Michael Zhang, MD[‡§]
Elizabeth Tong, MD[§]
Forrest Hamrick, BS[¶]
Edward H. Lee, PhD[§]
Lydia T. Tam, BS[‖]
Courtney Pendleton, MD[#]
Brandon W. Smith, MD[#]
Nicholas F. Hug, BA[‖]
Sandip Biswal, MD[§]
Jayne Seekins, DO[§]
Sarah A. Mattonen, PhD[**]
Sandy Napel, PhD[§]
Cynthia J. Campen, MD[‡‡]
Robert J. Spinner, MD[#]
Kristen W. Yeom, MD[§]
Thomas J. Wilson, MD[‡*]
Mark A. Mahan, MD[¶*]

[‡]Department of Neurosurgery, Stanford University, Stanford, California, USA; [§]Department of Radiology, Stanford University, Stanford, California, USA; [¶]Department of Neurosurgery, Clinical Neurosciences Center, University of Utah, Salt Lake City, Utah, USA; [‖]Stanford School of Medicine, Stanford University, Stanford, California, USA; [#]Department of Neurosurgery, Mayo Clinic, Rochester, Minnesota, USA; [**]Department of Medical Biophysics, Western University, London, Canada; [‡‡]Department of Neurology and Neurological Sciences, Stanford University, Stanford, California, USA

*Thomas J. Wilson and Mark A. Mahan contributed equally to this work.

Correspondence:
Mark A. Mahan, MD,
Department of Neurosurgery,
Clinical Neurosciences Center,
University of Utah,
175 North Medical Dr E,
Salt Lake City, UT 84106, USA.
Email: neuropub@hsc.utah.edu

Received, September 29, 2020.
Accepted, April 27, 2021.
Published Online, June 15, 2021.

Machine-Learning Approach to Differentiation of Benign and Malignant Peripheral Nerve Sheath Tumors: A Multicenter Study

BACKGROUND: Clinicoradiologic differentiation between benign and malignant peripheral nerve sheath tumors (PNSTs) has important management implications.
OBJECTIVE: To develop and evaluate machine-learning approaches to differentiate benign from malignant PNSTs.
METHODS: We identified PNSTs treated at 3 institutions and extracted high-dimensional radiomics features from gadolinium-enhanced, T1-weighted magnetic resonance imaging (MRI) sequences. Training and test sets were selected randomly in a 70:30 ratio. A total of 900 image features were automatically extracted using the PyRadiomics package from Quantitative Imaging Feature Pipeline. Clinical data including age, sex, neurogenetic syndrome presence, spontaneous pain, and motor deficit were also incorporated. Features were selected using sparse regression analysis and retained features were further refined by gradient boost modeling to optimize the area under the curve (AUC) for diagnosis. We evaluated the performance of radiomics-based classifiers with and without clinical features and compared performance against human readers.
RESULTS: A total of 95 malignant and 171 benign PNSTs were included. The final classifier model included 21 imaging and clinical features. Sensitivity, specificity, and AUC of 0.676, 0.882, and 0.845, respectively, were achieved on the test set. Using imaging and clinical features, human experts collectively achieved sensitivity, specificity, and AUC of 0.786, 0.431, and 0.624, respectively. The AUC of the classifier was statistically better than expert humans ($P = .002$). Expert humans were not statistically better than the no-information rate, whereas the classifier was ($P = .001$).
CONCLUSION: Radiomics-based machine learning using routine MRI sequences and clinical features can aid in evaluation of PNSTs. Further improvement may be achieved by incorporating additional imaging sequences and clinical variables into future models.

KEY WORDS: Machine learning, Magnetic resonance imaging, Malignant peripheral nerve sheath tumor, Peripheral nerve sheath tumor, Radiomics, Sensitivity, Specificity

Neurosurgery 89:509–517, 2021 https://doi.org/10.1093/neuros/nyab212 www.neurosurgery-online.com

Accurate differentiation between benign (BPNST) and malignant (MPNST) peripheral nerve sheath tumors is critical because the treatment paradigms differ greatly. BPNSTs are often managed nonoperatively with serial imaging, whereas MPNSTs are rapidly progressive and require aggressive, multidisciplinary treatment. Even with aggressive treatment, the 5-yr survival rate for patients with MPNSTs is ~18% to 50%.[1] Early identification and appropriate management offers the best chance of survival.

ABBREVIATIONS: BPNST, benign peripheral nerve sheath tumor; **gad,** gadolinium; **GLCM,** Gray-Level Co-occurrence Matrix; **GLDM,** Gray-Level Dependence Matrix; **GLRLM,** Gray-Level Run-Length Matrix; **GLSZM,** Gray-Level Size Zone Matrix; **LASSO,** least absolute shrinkage and selection operator; **MPNST,** malignant peripheral nerve sheath tumor; **NGTDM,** Neighboring Gray-Tone Difference Matrix; **NIR,** no-information rate; **PNST,** peripheral nerve sheath tumor; **QIFP,** Quantitative Imaging Feature Pipeline

Supplemental digital content is available for this article at www.neurosurgery-online.com.

Differentiation between BPNSTs and MPNSTs with conventional imaging and clinical data remains error-prone.[1-3] 18-Fluoro-deoxyglucose positron emission tomography (FDG-PET) and percutaneous biopsy are alternatives for diagnosis but are also beset with potential risks and diagnostic imprecision.[4-8] Thus, there remains a substantial need for improved discrimination between BPNSTs and MPNSTs using noninvasive modalities.

Machine-learning approaches to image analysis can add quantitative insights to existing qualitative interpretation. Radiomics evaluates at a voxel level to identify significant quantitative image features that can be used to develop artificial intelligence-based prediction models. The availability of large digital image data offers the potential to clinically translate radiomics techniques, which have been successfully applied to other tumors to aid diagnosis.[9-13] The same approach could provide another lens for distinguishing BPNSTs from MPNSTs.

We used a multi-institutional cohort of patients with PNSTs to develop and evaluate radiomics-based classifiers to distinguish between benign and malignant lesions using basic clinical data and conventional imaging.

