Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: a systematic review and meta-analysis

Anthony V. Nguyen, BA,¹ Elizabeth E. Blears, MD,² Evan Ross, MD,² Rishi R. Lall, MD,³ and Juan Ortega-Barnett, MD³

¹School of Medicine, ²Department of Surgery, and ³Department of Neurosurgery, The University of Texas Medical Branch, Galveston, Texas

OBJECTIVE Glioblastoma (GBM) and primary central nervous system lymphoma (PCNSL) are common intracranial pathologies encountered by neurosurgeons. They often may have similar radiological findings, making diagnosis difficult without surgical biopsy; however, management is quite different between these two entities. Recently, predictive analytics, including machine learning (ML), have garnered attention for their potential to aid in the diagnostic assessment of a variety of pathologies. Several ML algorithms have recently been designed to differentiate GBM from PCNSL radiologically with a high sensitivity and specificity. The objective of this systematic review and meta-analysis was to evaluate the implementation of ML algorithms in differentiating GBM and PCNSL.

METHODS The authors performed a systematic review of the literature using PubMed in accordance with PRISMA guidelines to select and evaluate studies that included themes of ML and brain tumors. These studies were further narrowed down to focus on works published between January 2008 and May 2018 addressing the use of ML in training models to distinguish between GBM and PCNSL on radiological imaging. Outcomes assessed were test characteristics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC).

RESULTS Eight studies were identified addressing use of ML in training classifiers to distinguish between GBM and PCNSL on radiological imaging. ML performed well with the lowest reported AUC being 0.878. In studies in which ML was directly compared with radiologists, ML performed better than or as well as the radiologists. However, when ML was applied to an external data set, it performed more poorly.

CONCLUSIONS Few studies have applied ML to solve the problem of differentiating GBM from PCNSL using imaging alone. Of the currently published studies, ML algorithms have demonstrated promising results and certainly have the potential to aid radiologists with difficult cases, which could expedite the neurosurgical decision-making process. It is likely that ML algorithms will help to optimize neurosurgical patient outcomes as well as the cost-effectiveness of neurosurgical care if the problem of overfitting can be overcome.

<https://thejns.org/doi/abs/10.3171/2018.8.FOCUS18325>

KEYWORDS machine learning; predictive analytics; primary central nervous system lymphoma; glioblastoma; radiological diagnosis

GLIOMAS (GBMs) and primary central nervous system lymphomas (PCNSLs) are two of the most common malignant primary brain tumors.²⁵ The two can be difficult to differentiate based on radiology alone, and treatment differs between the conditions.^{3,15}

Although histology is the gold standard for diagnosis, patients with PCNSL often present symptomatically, and use of steroids to manage symptoms decreases the reliability of subsequent biopsy procedures.^{3,15} Current guidelines for the treatment of GBM and PCNSL differ, with the former

ABBREVIATIONS AUC = area under the receiver operating curve; FN = false negative; FP = false positive; GBM = glioblastoma; ML = machine learning; PCNSL = primary central nervous system lymphoma; PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analysis; QUADAS-2 = Quality Assessment of Diagnostic Accuracy Studies-2; SVM = support vector machine; TN = true negative; TP = true positive.

SUBMITTED June 23, 2018. **ACCEPTED** August 2, 2018.

INCLUDE WHEN CITING DOI: 10.3171/2018.8.FOCUS18325.

being treated by aggressive resection, while patients with the latter condition undergo chemotherapy, targeted therapies, and whole brain radiation treatment.^{2,26} Stereotactic biopsy procedures have an estimated 19.1% complication rate, and in patients with PCNSL, open biopsy has a comparable complication rate.^{20,29} Accurate radiological diagnosis can therefore guide neurosurgical decision-making while minimizing invasive procedures that may not be warranted, thus optimizing patient outcomes, quality of care, and cost-effectiveness.⁴⁰

Machine learning (ML) is the application of probabilistic algorithms to train a computational model to make predictions.^{9,11} ML is a form of artificial intelligence with the potential to become a valuable clinical tool as it can be utilized in diagnosis, treatment optimization, and prediction of outcomes.^{4,6,14,21,22} Of particular interest here is the application of ML to the interpretation of MRI data to correctly diagnose GBM and PCNSL.

