International Journal of Computer Assisted Radiology and Surgery (2020) 15:1445–1455
https://doi.org/10.1007/s11548-020-02222-y

ORIGINAL ARTICLE

Manual segmentation versus semi-automated segmentation for quantifying vestibular schwannoma volume on MRI

Hari McGrath[1,2] · Peichao Li[2] · Reuben Dorent[2] · Robert Bradford[3,4] · Shakeel Saeed[4,5,6] · Sotirios Bisdas[7] · Sebastien Ourselin[2] · Jonathan Shapey[2,4,8] · Tom Vercauteren[2]

Received: 20 January 2020 / Accepted: 20 June 2020 / Published online: 16 July 2020

Abstract

Purpose Management of vestibular schwannoma (VS) is based on tumour size as observed on T1 MRI scans with contrast agent injection. The current clinical practice is to measure the diameter of the tumour in its largest dimension. It has been shown that volumetric measurement is more accurate and more reliable as a measure of VS size. The reference approach to achieve such volumetry is to manually segment the tumour, which is a time intensive task. We suggest that semi-automated segmentation may be a clinically applicable solution to this problem and that it could replace linear measurements as the clinical standard.

Methods Using high-quality software available for academic purposes, we ran a comparative study of manual versus semi-automated segmentation of VS on MRI with 5 clinicians and scientists. We gathered both quantitative and qualitative data to compare the two approaches; including segmentation time, segmentation effort and segmentation accuracy.

Results We found that the selected semi-automated segmentation approach is significantly faster (167 s vs 479 s, $p < 0.001$), less temporally and physically demanding and has approximately equal performance when compared with manual segmentation, with some improvements in accuracy. There were some limitations, including algorithmic unpredictability and error, which produced more frustration and increased mental effort in comparison with manual segmentation.

Conclusion We suggest that semi-automated segmentation could be applied clinically for volumetric measurement of VS on MRI. In future, the generic software could be refined for use specifically for VS segmentation, thereby improving accuracy.

Keywords Segmentation · Vestibular schwannoma · Neuroimaging · Machine learning · Imaging

Introduction

Vestibular schwannoma (VS) is a benign tumour of the vestibulocochlear nerve arising within the cerebellopontine angle, deep inside the cranium. It accounts for approximately 6–8% of all intracranial neoplasms and has a prevalence of around 0.02% of the population [21]. Patients may present with a variety of symptoms including hearing loss, balance problems, vertigo, dizziness and headache among others [29]. Diagnosis is usually made on a Magnetic Resonance Imaging (MRI) scan with intravenous contrast demonstrating a homogeneously enhancing lesion within the internal acoustic canal that may also extend into the intracranial cavity [28]. Grading of tumours is performed according to

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11548-020-02222-y) contains supplementary material, which is available to authorized users.

✉ Hari McGrath
hari.mcgrath@kcl.ac.uk

1 GKT School of Medical Education, King's College London, London, UK

2 School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

3 Queen Square Radiosurgery Centre (Gamma Knife), National Hospital for Neurology and Neurosurgery, London, UK

4 Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, UK

5 The Ear Institute, UCL, London, UK

6 The Royal National Throat Nose and Ear Hospital, London, UK

7 Neuroradiology Department, National Hospital for Neurology and Neurosurgery, London, UK

8 Wellcome/EPSRC Centre for Interventional and Surgical Sciences, UCL, London, UK

radiographic characteristics indicating tumour extent and size and is used to guide treatment [19]. Patients with small or asymptomatic tumours are usually managed conservatively with serial surveillance scans. Small- or medium-sized tumours deemed suitable for treatment can be treated effectively and safely with stereotactic radiosurgery (SRS) [22], but larger tumours are usually managed with surgery.

Measuring the size of a VS on MRI is important in guiding treatment or monitoring growth patterns. There are several methods for measuring tumour size, but the most common technique is to measure diameter at the tumour's widest point [16,31,43]. However, this approach is prone to measurement inaccuracies. Volumetric measurement is a solution to this problem [37]. Volumetric analysis offers a more accurate representation of the tumour [38] and could significantly aid the management of these patients. Segmentation (contouring) is already used in the planning of gamma knife SRS treatment. Segmentation also provides a means of performing volumetric measurement of the tumour. Compared with two-dimensional measurements, it may be used more accurately for the active surveillance of VS. Volumetric measurement has been used to predict recurrence in patients with residual tumours following surgical intervention [35], to measure change in tumour size following SRS treatment [44] and to predict hearing preservation following SRS treatment [11]. There are three main methods of volumetric analysis: manual segmentation, semi-automated segmentation and automated segmentation. Manual segmentation involves comprehensively labelling the 3D structure in each 2D slice. It is a time-intensive task with relatively low inter- and intra-individual reliability and has not been widely employed in clinical practice.

Automated segmentation has been applied successfully to MR imaging for a wide range of brain tumours [46]. Automated segmentation may be accurate in the assessment of tumour progression and in overall survival prediction in glioma [1,26] as well as for the clinical assessment of biomarkers in glioma [4]. For VS imaging, automated segmentation has been applied with positive results [32,40] and there is growing interest in the field [10]. An automated segmentation tool could also improve clinical workflow and operational efficiency during the planning of stereotactic radiosurgery (SRS) by using the tool as an initialisation step in the process. However, automated approaches are, for the most part, not fully validated and are confined to academic use. Furthermore, some tumours display heterogeneous enhancement including the 4% of VS tumours that may be cystic, which can lead to inaccurate segmentation when automated methods are applied [25].

