained to independently assess confidence in autosegmentations generated with independent deep-learning models.

5.3.3.1 General performance of the ACo model trained with synthetic errors on internal and external sources of autosegmentation

The ACo model showed generally excellent performance on the internal autosegmentations produced from the generator (IAS) (Chapter 4, figure 4.3a, table S4) and external DL-AS from RayStation (MRIeMRI and MRIu) (Chapter 4, figure 4.3a, tables S5-S6) for all brain OARs except lenses. Lenses showed more motion artefacts and variability compared to other OARs in the brain.

5.3.3.2 Importance of synthetic errors and post-processing

The ACo model trained with several synthetic errors outperformed the ACo model trained without synthetic errors (Chapter 4, table 2S-3S), indicating the importance of including such errors during training. These errors allowed the ACo network to maintain its ability to detect significant errors even as the quality of generated segmentations improved during adversarial training.

Intelligent edge removal (IER) was required to achieve good clinical utility (by removing small, distracting errors along boundaries). Geometric distance correction (GDC) was also required to account for regions of uncertainty overprediction, relative to ground-truth error. AutoConfidence was designed to detect uncertainty, as well as error, but was validated against error, as ground truth for ‘uncertainty’ does not exist. Hence, overprediction was common but clinically acceptable (even desirable).

5.3.3.3 Visual assessment of the confidence map

The visual assessment of the resulting confidence maps demonstrated that ACo could highlight inaccurate or missing segmentations and define areas of low confidence related to poor image contrast or artifact (except for lenses, as discussed above).

5.3.4 Summary of implications of results

The findings presented in chapters 2 and 3 provide valuable insights into the requirements for developing a robust MRI autosegmentation model for brain OARs. The importance of defining and editing the training data to achieve clinically acceptable segmentation quality for all brain OARs was demonstrated, enabling acceptable performance even when using a limited dataset. Moreover, the value of the comprehensive assessment of geometry and dosimetry measures to determine the clinical usability of DL-AS models for radiotherapy planning and delivery was clearly highlighted.

The findings reported in Chapter 4 provide vital insights into the need for, and an approach to, developing an independent quality assurance (QA) tool. This tool can be used clinically to improve the efficiency, confidence, and safety of DL-AS in the clinic.

5.4 Overall discussion

The subsequent discussion considers the importance of integrating several geometric and dosimetric evaluation methods with the newly developed ACo AI-QA tool to assess contour accuracy and clinical utility in practice.

5.4.1 Geometric evaluation

5.4.1.1 Geometric evaluation alone

5.4.1.1.1 Assessment of contour accuracy using different metrics

Geometric evaluation is the most common method to assess DL-AS quality relative to the gold standard contours (Baroudi et al., 2023, Mackay et al., 2023). Based on the latest review study published in 2023 on the metrics applied to evaluate autocontouring tools in Radiotherapy, it was found that (115/117) 98.3% of studies used geometric evaluation to evaluate contouring accuracy (Mackay et al., 2023). Geometry is easy to calculate and generalise (Baroudi et al., 2023,

Harrison et al., 2022), but may be of limited clinical relevance in many cases. The evaluation is accomplished by assessing how well the DL-AS model outlines OARs on the medical image relative to the gold standard.

In this project, three different metrics were used to evaluate the geometric accuracy, as described in the method section of Chapter 2 (Alzahrani et al., 2023): the Dice Similarity Coefficient, mean distance to agreement, and sensitivity. The dice similarity coefficient (DSC) was used to evaluate the similarity between the gold standard contours and the automatic contours from each model by calculating the overlap volume between them (Wong et al., 2020). The sensitivity was used to evaluate the ability of the autosegmentation model to correctly predict the pixels located within the OAR gold standard contour (van Rooij et al., 2019). The mean distance to agreement (MDA) is the mean of the pointwise distance required to make the outline points of the automatic segmentation and outline points of the gold standard contour match perfectly. All three methods were used, to overcome the limitations inherent in each metric, ensuring they complement each other effectively to be able to assess the contouring accuracy robustly.

5.4.1.1.2 Importance of geometric metric results

The outputs of geometric metrics highlighted the number of failed segmentations for each model (table S7) (Alzahrani et al., 2023). It showed that editing clinical contours prior to training the model reduced the number of failed segmentations compared to the unedited models in both modalities (CT and MRI).

This result emphasizes the value of editing the clinical structures before model training, to ensure accuracy and standardisation. Moreover, the outputs of geometric assessment provided a comprehensive overview of the accuracy of each autosegmented structure, across different models and modalities. This highlighted the strengths and limitations of each model. For example, the geometric assessment results revealed that the MRI DL-AS models have a limitation in delineating the lacrimal glands, while the CT DL-AS models have a limitation in delineating the optic chiasm (table S1-6, Chapter 2).

Also, the results suggested structures that benefited from pre-training editing of the clinical contours on MRI. Editing on MRI of pituitary, orbits, optic nerves, lenses, and optic chiasm was key to improving the geometric performance (table

2.1) (Alzahrani et al., 2023). However, there was no apparent benefit in accuracy from editing the clinical contours prior to training CT models to improve their geometric outputs, beyond the increased number of successful segmentations (table S7-8) (Alzahrani et al., 2023).

Accordingly, the geometric results made it easier to determine which structures needed to be edited pre-training to build a robust DL-AS model.

5.4.1.2 Integration of geometric output with the visual inspection

Performing visual assessment combined with the geometric output guided us in identifying where the geometric discrepancies between the gold standard contour and generated segmentation most frequently occurred. This assessment also enables editing guidance for clinicians working with contours from these models to be produced during clinical commissioning.

Regarding MRI DL-AS models, in the case of the brainstem, it was necessary to check the superior part of the generated contours, which were often inconsistent and deviated from contouring guidelines. In some cases, the model may produce incomplete segmentation of some slices, such as in the optic chiasm and pituitary. The geometric differences for the optic nerves mainly concerned their posterior limit, while for cochlea and orbits, these were mostly occurred at the superior and inferior limits.

This information will guide staff to know where and what they are specifically required to verify or edit, on autogenerated contours, alongside their overall assessment. Notably, the frequency of these errors from the highest quality (MRIeMRI) model was less than other MRI DL-AS models based on both the geometric assessment and visual inspection, except for the lacrimal glands.

5.4.1.3 Limitations of geometric evaluation

Even though several geometry metrics with different perspectives (volumetric and distance metrics) were used, outputs did not fully demonstrate the clinical acceptability of autosegmentations. Geometric evaluations did not consider the proximity of the structure to the target volume (dose gradient) or the structure size (Sherer et al., 2021). Moreover, there is no standard cut-off value for geometric metrics that is associated with clinical acceptability as it would vary based on the structure size and treatment site (Sherer et al., 2021).

These limitations imply that, even with the high geometric similarity between the gold standard contour and autosegmentation, the clinical applicability is not guaranteed. Therefore, this study was not limited to the geometric assessment. The dosimetric evaluation was also needed to investigate the clinical suitability of the autosegmentation model.

5.4.2 Dosimetric evaluation

5.4.2.1 Dosimetric evaluation alone

5.4.2.1.1 Assessment of clinical relevance using different approaches

Dosimetric evaluation was performed by using the previous, clinically delivered, treatment plan to determine the impact of a geometric discrepancy between the gold standard contour and autosegmentation on dose (as described in Chapter 3, method section 3.2.3). Ideally, dosimetric discrepancy would have been evaluated using dose reoptimized using the autosegmented OARs, but the RT plan data was unavailable.

Dosimetric analysis is less popular than geometric assessment, based on a recent review article about metrics for the assessment of autocontouring accuracy. Dosimetric analysis was used only in 23.1% (27/117) of previous studies (Mackay et al., 2023). It is challenging for many reasons: There is no agreed threshold for clinical acceptance; a treatment plan is required, and results are affected by the beam arrangement, proximity of OARs to the high dose region, and dose constraints. All these factors, therefore, could impact the overall assessment (Sherer et al., 2021).

In our work, I investigated the impact on dosimetry of editing the clinical contours to meet a gold standard definitions pre-training. I did this by comparing the edited and unedited models. I also assessed the clinical importance of the dosimetric errors for brain OARs in each DL-AS model, using a pragmatic approach as described in the method section 3.2.4 of Chapter 3.

The outputs of dosimetric evaluations directly showed the impact of contour accuracy on the treatment plan. Furthermore, the clinical importance method for assessing dosimetric variation examined alignment with the clinical dosimetric constraints and thus potential organ toxicity. This method guided us in detecting

the dosimetric variations of greatest clinical significance (result section 3.3.5, Chapter 3), even when the geometric performance was good.

5.4.2.2 Integration of dosimetric, geometric outputs and visual inspection

In the analysis reported in results section 3.3.4, table 3.4, Chapter 3, the Pearson correlation coefficient was used to find the correlation between the geometric and dosimetric results. The results illustrated a weak correlation between the geometric error and absolute dosimetric change. In general, this was related to the several factors affecting dosimetric results that do not affect geometric analysis, such as the location of the high dose region and high dose gradients. This location is patient specific and not correlated with the regions prone to geometric errors. Herein, we examined some of the possible scenarios and how they affect the geometric to dosimetric result correlation.

Combining the dosimetric, geometric outputs, and visual inspection helps to investigate the reasons for the weak correlations between them, and the situations in which one may be more clinically useful than the other. One can consider this analysis as an analogue of the well-known 'confusion matrix' of statistical testing. Instead of true positive, false positive, true negative, and false negative, the four categories are based on agreement (or otherwise) of the geometric and dosimetric tests, as shown in figure 5.1.

Each of these categories was examined as a scenario, using typical and illustrative examples of each situation found via visual assessment during geometric analysis.

Scenario 1 **LARGE Geometric error** **LARGE Dosimetric error**	**Scenario 2** **LARGE Geometric error** **SMALL Dosimetric error**
Scenario 3 **SMALL Geometric error** **LARGE Dosimetric error**	**Scenario 4** **SMALL Geometric error** **SMALL Dosimetric error**

Figure 5.1: The four scenarios described in this chapter, based on the relative magnitude of geometric and dosimetric errors.

5.4.2.2.1 Scenario 1: Large geometric error with large dosimetric impact

There were two ways in which large geometric errors were found to be correlated with large dosimetric errors. In some cases, poor geometric agreement which led to significant changes in dose (figure 5.2: a and b) occurred at steep dose gradient regions. For example, the error in delineation of the optic nerve shown in figure 5.2 crossed the dose gradient of the patient plan, and hence correlated with high dosimetric error, and with the predicted geometric error from AutoConfidence.

In other cases, the large relative change in dosimetric statistics was due to the complete or partial failure to segment an organ which did not lie close to a high dose gradient region on the gold standard segmentation. This was mainly observed for some structures delineated by the CTu DL-AS model. The geometric overlap between the gold standard and autosegmentation was low or zero due to missing autosegmentation on some or all slices.

