Deep Learning for Automatic Segmentation of Vestibular Schwannoma: A Retrospective Study from Multi-Centre Routine MRI

Aaron Kujawa[1], PhD, Reuben Dorent[1], PhD, Steve Connor[1,2,3], MRCP, FRCR, Suki Thomson[2], FRCR, Marina Ivory[1], MD, Ali Vahedi[2], FRCR, Emily Guilhem[2], FRCR, Robert Bradford[5,6], MD, FRCS, Neil Kitchen[5,6], MD, FRCS, Sotirios Bisdas[7] PhD, MD, Sebastien Ourselin[1], PhD, Tom Vercauteren[1], PhD, Jonathan Shapey[1,4], PhD, FRCS

[1]School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom

[2]Department of Neuroradiology, King's College Hospital, London, United Kingdom

[3]Department of Radiology, Guy's and St Thomas' Hospital, London, United Kingdom

[4]Department of Neurosurgery, King's College Hospital, London, United Kingdom

[5]Queen Square Radiosurgery Centre (Gamma Knife), National Hospital for Neurology and Neurosurgery, London, United Kingdom

[6]Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, United Kingdom

[7]Department of Neuroradiology, National Hospital for Neurology and Neurosurgery, London, United Kingdom

Corresponding Author: Aaron Kujawa[1], King's College London
Email: aaron.kujawa@kcl.ac.uk

Key Words: Vestibular Schwannoma, Segmentation, Deep Learning, Convolutional Neural Network, Volumetry, Surveillance MRI

Abstract

Objective:
Automatic segmentation of vestibular schwannoma (VS) from routine clinical MRI can improve clinical workflow, facilitate treatment decisions, and assist patient management. Previously, excellent automatic segmentation results were achieved on datasets of standardised MRI images acquired for stereotactic surgery planning. However, diagnostic clinical datasets are generally more diverse and pose a larger challenge to automatic segmentation algorithms. Here, we show that automatic segmentation of VS on such datasets is also possible with high accuracy.

Methods:
We acquired a large multi-centre routine clinical (MC-RC) dataset of 168 patients with a single sporadic VS who were referred from 10 medical sites and consecutively seen at a single centre. Up to three longitudinal MRI exams were selected for each patient. Selection rules based on image modality, resolution orientation, and acquisition timepoint were defined to automatically select contrast-enhanced T1-weighted (ceT1w) images (n=130) and T2-weighted images (n=379). Manual ground truth segmentations were obtained in an iterative process in which segmentations were: 1) produced or amended by a specialized company; and 2) reviewed by one of three trained radiologists; and 3) validated by an expert team. Inter- and intra-observer reliability was assessed on a subset of 10 ceT1w and 41 T2w images. The MC-RC dataset was split randomly into 3 nonoverlapping sets for model training, hyperparameter-tuning and testing in proportions 70/10/20%. We applied deep learning to train our VS segmentation model, based on convolutional neural networks (CNN) within the nnU-Net framework.

Results:
Our model achieved excellent Dice scores when evaluated on the MC-RC testing set as well as the public testing set. On the MC-RC testing set, Dice scores were 90.8±4.5% for ceT1w, 86.1±11.6% for T2w and 82.3±18.4% for a combined ceT1w+T2w input.

Conclusions:
We developed a model for automatic VS segmentation on diverse multi-centre clinical datasets. The results show that the performance of the framework is comparable to that of human annotators. In contrast, a model trained a publicly available dataset acquired for Gamma Knife stereotactic radiosurgery did not perform well on the MC-RC testing set. The application of our model has the potential to greatly facilitate the management of patients in clinical practice. Our pre-trained segmentation models are made available online. Moreover, we are in the process of making the MC-RC dataset publicly available.

Introduction

Vestibular Schwannoma (VS) is a slow growing, benign tumour that develops in the internal auditory canal. It originates from an abnormal multiplication of Schwann cells within the insulating myelin sheath of the vestibulo-cochlear nerve. It typically presents with hearing loss but also frequently causes tinnitus and balance disturbance. Larger tumours may also cause headaches, cranial neuropathies, ataxia, and hydrocephalus. It is estimated that 1 in 1000 people will be diagnosed with a VS in their lifetime[1]; however, improvements in magnetic resonance imaging (MRI) that facilitate the detection of smaller VS have led to an increased incidence of VS in recent years.[2] Treatment options include conservative management, radiosurgery, radiotherapy, and microsurgery for tumours that are growing or exhibit mass effect.[3]

Previous studies have demonstrated that a volumetric measurement is more accurate than linear measurements and smaller interval changes in VS size may be detected.[4,5] Implementing routine volumetric measurements would enable clinicians to more reliably demonstrate tumour growth and potentially offer earlier interventions. However, available tools make calculating tumour volume assessment a labour-intensive process, prone to variability and subjectivity. Consequently, volumetric methods of measuring tumour size have not been widely implemented in routine clinical practice.[5]

To reduce the workload for clinical staff and free resources, deep learning models have recently been developed to automate this time-consuming and repetitive task. Shapey et al.[6] and Wang et al.[7] recently presented a deep learning framework for automatic segmentation of VS that achieved high accuracy on a large publicly available dataset of MR images acquired for Gamma Knife (GK) stereotactic radiosurgery.[8] As Shapey et al.[6] point out, "the main limitation of [their] study is […] that it was developed using a uniform dataset and consequently may not immediately perform as well on images obtained with different scan parameters."[6]

Such scan parameters include the type of pulse sequence and hardware-specific parameters relating to the MRI scanner and radiofrequency coil such as the magnetic field strength and field inhomogeneities as well as the use and type of contrast agent. These parameters influence the degree of T1-weighted (T1w) and T2-weighted (T2w) contrast, determine the image resolution and field-of-view (FOV) and regulate the image noise and other acquisition artifacts.

As there are no official national guidelines for MRI acquisition protocols for VS in the UK imaging centres choose and optimize pulse sequences independently from each other so that images from different centres are rarely equivalent with scan parameters varying widely.

Although deep-learning models have pushed performance in medical image segmentation to new heights, they are particularly sensitive to shifts between the training and testing data.[9,10] Thus, a model that is trained on data from a single scanner with a fixed set of settings might perform well on images acquired with the same scanner/settings but fail on images acquired differently. However, most 3D-segmentation models for VS published to date rely on standardized radiosurgery treatment planning data from individual institutions with minimal differences in acquisition parameters.[6,7,11–13] The advantage of radiosurgery data is that high quality manual segmentations, which are required for model training and evaluation, are readily available as a part of the treatment plan. The disadvantages are the lack of data variability as well as a bias towards tumours with those characteristics which suggest treatment by radiosurgery. Consequently, these models are unlikely to be robust in a general clinical setting.