Semi-automated segmentation has been shown to be a more reliable option for the analysis of VS on MRI scans [24]. However, there has been no previous analysis of cognitive load or user experience of VS segmentation. When using semi-automated methods, segmentation time and repeatability may be improved when compared with manual segmentation [2,6,39,41]. Compared with fully automatic segmentation, results may be more accurate [1] and are more acceptable to clinicians due to increased transparency in the segmentation process [12]. Currently proposed methods require user input for one or more of the following steps: segmentation parameters, feedback or evaluation, including refinement and validation of the segmentation. There is little material in the literature regarding user experience of interactive segmentation in brain imaging, despite the intention to pursue clinical translation in the field [18,33].

A number of software packages are academically available for medical image segmentation spanning a variety of different methods. For manual segmentation, ITK-SNAP[1] [45] is a widely used open-source software library with manual, semi-automated and automated segmentation offerings. 3D slicer[2] has the standard offerings of image viewing and analysis tools, along with a variety of downloadable packages for semi-automated and automated segmentation [8]. MRIcron[3] is a package of image viewing and manual segmentation tools. For semi-automated segmentation, ImFusion Labels (ImFusion, Munich, Germany) is a recent commercial-grade package with academic licensing options.

We present the findings of a proof of concept study using combined quantitative and qualitative analysis, comparing manual segmentation with semi-automated segmentation of VS on MRI. We hypothesise that semi-automated segmentation is faster than manual segmentation with a comparable performance. In this study, we also compare the user experience of two software suites, including that of clinicians and senior researchers.

Materials and method

We selected four tumours from our database for the study (see Table 1). All four patients had previously undergone Gamma Knife SRS treatment [3]. The images were representative of a variety of tumour sizes and shapes encountered in clinical practice. We selected two small and two moderate-sized tumours (see Table 1). The ground truth measurements were made prior to the study by the treating skull base neurosurgeon and stereotactic radiosurgery physicist using Gamma Knife planning software (Leksell GammaPlan, Elekta, Sweden). The images used in this study were all contrast-enhanced T1-weighted scans with 0.4 mm × 0.4 mm in-plane resolution, in-plane matrix of 512 × 512 and 1.5 mm slice thickness. All cases included

[1] http://www.itksnap.org.

[2] http://www.slicer.org.

[3] https://people.cas.sc.edu/rorden/mricron/.

Table 1 Tumour characteristics according to commonly used criteria for representing tumour size and extent

Tumour identifier	Volume (mm^3)	Largest diameter (mm)
VS_1	623	15.1
VS_2	1050	20.5
VS_3	3590	25
VS_4	975	17

an extracanalicular (intracranial) component, and none of the tumours had a cystic component. Patients with multiple tumours were excluded.

We selected ITK-SNAP for manual segmentation since this offered the most intuitive user interface. In our group, it was also the most widely used library for manual segmentation. We selected ImFusion Labels for semi-automated segmentation since this was a recent software with a good selection of machine learning tools and a high-quality user interface. It was made available to our group through an academic license.

Five observers, including two medical students, two biomedical engineers and one neurosurgeon, performed manual and semi-automated segmentation on each of the four scans. The participants had a variety of experience with segmentation. Three participants were inexperienced segmenters (with no or limited previous experience), and two were experts in medical image segmentation, with multiple years experience of medical image segmentation. Three had previous experience using ITK-SNAP, one of whom had limited experience of using ImFusion Labels.

Study design

A training period was included for each study participant at the start of the study and for each software library, using a training data set which was not part of the study. This training period was standardised to 10 min for each participant and included an initial demonstration from the study lead followed by a trial run for each participant. During the training period, participants were free to ask questions relating to the segmentation. The trial runs were not included in the results or the analysis. Participants were advised on the optimal tools to use in each software library. This training period was adapted based on the needs and previous experience of the participant, such that no demonstration was given for those participants well-versed in the use of the software library.

In ITK-SNAP, participants used the polygon drawing tool to outline tumour boundaries in each slice and fill in the tumour volume (see Fig. 1). The paintbrush tool was used to make small alterations as needed. A time limit of ten minutes per segmentation was provided in order to standardise the process according to arbitrary mock-clinical parameters.

In ImFusion Labels, participants used the 'Interactive Segmentation' module (see Fig. 2). They were advised to first draw background labels which included structures of a variety of intensities (e.g. bone, dura, healthy brain). After the first iteration of the segmentation, participants were advised to only undertake two alterations in the segmentation. This was determined to produce optimum results while creating an incentive to complete the task in a time-pressured manner.

A document containing participant instructions is included as Online Resource 1. A video depicting segmentation in ITK-SNAP is included as Online Resource 2. A video depicting segmentation in ImFusion Labels is provided as Online Resource 3.

Qualitative data collection

The NASA Task Load Index (TLX) [14] questionnaire was performed at the end of the study to quantify user effort for each method of segmentation. The TLX scores different aspects of a task on a graded scale from 1–21, including effort, frustration and performance. It can be found as "Table 2 in the Appendix". The TLX was used as a relative comparator of the libraries, rather than as an absolute scale. For data analysis, we processed the raw TLX data. This may be a more reliable use of the TLX compared with using part two to calculate an overall weight-adjusted score [5].

We performed short post-segmentation interviews to explore the participants' experiences of the different toolboxes. The questions were based around themes, which included 'segmentation experience', 'toolbox' and 'study design'. "Table 3 in the Appendix" details the questions asked of each participant. Participants were asked about each software library separately. Data were collected in shorthand form by the study lead during the interview and then expanded following the interview.

