Universiteit
Leiden
The Netherlands

Value-based and data-driven vestibular schwannoma care

Neve, O.M.

Citation
Neve, O. M. (2024, November 6). *Value-based and data-driven vestibular schwannoma care*. Retrieved from https://hdl.handle.net/1887/4107527

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4107527

Note: To cite this publication please use the final published version (if applicable).

Fully Automated 3D Vestibular Schwannoma Segmentation with and without Gadolinium Contrast: A Multicenter, Multivendor Study

Olaf Neve*
Yunjie Chen*
Qian Tao
Stephan Romeijn
Nick de Boer
Mark Kruit
Bouwdewijn Lelieveldt
Jeroen Jansen
Erik Hensen
Berit Verbist
Marius Staring

* contributed equally to this work.

Radiol Artif Intell. 2022 Jun 22;4(4):e210300
doi: 10.1148/ryai.210300

ABSTRACT

Purpose

To develop automated vestibular schwannoma measurements on contrast-enhanced T1- and T2-weighted MRI.

Material and methods

MRI data from 214 patients in 37 different centers was retrospectively analyzed between 2020-2021. Patients with hearing loss (134 vestibular schwannoma positive [mean age ± SD, 54 ± 12 years; 64 men], 80 negative) were randomized to a training and validation set and an independent test set. A convolutional neural network (CNN) was trained using five-fold cross-validation for two models (T1 and T2). Quantitative analysis including Dice index, Hausdorff distance, surface-to-surface distance (S2S), and relative volume error were used to compare the computer and the human delineations. Furthermore, an observer study was performed in which two experienced physicians evaluated both delineations.

Results

The T1-weighted model showed state-of-the-art performance with a mean S2S distance of less than 0.6 mm for the whole tumor and the intrameatal and extrameatal tumor parts. The whole tumor Dice index and Hausdorff distance were 0.92 and 2.1 mm in the independent test set. T2-weighted images had a mean S2S distance less than 0.6 mm for the whole tumor and the intrameatal and extrameatal tumor parts. Whole tumor Dice index and Hausdorff distance were 0.87 and 1.5 mm in the independent test set. The observer study indicated that the tool was comparable to human delineations in 85-92% of cases.

Conclusion

The CNN model detected and delineated vestibular schwannomas accurately on contrast-enhanced T1 and T2-weighted MRI and distinguished the clinically relevant difference between intrameatal and extrameatal tumor parts.

1. INTRODUCTION

Vestibular schwannomas are rare, benign intracranial tumors arising from the neurilemma of the vestibular nerve. Initial symptoms usually comprise hearing loss, tinnitus, and balance disturbance. Approximately 60% of tumors show no or minimal progression over time, while 40% are either very large at presentation or show progression during follow-up.[1] Small to medium-sized tumors are not life-threatening and are generally conservatively managed, at least initially, using surveillance with repeated MRIs. Conversely, patients with large tumors at presentation or with tumors that progress during follow-up may need intervention through either radiotherapy or surgery. Currently, there are no reliable predictors for tumor progression.

Currently, tumor progression is determined based on the extrameatal manual diameter measurements on subsequent MRIs.[2] However, these two-dimensional (2D) measurements have considerable error, resulting in inter- and intraannotator differences of 10-40%.[3-5] The more accurate three-dimensional (3D) volume measurements have not been widely applied in clinical practice since these measurements are time-consuming.[3-6]

To address this problem, several automated segmentation tools have been developed in recent years.[7, 8, 9] The reported tools were trained for volume measurement of vestibular schwannoma on gadolinium-enhanced T1-weighed MRIs and sometimes additional T2-weighted MRIs. These tools are increasingly based on deep learning methods, which yield state-of-the-art performance in many vision tasks including medical image segmentation. Deep convolutional neural networks (CNNs), particular the UNet architecture, can reach expert-level performance in various organ segmentation tasks from clinical MRI.[8] Although many variants of the UNet have been proposed and demonstrated task-specific improvements, recent insights suggest that rather than the architecture, careful selection of the hyperparameters and training strategy can have an important effect on performance.[9] The no-new-UNet framework, abbreviated nnUNet, indeed demonstrated this for several organs and imaging modalities.[10, 11] As such, we propose application of nnUNet to address vestibular schwannoma segmentation in our clinical setting.

This study aimed to develop a deep learning CNN model to automatically detect and segment vestibular schwannoma in 3D from T2-weighted and gadolinium-enhanced T1-weighted MRI, acquired from multiple centers using different MRI scanners and scan protocols. We additionally carried out a carefully designed observer study, based on the concept that the radiologists' visual observation of the segmentation results can be a direct, important evaluation of segmentation quality. In addition to conventional metrics, the observer study highlights the applicability of our model in a clinical setting.

1. MATERIALS AND METHODS

This retrospective study was performed at the Leiden university Medical Center, a tertiary referral center for vestibular schwannoma in 2020-2021. The institutional review board approved the study protocol (G19.115) and waived the obligation to obtain informed consent.