METHODS

Study Population

Patients with PNSTs were identified at 3 participating institutions. The surgical pathology was used as ground truth. Exclusion criteria were lack of preoperative magnetic resonance imaging (MRI) and poor-quality, nondiagnostic MRI. For PNST imaging that passed quality control, axial T1-weighted, gadolinium-enhanced MRI (T1-gad) was identified as the most commonly acquired imaging across the participating centers. Many available studies either lacked T2-weighted images or, if these were available, they comprised heterogeneous T2-weighted imaging protocols (eg, T2 short tau inversion recovery, T2 fast spin echo, and T2 iterative decomposition of water and fat with echo asymmetry and least-squares estimation) and imaging planes (sagittal, coronal, or axial). Features used for MRI and gradient boost modeling are included in the **Supplemental Digital Content**. Thus, T1-gad was chosen for computational image feature extraction and machine-learning model development (see **Table S1** in **Supplemental Digital Content**).

Patient demographic and clinical variables were abstracted via chart review. The institutional review boards at all 3 institutions approved the study, with waiver of consent. The report was prepared according to the Strengthening the Reporting of Observational Studies in Epidemiology guidelines.

Clinical Variables

Clinical variables abstracted for analysis included age at operation, sex, neurogenetic diagnosis (NF1 or NF2 or schwannomatosis), and presence of spontaneous pain or preoperative motor deficit.

Image Segmentation, Preprocessing, and Feature Extraction

The volumetric regions of interest for each PNST were delineated and verified by 2 board-certified neuroradiologists using ITK-SNAP (University of Pennsylvania, Philadelphia, Pennsylvania). A total of 900 image features were automatically extracted using the PyRadiomics package on the Quantitative Imaging Feature Pipeline (QIFP),[14] including first-order statistics, 2D/3D Shape, Gray-Level Co-occurrence Matrix (GLCM), Gray-Level Run-Length Matrix (GLRLM), Gray-Level Size Zone Matrix (GLSZM), Neighboring Gray-Tone Difference Matrix (NGTDM), and Gray-Level Dependence Matrix (GLDM), as defined by the Imaging Biomarker Standardization Initiative (see **Table S2** in **Supplemental Digital Content**).[15] First-order features depend upon the individual value of voxels, without spatial relationship to other voxels in the image. Matrix evaluations consider spatial relationships between voxels. MRI studies were normalized for voxel size ($1 \times 1 \times 1$ mm) and intensity (scale factor of 100). A fixed bin width (10) was used for gray-value discretization. Preprocessing filters included wavelet (8 coefficients) and Laplacian of Gaussian (3 sigma). Feature extraction was calculated for classes including first-order statistics and gray-level derivatives.

Feature Selection and Validation

Patients were randomly allocated into training and test sets in a 70:30 ratio. Feature selection for the allocated training set was performed using sparse regression analysis by a least absolute shrinkage and selection operator (LASSO), performed with 10-fold cross-validation and repeated for 1000 cycles. The mean squared error was calculated for 100 lambdas in each cycle. The optimal lambda was identified as the lowest mean squared error value and used for feature reduction and coefficient calculations. Both radiologic and clinical variables were incorporated into the primary model. Selected features represented in >80% of the cycles were retained for subsequent classifier optimization.

Retained features were further refined by gradient boost modeling using the caret package in R.[16] Training was performed with 10-fold cross-validation, repeated for 3 cycles. Final tuning was performed for interactions, tree depth, minimal terminal node size, and shrinkage (see **Figure S1** in **Supplemental Digital Content**). The final radiomic classifier was guided by maximizing the area under the curve (AUC). The same process was repeated to generate separate secondary models using only the imaging features or only the clinical variables. All classifiers were then applied to the test cohort, and the predicted pathology was evaluated against the pathological diagnoses. The relative influence of the clinical and radiologic features was calculated as described previously.[17] All modeling was performed using RStudio version 1.2.5033 (PBC, Boston, Massachusetts).

Test Set Evaluation by Human Evaluators

For comparison, the same test set was evaluated by human readers, who were provided the T1-gad images plus T2- or proton-density-weighted images when available and asked to classify the tumor as a BPNST or MPNST. The evaluators were then provided the clinical variables associated with the images and again asked to classify the tumor. The human evaluators included 2 medical students, 2 peripheral nerve surgery fellows, 2 attending peripheral nerve surgeons, and 2 attending radiologists (1 general radiologist and 1 musculoskeletal radiologist). The attending peripheral nerve surgeons and radiologists were grouped as expert human evaluators for analysis. To compare against the classifier, a human expert score was generated for each tumor (1 point for each malignant attribution and 0 for benign). With 4 experts, the maximum score was 4 and the minimum score was 0. A receiver operating characteristic (ROC) curve was generated for the human expert score and the optimized threshold chosen to maximize AUC.

TABLE 1. Comparison of Clinical Variables Between the Benign and Malignant Peripheral Nerve Sheath Tumor Groups

	Benign (N = 171)	Malignant (N = 95)	*P* value
Mean age, yr (SD)	45.5 (15.3)	43.3 (18.2)	.320
Sex			
Male	75 (44%)	54 (57%)	
Female	96 (56%)	41 (43%)	.042
Spontaneous pain	41 (24%)	71 (75%)	<.001
Motor deficit	45 (26%)	31 (33%)	.275
NF1	38 (22%)	41 (43%)	<.001
NF2	10 (6%)	0 (0%)	.016
Schwannomatosis	5 (3%)	0 (0%)	.164

Values indicate number of patients (%), unless otherwise indicated.
SD = standard deviation.