In this work we present a deep-learning model for automatic 3D-segmentation that performs well on routine clinical scans performed for diagnosis and surveillance that generalizes to a wide range of scan parameters. In contrast to previous studies, we acquired a large multi-centre routine clinical (MC-RC) longitudinal dataset with images from 10 medical centres. At the time of writing this manuscript, a single study similar to ours was published that employed multi-centre data to develop VS segmentation models.[14] Similarities and differences between both studies are highlighted in the Discussion section of our work. Moreover, we devised a multi-stage annotation pipeline that involved a detailed review of each VS segmentation, to obtain high-quality manual VS segmentations for all 3D-images. The final manual segmentations and images were used to train the model and assess its performance. Our results show that the model performs robustly on most images of the MC-RC testing set. In contrast, models trained on images of a singular scanner dataset acquired for Gamma Knife (GK) stereotactic radiosurgery perform significantly worse on our routine clinical MRI dataset.

Methods

Ethics statement

This study was approved by the NHS Health Research Authority and Research Ethics Committee (18/LO/0532). Because patients were selected retrospectively and the MR images were completely anonymised before analysis, no informed consent was required for the study.

Multi-Centre Routine Clinical (MC-RC) dataset

Study population

This study included adult patients, diagnosed with a unilateral sporadic VS and seen consecutively in the skull base clinic at the National Hospital of Neurology and Neurosurgery, London UK, over an approximate two period from April 2012–May 2014. All adult patients aged 18 years and above with a single unilateral VS were eligible for inclusion in the study, including patients who had previously undergone previous surgical or radiation treatment. Patients with Neurofibromatosis type 2 (NF2) were excluded. All patients had a minimum 5-year surveillance period.

Uncurated dataset

Imaging data from 168 patients referred to the single centre from 10 medical sites with dates of imaging ranging between February 2006 and September 2019 were screened for the study. The median number of timepoints at which each patient underwent an MRI examination was 4 (interquartile range (IQR) 3-7), and the median number of MRI sequences acquired per session was 7 (IQR 4-9). The complete uncurated image dataset comprised MRI sessions from 868 timepoints with 5805 MRI scans.

Automatic image selection

To select the images most suitable for VS delineation and volumetry, an automatic selection pipeline was employed (Figure 1). For each patient, images from at most 3 timepoints were included in the final dataset to limit the number of manual segmentations required. If more than 3 timepoints were available the first, last, and the point closest to the middle were included, while data from all other timepoints were discarded. Consequently, initial diagnostic as well as post-operative images were included in the final dataset. Subsequently, all images with a slice thickness of more than 3.9mm were discarded since such images tend to be unsuitable for volumetry. Finally, at each of the remaining timepoints, images were selected subject to the following selection rules which were designed to automatically select the most suitable MRI scans for manual segmentation.

1) If a high-resolution contrast enhanced T1 (ceT1w) image was available, the image was selected. High-resolution was defined as a voxel spacing of less than 1mm in the three directions of the voxel grid. (19 cases)
2) If a low-resolution (defined as not high-resolution) ceT1w image and a high-resolution T2w image (hrT2w) were available, both were selected. The low-resolution ceT1w image was selected with preference for axial orientation. (63 cases)
3) If a low-resolution ceT1w image was available, but no hrT2w image, the low-resolution ceT1w image was selected. (48 cases)
4) If no ceT1w image was available, a T2w image was selected with preference for high-resolution. (316 cases)

Finally, for cases where multiple images of the same modality passed the selection process, the image with the smallest average voxel spacing was chosen.

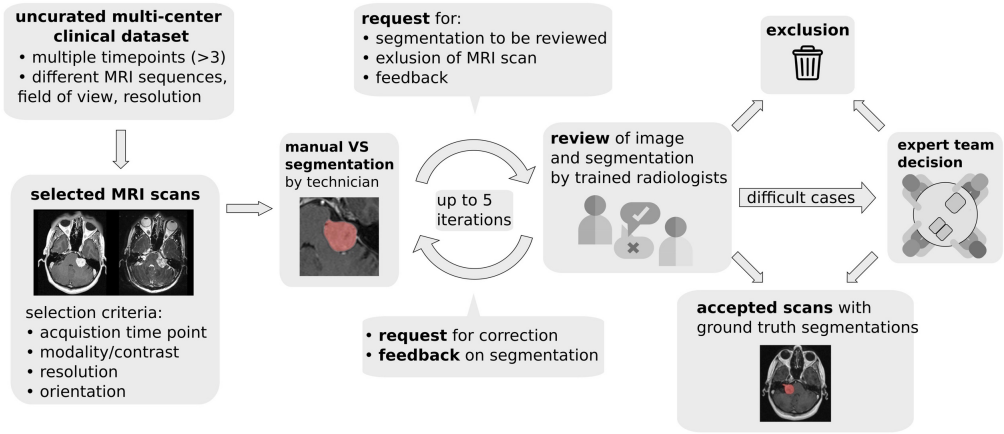

Figure 1: Pipeline for data curation, iterative generation of manual vestibular schwannoma ground truth segmentations, and review of annotated multi-centre clinical (MC-RC) dataset. Focus was placed on the accuracy of segmentation edges, the brain/tumour interface, tumour within the internal acoustic meatus, the exclusion of obvious neurovascular structures from the segmentation, and for post-operative images the exclusion of scar tissue and fat. All segmentations were created, edited, and reviewed using the segmentation tool ITK-SNAP.

Exclusion of imaging data

Subsequently, during the manual annotation process, 39 timepoints were excluded either because parts of the tumour were outside the FOV (n=25), or because a different tumour type (meningioma) was identified instead of a VS (n=2). Post-operative images in which no residual tumour could be identified (n=12) were excluded from the MC-RC dataset, because the corresponding ground truth segmentations without foreground pixels complicate the model training and evaluation in terms of Dice Score Coefficient (DSC). However, model performance on these images was considered in a separate evaluation.

Demographic data

The final MC-RC dataset included 165 patients (males/females 73:92; median age 58 years, IQR 49–67 years). 9 patients had imaging data from a single timepoint, 31 patients from two timepoints, and 125

patients from three timepoints, resulting in a total of 446 timepoints and 509 3D-images. The average time between the first two timepoints was 2.4±1.6 years and between the first and third timepoint 4.9±2.7 years. With respect to the image modality, 67 timepoints included only a ceT1w image, 315 timepoints only a T2w image, and 64 timepoints included both. The ceT1w images were composed of 19 high and 111 low resolution images, while the T2w images comprised 364 high and 15 low resolution images.

