Article

CLINICAL and TRANSLATIONAL NEUROSCIENCE

Clinical & Translational Neuroscience
July-December 2018: 1–11

DOI: 10.1177/2514183X18786602
journals.sagepub.com/home/ctn
SAGE

Computer-aided radiological diagnostics improves the preoperative diagnoses of medulloblastoma, pilocytic astrocytoma, and ependymoma: A reproducibility study

Nicole Porz[1], Urspeter Knecht[1], Beate Sick[2], Elvis Murina[3], Nuno Barros[1], Philippe Schucht[1], Evelyn Herrmann[1], Jan Gralla[1], Roland Wiest[1], Marwan El-Koussy[1], and Johannes Slotboom[1]

Abstract

Introduction: Imaging-based diagnosis of intra-axial contrast-enhancing brain tumors is frequently challenging. We show that the diagnosis of medulloblastoma (MDB) versus pilocytic astrocytoma (PA) and ependymoma (EPM) profit from computational analyses, based on quantitative image properties (i.e. textural features from apparent diffusion coefficient (ADC)-maps) and an automated machine learning classification (random forests (RF)). **Methods:** Forty patients who were diagnosed with three types of brain tumors were included in this study: 16 with MDB, 4 with PA, and 10 EPM. Based on the analysis of multi parametric preoperative magnetic resonance images, neuroradiologists gave a clear-cut diagnosis if they were sure of the diagnosis; however, most diagnoses comprise several possible tumor types. To distinguish between the named tumor types, a computer-based differential diagnosis (DD) tool was developed. Tumor lesion volumes were manually defined using ADC-maps only. From the demarked ADC-map, texture-parameters were extracted to train RF classifiers for pairwise DD. Performance of the RF models and reproducibility of the manual segmentation were evaluated. **Results:** Neuroradiologists gave correct and clear-cut diagnoses for 31% of MDB, 14.3% of PA, and 10% of EPM. Most diagnoses comprised several tumor types and altogether diagnoses containing the right tumor were given in 69% of true MDB, 64% of true PA, and 30% of true EPM. Ambiguous diagnoses could be improved by RF classifiers showing the following PA versus MDB performance: *sensitivity* 0.888 ± 0.031, *specificity* 0.886 ± 0.036; EPM versus MDB: *sensitivity:* 0.938 (95% CI = (0.677, 0.997)) and *specificity:* 0.7 (95% CI = (0.354, 0.919)); EPM versus PA: *sensitivity*: 0.786 (95% CI = (0.488, 0.942) and *specificity:* 0.100 (95% CI = (0.005, 0.458). An inter- and intra-rater analysis (three human raters) was performed and the Fleiss' kappa test revealed high *inter-rater* agreement of $\kappa = 0.821$ (p value << 0.001) and an *intra-rater* agreement of $\kappa = 0.822$ (p value << 0.001). **Conclusion:** In the frequent case of ambiguous neuroradiologist diagnoses, a subsequent differential RF classification improves the diagnoses in all cases. The largest benefit is gained for the discrimination PA versus MDB with an accuracy of $88.0 \pm 3.0\%$ followed by EPM versus MDB with an accuracy of 84.6%.

Keywords

MRI, radiomics, computer-aided neuroradiological diagnostics, medulloblastoma, pilocytic astrocytoma, ependymoma, texture analysis, brain cancer

[1] University Hospital Bern and Inselspital, Bern, Switzerland
[2] University of Zurich, Zürich, Switzerland
[3] Zurcher Hochschule fur Angewandte Wissenschaften Rosenstrasse, Winterthur, Switzerland

Corresponding author:
Johannes Slotboom, Inselspital Universitatsspital Bern, Freiburgstrasse 10, Bern, Switzerland.
Email: johannes.slotboom@insel.ch

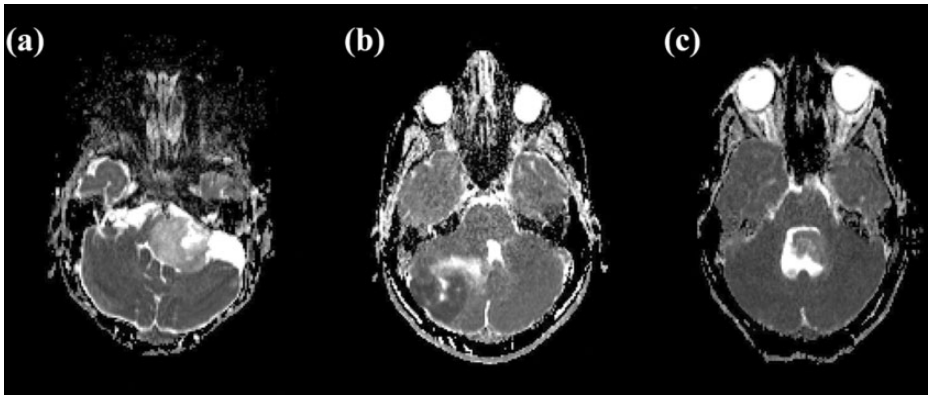

Figure 1. Illustrative ADC-maps of three patients suffering from (a) PA, (b) MDB, and (c) EPM. By only visual inspection, it is very difficult to distinguish these brain tumors from each other. As shown in this article, quantitative texture parameter analysis, combined with machine learning can improve diagnostic accuracy substantially. PA: pilocytic astrocytoma; MDB: medulloblastoma; EPM: ependymoma.