TABLE 2. Relative Influence of the Top 4 Most Influential Features Contributing to the Final Radiomics Classifier

Feature	Relative influence
Texture: log-sigma-3-mm-3D-glszm_ZoneEntropy	19.8%
Clinical: pain	13.5%
Texture: wavelet-LHH_glrlm_RunVariance	11.6%
Shape: original_shape_Maximum2DDiameterSlice	11.5%

Statistical Analysis

Categorical variables were analyzed using the χ^2 test or Fisher exact test, as appropriate, and continuous variables were analyzed using Student's *t*-test. A *P*-value $< .05$ was considered statistically significant for all analyses. Sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and AUC for the ROC curve were calculated for each of the classifiers, the best-performing human evaluator, and the human expert score. For these calculations, malignant was designated as "positive." Accuracy was compared against the no-information rate (NIR), a statistical evaluator based on comparison with random chance within the known distribution of outcomes.[18] ROC curves were compared against one another using the method of DeLong.[19] All data analyses were performed using StataSE.

RESULTS

Comparison of Clinical Variables

Of the 266 patients with T1-gad imaging, 171 (64%) had BPNSTs and 95 (36%) had MPNSTs (Table 1). Patients with an MPNST were more likely to be male ($P = .042$), to have spontaneous pain ($P < .001$), and to have an NF1 diagnosis ($P < .001$). Patients with a BPNST were more likely to have an NF2 diagnosis ($P = .016$). There was no significant difference in age.

Primary Model: Imaging and Clinical Features

The primary model was created using both radiologic and clinical features. After feature reduction, 19 textural features and 2 clinical features were retained. The textural features included 2 shape, 4 first-order, 3 GLCM, 6 GLSZM, and 3 GLRLM features (see **Table S2** in **Supplemental Digital Content**). The clinical features included presence of spontaneous pain and NF2 diagnosis.

The 21 selected features were used to train a gradient boost model. The imaging features that were most influential for classification were zone entropy, run variance, and diameter, and the most influential clinical feature was spontaneous pain (Table 2; Figure 1). Zone entropy is a measure of the randomness of features within the image, wherein a higher value indicates greater heterogeneity in patterns. Run variance is the variance in lengths of runs of consecutive voxels that have the same gray value. Diameter is the 2-dimensional maximal diameter. A drop-off in relative influence was seen for the remaining imaging and clinical features (Figure 2). A correlation matrix was constructed to assess for any redundant textural features. One pair of features (difference variance and autocorrelation) had a correlation >0.8, but neither was an important contributor to the final model (see **Figure S2** in **Supplemental Digital Content**).

When the final classifier was applied to the training and test sets (see **Table S3** in **Supplemental Digital Content**), the final AUCs were 0.940 and 0.845, respectively (Figure 3; Table 3). The accuracy significantly exceeded the NIR ($P = .001$).

Secondary Model: Imaging Features Alone

A secondary model was constructed using only imaging features. Eight features were retained, with 7 contributing to the final classifier (see **Figure S3A** in **Supplemental Digital Content**). The final AUC for the test set was 0.773 (see **Figure S3B** in **Supplemental Digital Content**; Table 3).

Secondary Model: Clinical Features Alone

A secondary model was constructed using only clinical features (see **Figure S4A** in **Supplemental Digital Content**). The final AUC for the test set was 0.749 (see **Figure S4B** in **Supplemental Digital Content**; Table 3).

Human Evaluation of Test Set

The human expert score, using both clinical variables and imaging, similar to the primary model classifier, had an accuracy of 0.557 and AUC 0.624 (Figure 3). The accuracy did not exceed the NIR (Table 3). When using imaging alone, the human expert score had an accuracy of 0.557 and AUC 0.595 (see **Figure S5** in **Supplemental Digital Content**). The accuracy did not exceed the NIR (Table 3). Performance of all evaluators is provided in **Table S4** in **Supplemental Digital Content**.

Comparison of ROC Curves

The AUC for the ROC curve for the primary classifier (imaging and clinical features) significantly exceeded the AUC for the best human expert using imaging and clinical features ($P = .048$) and using only imaging ($P = .048$) and for the human expert score