Quantitative data collection and analysis

The time taken to perform the segmentation was measured from the time of launching the software to the time of closing the software following the segmentation. A paired *t*-test was performed on this data to calculate the *p*-value as well as the confidence intervals. We quantified segmentation accuracy by comparing the segmentations in each software with the ground truth data in order to establish a comparative analysis. We calculated the Dice coefficient (Dice) since this is a standard comparative measure of radiological data [26,27]. We also calculated relative volume error (RVE) and average symmetric surface distance (ASSD) for each segmentation. We performed subgroup analysis on both the time and accuracy data. We took the two more experienced segmenters and

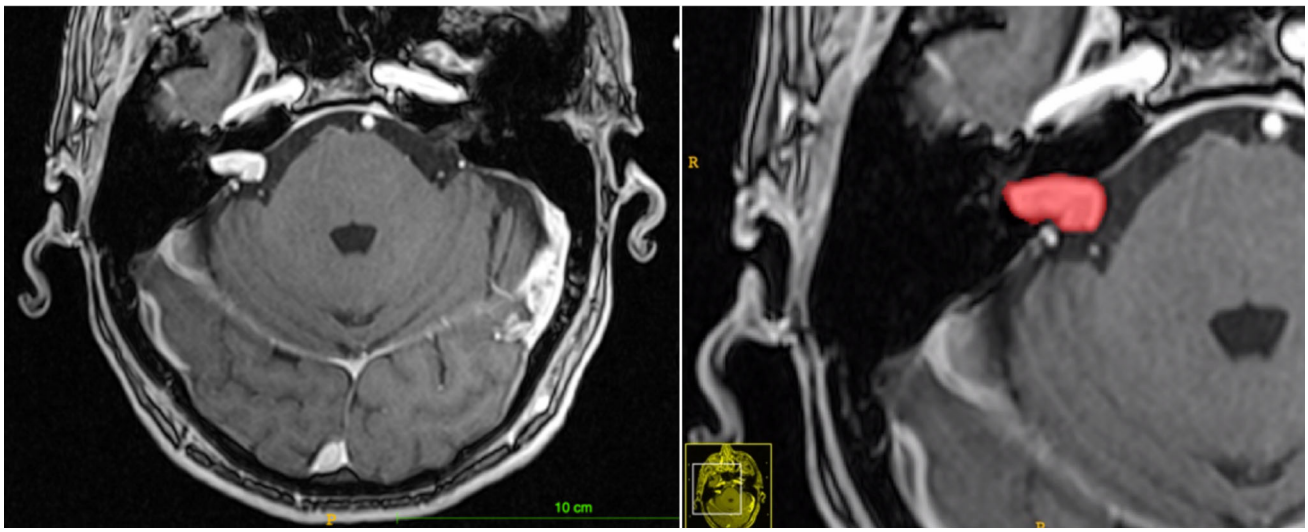

Fig. 1 Example of a tumour, pre- and post-segmentation, represented in ITK-SNAP. This was tumour 'VS_1', classed as a small tumour with limited extracanalicular extension

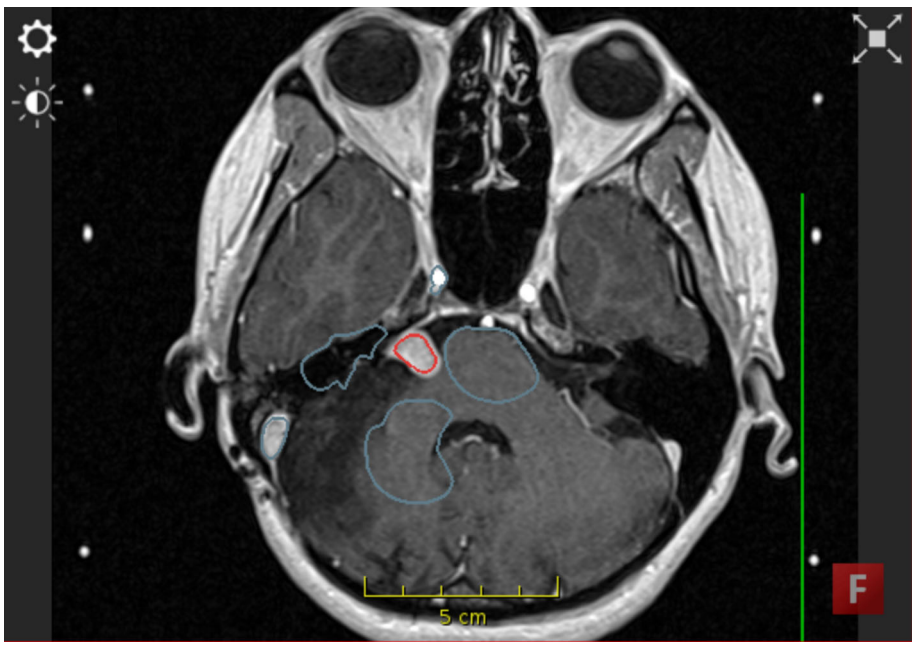

Fig. 2 Segmentation in ImFusion Labels using background labels (blue) and foreground labels (red) to demarcate tumour and non-tumour tissue

compared results from these individuals against the three less experienced segmenters.

Results

Segmentation time was significantly faster in ImFusion Labels. In terms of TLX data, ITK-SNAP was more time demanding and physically demanding, whereas ImFusion was more mentally demanding and frustrating. The performance, in terms of accuracy, and overall effort of the libraries were comparable. Qualitatively, participants preferred the control that ITK-SNAP offered; however, some did not like the time demand. ImFusion was a good tool for rapidly estimating tumour volume, but there were frustrating errors produced in complex tumour segmentation.