2.1 Patients and Data

In total, 214 patients who underwent an MRI examination because of hearing loss were included in the study, with 134 patients who were vestibular schwannoma-positive (mean age, 54 ± [SD] 12 years; 64 men) and 80 who were vestibular schwannoma-negative. Vestibular schwannoma patient selection included a wide spectrum of patient and tumor characteristics such as patient age, sex, tumor size and tumor consistency. All positive patients were adults with a unilateral vestibular schwannoma, and at least one gadolinium-enhanced T1-weighted MRI. High-resolution T2-weighted images were available in 112 patients. MRIs post-surgery or irradiation were excluded. Available MRI examinations were originally acquired in 37 different hospitals on 12 different MRI scanners from 3 major MRI vendors. The MRIs of negative cases, included to optimize detection performance, were solely acquired at the LUMC of adult patients with hearing loss prior to cochlear implantation, and had no demographic data available due to prior anonymization. Patients' characteristics and technical information is shown in Table 1. In positive cases, the intra and extrameatal components[2] and the whole tumor were manually delineated by two annotators independently (ON M.D. 3 years of experience and SR technical physician, 2 years of experience) on the gadolinium enhanced T1-weighted MRI, supervised and when necessary corrected by a senior head-and-neck radiologist (BV). Two senior radiologists with 18 (MK) and 21 (BV) years of experience trained both annotators. Delineation was performed using Vitrea software v7.14.2.227 (Vital Images Inc., Minnetonka, MN, USA). The delineation was automatically propagated to T2-weighted MRI after rigid image registration using elastix.[12, 13] The complete data set was split into a training and validation set (80% from 26 centers), and an independent test set (20% from 11 different centers) on which the model was not trained, see Figure 1 for details. This was done to mimic clinical deployment where new cases may be slightly different from the data seen in the training phase and possibly bear an unknown distribution shift.[14]

Furthermore, the publicly available data set by Shapey et al. was used as additional external test of the contrast-enhanced T1-weighted model (n=242).[15] This dataset contained 47 post-surgery scans, which were omitted from the analysis.

Table 1. Patient and Technical Characteristics

Patients with vestibular schwannoma	**Value**	
N	134	
Age (y), mean (SD)	54 (12)	
Sex, men	64 (48%)	
Cystic component	63 (47%)	
Tumor size		
Intrameatal	28 (21%)	
Small (0-10mm)	19 (14%)	
Medium (11-20mm)	26 (19%)	
Moderately large (21-30mm)	24 (18%)	
Large (31-40mm)	24 (18%)	
Giant (>40mm)	13 (10%)	
Technical MRI features	**Contrast-enhanced T1**	**T2**
	Median (range)	**Median (range)**
N	134	112
In-plane resolution (mm)	0.35x0.35 (0.27x0.27 - 1.0x1.0)	0.29x0.29 (0.23x0.23 - 0.70x0.70)
In-plane matrix	400x400 (256x208 - 560x560)	512x512 (256x192 - 768x652)
TE (ms)	9 (2.38 - 20)	200 (1.53 - 297)
TR (ms)	602.10 (8.76 - 2200)	2400 (4.47 - 5000)
Slice thickness (mm)	1.0 (0.9 - 5.0)	0.6 (0.5 – 1.8)

Note.—Data presented as number of patients (percentage), unless otherwise noted. TE = echo time, TR = repetition time, SD = standard deviation

2.2 CNN Architecture and Training

NnUNet is a deep learning-based segmentation method that automatically selects one of three network architectures, includes pre-processing and post-processing methods, and performs automatic tuning of hyperparameters.[10] In this study, a 3D U-net with five encoder and decoder layers was selected, using randomly cropped 3D image patches of size 320x320x20 voxels as network input during training. The network was trained as a multi-class segmentation task to automatically segment both the intra and extrameatal component of the tumor. Two 3D nnUNets were trained, one for contrast-enhanced T1, and one for T2-weigthed MRI, from scratch with He initialization. Five-fold cross-validation was used, generating five models that were merged by averaging the softmax scores. To deal with multi-center settings, z-scoring normalization was performed to each image independently. All the training images were then resampled to the median spacing of the training dataset using third-order spline interpolation. Training was performed on an NVIDIA Tesla V100 graphics processing unit with 16GB memory using the PyTorch (v1.7.1) library.

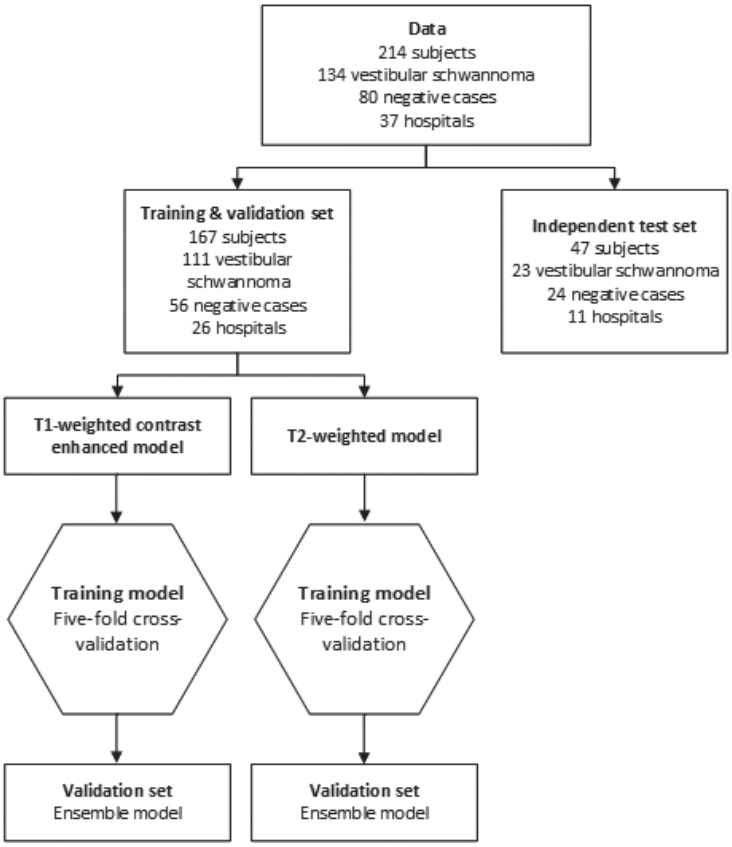

Figure 1. Flowchart of data. Patients were randomized to the training and validation set (80%) and the independent test set (20%). Positive cases were randomized based on the hospital where the scan was acquired, so the independent test set contained data of 11 hospitals that were not used to train the algorithm. For training and validation, five-fold cross-validation was used. The average of the five models is the ensemble model. This ensemble model was evaluated in the independent test set.