A previous study reviewing the performance of ML compared to expert clinicians in the field of neurosurgery determined that more often than not, ML algorithms performed better than clinicians as measured by accuracy, area under the receiver operating curve (AUC), and other performance measures such as sensitivity and specificity.²⁹ Researchers have predicted that the incorporation of ML into clinical practice may begin with pattern recognition in radiological imaging and diagnosis.³¹ Thus, we sought to systematically review the literature and conduct a meta-analysis of ML in creating classifiers capable of differentiating GBM from PCNSL on radiological imaging.

Methods

Literature Review

This study was registered with the International Prospective Register of Systematic Reviews (PROSPERO, no. CRD42018098563) and conducted in concordance with Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines.¹⁸ The primary literature search involved querying the PubMed database to identify neurosurgical literature published in English between January 1, 2008, and May 1, 2018, regarding ML algorithms trained to differentiate GBM from PCNSL on radiological imaging. To accomplish this, one set of terms for ML and one set of terms for brain tumors were overlapped to search the database (Table 1). These broad search criteria generated an extensive list of relevant literature.

Articles were included if they utilized ML algorithms and aimed to differentiate between GBM and PCNSL on radiological imaging. Articles were excluded if they were commentaries, editorials, letters, or case reports. Two authors (A.V.N. and E.E.B.) independently reviewed the articles returned by the PubMed query using the aforementioned inclusion and exclusion criteria to narrow down the relevant literature. The secondary literature search consisted of a combined reference list from the articles included in the primary literature search, and these papers were screened using the same initial inclusion and exclusion criteria. A third author (E.R.) served as the tiebreaker to resolve disagreements of article inclusion.

The qualifying papers were then reviewed, and various components of the studies were entered into tables prior to analysis. Elements of interest included study population characteristics; imaging modalities; ML algorithms used; methods of model training; performance measures such as accuracy, sensitivity, specificity, and AUC; and direct comparison to radiologist performance.

Risk of Bias Assessment

Risk of bias assessment for the systematic review was performed independently by two reviewers (E.E.B. and E.R.) using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool. This assessment tool is recommended by the Agency for Healthcare Research and Quality, the Cochrane Collaboration, and the United Kingdom National Institute for Health and Clinical Excellence in order to assess risk of bias along four domains: “patient selection,” “index tests,” “reference standard,” and “flow and timing.”^{24,27,35} Each of these domains contains 2–3 questions to answer the following important questions: could the selection of patients have introduced bias; could the conduct or interpretation of the index test have introduced bias; could the reference standard, its conduct, or its interpretation have introduced bias; and could the patient flow have introduced bias? Each of these domains is ranked as “low,” “high,” or “unclear,” based on a composite of the 2–4 questions prompted for each domain. Answers were assessed in each domain for all studies and scores by the two reviewers were compared after independent assessment. In articles in which there was both high and low risk of bias in a given domain (“yes” and “no” answers to questions), a conservative approach was taken, and the overall domain was rated as having a “high” risk of bias. If there was uncertain degree of bias (“unclear”) and low risk of bias (“yes”), again, a conservative approach was taken, and the ranking of bias was given as “unclear.” Discrepancies were resolved by consensus discussion after independent ranking. A study was considered to have random sampling of patients if this was specified by the authors and was ranked “unclear” if the randomization process was not specified. “Inappropriate exclusions” within a paper were determined by clinical judgment of the reviewers, and if there was uncertainty as to whether or not it was appropriate to exclude particular images, such as images with necrosis in GBM, then this question was ranked “unclear.” All patients were considered to be included in the analysis only if all data from the images from the included patients were used to train the data during ML.^{27,35}

Statistical Analysis

We also performed a meta-analysis of the performance of ML algorithms in differentiating PCNSL from GBM and compared ML to radiologists. All statistical analysis was completed using the statistical program in R (version 3.4.2) using the mada package, a freely available package that can be used to construct either bivariate or hierarchical summary receiver operating characteristic models, as recommended by the Cochrane Collaboration for meta-analyses of diagnostic tests.¹⁶

Sensitivities and specificities of studies that prospectively included radiologists were used to back-calculate the 2×2 tables for radiologists and ML models, which report the number of true positives (TPs), false negatives (FNs), false positives (FPs), and true negatives (TNs). If data provided in the paper was insufficient for analysis, we contacted the authors asking for their definition of TP and for the number of TPs, FNs, FPs, and TNs produced by each classifier. If papers described performance using receiver operating characteristic curves, we back-calculated possible sensitivities and specificities. We then calculated Youden's J statistic to find optimal cutoffs for the diagnostic entity, and the sensitivity and specificity at the cutoff was used to generate the 2×2 table. For studies describing multiple classifiers in either the ML or radiologist group, their performance was averaged.