Introduction

Noninvasive differentiation between medulloblastoma (MDB), pilocytic astrocytoma (PA), and ependymoma (EPM) using conventional magnetic resonance imaging (MRI) techniques is frequently prone to misinterpretation, since these tumors may have similar appearance on, for example, diffusion-weighted imaging (DWI), T2w/Fluid-attenuated inversion recovery (FLAIR), and T1 contrast-enhancing images. To illustrate the difficulty to distinguish these diseases by visual inspection, Figure 1 displays the apparent diffusion coefficient (ADC)-maps of three typical patients suffering from the above mentioned tumor types. A major difference between the tumor classes is achieved by the cellular density and organization patterns, which directly influence the diffusivity of protons in the extracellular space, which is characterized by the so-called ADC. Many studies using ADC-map information in diagnosing brain tumors have been performed. For instance, Rumboldt et al.[1] found *significant* differences in ADC mean value between PA, MDB, and EPMs. Schneider et al.[2] used the combined value of DWI and proton magnetic resonance spectroscopy (1H-MRS) for the same differential diagnosis (DD). Yamashita et al.[3] found that minimum apparent diffusion coefficient is significantly correlated with cellularity and found differences between the tumor types. Jaremko et al.[4] showed that MDB and PA could be differentiated but EPM could *not* be reliably differentiated from MDB or PA. Gimi et al.[5] used a tumor/normal brain ADC-ratio threshold and used ratio thresholds only for tumor discrimination. Bull et al.[6] used more sophisticated ADC-based histogram parameters to discriminate the tumors, however, on a very small number of patients. Koral et al.[7] studied the impact of diffusion MRI on accuracy of visual diagnoses, and concluded that ADC-maps help to improve the diagnosis. Pierce and Provenzale[8] also found that minimal ADC values can be used to differentiate brain tumors. Another method for differentiating these brain tumor types was proposed by Gutierrez et al.,[9] using support vector machine-based classifiers using ADC histogram features that yielded very good discrimination among pediatric posterior fossa tumor types, and ADC-extracted textural-features that show promising results for further subtype discrimination. The approach of Gutierrez et al.[9] uses classification algorithms applied to radiological image data for diagnosis. Such an approach in diagnostics can be regarded as *computer-aided radiological diagnostics* (CARD).

In this article, we describe a different type of CARD method aiming at application in a clinical routine setting. This approach is based on random forests (RF) of Breiman.[10] Our novel semiautomatic CARD method should enable the neuroradiologist in daily clinical routine to obtain support for choosing the most likely diagnosis (in this case PA, MDB, or EPM). The presented method is also based on ADC-map features as MRI surrogate markers to tumor-specific molecular processes. The method combines expert-based segmentation of the *complete* tumor-affected volume, with RF classification for diagnosis. Since the contouring of the tumors is performed manually, it's inherent that the reproducibility cannot be 100%. Hence it is to be conceivable that the RF classification performance, that is diagnostic performance, will be randomly influenced by this. One aim of this study is to evaluate how big the inter- and intra-rater influence of manual segmentation is on the RF-classification performance, in order to test the robustness of the classifiers obtained.

The following research questions were investigated: (1) how does the CARD method performs compared to expert-based diagnosis, (2) does the intra- and inter-rater variability in segmentation affect the RF prediction, and (3) is it possible to improve the *individual* clinical diagnosis, without the need of additional image data, that is, by utilizing the available digital MRI information and previously confirmed diagnoses in a quantitative way.

Methods

Included patients

Patients with newly diagnosed and histologically confirmed PA, MDB, or EPM, preoperatively submitted to our institution between January 2009 and July 2015, were included in this retrospective study. Exclusion criteria were incomplete image acquisition and previous cranial neurosurgery. Quantitative brain tumor textural information of a total of 40 patients was extracted from ADC-maps only. The ADC-map data were either acquired in each of our standard brain tumor protocols or from ADC-maps outside our institution. Therefore, except for one patient, all ADC-maps were, however, acquired on the scanners of the same manufacturer. All clinical diagnoses, which served as ground truth, were histologically certified.

Ethics

This retrospective study was performed conform to the Swiss Human Research Act and was approved by the Bernese Cantonal Ethics Committee (KEK-Berne, Switzerland).

Magnetic resonance imaging

Several different 1.5 T MR-scanners (Siemens Erlangen, Germany) from the same manufacturer have been used to record the apparent diffusion ADC-maps. The images were recorded typically on a 128 × 128 image *k*-space matrix and interpolated by zero-filling to a 256 × 256 image reconstruction matrix. The manufacturers' standard product EPI-pulse sequences with a typical TE = 89 ms and TR = 3000 ms was used. The slice thickness was 5.0 mm with a gap of 1.5 mm between the slices. The interpolated pixel size was typically in the order of 1.2 × 1.2 × 5.0 mm^3. Since data from various scanners and hospitals were over a time period of more than 10 years, the MR-acquisition parameters were not identical in all cases.

Extraction of image features

The CARD method which is used in this article is a radiomics variant (see e.g. Lambin et al.[11]), due to the fact that it combines *image feature extraction* with *machine learning*. Please refer to the Online Supplementary Material for more details on the method variant of this article, more specifically, how exactly the *image feature extraction* was performed.

RF classifiers in diagnostics

Classifiers can be viewed as algorithms that can decide between several possible diagnoses, that is, they return the most likely diagnosis class. The main requirement is that the patient data used for the training of the classifier are representative for *de novo* patients. In this article, RF classifiers as proposed by Breiman[10] were used. This type of classification has been used in a large number of studies performed in different fields of application[12,13] and carry a high prediction accuracy (see e.g., Breiman,[10] Liaw and Wiener,[14] Strobl et al.[15]). The RF consist of several hundred different decision trees. Each decision tree corresponds to a set of rules that leads for each feature set to a clear-cut diagnosis along with a probability measure (in case of equal probabilities for each diagnosis, the decision is taken at random). Each tree is trained on a different bootstrap sample of the training data.

During training, the split rules are optimized such that the training observations with known diagnosis label get most possible unmixed with regard to their diagnosis labels. If a new observation follows the decision rules and ends up in one leaf of the tree, then the probability for a certain diagnosis is given by the proportion of this diagnosis among all training observations corresponding to this leaf. That diagnosis which gets the majority of the votes of the RF-classifier is the predicted clinical diagnosis.

The diagnosis of the whole RF is derived by letting the trees majority prediction vote or by averaging the probabilities over all trees and pick the diagnosis with the highest probability. Aggregation of many independent and unbiased predictions yields in general a highly accurate prediction since the variance of the individual classifications is averaged out. The performance is given as out-of-bag (OOB) error, which corresponds to the misclassification rate of the classifier when each observation is only classified with that subset of trees that did not have this observation in the bootstrap training set. Therefore, we expect the OOB error rate to resemble the test error rate when classifying completely new patients with the RF.