2.3 Observer Study

An observer study was performed to test whether the CNN could perform as well as human delineation on contrast-enhanced T1-weighted images. The T1-weighted annotations were propagated to T2-weighted MRI; therefore, the observer study was only conducted for the T1-weighted images. A user interface was created (Fig. 2), showing a gadolinium-enhanced T1-weighted image and the registered T2-weighted image in the top row and the human and automatic delineation in random order on the bottom row, projected on the gadolinium-enhanced T1-weighted MRI. Observers were able to scroll through the MRI, manually adjust its brightness and contrast, and toggle the segmentations on and off for optimal assessment. The observers were a head-and-neck radiologist (BV) and a skull base otorhinolaryngologist (EH, 18 years of experience), blinded for case information and delineation type (human or automated). The observers were

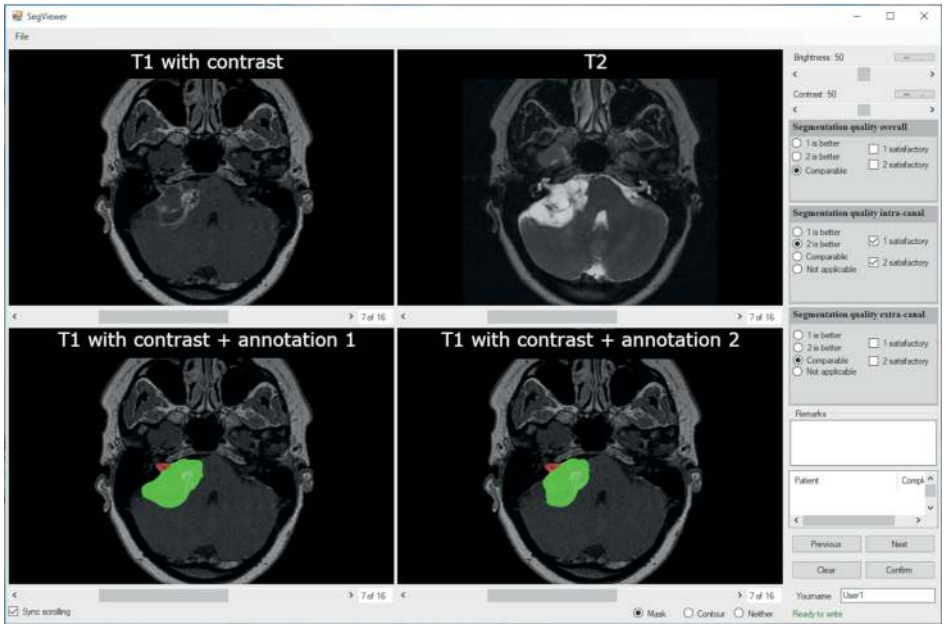

Figure 2. Observer study interface. The top row shows the clean, gadolinium-enhanced T1-weighted MRI and T2-weighted MRI. The bottom row shows the convolutional neural network and human annotations, randomized to left and the right pane, respectively. The multiple-choice questions for each observer are shown at the right side of the interface. The observers could additionally add free text comments.

asked to rate and compare the two delineations by answering two separate questions about the intra- and extrameatal part and the whole tumor: (1) Which delineation is better (annotation 1, annotation 2, or comparable), (2) Is the annotation quality satisfactory (yes or no). In a consensus meeting, cases in which observers did not agree were discussed. The consensus results are presented in section 3.5.

2.4 Testing and Statistical Analysis

All test images were resampled in the same way as the training data, and a sliding window approach was used to predict images with a window size of 320x320x10 voxels, which is the same as the network's input size. The step size is half of the window size, and a Gaussian weighted function was applied in aggregating the predictions. To eliminate false detection, connected component-based post-processing was performed. Only the largest connected component in the predictions was kept. Tumor detection by the CNN was defined as at least one voxel being detected. The performance was evaluated using the Dice index measuring overlap of the delineations, 95th percentile Hausdorff distance indicating the maximum distance between delineations, surface-to-surface (S2S) distance indicating the mean distance between delineations, and the relative volume error (RVE) indicating the difference in volume in percentage. One of the annotator's

(ON annotator 1) delineations were used for training and quantitative evaluation. The results were plotted in box-and-whisker plots. Furthermore, inter-annotator variability was investigated. Differences between the prediction performance of each annotator and the inter-annotator variabilities were tested using Wilcoxon signed-rank test. In addition, a post hoc analysis was conducted of T1-model performance with respect to tumor size, according to the classification by Kanzaki et al.2 To avoid group sizes that were too small per category, the validation and test set were pooled and a Kruskal Wallis test was performed. P-values < .05 were considered statistically significant. Observer agreement before the consensus meeting on satisfactory degree for segmentation and human delineation was expressed as percentage agreement. All analyses were performed in Python (v3.8.2) with NumPy (v1.20.2), SciPy (v1.3.3) and the sklearn (v0.23.2) library.