We set "PCNSL" as the definition of TP and generated two matrices where each row was the average 2×2 table for a group per study. One matrix consisted of average radiologist performance per study, and the other matrix consisted of average ML-trained model per study. Because we used sensitivities and specificities, we constructed a bivariate model and conducted meta-regression. Meta-regression generates a logit-transformed sensitivity and FP rate for the ML group and radiologist group while also evaluating whether a statistically significant difference exists.

Results

Systematic Review

The primary literature search returned 951 works of interest. An additional 175 articles were identified in the secondary literature search. Of these articles, 8 were determined to adequately meet inclusion and exclusion criteria (Fig. 1).^{1,5,12,19,33,36–38}

Overall, 6 (75%) of the studies exclusively compared GBM to PCNSL, and 2 of the studies included other tumors as well (Table 2). All studies only investigated de novo tumors that were imaged using T1-/T2-weighted MRI, FLAIR, and/or diffusion-weighted MRI prior to steroid administration, biopsy, or any other treatment. Each study differed in its imaging modality of choice. Five of the studies were published in the last 5 years, and the first author's primary institution of all but 1 study was located in Asia. Three studies did not define what they considered to be a TP diagnosis. All studies reported at least one of the following: accuracy, sensitivity, specificity, or AUC (Table 3). The most commonly used ML algorithm (75% of the studies) was a support vector machine (SVM). All of the studies performed some version of cross-validation to train their ML algorithms, but only 1 of the studies externally validated their model.¹²

Yamashita et al. were the only ones to evaluate how ML results could aid a radiologist's diagnosis. Their data showed that through the use of ML, radiologists were able to significantly improve diagnostic accuracy, sensitivity, and specificity ($p < 0.005$).³⁷ A total of 4 studies prospectively assessed the ability of radiologists to distinguish between GBM and PCNSL, providing a comparison group for ML.^{1,12,33,37} In all 4 of these studies, ML algorithms performed as well as or better than radiologists (Table 3).

TABLE 1. Overview of terms used in systematic review search strategy

Subject	Key Terms
Machine Learning	MeSH Terms
	Machine learning
	Artificial intelligence
	Natural language processing
	Neural networks (computer)
	Support vector machine
	Within title or abstract
	Artificial intelligence
	Bayesian learning
	Boosting
	Computational intelligence
	Computer reasoning
	Deep learning
	Machine intelligence
	Machine learning
	Naive Bayes
	Neural network
	Neural networks
	Natural language processing
	Support vector*
	Random forest*
Brain Tumors	MeSH Terms
	Central nervous system neoplasms
	Neoplasms, neuroepithelial
	Nerve sheath neoplasms
	Skull base neoplasms
	Within title or abstract
	Brain metastases
	Brain metastasis
	Brain tumor
	Brain tumour
	Glioblastoma
	Glioma
Restrictions	Exclude publication type
	Case report
	Comment
	Editorial
	Letter

Neurosurgical literature published in English between January 1, 2008, and May 1, 2018.

* Represents truncation in PubMed so that all phrases starting with the term prior to the asterisk are included in the search.

The included studies were judged to have a high risk of bias overall, primarily due to the potential bias introduced by the retrospective nature of the methods used for patient inclusion. These results are summarized in Table 4, with detailed descriptions of bias assessment included in Table 5. Specifically, the patient selection for these trials was based on a case-control design because outcomes were known prior to implementation of ML. This feature of study design may be a bias shared by all ML ap-

FIG. 1. Flowchart depicting the systematic review PubMed search strategy and study selection in accordance with PRISMA guidelines.

