Prediction of transient tumor enlargement using MRI tumor texture after radiosurgery on vestibular schwannoma

Patrick P. J. H. Langenhuizen[a)]
Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
Gamma Knife Center Tilburg, Department of Neurosurgery, ETZ Hospital, Tilburg, The Netherlands

Sander H. P. Sebregts, and Svetlana Zinger
Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

Sieger Leenstra
Department of Neurosurgery, Erasmus Medical Center, Rotterdam, The Netherlands

Jeroen B. Verheul
Gamma Knife Center Tilburg, Department of Neurosurgery, ETZ Hospital, Tilburg, The Netherlands

Peter H. N. de With
Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

(Received 9 July 2019; revised 13 January 2020; accepted for publication 16 January 2020; published 18 February 2020)

Purpose: Vestibular schwannomas (VSs) are uncommon benign brain tumors, generally treated using Gamma Knife radiosurgery (GKRS). However, due to the possible adverse effect of transient tumor enlargement (TTE), large VS tumors are often surgically removed instead of treated radiosurgically. Since microsurgery is highly invasive and results in a significant increased risk of complications, GKRS is generally preferred. Therefore, prediction of TTE for large VS tumors can improve overall VS treatment and enable physicians to select the most optimal treatment strategy on an individual basis. Currently, there are no clinical factors known to be predictive for TTE. In this research, we aim at predicting TTE following GKRS using texture features extracted from MRI scans.

Methods: We analyzed clinical data of patients with VSs treated at our Gamma Knife center. The data was collected prospectively and included patient- and treatment-related characteristics and MRI scans obtained at day of treatment and at follow-up visits, 6, 12, 24 and 36 months after treatment. The correlations of the patient- and treatment-related characteristics to TTE were investigated using statistical tests. From the treatment scans, we extracted the following MRI image features: first-order statistics, Minkowski functionals (MFs), and three-dimensional gray-level co-occurrence matrices (GLCMs). These features were applied in a machine learning environment for classification of TTE, using support vector machines.

Results: In a clinical data set, containing 61 patients presenting obvious non-TTE and 38 patients presenting obvious TTE, we determined that patient- and treatment-related characteristics do not show any correlation to TTE. Furthermore, first-order statistical MRI features and MFs did not significantly show prognostic values using support vector machine classification. However, utilizing a set of 4 GLCM features, we achieved a sensitivity of 0.82 and a specificity of 0.69, showing their prognostic value of TTE. Moreover, these results increased for larger tumor volumes obtaining a sensitivity of 0.77 and a specificity of 0.89 for tumors larger than 6 cm^3.

Conclusions: The results found in this research clearly show that MRI tumor texture provides information that can be employed for predicting TTE. This can form a basis for individual VS treatment selection, further improving overall treatment results. Particularly in patients with large VSs, where the phenomenon of TTE is most relevant and our predictive model performs best, these findings can be implemented in a clinical workflow such that for each patient, the most optimal treatment strategy can be determined. © *2020 The Authors. Medical Physics published by Wiley Periodicals, Inc. on behalf of American Association of Physicists in Medicine.* [https://doi.org/10.1002/mp.14042]

Key words: Gamma Knife radiosurgery, MRI tumor texture, pseudoprogression, transient tumor enlargement, vestibular schwannomas

1. INTRODUCTION

Vestibular schwannomas (VSs) are relatively rare benign brain tumors, originating from the Schwann cells of the eighth cranial nerve. These tumors make up 8% of the primary brain tumors diagnosed in the United States[1] and have an incidence of 37.5 per million inhabitants in the Netherlands.[2] In the last few decades, the main treatment goal for VS has shifted from complete removal of the tumor to functional preservation of the facial, trigeminal, and cochlear nerves.[3] Especially the introduction of less invasive treatment options and their reduced risks at post-treatment morbidities has led to this substantial shift.[4] In a systematic review by Wolbers et al., the authors[5] determined that for small- to medium-sized VSs, radiosurgery is nowadays generally preferred over microsurgery. The reason for this preference is the highly invasive nature of microsurgery, which results in: (a) a higher risk at mortality, (b) an inferior preservation of the facial nerve function, (c) a decreased hearing preservation and (d) a lower quality of life. Moreover, microsurgical treatments invoke a fourfold larger overall cost on average, compared to radiosurgery.[6]

However, for large VSs the discussion concerning the best treatment strategy is still ongoing. Most medical centers consider microsurgical resection as the optimal treatment strategy for these large tumors, as it effectively averts the compression of surrounding critical brain structures, such as the brainstem, the cerebellum and the previously mentioned cranial nerves. Since the risks involved in microsurgery can be contra-indicative for this strategy, less invasive treatments such as radiosurgery and radiotherapy have been considered increasingly in the last decade. These strategies have shown good results for large VSs and obtained acceptable radiation-induced morbidities.[7–14]

Nevertheless, radiosurgical treatments of large VSs remain controversial due to the possible transient tumor enlargement (TTE). This radiation-induced swelling of the tumor, also known as pseudoprogression, occurs in a broad range of 11%–74% of all VS patients in the two to three years following treatment and can cause a temporary increase in cranial nerve morbidities.[15–30] For large VSs, where the tumor already exhibits a mass effect on the brainstem, this post-radiation effect may cause severe, and in some cases, life-threatening morbidities. This adverse effect necessitates salvage treatment, further increasing the risk of surgical complications.

As TTE is one of the major contra-indicators for radiosurgical treatment of large VSs, it would be extremely beneficial if this effect can be predicted *a priori*. This will enable physicians to select the most optimal treatment strategy on an individual basis. However, it remains unclear why some patients exhibit TTE, while others do not show TTE but exhibit an arbitrary volumetric response. Several investigations into the correlation of tumor- and treatment-related characteristics to this effect have been reported.[15–31] However, their results remain inconclusive. Treatment-related characteristics, such as marginal radiation dose and maximum tumor dose, were found not to correlate with TTE occurrence in all but one study.[27] Some papers describe that tumor volume is significantly different between VSs presenting TTE and those that do not,[19,27,30] while others could not find this correlation.[15–18,24,26,28,29] Also tumor appearance on MRI, classifying a VS tumor as cystic or solid, has been investigated. Shirato et al. determined that cystic tumors are more likely to exhibit TTE.[31] However, others did not find this correlation[17,19,25,28,32] or even observed that cystic tumors are less likely to exhibit TTE.[21]

The assumed biological effect of radiosurgery on VS cells is a combination of acute inflammation and vascular occlusion.[33,34] Because of this and the previously described contradicting results, it is hypothesized that differences in tumor biology may be the cause of TTE in a subset of patients. Ideally, a biopsy is performed to analyze tumor tissue. However, this is an undesired procedure as post-biopsy hemorrhage is one the most frequently encountered complications, which can cause even death due to the VS location close to the brainstem. The more readily available source of biological information is through imaging techniques, such as magnetic resonance imaging (MRI). These scans are already obtained for diagnostics and may contain information of the biological tumor features.