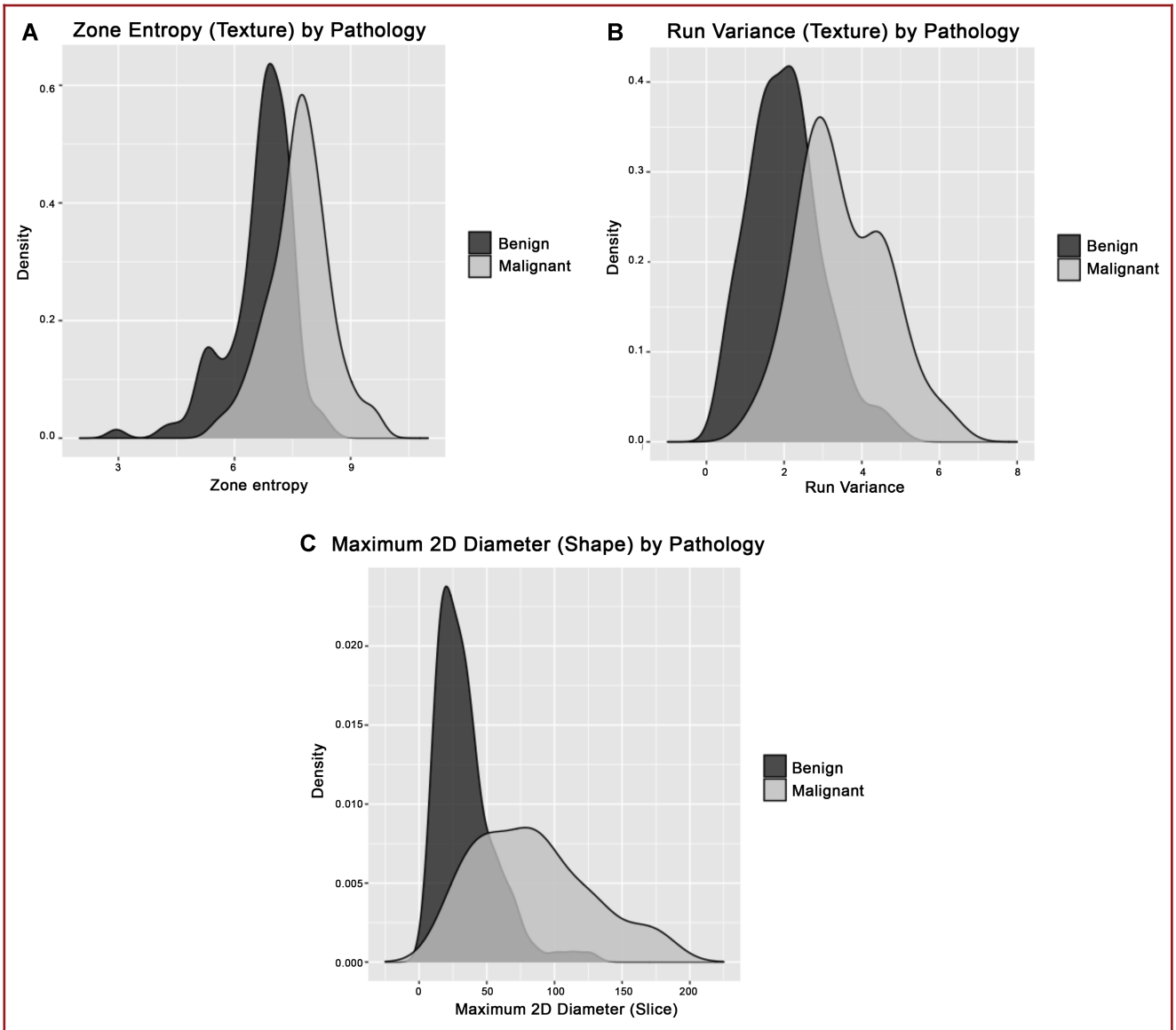

FIGURE 1. *Density plots of the top 3 radiologic features contributing to the final radiomics classifier, including* **A**, *zone entropy with 3 sigma filter (GLSZM),* **B**, *run variance with LHH filter (GLRLM), and* **C**, *maximum 2D diameter.*

using imaging and clinical features ($P = .002$) and using only imaging ($P < .001$) (Table 4).

DISCUSSION

Accurate differentiation between BPNSTs and MPNSTs is critical because the treatment paradigms are markedly different. Unfortunately, precise differentiation on radiologic and clinical features remains deficient. Even for experienced radiologists, accuracy may only be ~50%.[2] Machine-learning radiomics may offer a method for improving diagnosis. In this study, we developed and evaluated radiomic classifiers to differentiate between BPNSTs and MPNSTs.

Our primary classifier model using both clinical and imaging parameters achieved an AUC on the ROC curve of 0.845. Comparatively, expert human evaluators achieved an AUC of 0.624. The AUC for the primary classifier significantly exceeded both the best-performing human expert and the human expert score. The expert human readers were not statistically better than having no information beyond the distribution of the binary choices, whereas the primary classifier's accuracy significantly exceeded this measure against random chance. This is particularly impressive when considering that the radiomic classifier

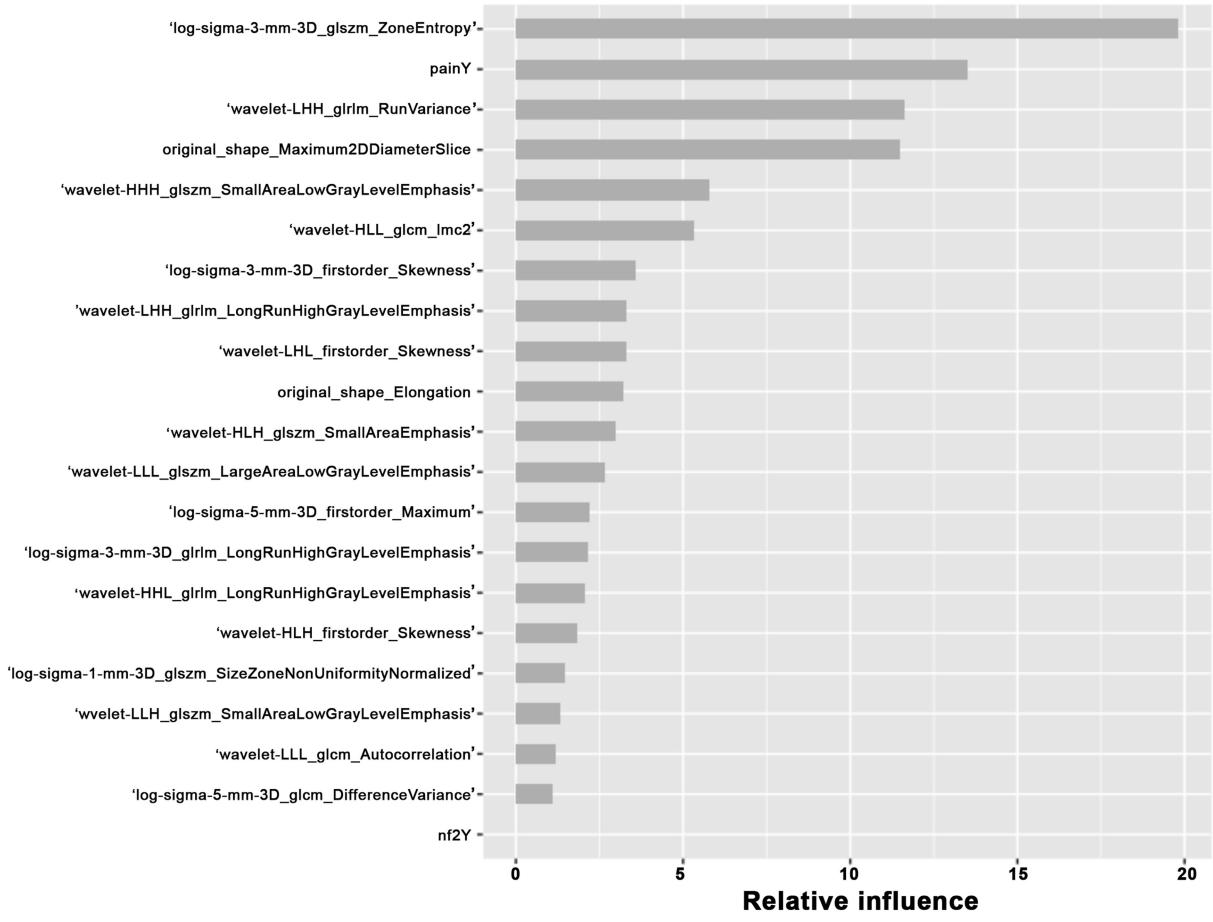

FIGURE 2. *Bar plot of relative influence of clinical and radiologic features included in the primary radiomic classifier.*

evaluated only T1-gad images, whereas the human readers were provided any available T2- or proton density-weighted images. T2-weighted images provide valuable information, such as perilesional edema and distal denervation. Thus, the human evaluators may have had a significant advantage over the radiomic classifier.