TABLE 2. Summary of studies including country of first author's institution, tumors evaluated, training sample sizes, ML algorithms used to train models, and study methods

Study	Country	Tumors	Training Sample		ML Algorithm	Compared vs Radiologists?	Externally Validated?
			GBM	PCNSL			
Suh et al., 2018	South Korea	GBM & PCNSL	23	54	Random forest	Yes	No
Kang et al., 2018	South Korea	GBM & PCNSL	70	42	k-NN, naive-Bayes, decision tree, LDA, random forest, Adaboost, linear-SVM, RBF-SVM	Yes	Yes*
Chen et al., 2018	China	GBM & PCNSL	44	20	SVM	No	No
Alcaide-Leon et al., 2017	Canada	Gliomas (including GBM) & PCNSL	69†	35	SVM	Yes	No
Yang et al., 2017	China	GBM & PCNSL	58	37	SVM	No	No
Yamasaki et al., 2013	Japan	GBM & PCNSL	20	20	SVM	No‡	No
Liu et al., 2012	China	GBM & PCNSL	10	8	SVM	No	No
Yamashita et al., 2008	Japan	Gliomas (high & low grade), lymphoma, & metastases	58§	12	ANN	Yes	No

ANN = artificial neural network; k-NN = k-nearest neighbor; LDA = linear discriminant analysis; RBF = radial basis function kernel.

* Kang et al. stated that the external validation set was composed of 42 patients from another tertiary medical center that was not identified.

† Full glioma training set consisted of 71 patients, 2 of whom had grade III gliomas and thus not GBM.

‡ A radiologist diagnosed radiological images using apparent diffusion coefficient cutoffs, an objective quantification.

§ The nature of this training set was not described in the study, so it was impossible to tell how many images were of patients diagnosed with GBM and how many were of patients with grade III gliomas.

TABLE 3. Performance evaluation of ML algorithms versus radiologists in distinguishing GBM from PCNSL

Study	TP Definition	Machine Learning				Radiologists			
		Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
Suh et al., 2018	GBM*	89.6%	91.3%	88.9%	0.921	62.3%*	75.4%†	58.0%†	0.717
Kang et al., 2018	PCNSL	92.9%‡	84.1%	98.4%	0.979	85.7%‡	76.7%	96.7%	0.867
Chen et al., 2018	PCNSL*	95.3%	85.0%	100%	0.991	—	—	—	—
Alcaide-Leon et al., 2017	PCNSL	84.9%‡	74.3%§	90.1%§	0.878	84.9%‡	77.1%§	88.7%§	0.874
Yang et al., 2017	—	96.8%	—	—	—	—	—	—	—
Yamasaki et al., 2013	PCNSL	95.4%	—	—	—	—	—	—	—
Liu et al., 2012	GBM¶	99.3%	100%	98.8%	—	—	—	—	—
Yamashita et al., 2008	Correct diagnosis	—	—	—	0.949	86.9%	78.5%	89.7%	0.899

* Inferred from the provided sample sizes and performance measures reported in the paper.

† Back-calculated using receiver operating characteristic curve.

‡ Calculated from 2 × 2 tables of averaged TPs, TNs, FPs, and FNs.

§ Provided by Dr. Alcaide-Leon via correspondence.

¶ Possible discrepancy in definition; this discrepancy was inferred from the provided sample sizes and performance measures reported in the paper.

proaches, and methods aimed at mitigating this potential bias should be considered in future study designs aimed at training diagnostic ML algorithms. Removing these sources of bias may help optimize the ability of the algorithm to make an accurate diagnosis when histological diagnoses are not available. Additionally, in the second domain (“index tests”), the study designs for the papers examined necessarily had prior knowledge of the reference standard prior to implementing the index test, which introduces a high risk of bias. As noted previously, future studies of ML should attempt to remove this risk of bias as much as possible, ideally by utilizing a prospective design and external validation. As judged in domain 3, however, the reference standard of histological diagnosis was considered to provide an accurate classification of the target condition, although this reporting could be improved if the authors provided details regarding how the histological samples were obtained and processed and the specific histological characteristics that determined the diagnosis. Finally, most of the studies were unclear as to whether all eligible patients were included in the analysis or had different inclusion and exclusion criteria for immunocompetency, leaving most studies with an unclear amount of bias in the fourth domain, “flow and timing.”