Scanner/acquisition settings

Out of 446 MRI exams, 213 were acquired on a SIEMENS, 116 on a Philips, 116 on a General Electrics, and 1 on a Hitachi MRI scanner. The magnetic field strength was 1.5T for 328 exams, 3.0T for 83 exams, 1.0 T for 34 exams and 1.16T for 1 exam. Figure 2 shows the distributions of slice thickness and voxel volume and the intensity distribution of the dataset in comparison to a Gamma-Knife dataset acquired for stereotactic radiosurgery (described in detail below).

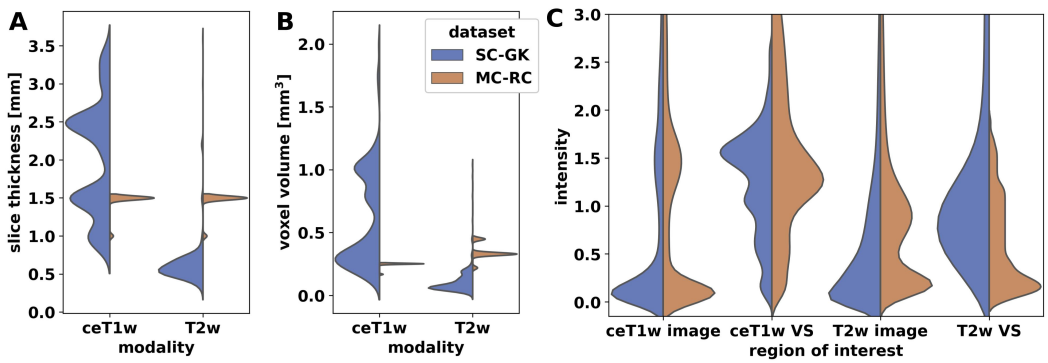

Figure 2: Comparison of multi-centre clinical (MC-RC) and single-centre Gamma Knife (SC-GK) datasets. The violin plots visualize the distribution of slice thickness (A) and image resolution in terms of voxel volume (B) across all ceT1w and T2w images. Parameter values of the MC-RC dataset vary significantly, while parameters of the SC-GK dataset are fixed to a small range of values. (C) compares the normalized voxel intensity distributions of the whole image and the voxels belonging to the vestibular schwannoma (VS). All intensities were normalized by the standard-deviation of the whole image. Images of the SC-GK dataset are similar to each other, leading to more pronounced peaks, while the larger variability in the MC-RC dataset leads to a larger spread of intensity values.

Production of ground truth segmentations

A multi-stage manual annotation pipeline (Figure 1) was designed to obtain high quality ground truth segmentations. The heart of the pipeline comprised an iterative process in which annotations were gradually improved and reviewed at each iteration.

Initial segmentations for each tumour in the remaining dataset were produced by a technician at a company specialized in providing brain measurement services based on MRI scans (Neuromorphometrics, Somerville, Massachusetts, USA) according to our specified guidelines. Capping cysts were included in the segmentation. If a timepoint included both ceT1w and T2w image, the segmentation was performed on the higher resolution image and additional visual assessment of the other image.

Subsequently, each segmentation was reviewed by one of three trained radiologists (MI, AV, EM) who either accepted the segmentation or provided suggestions for improvement in the form of written comments. Alternatively, reviewers had the option to exclude scans that did not fulfil inclusion criteria or refer ambiguous cases to an expert team.

During each iteration, Neuromorphometrics improved segmentations based on the reviewer feedback until each segmentation was accepted, or the corresponding image excluded from the dataset. Finally, a subset of segmentations that had either been flagged by the reviewers as problematic, or had not been accepted after 5 iterations, was reviewed and jointly annotated by the team of experts including two consultant neuroradiologists (SC + ST) and a consultant neurosurgeon (JS).

Inter- and intra-observer reliability

Inter- and intra-observer reliability was assessed on a subset of 10 ceT1w (5 high and 5 low resolution) and 41 T2w images (39 high and 2 low resolution). For intra-observer reliability assessment, 2 sets of segmentations were provided by Neuromorphometrics at two timepoints, approximately 5 months apart. The first set of segmentations was reviewed according to the described iterative process, while the second set produced at the end of the learning curve for the technician was not reviewed. Thus, the measured intra-observer reliability reflects the annotators capability to recreate the first set of validated annotations.

For inter-observer reliability, the first set of validated segmentations provided by Neuromorphometrics and independent, unreviewed segmentations generated by the three reviewers (trained radiologists) were compared. For each pair of annotators and for both modalities, the mean DSC over all images in the subset was calculated and the averaged results reported.

Single-Centre Gamma Knife (SC-GK) dataset

This dataset was chosen as an example of a GK dataset acquired on a single scanner with little variation in sequence parameters. It enables a comparison of models trained on this dataset with models trained on our MC-RC dataset. The SC-GK dataset is a publicly available collection[15,16] of 484 labelled MR image sets of 242 consecutive patients with a unilateral VS undergoing GK Stereotactic Radiosurgery. 51 patients had previously undergone surgery. Images were acquired with a 1.5T MRI scanner (Avanto Siemens Healthineers). For further details we refer to the dataset publication.[20]

Model training and testing

Training

All models were trained and evaluated with nnU-Net, a framework for biomedical image segmentation that yields state-of-the-art results for a wide range of public datasets used in international biomedical segmentation competitions.[17] Based on the training set, the framework automatically determines the architecture of a 3D-U-Net, a well-established type of Convolutional Neural Network (CNN) in the field of medical image segmentation.[18]

A five-fold cross-validation strategy was applied for model training, resulting in five networks, each trained on a slightly different subset of the training data. For inference, a model ensemble was used by averaging the outputs of the five networks prior to generating the segmentation map via an argmax operation. The segmentation networks were trained for 1000 epochs where one epoch is defined as an iteration over 250 mini-batches. The mini-batch size was 2. The optimizer was stochastic gradient descent with Nesterov momentum ($\mu = 0.99$). The initial learning rate of 0.01 was decayed during training according to the "poly" learning rate policy.[19] The loss function was the sum of cross-entropy and Dice loss.[20] For training scripts and the full list of hyperparameters we refer to the publicly available nnU-Net source code. (https://github.com/MIC-DKFZ/nnUNet)

Evaluation

For each of the datasets (MC-RC and SC-GK), three models were trained for the different clinical scenarios in which either only a ceT1w or T2w image, or both modalities are available (ceT1w+T2w).

For each scenario, the available images were split in a ratio 70:10:20 for training, hyperparameter optimization and testing. Each model's performance was evaluated on the corresponding testing sets of both datasets. The sample sizes for each experiment are shown in Table 1. The trained segmentation models, example images, and an inference script were made available online.[21]

Evaluation metrics

The main metric applied to assess and compare the models' segmentation performances was the commonly reported Dice score coefficient, which is the recommended evaluation metric for semantic segmentation.[22] It is defined as:

$$\mathrm{DSC} = \frac{2\sum_i S_i G_i}{\sum_i S_i + \sum_i S_i}$$

where S and G represent the binary segmentation masks of model prediction and ground truth segmentation, respectively. The DSC ranges from 0 (no overlap between model prediction and ground truth) to 1 (perfect overlap).