We have studied literature to find out as to how far image analysis techniques were explored for determining features describing tumor biological properties. In a review by Gillies et al., the authors reported on the potential power of medical image analysis using radiomics to facilitate improved clinical decision making.[35] Indeed, numerous studies describe the ability of employing computer-aided diagnosis using medical imaging for classifying disease and treatment response. Yang et al. evaluated tumor-derived MRI-texture features for discriminating molecular subtypes of glioblastomas and the corresponding 12-month survival status.[36] Their study obtained area under the receiver operating characteristic (AUC) values of 0.70 to 0.82 for the specific subtypes, and 0.69 for the 12-month survival status. Moreover, specifically for radiosurgical treatment responses, several authors evaluated the possibility to distinguish true tumor growth from radionecrosis in primary malignant brain tumors and brain metastases. Utilizing computer-extracted texture features, Tiwari et al. were able to distinguish cerebral radionecrosis from recurrent brain tumors on multi-parametric MRI.[37] Their method obtained AUC values of 0.79 on fluid-attenuated inversion recovery MRI images, both for primary malignant brain tumors and for brain metastases, thereby outperforming the diagnosis made by the medical experts. Zhang et al. evaluated 285 texture features calculated on four different MRI sequences to find a predictive model distinguishing radionecrosis from true tumor progression following radiosurgery on brain metastases.[38] They obtained an AUC value of 0.73 using so-called delta feature values, which represented the change in feature values from one time-point to the next. Peng et al. obtained an AUC value of 0.79 on distinguishing radionecrosis from tumor progression, using tenfold cross-validation of their prediction model.[39] Wang et al. demonstrated that multi-modality MRI imaging and radiomics analysis have potential to identify early treatment response of malignant gliomas

treated with concurrent radiosurgery and bevacizumab.[40] These studies all show the potential of distinguishing different radiosurgical treatment responses in malignant brain tumors. However, the ability to predict such a treatment response *prior* to treatment is, in the case of large VS tumors, crucial as this can lead to a well-informed treatment selection based on quantitative analysis. Since no clinical- or treatment-related parameters have shown their prognostic values, it is hypothesized that quantitatively analyzing the tumor appearance on readily available MRI scans can facilitate pre-treatment prediction of the TTE effect.

Therefore, the objective of this research is to explore whether TTE after radiosurgery, specifically Gamma Knife radiosurgery (GKRS) on VS, can be predicted from the measured MRI tumor texture characteristics. We analyze several texture features and apply machine learning as a technique for classifying the MRI observations. Our results show that the MRI tumor texture data of VS correlate to TTE, thereby enabling the prediction of this adverse effect.

2. MATERIALS AND METHODS

In this section, a description of the available data is presented and the proposed approach is discussed. Figure 1 depicts the flow diagram of the proposed approach, of which each element is discussed below. The texture feature extraction methods applied are first-order statistics and second-order statistics, based on gray-level co-occurrence matrices (GLCMs) and Minkowski functionals (MFs).

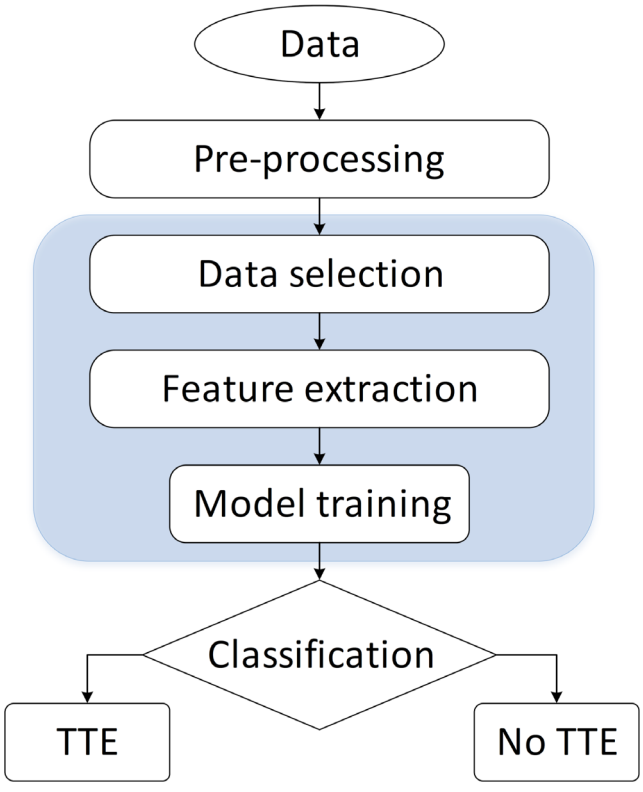

FIG. 1. Flow diagram of the proposed transient tumor enlargement prediction approach. [Color figure can be viewed at wileyonlinelibrary.com]

2.A. Data

The data employed in this research consist of prospectively collected patient- and treatment-related information and clinical MRI image data. The included patients have been selected based on a volumetric threshold derived from the Koos grade. This grade is used in a clinical setting to classify the VS tumor size. This classification is based on the tumor extent, where Grade I is selected for tumors only present in the auditory canal and Grade IV for tumors displacing the brainstem. Koos Grade IV is in medical terms considered a large tumor. For these large tumors, the adverse effects of TTE can cause severe complications because of the already caused displacement of critical brain structures and cranial nerves. Koos Grade IV tumors have a corresponding tumor volume of 4.17 ± 2.75 cubic centimeter (cm^3).[41] In this research, we therefore selected a lower bound of 1.42 cm^3 as the minimum inclusion threshold. Furthermore, as TTE will occur between 6 and 18 months after treatment, all included patients had at least an available MRI scan at 6 months following treatment and were followed-up for at least 18 months. These follow-up scans were employed for calculating tumor volume changes, needed for determining whether a TTE has occurred or not. This resulted in the inclusion of 99 patients.

The obtained patient- and treatment-related information included age at treatment, tumor volume at treatment [gross target volume (GTV)], radiation dose to 99% of the GTV, coverage (ratio between GTV within prescription isodose volume GTV_{PIV} and GTV), selectivity (ratio between GTV_{PIV} and PIV), gradient index (ratio between volume enclosed by half the prescription isodose and PIV), Paddick conformity index (coverage multiplied by selectivity), number of iso-centers and the beam-on time. We employed Student's *t*-tests to evaluate differences in these patient- and treatment-related characteristics between patients that suffered from TTE and those that did not.