Time

Between the two libraries, segmentation in ImFusion Labels was significantly faster than ITK-SNAP. The mean segmentation time (ST) in ITK-SNAP was 479 s (95% CI 439–519), while the mean ST in ImFusion Labels was 168 s (95% CI 168–249), with a p value of < 0.001 (see Fig. 3a). There

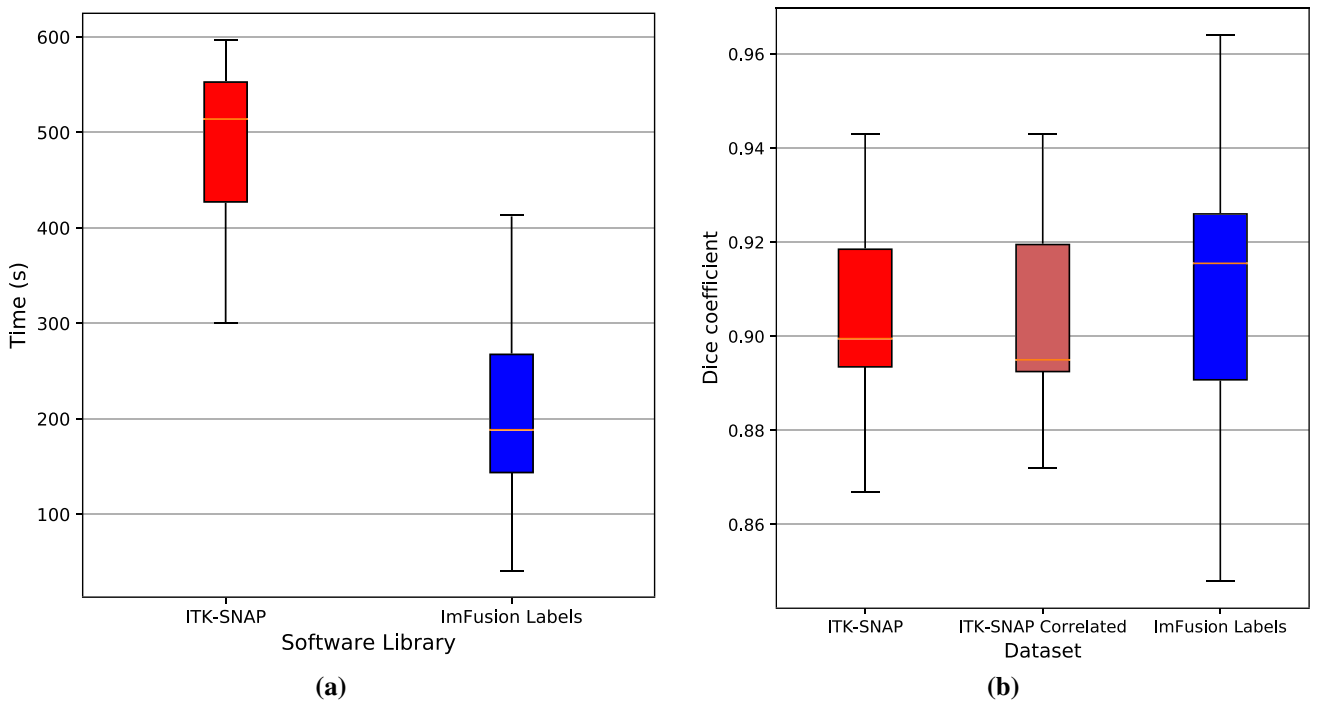

Fig. 3 **a** Comparison of segmentation time between the two software libraries; **b** spread of Dice scores in ITK-SNAP as compared to ImFusion Labels. The "ITK-SNAP Correlated" plot only takes into account the data which corresponds to the one from ImFusion labels that we still had access to (after data loss had occurred)

was no observed difference in segmentation time between the less experienced individuals and the more experienced individuals.

Accuracy

The user-generated segmentation dataset was compromised during the study, resulting in half of the ImFusion data being unavailable for analysis of segmentation accuracy. On the remaining data "Table 4 in the Appendix", we observed comparable accuracy between the two libraries, with a Dice score range of 0.848–0.964 for ImFusion compared with a range of 0.867–0.943 for ITK-SNAP. Compared with segmentations in ITK-SNAP, segmentations in ImFusion Labels were more similar to the ground truth data in terms of Dice (0.913 vs 0.902, $p = 0.301$), RVE (0.0723 vs 0.124, $p = 0.245$) and ASSD (0.381 vs 0.419, $p = 0.349$) as illustrated in Fig. 3b. In our subgroup analysis, the two cohorts achieved similar levels of accuracy for manual segmentation in ITK-SNAP. The experienced cohort achieved more accurate Dice scores (0.901 vs 0.899, $p = 0.533$), and RVD scores (0.155 vs 0.104, $p = 0.312$), while the inexperienced cohort achieved more accurate ASSD scores (0.417 vs 0.420, $p = 0.936$) when compared with ground truth data. However, none of these differences were statistically significant.

NASA TLX score

The TLX scores showed a trend towards ITK-SNAP being the more physically and temporally demanding approach (+6 and +3.4-point scores on average, respectively), while ImFusion tended to be more mentally demanding and worse in terms of perceived performance (−7.8 and −2.4 points on average, respectively). All participants graded ImFusion as being more frustrating, with a +7.4-point greater score on average. All participants also graded ImFusion as being more mentally demanding, with a +7.8 greater score on average. ITK-SNAP was graded as being more physically demanding by all but one participant. Less experienced raters tended to score the segmentation performance of ImFusion higher than more experienced raters. Overall effort was slightly greater (+2.4 points on average) in ImFusion (Fig. 4).