Additionally, we report the following metrics: Average Symmetric Surface Distance (ASSD), undirected Hausdorff distance (HD), Relative absolute Volume Error (RVE), and distance between the Centres of Mass (COM).

Post-operative cases without residual tumour

Post-operative cases tend to be the most difficult to segment since the residual tumour is often small, and obscured by scar tissue, fat, and an accumulation of CSF. In cases where no residual tumour is present, the DSC is less meaningful. For example, the classification of a single voxel (or more) as a voxel belonging to the tumour would lead to a DSC of 0. Moreover, the other metrics described above are not defined for cases without residual tumour. Therefore, we examined the performance of our model on these cases in a separate evaluation by reporting whether residual tumour was predicted.

Results

Exemplary segmentation results generated by the deep learning models trained on the different input modalities of the MC-RC dataset are shown in Figure 3. The selected example cases have DSCs that are close to the median DSC achieved on the respective testing sets.

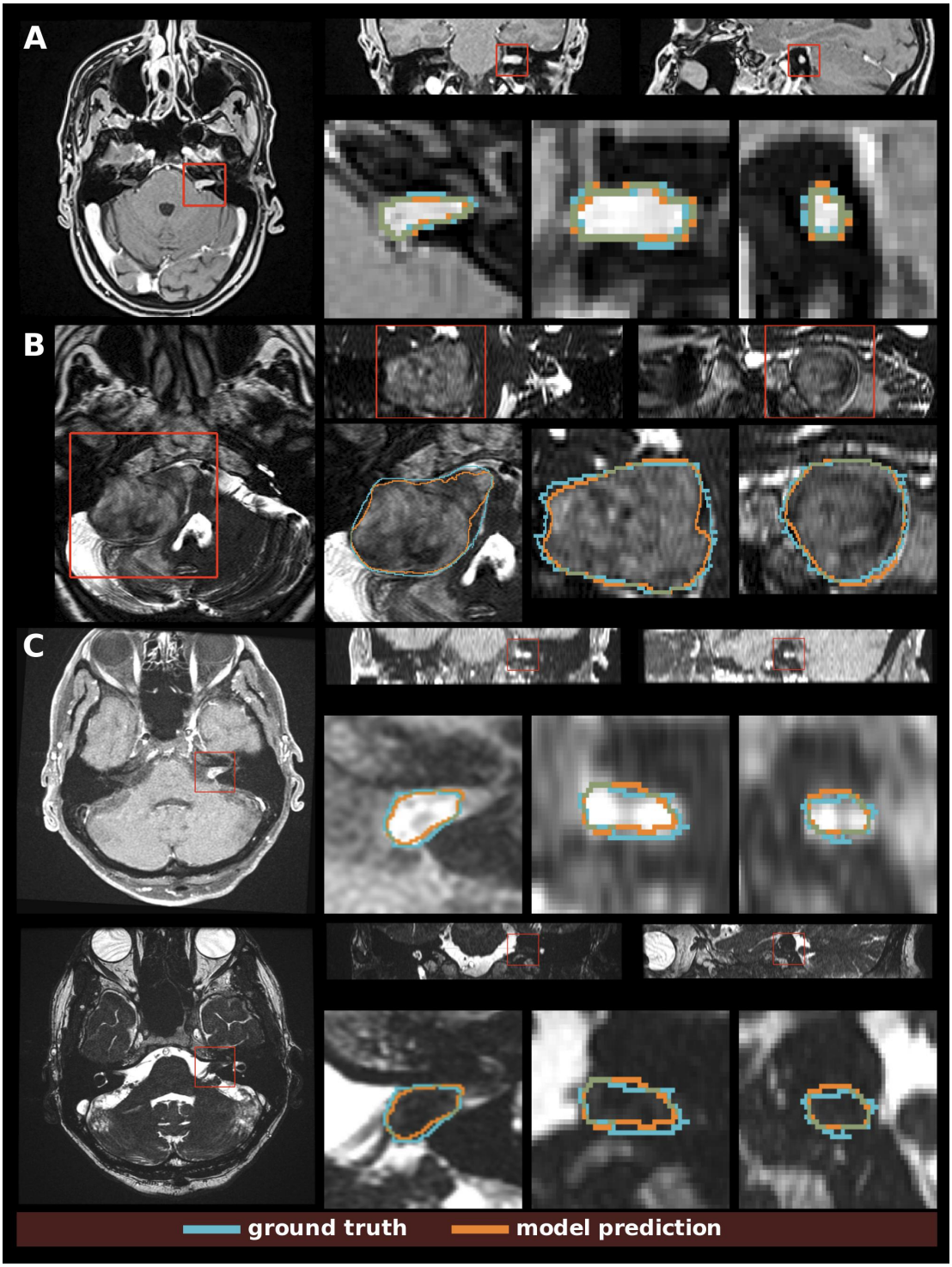

Figure 3: Exemplary segmentation results generated with the 3 models trained on the multi-centre clinical (MC-RC) training sets. Predictions in A) were obtained with the ceT1wmodel, predictions in B) with the T2w model and predictions in C) with the ceT1w+T2w model. Each example shows axial, coronal, and sagittal views of the full MRI image and magnified images of the tumour region. The magnified region is indicated by a red bounding box. Dice scores coefficients were 90.1%, 90.3%, and 83.6% respectively, which is close to the median Dice score coefficient achieved by each model on the MC-RC testing set.

Due to the reduced contrast between tumour and surrounding tissues, tumour boundaries are less pronounced and can be more ambiguous in T2w images than in ceT1w images. This is reflected in the measured inter- and intra-observer reliability which was higher for ceT1w images than for T2w images. The average DSC between two annotators was 88.1±3.4% (minimum: 87.5±4.3%, maximum: 89.1±3.4%) when the segmentation was performed on ceT1w images and 84.5±7.8 (minimum: 82.5±13.9%, maximum: 85.8±7.9%) when performed on T2w. Similarly, intra-observer reliability was higher for ceT1w (87.8±4.4%) than for T2w (84.4±11.5%).