Perhaps not surprisingly, experienced clinicians tend to grade in favor of a higher sensitivity at the expense of a lower specificity to avoid missing a malignancy (Figure 4). Our primary classifier has the opposite properties, with a higher specificity and lower sensitivity in part because of the methodology, which was developed to maximize AUC, a balance between sensitivity and specificity, in a training set enriched with BPNSTs. There is risk to favoring either side. Favoring sensitivity over specificity will lead to over-calling malignancies and potentially overaggressive therapies.[8] Conversely, favoring specificity over sensitivity will lead to missing the rare diagnosis of malignancy. Aiming for accuracy is an appropriate approach to balance risks; however, radiomic analysis should not supplant expert evaluation and is not sufficiently accurate to eliminate the potential need for further preoperative diagnostic testing.

Malignancy in PNSTs has been associated with large size (>5 cm), perilesional edema, irregular or peripheral enhancement, intratumoral cystic components on imaging, rapid tumor growth, spontaneous pain, and neurological deficits clinically.[20-25] Our machine-learning classifier identified features similar to these qualitative imaging characteristics as providing strong correlation to final diagnosis. For example, slice diameter emerged as a strong predictor; this was similar to our prior demonstration that increasing maximal dimension is associated with malignancy but with considerable overlap.[2,24,26] Similarly, multiple radiomic textural features, including run variance and zone entropy, match the expectation for greater heterogeneity among MPNSTs on postgadolinium sequences.[23,24,27] Also importantly, spontaneous pain demonstrated significant influence in the development of the final classifier. Compared

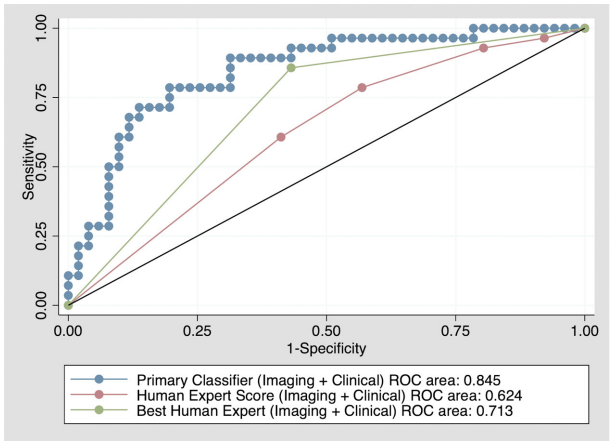

FIGURE 3. *Comparison of ROC curves for the primary classifier, human expert score, and best human expert, all using both imaging and clinical features. The AUC for the primary classifier significantly exceeded both the human expert score and best human expert.*

with prior qualitative associations, a radiomic approach allows a more precise characterization and comparison in a quantified manner, which allows for more discriminatory power. Furthermore, radiomics is not dependent upon years of human training and experience. A substantial advantage of the current classifier model is that it uses a conventional MRI sequence (ie, T1-gad), which provides potential for clinical use.

Diffusion-weighted imaging (DWI) has recently been proposed as an additional sequence on conventional MRI that can be discriminatory.[28] BPNSTs tend to have a higher apparent diffusion coefficient than MPNSTs.[20,29] Although these initial results are promising, the typical DWI sequence uses a single-shot echo-planar imaging approach, mainly because of its speed. Unfortunately, these images demonstrate susceptibility artifacts that result in geometric distortion, signal drop-out, and image blurring, which can pose challenges in evaluating PNSTs in susceptibility-prone regions (eg, bone, air, and soft-tissue interfaces). There are several, typically vendor-specific, solutions to these challenges, but divergent protocols limit multi-institutional comparison with radiomics.[30-36] Standardized and reduced susceptibility DWI protocols could be included in subsequent radiomics models.

FDG-PET offers another potential imaging modality to differentiate benign from malignant nerve tumors.[21,25,37,38] Although there is certainly value in PET imaging, there are concerns with using this technique to universally evaluate nerve tumors. First, schwannomas, in particular, can demonstrate elevated maximum standardized uptake value without clinical implications for malignancy. An elevated value may prompt unnecessary biopsy or patient concern if PET were indiscriminately used to evaluate all PNSTs. Second, FDG-PET exposes the patient to radiation, which is particularly concerning for tumor syndromes and for serial imaging.

Overall, a radiomic classifier may provide the most accurate option to categorize patients for further testing for MPNST, such as PET imaging or biopsy, that could be used along with expert opinion for increased sensitivity. A sequential strategy may prove to reduce overall healthcare costs and reduce risks.

Limitations and Future Directions

Our classification algorithm shares many of the common challenges in radiomics that limit its performance. Heterogeneity in MRI acquisition technique as a result of institutional variations in machine technology and sequence selection can affect the assignment of gray-level intensities and higher-level feature calculations. These variations can also lead to differences in how motion, fat-saturation, and contrast quality are captured, which have downstream implications on how features are calculated. The presence of lesions from various anatomic locations may preclude reconciliation of features.