Interview data

ITK-SNAP was the preferred choice for highly accurate segmentation, while one participant recommended ImFusion as a 'rough volumetric estimate'. All participants cited the improved performance of the ImFusion algorithm with 'simple' tumours, i.e. those which were highly contrast-enhancing, homogeneous with well-defined boundaries and no or minimal adjacent high contrast structures, such as blood

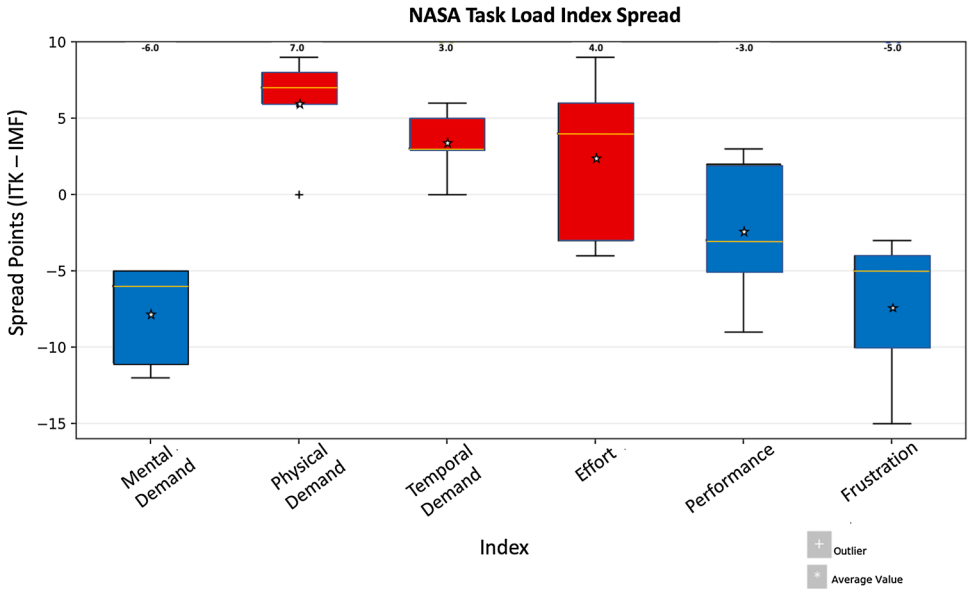

Fig. 4 Relative NASA TLX spread data. The ImFusion score was subtracted from the ITK-SNAP score for each participant and combined for each index to show spread of data across participants. Positive values represent a greater score for ITK-SNAP, while negative values were greater for ImFusion. The scores at the top indicate the median value, while the colours represent the software which the mean value favoured. Blue indicates a mean which favoured ImFusion labels, while red indicates a mean value which favoured ITK-SNAP

vessels or dura. However, for complex tumours the algorithm often made small, but frustrating, errors in segmentation—"[the algorithm] threw up errors which required a complete restart". Occasionally, non-tumour areas were included, and tumour areas were not included. There was generally no way to fix this using the tool. One participant complained that in these more challenging cases, the algorithm was "a one-trick pony…if you make alterations to the initial segmentation you may worsen it". Participants commented on the 'unpredictability' of the algorithm and the lack of transparency as being a significant problem in solving these issues. In ITK-SNAP, the majority of participants cited the need to compromise between thoroughness and timing of segmentation. One stated "I am a perfectionist…if we were not timed, [the segmentation] would take me much longer". In terms of study design, participants found the instructions clear and found it "helpful to have someone here to explain and provide feedback [during the training period]". A full breakdown of the qualitative data taken from interviews is provided in the appendix (see "Table 5 in the Appendix").

Discussion

In this paper, we sought to compare manual segmentation to semi-automated segmentation on several variables, both quantitative and qualitative, for segmentation of VS. It is widely published that semi-automated segmentation may reduce the time taken to perform segmentation [9,23,30]. We showed that semi-automated segmentation is significantly faster and has comparable performance when compared with manual segmentation for volumetric analysis of VS. This would suggest good viability for this approach in clinical practice, where time constraints may restrict which methods are used. However, this study does have some limitations.

In terms of performance, both semi-automated and manual segmentation were highly accurate when compared with ground truth data and there was no statistically significant difference between the two methods. In terms of clinical applicability, any differences between the two may also be clinically insignificant, thereby making semi-automated segmentation a desirable option. The involvement of inexperienced segmenters may reduce the validity of the conclusions we can draw. However, we observed a high degree of similarity in accuracy data for the experienced segmenters when compared with the inexperienced segmenters, suggesting that there was no compromise on data quality due to the inclusion of less-experienced participants.

In interview, some participants suggested that the segmentation in ImFusion produced significant errors in complex tumours. The Dice scores, however, indicated a high degree of accuracy in these segmentations. One explanation for this inconsistency in perception versus result may be attributable to a finer margin for error applied to the analysis of segmentations in ImFusion. Participants spent, on average, 479 s on each segmentation in ITK-SNAP, compared with 168 s in ImFusion. This time discrepancy may have led to a higher acceptance threshold for the segmentation in ImFusion, and small mistakes may have been picked up more readily.

In terms of effort measures, the NASA TLX was a useful tool. However, one limitation is that the system was used as a relative measure of effort between the different software libraries used for the study. Therefore, the absolute values offered by participants may not be an accurate measure of absolute effort and would therefore provide unreliable data for inter-rater comparison. We compared the inter-rater

scores by subtracting the ImFusion scores from the ITK data for each participant. We would therefore suggest the use of the full TLX as opposed to the Raw TLX to overcome these issues.

We chose to state the segmentation goal as what would be clinically, or personally, acceptable to the participants. In this way, we felt that participants would apply the same requirements to both libraries. In some cases, the opposite was true. A very thorough approach was employed by some participants in ITK-SNAP, but in ImFusion Labels they used a crude approach. This difference in perceived goals may have introduced bias in the time and effort of segmentation. This challenge could be avoided in future by clearly stating the goals of the segmentation, whether targeting accuracy or speed.