Application of computer algorithms in neuroradiology to aid the neuroradiologist in finding the most likely diagnosis can be called CARD. A software was developed to make CARD practically feasible in a clinical setting offering the following functionality: (i) a simple way to create novel disease specific databases; (ii) add the relevant radiological data of patients for which histological asserted findings are available into these databases; (iii) automatic training of RF classifiers based on this data; and (iv) extracting the same type of image data for new, for which the diagnosis is unclear; (v) performing the diagnosis, based on a given *DD* formulated by the neuroradiologist. In Figure 2, the principle of CARD is displayed. A prototype software for these purposes was developed in our institute in the programming language JAVA (version 1.7), using the RCaller-class (version 2.0.7)[16] to enable the usage of the R-implementation Breimans' RF algorithm[10] to perform the classification.

Computed texture features

In the first step, the developed computer program computed for each ADC-map a total of 17 derived texture maps (see Figure 3 for an illustrative example for a PA). A graphical interpretation of the relationship between the original ADC-map and its associated texture parameter maps and

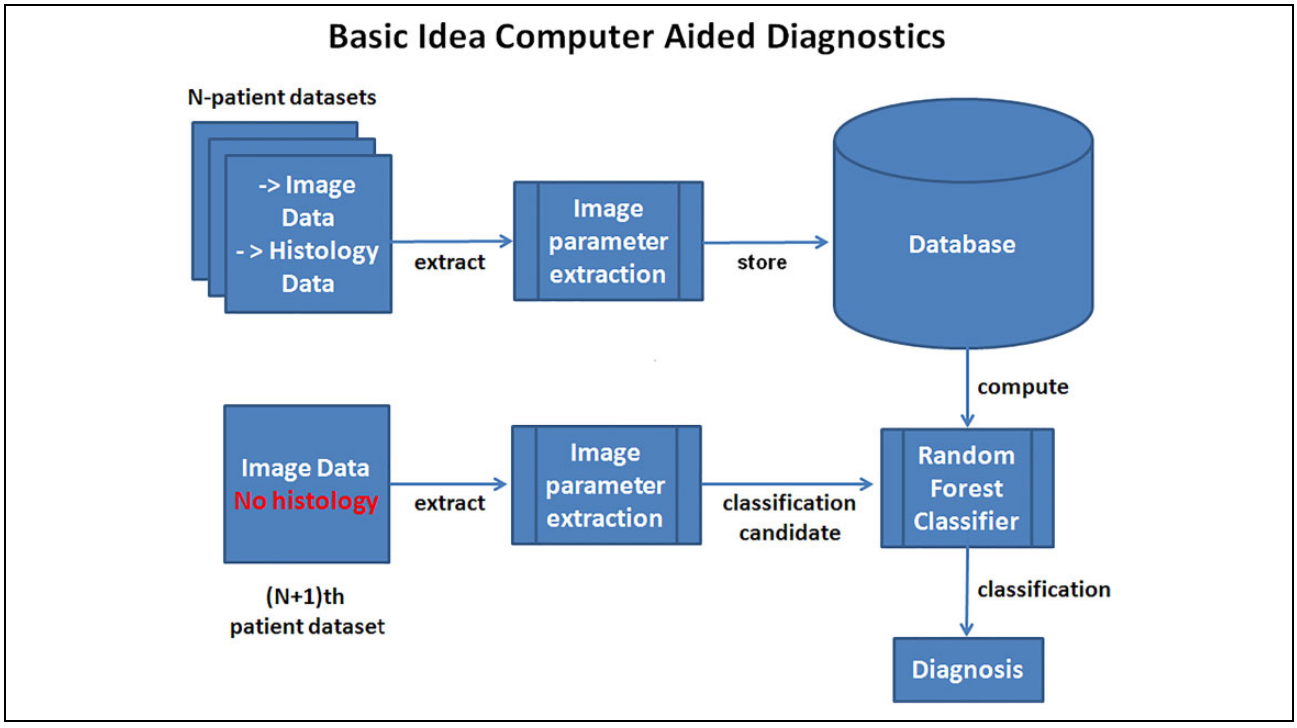

Figure 2. Principle of CARD. The extracted image and clinical data of *N* patients having histologically certified tumor diagnosis are stored into a database. With these data, RF classifiers can be computed for any given DD. For the $(N + 1)$th patient, for which only a DD can be defined, the same image-related parameters are extracted. With these data and the DD, the RF classifier computes the most likely diagnosis. CARD: computer-aided radiological diagnostics; DD: differential diagnosis; RF: random forests.

texture parameters is displayed in Figure 4. From these texture maps, a total of 94 texture parameters are computed. More details on the computation of the used texture parameters are given in the Online Supplementary Material.

Manual tumor segmentation and reproducibility of classifiers

Three independent raters segmented the tumor volume slice by slice in the ADC-maps of all 40 included PA, MDB, and EPM patients, by drawing manually contours that surround the *complete* tumor-affected tissue (i.e. solid parts *and* edema). Per slice, per contour, and per texture map 94 texture parameters are computed. The number of pixels within one contour defines the weighting factor for computation of the *averaged mean value* of the parameter over all slices. In this fashion, 94 texture parameters are obtained per patient to characterize the tumor. These values are the input features of the RF algorithm.

Measures for neuroradiological diagnostic performance

To investigate the diagnostic neuroradiological performance in our department, we retrospectively analyzed the neuroradiological diagnostic texts, stored in our institutes' Radiological Information System (RIS)-system for all study patients. All diagnostic texts in our institute are based upon the four-eyes principle, where a junior neuroradiologist and a senior neuroradiologist analyze all images of the examination. Additionally, the final diagnostic text is approved by a senior neuroradiologist.

To evaluate the diagnostic performance, we created a diagnostic score (DS). Three possible situations were discriminated: (1) the neuroradiological diagnosis was correct and *identical* with the histological and the best possible $DS = 100\%$ is assigned; (2) the *correct* DD was within the formulated stating N-possible DD set and therefore a $DS = (100/N)\%$ is assigned (see Tables 7 to 9 for all DDs mentioned). It should be noted that this definition possibly underestimates the true neuroradiological performance, since the order in which the DDs were formulated is not taken into account; and (3) the radiological (differential-) diagnosis was wrong and a $DS = 0\%$ is assigned. Finally, an average DS_{mean} was calculated per disease by the summation of all patient *DS* values and divided by the total number of patients in this group. The written neuroradiological diagnostic findings of patients who received their initial preoperative MRI scan in other institutions than ours were requested.