If the segmentation had completely failed, this would be associated with a significant relative change in dose, even though the structure was far from the planning target volume (PTV) (e.g., lens). In this scenario, the relative change in dose was associated strongly with the geometric error, although the clincial significance would be low, due to the low absolute dose.

However, the CTu model occasionally produced a segmentation in a completely inappropriate location (figure 5.3: a-c). Here, the incorrect segmentation for the right lens was in a ventricle, and close to the PTV boundary, even though the physical lens (and gold-standard segmentation) was not. As the incorrect segmentation was near a high dose gradient, this led to a large absolute dose change. In general, this would not be the case, for a simple missing OAR, but this extreme example demonstrates how dosimetric results can be misleading without geometric context.

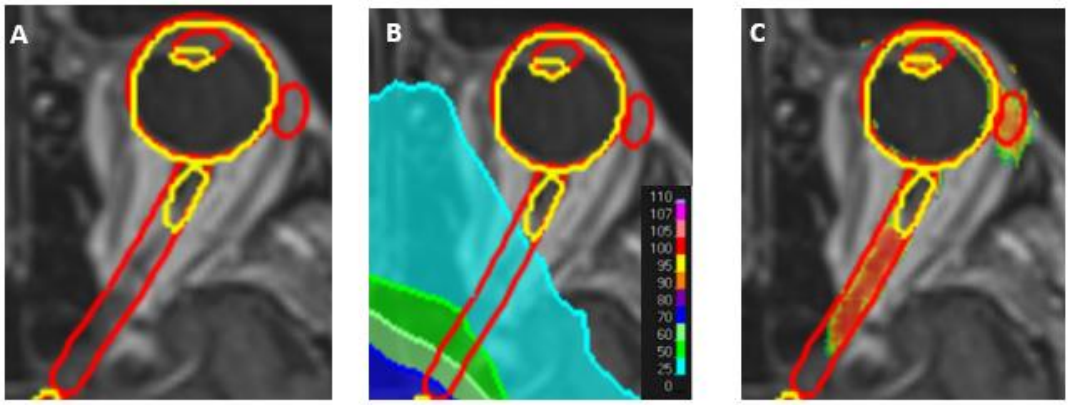


Figure 5.2: a, b, and c axial T1w-Gd MRI, showing examples of scenario 1 for left optic nerve. a) axial scans representing the performance of the MRI DL-AS model in the delineation of the left optic nerve (yellow) relative to the gold standard contours (red). The geometric evaluation showed the geometric difference between the MRI autosegmentations (yellow) and the gold standard (red) for the left optic nerve is high (DSC=0.21, MDA = 0.38 cm, sensitivity= 0.12). b) axial T1w-Gd MRI showing the segmentations with overlying dose distribution. The colourwash represents the percentage dose distribution relative to the prescription dose, according to the inset colorbar. The relative dosimetric impact for this particular geometric error is high-36% (Δ dose= -1820.88 cGy), as a result of the incomplete segmentation. c) Uncertainty prediction of the ACo model. The ACo map successfully indicates high uncertainty for the incorrect segmentation region, indicating that this area needs attention from the user.

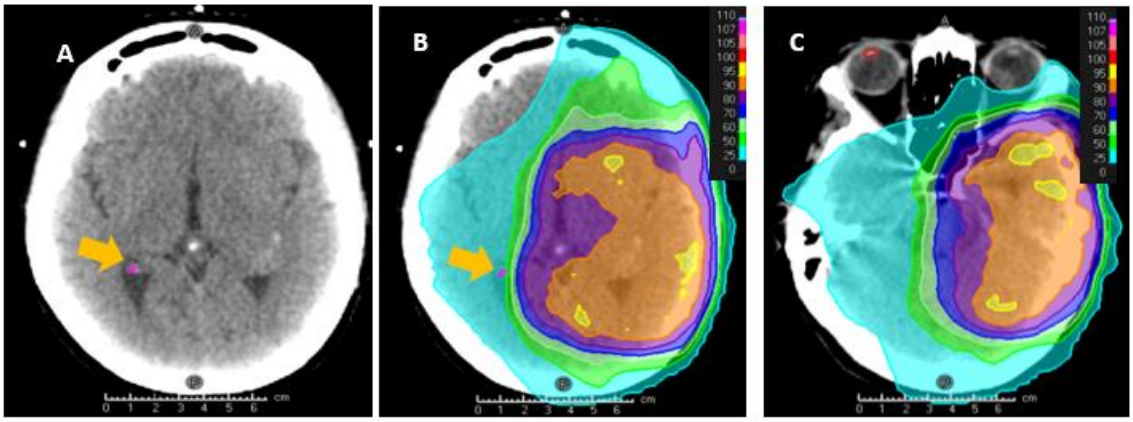


Figure 5.3: Example of complete segmentation failure for the right lens. a) axial CT slice showing the incorrect segmentation by the CTu DL-AS model of part of the ventricle as right lens (pink). b) axial CT slice showing the incorrect autosegmentation of the right lens (pink) with overlying dose distribution. c) axial CT slice showing the gold standard contour of the right lens (red) with overlying dose distribution, demonstrating that the right lens is outside the high dose region. The colourwash represents the percentage dose distribution relative to the prescription dose, according to the inset colorbar. The CTu DL-AS had no geometric overlap with the gold standard contour for the right lens (DSC= 0, sensitivity= 0), resulting in a 446% change in dose (Δ dose= 2536.60 cGy). This significant change in dose was caused by the dose in the brain being approximately 4.5 times greater (given its proximity to the PTV boundary) than in the right lens delineated as the gold standard.

This scenario of a large geometric error being associated with large dosimetric consequences is the most obvious one, which careful commissioning and QA of DL-AS would be expected to identify. However, it is important to note that there are two ways in which this can happen. One involves clinically significant errors occurring on steep dose gradients, and the other involves gross geometric errors occurring in relatively low gradient dose regions. This highlights the importance of understanding whether the dosimetric tests are relative or absolute, and of considering whether the dosimetric changes are likely to be clinically significant (e.g. using the pragmatic approach described in the method section 3.2.4 of Chapter 3). ACo shows its value by highlighting the geometric errors localised

onto the organ (figure 5.2c), which allows easy visual assessment of their likely dosimetric and clinical consequence. This is particularly useful once the PTV delineation is available, giving an indication of the likely location of dose gradients, and enabling a human operator to make good decisions about the need for OAR contour editing, prior to dose planning.

5.4.2.2.2 Scenario 2: Large geometric error with small dosimetric impact

In some cases, even if the autosegmentation of the structure was poor geometrically, it had little effect on the dosimetry. In other words, the dosimetric statistics for the manual and automatic contours were similar, regardless of geometric discrepancies. This was observed when autosegmentation delineated structures were situated in low dose regions with shallow gradients (far from the PTV), or when the (second order) OAR was within the homogeneous high dose region, again with shallow gradients (figure 5.4). Also, this result may be related to the known limitations of geometry metrics for small and complex boundaried structures (e.g. optic chiasm). This can lead to apparently large geometric errors, which do not result in significant dosimetric impact (figure 5.4).

In these scenarios, both geometric and dosimetric evaluations are again necessary to evaluate the clinical applicability of the DL-AS model. The dosimetric evaluation alone may fail to identify the underlying geometric error, which may recur and cause clinically significant errors for another patient with a different dose distribution, leading to a suboptimal treatment plan and potentially affecting patient outcomes. Integrating both evaluations improves overall evaluation of autosegmentation accuracy, ensuring both geometric accuracy and clinical relevance for optimal treatment planning and delivery.

Of course, in clinical use, gold standard contours are not available, and geometric error metrics cannot be computed. However, ACo provides a critical mechanism for predicting such geometric errors and highlighting them to clinicians. In this case, ACo is likely to have prompted editing of the pituitary (figure 5.4). In addition, in this case, the combination of a near-maximal dose statistic and the location of the pituitary within the high dose region prevented the severe geometric error having a large dosimetric impact.

Whilst this editing would not have significantly affected the reported dose statistic, it would have provided an accurate and robust dosimetric assessment for

treatment planning, rather than one based on a fortunate and fragile set of coincidences.

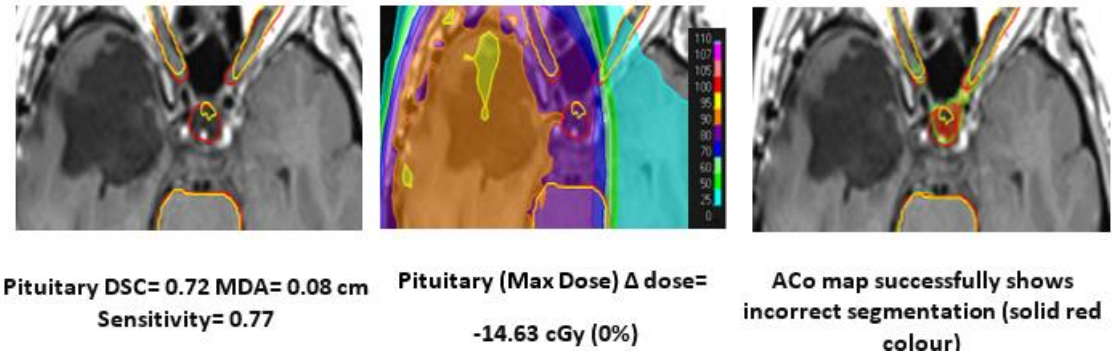


Figure 5.4: Axial T1w-Gd MRI images, providing an example of scenario 2. a) slice showing MRI DL-AS delineation of pituitary (yellow) relative to the gold standard (red). The geometric evaluation showed the geometric difference between the MRI autosegmentations, and the gold standard is high (DSC= 0.72 MDA= 0.08 cm Sensitivity= 0.77). b) as a) with overlying dose distribution. The colourwash represents the percentage dose distribution relative to the prescription dose, according to the inset colorbar. The dosimetric impact for this geometric error was very low relative to the gold standard, which is 0% (Δ max dose = -14.63 cGy); the very low change in dose mainly resulted from incorrect segmentation within the homogeneous high dose region, with shallow gradients. c) the uncertainty estimates of the ACo model. The ACo map successfully indicates incorrect segmentation in red, as high uncertainty, requiring attention from the user.