One constraint on semi-automated segmentation lies in usability of the tools. In this study, a common point of feedback was that the algorithm was inconsistent and unpredictable in its segmentation. Some users found this tedious and had to restart when the algorithm produced errors. In the literature, a commonly cited limitation in clinical application is algorithmic transparency [17]. Users did not understand what the algorithm did and why. ImFusion Labels is a generic library and has wide applicability in medical imaging. A solution to this issue may be to refine an algorithm specifically for VS segmentation.

There is very little qualitative data in the literature on the use of segmentation tools. Qualitative data are particularly important given the current interest in clinical translation of AI tools, which must be robust, easy to use and accurate [17]. As far as we can see, this is the first paper to use a mixed quantitative and qualitative format to compare semi-automated segmentation with manual segmentation in medical imaging. The small sample size of this study, in terms of participants and scans segmented, may limit the validity of the conclusions we can draw. One further challenge was in data representation for qualitative analysis, since none of the research team had previous experience of handling interview data. It may be useful to recruit this expertise in future studies.

In terms of applicability to the current clinical workflow, semi-automated segmentation may assist in monitoring VS growth, especially in those patients with small tumours being managed conservatively with serial imaging [13,15,31]. It has been established that volumetric measurement is superior to single-dimension diametric measurements for quantifying growth [24,36]. Manual segmentation is not feasible in routine clinical practice due to the time-demanding nature of the task. Thresholding is an additional tool that may help an experienced user to segment VS and could make for an interesting comparator in future work. When compared with manual segmentation, we showed that semi-automated segmentation is less time-demanding, less physically demanding and of comparable performance.

In the future, it is hoped that further algorithmic developments could support the practice of radiology among other specialities [34]. Deep learning is a sub-type of artificial intelligence that utilises multiple layers of analysis to process an image. A variety of applications of deep learning are postulated [7,20,42], and one study has shown this to be a useful approach in automated VS segmentation [32] in terms of both time and accuracy. Despite the accuracy of automated approaches, interactive corrections may continue to play a role even with deep learning due to the lack of adaptability of automated methods to the specific imaging sequences and protocols used clinically [39]. The next steps are to further analyse this methodology and work towards clinical translation.

The findings of this study may also be applied more widely to semi-automated segmentation of other neuroimaging data. Some participants felt that manual segmentation could not be matched in terms of performance if plenty of time was spent. The participants did not have specific expertise in the diagnosis or management of VS, aside from the neurosurgeon. We would expect that similar results, in terms of qualitative findings, may be present in other applications; for instance tumour segmentation for glioma. We would recommend that semi-automated segmentation is used as a supportive measure to other standard approaches in neuroimaging segmentation.

Conclusion

Gains are being made in the machine learning and medical imaging fields. Machine learning applications are now performing comparably with their manual counterparts. However, a finding of this study was that even the state-of-the-art machine learning tools may not yet be fully ready for clinical roll out in segmentation of vestibular schwannoma. Users found the tools to be fast and accurate, but at times unpredictable and frustrating to use. There were limitations in the study, including the small sample size in terms of participants, particularly those with experience in segmentation, and in the number of scans segmented. This makes conclusions difficult to draw. The strengths of this study lie in the joint use of both qualitative and quantitative methods, which were employed to address the clinical applicability of algorithms. Unpredictability of algorithm behaviour and lack of transparency with algorithmic methods are cited as being key issues. To remedy this, developers should focus on involving groups with a variety of backgrounds and expertise in the development process, to ensure clinical and research applicability.

Acknowledgements This work was supported by Wellcome [203145Z-/16/Z, 203148/Z/16/Z, WT106882] and EPSRC [NS/A000050/1, NS-/A000049/1] funding. TV is supported by a Medtronic/Royal Academy of Engineering Research Chair [RCSRF1819\7\34].

Compliance with ethical standards

Conflicts of interest An academic license was provided by ImFusion for the use of ImFusion Labels. Besides this, the authors have no conflicts of interest to declare.

Human and animal rights There were no human or animal studies conducted in this work.

Informed consent There was no informed consent or IRB study required for the work reported in this manuscript.

Appendix

See Tables 2, 3, 4 and 5.

Table 2 NASA Task Load Index. Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales

How mentally demanding was the task?
How physically demanding was the task?
How hurried or rushed was the pace of the task?
How successful were you in accomplishing what you were asked to do?
How hard did you have to work to accomplish your level of performance?
How insecure, discouraged, irritated, stressed, and annoyed were you?

Table 3 Interview questions for qualitative comparison of the two software libraries

Was the segmentation in each software to your satisfaction?
Overall, how did you find each software?
What would you add or remove from each software to improve them?
How did you find the study?

Table 4 Mean segmentation accuracy values for each scan in ITK-SNAP and ImFusion

Tumour identifier	ITK-SNAP			ImFusion		
	Dice	RVE	ASSD	Dice	RVE	ASSD
VS_1	0.882	0.094	0.457	0.885	0.114	0.424
VS_2	0.893	0.110	0.398	0.890	0.043	0.422
VS_3	0.929	0.115	0.441	0.945	0.085	0.357
VS_4	0.903	0.178	0.379	0.925	0.056	0.311