Mean DSCs achieved by deep learning models on MC-RC and SC-GK testing sets are presented in Table 1. The main result of this paper is that the models which were trained on our MC-RC training sets performed well on both the MC-RC and the SC-GK testing sets. On both testing sets, the average DSCs were comparable (MC-RC) or exceeded (SC-GK) the DSCs of our inter- and intra-observer experiments. In contrast, the models trained on the SC-GK training sets performed well only on the SC-GK testing sets but poorly on the MC-RC testing sets. This highlights that variability in the training data is paramount to obtaining robust segmentation results in a clinical setting.

model training dataset	**testing set**	**input modality**	**split sample sizes training/optimization/testing**	**DSC [%]**
MC-RC	MC-RC	ceT1w	47 / 7 / 13	90.8 ± 4.5
		T2w	261 / 43 / 75	86.1 ± 11.6
		ceT1w+T2w	44 / 7 / 12	82.3 ± 18.4
	SC-GK	ceT1w	47 / 7 / 46	91.0 ± 2.6
		T2w	261 / 43 / 46	89.6 ± 4.1
		ceT1w+T2w	44 / 7 / 46	89.9 ± 5.8
SC-GK	MC-RC	ceT1w	176 / 20 / 13	67.1 ± 37.4
		T2w	176 / 20 / 75	48.5 ± 37.9
		ceT1w+T2w	176 / 20 / 12	49.3 ± 37.1
	SC-GK	ceT1w	176 / 20 / 46	95.1 ± 2.2
		T2w	176 / 20 / 46	92.1 ± 3.1
		ceT1w+T2w	176 / 20 / 46	95.2 ± 2.2

Table 1: Dice score coefficients (DSC) achieved by deep-learning models
Dice score coefficients (DSC) achieved by deep-learning models trained on multi-centre clinical (MC-RC) and single-centre Gamma Knife (SC-GK) training sets and evaluated on the MC-RC and SC-GK testing sets. The DSC values correspond to the mean DSC over all images in the testing sets, and the errors correspond to the standard deviation. A boxplot with these results is shown in Figure 4. Moreover, the split sample sizes for training sets, hyperparameter optimization sets and testing sets for each experiment are shown. For a fair comparison between models trained on MC-RC data and models trained on SC-GK data of the same input modality, the respective testing sets were comprised of the same images.

The spread of DSCs is shown in the box and whisker plots of Figure 4. Similar to the results for inter- and intra-observer reliability, performance on ceT1w images was generally better than on T2w.

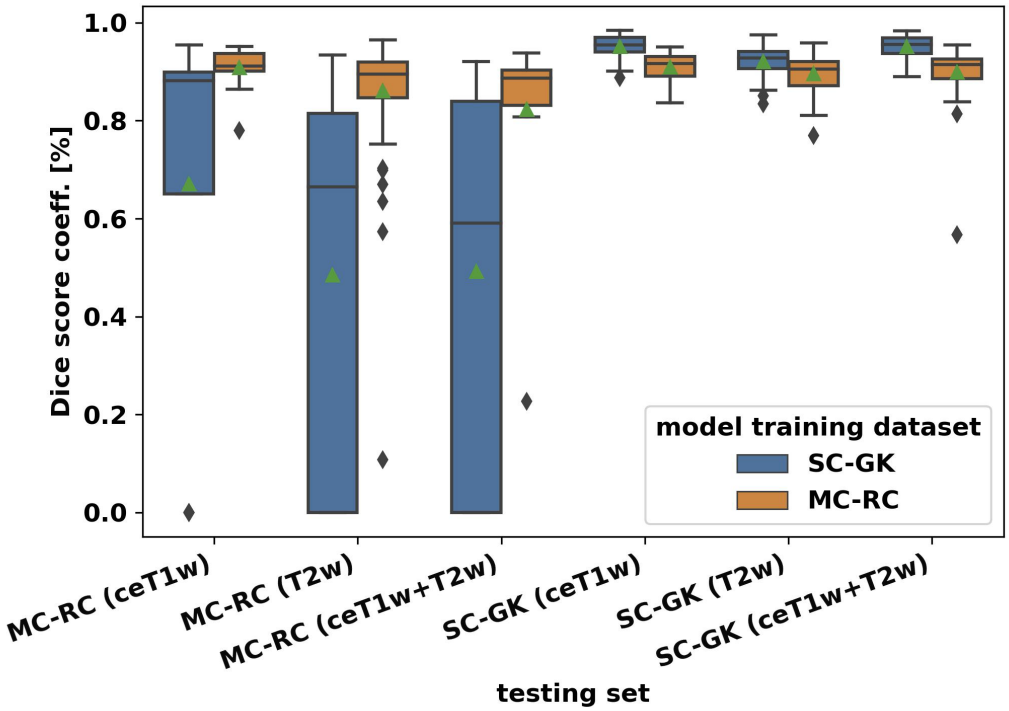

Figure 4: Dice score coefficients achieved by deep learning models on multi-centre clinical (MC-RC) and single-centre Gamma Knife (SC-GK) datasets. The x-axis indicates on which testing set the models were evaluated. The centre vertical line indicates the median and the green triangle indicates the mean. The boxes extend from the lower quartile Q_1 to the upper quartile Q_3, the whiskers extend from $Q_1 - 1.5(Q_3 - Q_1)$ to $Q_3 + 1.5(Q_3 - Q_1)$. Data beyond the whiskers are considered outliers and shown as black diamonds.

When evaluated on the MC-RC testing set, the ceT1w+T2w model trained on the MC-RC training set performed worse than the models trained separately on each modality. This is likely due to the small training set available for the ceT1w+T2w model (Table 1) and the presence of a tumour with a large cyst in the test set which the ceT1w+T2w model did not include in the segmentation (Figure 5c). Although the models trained on the MC-RC training sets generated some outliers there was always a partial overlap between ground truth and segmentation (DSC > 0) and the tumours were never missed completely. The worst cases of each of these models are addressed in the Discussion section. In contrast, the few outliers generated by models trained on the SC-GK training set on the SC-GK testing set have high DSC (>80%) but the same models lead to several complete misses (DSC = 0) on the MC-RC testing sets.

Post-operative cases without residual tumour were assessed separately. The ceT1w model correctly predicted no residual tumour in 6 out of 7 cases. In one case a small tumour residual of 2.5 mm^3 was predicted. The T2w model correctly predicted no residual tumour in 6 out of 9 cases, while predicting volumes of 4.8 mm^3, 8.1 mm^3 and 30.8 mm^3 for the remaining cases. The ceT1w+T2w model correctly predicted no residual in 2 out of 4 cases, while predicting volumes of 0.7 mm^3 and 49.4 mm^3 for the remaining cases.