5.4.2.2.3 Scenario 3: Small geometric error with large dosimetric impact

Some structures, such as the superior part of the brainstem, posterior limit of optic nerves, and superior and inferior limit of cochleae, occasionally exhibited high or acceptable geometrical agreement between the gold standard contour and autosegmentation, but there was significant change in OAR dose. Based on visual inspection, this scenario typically occurred on a steep dose gradient (near the PTV boundary) and led to a substantial dosimetric change for relatively little geometric error on a large organ (figure 5.5 a and b). Geometric analysis with

DSC and MDA is known to be insensitive for small errors on large organs. Using the brainstem as an example, AutoConfidence identified a small region of geometric uncertainty at the superior limit of the organ, close to the high dose gradient, which resulted in a large dosimetric impact. This is probably the most concerning scenario, from a clinical perspective and it is encouraging to see the ability of ACo to identify this critical region of DL-AS error.

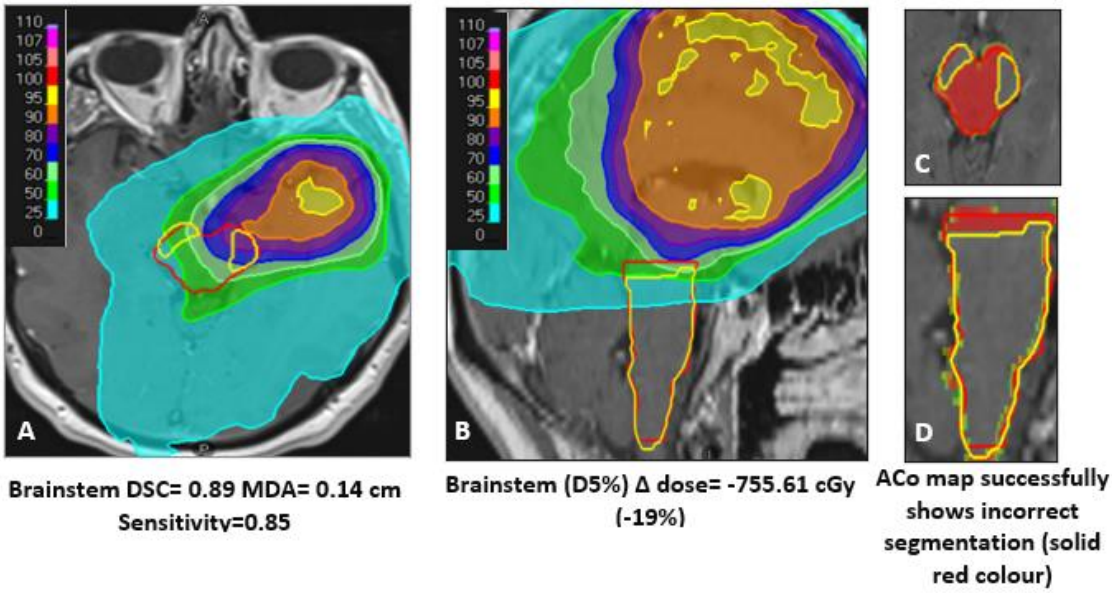


Figure 5.5: a) axial and b) sagittal T1w-Gd MRI with overlying dose distribution, showing examples of scenario 3. The colourwash represents the percentage dose distribution relative to the prescription dose, according to the inset colour bar. The low geometric discrepancy was found between the MRI autosegmentations (yellow) and the gold standard (red) (DSC= 0.89 MDA= 0.14 cm, sensitivity= 0.85), while the dosimetric impact for this geometric error was high relative to the gold standard, which is -19% (ΔD5% dose= -755.61 cGy). c) axial and d) sagittal T1w-Gd MRI with ACo uncertainty map (red = high uncertainty), showing the successful performance from ACo in detecting segmentation error as high uncertainty requiring attention from the user.

This scenario highlights the major reasons for inconsistency between geometric and dosimetric analysis, which is the location of the high dose region and steep dose gradients. Due to the nature of RT dose distributions, the dosimetric impact

of contouring errors is critically dependent on their location relative to dose gradients. This is the primary reason that geometric analysis alone is insufficient for DL-AS. Small geometric errors can be highly significant, but only if they occur on a steep dose gradient. As these gradients are highly patient specific, a very large cohort would be required in order to find cases in which they align with geometric contouring errors. Even if such a cohort were analysed, it would indicate that in the majority of cases small geometric errors were unimportant clinically. However, in RT we are interested in the outlier cases, in which a patient may suffer toxicity and undesirable side effects. Hence it is important to take a conservative approach.

Geometric analysis can be insensitive to this type of error (small region of error on a large organ) as DSC is a volumetric measure and the change in overlap volume would be minimal. Also, MDA, as a mean measure of disagreement is also insensitive to small regions of error which may be dosimetrically and clinically crucial. Whilst max or near-max distance metrics can be sensitive to such errors they are typically very susceptible to noise and hence are rather non-specific, limiting their utility in DL-AS assessment.

ACo can be of vital assistance here, by highlighting the regions of potential geometric error to the operator. The operator can then apply their own judgement as to whether the uncertain region is likely to be near to a steep dose gradient as in the example of the superior aspect of brainstem shown in figure 5.5 c-d.

5.4.2.2.4 Scenario 4: Small geometric error with small dosimetric impact

Many ROIs showed high geometric agreement (low error) accompanied by minimal changes in dosimetric statistics. This was by far the most common scenario, as would be expected for high quality segmentation models. For example, figure 5.6 shows a high agreement geometrically between the gold standard contour and autosegmentation for the left and right orbits with minimal dosimetric impact. AutoConfidence demonstrated a high level of confidence in the autosegmentation with only small regions of uncertainty close to the boundaries. The user could either edit these minor errors if they considered them important, or more likely choose to leave them as they were produced, due to the distance from the high-dose region.

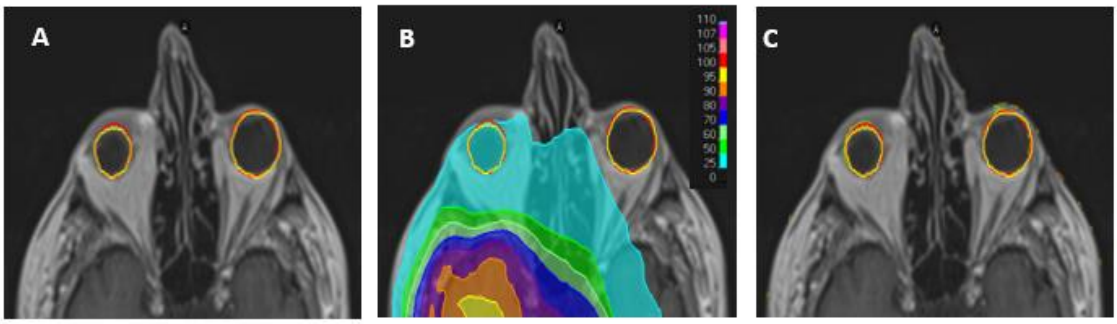


Figure 5.6: a, b, and c axial T1w-Gd MRI, showing examples of scenario 4 for the left and right orbits. a) axial scans representing the performance of the MRI DL-AS model in the delineation of the left and right orbits (yellow) relative to the gold standard contours (red). The geometric evaluation showed the geometric difference between the MRI autosegmentations (yellow) and the gold standard (red) is low (DSC= 0.92,0.91, MDA= 0.05, 0.07 cm, sensitivity= 0.91, 0.85 for the left and right orbits, respectively). b) axial T1w-Gd MRI showing the segmentations with overlying dose distribution. The colourwash represents the percentage dose distribution relative to the prescription dose, according to the inset colorbar. The dosimetric impact for this geometric error is low relative to the gold standard, which is -1% (ΔD1% dose= -16.52 cGy, -19.76 cGy for the left and right orbits, respectively).c) The ACo map showed minimal regions of low confidence, very close to the segmentation boundary, indicating a high quality autosegmentation.

In summary, I demonstrate the complex relationship between geometric and dosimetric impact of autosegmentation errors and highlights the reasons for the weak correlation between the geometric and dosimetric outputs. Through four different scenarios, we discuss the possible outcomes, from large geometric errors with significant dosimetric impact to small geometric change with minimal dosimetric influence. Each scenario showed the importance of combining geometric and dosimetric evaluations to assess the clinical applicability of autosegmentation. Depending only on either metric alone was not sufficient to capture the full picture of autosegmentation reliability.

Furthermore, these scenarios illustrate the importance of the ACo. The ACo uncertainty maps identify the geometric errors, even in the absence of ground truth segmentations, and potential dosimetric and hence clinical implications. Accordingly, ACo is able to significantly add value, for two reasons. First, it can be used in clinical practice for real patients where gold standard contours are not available. Second, it can provide localised estimates of geometric error, which traditional organ-wise geometric analyses cannot do. These geometric uncertainty predictions need to be interpreted in the light of potential dosimetric consequences. In practice this could be done by a human operator with knowledge of the PTV location, as a surrogate for the high dose region.

Overall, these results highlight the importance of a comprehensive analysis to evaluate autosegmentation accuracy, ensuring the safe utilization of this tool in a clinical setting.

5.4.2.3 Comparison to literature

5.4.2.3.1 Comparison to Turcas et.al 2023

Our conclusions align with the latest published work assessing the performance of a commercial MRI DL-AS on brain OARs from both geometry and dosimetry perspectives, using T1-w with gadolinium MRI scan (Turcas et al., 2023).

Their assessment was based on comparing the generated contour to the reference gold standard contour. Our reference gold standard contour was based only on MRI scan (Alzahrani et al., 2023), while their reference gold standard contour was based on the planning CT scan that was rigidly registered with the MRI scan (Turcas et al., 2023). Regarding training and evaluation of the model, our MRI DL-AS model was trained in-house using 32 glioma cases based on the edited and unedited clinical contour as described in the method section 2.2.4 (Alzahrani et al., 2023). In contrast, in the published work, the researchers did not train the model, while they used the MVision GBS™ MRI DL-AS model (Turcas et al., 2023).

Geometric metrics such as DSC and MDA for examination of the accuracy of generated segmentations relative to the reference contours were used by Turcas et al and ourselves. Similarly, dose volume histogram (DVH) comparisons were used to assess the impact of geometric differences on the treatment plan.

Statistical analyses metrics and correlation were applied to identify the significance of findings. However, there are some differences in the evaluation methods based on the particular focus of the investigation. First is the number of test cases. I tested the model from the geometry and dosimetry perspectives on 9 test cases with more emphasis on the impact of pre-training editing of the clinical contour on the quality of the model using different modalities (Alzahrani et al., 2023). Conversely, Turcas et al tested their model on 30 cases for geometric evaluation, while 15 of them were also used for dosimetric evaluation, and here dose plans were reoptimized before assessment (Turcas et al., 2023). The focus of Turcas et al was on comparing the reference contours with DL_AS-generated contours and then with manual editing of DL-AS contours by two radiation oncologists, to assess the impact of different delineation methods (i.e. AI-generated contours and post-editing AI-generated contours) on treatment planning (Turcas et al., 2023). Therefore, their study assessed the performance of an AI-human combined system, rather than the model itself.