Results

Study population

A total of 40 patients were retrospectively analyzed. All patients received their preoperative MRI scan from January

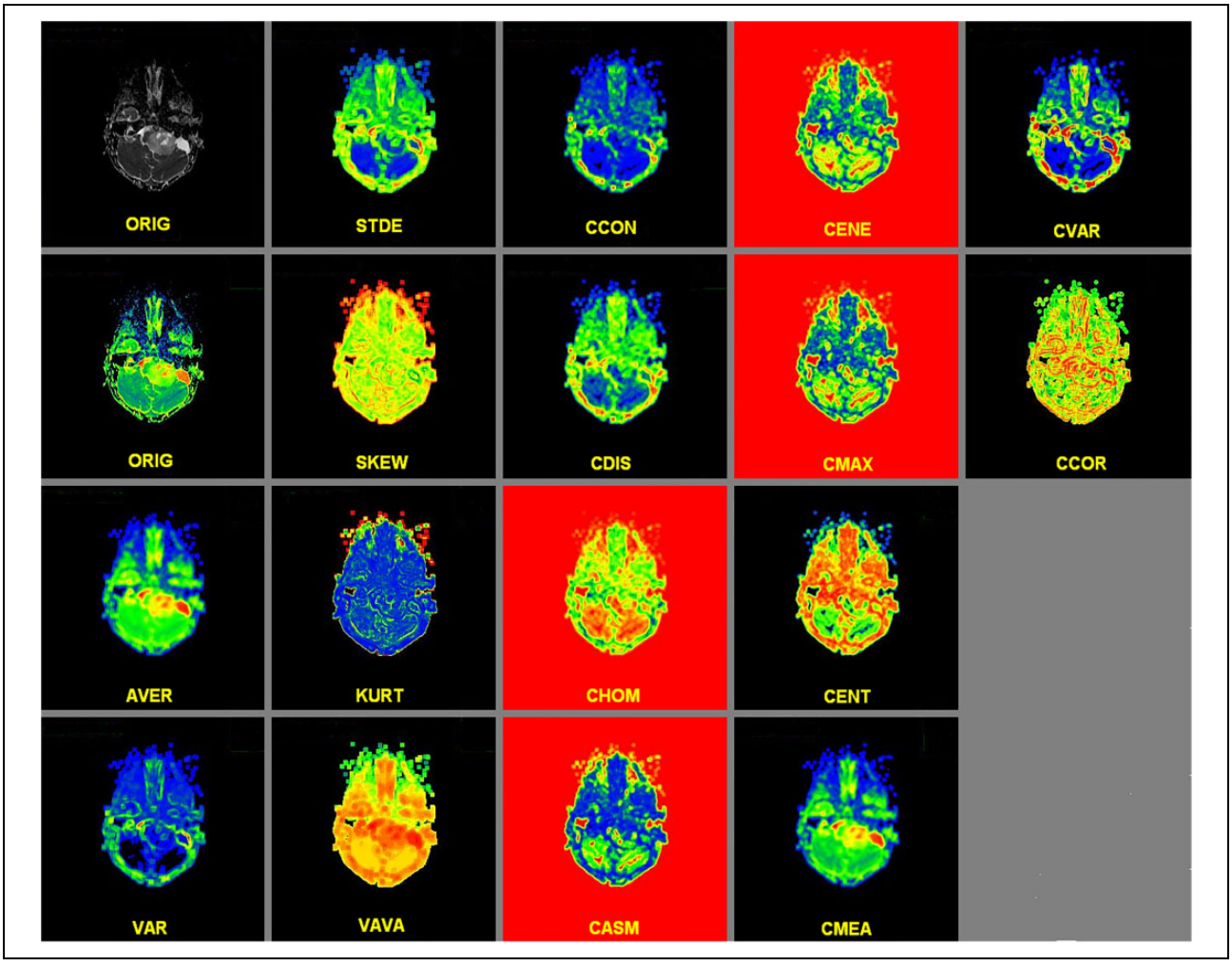

Figure 3. Texture maps of the ADC-values of a patient suffering from PA. The image shows the ADC-map itself, all 17 texture parameter maps that are output of the developed software. All maps were color-coded using a rainbow color map lookup table: highest values in red, lowest values in blue. PA: pilocytic astrocytoma; ADC: apparent diffusion coefficient.

2005 until July 2015. A total of 14 PAs, 16 MDBs, and 10 EPMs were included. Quantitative textural information was extracted from ADC-maps described above. For demographics, please see Table 1.

Comparison of average tumor ADC values

In addition to the textural analysis, the group-mean ADC values for the three different tumor types were computed and are listed in Table 2. For the three different tumor types, a significant difference between group-means was found regarding the tumor average ADC values. Equivalent values are reported by other authors in the past.[1,9,17]

To further assess where the ADC differences between the three tumor types arise, we used the Wilcoxon rank sum test for two-group comparisons. The Benjamini–Hochberg method to correct for multiple testing was used. Strong evidence for differences in ADC group-means when comparing PA versus MDB or EPM versus MDB with $p < 0.001$ (see Table 3) was found.

Intra- and inter-rater reliability

We base this reliability analysis on the two-group comparison of PA versus MDB.

Inter-rater variability. Three raters (R.1, R.2, and R.3) have segmented the tumors and used the RF classification for a CARD diagnosis. The Fleiss' kappa test used to quantify the agreement of the diagnosis results revealed a $\kappa = 0.821$ with $z = 7.79$, and a p value $= 6.88 \times 10^{-15}$ (Subjects $=$ 30). For a graphical interpretation of the classifier performance as a function of the rater, the confusion matrix is displayed in Table 4.

