Machine learning analyses can differentiate meningioma grade by features on magnetic resonance imaging

Andrew T. Hale, MS,1–3 David P. Stonko, MD,3 Li Wang, MS,4 Megan K. Strother, MD,5 and Lola B. Chambless, MD1,3

1Department of Neurosurgery, Vanderbilt University Medical Center; 2Medical Scientist Training Program, Vanderbilt University School of Medicine; 3Vanderbilt University School of Medicine; 4Department of Biostatistics, Vanderbilt University; and 5Department of Radiology, Vanderbilt University Medical Center, Nashville, Tennessee

OBJECTIVE Prognostication and surgical planning for WHO grade I versus grade II meningioma requires thoughtful decision-making based on radiographic evidence, among other factors. Although conventional statistical models such as logistic regression are useful, machine learning (ML) algorithms are often more predictive, have higher discriminative ability, and can learn from new data. The authors used conventional statistical models and an array of ML algorithms to predict atypical meningioma based on radiologist-interpreted preoperative MRI findings. The goal of this study was to compare the performance of ML algorithms to standard statistical methods when predicting meningioma grade.

METHODS The cohort included patients aged 18–65 years with WHO grade I (n = 94) and II (n = 34) meningioma in whom preoperative MRI was obtained between 1998 and 2010. A board-certified neuroradiologist, blinded to histological grade, interpreted all MR images for tumor volume, degree of peritumoral edema, presence of necrosis, tumor location, presence of a draining vein, and patient sex. The authors trained and validated several binary classifiers: k-nearest neighbors models, support vector machines, naïve Bayes classifiers, and artificial neural networks as well as logistic regression models to predict tumor grade. The area under the curve–receiver operating characteristic curve was used for comparison across and within model classes. All analyses were performed in MATLAB using a MacBook Pro.

RESULTS The authors included 6 preoperative imaging and demographic variables: tumor volume, degree of peritumoral edema, presence of necrosis, tumor location, patient sex, and presence of a draining vein to construct the models. The artificial neural networks outperformed all other ML models across the true-positive versus false-positive (receiver operating characteristic) space (area under curve = 0.8895).

CONCLUSIONS ML algorithms are powerful computational tools that can predict meningioma grade with great accuracy.

https://thejns.org/doi/abs/10.3171/2018.8.FOCUS18191

KEYWORDS artificial intelligence; machine learning; meningioma; predictive modeling

Throughout the medical and surgical literature, it has become increasingly important to develop methods to predict outcomes, complications, and prognoses for given disease states. Specifically, it is often critical to be able to discriminate between binary characteristics or outcomes in the clinical setting (e.g., death or survival, benign or malignant, etc.).17,18,32–34 Over the past several decades, as machine learning (ML) algorithms and artificial intelligence (AI) have improved, the modeling choices available to overcome these diagnostic and prognostic dilemmas have evolved. Now a number of binary classifiers exist that can offer unique statistical insights that are often superior to traditional statistical methods, including logistic regression.8,10,15 AI and ML methods are not entirely new to the neurosurgical literature,16,25–27 but the nuances of their capabilities and limitations may be unfamiliar to many neurosurgeons. ML and AI methods may provide physicians with much more powerful evidence-based predictive modeling tools to help guide clinical practice.

One advantage of ML is that it can facilitate construc-
tion of accurate statistical models even with relatively small data sets. This is critical to the study of rare diseases like meningiomas. Although approximately 90% of meningiomas are WHO grade I, higher-grade (WHO grades II and III) meningiomas occur with some frequency and have a substantially worse prognosis. Preoperative management of meningioma can be challenging as the surgeon and patient weigh the potential of obtaining gross-total resection against the risk of recurrence or progression to higher-grade pathology if a less aggressive surgical approach is used. The advantage of preoperative diagnosis of WHO I versus WHO II is 2-fold: 1) assisting the surgeon’s decision to operate sooner versus conservative management, and 2) helping the surgeon determine how aggressive to be during the resection.

We previously described a series of patients with meningioma who underwent preoperative MRI, and we used traditional statistical methods to define radiographic features including tumor volume, presence of necrosis, degree of peritumoral edema, and location of the tumor that were significantly associated with higher-grade meningioma. Here, we test a multitude of ML algorithms with the goal of building a model to predict meningioma grade with greater accuracy. For this reason, and because of the clinical utility that such a predictive tool would provide, we tested a wide array of applicable binary classification modeling techniques (both conventional statistics and ML algorithms) and built a clinical tool with the highest attainable predictive ability by using our cohort of patients.