Secondly, there were differences in the selection of evaluation metrics. In our work, DSC, sensitivity, and MDA (cm) were used for geometric assessment, and the DVH comparisons for dosimetric evaluation (Alzahrani et al., 2023). In the work by Turcas et al, DSC, median distance (mm), and maximum Hausdorff Distance (mm) were used, along with DVH comparisons and gamma analysis with reoptimization of the plan (Turcas et al., 2023).

Moreover, regarding dosimetric assessment, our study used a novel approach for clinical evaluation based on the average metric threshold and worst-case scenario threshold to identify clinically significant cases (described in Chapter 3, method section 3.2.4). In contrast, Turcas et al used a gamma analysis approach with a pass rate of 3%/3mm and a 10% dose voxel failure cut-off (Turcas et al., 2023).

Regardless of the differences, the findings of Turcas et al broadly align with our findings. Both investigations proved the superior performance of the MRI deep learning models in delineating large structures, such as the brainstem, while both studies confirmed that the MRI DL-AS models perform less well in delineating small structures, such as the lacrimal gland (Turcas et al., 2023, Alzahrani et al., 2023). Both works confirmed that a CT scan was needed for better visualization of the lacrimal glands (Turcas et al., 2023, Alzahrani et al., 2023). Table 5.1

illustrates the median DSC and MDA for the investigated structures in both studies (Turcas et al., 2023, Alzahrani et al., 2023).

Table 5.1: Comparison of the Median DSC and MDA for selected brain OARs investigated by Turcas et al and the current study using MRI DL-AS models. Bold values highlight superior geometric performance in the study comparisons.

Median DSC		OARs	Distance to agreement (cm)	
(Alzahrani et al., 2023)	(Turcas et al., 2023)		(Alzahrani et al., 2023)	(Turcas et al., 2023)
0.89	0.89	Brainstem	**0.08**	0.09
0.69	0.55	Optic Chiasm	0.09	**0.06**
0.04	**0.45**	Lacrimal Gland L	0.18	**0.14**
0.15	**0.55**	Lacrimal Gland R	0.25	**0.13**
0.62	0.49	Optic Nerve L	0.18	**0.15**
0.67	0.56	Optic Nerve R	0.10	**0.09**
0.72	0.61	Pituitary	**0.08**	0.09

Regarding the visual assessment, both works illustrate that the MRI DL-AS model has frequent geometric discrepancies in the superior borders of the brainstem (Chapter 3, discussion section). However, in their work, Turcas et al found geometric discrepancies in the inferior border of the brainstem as well, while I did not experience that as when the model was trained, the inferior border was always at the level of C2 (Alzahrani et al., 2023).

Regarding dosimetric evaluation, both studies show that dosimetry metrics outputs are strongly impacted by tumour location in relation to OARs, and that variations in dose between autosegmentation and gold standard contours do not exactly reflect discrepancies in geometry (Turcas et al., 2023) (Chapter 3, discussion section). Both studies show a negative correlation between the DSC and dosimetry outputs (Turcas et al., 2023) (Chapter 3, discussion section). Moreover, both studies confirm that the correlation is not strong and indicate that both geometric and dosimetric analyses are needed. Both studies confirmed that using the commercial MRI DL-AS tool to delineate brain OARs is a promising tool

to improve clinical workflow, quality, and consistency, but some editing for small structures is still required (Turcas et al., 2023, Alzahrani et al., 2023).

5.4.2.3.1.1 OARs that clinicians tend to delineate them on CT

Based on the international consensus guidelines, clinical oncologists indicated that they tend to contour the following structures based on CT imaging: orbits, lenses, lacrimal glands, cochlea, and optic nerve until the bony optic canal (Eekers et al., 2018). For orbits, lenses, cochleae and the extra-cranial portions of the optics nerves, this is largely because, at present, treatments are planned based on CT, and so the position of these organs on CT is considered most meaningful from a planning perspective. For lacrimal glands, this partly relates to ease of identifying these structures on CT compared to MRI. It is, however, essential to investigate the performance of these structure using the MRI DL-AS model to ensure the model can produce acceptable segmentation without a CT scan, especially in the era of the MRI-only pathway.

Orbits, cochlea, and lenses were not investigated in the published work of Turcas et al., 2023. The authors stated that they did not include them as manual contouring is needed based on the CT scan (Turcas et al., 2023).

Moreover, the authors also suggested that delineating the lacrimal glands and optic nerves using CT scans was preferable, based on comparison of their results (based on MRI DL-AS) with previous literature investigating CT-based segmentation (Turcas et al., 2023).

However, in our work, I investigated both modalities on the delineation of these structures to identify if the MRI DL-AS model alone is enough to delineate these structures or if a CT scan is needed (Alzahrani et al., 2023) . More information can be found in the following section.

5.4.2.3.1.1.1 Comparing CT DL-AS to MRI DL-AS

Based on our findings, except for the lacrimal glands and optic nerves, the general performance of the CTeCT DL-AS model vs MRIeMRI DL-AS model on the delineation of these structures was similar (table 5.2) (Alzahrani et al., 2023).

It is important to note that the performance of the CTu model trained with the original contours was lower than that of the CTeCT and MRIeMRI DL-AS models and produced more failed segmentations (Table S1-S7) (Alzahrani et al., 2023).

This suggests that similar performance could result from consistent clinical contours, which the CTeCT and MRIeMRI models were trained with.

However, the lacrimal gland is better visualized on CT vs T1w-Gd MRI scan (table 6.2) (Alzahrani et al., 2023). Fat saturation is needed to visualize the lacrimal gland on an MRI scan based on the international consensus guidelines (Scoccianti et al., 2015).

Regarding the delineation of the optic nerve, based on our findings, the MRIeMRI model performed better than the CTeCT model (table 5.2) (Alzahrani et al., 2023). This could relate to the fact that MRI is essential for contouring the intra-cranial portions of the optic nerves, and these cannot be seen clearly based on CT. The extra-cranial portion of the optic nerves can be well visualised on both CT and MRI and the clinician tendency to use CT is based on the fact that CT is routinely used for planning, and not increased ease of identification of the structure on CT.

Accordingly, the MRIeMRI DL-AS is sufficient to produce segmentations for the structures stated in Table 5.2, with the exception of the lacrimal gland, and so could be used in an MRI-only pathway for brain OARs.

Table 5.2: Comparison of CTeCT and MRIeMRI DL-AS models for structures that clinicians tend to contour using CT.

OARs	CTeCT (n= 10 cases)			MRIeMRI (n= 9 test cases)		
	Average DSC	Average MDA (cm)	Missing segmentation	Average DSC	Average MDA (cm)	Missing segmentation
Lacrimal Gland L	0.40	0.18	0	0.10	0.23	7
Lacrimal Gland R	0.44	0.29	0	0.15	0.18	3
Cochlea L	0.51	0.08	0	0.57	0.07	0
Cochlea R	0.63	0.06	1	0.49	0.10	2
Lens L	0.71	0.05	2	0.68	0.07	0
Lens R	0.72	0.05	1	0.67	0.09	1
Optic Nerve L	0.50	0.15	0	0.65	0.09	0
Optic Nerve R	0.49	0.15	0	0.68	0.09	0
Orbit L	0.90	0.07	0	0.90	0.06	0
Orbit R	0.90	0.08	0	0.91	0.06	0

5.4.2.3.2 Comparison to Wong et.al 2020

Here I compare the results of our own MRI DL-AS work with that of Wong et al.,2020 to further evaluate the performance of our commercial MRI DL-AS compared to a commercial CT DL-AS for orbits and optic nerves.

Wong et al., 2020 investigated the performance of a commercial CT DL-AS model (Limbus AI, version 1.0.22) based on U-net architecture and assessed model delineation on orbits, and optic nerves (Wong et al., 2020). Twenty test cases were used to compare the autosegmentation with manual reference contours, using DSC and 95% Hausdorff distance (HD) (Wong et al., 2020).

For the optic nerve, the output of our MRIeMRI model (table 6.2, average DSC =0.65, 0.68 for optic nerve L and R, respectively) was similar to their finding of the CT DL-AS model average DSC=0.6 (Wong et al., 2020). However, this result was based on one side of the optic nerve. As they stated in their research method, the bilateral structures were delineated only on one side (Wong et al., 2020). However, they reported variability in the optic nerve junction with optic chiasm, which will impact individual contouring assessments (Wong et al., 2020). However, based on our brain OARs atlas, which is based on international consensus guidelines, MRI is recommended for the delineation of the optic nerve after passing beyond the bony optic canal, and the delineation of the optic chiasm should be contiguous with the optic nerve, using MRI scan to reduce variability.

Lastly, our DSC for orbits is improved compared to that of Wong et al (Average DSC=0.91 vs Average DSC=0.85) (Wong et al., 2020, Alzahrani et al., 2023).

In summary, except for the lacrimal glands, the MRI DL-AS can produce acceptable contouring for all brain OARs, including for the structures for which the clinician tends to use the CT scan for delineation. However, the MRI model needs to be trained with consistent contouring (Alzahrani et al., 2023) or with a large number of cases.

There is limited literature on MRI brain OARs DL-AS delineation to compare our work with, and of the literature that does exist, this mainly concerns the use of non-clinical algorithms and relies only on geometric assessment (Chen et al., 2019, Mlynarski et al., 2020, Wiesinger et al., 2021). Comparisons with this literature can be found in Chapter 2 (Alzahrani et al., 2023).

5.4.2.4 Factors affecting the quality of DL-AS models, measured by geometric and dosimetric tests

The output of both assessments could be affected by several factors. First is the quality of the DL-AS model. The MRIeMRI DL-AS model was our preferred model. This model was trained with high-quality labelled data as all the clinical contours were edited to be consistent based on specific guidelines in the institution brain OAR atlas (Chapter 2, method section 2.2.2). Moreover, this model was trained using MRI scans, which have better contrast for the soft tissue delineation than CT scan. Using MRI scans helped the model to learn the location and features of the structure as it was easy to detect the boundaries of the structures except for the lacrimal glands (reasons for failing to delineate the lacrimal glands were highlighted in the discussion section, Chapter 2). Therefore, generally, less clinically significant cases, less dosimetric change, low failed segmentations, and more geometry agreement were achieved with the MRIeMRI DL-AS model compared to other models (result section 2.3.2, table 2.1, Chapter 2, and result section 3.3.2, table 3.5 and 3.6, Chapter 3).

Other factors that could affect the evaluations are the images' quality, tumour location and its deformation for the structures and the gold standard contour delineation (Claessens et al., 2022). In our work, I did not exclude any cases based on any of these factors, in order to reflect the real-world clinical setting and so the model can be used in high- or low-quality image and regardless of tumour location and deformation.