TABLE 3. Comparison of Metrics for the Human Evaluators and the Machine-Learning Classifiers

Evaluator	Sensitivity	Specificity	PPV	NPV	F1 Score	Accuracy (*P*-value)	AUC (95% CI)
Imaging and clinical							
Overall human evaluators	0.684	0.742	0.589	0.823	0.625	0.722 ($P = .303$)	0.704 (0.643-0.765)
Expert human evaluators	0.833	0.673	0.583	0.888	0.684	0.730 ($P = .254$)	0.746 (0.700-0.792)
Primary model classifier	0.676	0.882	0.760	0.833	0.717	0.810 ($P = .001$)	0.845 (0.823-0.979)
Imaging only							
Overall human evaluators	0.704	0.723	0.582	0.826	0.632	0.716 ($P = .347$)	0.702 (0.655-0.749)
Expert human evaluators	0.833	0.686	0.594	0.890	0.691	0.738 ($P = .241$)	0.750 (0.708-0.792)
Secondary model classifier	0.607	0.784	0.607	0.784	0.717	0.722 ($P = .096$)	0.773 (0.693-0.894)
Clinical only							
Secondary model classifier	0.643	0.804	0.643	0.804	0.643	0.747 ($P = .036$)	0.749 (0.630-0.867)

Accuracy was compared against the no-information rate.
PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; 95% CI, 95% confidence interval.

TABLE 4. Pairwise Comparison of the Area Under the Curve for the Receiver Operating Characteristic Curves Using the Method of DeLong

		Best expert		Human expert		Classifier		
		Imaging + clinical	Imaging	Imaging + clinical	Imaging	Imaging + clinical	Imaging	Clinical
Best expert	Imaging + clinical	–	$P = .999$	$P = .305$	$P = .180$	**$P = .048$**	$P = .396$	$P = .458$
	Imaging	$P = .999$	–	$P = .305$	$P = .180$	**$P = .048$**	$P = .396$	$P = .458$
Human expert	Imaging + clinical	$P = .305$	$P = .305$	–	$P = .459$	**$P = .002$**	$P = .072$	**$P = .049$**
	Imaging	$P = .180$	$P = .180$	$P = .459$	–	**$P < .001$**	**$P = .030$**	**$P = .015$**
Classifier	Imaging + clinical	**$P = .048$**	**$P = .048$**	**$P = .002$**	**$P < .001$**	–	$P = .057$	$P = .209$
	Imaging	$P = .396$	$P = .396$	$P = .072$	**$P = .030$**	$P = .057$	–	$P = .951$
	Clinical	$P = .458$	$P = .458$	**$P = .049$**	**$P = .015$**	$P = .209$	$P = .951$	–

Bold indicates statistical significance ($P < .05$).

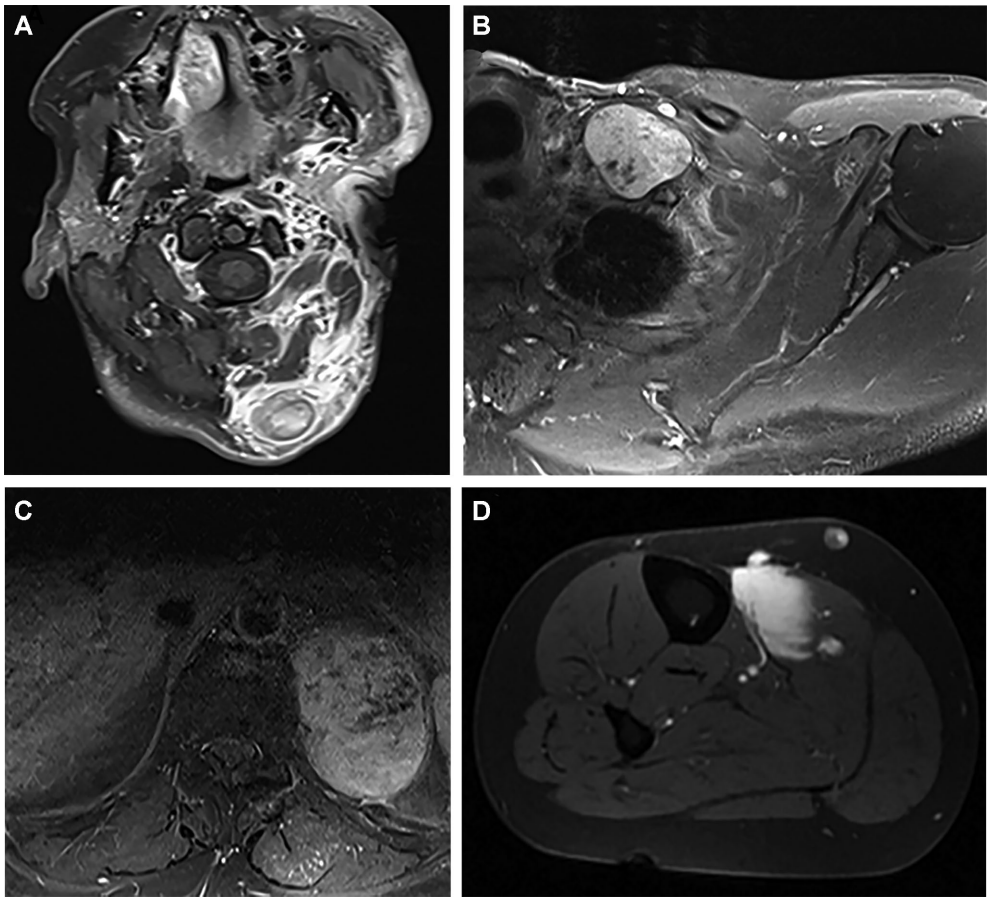

FIGURE 4. **A**, *Imaging from a 26-yr-old woman with spontaneous pain, a neurological deficit on examination, and NF1 for whom the expert human consensus was MPNST, but the primary classifier predicted benign. The tumor was a benign neurofibroma.* **B**, *Imaging from a 62-yr-old man with no spontaneous pain, a neurological deficit on examination, and no neurogenetic diagnosis. The expert human consensus in this case was malignant but the primary classifier predicted benign. The tumor was a benign schwannoma.* **C**, *Imaging from a 74-yr-old woman with spontaneous pain, no neurological deficit on examination, and no neurogenetic diagnosis. In this case, the expert human consensus was malignant, and the primary classifier predicted benign. The tumor was a benign schwannoma.* **D**, *Imaging from a 55-yr-old man with no spontaneous pain, no neurological deficit on examination, and no neurogenetic diagnosis for whom the expert human consensus was benign, but the primary classifier predicted malignant. Tumor was a malignant peripheral nerve sheath tumor.*

In the future, we intend to gather a larger multi-institutional series so that we can include T2-weighted imaging in the next iteration. Furthermore, a larger cohort may allow for other machine-learning techniques, such as deep learning, that may also improve the classifier. Future iterations will again be set up to maximize AUC, but it is not clear how incorporating a larger cohort or incorporating T2-weighted images will affect sensitivity vs specificity.