3. RESULTS

The CNN detected tumors with 100% sensitivity and 99.1% specificity for the validation set and 100% sensitivity and 100% specificity for the test set. The algorithm was able to calculate the segmentation with a median runtime of 78 seconds per patient.

3.1 Performance with Contrast-enhanced T1-weighted MRI

The results of the CNN on contrast-enhanced T1-weighted MRI are shown in Table 2 and Figure 3A. S2S distance of the whole tumor is 0.31mm ± [SD] 0.36 and 0.47 mm ± 0.67 in the validation set and independent test set, respectively. These S2S distances are around the in-plane voxel size and lower than the slice thickness. The whole tumor Hausdorff distance in the independent test set was 2.10mm ± 3.34, and 1.34mm ± 0.84 and 2.18mm ± 3.43, in the intra- and extrameatal parts, respectively. All the median Hausdorff distances were below the 2 mm threshold, which is often used in clinical practice to define 2D growth.[1] T1 model performance on the independent test set was comparable to the results in the validation set, indicating robust external validity. Remarkably, the independent test set had higher mean Hausdorff properties compared to the median due to two outliers (cystic tumor) in the test set which influenced the Hausdorff distance and its standard deviation. Dice indices for the whole tumor were above 0.91±0.10 and 0.92±0.05 in both sets, and RVE 7.6±4.9 and 10.2±9.1 , with lower values for the intra- and extrameatal parts of the tumor due to the sensitivity of Dice and RVE to small volumes. Figure 4 shows some examples of the T1 model compared with annotator 1.

Table 2. Quantitative Results of the Contrast-enhanced T1-weighted Model

(a) Validation set								
	Dice		**95% Hausdorff (mm)**		**S2S (mm)**		**RVE (%)**	
	mean ± SD	**median**	**mean ± SD**	**median**	**mean ± SD**	**Median**	**mean ± SD**	**median**
Whole tumor	0.91 ± 0.10	0.93	1.13 ± 1.45	1.00	0.31 ± 0.36	0.24	7.59 ± 8.10	4.88
Intrameatal	0.78 ± 0.21	0.85	1.26 ± 0.78	1.00	0.31 ± 0.20	0.26	19.7± 43.5	11.5
Extrameatal	0.83 ± 0.26	0.93	1.43 ± 1.67	1.00	0.41 ± 0.43	0.31	12.0 ± 21.6	4.94
(b) Independent test set								
	Dice		**95% Hausdorff (mm)**		**S2S (mm)**		**RVE (%)**	
	mean ±SD	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**
Whole tumor	0.92 ± 0.05	0.93	2.10 ± 3.34	1.00	0.47 ± 0.67	0.36	10.2 ± 9.1	7.1
Intrameatal	0.81 ± 0.08	0.81	1.34 ± 0.84	1.12	0.37 ± 0.23	0.32	14.7 ± 14.8	6.8
Extrameatal	0.89 ± 0.12	0.93	2.18 ± 3.43	1.00	0.52 ± 0.68	0.37	12.1 ± 16.9	6.5
(c) Publicly available dataset by Shapey et al.								
	Dice		**95% Hausdorff (mm)**		**S2S (mm)**		**RVE (%)**	
	mean ±SD	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**
Whole tumor	0.88 ± 0.04	0.88	1.31 ± 0.22	1.30	0.39 ± 0.12	0.37	27.6 ± 11.9	26.1

Dice index, Hausdorff distance, surface-to-surface distance (S2S) and relative volume error (RVE) of the model compared with annotator 1 in the (a) validation set, (b) independent test set, and (c) publicly available data set by Shapey et al. The publicly available data set seems to have structurally smaller ground truths, as can be seen in Fig. D in the supplemental material. SD = standard deviation

The CNN model, when applied to the publicly available dataset of Shapey et al., performed at the same level as with the independent test set, with a mean Dice index of 0.88±0.04, a mean Hausdorff distance of 1.31mm±0.22, a mean S2S distance of 0.39 mm±0.12, and an RVE of 26%±11.9.

3.2 Performance with T2-weighted MRI

The results of the whole tumor and the intra- and extrameatal parts are summarized in Table 3 and Figure 3B. S2S distances ranged between 0.46±0.28 and 1.00 mm ±3.75for all tumor parts in both data sets. Hausdorff distance of the whole tumor in the validation set was 3.12 mm ±9.28, with a smaller value in the independent test set (1.52 mm ±0.76). Whole tumor Dice indices were 0.82±0.19 and 0.87± 0.06 and RVE values ranged from 12.1% ± 10.8 and 24.5% ±98.8 in both data sets. Intrameatal tumors had worse Dice indices and RVE0.69±0.23 and 0.74±0.08 and 14.5%± 18.7 and % ± 21.2, respectively, likely due to the low contrast between the tumor and adjacent petrous bone in T2-weighted images. Overall T2 performance was slightly degraded compared to post-contrast T1. However, S2S distances below 1 mm indicate acceptable performance.