Medians and IQRs of Dice scores and other commonly reported metrics for segmentation tasks are reported in Table 2.

model training dataset	testing set	input modality	DSC [%]	ASSD [mm]	RVE [%]	HD [mm]	COM [mm]
MC-RC	MC-RC	ceT1w	91.1 ± 3.6	0.2 ± 0.3	5.8 ± 2.8	1.7 ± 2.2	0.2 ± 0.1
		T2w	89.4 ± 7.4	0.3 ± 0.2	6.9 ± 12.9	2.4 ± 2.1	0.3 ± 0.4
		ceT1w+T2w	88.6 ± 7.1	0.4 ± 0.2	12.1 ± 8.2	2.6 ± 2.2	0.5 ± 0.6
	SC-GK	ceT1w	91.6 ± 4.0	0.2 ± 0.1	15.6 ± 10.9	2.0 ± 1.4	0.2 ± 0.2
		T2w	90.5 ± 5.1	0.3 ± 0.2	7.1 ± 10.4	2.5 ± 1.9	0.4 ± 0.4
		ceT1w+T2w	91.4 ± 4.0	0.3 ± 0.2	6.1 ± 9.8	2.5 ± 2.8	0.4 ± 0.3
SC-GK	MC-RC	ceT1w	88.2 ± 24.8	0.4 ± 0.4	19.6 ± 33.2	3.1 ± 5.1	0.4 ± 1.1
		T2w	66.5 ± 81.6	1.1 ± 59.9	32.3 ± 85.4	9.6 ± 69.2	1.9 ± 64.5
		ceT1w+T2w	59.1 ± 83.5	0.8 ± 15.2	62.8 ± 57.5	4.2 ± 21.9	1.3 ± 21.4
	SC-GK	ceT1w	95.4 ± 3.0	0.1 ± 0.1	4.2 ± 5.3	1.5 ± 0.4	0.2 ± 0.2
		T2w	92.8 ± 3.4	0.2 ± 0.2	5.7 ± 8.0	2.0 ± 1.3	0.3 ± 0.3
		ceT1w+T2w	95.6 ± 3.2	0.1 ± 0.1	4.8 ± 5.1	1.5 ± 0.4	0.2 ± 0.2

Table 2: Commonly reported segmentation metrics.
The values represent the median and interquartile range of all images in the testing sets. The metrics are Dice score coefficient (DSC), Average Symmetric Surface Distance (ASSD), Relative Absolute Volume Error (RVE), undirected Hausdorff distance (HD), and distance between the centres of mass (COM).

Discussion

Summary of Contributions

For an automatic VS segmentation model to be useful in a routine clinical setting, accurate and reliable performance irrespective of the set of acquisition parameters is essential. In this work we developed a model for automatic VS segmentation whose application is not limited to MRI images from a specific scanner and acquisition protocol. Rather, by collecting a large multi-centre dataset and providing labour-intensive high-quality annotations it was possible to train and evaluate a model that generalizes well under a wide range of settings. Specifically, the generated automatic segmentations had average DSCs comparable to those of human annotators as measured by inter- and intra-observer experiments. Therefore, this work represents a key step toward the incorporation of automated segmentation algorithms in the clinical workflow and management of VS patients.

For example, based on the model segmentation, automatic surveillance of the patient's tumour growth through longitudinal scans can be performed.[23] Currently, tumour size is assessed by determination of the maximum extrameatal linear tumour dimension, although several studies have shown that tumour volume is a more reliable and accurate metric to measure tumour size.[5,24–26] Based on the model's segmentation the automatic calculation of the tumour volume is a simple task. Moreover, our model could be used to generalize a recent method for automatic classification of VS according to the Koos scale which requires accurate tumour segmentations as an initial step.[27,28] Finally, the model could be used as the initialization step of an interactive segmentation approach.[29,30]

A study similar to ours was published at the time of writing this manuscript. The authors employ ceT1w and T2w imaging data from 37 different centres.[14] In contrast to our study, no longitudinal imaging data but only one time point per patient were considered. Furthermore, all post-operative images were excluded from the model training and analysis. Residual tumour tissue is often difficult to segment,

however, for post-operative surveillance and detection of tumour recurrence, the inclusion of these cases is essential. A comparison of segmentation accuracy in terms of DSC is presented in Table 3.

Comparison with state-of-the-art

Like the model presented here, state-of-the-art methods for VS segmentation employ CNNs based on a U-Net architecture. In comparison with these methods the average DSCs achieved on the MC-RC testing set are slightly lower (Table 3) which was expected due to the increased variability of the MC-RC dataset compared to the relatively homogenous GK datasets used in the respective studies. This interpretation is underlined by the reduced inter-observer reliability (DSC= 88.1±3.4% on ceT1w images and 84.5±7.8 on T2w images) compared to the inter-observer reliability reported on the SC-GK dataset (DSC=93.82±3.08 based on ceT1w and T2w images).[6] For ceT1w+T2w input, another contributing factor is the relatively small number of available cases (Table 1).

reference	dataset	input modality		
		ceT1w	T2w	ceT1w+T2w
Shapey et al. [6]	Gamma Knife	93.4 ± 4.0	88.3 ± 3.9	93.7 ± 2.8
Shapey et al. [15]	Gamma Knife	94.5 ± 2.2	90.7 ± 3.6	-
Wang et al. [7]	Gamma Knife	-	87.3 ± 4.9	-
Lee et al. [12]	Gamma Knife	-	-	90 ± 5
Neve et al. [14]	Multi-Centre (pre-operative)	92 ± 5	87 ± 6	-
This work	Multi-Centre (MC-RC)	90.8 ± 4.5	86.1 ± 11.6	82.3 ± 18.4

Table 3: Comparison with state-of-the-art

Comparison with state-of-the-art results in terms of mean Dice score coefficients and standard deviation. Prior results are based on data extracted from treatment plans for Gamma Knife stereotactic radiosurgery. Another study employs a multi-centre dataset with pre-operative images.

Worst cases

Figure 5 shows the worst model prediction for each of the three inputs. While the worst ceT1w prediction (Figure 5A) has an acceptable DSC of 78.0% it misses to include parts of the tumour with lower contrast agent uptake in the inferior slices. The worst T2w prediction (Figure 5B) was generated in a post-operative case (resection of VS with retrosigmoid approach) with small intrameatal residual. The increased signal of cerebrospinal fluid in the former tumour cavity makes the detection of residuals difficult. The worst ceT1w+T2w prediction (Figure 5C) delineates the solid component of the tumour accurately but misses the large cystic component. It is likely that the small training set available for ceT1w+T2w input did not contain a sufficiently large number of cystic tumours to train the model with respect to their inclusion.