The clinical MRI data employed in this research for texture analysis consisted of the MRI scans that were already acquired for treatment planning. These included T1-weighted, T2-weighted, and contrast-enhanced T1-weighted (T1CE) MRI scans and were obtained on the day of treatment. Ideally, for discriminating TTE from non-TTE, a histopathological evaluation of tumor tissue is employed. However, in current clinical practice, surgical intervention is highly unwanted due to the significantly increased inherent risks. This is why the medical team at our center opted for GKRS in the first place and also has the protocol to only intervene when the tumor expansion becomes life-threatening. In all other cases, watchful waiting is preferred for the first 2–3 yr following GKRS. As such, the presence or absence of TTE needs to be determined from the MRI data obtained at follow-up visits. To this end, tumor volumes were calculated on each available follow-up MRI, by segmenting the tumor using the treatment planning software (GammaPlan version 11, Elekta AB, Stockholm, Sweden). These tumor segmentations were created by authors PL and JV. Several publications report that the maximum TTE is observed

between 6 to 15 months after treatment, followed by volumetric reduction.[16,20,26,28,42] For this reason, the TTE effect is defined as a volumetric increase of at least 10% within the first 12 months after treatment, followed by volumetric reduction to at least the tumor volume at treatment. This threshold for volumetric increase was chosen based on inter- and intra-observer variability analysis of the tumor contouring in our center. Examples of the treatment and follow-up MRI scans of a VS tumor that exhibited TTE are shown in Fig. 2. If tumor expansion was less than 10% during the first 2 yr, the VS was considered to be stable or shrinking and consequently classified as non-TTE. Using these definitions for TTE and non-TTE, 38 out of the included 99 patients experienced a TTE after GKRS treatment. The remaining 61 patients were classified as non-TTE.

The treatment MRI scans, including the tumor delineations created by the neurosurgeon on the day of treatment, were extracted from the database of the Gamma Knife treatment system. The data from which image features are extracted consist of the MRI volume elements (voxels) within the tumor delineations.

2.B. Pre-processing

Whereas data from other medical imaging modalities are measured in absolute units, MRI data provides relative values. To support comparison between subjects, MRI intensities need to be normalized. To this end, we employ a multi-landmark intensity normalization (MLIN), which is based on the work by Madabhushi and Udupa.[43] This method aims to find a generalized intensity scale, such that MRI scanning parameters have a limited influence on the image analysis techniques. It performs normalization utilizing tissue-specific landmarks. For T1 and T1CE MRI scans, the utilized landmarks are the brainstem and the fiducial markers. For the T2 scans, the selected landmarks are the brainstem, the fiducial markers, and the cerebrospinal fluid. Examples of the landmarks are given in Fig. 3.

2.C. Feature extraction

For each tumor, the following first-order statistics (FOS) are computed from the tumor MRI voxels for each individual

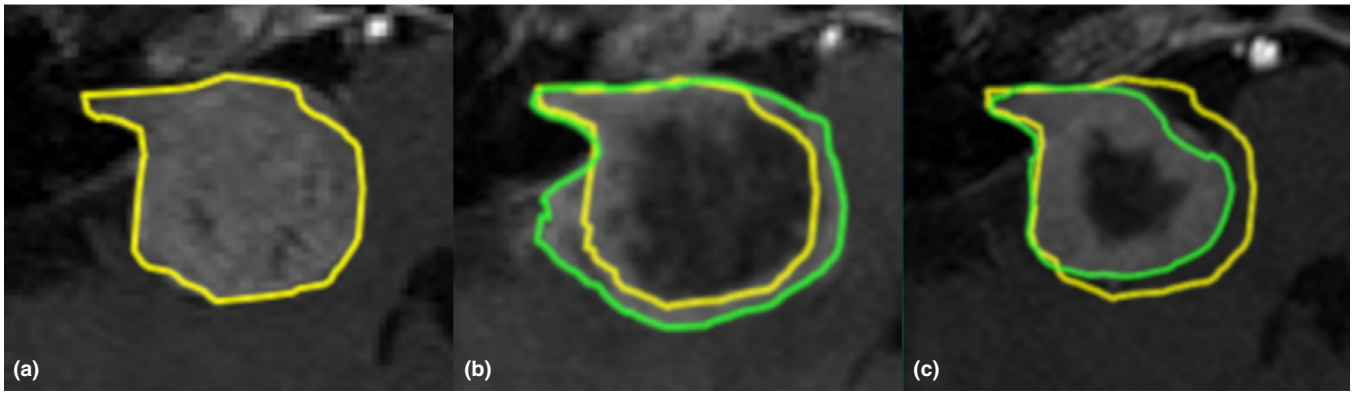

FIG. 2. T1-weighted, contrast-enhanced magnetic resonance images of a vestibular schwannoma that exhibited transient tumor enlargement after Gamma Knife radiosurgery. In each part of the figure, the yellow delineation depicts the tumor at time of treatment. Part a: tumor at time of treatment, with a volume of 12.8 cm^3. Part b: in green, the tumor 6 months after treatment, with a volume of 17.7 cm^3. Part c: in green, the tumor 24 months after treatment, with a volume of 8.7 cm^3.

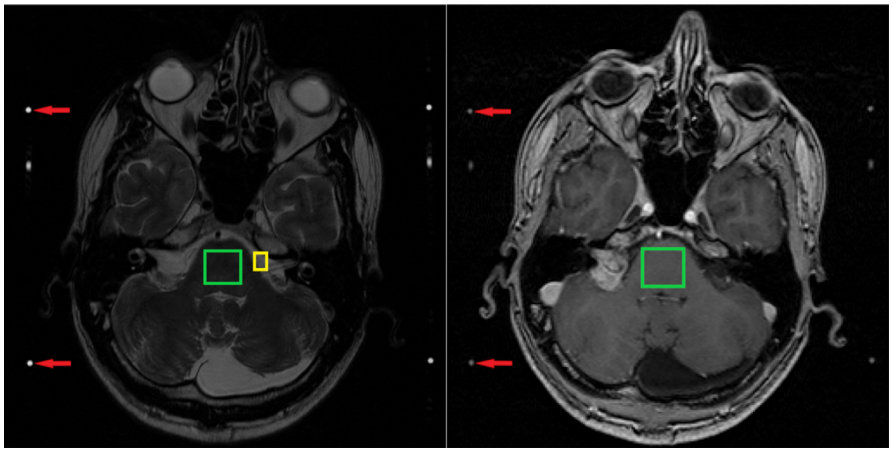

FIG. 3. Examples of the landmarks used in the multi-landmark intensity normalization method for the T2-weighted magnetic resonance imaging (MRI) scans (left) and T1-weighted, contrast-enhanced MRI scans (right). For the T1-weighted MRI scans, the same landmarks are used as shown in the right image. Highlighted in these images are the areas for the cerebrospinal fluid (yellow), brainstem (green) and the fiducial markers (red arrows).