The gold standard contour used in our evaluation was consistent as it was done by a single person after being trained by an expert radiation oncologist who then reviewed the delineations. Manual editing was based on the guidance contained in the institution brain OARs atlas (More information can be found in Chapter 2, method section 2.2.2.

It is also essential to consider potential confounding factors, such as the image quality of testing data, tumour location and gold standard contour quality since these factors will also impact the outputs of the evaluations. Importantly, to track the performance of the models in the clinical setting for such cases, an independent QA model is needed to improve the safety of using autosegmentation tools.

5.4.3 QA-AI model (ACo)

5.4.3.1 Importance of the ACo model: standardized and expediate the process of autosegmentation verification

Despite all the efforts to evaluate DL-AS models in the pre-clinical implementation setting, it is still challenging to use these models clinically (Mackay et al., 2023, Claessens et al., 2022), due to the possibility of errors occurring which are unexpected. Several groups have developed successful autosegmentations models, but they are not usually used clinically, as the performance can be variable with individual patients (Mackay et al., 2023). Moreover, inadequate guidance is available on how to edit DL-AS contours and be more confident in their clinical use (Mackay et al., 2023).

Concerns have been raised regarding how staff will deal with autosegmentation outputs (Claessens et al., 2022). It is expected that some clinicians will rely on them and perform insufficient editing (suffering automation bias), while others might do too much editing of clinically unimportant regions of the segmentation. In either case, this will introduce variability and inconsistency in the segmentation and diminish the benefits of automation.

Quality assurance (QA) for DL-AS is still developing (Claessens et al., 2022); however, recommendations have been made to use an independent QA tool to estimate uncertainty and/or potentially incorrect segmentations. The idea is not to directly improve contours, but rather to indicate to the operator that a specific contour needs review and possible adjustment.

Thus, in this project, an independent QA-AI model was developed to detect uncertainty in the segmentation and serve as an assistance tool to highlight areas to operators that need potential editing to increase the safety of using the DL-AS tool. ACo was based on adversarial learning of a discriminator, trained alongside an internal segmentation model. Importantly, following training, it requires only the scan and proposed segmentation to provide a localised map of uncertainty, on a patient specific, reference free basis.

Using this approach should also help to reduce bias from the operator while performing their manual verification. Moreover, it will potentially speed up the segmentation evaluation, which is needed with the increase of cancer patients to avoid any delay in treatment (Soomro et al., 2023). It should also increase confidence in utilizing the autosegmentation DL-AS model.

The ACo model generated a ‘confidence map’ with a range of colours from red (to highlight low confidence in the segmentation) to green (to highlight moderate confidence in the segmentation), with high confidence regions transparent, allowing overlay on the segmentation itself. It is possible to assess any segmentation (e.g., manual, or autosegmentation) given the underlying scan.

5.4.3.2 Comparison to previous literature

Two previous studies were conducted investigating deep learning methods to evaluate autosegmentation quality on breast and salivary glands using CT scans (van Rooij et al., 2021, Chen et al., 2020). More information about both studies can be found in the discussion section of Chapter 4. The recent study investigating the QA of DL-AS was conducted on head and neck CT (Luan et al., 2023). The researchers predicted DSC scores based on radiomics features near contour boundaries. However, predicting DSC is not particularly valuable as it will not provide information about the location of the error in contour boundaries, and DSC also depends on the structure size. Also, setting a threshold for clinical suitability is very difficult based on the DSC value alone. Based on our finding in Chapter 3 (table 3.4), a weak correlation was found between the DSC and dosimetry outputs. So, DSC alone is not a useful metric, even if the DSC score is perfect. This is because the DSC score did not give information about the clinical perspective.

5.4.3.3 Integration of the QA-AI model (Confidence estimation) with geometric and dosimetric evaluations

Based on the evaluation of the performance of the ACo model previously mentioned in Chapter 4, figure 4.3a, the ACo model was able to detect uncertainty in areas with geometric discrepancies between the gold standard contour and autosegmentation for all brain OARs except lenses.

This finding was based on model performance on different sources of autosegmentation (as described in the method section 4.2.3 and 4.2.4, Chapter 4). One was the internal segmentations generated from the ACo generator. The second was the external segmentations generated from our previously developed and tested MRI DL-AS models. Different segmentation qualities were used to establish that the ACo model could highlight the uncertainty in high or low-quality autosegmentations.

ACo performance on MRI DL-AS models was examined in all the scenarios previously described in Chapter 5. The confidence maps successfully identified segmentation errors and missing structures, which generally correlated well with errors when comping ACo output to the gold standard, highlighting these appropriately for clinical operator review.

However, the ACo model was not trained with the PTV or dose information, so the model highlighted errors only based on the geometric accuracy of the segmentation, regardless of the dosimetry or location of the PTV. In general, this behaviour is desirable, as it provides an unbiased assessment of the autosegmentation quality. However, for clinical use, it may be helpful to overlay the PTV, once it has been defined, to estimate the dosimetric and clinical importance of highlighted geometric uncertainty and errors.

Regardless of the location of the PTV, the investigated ACo model was shown to be safe and can be used in the clinical setting to estimate the accuracy of segmentation for brain OARs other than lenses. It is then the responsibility of the human operator to consider the likelihood of a geometric error correlating with a dosimetric error, based on their knowledge of the target location. This is not particularly challenging for a human operator but could be investigated for automation in future. However, this would likely require prediction of dose distributions, which is a separate challenge.

ACo has the potential to enable clinics to finally realise the promised benefits of autosegmentation technologies, saving clinician time and improving consistency and quality of treatment for patients.

5.5 Limitations

5.5.1 Geometry and dosimetry assessment (Chapter 2 and 3)

5.5.1.1 Limited number of training and test cases

The investigation of the accuracy of DL-AS was trained and tested on a limited number of cases for two reasons: the clinician time involved in generating gold standard contours and the limited availability of MRI and paired CT scans. However, even with a small cohort, pre-training curation of the clinical contours helped to produce acceptable segmentation from MRI DL-AS, demonstrating this is highly desirable (Alzahrani et al., 2023).

5.5.1.2 One commercial DL-AS model (RayStation) and a single MRI sequence (T1-w spin echo (SE) with gadolinium)

The investigated models were based on only one commercial algorithm, for one tumour site. Whilst the findings that editing, and a combination of geometric and dosimetric analyses are required are likely generalisable, the numerical results and clinical utility of the specific model are not. This implies that a similar analysis should be performed for any model prior to clinical implementation. Our work therefore forms a basis for clinical commissioning design.

Moreover, our investigation depends on using one MR sequence, which is T1-w spin echo (SE) with gadolinium, as used locally in routine care. It is worth replicating this investigation with other sequences as some structures, such as the lacrimal glands, can be better visualized using T2-w MRI or fat-saturated imaging. This finding will help build a multi-modality MRI model to delineate all brain OARs using one model (Alzahrani et al., 2023). However, it was challenging in this study to use both T1-w and T2-w MRI, as several cases had T1-w scans only. Even when both modalities were available, the lack of inherent registration between T1-w and T2-w images made it difficult to use both data types in a single model due to potential misregistration.

5.5.1.3 Generalisability across centres

Our preferred MRI DL-AS model, which was trained using data from several scanners with matched protocols, may operate poorly with similar data from other institutions due to the scanner harmonization issues. However, in this project, the main aim is to evaluate the accuracy of the MRI DL-AS model instead of building

a model that can operate with any data from other organizations (Alzahrani et al., 2023). Future technological developments, including harmonisation of MR images (Fatania et al., 2022), may enable multi-scanner models to be built with sufficient performance, but until that is achieved, clinical implementation of MRI based DL-AS may be limited to centres which can train and validate their own models.

5.5.1.4 Time-saving evaluation and clinician feedback

The clinician's time editing contours and their feedback on the performance of autosegmentation were beyond the scope of this thesis and will be investigated in future work. Of particular interest is the potential benefit of incorporating uncertainty estimates into the clinical pathway, and the impact on consistency and efficiency this may yield. However, it is essential to note that the atlas created for curating the data before training was based on international consensus guidelines and was reviewed and approved by the CNS oncology team.

5.5.1.5 Using the original dose plan for DL-AS dosimetric evaluation

For dosimetric assessment, the original dose plan was used to compare the dosimetric changes between the gold standard contour and the DL-AS. I did not intend to re-optimise plans based on DL-AS segmentations and then assess dosimetry, as our main interest was to explore the changes in OAR geometry and dose.

Re-optimisation of plans will provide different results and may be of interest for future work. If the DL-AS contours were to be used clinically, the plans would be optimised based on these, potentially leading to different dose distributions. An ‘inverse’ analysis of the DL-AS optimised plan, based on the gold standard contours could give an indication of how the DL-AS contours would affect OAR doses in clinical practice. The effects should be similar to the forward analysis performed herein but could be a valuable extension of this work. However, significant resource would be required for re-planning.

5.5.2 The ACo model assessment (Chapter4)

5.5.2.1 Using a 2D recurrent U-net architecture

A 2D recurrent U-net architecture was used for the ACo model instead of a 3D model to address the limitation of available GPU memory. As a result, over-

and/or under-prediction of errors at superior and inferior limits of OARs were produced from using a 2D approach to 3D contouring. Moreover, this issue could not be resolved even when using IER for the post processing, as IER also employed a 2D algorithm. Using a 3D architecture has the potential to enhance the performance of the ACo model, particularly for defining the superior and inferior limits of the structures, although significant compromises on resolution would be required to make this practical when using current hardware.

5.5.2.2 Limited number of cases

The ACo model was trained and tested on limited data again due to the limited availability of MR imaging. However, the consistent anatomy of brain OARs supported the output of the ACo model, even with the small dataset. More data may improve performance for the more mobile structures, such as lenses, and for other body sites, where more anatomical variation and motion are expected.

5.5.2.3 Time-saving and ACo clinical utility evaluation

The human interaction with the outputs of the ACo model was unable to be investigated in this project due to time constraints. However, there is an ongoing project built on this work, investigating different ways to present the outputs of the ACo to the clinician. More information about future work can be found in the next section.

All these limitations provide interesting opportunities to extend the scope of the evaluation of DL-AS and optimization of the performance of the ACo model.

Stating these limitations could motivate other researchers to address these constraints to enhance the robustness and reliability of the created models. However, the output of this project is a significant contribution in the field and can be further optimized in the future.

5.6 Clinical implementation for MRI DL-AS brain OARs and ACo model and future work

The work involved in this thesis extends beyond research; it is part of service development and the clinical implementation programme within the radiotherapy department at Leeds Cancer Centre. The department aims to utilize an MRI simulator for brain MRI-only radiotherapy. Accordingly, there is an intention to implement MRI DL-AS for brain OARs and AutoConfidence models clinically.