Table 3. Quantitative Results of the T2-weighted Model

	Dice		95% Hausdorff (mm)		S2S (mm)		RVE (%)	
(a) Validation set								
	mean ± SD	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**
Whole tumor	0.82 ± 0.19	0.87	3.12 ± 9.28	1.27	1.00 ± 3.75	0.42	24.5 ± 98.9	7.60
Intrameatal	0.69 ± 0.23	0.78	1.60 ± 0.95	1.20	0.46 ± 0.28	0.40	14.5 ± 18.7	8.39
Extrameatal	0.77 ± 0.28	0.88	2.70 ± 3.19	1.67	0.82 ± 1.01	0.54	30.9 ± 73.3	18.5
(b) Independent test set								
	Dice		**95% Hausdorff (mm)**		**S2S (mm)**		**RVE (%)**	
	mean ± SD	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**	**mean ± SD**	**median**
Whole tumor	0.87 ± 0.06	0.89	1.52 ± 0.76	1.21	0.54 ± 0.31	0.47	12.1 ± 10.8	9.01
Intra meatal	0.74 ± 0.08	0.74	1.64 ± 0.59	1.50	0.52 ± 0.20	0.50	12.6 ± 21.2	5.27
Extrameatal	0.85 ± 0.17	0.89	1.60 ± 0.92	1.14	0.56 ± 0.33	0.42	22.3 ± 14.9	20.0

Dice index, Hausdorff distance, surface-to-surface distance (S2S) and relative volume error (RVE) of the model compared with the annotator 1 in the (a) validation set and (b) independent test set. SD = standard deviation

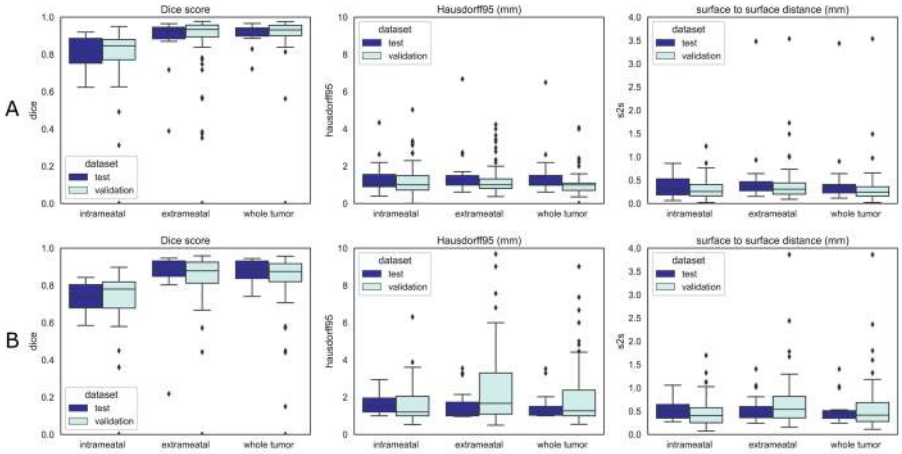

Figure 3. Quantitative boxplots of convolutional neural network tumor segmentation performance. The Dice 95% Hausdorff (Hausdorff95) distance, and surface-to-surface distance (S2S) measures are shown from left to right. (A) Boxplots of the contrast-enhanced T1 model. (B)Results of the T2-weighted model. Validation set results are shown in sky blue and independent test set results in dark blue.

3.3 Inter-annotator Variability

Comparisons between the T1-weighted model and the two annotators and between the two annotators are shown in Table 4 and Figure 5. The comparison between both annotators shows the whole tumor inter-annotator variability, resulting in a Dice index around 0.91 and RVE of 7-9%. When the model was compared to each annotator in both datasets, S2S distances were similar and below 0.5 mm. The model was trained on annotator 1, but the results compared with annotator 2 are similar for all quantitative measures.

Table 4. Comparison of the Model with Annotators and Inter-annotator Variability

	(a) Validation set											
	Dice			95% Hausdorff (mm)			S2S (mm)			RVE(%)		
	mean ± SD	p-value	median	mean ± SD	p-value	median	mean ± SD	p-value	median	mean ± SD	p-value	median
CNN – ann 1	0.91 ± 0.10	<.001	0.93	1.13 ± 1.45	<.001	1.00	0.31 ± 0.36	<.001	0.24	7.59 ± 8.10	.21	4.88
CNN – ann 2	0.90 ± 0.11	.40	0.92	1.33 ± 1.52	.18	1.00	0.36 ± 0.36	.58	0.31	10.1 ± 9.8	.35	7.1
ann 1 – ann 2	0.91 ± 0.05		0.92	1.27 ± 0.82		1.00	0.34 ± 0.20		0.31	9.01 ± 9.14		6.40
	(b) Independent test set											
	Dice			95% Hausdorff (mm)			S2S (mm)			RVE(%)		
	mean ± SD	p-value	median	mean ± SD	p-value	median	mean ± SD	p-value	median	mean ± SD	p-value	median
CNN – ann 1	0.92 ± 0.05	.56	0.93	2.10 ± 3.34	.83	1.00	0.48 ± 0.67	.67	0.35	10.2 ± 9.1	.28	7.1
CNN – ann 2	0.91 ± 0.05	.69	0.93	2.08 ± 3.41	.94	1.07	0.50 ± 0.68	.96	0.35	9.69 ± 9.19	.57	7.72
ann 1 – ann 2	0.92 ± 0.04		0.93	1.20 ± 0.65		1.00	0.34 ± 0.19		0.36	6.93 ± 5.32		4.53

Note.—Dice index, Hausdorff distance, surface-to-surface distance (S2S), and relative volume error (RVE) of the model compared withannotator (ann) 1, annotator 2 and both annotators of the contrast-enhanced T1-weighted model. Results of the (a) validation set and (b) independent test set are shown.CNN – convolutional neural network. P-values denotes Wilcoxon signed ranks test between this quantitative score and corresponding score of annotator1-annotator 2 (the third row).