MRI sequence: mean, standard deviation, skewness, kurtosis and a 16-bin histogram.

Next, we calculate the MFs as defined by Hadwiger,[44] again for each individual MRI sequence. In mathematical morphology, these functionals represent geometric measurements of shapes. These shapes are obtained by transforming gray-scale images to binary images using a threshold value. Varying this threshold value will result in multiple instances from which MFs can be computed. These variations allow for extracting texture information. From the binarized data, thresholded at level T, the following elementary geometric shape objects are extracted: (a) number of cubes N_c, (b) number of open faces N_f, (c) number of open edges N_e, and (d) number of open vertices N_v.[45] These objects are employed in the calculation of the following 4 functionals: foreground volume M_0^T, surface area M_1^T, curvature M_2^T, and Euler number M_3^T. These functionals are specified by

$$M_0^T = N_c, \tag{1}$$

$$M_1^T = -6N_c + 2N_f, \tag{2}$$

$$M_2^T = 3N_c - 2N_f + N_e, \tag{3}$$

$$M_3^T = -N_c + N_f - N_e + N_v. \tag{4}$$

The MFs are highly scale-dependent. Since the VS tumors in our data set have a ratio between minimum and maximum volume of 13:1, the MFs need to be normalized with respect to the tumor volume. This is performed by dividing the functionals by the maximum tumor volume in the data set.

Finally, for each individual MRI sequence, the GLCMs are computed. For the GLCM $\mathbf{P}_{\theta,d,l}$, each matrix element $\mathbf{P}_{\theta,d,l}(i,j)$ denotes the number of times a pixel with intensity i occurs together with a pixel of intensity j, at angle θ, distance d, and quantization level l. Each element is normalized with respect to the total number of elements in the GLCM. From these matrices, the following 4 features are calculated: entropy (H), contrast (Cn), energy (E), and correlation (Co).[46] These features are specified by:

$$H = \sum_{i,j} -\mathbf{P}_{\theta,d,l}(i,j)\log\left(\mathbf{P}_{\theta,d,l}(i,j)\right), \tag{5}$$

$$Cn = \sum_{i,j} |i-j|^2 \mathbf{P}_{\theta,d,l}(i,j), \tag{6}$$

$$E = \sum_{i,j} \mathbf{P}^2_{\theta,d,l}(i,j), \tag{7}$$

$$Co = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)\mathbf{P}_{\theta,d,l}(i,j)}{\sigma_i \sigma_j}. \tag{8}$$

here, i and j are the row and column indices of each GLCM element, respectively. Parameters μ_i and σ_i denote the mean and standard deviation of row i and μ_j and σ_j the mean and standard deviation of column j, respectively.

2.D. Classification

The final step is to train a classifier for binary prediction of TTE. The implemented machine learning method in this research is support vector machines (SVM), since it has proven to be effective in binary classification problems without requiring large amounts of data. As our data set is relatively small, containing only 99 tumors, advanced algorithms such as neural networks are not well-suited for classification. Considered SVM types include linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian, which are implemented in MATLAB (MathWorks, Natick, Massachusetts, USA). Validation is performed by tenfold cross-validation. The performance metrics for determining the optimal model are the sensitivity and specificity of the trained model. Additionally, models are evaluated by the area under the curve (AUC) of the receiver operating characteristic (ROC). The best models are selected based on their sensitivity and specificity. In case the multiple models perform equally well, preference is given to a higher specificity: false positives can be related to miss-predicted occurrence of TTE, necessitating salvage treatment in a patient. Alternatively, false negatives are related to a miss-predicted absence of TTE, resulting in the selection for microsurgical treatment. Because the first situation has a larger impact on the well-being of the patient, larger emphasis is placed on specificity.

Training of the models is based on different data inclusion criteria. The first criterion is derived from the small imbalance in the available data, where the majority class contains 61 patients (non-TTE) and the minority class 38 patients (TTE). Due to this imbalance, training can lead to a model that is skewed towards the majority class. This way a classification algorithm can obtain a reasonable accuracy, at the cost of a low specificity. To evaluate whether this imbalance impacts the results, training is performed in two different ways. First, all available data points are employed in training the SVM models. Second, a balanced training set is used, in which each cohort is equally sized. Balancing the data is performed by random subsampling of the majority class. To account for possible data biases, a further validation loop is employed. In this loop, models are trained using n resampled subsets of the majority class. The resulting model is the average of these n models, indicating the combined model performance.

The second data inclusion criterion is based on the MRI sequence. In this research, we have T1-, T1CE- and T2-weighted MRI data available. In the classification approach, we have evaluated each individual MRI sequence as well as the combination of all three sequences.

The third data inclusion criterion is based on the tumor volume. The data from which the MRI image features are extracted consist of the MRI voxels within the tumor delineations. Due to the employed scanning method and parameters, each tumor is scanned using the same voxel dimensions. Thus, MRI scans of large tumors contain more tumor voxels than scans of small tumors. If the number of tumor voxels increases, the amount of texture information also expands.

Therefore, we also explored the impact of the tumor volume, by imposing various volume thresholds for specific selection of the data (volume filtering). The selected volume thresholds were 2, 3, 4, 5, 6, and 7 cm^3, as higher thresholds resulted in too few number of patients in the minority class.

3. RESULTS

In this section we first describe the statistical analyses of the patient- and treatment-related characteristics. Next, the feature extraction parameters and results are presented. Finally, the classifier performances with regards to the balancing of the training data, the employed features and the tumor volume filtering are given.

3.A. Statistical analysis

For the statistical analysis, all patient- and treatment-related characteristics of the included patients were obtained from a prospectively collected database. A summary of the resulting characteristics can be found in Table I.

First, Student t-tests are employed for evaluating differences in patient- and treatment-related characteristics between patients suffering from TTE and those that do not show TTE. These tests are also performed after implementing the additional volume thresholds. The resulting p-values are presented in Table II. None of the tests obtained statistical significance ($P < 0.05$), showing that the patient- and treatment-related characteristics have no prognostic value of the occurrence of TTE. This is fully in agreement with the found literature on this subject.

3.B. Classification performance

This section presents the implemented feature parameters and results obtained per feature extraction method. For each extractor, we evaluate the impact of the volume thresholding, as discussed above. First, the FOS results are discussed. Next, the results of the MFs are presented and finally, the GLCM-based results are given.

TABLE I. Patient- and treatment-related characteristics for the complete patient cohort.