5.6.1 Pre-clinical implementation assessment – ongoing work

Some future work needs to be considered to determine the potential of implementing the MRI DL-AS model and ACo model within the department. This work is currently ongoing, based on the finding of this thesis.

5.6.1.1 Human interactions with the outputs of DL-AS model

Studies are necessary to determine how the clinician will interact with the output generated by both models. Moreover, clinician editing time for DL-AS model outputs should be considered.

Several neuro-oncology clinicians with different levels of expertise in the department will review and edit the generated contours. Geometric and dosimetric evaluations based on post-editing of DL-AS contours is required, including in comparison to both gold standard contours and the unedited contours generated from autosegmentation models.

The findings will enhance understanding of the practical process and potential benefits of utilizing these models in the context of clinical workflow. Moreover, our findings will help the department to provide the clinicians with essential guidance on how to deal with generated segmentations to reach optimal segmentation efficiently.

5.6.1.2 Human interactions with the outputs of ACo model

The department is conducting a parallel study based on this work. It aims to investigate how the ACo output should be presented to the clinician. This study will involve four clinicians inspecting and editing the 9 test cases with and without the benefit of ACo confidence map data, to establish the most efficient and robust clinical implementation approach. ACo data will be presented in several different ways, to establish the optimal method. The impact of ACo on editing quality will be assessed using a combination of geometric and dosimetric test as above. The efficiency of editing will also be investigated to determine the impact of ACo on time-saving.

The output of this study will provide us with a better understanding of which method is ideal for clinicians in terms of editing accuracy, speed, clinician perception, and confidence.

In summary, the use of DL-AS models in the radiotherapy department needs to be comprehensively examined. Whilst clinical safety is relatively easy to

determine and ensure, given good training and robust processes, benefit is harder to quantify, particularly where humans and algorithms interact. The human operator is required to deal with the DL-AS and ACo outputs appropriately. If this process is performed well and efficiently, contour consistency and productivity in the department could be significantly improved. Moreover, consistent contouring will aid future research if correlations between dose received, and side effects are to be investigated.

5.7 Conclusions

Patient safety in radiotherapy is a priority, and requires effective treatment planning based on accurate contouring, to concentrate the dose on the tumour while reducing radiation exposure to OARs. This is essential to minimize treatment-related toxicity and try to maintain patient quality of life, whilst maximising clinical benefit.

Traditionally, delineation of the brain OARs in radiotherapy is a manual and time-consuming process and exhibits high variability between operators. Accordingly, the goal of this thesis was to take advantage of the recent technology of deep learning autosegmentation tools for brain OARs, to alleviate these challenges. I aimed to assess the performance of three deep learning models (CT DL-AS, MRI DL-AS models, and AutoConfidence QA-AI model).

From quantitative and qualitive assessments, it can be concluded that more than one evaluation method is necessary to verify the clinical usability of the DL-AS models. Based on the weak correlations observed between the geometry and dosimetry outputs, it can be stated that geometric assessment alone is insufficient to fully capture the limitations and strengths of the model from different perspectives, including the accuracy of the segmentation and relevance to clinical goals. Thus, dosimetric evaluation is also needed to assess model performance, which is mainly due to the variable location of dose gradients relative to OARs on a per-patient basis.

Moreover, editing to improve the consistency of the clinical contours in the training data has a positive impact on the performance of the model. Except for lacrimal gland in MRI, editing the clinical contours pre-training aids the performance of the model, producing more successful segmentations, and less geometric and dosimetric error in relation to the gold standard contour.

I conclude that of all the investigated DL-AS models, the MRIeMRI DL-AS model could best be used clinically for treatment planning. The model could segment all brain OARs, despite some structures requiring manual contour editing, producing the fewest clinically significant dosimetric errors and high geometric agreement relative to the gold standard (except for the lacrimal glands). A CT scan is currently still needed for the lacrimal gland (or potentially a different MRI sequence). MRI DL-AS in radiotherapy for brain OARs can potentially improve efficiency, productivity, and quality for radiotherapy planning and delivery.

However, a comprehensive pre-clinical implementation evaluation is needed to investigate its clinical applicability, and it needs to be integrated with a QA-AI tool, for in use routine monitoring and QA.

The outputs of the ACo model demonstrated that ACo was able to successfully predict regions of low confidence, including errors relative to the gold standard or missing segmentations, without relying on the gold standard segmentations as a reference. ACo confidence maps can be used as a per-patient, reference-free segmentation QA tool, aiding human operators in editing and validating autosegmentations.

Integrating an independent QA model with DL-AS tools will help to improve its safety, reliability, accuracy, and clinical usefulness, as the performance of the DL-AS model can change with the individual patients, especially if they are significantly different from training data. The ACo model has the potential to make the manual verification process for autosegmentation more robust, without significantly reducing the efficiency benefits.

Future studies are needed to investigate human interactions with the outputs of MRI DL-AS and ACo models to estimate the value of utilizing the MRI DL-AS models and ACo model in routine clinical practice.

In summary, I have determined that DL-AS segmentation of OARs in the brain can be validated using a combination of geometric evaluation, dosimetric evaluation, clinical significance analysis and visual assessment. In our experience, MRI models particularly benefit from editing of contours prior to model training, due to the limited data available and high detail within the images. The question of human interaction with automated models is a challenging one. However, use of novel tools like AutoConfidence can improve efficiency, accuracy, and confidence, enabling robust human-in-the-loop use of AI in healthcare.

With these foundations in place, the routine clinical use of DL-AS in radiotherapy is now possible, potentially yielding significant benefits for hospitals and patients at a critical time for healthcare.

5.8 References

ABDAR, M., POURPANAH, F., HUSSAIN, S., REZAZADEGAN, D., LIU, L., GHAVAMZADEH, M., FIEGUTH, P., CAO, X., KHOSRAVI, A., ACHARYA, U. R., MAKARENKOV, V. & NAHAVANDI, S. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion,* 76**,** 243-297.

AINSLIE, A. P., KLAVER, M., VOSHART, D. C., GERRITS, E., DEN DUNNEN, W. F. A., EGGEN, B. J. L., BERGINK, S. & BARAZZUOL, L. 2024. Glioblastoma and its treatment are associated with extensive accelerated brain aging. *Aging Cell,* 23**,** e14066.

ALDAPE, K., BRINDLE, K. M., CHESLER, L., CHOPRA, R., GAJJAR, A., GILBERT, M. R., GOTTARDO, N., GUTMANN, D. H., HARGRAVE, D., HOLLAND, E. C., JONES, D. T. W., JOYCE, J. A., KEARNS, P., KIERAN, M. W., MELLINGHOFF, I. K., MERCHANT, M., PFISTER, S. M., POLLARD, S. M., RAMASWAMY, V., RICH, J. N., ROBINSON, G. W., ROWITCH, D. H., SAMPSON, J. H., TAYLOR, M. D., WORKMAN, P. & GILBERTSON, R. J. 2019. Challenges to curing primary brain tumours. *Nat Rev Clin Oncol,* 16**,** 509-520.

ALZAHRANI, N., HENRY, A., CLARK, A., MURRAY, L., NIX, M. & AL-QAISIEH, B. 2023. Geometric evaluations of CT and MRI based deep learning segmentation for brain OARs in radiotherapy. *Phys Med Biol,* 68.

AN, L., CHEN, J., CHEN, P., ZHANG, C., HE, T., CHEN, C., ZHOU, J. H., YEO, B. T. T., ALZHEIMER'S DISEASE NEUROIMAGING, I., AUSTRALIAN IMAGING, B. & LIFESTYLE STUDY OF, A. 2022. Goal-specific brain MRI harmonization. *Neuroimage,* 263**,** 119570.

ANGOM, R. S., NAKKA, N. M. R. & BHATTACHARYA, S. 2023. Advances in Glioblastoma Therapy: An Update on Current Approaches. *Brain Sci,* 13.

APOLLE, R., APPOLD, S., BIJL, H. P., BLANCHARD, P., BUSSINK, J., FAIVRE-FINN, C., KHALIFA, J., LAPRIE, A., LIEVENS, Y., MADANI, I., RUFFIER, A., DE RUYSSCHER, D., VAN ELMPT, W. & TROOST, E. G. C. 2019. Inter-observer variability in target delineation increases during adaptive treatment of head-and-neck and lung cancer. *Acta Oncol,* 58**,** 1378-1385.

ASGHARNEZHAD, H., SHAMSI, A., ALIZADEHSANI, R., KHOSRAVI, A., NAHAVANDI, S., SANI, Z. A., SRINIVASAN, D. & ISLAM, S. M. S. 2022. Objective evaluation of deep uncertainty predictions for COVID-19 detection. *Sci Rep,* 12**,** 815.

BAROUDI, H., BROCK, K. K., CAO, W., CHEN, X., CHUNG, C., COURT, L. E., EL BASHA, M. D., FARHAT, M., GAY, S., GRONBERG, M. P., GUPTA, A. C., HERNANDEZ, S., HUANG, K., JAFFRAY, D. A., LIM, R., MARQUEZ, B., NEALON, K., NETHERTON, T. J., NGUYEN, C. M., REBER, B., RHEE, D. J., SALAZAR, R. M., SHANKER, M. D., SJOGREEN, C., WOODLAND, M., YANG, J., YU, C. & ZHAO, Y. 2023. Automated Contouring and Planning in Radiation Therapy: What Is 'Clinically Acceptable'? *Diagnostics (Basel),* 13.

BIBAULT, J. E. & GIRAUD, P. 2024. Deep learning for automated segmentation in radiotherapy: a narrative review. *Br J Radiol,* 97**,** 13-20.

BISHOP, C. M. Neural networks for pattern recognition. 1995.

BROUWER, C. L., BOUKERROUI, D., OLIVEIRA, J., LOONEY, P., STEENBAKKERS, R., LANGENDIJK, J. A., BOTH, S. & GOODING, M. J. 2020a. 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.

BROUWER, C. L., DINKLA, A. M., VANDEWINCKELE, L., CRIJNS, W., CLAESSENS, M., VERELLEN, D. & VAN ELMPT, W. 2020b. Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. *Physics and Imaging in Radiation Oncology,* 16**,** 144-148.

BRUNESE, L., MERCALDO, F., REGINELLI, A. & 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. & BROCK, K. B. 2019. Advances in Auto-Segmentation. *Semin Radiat Oncol,* 29**,** 185-197.

CHEN, H., LU, W., CHEN, M., ZHOU, L., TIMMERMAN, R., TU, D., NEDZI, L., WARDAK, Z., JIANG, S., ZHEN, X. & GU, X. 2019. A recursive ensemble organ segmentation (REOS) framework: application in brain radiotherapy. *Phys Med Biol,* 64**,** 025015.