**Methods**

We performed a retrospective review of patients 18–65 years of age with histologically confirmed WHO grade I (n = 94) and II (n = 34) meningioma, who had undergone preoperative MRI at Vanderbilt University Medical Center between 1998 and 2010. All MRI studies were interpreted in blinded fashion by a board-certified neuroradiologist. Using conventional statistics, we had previously determined that tumor volume calculated from preoperative MRI, degree of peritumoral edema seen on MRI, presence of tumor necrosis seen on MRI, and tumor location were statistically significant predictors of WHO grade II pathology as set a priori at p < 0.05. Peritumoral edema was evaluated according to the following scale. 0) No edema: absence of increased T2 signal surrounding the meningioma. 1) Mild edema: rim or crescent of increased T2 signal surrounding the meningioma without mass effect. 2) Moderate edema: more extensive increased T2 signal surrounding the meningioma without mass effect. 3) Severe edema: mass effect from edema and/or tongues of advancing edema. Tumor location was stratified into the following groups: 1) convexity (including frontal, parietal, temporal, lateral sphenoid wing, and petrous); 2) skull base (including sellar/suprasellar/anterior clinoid, cavernous/medial sphenoid wing, and fovea ethmoidalis); 3) falx; and 4) posterior fossa (including tentorium, cerebellum, and cerebellopontine angle). Tumor volume was calculated using manual tracing, considering the largest diameter in the anterior-posterior, craniocaudal, or transverse directions seen on T1-weighted MRI as previously described. Only patients with complete data were included in these analyses; no data imputation was performed. Furthermore, female sex (p = 0.072) and the presence of a draining vein on preoperative MRI (p = 0.079) approached statistical significance.

All aforementioned preoperative variables (tumor volume, edema, necrosis, location, sex, and draining vein) were included in our ML analyses to avoid bias toward any particular modeling technique. The models analyzed here include the following: 1) k-nearest neighbors (k-NN) by minimization of expected classification cost; 2) support vector machines (SVMs) via sequential minimal optimization; 3) naive Bayes classifier; 4) artificial neural network (ANN) for binary classification, which was trained by scaled conjugate gradient method; as well as 5) logistic regression models, with main effect terms, interaction terms, and quadratic terms.

All models were subjected to hyperparameter optimization individually, and cross-validation was performed. Specifically, k-NN underwent optimization for distance metric and number of neighbors included, SVMs were optimized for kernel scale and function, and naïve Bayes classifier was optimized for distribution type and width. Logistic regression models were fit several ways: manually, allowing for interactions and quadratics, but also with stepwise variable incorporation and removal. Chi-square statistics were calculated for logistic regression curves to assess for difference between the test model and a constant model. ANNs underwent supervised training, with 70% of data allocated for training, 15% for validation, and 15% for testing. Varying numbers of nodes were used while monitoring performance, and during monitoring the receiver operating characteristic (ROC) curves and area under the curves (AUCs) were calculated by the trapezoid method. These were also computed for all other models and were one of the instruments used for comparison. Additionally, the likelihood ratio and Wald test were used for comparison of logistic regression models, cross-entropy minimization was used for comparison of ANNs, and 5-fold cross-validation was used for other ML models. All mathematical analysis was done using offline MATLAB (R2016b (9.1.0.441655) with an academic license on a 64-bit MacBook Pro running version 10.11.6. All aspects of this study were preapproved by the Vanderbilt University Medical Center Institutional Review Board.

**Results**

Our cohort included 128 patients diagnosed with WHO grade I (n = 94) or WHO grade II (n = 34) meningioma by histological grading after resection. The mean age at diagnosis was 55 years (range 18–65 years). The following preoperative variables were included in constructing all ML models: 1) tumor volume; 2) degree of peritumoral edema; 3) tumor location; 4) presence of tumor necrosis; 5) presence of draining vein; and 6) patient sex (Table 1). Selection of these variables was based on their statistically significant association by regression analyses with grade II meningioma. These 6 variables were used to train several ML algorithms and statistical tests, where each was individually optimized and then compared across model-
ing subtypes. ROC curves and AUCs were calculated for each method (Table 2). The ROC curves correspond to predictions on both training and testing data.