Intra-rater variability. One rater has segmented the tumors on three different days one time and used the RF classification of each segmentation to obtain three times a separate CARD diagnosis for the same case. The Fleiss' kappa test used to quantify the agreement of the diagnosis results revealed a $\kappa = 0.822$ with $z = 7.15$ and p value $= 6.22$

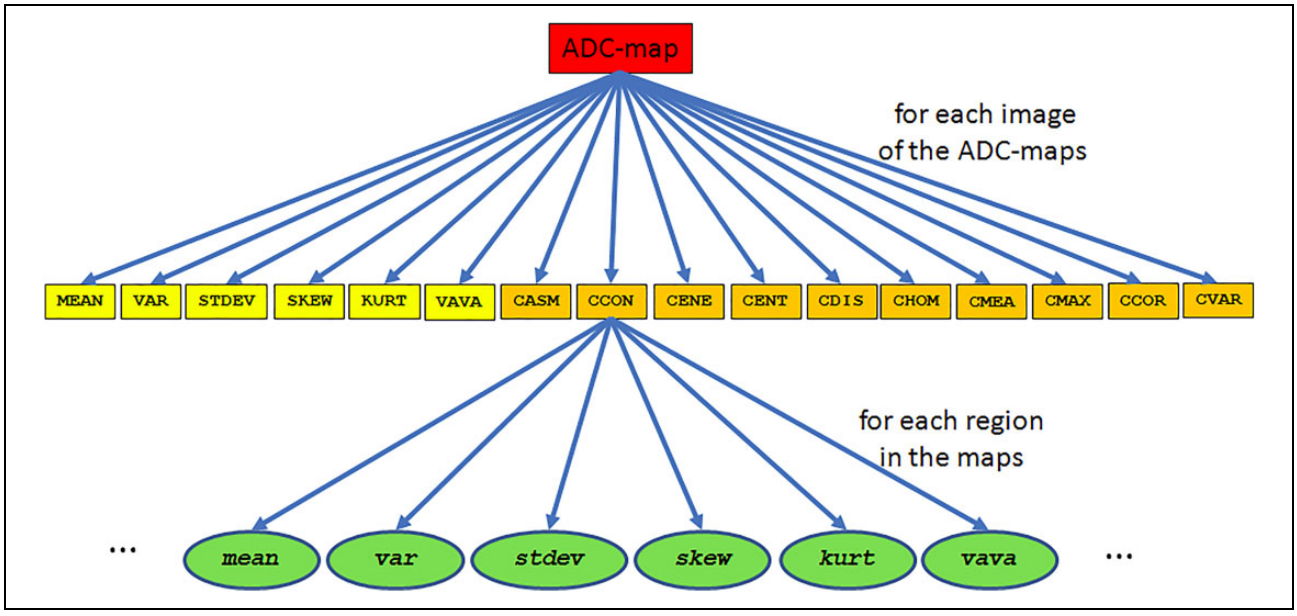

Figure 4. Graphical model of the relationship between the measured ADC-map (red box, also denoted as ORIG), the texture parameter maps computed from it (yellow boxes), and finally the texture parameters (green ellipsoid) that were taken as feature inputs of the RF classification algorithm. ADC: apparent diffusion coefficient; RF: random forest.

Table 1. Demographical data: sex (male:female), and mean age in years ± standard error of patients with different tumor types.

Feature	PA	EPM	MDB
Sex (M:F)	5:9	5:6	10:6
Age in years (mean ± std error)	9.1 ± 4.1	45.2 ± 15.5	13.1 ± 7.5

PA: pilocytic astrocytoma; MDB: medulloblastoma; EPM: ependymoma.

$\times 10^{-15}$ (subjects = 30). The confusion matrix is also given in Table 4.

From this, we can conclude that the inter-rater as well as the intra-rater agreement is very high.

Classification performance

The overall classification-error rate performance for five times repetitive contouring (thus averaging over inter- and intra-rater results) was 11.3 ± 2.7%. The average sensitivity was 0.888 ± 0.031 and the average specificity was 0.886 ± 0.036 and together with the individual scores they are listed in Table 5.

Diagnostic performance of the RF-based CARD method

In Tables 5 and 6, the performance of the classifiers to distinguish between MDBs, PAs, and EPM are

- For the MDB versus PA DD, an average *sensitivity* of 0.888 ± 0.031, an average *specificity* of 0.886 ± 0.036, and an *accuracy* of 88.0 ± 3.0% was obtained, with an OOB error rate of 11.3 ± 2.7% (Table 5). The intra-rater variability alone was additionally computed, and a mean sensitivity of 0.896 ± 0.042 was obtained; for inter-rater variability of the sensitivity, a value 0.882 ± 0.036 was found. For the intra-rater variability of the specificity, a value of 0.929 ± 0.041, and for inter-rater specificity, a value of 0.893 ± 0.040 was found.
- The EPM versus MDB DD was not part of the reproducibility study, and the classifier was determined only once. A *sensitivity* of 0.938 with 95% CI range of (0.677, 0.996), and a *specificity* of 0.70 with 95% CI range of (0.353, 0.919), the mean *accuracy* was 84.6% and the mean OOB error rate was 15.38% (Table 6).
- Also for the PA versus EPM, DD was not part of the reproducibility study but the performance of the classifier was determined. Here a *sensitivity* of 0.786 with 95% CI (0.488, 0.943), specificity was only 0.100 with a 95% CI (0.005, 0.459) (see Table 6), and the mean accuracy as well as the OOB error were both 50%.

Clinical differential diagnostic performance

Tables 7 to 9 list the diagnostic performance of neuroradiologist to diagnose the tumor types correctly. The neuroradiologists had access to all multiparametric images of the complete MR-examination (including at least T_1, $T_{1,c}$, T_2, FLAIR, and ADC imaging). However, the neuroradiologists had to decide between all possible tumor types, which is a much more demanding task than CARD as

Table 2. ADC values (in 10^{-6} mm^2/s) information grouped by disease type and a *p* value corresponding to an ANOVA testing for differences between all three groups.

ADC	PA	EPM	MDB	ANOVA *p* value
Mean $\pm$ standard error	1575.5 $\pm$ 326.7	1433.3 $\pm$ 285.3	910.5 $\pm$ 131.9	2.74×10^{-5}
Range (min–max)	751.1 – 2319.1	938.8 – 2244.3	572.9 – 1264.7	

ADC: apparent diffusion coefficient; PA: pilocytic astrocytoma; MDB: medulloblastoma; EPM: ependymoma; ANOVA: analysis of variance.

Table 3. Results of two-group comparison of ADC values given as multiple testing corrected *p* values from Wilcoxon rank sum tests.

Differential diagnosis	*p* value
PA versus MDB	0.000028
EPM versus MDB	0.00021
EPM versus PA	0.28468

ADC: apparent diffusion coefficient; PA: pilocytic astrocytoma; MDB: medulloblastoma; EPM: ependymoma.

described here, namely to decide between given disease alternatives.

The best performance was obtained for the diagnosis of MDB, for which in 31.25% of the cases there was a single correct diagnosis, and in 37.5% of the cases a correct DD was formulated: therefore in a total of 68.75% of the cases, the diagnosis contained MDB as alternative. The DD score DS_{mean} was 46.9%, weighting the DS with the amount of differentials stated by the neuroradiologist (see the definition above).