Meta-Analysis

A bivariate model was successfully generated for 3 of the 4 studies comparing ML to radiologists (Fig. 2). The fourth study was excluded because a 2 × 2 table could not be generated with PCNSL as the definition of TP. We inferred that the definition of TP in the study of Suh et al. was GBM and thus had to switch sensitivity and specificity to redefine TP for our calculations. The bivariate model demonstrates significant overlap between the ML group and the radiologists group. However, the total area of the ML group is less than that of the radiologists group, which indicates less interrater variability of the ML-trained models compared to the radiologists. The bivariate model also allows for performance evaluation of each group separately, and the overall sensitivity and specificity of the

ML-trained models in diagnosing PCNSL are approximately 0.8 and 0.9, respectively (Fig. 2).

Comparing the logit-transformed sensitivity and FP rate between the ML group and the radiologists group, we found no statistically significant difference for either measure ($p = 0.252$ and $p = 0.257$, respectively). However, while the point estimate for FP rate was essentially 1, indicating a trend toward no effect, the point estimate for sensitivity was -0.423 when comparing radiologists to ML, indicating a trend toward ML performing with better sensitivity compared to radiologists.

Discussion

Senders et al. have evaluated ML application in neurosurgery as a general concept.^{29–31} In the present study we assessed the viability of ML utilization in a specific clinical scenario: discriminating between PCNSL and GBM on radiological imaging. We identified 8 studies that trained predictive models using ML to make a diagnosis

TABLE 4. Summary of QUADAS-2 tool assessment for all reviewed studies

Authors & Year	Risk of Bias			
	Patient Selection	Evaluation of Index Test	Evaluation of Reference Standard	Flow & Timing
Suh et al., 2018	+	+	–	+
Kang et al., 2018	+	?	–	?
Chen et al., 2018	+	+	?	?
Alcaide-Leon et al., 2017	+	+	?	–
Yang et al., 2017	+	?	?	?
Yamasaki et al., 2013	?	?	?	?
Liu et al., 2012	+	?	–	?
Yamashita et al., 2008	+	?	–	–

Risk of bias: + = high, – = low, ? = unclear.

TABLE 5. Specific breakdown of bias assessment domains for each study evaluating ML diagnostic tools in differentiating GBM from PCNSL