Table 5 Interview answers grouped by theme

Theme	Software	Quotes	Prevalence
Performance discrepancy across tumours	ImFusion	'Very good for clear-cut, simple tumours... [those which were] highly contrast enhancing, homogeneous, with well-defined boundaries and minimal adjacent blood vessels.'	All five participants (100%)
		'Complex tumours threw up errors which required a complete restart.'	
Compromise between thoroughness and timing	ITK-SNAP	'I am a perfectionist... if we weren't timed it would take me much longer.'	Four out of five (80%)
		'I made lots of small mistakes... but it would have taken too long to correct.'	
		'It was very fiddly.'	
Unpredictable outcome after drawing labels	ImFusion	'a one-trick pony…if you make alterations to the initial segmentation you may worsen it.'	Three out of five (60%)
		'if we wanted perfection…we would have to go back again and again.'	
		'I do not know if the changes I make will improve of worsen the segmentation.'	
Speed of segmentation	ImFusion	'Much faster so it would be great for my work.'	Four out of five (80%)
		'The algorithm works very quickly.'	
UI and tools	Both	'[Using ImFusion] was a much nicer experience…and a sleek UI.'	-
		'[ImFusion] is better for visualization.'	
		'[In ImFusion] I would like to have a paintbrush tool which draws and erases exactly what I want it to…there is too much prediction required…scribbles I make should not affect the whole segmentation.'	
Study design	–	'It was helpful to have someone here to explain and provide feedback.'	-
		'Would have been good to define the goal more clearly…do we want a very accurate segmentation or a rough volume estimate.'	
		'You could have gone through all the tools I might need during the training phase.'	

References

1. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, Shinohara RT, Berger C, Ha SM, Rozycki M, Prastawa M, Alberts E, Lipkova J, Freymann J, Kirby J, Bilello M, Fathallah-Shaykh H, Wiest R, Kirschke J, Wiestler B, et al (2018) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv:1811.02629
2. Birkbeck N, Cobzas D, Jagers M, Murtha A, Kesztyues T (2009) An interactive graph cut method for brain tumor segmentation. In: 2009 Workshop on applications of computer vision (WACV), pp 1–7. IEEE
3. Boari N, Bailo M, Gagliardi F, Franzin A, Gemma M, del Vecchio A, Bolognesi A, Picozzi P, Mortini P (2014) Gamma knife radiosurgery for vestibular schwannoma: clinical results at long-term follow-up in a series of 379 patients. J Neurosurg 121(Suppl-2):123–142
4. Booth TC, Williams M, Luis A, Cardoso J, Ashkan K, Shuaib H (2020) Machine learning and glioma imaging biomarkers. Clin Radiol 75(1):20–32. https://doi.org/10.1016/j.crad.2019.07.001
5. Bustamante EA, Spain RD (2008) Measurement invariance of the NASA TLX. Proc Human Factors Ergon Soc Annu Meet 52(19):1522–1526
6. Chae SY, Suh S, Ryoo I, Park A, Noh KJ, Shim H, Seol HY (2017) A semi-automated volumetric software for segmentation and perfusion parameter quantification of brain tumors using 320-row multidetector computed tomography: a validation study. J Neuroradiol 59(5):461–469
7. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, Mahajan V, Rao P, Warier P (2018) Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 392(10162):2388–2396. https://doi.org/10.1016/S0140-6736(18)31645-3

8. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J, Aylward S, Miller VJ, Pieper S, Kikinis R (2012) 3D slicer as an image computing platform for the quantitative imaging network. Magn Reson Imaging 30(9):1323–1341
9. Fernquest S, Park D, Marcan M, Palmer A, Voiculescu I, Glyn-Jones S (2018) Segmentation of hip cartilage in compositional magnetic resonance imaging: a fast, accurate, reproducible, and clinically viable semi-automated methodology. J Orthop Res 36(8):2280–2287
10. George-Jones N, Wang K, Wang J, Hunter JB (2020) An automated method for determining vestibular schwannoma size and growth. Journal of Neurological Surgery, Part B Skull Base. In: Conference: 30th annual meeting North American skull base society. United States 81(Supplement 1)
11. Gjuric M, Mitrecic MZ, Greess H, Berg M (2007) Vestibular schwannoma volume as a predictor of hearing outcome after surgery. Otol Neurotol 28(6):822–827. https://doi.org/10.1097/MAO.0b013e318068b2b0
12. Gordillo N, Montseny E, Sobrevilla P (2013) State of the art survey on MRI brain tumor segmentation. Magn Reson Imaging 31(8):1426–1438
13. Halliday J, Rutherford SA, McCabe MG, Evans DG (2018) An update on the diagnosis and treatment of vestibular schwannoma. Expert Rev Neurother 18(1):29–39
14. Hart SG, Staveland LE (1988) Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock PA, Meshkati N (eds) Human mental workload, advances in psychology, vol 52. Elsevier, North-Holland, pp 139–183
15. Hughes M, Skilbeck C, Saeed S, Bradford R (2011) Expectant management of vestibular schwannoma: a retrospective multivariate analysis of tumor growth and outcome. Skull Base 21(05):295–302
16. Kanzaki J, Tos M, Sanna M, Moffat DA (2003) New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma. Otol Neurotol 24(4):642–649
17. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17(1):195
18. Khademi A, Reiche B, DiGregorio J, Arezza G, Moody AR (2019) Whole volume brain extraction for multi-centre, multi-disease FLAIR MRI datasets. Magn Reson Imaging. 66:116–130
19. Koos WT, Day JD, Matula C, Levy DI (1998) Neurotopographic considerations in the microsurgical treatment of small acoustic neurinomas. J Neurosurg 88(3):506–512
20. Lieman-Sifry J, Le M, Lau F, Sall S, Golden D (2017) FastVentricle: cardiac segmentation with ENet. In: Pop M, Wright G (eds) Functional imaging and modelling of the heart. FIMH 2017, vol 10263. Lecture notes in computer science. Springer, Cham
21. Lin D, Hegarty JL, Fischbein NJ, Jackler RK (2005) The prevalence of "incidental" acoustic neuroma. Arch Otolaryngol 131(3):241–244
22. Lunsford LD, Niranjan A, Flickinger JC, Maitz A, Kondziolka D (2005) Radiosurgery of vestibular schwannomas: summary of experience in 829 cases. J Neurosurg 102(Special–Supplement):195–199
23. Ma C, Zhang L, Wei C, Wang D, Wang D (2010) Repeatability and accuracy of quantitative knee cartilage volume measurement using semi-automated software at 3.0T MR. Chin J Med Imaging Technol 26(4):760–763
24. MacKeith S, Das T, Graves M, Patterson A, Donnelly N, Mannion R, Axon P, Tysome J (2018) A comparison of semi-automated volumetric versus linear measurement of small vestibular schwannomas. Eur Arch Oto Rhino L 275(4):867–874
25. Mazzara GP, Velthuizen RP, Pearlman JL, Greenberg HM, Wagner H (2004) Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation. Int J Radiat Oncol 59(1):300–312
26. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L, Gerstner E, Weber M, Arbel T, Avants BB, Ayache N, Buendia P, Collins DL, Cordier N, Corso JJ et al (2015) The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans Med Imaging 34(10):1993–2024. https://doi.org/10.1109/TMI.2014.2377694
27. Milletari F, Navab N, Ahmadi S (2016) V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth international conference on 3D vision (3DV), pp 565–571. https://doi.org/10.1109/3DV.2016.79
28. Nguyen D, de Kanztow L (2019) Vestibular schwannomas: a review. Appl Radiol 48(3):22–27
29. Pinna MH, Bento RF, de Brito Neto RV (2012) Vestibular schwannoma: 825 cases from a 25-year experience. Int Arch Otorhinolaryngol 16(04):466–475
30. Seuss H, Janka R, Prümmer M, Cavallaro A, Hammon R, Theis R, Sandmair M, Amann K, Bäuerle T, Uder M, Hammon M (2017) Development and evaluation of a semi-automated segmentation tool and a modified ellipsoid formula for volumetric analysis of the kidney in non-contrast T2-weighted MR images. J Digit Imaging 30(2):244–254
31. Shapey J, Barkas K, Connor S, Hitchings A, Cheetham H, Thomson S, U-King-Im J, Beaney R, Jiang D, Barazi S, Obholzer R, Thomas N (2018) A standardised pathway for the surveillance of stable vestibular schwannoma. Ann R Coll Surg Engl 100(3):216–220
32. Shapey J, Wang G, Dorent R, Dimitriadis A, Li W, Paddick I, Kitchen N, Bisdas S, Saeed SR, Ourselin S, Bradford R, Vercauteren T (2019) 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 1(aop):1–9
33. Shaver MM, Kohanteb PA, Chiou C, Bardis MD, Chantaduly C, Bota D, Filippi CG, Weinberg B, Grinband J, Chow DS, Chang P (2019) Optimizing neuro-oncology imaging: a review of deep learning approaches for glioma imaging. Cancers 11(6):829
34. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7
35. Vakilian S, Souhami L, Melançon D, Zeitouni A (2012) Volumetric measurement of vestibular schwannoma tumour growth following partial resection: predictors for recurrence. J Neurol Surg B 73(02):117–120. https://doi.org/10.1055/s-0032-1301395
36. Varughese JK, Breivik CN, Wentzel-Larsen T, Lund-Johansen M (2012) Growth of untreated vestibular schwannoma: a prospective study. J Neurosurg 116(4):706–712
37. van de Langenberg R, de Bondt BJ, Nelemans PJ, Baumert BG, Stokroos RJ (2009) Follow-up assessment of vestibular schwannomas: volume quantification versus two-dimensional measurements. J Neuroradiol 51(8):517
38. Lees KA, Tombers NM, Link MJ, Driscoll CL, Neff BA, Van Gompel JJ, Lane JI, Lohse CM, Carlson ML (2018) Natural history of sporadic vestibular schwannoma: a volumetric study of tumor growth. Otolaryngol Head Neck Surg 159(3):535–542
39. Wang G, Li W, Zuluaga MA, Pratt R, Patel PA, Aertsen M, Doel T, David AL, Deprest J, Ourselin S, Vercauteren T (2018) Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE T Med Imaging 37:1562–1573
40. Wang G, Shapey J, Li W, Dorent R, Demitriadis A, Bisdas S, Paddick I, Bradford R, Zhang S, Ourselin S, Vercauteren T (2019) Automatic segmentation of vestibular schwannoma from T2-weighted MRI by deep spatial attention with hardness-

weighted loss. Med Image Comput Comput Assist Interv MICCAI 2019:264–272
41. Wang G, Zuluaga MA, Li W, Pratt R, Patel PA, Aertsen M, Doel T, David AL, Deprest J, Ourselin S, Vercauteren T (2019) Deepigeos: a deep interactive geodesic framework for medical image segmentation. IEEE T Pattern Anal 41:1559–1572
42. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 3462–3471. https://doi.org/10.1109/CVPR.2017.369
43. Yoshimoto Y (2005) Systematic review of the natural history of vestibular schwannoma. J Neurosurg 103(1):59–63
44. Yu CP, Cheung JYC, Leung S, Ho R (2000) Sequential volume mapping for confirmation of negative growth in vestibular schwannomas treated by gamma knife radiosurgery. J Neurosurg 93(supplement_3):82 https://doi.org/10.3171/jns.2000.93.supplement_3.0082
45. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G (2006) User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31(3):1116–1128
46. Zou KH, Wells WM, Kikinis R, Warfield SK (2004) Three validation metrics for automated probabilistic image segmentation of brain tumours. Stat Med 23(8):1259–82. https://doi.org/10.1002/sim.1723