Second best diagnostic performance was obtained for PA, with only 14.3% correct diagnoses and 50.0% stating the correct DD (making 64.3% correct or correct DD). However for PA, a DD score DS_{mean} of 30.5% was obtained.

For radiologists, the most difficult diagnosis seems to be that of EPM. Here only a correct radiological diagnosis score of 10% was obtained and in 20%, a correct DD. This makes a total of only 30% for correct or correct DD. With a DD score DS_{mean} of just 17.6%, it is clear that diagnosis of this type of brain tumor is the most difficult to diagnose out of the three types examined in this article.

Discussion

Observed ADC values

We have determined the ADC-mean values and standard deviations for PA, EPM, and MDB and obtained results which are comparable to the ADC values published earlier in a pediatric cohort study and other studies[2,9,1,18] despite the high age variation in our group. MDB showed significantly lower ADC values than PA and EPM, whereas EPM and PA showed similar ADC values not finding evidence for significant different mean values.

Clinical routine tumor DD performance

Tables 7 to 9 give insight into the performance of tumor diagnostics in daily routine. Neuroradiologists give infrequently preoperative clear-cut diagnosis. In our study with 40 patients, correct and clear-cut diagnosis was given for 31% of 16 MDB tumors, 14.3% of 14 PA tumors, and 10% of 10 EPM tumors. Most of the diagnosis comprises several tumor types and together with the clear-cut diagnosis the neuroradiologists diagnosis contained the right tumor in 69% of true MDB, 64% of true PA, and 30% of true EPM. DD scores DS_{mean} for the three tumor types were determined and are listed in Table 6. With a DS_{mean} of 57% for MDBs, this tumor type was best distinguished and most frequently correct, on average the DD contained less than two alternatives. For PAs, on average nearly three possible tumor types were formulated in the DDs, and for EPMs even more than three tumor type differentials were formulated. These numbers show that visual discrimination of these tumor entities is an extremely challenging task for the human visual system, even having access to multiple MR modalities beside the ADC-maps.

Computer-aided radiological diagnostics

It should be noted that the CARD-algorithms starts with a DD formulated by a neuroradiologist. For the clinical important DDs of MDB versus PA and EPM versus MDB, useful RF classifiers could be developed. For the DD between MDB and PA, the best performance was observed, namely an average *sensitivity* of 0.888 $\pm$ 0.031 and specificity of 0.886 $\pm$ 0.036 with an average classifier OOB error of 11.3 $\pm$ 2.7%. This means that for this DD, application of CARD could substantially improve radiological diagnostic quality. The DD between EPM and MDB, however, was little less performing, with a sensitivity of 0.938 and specificity of 0.700 together with accuracy of 84.6%.

For the DD between PA and EPM, which has a sensitivity of 78.6%, a specificity of only 10% was found. With such a poor performance, it is clear that such a classifier cannot be used in practice. One has to conclude from this that PA and EPM do not differ sufficiently in ADC-heterogeneity parameters in such a way that they could be used to distinguish these diseases from each other's in a meaningful way. A substantial improvement is expected in case more data and additional image series of the MRI-examination (e.g. perfusion imaging) are provided for the RF training and classification.

Table 4. Confusion matrix for the inter-rater and intra-rater reproducibility of the CARD method applied to the DD of MDB versus PA.

	Inter-rater comparison				Intra-rater comparison		
Patient number	Rater 1	Rater 2	Rater 3	Ground truth	First	Second	Third
1	MDB	MDB	MDB	MDB	MDB	MDB	MDB
2	MDB	MDB	PA	MDB	MDB	PA	MDB
3	MDB	MDB	MDB	MDB	MDB	MDB	MDB
4	MDB	MDB	MDB	MDB	MDB	MDB	MDB
5	MDB	MDB	MDB	MDB	MDB	MDB	MDB
6	MDB	MDB	MDB	MDB	MDB	MDB	MDB
7	MDB	PA	PA	MDB	MDB	PA	MDB
8	MDB	MDB	MDB	MDB	MDB	MDB	MDB
9	MDB	MDB	MDB	MDB	MDB	MDB	MDB
10	MDB	MDB	MDB	MDB	MDB	MDB	MDB
11	MDB	MDB	MDB	MDB	MDB	MDB	MDB
12	MDB	MDB	MDB	MDB	MDB	MDB	MDB
13	MDB	MDB	MDB	MDB	MDB	MDB	MDB
14	PA	PA	PA	MDB	PA	PA	PA
15	MDB	MDB	MDB	MDB	MDB	MDB	MDB
16	MDB	MDB	MDB	MDB	MDB	MDB	MDB
17	PA	PA	PA	PA	PA	PA	PA
18	PA	PA	PA	PA	PA	MDB	PA
19	PA	PA	PA	PA	PA	PA	PA
20	MDB	MDB	MDB	PA	PA	MDB	MDB
21	PA	MDB	MDB	PA	PA	PA	PA
22	PA	PA	PA	PA	PA	PA	PA
23	PA	MDB	MDB	PA	PA	PA	PA
24	PA	PA	PA	PA	PA	PA	PA
25	PA	PA	PA	PA	PA	PA	PA
26	PA	PA	PA	PA	PA	PA	PA
27	PA	PA	PA	PA	PA	PA	PA
28	PA	PA	PA	PA	PA	PA	PA
29	PA	PA	PA	PA	PA	PA	PA
30	PA	PA	PA	PA	PA	PA	PA

CARD: computer-aided radiological diagnostics; PA: pilocytic astrocytoma; MDB: medulloblastoma; EPM: ependymoma.

Table 5. The variability on the classification performance parameters (sensitivity, specificity with their CI boundaries CI-min and CI-max) due to inter-rater and intra-rater differences in contouring was examined for the DD of MBD versus PA.[a]

DD, MDB versus PA	R1.1	R1.2	R1.3	R2	R3	Overall classifier performance (Average $\pm$ Standard error)
Sensitivity	0.938	0.813	0.938	0.938	0.813	0.888 $\pm$ 0.031
CI-min	0.677	0.537	0.667	0.677	0.537	0.619 $\pm$ 0.033
CI-max	0.997	0.950	0.997	0.997	0.950	0.978 $\pm$ 0.012
Specificity	1.000	0.857	0.929	0.786	0.857	0.886 $\pm$ 0.036
CI-min	0.699	0.562	0.642	0.488	0.562	0.591 $\pm$ 0.036
CI-max	1.000	0.950	0.996	0.943	0.975	0.973 $\pm$ 0.012
OOB estimate of error rate (%)	3.33	6.67	16.67	13.3	16.7	11.3 $\pm$ 2.7

CI: confidence interval; DD: differential diagnosis; PA: pilocytic astrocytoma; MDB: medulloblastoma; OOB: out-of-box; CI-min: minimal 95% confidence range; CI-max: maximal 95% confidence range.