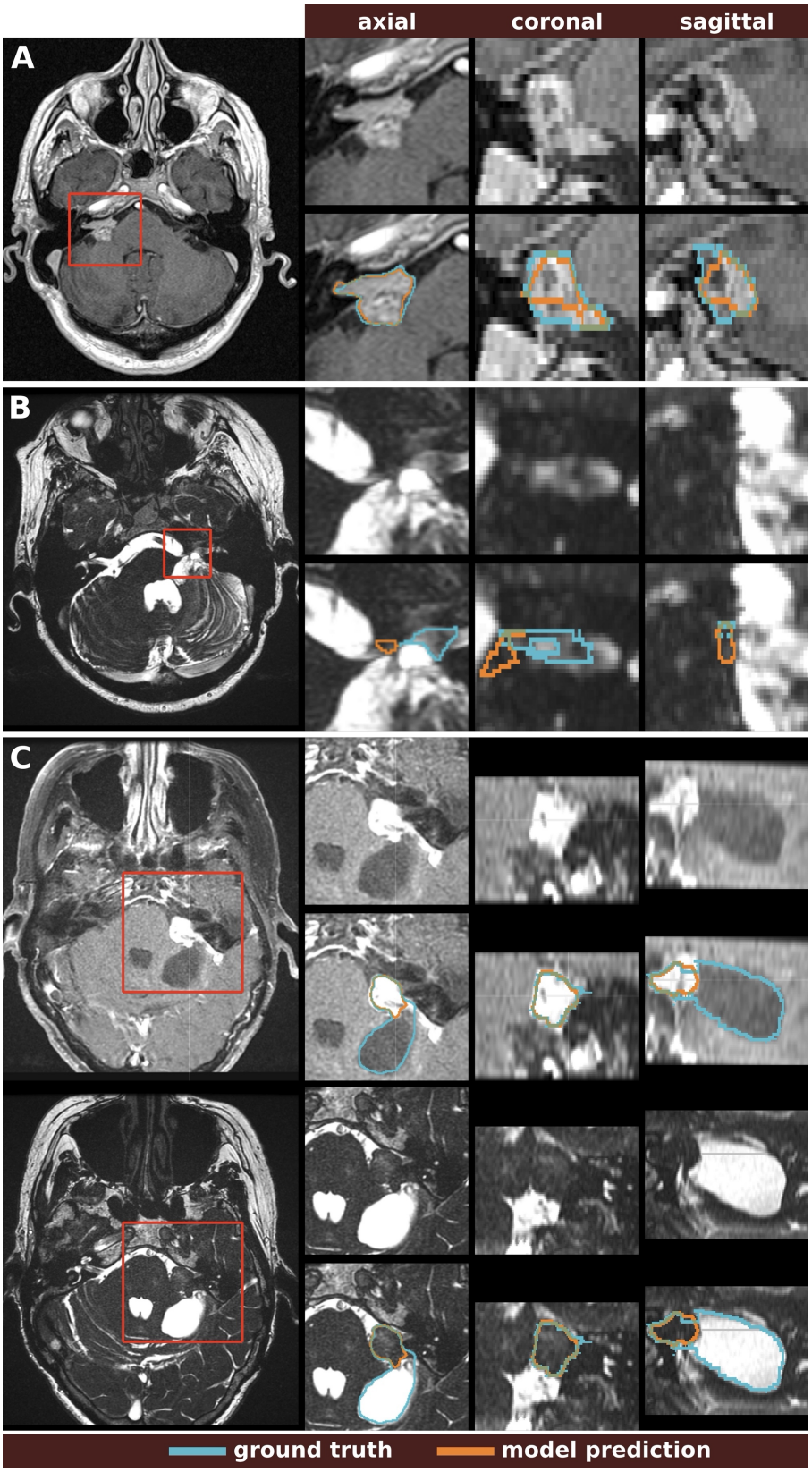

Figure 5: Comparison of worst model predictions for each input modality with the manual segmentation ground truth. The models were trained on the multi-centre clinical dataset using ceT1w images (A), T2w images (B), or their combination ceT1w+T2w (C). Each example shows an axial slice of the full MRI image and magnified images of the tumour region. The magnified region is indicated by a red bounding box. Dice scores were 78.0%, 10.9%, and 22.7% respectively.

Post-operative cases without residual tumour

Cases without residual tumour are edge cases that can be particularly challenging for segmentation algorithms. While our models correctly predicted no residual tumour in 14 out of 20 cases, small residuals were predicted in the remaining cases. A frequently applied strategy to improve model performance is to remove segments below a fixed volume threshold.[31] For example, a reasonable volume threshold can be based on an assumed detection limit for VS in routine clinical MRI of 2mm, which corresponds to a cubic volume of $8mm^3$. In comparison, the smallest predicted volume on the MC-RC dataset was $30mm^3$. Application of this strategy improves the number of correct predictions to 17 out of 20 cases without affecting the other results presented above.

Limitations and future work

While the dataset acquired in this work is among the first large multi-centre datasets for VS segmentation, sample sizes for ceT1w and ceT1w+T2w images were small compared to T2w. This is because the slice thickness of the majority of ceT1w images in the uncurated dataset exceeded the threshold of the inclusion criterion (3.9mm). On such images the annotation of small tumours is difficult and generally not sufficiently accurate for volumetric measurements. We expect that segmentation performance could be further improved by increasing the number of ceT1w images in the dataset.

Furthermore, since this study focuses on sporadic unilateral VS, bilateral tumours from patients with the hereditary condition NF2 were excluded. Due to the simultaneous presence of multiple schwannomas and meningiomas, the segmentation task is disproportionately more difficult. In the future, we plan to integrate NF2 cases into the model development.

Finally, we aim to apply the model for the automatic generation of case reports for multidisciplinary team meetings (MDM). The reports will include multiple automatically generated views of the tumour and the model segmentation and frequently reported tumour measures, such as volume and extrameatal dimensions. We will assess how the reports might facilitate, on the one hand, MDM preparation, and on the other hand, the treatment decision process during the meeting itself.

Conclusion

We developed a model for automatic VS segmentation for diverse clinical images acquired at different medical centres with a wide range of scan protocols and parameters. In clinical practice, the application of this model has the to demonstrate tumour growth more accurately and efficiently, hence facilitating the management of patients.

Acknowledgements

The authors would like to thank Dr Andrew Worth and Gregory Millington for their contributions to the generation of the segmentation ground truth.

Disclosures:
This work was supported by Wellcome Trust (203145Z/16/Z, 203148/Z/16/Z, WT106882), EPSRC (NS/A000050/1, NS/A000049/1) and MRC (MC/PC/180520) funding. Additional funding was provided by Medtronic. TV is also supported by a Medtronic/Royal Academy of Engineering Research Chair (RCSRF1819/7/34). SO is co-founder and shareholder of BrainMiner Ltd, UK.