	Mean	Interquartile range	Range
Age (yr)	58	47–66	24–84
Tumor volume at treatment (cm^3)	6.54	3.10–6.04	1.44–18.72
Dose to 99% of the tumor volume (Gy)	12.36	11.80–13.00	11.10–13.20
Coverage (%)	95.74	91.00–99.00	86.00–100.00
Selectivity	0.89	0.85–0.90	0.71–0.99
Gradient index	2.74	2.58–2.82	2.45–3.60
Paddick conformity index	0.84	0.84–0.89	0.17–0.93
Number of iso-centers	24	17–31	1–53
Beam-on time (min)	60.27	42.18–75.03	22.80–144.80

TABLE II. Resulting *P*-values of the student's *t*-tests per volume threshold. In the second row, the number of patients after each volume threshold is given. None of the *P*-values reach statistical significance.

Volume threshold	–	2 cm^3	3 cm^3	4 cm^3	5 cm^3	6 cm^3	7 cm^3
Number of patients (TTE — non-TTE)	38–61	34–58	31–45	25–41	24–37	19–32	17–26
Age	0.315	0.514	0.696	0.604	0.614	0.643	0.149
Tumor volume at treatment	0.527	0.513	0.142	0.332	0.191	0.254	0.121
Dose to 99% of the tumor volume	0.152	0.145	0.094	0.126	0.204	0.202	0.301
Coverage	0.581	0.739	0.590	0.681	0.672	0.993	0.782
Selectivity	0.909	0.908	0.919	0.980	0.761	0.910	0.739
Gradient index	0.383	0.443	0.248	0.280	0.225	0.378	0.595
Paddick conformity index	0.961	0.989	0.954	0.932	0.774	0.740	0.757
Number of iso-centers	0.792	0.645	0.786	0.499	0.687	0.768	0.819
Beam-on time	0.548	0.630	0.550	0.853	0.504	0.611	0.990

3.B.1. FOS

The calculated first-order-statistics of the MRI scans are the mean, standard deviation, skewness and kurtosis. Furthermore, a 16-bin histogram is included, resulting in a total of 20 features per MR image sequence. In Table III, the performances for both training strategies, including all available data and including balanced data, of the best FOS-based models are presented for the various volume thresholds.

For the FOS-based features, the model based on balanced training data achieved a sensitivity and specificity of 0.72 and 0.40, respectively. Excluding tumors smaller than 7 cm^3 from the training data improves this performance slightly to values of 0.66 and 0.58, respectively. Training the SVM models on all available data resulted in models that were skewed towards the majority class. This is clearly visible in Table III, where the values for the specificity are all significantly below 0.50 in all but one of the best-performing models. The model of

TABLE III. Highest-performing first-order statistics-based models for various volume thresholds and data selection methods. Training data is either a balanced subset (Balanced) or the entire set (Full).

	Balanced		Full	
Volume threshold	Sensitivity	Specificity	Sensitivity	Specificity
–	0.72	0.44	0.84	0.34
2 cm^3	0.48	0.63	0.87	0.29
3 cm^3	0.47	0.70	0.73	0.52
4 cm^3	0.63	0.50	0.83	0.40
5 cm^3	0.46	0.65	0.95	0.25
6 cm^3	0.67	0.55	0.94	0.32
7 cm^3	0.66	0.58	1.00	0.35

exception obtained sensitivity and specificity values of 0.73 and 0.52, respectively. It becomes clear that, when including the results after balancing the training data, FOS-based features are not well-suited for predicting TTE.

3.B.2. MFs

The MFs are computed as a function of the binarization threshold *T*. These computations can be performed for all available discrete levels, though the functionals show high correlation between subsequent thresholds when the difference between thresholds is small. Therefore, we employ 9 threshold levels equally spaced between 0 and 1. The resulting 36 MF features are computed per MRI sequence. For training the SVM models, we employed each functional M_i^T for $i = 0,\ldots,3$ individually, as well as combined. The performance metrics of the best MF-based models are given in Table IV. The performance of the best model employing MF features, combined with a balanced training set, results in sensitivity and specificity values of 0.69 and 0.53, respectively. Implementation of the volumetric threshold slightly increases these metrics to 0.64 and 0.61, respectively, for 7 cm^3. The impact of the imbalance in the dataset is less present for the MF-trained models, compared to the FOS-trained models, although some models still are skewed towards the majority class. The highest-performing MF-based model trained on all available data obtained sensitivity and specificity values of 0.80 and 0.60, respectively. These values were obtained with a minimum volumetric inclusion criterion of 4 cm^3.

3.B.3. GLCM

Generally, GLCM matrices are evaluated for the four unique two-dimensional (2D) directions, chosen as $\theta \in \{0°, 45°, 90°, 135°\}$. Specific subsets may be chosen based on existing clinical knowledge. However, such clinical information is unavailable, due to the unknown factors that influence TTE. As 3D MRI scans are available, we employ the 3D extension of the GLCM directions. Each direction is separated by a 45° rotational offset on the cardinal planes, resulting in 13 unique GLCM directions per MRI scan. The second parameter of the GLCM, distance *d*, is evaluated for the integer values 1,2,...,6. The upper bound of 6 has been selected according to half the size of the smallest tumor dimension in the data set. The third parameter, the maximum number of quantization levels *l*, affects the fine details retained in the input image. This parameter is evaluated for values with power of 2: 2^2,2^3,...,2^6. Implementing all parameter combinations gives a total of 390 unique GLCMs per MRI sequence.

The GLCM features employed in training a single SVM model, are composed of the entropy, contrast, energy and correlation, calculated from a single GLCM. Given the number of GLCMs per MRI sequence and the number of SVM types, a total of 9360 GLCM-based models are trained. Table V shows the results of the top-performing model for each data inclusion setting.

Initial tests with GLCM-based features were performed with a balanced training set. Without volumetric exclusion of data, sensitivity and specificity values of 0.69 and 0.75 are obtained, respectively. Applying volumetric thresholds on the data improves model performance. A sensitivity and specificity of 0.79 and 0.75 are obtained, respectively, when implementing the maximum volume threshold.

Next, the effect of data balancing is explored. Utilizing the full data set, in contrast to a balanced subset, increases the number of included samples by more than 20%. The effect of these additional data yields a performance improvement, increasing sensitivity and specificity to 0.82 and 0.69, respectively. For a minimum volume inclusion criterion of 6 cm^3, the highest sensitivity and specificity values of 0.77 and 0.89 are obtained, respectively. From these results, it can be concluded that GLCM features contain the most predictive information of TTE. Application of all training data compared to a balanced training set only slightly improves the performance of these GLCM-based models.

However, the largest effect on the performance for these models is imposing more strict volumetric data thresholds. Table V shows this effect on the highest obtained sensitivity and specificity. A different metric employed to evaluate model performance is the area under the curve (AUC) of the

TABLE IV. Highest-performing Minkowski functionals-based models for various volume thresholds and data selection methods. Training data is either a balanced subset (Balanced) or the entire set (Full).