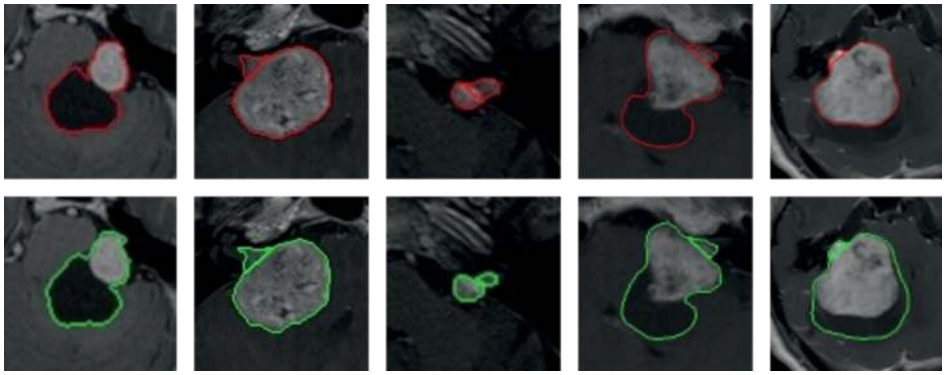

Figure 4. Examples of different (cystic, large, small) vestibular schwannoma whole tumor annotations, including the separation between the intra- and extrameatal tumor parts, of contrast-enhanced T1-weighted MRIs. The first row shows the convolutional neural network (CNN) predictions in red, and the second row shows the delineation of annotator 1 in green. The first, fourth and fifth tumors are potentially hard to delineate for the CNN due to the large peripheral cystic tumor parts. The Dice scores of these patients were 0.96, 0.96, 0.91, 0.93 and 0.72, respectively, and the surface-to-surface distances (mm) were 0.39, 0.21, 0.24, 0.35 and 3.44, respectively.

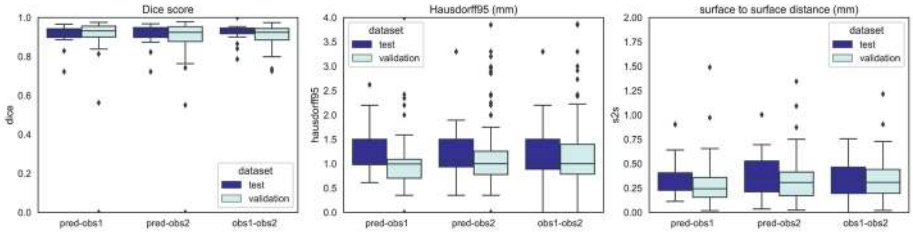

Figure 5. Quantitative measures of whole tumor convolutional neural network performance compared with the two annotators on contrast enhanced T1-weighted MRIs. Inter-annotator variability is also shown (obs 1-obs 2). From left to right the Dice indices, 95% Hausdorff distance (Hausdorff95) and surface-to-surface (S2S) distance boxplots are shown. The validation set results are shown in sky blue and the independent test set in dark blue.pred = CNN prediction, obs = observer.

3.4 Performance by Tumor Size

In the supplemental material (Fig. C) the results of the performance per size category are shown. Whole tumor results show a pattern of higher Dice indices for larger tumors, which was expected since the Dice index is sensitive to size. S2S was very similar in all size groups (<0.5mm), although S2Swere slightly larger in larger tumors ($p<0.001$). Results of intra- and extrameatal tumor parts show stable performance, except for four outliers in the small tumors (inaccurate extrameatal segmentation) and three outliers in giant tumors (false intrameatal tumor detection). In these tumors, there were some differences between model and human delineation for a completely intrameatal tumor with or without a tiny extrameatal part (small) or an extrameatal tumor with or without an intrameatal part (giant).

3.5 Outcomes of Observer Study

Agreement between the two observers before the consensus meeting on whole tumor segmentation quality was 131/134 (98%) for the human annotators and 127/134 (95%) for the CNN.

CNN segmentations of the whole tumor were considered comparable to the human segmentations in 103/111 (93%) of cases in the validation set and 20/23 (87%) in the test set. The CNN segmentations were rated better than the human segmentations in 2/111 (2%) and 2/23 (9%) of cases in the two datasets, respectively. Intrameatal segmentations were rated comparable to or better than human segmentations in 100/106 (94%) and 22/23 (96%) in the validation and test sets, respectively. For extrameatal segmentations, these percentages were 83/97 (86%) and 18/22 (82%).

In addition, the observers considered 104/111 (94%, validation set) and 20/23 (87%, test set) of whole tumor CNN segmentations satisfactory. Intrameatal tumor parts were considered satisfactory in 100/104 (94%, validation set) and 22/23 (96%, test set) of segmentations. Extrameatal tumor parts were considered satisfactory in 90/97 (93%, validation set) and 18/22 (82%) (test set) of segmentations. For human segmentations of the intrameatal tumor, 98/104 (94%) in the validation and 23/23 (100%) in the test set were rated satisfactory. Other satisfaction levels of the human segmentations were 110/111 (99%,validation set) and 22/23 (96%) (test set) for the whole tumor and 89/97 (92%, validation set) and 21/22 (95%, test set) for the extrameatal tumor part.