References

1. Marinelli JP, Lohse CM, Carlson ML. Incidence of intralabyrinthine schwannoma: A population-based study within the United States. *Otol Neurotol*. 2018;39(9):1191-1194. doi:10.1097/MAO.0000000000001875
2. Stangerup SE, Caye-Thomasen P, Tos M, Thomsen J. The natural history of vestibular schwannoma. *Otol Neurotol*. 2006;27(4):547-552. doi:10.1097/00129492-200606000-00018
3. Carlson ML, Habermann EB, Wagie AE, et al. The Changing Landscape of Vestibular Schwannoma Management in the United States - A Shift Toward Conservatism. In: *Otolaryngology - Head and Neck Surgery (United States)*. Vol 153. SAGE PublicationsSage CA: Los Angeles, CA; 2015:440-446. doi:10.1177/0194599815590105
4. Varughese JK, Breivik CN, Wentzel-Larsen T, Lund-Johansen M. Growth of untreated vestibular schwannoma: A prospective study - Clinical article. *J Neurosurg*. 2012;116(4):706-712. doi:10.3171/2011.12.JNS111662
5. MacKeith S, Das T, Graves M, et al. A comparison of semi-automated volumetric vs linear measurement of small vestibular schwannomas. *Eur Arch Oto-Rhino-Laryngology*. 2018;275(4):867-874. doi:10.1007/s00405-018-4865-z
6. Shapey J, Wang G, Dorent R, et al. An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI. *J Neurosurg*. 2021;134(1):171-179. doi:10.3171/2019.9.JNS191949
7. Wang G, Shapey J, Li W, et al. Automatic Segmentation of Vestibular Schwannoma from T2-Weighted MRI by Deep Spatial Attention with Hardness-Weighted Loss. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2019;11765 LNCS:264-272. doi:10.1007/978-3-030-32245-8_30
8. Shapey J, Kujawa A, Dorent R, et al. Segmentation of vestibular schwannoma from MRI — An open annotated dataset and baseline algorithm. *medRxiv*. 2021:2021.08.04.21261588. https://www.medrxiv.org/content/10.1101/2021.08.04.21261588v1%0Ahttps://www.medrxiv.org/content/10.1101/2021.08.04.21261588v1.abstract.
9. Van Opbroek A, Ikram MA, Vernooij MW, De Bruijne M. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Trans Med Imaging*. 2015;34(5):1018-1030. doi:10.1109/TMI.2014.2366792
10. Donahue J, Jia Y, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition. In: *International Conference on Machine Learning*. ; 2014:647-655.
11. Shapey J, Kujawa A, Dorent R, et al. Segmentation of vestibular schwannoma from MRI, an open annotated dataset and baseline algorithm. *Sci Data 2021 81*. 2021;8(1):1-6. doi:10.1038/s41597-021-01064-w
12. Lee C, Lee W-K, Wu C-C, et al. Applying artificial intelligence to longitudinal imaging analysis of vestibular schwannoma following radiosurgery. *Sci Rep*. 2021;11(1):3106. doi:10.1038/s41598-021-82665-8
13. Dorent R, Kujawa A, Ivory M, et al. CrossMoDA 2021 challenge: Benchmark of Cross-Modality Domain Adaptation techniques for Vestibular Schwnannoma and Cochlea Segmentation. *arXiv Prepr arXiv220102831*. 2022.
14. Neve OM, Chen Y, Tao Q, et al. Fully Automated 3D Vestibular Schwannoma Segmentation with and without Gadolinium Contrast: A Multicenter, Multivendor Study. *Radiol Artif Intell*. June 2022. doi:10.1148/ryai.210300
15. Shapey J, Kujawa A, Dorent R, et al. Segmentation of vestibular schwannoma from MRI, an open annotated dataset and baseline algorithm. *Sci Data*. 2021;8(1):286. doi:10.1038/s41597-021-01064-w

16. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging*. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7
17. Isensee F, Petersen J, Klein A, et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. *Inform aktuell*. September 2019:22. doi:10.1007/978-3-658-25326-4_7
18. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 9351. Springer, Cham; 2015:234-241. doi:10.1007/978-3-319-24574-4_28
19. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans Pattern Anal Mach Intell*. 2018;40(4):834-848. doi:10.1109/TPAMI.2017.2699184
20. Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, Pal C. The importance of skip connections in biomedical image segmentation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 10008 LNCS. Springer, Cham; 2016:179-187. doi:10.1007/978-3-319-46976-8_19
21. Kujawa A. Deep Learning for Automatic Segmentation of Vestibular Schwannoma: A Retrospective Study from Multi-Centre Routine MRI -- Deep learning models. July 2022. doi:10.5281/ZENODO.6827679
22. Maier-Hein L, Reinke A, Glocker B, et al. *Metrics Reloaded: Pitfalls and Recommendations for Image Analysis Validation; Metrics Reloaded: Pitfalls and Recommendations for Image Analysis Validation*.; 2022. https://arxiv.org/pdf/2206.01653.pdf. Accessed July 22, 2022.
23. Shapey J, Kujawa A, Dorent R, et al. Artificial Intelligence Opportunities for Vestibular Schwannoma Management Using Image Segmentation and Clinical Decision Tools. *World Neurosurg*. 2021;149:269-270. doi:10.1016/j.wneu.2021.03.010
24. Walz PC, Bush ML, Robinett Z, Kirsch CFE, Welling DB. Three-dimensional segmented volumetric analysis of sporadic vestibular schwannomas: Comparison of segmented and linear measurements. *Otolaryngol - Head Neck Surg (United States)*. 2012;147(4):737-743. doi:10.1177/0194599812447766
25. Tang S, Griffin AS, Waksal JA, et al. Surveillance after resection of vestibular schwannoma: Measurement techniques and predictors of growth. *Otol Neurotol*. 2014;35(7):1271-1276. doi:10.1097/MAO.0000000000000459
26. Roche PH, Robitail S, Régis J. Two- and three dimensional measures of vestibular schwannomas and posterior fossa - Implications for the treatment. *Acta Neurochir (Wien)*. 2007;149(3):267-273. doi:10.1007/s00701-006-1093-x
27. Kujawa A, Dorent R, Connor S, et al. Automated Koos Classification of Vestibular Schwannoma. *Front Radiol*. 2022;2:4. doi:10.3389/fradi.2022.837191
28. Koos WT, Day JD, Matula C, Levy DI. Neurotopographic considerations in the microsurgical treatment of small acoustic neurinomas. *J Neurosurg*. 1998;88(3):506-512. doi:10.3171/jns.1998.88.3.0506
29. Wang G, Zuluaga MA, Li W, et al. DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(7):1559-1572. doi:10.1109/TPAMI.2018.2840695
30. Wang G, Li W, Zuluaga MA, et al. Interactive Medical Image Segmentation Using Deep Learning with Image-Specific Fine Tuning. *IEEE Trans Med Imaging*. 2018;37(7):1562-1573. doi:10.1109/TMI.2018.2791721
31. Antonelli M, Reinke A, Bakas S, et al. The Medical Segmentation Decathlon. *Nat Commun*. 2022;13(1):4128. doi:10.1038/s41467-022-30695-9