CHEN, X., MEN, K., CHEN, B., TANG, Y., ZHANG, T., WANG, S., LI, Y. & DAI, J. 2020. CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy. *Front Oncol,* 10**,** 524.

ÇIÇEK, Ö., ABDULKADIR, A., LIENKAMP, S. S., BROX, T. & RONNEBERGER, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International conference on medical image computing and computer-assisted intervention, 2016. Springer, 424-432.

CLAESSENS, M., ORIA, C. S., BROUWER, C. L., ZIEMER, B. P., SCHOLEY, J. E., LIN, H., WITZTUM, A., MORIN, O., NAQA, I. E., VAN ELMPT, W. & VERELLEN, D. 2022. Quality Assurance for AI-Based Applications in Radiation Therapy. *Semin Radiat Oncol,* 32**,** 421-431.

DEMIRCIOGLU, A. 2024. The effect of feature normalization methods in radiomics. *Insights Imaging,* 15**,** 2.

DU, G., CAO, X., LIANG, J., CHEN, X. & ZHAN, Y. 2020. Medical Image Segmentation based on U-Net: A Review. *Journal of Imaging Science and Technology,* 64**,** 20508-1-20508-12.

EDMUND, J. M. & NYHOLM, T. 2017. A review of substitute CT generation for MRI-only radiation therapy. *Radiation Oncology,* 12**,** 1-15.

EEKERS, D. B., IN 'T VEN, L., ROELOFS, E., POSTMA, A., ALAPETITE, C., BURNET, N. G., CALUGARU, V., COMPTER, I., COREMANS, I. E. M., HOYER, M., LAMBRECHT, M., NYSTROM, P. W., MENDEZ ROMERO, A., PAULSEN, F., PERPAR, A., DE RUYSSCHER, D., RENARD, L., TIMMERMANN, B., VITEK, P., WEBER, D. C., VAN DER WEIDE, H. L., WHITFIELD, G. A., WIGGENRAAD, R., TROOST, E. G. C. & EUROPEAN PARTICLE THERAPY NETWORK" OF, E. 2018. The EPTN consensus-based atlas for CT- and MR-based contouring in neuro-oncology. *Radiother Oncol,* 128**,** 37-43.

EL NAQA, I., RUAN, D., VALDES, G., DEKKER, A., MCNUTT, T., GE, Y., WU, Q. J., OH, J. H., THOR, M. & SMITH, W. 2018. Machine learning and modeling: Data, validation, communication challenges. *Medical physics,* 45**,** e834-e840.

FAGHANI, S., MOASSEFI, M., ROUZROKH, P., KHOSRAVI, B., BAFFOUR, F. I., RINGLER, M. D. & ERICKSON, B. J. 2023. Quantifying Uncertainty in Deep Learning of Radiologic Images. *Radiology,* 308**,** e222217.

FATANIA, K., CLARK, A., FROOD, R., SCARSBROOK, A., AL-QAISIEH, B., CURRIE, S. & NIX, M. 2022. Harmonisation of scanner-dependent contrast variations in magnetic resonance imaging for radiation oncology, using style-blind auto-encoders. *Phys Imaging Radiat Oncol,* 22**,** 115-122.

GIBBONS, E., HOFFMANN, M., WESTHUYZEN, J., HODGSON, A., CHICK, B. & LAST, A. 2023. Clinical evaluation of deep learning and atlas-based auto-segmentation for critical organs at risk in radiation therapy. *J Med Radiat Sci,* 70 Suppl 2**,** 15-25.

HANSEN, C. R., HUSSEIN, M., BERNCHOU, U., ZUKAUSKAITE, R. & THWAITES, D. 2022. Plan quality in radiotherapy treatment planning - Review of the factors and challenges. *J Med Imaging Radiat Oncol,* 66**,** 267-278.

HARRISON, K., PULLEN, H., WELSH, C., OKTAY, O., ALVAREZ-VALLE, J. & JENA, R. 2022. Machine Learning for Auto-Segmentation in Radiotherapy Planning. *Clin Oncol (R Coll Radiol),* 34**,** 74-88.

HEILEMANN, G., BUSCHMANN, M., LECHNER, W., DICK, V., ECKERT, F., HEILMANN, M., HERRMANN, H., MOLL, M., KNOTH, J., KONRAD, S., SIMEK, I. M., THIELE, C., ZAHARIE, A., GEORG, D., WIDDER, J. & TRNKOVA, P. 2023. Clinical Implementation and Evaluation of Auto-Segmentation Tools for Multi-Site Contouring in Radiotherapy. *Phys Imaging Radiat Oncol,* 28**,** 100515.

HESAMIAN, M. H., JIA, W., HE, X. & KENNEDY, P. 2019. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *J Digit Imaging,* 32**,** 582-596.

HINDOCHA, S., ZUCKER, K., JENA, R., BANFILL, K., MACKAY, K., PRICE, G., PUDNEY, D., WANG, J. & TAYLOR, A. 2023. Artificial Intelligence for Radiotherapy Auto-Contouring: Current Use, Perceptions of and Barriers to Implementation. *Clin Oncol (R Coll Radiol),* 35**,** 219-226.

ISOLA, P., ZHU, J.-Y., ZHOU, T. & EFROS, A. A. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 1125-1134.

JENA, R., KIRKBY, N. F., BURTON, K. E., HOOLE, A. C., TAN, L. T. & BURNET, N. G. 2010. A novel algorithm for the morphometric assessment of radiotherapy treatment planning volumes. *Br J Radiol,* 83**,** 44-51.

JOHNSTON, N., DE RYCKE, J., LIEVENS, Y., VAN EIJKEREN, M., AELTERMAN, J., VANDERSMISSEN, E., PONTE, S. & VANDERSTRAETEN, B. 2022. Dose-volume-based evaluation of convolutional neural network-based auto-segmentation of thoracic organs at risk. *Phys Imaging Radiat Oncol,* 23**,** 109-117.

JOSEPH, V. R. 2022. Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal,* 15**,** 531-538.

KALET, A. M., LUK, S. M. & PHILLIPS, M. H. 2020. Radiation therapy quality assurance tasks and tools: the many roles of machine learning. *Medical physics,* 47**,** e168-e177.

KAMEPALLI, H., KALAPARTI, V. & KESAVADAS, C. 2023. Imaging Recommendations for the Diagnosis, Staging, and Management of Adult Brain Tumors. *Indian Journal of Medical and Paediatric Oncology,* 44**,** 026-038.

KARALIS, V. D. 2024. The Integration of Artificial Intelligence into Clinical Practice. *Applied Biosciences,* 3**,** 14-44.

KAZEMIFAR, S., MCGUIRE, S., TIMMERMAN, R., WARDAK, Z., NGUYEN, D., PARK, Y., JIANG, S. & 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. & OLSSON, L. E. 2021. Prospective Clinical Feasibility Study for MRI-Only Brain Radiotherapy. *Front Oncol,* 11**,** 812643.

LIESBETH, V., MICHAËL, C., ANNA, M. D., CHARLOTTE, L. B., WOUTER, C. & DIRK, V. 2020. Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. *Radiotherapy and Oncology*.

LIU, F., YADAV, P., BASCHNAGEL, A. M. & MCMILLAN, A. B. 2019. MR-based treatment planning in radiation therapy using a deep learning approach. *Journal of applied clinical medical physics,* 20**,** 105-114.

LIU, J., LIN, Z., PADHY, S., TRAN, D., BEDRAX WEISS, T. & LAKSHMINARAYANAN, B. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems,* 33**,** 7498-7512.

LONG, J., SHELHAMER, E. & DARRELL, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 3431-3440.

LU, Y., PATEL, M., NATARAJAN, K., UGHRATDAR, I., SANGHERA, P., JENA, R., WATTS, C. & SAWLANI, V. 2020. Machine learning-based radiomic, clinical and semantic feature analysis for predicting overall survival and MGMT promoter methylation status in patients with glioblastoma. *Magn Reson Imaging,* 74**,** 161-170.

LUAN, S., XUE, X., WEI, C., DING, Y., ZHU, B. & WEI, W. 2023. Machine Learning-Based Quality Assurance for Automatic Segmentation of Head-and-Neck Organs-at-Risk in Radiotherapy. *Technol Cancer Res Treat,* 22**,** 15330338231157936.

MACKAY, K., BERNSTEIN, D., GLOCKER, B., KAMNITSAS, K. & TAYLOR, A. 2023. A Review of the Metrics Used to Assess Auto-Contouring Systems in Radiotherapy. *Clin Oncol (R Coll Radiol),* 35**,** 354-369.

MAYO, C. S., MORAN, J. M., BOSCH, W., XIAO, Y., MCNUTT, T., POPPLE, R., MICHALSKI, J., FENG, M., MARKS, L. B., FULLER, C. D., YORKE, E., PALTA, J., GABRIEL, P. E., MOLINEU, A., MATUSZAK, M. M., COVINGTON, E., MASI, K., RICHARDSON, S. L., RITTER, T., MORGAS, T., FLAMPOURI, S., SANTANAM, L., MOORE, J. A., PURDIE, T. G., MILLER, R. C., HURKMANS, C., ADAMS, J., JACKIE WU, Q. R., FOX, C. J., SIOCHI, R. A., BROWN, N. L., VERBAKEL, W., ARCHAMBAULT, Y., CHMURA, S. J., DEKKER, A. L., EAGLE, D. G., FITZGERALD, T. J., HONG, T., KAPOOR, R., LANSING, B., JOLLY, S., NAPOLITANO, M. E., PERCY, J., ROSE, M. S., SIDDIQUI, S., SCHADT, C., SIMON, W. E., STRAUBE, W. L., ST JAMES, S. T., ULIN, K., YOM, S. S. & YOCK, T. I. 2018. American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. *Int J Radiat Oncol Biol Phys,* 100**,** 1057-1066.

MCCAGUE, C., MACKAY, K., WELSH, C., CONSTANTINOU, A., JENA, R., CRISPIN-ORTUZAR, M. & IMAGING, A. I. E. C. G. 2023. Position statement on clinical evaluation of imaging AI. *Lancet Digit Health,* 5**,** e400-e402.

MEKKI, L., ACHARYA, S., LADRA, M. & LEE, J. 2024. Deep learning segmentation of organs-at-risk with integration into clinical workflow for pediatric brain radiotherapy. *J Appl Clin Med Phys,* 25**,** e14310.

MILLER, K. D., OSTROM, Q. T., KRUCHKO, C., PATIL, N., TIHAN, T., CIOFFI, G., FUCHS, H. E., WAITE, K. A., JEMAL, A., SIEGEL, R. L. & BARNHOLTZ-SLOAN, J. S. 2021. Brain and other central nervous system tumor statistics, 2021. *CA Cancer J Clin,* 71**,** 381-406.

MILLETARI, F., NAVAB, N. & AHMADI, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision (3DV), 2016. IEEE, 565-571.