	Balanced		Full	
Volume threshold	Sensitivity	Specificity	Sensitivity	Specificity
–	0.69	0.53	0.82	0.50
2 cm^3	0.74	0.50	0.80	0.50
3 cm^3	0.63	0.59	0.93	0.42
4 cm^3	0.61	0.60	0.80	0.60
5 cm^3	0.63	0.56	0.73	0.63
6 cm^3	0.60	0.67	0.87	0.47
7 cm^3	0.65	0.65	0.69	0.64

TABLE V. Highest-performing gray-level co-occurrence matrices-based models for various volume thresholds and data selection methods. Training data is either a balanced subset (Balanced) or the entire set (Full).

	Balanced		Full	
Volume threshold	Sensitivity	Specificity	Sensitivity	Specificity
–	0.69	0.75	0.82	0.69
2 cm^3	0.64	0.76	0.76	0.65
3 cm^3	0.68	0.73	0.84	0.61
4 cm^3	0.70	0.75	0.88	0.64
5 cm^3	0.67	0.75	0.89	0.67
6 cm^3	0.71	0.79	0.77	0.89
7 cm^3	0.79	0.75	0.85	0.75

receiver operating characteristic (ROC). In Fig. 4, the ROC curves of the best GLCM-based models for the various volume thresholds are depicted. Here, all thresholds obtain similar performance. The models obtain AUC values of approximately 0.90–0.95. These results are validated by performing bootstrapping, resulting in confidence intervals of approximately 0.80 up to 0.99.

Furthermore, the models obtaining these results show large variations in their parameters. Among the seven best models, one for each volume threshold, all three image modalities perform best at least once. Additionally, all quantization levels show the same effect, where each level is implemented at least once in the highest-performing models. As the data mainly differ in volume between these models, the large variations between parameters indicate the presence of information on various levels. A combination of models and features may prove to further enhance the results.

4. DISCUSSION

This research was performed in order to find predictive features of TTE after GKRS treatment of VS. Previous studies investigated this problem from a clinical point-of-view.[15–31] These studies did not find decisive correlations. Moreover, several studies contradict the results previously found in other studies. Therefore, it remains unknown if prediction of TTE is possible. We were able to achieve a classification sensitivity and specificity of 0.82 and 0.69, respectively. When employing volume thresholding, we obtained improved performances for increasing volumes. For tumors larger than 6 cm^3, a sensitivity and specificity of 0.77 and 0.89 were realized, respectively. These results were obtained by employing features from individual GLCMs and represent the highest-scoring models. Additionally, multiple models based on individual GLCMs achieved promising classification results. Combining features from these individual GLCMs may improve the presented results and enable prediction of TTE with even higher accuracy, sensitivity, and specificity. Furthermore, we determined that features calculated from different MR sequences also show promising results. Thus, next to combining features from individual GLCMs, combining features from different MR sequences can improve these results also.

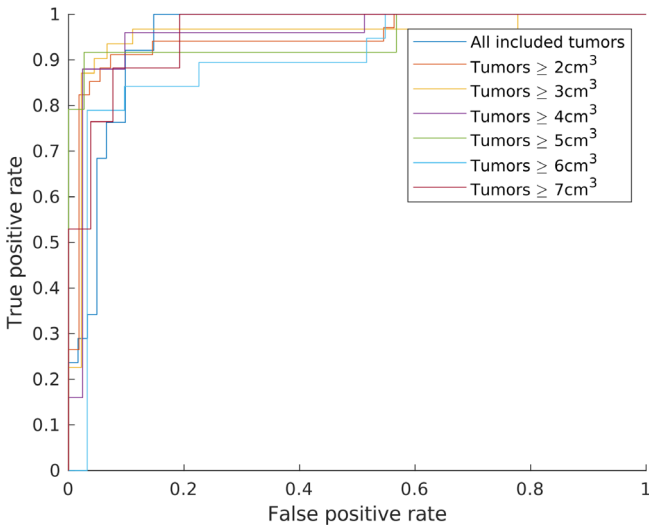

FIG. 4. Receiver operating characteristic curves of the best performing model per volume threshold setting. Model features derived from gray-level co-occurrence matrices, based on a balanced training set.

The three feature extractors implemented in this study were selected both on technical and on clinical aspects. Technically speaking, the implemented features and classification method have a proven track record in other healthcare image analysis applications, for instance for oncology. Both GLCM and MFs attempt to measure local changes in gray-level texture within the MR image, thereby addressing heterogeneous properties of the tissue. Furthermore, SVM has shown to be effective in binary classification problems without requiring large amounts of data. We are aware that these techniques are at present outperformed by machine learning using convolutional neural networks (CNNs). However, since this work is the first to explore this data and research question, the dataset was inherently limited at the beginning, which prevented straightforward application of this new machine learning technology. In this view, the current work can serve as a good baseline benchmark. Furthermore, we consider that starting with such an advanced technique, clinical application would only be accepted if the learning network would provide what is actually learned from the data. Since our exploration has indicated important features to be used as a reference, we have learned what is important in the images and this knowledge can be further exploited in developing so-called explainable artificial intelligence.

Coming back on the employed feature extractors, but now from a clinical point of view, we remark that they are based on the supposition of the neurosurgeons that perform the GKRS treatment of VS tumors in our center. They surmise that enhancing tumors with inhomogeneous texture properties show different behavior than the homogeneously enhancing tumors. More specifically, inhomogeneity in the form of dark streaks and dark areas within the enhancing lesion are considered to be the most informative visual properties. Thus, we selected the three described feature extractors, since these can adequately quantify such forms of heterogeneity. However, the results in this study may further improve by investigating other texture features employed in radiomics analysis of medical images, which is a point of further research.

A significant confounder in this research is its retrospective character. One of the disadvantages of the retrospectively analyzed data is that MR image intensities can vary between subjects, because MR protocols and scanners may have changed in the course of time. Despite our attempt to minimize the impact of the inter-subject MR intensity variations by implementing an advanced normalization method, these variations may still be present in the prediction approach, albeit at reduced level.

Another confounder is the applied definition of TTE. As stated by Marston et al.,[42] TTE is difficult to differentiate from true tumor growth. Ideally, a histopathological examination of tumor tissue obtained from resection is employed for

determining the ground-truth labeling. However, in the case of a VS, surgical intervention is only warranted if the mass effect of the tumor causes life-threatening issues. In all other cases, the transient swelling is accepted and carefully followed-up. Thus, only the volumetric data obtained from the follow-up MRI scans can be used for determining TTE. Nevertheless, this may have caused an incorrect labeling of the data, creating uncertainty in the final classification results.