CONCLUSION

Optimal management of PNSTs depends on the clinical distinction between benign and malignant lesions. Currently, this distinction is drawn from a mixture of clinical art and imaging interpretation that remains lacking in precision; however, as radiomics becomes more integrated clinically, these tools may offer greater precision and serve as an adjunct to our clinical practice. Our machine-learning–based algorithm optimized for AUC using axial, T1-weighted, postgadolinium images demonstrated success at least equivalent, if not superior, to human interpretation, even when humans were presented with more data. Further refinement using T2-weighted imaging or DWI, larger datasets, and deeper machine-learning techniques may provide further discriminatory power and will likely become a critical aspect of clinical evaluation.

Funding

This study was funded in part by the Medtronic Young Clinician Investigator Award from the Neurosurgery Research & Education Foundation (NREF).

Disclosures

The authors have no personal, financial, or institutional interest in any of the drugs, materials, or devices described in this article. Dr Zhang is funded by a T32 award from the National Institutes of Health 5T32CA009695-27 (MPI). Dr Napel is funded by NIH U01 CA187947. Dr Mahan has served as a consultant for AxoGen and Joimax.

REFERENCES

1. Ferner RE, Gutmann DH. International consensus statement on malignant peripheral nerve sheath tumors in neurofibromatosis. *Cancer Res*. 2002;62(5):1573-1577.
2. Karsy M, Guan J, Ravindra VM, Stilwill S, Mahan MA. Diagnostic quality of magnetic resonance imaging interpretation for peripheral nerve sheath tumors: can malignancy be determined? *J Neurol Surg A Cent Eur Neurosurg*. 2016;77(6):495-504.
3. Broski SM, Johnson GB, Howe BM, et al. Evaluation of 18F-FDG PET and MRI in differentiating benign and malignant peripheral nerve sheath tumors. *Skeletal Radiol*. 2016;45(8):1097-1105.
4. Tovmassian D, Abdul Razak M, London K. The role of [(18)F]FDG-PET/CT in predicting malignant transformation of plexiform neurofibromas in neurofibromatosis-1. *Int J Surg Oncol*. 2016;2016:6162182.
5. Bai X, Wang X. Solitary benign schwannoma mimics residual malignancy on FDG PET/CT. *Clin Nucl Med*. 2018;43(10):782-784.
6. Lieber B, Han B, Allen J, et al. Utility of positron emission tomography in schwannomatosis. *J Clin Neurosci*. 2016;30:138-140.
7. Miyake KK, Nakamoto Y, Kataoka TR, et al. Clinical, morphologic, and pathologic features associated with increased FDG uptake in schwannoma. *AJR Am J Roentgenol*. 2016;207(6):1288-1296.
8. Perez-Roman RJ, Shelby Burks S, Debs L, Cajigas I, Levi AD. The risk of peripheral nerve tumor biopsy in suspected benign etiologies. *Neurosurgery*. 2020;86(3):E326-E332.
9. Lo CM, Weng RC, Cheng SJ, Wang HJ, Hsieh KL. Computer-aided diagnosis of isocitrate dehydrogenase genotypes in glioblastomas from radiomic patterns. *Medicine*. 2020;99(8):e19123.
10. Qian J, Herman MG, Brinkmann DH, et al. Prediction of MGMT status for glioblastoma patients using radiomics feature extraction from (18)F-DOPA-PET imaging. *Int J Radiat Oncol Biol Phys*. 2020;108(5):1339-1346.
11. Shboul ZA, Alam M, Vidyaratne L, Pei L, Elbakary MI, Iftekharuddin KM. Feature-guided deep radiomics for glioblastoma patient survival prediction. *Front Neurosci*. 2019;13:966.
12. Wang K, Qiao Z, Zhao X, et al. Individualized discrimination of tumor recurrence from radiation necrosis in glioma patients using an integrated radiomics-based model. *Eur J Nucl Med Mol Imaging*. 2020;47(6):1400-1411.
13. Uthoff J, De Stefano FA, Panzer K, et al. Radiomic biomarkers informative of cancerous transformation in neurofibromatosis-1 plexiform tumors. *Journal of Neuroradiology*. 2019;46(3):179-185.
14. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107.
15. Zwanenburg A, Vallieres M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338.
16. Kuhn M. Building predictive models in R using the caret package. *J Stat Soft*. 2008;28(5):1-26.
17. Friedman JH, Meulman JJ. Multiple additive regression trees with application in epidemiology. *Statist Med*. 2003;22(9):1365-1381.
18. Patro V, Patra M. Augmenting weighted average with confusion matrix to enhance classification accuracy. *Trans Mach Learn Artif Intel*. 2014;2(4):77-91.
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.
20. Demehri S, Belzberg A, Blakeley J, Fayad LM. Conventional and functional MR imaging of peripheral nerve sheath tumors: initial experience. *Am J Neuroradiol*. 2014;35(8):1615-1620.
21. Benz MR, Czernin J, Dry SM, et al. Quantitative F18-fluorodeoxyglucose positron emission tomography accurately characterizes peripheral nerve sheath tumors as malignant or benign. *Cancer*. 2010;116(2):451-458.
22. Salamon J, Mautner VF, Adam G, Derlin T. Multimodal imaging in neurofibromatosis type 1-associated nerve sheath tumors. *Rofo*. 2015;187(12):1084-1092.
23. Matsumine A, Kusuzaki K, Nakamura T, et al. Differentiation between neurofibromas and malignant peripheral nerve sheath tumors in neurofibromatosis 1 evaluated by MRI. *J Cancer Res Clin Oncol*. 2009;135(7):891-900.
24. Wasa J, Nishida Y, Tsukushi S, et al. MRI features in the differentiation of malignant peripheral nerve sheath tumors and neurofibromas. *AJNR Am J Neuroradiol*. 2010;194(6):1568-1574.
25. Derlin T, Tornquist K, Munster S, et al. Comparative effectiveness of 18F-FDG PET/CT versus whole-body MRI for detection of malignant peripheral nerve sheath tumors in neurofibromatosis type 1. *Clin Nucl Med*. 2013;38(1):e19-e25.
26. Soldatos T, Fisher S, Karri S, Ramzi A, Sharma R, Chhabra A. Advanced MR imaging of peripheral nerve sheath tumors including diffusion imaging. *Semin Musculoskelet Radiol*. 2015;19(2):179-190.
27. Friedrich RE, Kluwe L, Fünsterer C, Mautner VF. Malignant peripheral nerve sheath tumors (MPNST) in neurofibromatosis type 1 (NF1): diagnostic findings on magnetic resonance images and mutation analysis of the NF1 gene. *Anticancer Res*. 2005;25(3a):1699-1702.
28. Mazal AT, Ashikyan O, Cheng J, Le LQ, Chhabra A. Diffusion-weighted imaging and diffusion tensor imaging as adjuncts to conventional MRI for the diagnosis and management of peripheral nerve sheath tumors: current perspectives and future directions. *Eur Radiol*. 2019;29(8):4123-4132.
29. Ahlawat S, Blakeley JO, Rodriguez FJ, Fayad LM. Imaging biomarkers for malignant peripheral nerve sheath tumors in neurofibromatosis type 1. *Neurology*. 2019;93(11):e1076-e1084.
30. Butts K, Pauly J, de Crespigny A, Moseley M. Isotropic diffusion-weighted and spiral-navigated interleaved EPI for routine imaging of acute stroke. *Magn Reson Med*. 1997;38(5):741-749.