4. DISCUSSION

To our knowledge, this is the first study which presents the results of a multicenter, multivendor automated vestibular schwannoma segmentation tool. The developed 3D CNN-tool measured tumor volume with very high accuracy on contrast-enhanced T1-weighted MRIs and T2-weighted MRIs. The S2S distances were between 0.4 and 0.9 mm, which was lower than the median slice thickness of 1.0 mm. The observer study suggests that the tool performs comparably to human delineation in 87-93% of the cases.

The contrast-enhanced T1-weighted MRI model provided excellent S2S distances and Dice indices. However, the standard deviations of the Hausdorff distances were remarkably large in the test set due to two outliers which contained peripheral cysts in the extrameatal part. The model did have difficulties with tumors containing large peripheral cysts (see supplemental material Fig. A for examples), which were sometimes partially included by the model.

Evaluation of the model on the publicly available dataset of Shapey et al. showed robust performance on contrast-enhanced T1-weighted images.[15] Interestingly, the ground-truth delineations of Shapey et al. are smaller than those used in the current study, as shown in supplemental Figure D, reducing Dice index from 0.93 to 0.88.[7] When erosion (3x3 kernel) was performed on model delineation, Dice index improved again to 0.93±0.03, supporting this observation. The delineations by Shapey et al. were used for radiotherapy purposes, where preventing damage to the surrounding tissue is important, warranting conservative delineation. We did not compare the T2-weighted images of the publicly available dataset to those in our dataset given differences in the imaging characteristics (echo time and repetition time) and region of interest (whole brain vs. cerebellopontine angle region).

In our study, CNN performance on T2-weighted MRI was slightly less accurate with more uncertainty compared with the contrast-enhanced T1-weighted images. This was particularly the case in polycystic tumors, where the tumor border was hard to distinguish from the cerebrospinal fluid solely on T2 (supplemental fig. B). In one case, the model could not distinguish a small tumor obliterating the internal meatus. In another single case, the model detected the contralateral eye as a false positive volume outside the region of interest.

The RVE values of the whole tumor ranged from 8-12%, compared to 9-10% inter-annotator volume differences. Only the T2 model in the validation set had a larger RVE of 25%.The performance of our CNN compared with human volume measurement is below previously reported inter-annotator variabilities ranging from 15-20%[3-5], and also below the generally accepted threshold of 20% before volume increase is considered growth. Two dimensional measurements are advised in the consensus guidelines but have high intra-observer variabilities ranging from 10-40%.[2-5] Volume measurement is more accurate, and the proposed tool can reduce the workload which has been a barrier for clinical adoption, enabling the shift from 2D measurement. Since documented detection and evaluation of tumor growth is one of the main factors that indicate the need for treatment, be it surgical removal or irradiation, this is of notable clinical relevance.

A unique attribute in vestibular schwannoma research is the integration of an observer study. Determining a ground truth is necessary in artificial intelligence imaging studies. The reliability of the ground truth is uncertain when human observer performance is suboptimal, as described above. Our observer study allowed evaluation of the comparability between CNN segmentation and human segmentation, the reference standard. Our results showed that the CNN tool performs comparably to human observers in the vast majority of cases, supporting the quantitative results that the tool is feasible and

robust for usage in clinical practice. Whole tumor delineations performed slightly better than the extrameatal delineations, which should be considered when using the tool in clinical practice as extrameatal tumor progression is of particular interest for treatment decisions.

Artificial intelligence tools for vestibular schwannoma segmentation that have been previously proposed were all performed on data from a single center.[5,6,7] In clinical practice, however, diagnostic and follow-up scans are often performed in different centers using a variety of scanners and MRI protocols. In addition to its documented performance in a multicenter, multivendor setting, our method contains three features that make the tool more suitable for clinical practice compared to previous automated vestibular schwannoma delineation methods. First, the tool can distinguish between the intra- and extrameatal parts of the tumor. This distinction is important for clinical decision-making, as extension and progression of the extrameatal part usually determines the need for intervention. For this reason, current tumor staging systems are based mainly on the extrameatal dimensions of the tumor, while the intrameatal part is not measured.[2, 16] Second, the proposed tool can also delineate on solely T2-weighted MRI. Given the ongoing debate on use of gadolinium, this is a valuable feature.[17] Third, unlike previous models, our network is a fully 3D network that enables complete use of intra-slice information.

This study has some inherent limitations. First, the study was performed using retrospective MRI data. While this is an accepted method for the development of a new tool, some bias may be introduced by using older MRIs with suboptimal image quality and resolution. Therefore, accuracy and efficacy should also be investigated in prospective studies before clinical implementation and use. Second, for training of the T2 model, the registered human T1 delineations were used. This might have resulted in a suboptimal ground truth for the T2 model, although the reported tumor size correlations between T1 and high-resolution T2 were high.[18, 19] Third, the model is only trained on data before treatment and cannot be used for follow-up after surgery or radiotherapy without retraining.

Implementation of the CNN tool in clinical practice could lead to more accurate volume measurements of vestibular schwannoma at diagnosis and during follow-up, while reducing the workload of radiologists. Tumor volume change over time is a decisive factor in clinical decision making, and future research should focus on the tool's performance in a prospective study and its impact on clinical practice. The tool might be improved using post processing to reduce the false positive volumes outside the region of interest. In addition, the algorithm used for development of the tool could be adapted to analyze

other slow-growing skull base pathologies that are typically approached by a wait and scan policy, such as meningiomas.[20]

The proposed CNN model delineated vestibular schwannoma from MRI with excellent accuracy, comparable to human performance in the majority of cases. The CNN tool made the clinically relevant distinction between intra- and extrameatal tumor parts. The study shows the feasibility of automatically detecting and evaluating vestibular schwannoma with or without contrast administration in large datasets acquired from multiple medical centers and MRI vendors.