[a]R1.1, R1.2, and R1.3 refer to one single rater doing three different evaluations and reveals information on the intra-rater variability. Raters R1, R2, and R3 are three different independent raters and reveals information on the inter-rater variability. The errors indicated are standard errors.

Dependency of CARD results on rater segmentation

Since the CARD method depends on manual segmentation of the *complete* tumor-affected area, the classification performance could, in principle, strongly depend on the individual segmentation of each individual rater. Therefore, a reliability analysis was performed to investigate the *reproducibility* of the CARD diagnosis. For the inter- and intra-rater reproducibility, a Fleiss' kappa test value of $\kappa = 0.821$, 0.822, respectively, was found, which means that there is a very high agreement in obtained classifiers.[19] Since the inter-rater and intra-rater reproducibility seems to

Table 6. RF classifier performance of the classifiers for the DD of PA versus MDB and EPM versus PA.[a]

DD	Value	CI-min	CI-max	Accuracy
EPM versus MDB				
Sensitivity	0.938	0.677	0.997	
Specificity	0.700	0.354	0.919	
OOB estimate of error rate				15.38%
Accuracy				84.6%
EPM versus PA				
Sensitivity	0.786	0.488	0.943	
Specificity	0.100	0.005	0.459	
OOB estimate of error rate				50%
Accuracy				50%

CI: confidence interval; DD: differential diagnosis; PA: pilocytic astrocytoma; MDB: medulloblastoma; OOB: out-of-box; RF: Random Forest; CI-min: minimal 95% confidence range; CI-max: maximal 95% confidence range.

[a]The best classification performance obtained is indicated. Apart from the sensitivity and specificity, the OOB error rate estimates are indicated.

Table 7. Histological diagnosis compared to the radiological diagnosis for PA.

PA	Absolute patient numbers	Percentage (%)
Total number of patients	14	100
Correct radiological diagnosis	2	14.3
Correct differential radiological diagnosis	7	50.0
Wrong radiological diagnosis	5	35.7
DD score DS_{mean}		30.5%
Posed radiological DDs by tumor type	**Number of times**	**Rel. % of DD**
PA	9	26.5
Glioma WHO Grade-II	4	11.8
Glioma WHO Grade-III	3	8.8
Glioma WHO Grade-IV	2	5.9
Craniopharyngioma	1	2.9
Germinoma	1	2.9
Schwannoma	1	2.9
PNET	1	2.9
MDB	1	2.9
EPM	2	5.9
Hemangioblastoma	1	2.9
Neurinoma	1	2.9
MDB	2	5.9
Neurofibromatosis	1	2.9
Hippel–Lindau syndrome	1	2.9
Epidermoid	1	2.9
Teratoma	1	2.9
Unknown lesion	1	2.9
Total number of differential diagnoses	34	100

PA: pilocytic astrocytoma; MDB: medulloblastoma; EPM: ependymoma; DD: differential diagnosis; DS: diagnostic score; PNET: Primitive NeuroEctodermal Tumor; WHO: World Health Organization.

Table 8. Histological diagnosis compared to the radiological diagnosis for EPM.

EPM	Absolute patient numbers	Percentage (%)
Total number of patients	10	100
Correct radiological diagnosis	1	10.0
Correct radiological differential diagnosis	2	20.0
Wrong radiological diagnosis	7	70.0
DD score DS_{mean}	24 differentials	17.6
Posed radiological DD by tumor types	**Number of times**	**Rel. % of DD**
EPM	3	12.5
MDB	1	4.2
PA	5	20.8
Glioma WHO Grade-II	1	4.2
Glioma WHO Grade-III	2	8.3
Glioma WHO Grade-IV	2	8.3
Subependymal giant cell astrocytoma	1	4.2
Plexus choroid papilloma	2	8.3
Ganglioglioma	1	4.2
Hemangioblastoma	1	4.2
Metastasis	3	12.5
Meningioma	2	8.3
Total number of differentials	24	100

PA: pilocytic astrocytoma; MDB: medulloblastoma; EPM: ependymoma; DD: differential diagnosis; DS: diagnostic score.

Table 9. Histological diagnosis compared to the radiological diagnosis for MDB.

MDB	Absolute patient numbers	Percentage (%)
Total number of patients	16	100
Correct radiological diagnosis	5	31.25
Correct differential radiological diagnosis	6	37.50
Wrong radiological diagnosis	5	31.25
DD score DS_{mean}		46.9
Posed radiological differential diagnoses by tumor type	**Number of cases**	**Rel.% of DDs**
MDB	11	31.4
PNET	3	8.6
PA	5	14.3
EPM	5	14.3
Meningioma	2	5.7
Lymphoma	2	5.7
Plexus papilloma	1	2.9
Metastasis	3	8.6
Glioma WHO-III	1	2.9
Glioma WHO-IV	1	2.9
Unknown lesion	1	2.9
Total number of differentials	35	100

PA: pilocytic astrocytoma; MDB: medulloblastoma; EPM: ependymoma; DD: differential diagnosis; DS: diagnostic score.

be nearly identical, one may conclude that small differences in the human contoured tumor volumes do not really affect classification performance. Additionally *one* rater can reproduce its contouring as good as *different* raters among each other's. It can be concluded that the task to segment to "whole tumor affected volume" can be reproduced to a high level of agreement.

Diagnostic performance of machine versus man

For the presented CARD method, it is not possible to directly compare the diagnostic performance of the machine to the human diagnostic performance. We want to emphasize that the success of *any* CARD method depends on the posed DD by the neuroradiologist, that is, a DD in which the correct diagnosis needs to belong to the solution set. The performance of the RF is quantified by the probability to find the correct diagnosis within a DD. This probability is given by one minus the OOB-class error. However, in practice, this performance needs to be down-corrected with the *probability* of a correctly posed DD by the neuroradiologist (see Tables 7 to 9). Moreover, neuroradiologists frequently indicate in their DD more than two options (DS_{mean} takes into account the number of options). In these cases, several trained RFs need to be applied to determine which tumor we are dealing with.