31. Liu C, Bammer R, Kim DH, Moseley ME. Self-navigated interleaved spiral (SNAILS): application to high-resolution diffusion tensor imaging. *Magn Reson Med*. 2004;52(6):1388-1396.
32. Pipe JG, Farthing VG, Forbes KP. Multishot diffusion-weighted FSE using PROPELLER MRI. *Magn Reson Med*. 2002;47(1):42-52.
33. Skare S, Newbould RD, Clayton DB, Bammer R. Propeller EPI in the other direction. *Magn Reson Med*. 2006;55(6):1298-1307.
34. Holdsworth SJ, Skare S, Newbould RD, Bammer R. Robust GRAPPA-accelerated diffusion-weighted readout-segmented (RS)-EPI. *Magn Reson Med*. 2009;62(6):1629-1640.
35. Holdsworth SJ, Skare S, Newbould RD, Guzmann R, Blevins NH, Bammer R. Readout-segmented EPI for rapid high resolution diffusion imaging at 3T. *Eur J Radiol*. 2008;65(1):36-46.
36. Porter DA, Heidemann RM. High resolution diffusion-weighted imaging using readout-segmented echo-planar imaging, parallel imaging and a two-dimensional navigator-based reacquisition. *Magn Reson Med*. 2009;62(2):468-475.
37. Schwabe M, Spiridonov S, Yanik EL, et al. How effective are noninvasive tests for diagnosing malignant peripheral nerve sheath tumors in patients with neurofibromatosis type 1? Diagnosing MPNST in NF1 patients. *Sarcoma*. 2019;2019:1-8.
38. Salamon J, Derlin T, Bannas P, et al. Evaluation of intratumoural heterogeneity on 18F-FDG PET/CT for characterization of peripheral nerve sheath tumours in neurofibromatosis type 1. *Eur J Nucl Med Mol Imaging*. 2013;40(5):685-692.

Supplemental digital content *is available for this article at www.neurosurgery-online.com.*

Supplemental Digital Content. Four tables and 5 figures. The Supplemental Digital Content provides MRI acquisition and gradient boost modeling features and hyperparameters and their interpretations. It also includes demographic data for the training and test sets of patients and comparison of the relative influence of features in the classifiers and the ROC curves of the performance of the classifiers. A comparison of the metrics for human evaluators of various skill levels is presented.

COMMENT

This approach comes to provide additional tool for preoperative differentiation between PNST and MPNST. The authors present a machine-learning classifier using computational radiomics (mostly T1 + Gad) along with limited clinical data collection and compare it to human evaluators showing a better diagnostic yield to the former when balancing between sensitivity and specificity.

While prognosis is fundamentally different and lies on the 2 extremities of the diagnosis spectrum, preoperative misdiagnosis does not account for a catastrophic outcomes as the authors suggest but should rather serve as an important decision-making point. Furthermore, cautioning against the current over utilization of unnecessary biopsies along with associated sampling error and possible neurological injury is an important point of discussion. Still, the certainty in which a surgeon can rely nowadays exclusively on image characteristics of tumors without putting into account other parameters remains controversial as the authors allude.

Current surge in data learning prediction models is gaining increased interest in both research and clinical utilization. Recent reviews consider its superiority by addressing significant lack of transparent reporting in 'standard' multivariable prediction model studies reporting and human evaluation. Yet, the main drawback of this approach for clinical utilization is by the non-linear and variable data sets it is provided with.

This approach is surely commendable and can definitely aid our decision making, but one should take caution concentrating on interpreting a single image modality albeit being a purely objective tool and caution overlooking neurological findings over time, extended demographics and expert surgeon decision making.

Daniel Umansky
Rajiv Midha
Calgary, Alberta, Canada