MLYNARSKI, P., DELINGETTE, H., ALGHAMDI, H., BONDIAU, P. Y. & AYACHE, N. 2020. Anatomically consistent CNN-based segmentation of organs-at-risk in cranial radiotherapy. *J Med Imaging (Bellingham),* 7**,** 014502.

MUKAKA, M. 2012. Statistics corner: a guide to appropriate use of correlation in medical research. *Malawi Med J,* 24**,** 69-71.

OH, S., KIM, J. H., CHOI, S.-W., LEE, H. J., HONG, J. & KWON, S. H. 2019. Physician confidence in artificial intelligence: an online mobile survey. *Journal of medical Internet research,* 21**,** e12422.

OSTROM, Q. T., PATIL, N., CIOFFI, G., WAITE, K., KRUCHKO, C. & BARNHOLTZ-SLOAN, J. S. 2020. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013-2017. *Neuro Oncol,* 22**,** iv1-iv96.

RAGHAVAPUDI, H., SINGROUL, P. & KOHILA, V. 2021. Brain Tumor Causes, Symptoms, Diagnosis and Radiotherapy Treatment. *Curr Med Imaging,* 17**,** 931-942.

RANTA, I., WRIGHT, P., SUILAMO, S., KEMPPAINEN, R., SCHUBERT, G., KAPANEN, M. & 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.

RONG, Y., CHEN, Q., FU, Y., YANG, X., AL-HALLAQ, H. A., WU, Q. J., YUAN, L., XIAO, Y., CAI, B., LATIFI, K., BENEDICT, S. H., BUCHSBAUM, J. C. & QI, X. S. 2023. NRG Oncology Assessment of Artificial Intelligence Deep Learning-Based Auto-segmentation for Radiation Therapy: Current Developments, Clinical Considerations, and Future Directions. *Int J Radiat Oncol Biol Phys*.

RONNEBERGER, O., FISCHER, P. & BROX, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, 2015. Springer, 234-241.

SCHIPAANBOORD, B., BOUKERROUI, D., PERESSUTTI, D., VAN SOEST, J., LUSTBERG, T., DEKKER, A., ELMPT, W. V. & GOODING, M. J. 2019. An Evaluation of Atlas Selection Methods for Atlas-Based Automatic Segmentation in Radiotherapy Treatment Planning. *IEEE Trans Med Imaging,* 38**,** 2654-2664.

SCHMIDT, M. A. & PAYNE, G. S. 2015. Radiotherapy planning using MRI. *Physics in Medicine & Biology,* 60**,** R323.

SCOCCIANTI, S., DETTI, B., GADDA, D., GRETO, D., FURFARO, I., MEACCI, F., SIMONTACCHI, G., DI BRINA, L., BONOMO, P., GIACOMELLI, I., MEATTINI, I., MANGONI, M., CAPPELLI, S., CASSANI, S., TALAMONTI, C., BORDI, L. & LIVI, L. 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.

SHERER, M. V., LIN, D., ELGUINDI, S., DUKE, S., TAN, L. T., CACICEDO, J., DAHELE, M. & GILLESPIE, E. F. 2021. Metrics to evaluate the

performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother Oncol,* 160**,** 185-191.

SIMON, J., SZUMOWSKI, J., TOTTERMAN, S., KIDO, D., EKHOLM, S., WICKS, A. & 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. & GAO, J. 2023. Image Segmentation for MR Brain Tumor Detection Using Machine Learning: A Review. *IEEE Rev Biomed Eng,* PP.

TANG, H., CHEN, X., LIU, Y., LU, Z., YOU, J., YANG, M., YAO, S., ZHAO, G., XU, Y. & CHEN, T. 2019. Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nature Machine Intelligence,* 1**,** 480-491.

THWAITES, D., MOSES, D., HAWORTH, A., BARTON, M. & HOLLOWAY, L. 2021. Artificial intelligence in medical imaging and radiation oncology: Opportunities and challenges. *J Med Imaging Radiat Oncol,* 65**,** 481-485.

TURCAS, A., LEUCUTA, D., BALAN, C., CLEMENTEL, E., GHEARA, C., KACSO, A., KELLY, S. M., TANASA, D., CERNEA, D. & ACHIMAS-CADARIU, P. 2023. Deep-learning magnetic resonance imaging-based automatic segmentation for organs-at-risk in the brain: Accuracy and impact on dose distribution. *Phys Imaging Radiat Oncol,* 27**,** 100454.

UWISHEMA, O., FREDERIKSEN, K. S., BADRI, R., PRADHAN, A. U., SHARIFF, S., ADANUR, I., DOST, B., ESENE, I. & ROSSEAU, G. 2023. Epidemiology and etiology of brain cancer in Africa: A systematic review. *Brain Behav,* 13**,** e3112.

VAASSEN, F., HAZELAAR, C., VANIQUI, A., GOODING, M., VAN DER HEYDEN, B., CANTERS, R. & VAN ELMPT, W. 2020. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology,* 13**,** 1-6.

VAN DEN BERG, C. A. T. & MELIADO, E. F. 2022. Uncertainty Assessment for Deep Learning Radiotherapy Applications. *Semin Radiat Oncol,* 32**,** 304-318.

VAN DER HEYDEN, B., WOHLFAHRT, P., EEKERS, D. B. P., RICHTER, C., TERHAAG, K., TROOST, E. G. C. & VERHAEGEN, F. 2019. Dual-energy CT for automatic organs-at-risk segmentation in brain-tumor patients using a multi-atlas and deep-learning approach. *Sci Rep,* 9**,** 4126.

VAN DER VEEN, J., WILLEMS, S., DESCHUYMER, S., ROBBEN, D., CRIJNS, W., MAES, F. & NUYTS, S. 2019. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiotherapy and Oncology,* 138**,** 68-74.

VAN DIJK, L. V., VAN DEN BOSCH, L., ALJABAR, P., PERESSUTTI, D., BOTH, S., R, J. H. M. S., LANGENDIJK, J. A., GOODING, M. J. & BROUWER, C. L. 2020. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiother Oncol,* 142**,** 115-123.

VAN ROOIJ, W., DAHELE, M., BRANDAO, H. R., DELANEY, A. R., SLOTMAN, B. J. & VERBAKEL, W. F. 2019. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. *International Journal of Radiation Oncology* Biology* Physics,* 104**,** 677-684.

VAN ROOIJ, W., VERBAKEL, W. F., SLOTMAN, B. J. & DAHELE, M. 2021. Using Spatial Probability Maps to Highlight Potential Inaccuracies in Deep Learning-Based Contours: Facilitating Online Adaptive Radiation Therapy. *Adv Radiat Oncol,* 6**,** 100658.

VINOD, S. K., JAMESON, M. G., MIN, M. & HOLLOWAY, L. C. 2016. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiother Oncol,* 121**,** 169-179.

WANG, G., LI, W., AERTSEN, M., DEPREST, J., OURSELIN, S. & VERCAUTEREN, T. 2019a. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing (Amst),* 335**,** 34-45.

WANG, Y., ZHAO, L., WANG, M. & SONG, Z. 2019b. Organ at risk segmentation in head and neck ct images using a two-stage segmentation framework based on 3D U-Net. *IEEE Access,* 7**,** 144591-144602.

WIESINGER, F., BYLUND, M., YANG, J., KAUSHIK, S., SHANBHAG, D., AHN, S., JONSSON, J. H., LUNDMAN, J. A., HOPE, T., NYHOLM, T., LARSON, P. & COZZINI, C. 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**,** 1440-1451.

WIESINGER, F., PETIT, S., HIDEGHÉTY, K., HERNANDEZ TAMAMES, J., MCCALLUM, H., MAXWELL, R., PEARSON, R., VERDUIJN, G., DARÁZS, B., KAUSHIK, S., COZZINI, C., BOBB, C., FODOR, E., PACZONA, V., KÓSZÓ, R., EGYÜD, Z., BORZASI, E., VÉGVÁRY, Z., TAN, T., GYALAI, B., CZABÁNY, R., DEÁK-KARANCSI, B., KOLOZSVÁRI, B., CZIPCZER, V., CAPALA, M. & RUSKÓ, L. 2021. Deep-Learning-based Segmentation of Organs-at-Risk in the Head for MR-assisted Radiation Therapy Planning. *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies.*

WILLEMS, S., CRIJNS, W., SAINT-ESTEVEN, A. L. G., VAN DER VEEN, J., ROBBEN, D., DEPUYDT, T., NUYTS, S., HAUSTERMANS, K. & MAES, F. 2018. Clinical implementation of DeepVoxNet for auto-delineation of organs at risk in head and neck cancer patients in radiotherapy. *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis.* Springer.

WONG, J., FONG, A., MCVICAR, N., SMITH, S., GIAMBATTISTA, J., WELLS, D., KOLBECK, C., GIAMBATTISTA, J., GONDARA, L. & 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**,** 152-158.

WONG, J., HUANG, V., WELLS, D., GIAMBATTISTA, J., GIAMBATTISTA, J., KOLBECK, C., OTTO, K., SAIBISHKUMAR, E. P. & ALEXANDER, A. 2021. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. *Radiation Oncology,* 16**,** 1-10.

YANG, R. & YU, Y. 2021. Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis. *Frontiers in Oncology,* 11**,** 573.

YEUNG, M., RUNDO, L., NAN, Y., SALA, E., SCHÖNLIEB, C. B. & YANG, G. 2023. Calibrating the Dice Loss to Handle Neural Network Overconfidence for Biomedical Image Segmentation. *J Digit Imaging,* 36**,** 739-752.

ZHU, J., CHEN, X., YANG, B., BI, N., ZHANG, T., MEN, K. & DAI, J. 2020. Evaluation of Automatic Segmentation Model With Dosimetric Metrics for Radiotherapy of Esophageal Cancer. *Front Oncol,* 10**,** 564737.

Appendix: Conferences participation

This is the list of conferences in which I have taken part:

Oral presentation

- A mini-oral presentation at the European Society for Radiotherapy and Oncology (ESTRO),3-7 May 2024,Glasgow.
- 'Artificial Intelligence' session at BIR Annual Radiotherapy and Oncology, 29 February–1 March 2024, London.
- Leeds Institute of Medical Research PGR Symposium, 16 March 2023, Leeds - prize awarded.

Poster presentation

- AI Congress 2024/augmentation and automation, 21-22 March 2024, London. Shortlisted abstract award.
- AI congress 2023/AI in action, 23-24 March 2023, London.
- The European Society for Radiotherapy and Oncology (ESTRO), 12-16 May 2023, Vienna.
- Research and Innovation annual conference, 24 May 2023, Leeds.
- CRUK RadNet PhD & Post Doc Symposium, 19 June 2023, London.