Practical benefit of CARD

In practice, knowing the tumor type is critical for the therapeutic option stratification, for example, deciding and planning the extent of resection. However, based on MRI scans, neuroradiologists can often give only a set of possible tumor types. In such cases, the presented CARD results are valuable. Since the presented semiautomatic statistical method for CARD can be performed time-efficiently, it is feasible for the usage in clinical practice. The accuracy of the method can be further improved by also taking other modalities into account, for example, textural information derived from FLAIR or T_{1c}-imaging. Combination with complete automatic segmentation algorithms like Porz et al.[20] could improve the reproducibility of the method further.

Limitations

Our patient collective is relatively small for further subclassification or investigation of genetic differences. Again one should realize that a classifier can only be used for the purpose it has been trained. The classifier can only provide more confidence about the diagnosis, if the DD includes the correct diagnosis. It should be noted that this is not only the case for the proposed CARD but is true for all machine learning-based classifiers.

Conclusions

Reproducible and highly accurate, sensitive and specific classifiers for CARD can be obtained by feeding texture parameters extracted of ADC-maps only into an RF classification algorithm for deciding which brain tumor is most likely in the DDs of PA versus MDB or MDB versus EPM. For the DD of PA versus EPM, the classifiers were less accurate, specific, and sensitive but still can be used to improve the clinical neuroradiological diagnostics. Interesting aspect of the presented CARD method is the fact that all data of the past can simply be used to enhance the diagnostics of future patients.

Authors' note

Nicole Porz and Urspeter Knecht are equally contributing first authors.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed the receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the Swiss National Foundation, grant number 140958.

ORCID iD

Urspeter Knecht http://orcid.org/0000-0003-2181-4508
Johannes Slotboom http://orcid.org/0000-0001-5121-9852

Supplementary material

Supplementary material for this article is available online.

References

1. Rumboldt Z, Camacho DLA, Lake D, et al. Apparent diffusion coefficients for differentiation of cerebellar tumors in children. *Am J Neuroradiol* 2006; 27(6): 1362–1369.
2. Schneider JF, Confort-Gouny S, Viola A, et al. Multiparametric differentiation of posterior fossa tumors in children using diffusion-weighted imaging and short echo-time 1H-MR spectroscopy. *J Magn Reson Imaging* 2007; 26(6): 1390–1398. doi:10.1002/jmri.21185.
3. Yamashita Y, Kumabe T, Higano S, et al. Minimum apparent diffusion coefficient is significantly correlated with cellularity in medulloblastomas. *Neurol Res* 2009; 31(9): 940–946. doi:10.1179/174313209X382520.
4. Jaremko JL, Jans LBO, Coleman LT, et al. Value and limitations of diffusion-weighted imaging in grading and diagnosis of pediatric posterior fossa tumors. *Am J Neuroradiol* 2010; 31(9): 794–800.
5. Gimi B, Cederberg K, Derinkuyu B, et al. Utility of apparent diffusion coefficient ratios in distinguishing common pediatric cerebellar tumors. *Acad Radiol* 2012; 19(7): 794–800. doi:10.1016/j.acra.2012.03.004.

6. Bull J, Saunders D and Clark C. Discrimination of paediatric brain tumours using apparent diffusion coefficient histograms. *Eur Radiol* 2012; 22: 447–457. http://discovery.ucl.ac.uk/1368038/ (accessed 26 April 2018).
7. Koral K, Zhang S, Gargan L, et al. Diffusion MRI improves the accuracy of preoperative diagnosis of common pediatric cerebellar tumors among reviewers with different experience levels. *Am J Neuroradiol* 2013; 34(12): 2360–2365. doi:10.3174/ajnr.A3596.
8. Pierce TT and Provenzale JM. Evaluation of apparent diffusion coefficient thresholds for diagnosis of medulloblastoma using diffusion-weighted imaging. *Neuroradiol J* 2014; 27(1): 63–74. doi:10.15274/NRJ-2014-10007.
9. Gutierrez DR, Awwad A, Meijer L, et al. Metrics and textural features of MRI diffusion to improve classification of pediatric posterior fossa tumors. *Am J Neuroradiol* 2014; 35(5): 1009–1015. doi:10.3174/ajnr.A3784.
10. Breiman L. Random forests. *Mach Learn* 2001; 45(1): 5–32. doi:10.1023/A:1010933404324.
11. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 2012; 48(4): 441–446. doi:10.1016/j.ejca.2011.11.036.
12. Cutler DR, Edwards TC, Beard KH, et al. Random forests for classification in ecology. *Ecology* 2007; 88(11): 2783–2792.
13. Díaz-Uriarte R and De Andrés SA. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 2006; 7(1): 3. doi:10.1186/1471-2105-7-3.
14. Liaw A and Wiener M. Classification and regression by randomforest. *R News* 2002; 2(3): 18–22. https://www.researchgate.net/profile/Andy_Liaw/publication/228451484_Classification_and_Regression_by_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression-by-RandomForest.pdf (accessed 26 April 2018).
15. Strobl C, Boulesteix AL, Zeileis A, et al. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 2007; 8(1): 25. doi:10.1186/1471-2105-8-25.
16. Satman MH. RCaller. (2017). https://github.com/jbytecode/rcaller (accessed 07 February 2018).
17. Pierce T, Kranz PG, Roth C, et al. Use of apparent diffusion coefficient values for diagnosis of pediatric posterior fossa tumors. *Neuroradiol J* 2014; 27(2): 233–244. doi:10.15274/NRJ-2014-10027.
18. Yamasaki F, Kurisu K, Satoh K, et al. Apparent diffusion coefficient of human brain tumors at MR imaging. *Radiology* 2005; 235(3): 985–991. doi:10.1148/radiol.2353031338.
19. Viera AJ and Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005; 37(5): 360–363.
20. Porz N, Bauer S, Pica A, et al. Multi-modal glioblastoma segmentation: man versus machine. *PLoS One* 2014; 9(5): e96873. doi:10.1371/journal.pone.0096873.