Our optimized ANN outperformed all other modeling types across most of the true-positive versus false-positive (ROC) space (AUC = 0.8895). The optimized logistic regression with quadratic terms (AUC = 0.8423, p = 0.00196) outperformed logistic regression with interactions (AUC = 0.7641, p = 0.00888), and simple logistic regression (AUC = 0.731, p = 0.004). Optimized k-NN (1 neighbor) minimized 5-fold cross-validation error with respect to other k-NN models (AUC = 0.8687, Table 2). Course gaussian SVM was the optimal SVM identified by hyperspace parameterization, and it had a relatively high AUC (0.8611). This method also outperformed all k-NN in 5-fold cross-validation (0.2547 vs 0.2738). Naïve Bayes classifier (AUC = 0.7096) and logistic regression with main effect terms (AUC = 0.7228) were the least predictive models when compared by AUC. These same data are summarized graphically in Fig. 1. Overall, these data suggest that ANN is the most robust methodology for predicting tumor grade in our cohort, when compared by AUC-ROC and when cross-validation scores are compared.

**Discussion**

Conventional statistics such as logistic regression have long been used for the process of binary classification and prediction. Logistic regression analysis serves as the de facto method by which one might approach clinical classification problems, such as the one presented here in predicting tumor grade preoperatively. This is often for good reason: these methods are familiar, provide interpretable and useful statistics, and often reach the predictive ability desired for clinical decision-making if enough data are analyzed. However, modern ML can often outperform logistic regression in predictive and discriminative ability despite using data from a limited sample size.17,20,24,28,32 These features make ML particularly useful in situations in which statistical power is relatively low, such as in outcomes in patients with relatively rare diseases like grade II meningioma.

Many binary classifiers have been described in the mathematics and statistics literature; ANNs are one such classifier. They are computational constructs that are trained on data inputs in order to predict an outcome, or set of outcomes, based on this pattern recognition.9,10,15 ANNs are often more useful than conventional statistics because 1) ANNs can take an arbitrary number of input variables and predict any number of target variables of either continuous or discrete nature, 2) ANNs are able to continue to “learn” or improve their predictive ability as they are exposed to new data, 3) ANNs benefit from internal validation and testing, and 4) ANNs tend to have stronger discriminant ability than conventional statistics.6–9

Logistic regression is a common method used for capturing and expressing the probability of a binary event based on a fit achieved with prior exposure to data. One benefit of logistic regression is the familiarity with this method among medical researchers, as well as the production of statistics (i.e., odds ratios) detailing strength of ef-
TABLE 2. Numerical results from model analysis

<table>
<thead>
<tr>
<th>Model Type</th>
<th>AUC</th>
<th>Secondary Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN (12 nodes, training by SCG)</td>
<td>0.8895</td>
<td>0.2305*</td>
</tr>
<tr>
<td>Logistic regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.8423</td>
<td>0.0012†</td>
</tr>
<tr>
<td>Main effect only</td>
<td>0.7228</td>
<td>0.0089†</td>
</tr>
<tr>
<td>k-NN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 neighbor</td>
<td>NA</td>
<td>0.2738‡</td>
</tr>
<tr>
<td>5 neighbors</td>
<td>0.8667</td>
<td>0.3223‡</td>
</tr>
<tr>
<td>10 neighbors</td>
<td>0.7641</td>
<td>0.3178‡</td>
</tr>
<tr>
<td>SVM: course gaussian</td>
<td>0.8611</td>
<td>0.2547‡</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.7096</td>
<td>0.2613‡</td>
</tr>
</tbody>
</table>

NA = not applicable; SCG = scaled conjugate gradient. AUC of the ROC was maximized under the optimal ANN (AUC = 0.8895).

Logistic regression with quadratic terms AUC and optimized (5 neighbors with euclidean distance measure) k-NN also had relatively high AUC: 0.8423 and 0.8667, respectively. Other performance metrics were useful for classification of models within subgroups.

* Performance calculation: cross-entropy averaged over 50 trainings with 12 nodes.
† p value of chi-square test indicating difference from constant model.
‡ Cross-validation: mean 5-fold cross-validation loss over 50 random partitions. Kernel scale = 8.915. AUCs shown here correspond to those shown on ROC curves in Fig. 1.