REFERENCES

1. Carlson ML, Link MJ. Vestibular Schwannomas. N Engl J Med 2021;384(14):1335-1348. doi: 10.1056/nejmra2020394
2. Kanzaki J, Tos M, Sanna M, Moffat DA, Monsell EM, Berliner KI. New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma. Otol Neurotol 2003;24(4):642-648; discussion 648-649. doi: 10.1097/00129492-200307000-00019
3. Varughese JK, Wentzel-Larsen T, Vassbotn F, Moen G, Lund-Johansen M. Analysis of vestibular schwannoma size in multiple dimensions: a comparative cohort study of different measurement techniques. Clin Otolaryngol 2010;35(2):97-103. doi: 10.1111/j.1749-4486.2010.02099.x
4. Mackeith S, Das T, Graves M, Patterson A, Donnelly N, Mannion R, Axon P, Tysome J. A comparison of semi-automated volumetric vs linear measurement of small vestibular schwannomas. Eur Arch Otorhinolaryngol 2018;275(4):867-874. doi: 10.1007/s00405-018-4865-z
5. van de Langenberg R, de Bondt BJ, Nelemans PJ, Baumert BG, Stokroos RJ. Follow-up assessment of vestibular schwannomas: volume quantification versus two-dimensional measurements. Neuroradiology 2009;51(8):517-524. doi: 10.1007/s00234-009-0529-4
6. Lees KA, Tombers NM, Link MJ, Driscoll CL, Neff BA, Van Gompel JJ, Lane JI, Lohse CM, Carlson ML. Natural History of Sporadic Vestibular Schwannoma: A Volumetric Study of Tumor Growth. Otolaryngology–Head and Neck Surgery 2018;159(3):535-542. doi: 10.1177/0194599818770413
7. Shapey J, Wang G, Dorent R, Dimitriadis A, Li W, Paddick I, Kitchen N, Bisdas S, Saeed SR, Ourselin S, Bradford R, Vercauteren T. 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 2019:1-9. doi: 10.3171/2019.9.JNS191949
8. Lee C-C, Lee W-K, Wu C-C, Lu C-F, Yang H-C, Chen Y-W, Chung W-Y, Hu Y-S, Wu H-M, Wu Y-T, Guo W-Y. Applying artificial intelligence to longitudinal imaging analysis of vestibular schwannoma following radiosurgery. Sci Rep 2021;11(1). doi: 10.1038/s41598-021-82665-8
9. George-Jones NA, Wang K, Wang J, Hunter JB. Automated Detection of Vestibular Schwannoma Growth Using a Two-Dimensional U-Net Convolutional Neural Network. The Laryngoscope 2021;131(2). doi: 10.1002/lary.28695
10. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 2021;18(2):203-211. doi: 10.1038/s41592-020-01008-z
11. Isensee F, Petersen J, Klein A, Zimmerer D, Paul, Kohl S, Wasserthal J, Gregor, Wirkert S, Klaus. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. arXiv pre-print server 2018. doi: None arxiv:1809.10486
12. Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2010;29(1):196-205. doi: 10.1109/tmi.2009.2035616
13. Shamonin DP, Bron EE, Lelieveldt BP, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front Neuroinform 2013;7:50. doi: 10.3389/fninf.2013.00050
14. Rai R, Holloway LC, Brink C, Field M, Christiansen RL, Sun Y, Barton MB, Liney GP. Multicenter evaluation of MRI-based radiomic features: A phantom study. Med Phys 2020;47(7):3054-3063. doi: 10.1002/mp.14173
15. Shapey J, Kujawa A, Dorent R, Wang G, Bisdas S, Dimitriadis A, Grishchuck D, Paddick I, Kitchen N, Bradford R, Saeed S, Ourselin S, Vercauteren T. Segmentation of Vestibular Schwannoma from

Magnetic Resonance Imaging: An Open Annotated Dataset and Baseline Algorithm [Data set]. The Cancer Imaging Archive2021.

16. 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
17. Buch K, Juliano A, Stankovic KM, Curtin HD, Cunnane MB. Noncontrast vestibular schwannoma surveillance imaging including an MR cisternographic sequence: is there a need for postcontrast imaging? J Neurosurg 2019;131(2):549-554. doi: 10.3171/2018.3.jns1866
18. Tolisano AM, Wick CC, Hunter JB. Comparing Linear and Volumetric Vestibular Schwannoma Measurements Between T1 and T2 Magnetic Resonance Imaging Sequences. Otol Neurotol 2019;40(5S):S67-S71. doi: 10.1097/mao.0000000000002208
19. Pizzini FB, Sarno A, Galazzo IB, Fiorino F, Aragno AMR, Ciceri E, Ghimenton C, Mansueto G. Usefulness of High Resolution T2-Weighted Images in the Evaluation and Surveillance of Vestibular Schwannomas? Is Gadolinium Needed? Otol Neurotol 2020;41(1):e103-e110. doi: 10.1097/mao.0000000000002436
20. Whittle IR, Smith C, Navoo P, Collie D. Meningiomas. The Lancet 2004;363(9420):1535-1543. doi: 10.1016/s0140-6736(04)16153-9