Moreover, inter- and intra-observer variations in measuring tumor volumes make determining true volume changes difficult. During treatment planning, the VS tumor is segmented by the treating neurosurgeon, using a semi-automatic contouring tool incorporated in the treatment planning software. In the course of time, a total of 6 different neurosurgeons treated VS tumors at our Gamma Knife center. Furthermore, the follow-up MRI scans were segmented using the same tool by one neurosurgeon (JV) and one researcher (PL). An in-house evaluation of the inter- and intra-observer variations demonstrated that these variations decreased for increasing tumor volumes. For volumes larger than 1 cm^3, this variation reduced to <10%. This, together with the confounding ground-truth labeling, motivates why we only included patients who presented obvious TTE and obvious non-TTE. This strict selection was implemented to create two distinct cohorts. However, this definition may have caused a selection bias that has influenced the obtained results.

Furthermore, the inter- and intra-observer variations also influence the amount of voxels included in the feature extraction algorithms. However, due to the employed method for tumor segmentation, the variations in contouring are found in the so-called partial-volume effects of the MRI scans. These variations are considered to have a limited impact, because they constitute <10% of the total amount of voxels and because features are calculated globally. Nevertheless, it could have influenced the calculated features and the obtained results.

These confounding factors may have influenced the obtained results and the robustness of it. Currently, we are the only Gamma Knife center in the Netherlands treating this type of brain tumor. As such, we assume that we have a good cross-section of all VS patients. Thus, the results found in this study are most likely applicable to other Gamma Knife centers as well. However, the obtained results need to be validated, preferably in a joint multi-center setting. This would ensure that these confounding factors are reduced, thereby improving the robustness of the obtained results. Furthermore, a prospective study could be designed to cope with the previously described problems. Nonetheless, the results achieved in this study strongly suggest the possibility of TTE-prediction for individual treatment selection, making an implementation of this in the clinical workflow conceivable.

5. CONCLUSIONS

At present, small-to-medium sized VSs are generally treated using GKRS, as the treatment goal for these tumors has shifted from complete removal with inherent risks for the cranial nerve functions to less invasive techniques such as GKRS. However, for large VS tumors, microsurgical excision remains the preferred treatment strategy. Since the risks involved in microsurgery can be contra-indicative for this strategy, less invasive treatments such as radiosurgery and radiotherapy have been considered increasingly in the last decade obtaining good results with acceptable radiation-induced morbidities. However, it remains a controversial alternative to microsurgery, since one of the major contra-indications for GKRS on large VSs is the adverse effect of TTE. Therefore, the possibility of predicting TTE would be extremely beneficial as this would enable the selection of the most optimal treatment strategy on an individual basis.

It is hypothesized that the origin of this phenomenon can be found in variations in individual tumor biology. We explored the idea that the various tumor appearances on MRI reflect variations in tumor biology. Therefore, we employed quantitative MRI texture features derived from conventional MR images in this research.

Using texture features extracted from MRI data, we were able to obtain classification sensitivity and specificity values of 0.77 and 0.89, respectively. These results clearly show that MRI tumor texture can provide information for enabling the prediction of TTE. This can form a basis for individual VS treatment selection, further improving overall treatment results. Particularly for patients with large VSs, where the phenomenon of TTE is most relevant and for which our predictive model performs best, these findings can lead to an implementation in a clinical workflow such that for each patient the most optimal treatment strategy can be determined.

FUNDING INFORMATION

This research and publication was funded by ZonMw [80-84200-98-15222].

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

Part of this work was presented at the 14th International Stereotactic Radiosurgery Society Congress 2019 (ISRS2019), June 9–13, Rio de Janeiro, Brazil.
[a)]Author to whom correspondence should be addressed. Electronic mails: p.p.j.h.langenhuizen@tue.nl, p.langenhuizen@etz.nl.

REFERENCES

1. Ostrom QT, Gittleman H, Liao P, et al. Statistical Report: primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. *Neuro Oncol*. 2017;19:v1–v88.
2. Kleijwegt M, Ho V, Visser O, Godefroy W, Van Der Mey A. Real incidence of vestibular schwannoma? Estimations from a national registry. *Otol Neurotol*. 2016;37:1411–1417.
3. Lin XEP, Crane XBT, Lin EP. The Management and Imaging of Vestibular Schwannomas.