FIG. 1. ROC curve for ML algorithms and statistical methods implemented for the prediction of atypical meningioma.

effect. ANNs do not have an analogous metric. Because the goal of this study was to achieve the greatest predictive ability and clinical utility possible, we also compared a number of other classification techniques, including k-NN models, which makes predictions by allowing training data close to new observations to “vote” on a prediction classifier, and naïve Bayes classifiers and SVMs, both of which use algorithms to perform supervised learning in order to classify new observations based on training data.

ANNs have been used for predicting outcomes in traumatic head injuries, postsurgical outcomes, and lengths of hospital stay in neurosurgery and other surgical and medical specialties.1,3–5,8,9,13,18,24,34 Recently, a number of systematic reviews have been published detailing the use of AI and ML in neurosurgery.25–27 These models are often designed by taking preoperative patient factors or disease characteristics, and using these variables to predict outcomes, complications, and/or hospital events. One limitation of much of this work is that algorithms are often chosen a priori, rather than by allowing quantitative “competition” between models to guide the choice of ML model.

Recently, a number of imaging-based ML studies for image segmentation have been developed,2,2,19,22,31 including for preoperative prediction of meningioma grade.7 These models, in theory, can capture much greater detail than radiologist-interpreted metrics.23 However, creation of these segmentation models is much more computationally challenging, requiring significant computer science and engineering skills. Our model, on the other hand, was built using data that can be collected during routine clinical care. The ANN trained here outperformed image-segmentation–based methods for predicting preoperative meningioma grade, in part due to the larger number of patients included in our study.7 Nonetheless, we are advocates for scaling up imaging-based segmentation tools for tumor grade prediction and incorporation of these models directly into imaging suites. We believe that this approach represents the most promising and practical way forward for incorporation of ML techniques into routine clinical care.

Ultimately, the optimized ANN outperformed all other modeling types tested when assessed by the area under the ROC curve (Table 2). It also provides the additional benefit of easy retraining, and the ability to learn from new data: as more data become available, each model can become more discriminative. However, this study does have limitations. The small sample size limits much more robust statistical prowess and predictive capability. In the future, it is possible to imagine that physicians could deposit their data in real time to an algorithm to increase the model’s predictive ability over time. However, the infrastructure for such an endeavor is not yet in place. In addition, in order to wield point-of-care diagnostic power, the imaging results will still need to be read and interpreted by a radiologist. Thus, additional studies accounting for interrater reliability would need to be conducted in order to quantify the amount of error between radiologists’ interpretation of the imaging metrics collected here. Furthermore, these imaging metrics represent a single MRI session at a single point in time, and thus do not take into account intratumoral changes over time, which have been shown to predict grade II meningioma.29 Despite these limitations, this proof-of-principle study demonstrates the feasibility of harnessing AI and ML techniques in predicting meningioma grade based on a radiologist-interpreted preoperative MR image. Further studies of this algorithm, including external validation across radiologists and institutions, are warranted.

Conclusions

We analyzed the preoperative imaging and demograph-
ic characteristics of 128 patients with grade I or grade II meningiomas in order to train several ML algorithms, with the goal of developing a predictive algorithm. Ultimately, we developed an ANN with strong predictive ability (AUC = 0.8895) for atypical meningioma based on preoperative imaging findings. Our ANN outperformed logistic regression and several other ML algorithms, highlighting the necessity of performing conventional statistics alongside a wide range of ML techniques for maximizing predictive capabilities. Until now, there have been no easy-to-use and accessible AI tools to identify grade I versus grade II meningioma. Given the advance in electronic medical records, bioinformatics, and computer science, we propose that models like those described here could be integrated into health data systems and imaging software, thereby becoming indispensable tools for meningioma surgeons.

References
Disclosures
The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

Author Contributions
Conception and design: Hale, Stonko, Chambless. Acquisition of data: Hale, Strother. Analysis and interpretation of data: Hale, Stonko, Chambless. Drafting the article: Hale. Critically revising the article: Hale, Stonko, Chambless. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Hale. Statistical analysis: Hale, Stonko, Wang. Study supervision: Strother, Chambless.

Correspondence
Andrew T. Hale: Vanderbilt University School of Medicine, Nashville, TN. andrew.hale@vanderbilt.edu.