4. Carlson ML, Habermann EB, Wagie AE, et al. The changing landscape of vestibular schwannoma management in the United States—a shift toward conservatism. *Otolaryngol–Head Neck Surg*. 2015;153:440–446.
5. Wolbers JG, Dallenga AH, Romero AM, Van Linge A. What intervention is best practice for vestibular schwannomas? A systematic review of controlled studies. *BMJ Open*. 2013;3:e001345.
6. Abou-Al-Shaar H, Azab MA, Karsy M, et al. Assessment of costs in open surgery and stereotactic radiosurgery for vestibular schwannomas. *J Neurosurg*. 2018;131:561–568.
7. Bailo M, Boari N, Franzin A, et al. Gamma Knife radiosurgery as primary treatment for large vestibular schwannomas: clinical results at long-term follow-up in a series of 59 patients. *World Neurosurg*. 2016;95:487–501.
8. Chung WY, Pan DHC, Lee CC, et al. Large vestibular schwannomas treated by Gamma Knife surgery: long-term outcomes. *J Neurosurg*. 2010;113:112–121.
9. Huang C-W, Tu H-T, Chuang C-Y, et al. Gamma Knife radiosurgery for large vestibular schwannomas greater than 3 cm in diameter. *J Neurosurg*. 2018;128:1380–1387.
10. Iorio-Morin C, AlSubaie F, Mathieu D. Safety and efficacy of Gamma Knife radiosurgery for the management of koos grade 4 vestibular schwannomas. *Neurosurgery*. 2016;78:521–530.
11. Milligan BD, Pollock BE, Foote RL, Link MJ. Long-term tumor control and cranial nerve outcomes following Gamma Knife surgery for larger-volume vestibular schwannomas: clinical article. *J Neurosurg*. 2012;116:598–604.
12. van de Langenberg R, Hanssens PEJ, Verheul JB, et al. Management of large vestibular schwannoma. Part II. Primary Gamma Knife surgery: radiological and clinical aspects.
13. Zeiler FA, Bigder M, Kaufmann A, et al. Gamma Knife radiosurgery for large vestibular schwannomas: a Canadian experience. *Can J Neurol Sci*. 2013;40:342–347.
14. Lefranc M, Da Roz LM, Balossier A, Thomassin JM, Roche PH, Regis J. Place of Gamma Knife stereotactic radiosurgery in grade 4 vestibular schwannoma based on case series of 86 patients with long-term follow-up. *World Neurosurg*. 2018;114:e1192–e1198.
15. Yu CP, Cheung JY, Leung S, Ho R. Sequential volume mapping for confirmation of negative growth in vestibular schwannomas treated by Gamma Knife radiosurgery. *J Neurosurg*. 2000;93:82–89.
16. Meijer OWM, Weijmans EJ, Knol DL, et al. Tumor-Volume Changes after Radiosurgery for Vestibular Schwannoma: Implications for Follow-Up MR Imaging Protocol.
17. Mohammed FF, Schwartz ML, Lightstone A, Beachey DJ, Tsao MN. Pseudoprogression of vestibular schwannomas after fractionated stereotactic radiation therapy. *J Radiat Oncol*. 2013;2:15–20.
18. Transient HL. Tumor Volume Increase in Vestibular Schwannomas After Radiotherapy. Cureus; 2012.
19. Aoyama H, Onodera S, Takeichi N, et al. Symptomatic outcomes in relation to tumor expansion after fractionated stereotactic radiation therapy for vestibular schwannomas: single-institutional long-term experience. *Int J Radiat Oncol Biol Phys*. 2013;85:329–334.
20. Nakamura H, Jokura H, Takahashi K, Boku N, Akabane A, Yoshimoto T. *Serial Follow-up MR Imaging after Gamma Knife Radiosurgery for Vestibular Schwannoma*. Vol 21; 2000.
21. Kim JH, Jung HH, Chang JH, Chang JW, Park YG, Chang WS. Predictive factors of unfavorable events after Gamma Knife radiosurgery for vestibular schwannoma. *World Neurosurg*. 2017;107:175–184.
22. Wowra B, Muacevic A, Jess-Hempen A, Hempel J-M, Müller-Schunk S, Tonn J-C. Outpatient Gamma Knife surgery for vestibular schwannoma: definition of the therapeutic profile based on a 10-year experience. *J Neurosurg*. 2005;102:114–118.
23. van Eck ATCJ, Horstmann GA. Increased preservation of functional hearing after Gamma Knife surgery for vestibular schwannoma. *J Neurosurg*. 2013;119:204–206.
24. Okunaga T, Matsuo T, Hayashi N, et al. Linear accelerator radiosurgery for vestibular schwannoma: measuring tumor volume changes on serial three-dimensional spoiled gradient-echo magnetic resonance images. *J Neurosurg*. 2005;103:53–58.
25. van de Langenberg R, Dohmen AJC, de Bondt BJ, Nelemans PJ, Baumert BG, Stokroos RJ. Volume changes after stereotactic LINAC radiotherapy in vestibular schwannoma: control rate and growth patterns. *Int J Radiat Oncol Biol Phys*. 2012;84:343–349.
26. Nagano O, Higuchi Y, Serizawa T, et al. Transient expansion of vestibular schwannoma following stereotactic radiosurgery. *J Neurosurg*. 2008;109:811–816.
27. Lee C-C, Wu H-M, Chung W-Y, Chen C-J, Pan DH-C, Hsu SPC. Microsurgery for vestibular schwannoma after Gamma Knife surgery: challenges and treatment strategies. *J Neurosurg*. 2014;121:150–159.
28. Hayhurst C, Zadeh G. Tumor pseudoprogression following radiosurgery for vestibular schwannoma. *Neuro Oncol*. 2012;14:87–92.
29. Pollock BE, Driscoll CLW, Foote RL, et al. Patient outcomes after vestibular schwannoma management: a prospective comparison of microsurgical resection and stereotactic radiosurgery. *Neurosurgery*. 2006;59:77–85.
30. Kim K-M, Park C-K, Chung H-T, Paek SH, Jung H-W, Kim DG. Long-term outcomes of Gamma Knife stereotactic radiosurgery of vestibular schwannomas. *J Korean Neurosurg Soc*. 2007;42:286–292.
31. Shirato H, Sakamoto T, Takeichi N, et al. Fractionated stereotactic radiotherapy for vestibular schwannoma (VS): comparison between cystic-type and solid-type VS. *Int J Radiat Oncol Biol Phys*. 2000;48:1395–1401.
32. van de Langenberg R, Hanssens PEJ, van Overbeeke JJ, et al. Management of large vestibular schwannoma. Part I. Planned subtotal resection followed by Gamma Knife surgery: radiological and clinical aspects. *J Neurosurg*. 2011;115:875–884.
33. Linskey ME, de Lunsford DL, Flickinger JC. Radiosurgery for acoustic neurinomas: early experience. *Neurosurgery*. 1990;26:736–745.
34. Witham TF, Okada H, Fellows W, et al. The characterization of tumor apoptosis after experimental radiosurgery. *Stereotact Funct Neurosurg*. 2005;83:17–24.
35. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278:563–577.
36. Yang D, Rao G, Martinez J, Veeraraghavan A, Rao A. Evaluation of tumor-derived MRI-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma. *Med Phys*. 2015;42:6725–6735.
37. Tiwari P, Prasanna P, Wolansky L, et al. Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric MRI: a feasibility study. *Am J Neuroradiol*. 2016;37:2231–2236.
38. Zhang Z, Yang J, Ho A, et al. A predictive model for distinguishing radiation necrosis from tumour progression after Gamma Knife radiosurgery based on radiomic features from MR images. *Eur Radiol*. 2018;28:2255–2263.
39. Peng L, Parekh V, Huang P, et al. Distinguishing true progression from radionecrosis after stereotactic radiation therapy for brain metastases with machine learning and radiomics. *Int J Radiat Oncol Biol Phys*. 2018;102:1236–1243.
40. Wang C, Sun W, Kirkpatrick J, Chang Z, Yin F-F. Assessment of concurrent stereotactic radiosurgery and bevacizumab treatment of recurrent malignant gliomas using multi-modality MRI imaging and radiomics analysis. *J Radiosurg SBRT*. 2018;5:171–181.
41. Mindermann T, Schlegel I. Grading of vestibular schwannomas and corresponding tumor volumes: ramifications for radiosurgery. *Acta Neurochir (Wien)*. 2013;155:71–74.
42. Marston AP, Jacob JT, Carlson ML, Pollock BE, Driscoll CLW, Link MJ. Pretreatment growth rate as a predictor of tumor control following Gamma Knife radiosurgery for sporadic vestibular schwannoma. *J Neurosurg*. 2017;127:380–387.
43. Madabhushi A, Udupa JK. New methods of MR image intensity standardization via generalized scale. *Med Phys*. 2006;33:3426–3434.
44. Hadwiger H. *Vorlesungen Über Inhalt, Oberfläche Und Isoperimetrie*. Berlin, Heidelberg: Springer; 1957.
45. Li X, Mendonça PRS, Bhotika R. Texture analysis using Minkowski functionals. In: Haynor DR, Ourselin S, eds. *Medical Imaging 2012: Image Processing*. Vol 8314. International Society for Optics and Photonics; 2012:83144Y.
46. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;SMC-3:610–621.