Authors & Year	Domains			
	1: Could the selection of patients have introduced bias?	2: Could the conduct or interpretation of the index test have introduced bias?	3: Could the reference standard, its conduct, or its interpretation have introduced bias?	4: Could the patient flow have introduced bias?
Suh et al., 2018	Neither a random sample of pts was enrolled, nor was a case-control design avoided, & there was uncertainty as to whether or not the study avoided inappropriate exclusions, thus a high risk of bias in this domain (N,N,U)	While it was unclear if a prespecified threshold was used in this study, the IT result was interpreted w/ prior knowledge of the ref standard, resulting in high risk of bias (N,U)	The ref standard likely correctly classified the target condition & the ref standard results were interpreted w/o knowledge of the results of the IT, resulting in low risk of bias (Y,Y)	While there was uncertainty as to whether or not there was an appropriate interval between the ITs & ref standards, all pts received a ref standard & all pts received the same ref standard; however, not all pts were included in the analysis, resulting in a high risk of bias (U,Y,Y,N)
Kang et al., 2018	Neither a random sample of pts was enrolled, nor was a case-control design avoided, but the study avoided inappropriate exclusions, thus a high risk of bias in this domain (N,N,Y)	While it was clear the IT was interpreted w/o knowledge of the ref standard, it was unclear whether or not a prespecified threshold was used in this study, resulting in unclear risk of bias (Y,U)	The ref standard likely correctly classified the target condition & the ref standard results were interpreted w/o knowledge of the results of the IT, resulting in low risk of bias (Y,Y)	While there was uncertainty as to whether or not there was an appropriate interval between the ITs & ref standards, all pts received a ref standard & all pts received the same ref standard; however, not all pts were included in the analysis, resulting in a high risk of bias (U,Y,Y,N)
Chen et al., 2018	It was unclear as to whether a random sample of pts was enrolled, or whether or not the study avoided inappropriate exclusions, & a case-control design was not avoided, thus a high risk of bias in this domain (U,N,U)	While it was unclear if a prespecified threshold was used in this study, the IT result was interpreted w/prior knowledge of the ref standard, resulting in high risk of bias (N,U)	It was unclear as to whether or not the ITs were interpreted w/o knowledge of the ref standard & if a threshold was used; thus, an overall assessment of unclear was given in this domain (U,U)	It was uncertain as to whether or not there was an appropriate interval between the IT & the ref standard; all of the pts received a ref standard, but it was unclear if they received the same ref standard; it was unclear if all of the pts were included in the analysis, thus this domain received an unclear risk of bias (U,Y,U,U)
Alcaide-Leon et al., 2017	While random sampling was provided for enrolled pts & the study avoided inappropriate exclusions, a case-control design was not avoided, thus a high risk of bias was given for this domain (Y,N,Y)	The IT results were interpreted w/prior knowledge of the ref standard, but a prespecified threshold was used to interpret the machine learning tests; thus, this domain received a high risk of bias (N,Y)	While the IT results were interpreted w/o prior knowledge of the ref standard, it was unclear as to whether or not the threshold was prespecified, giving this domain an unclear risk of bias (Y,U)	There was an appropriate interval between the IT & the ref standard, all of the pts received the same ref standard & all of the pts were included in the analysis, giving this domain a low risk of bias (Y,Y,Y,Y)
Yang et al., 2017	It was unclear as to whether a random sample of pts was enrolled, or whether or not the study avoided inappropriate exclusions, & a case-control design was not avoided, thus a high risk of bias in this domain (U,N,U)	While it was unclear as to whether or not the IT results were interpreted w/o prior knowledge of the ref standard, a prespecified threshold was used, giving this domain an overall unclear risk of bias (U,Y)	It was unclear as to whether or not the ITs were interpreted w/o knowledge of the ref standard & if a threshold was used; thus, an overall assessment of unclear was given in this domain (U,U)	It was uncertain as to whether or not there was an appropriate interval between the IT & the ref standard; all of the pts received a ref standard, but it was unclear if they received the same ref standard; it was unclear if all of the pts were included in the analysis; thus, this domain received an unclear risk of bias (U,Y,U,U)
Yamasaki et al., 2013	It was unclear as to whether a random sample of pts was enrolled, or whether or not the study avoided inappropriate exclusions, or whether case-control design was avoided, thus an unclear risk of bias in this domain was given (U,U,U)	While it was unclear as to whether or not the IT results were interpreted w/o prior knowledge of the ref standard, a prespecified threshold was used, giving this domain an overall unclear risk of bias (U,Y)	While the IT results were interpreted w/o prior knowledge of the ref standard, it was unclear as to whether or not the threshold was prespecified, giving this domain an unclear risk of bias (Y,U)	It was uncertain as to whether or not there was an appropriate interval between the IT & the ref standard, whether there was a ref standard, or if the same ref standard was used; all pts were included in the study, so this domain received an unclear risk of bias (U,U,U,Y)

CONTINUED ON PAGE 7 »

» CONTINUED FROM PAGE 6

TABLE 5. Specific breakdown of bias assessment domains for each study evaluating ML diagnostic tools in differentiating GBM from PCNSL

Authors & Year	Domains			
	1: Could the selection of patients have introduced bias?	2: Could the conduct or interpretation of the index test have introduced bias?	3: Could the reference standard, its conduct, or its interpretation have introduced bias?	4: Could the patient flow have introduced bias?
Liu et al., 2012	It was unclear whether or not pts were randomly enrolled, but a case-control design was not avoided; the study also avoided inappropriate exclusions; overall the study received an assessment of high bias in this domain (U,N,Y)	While it was unclear as to whether or not the IT results were interpreted w/o prior knowledge of the ref standard, a prespecified threshold was used, giving this domain an overall unclear risk of bias (U,Y)	The ref standard likely correctly classified the target condition & the ref standard results were interpreted w/o knowledge of the results of the IT, resulting in low risk of bias (Y,Y)	It was uncertain as to whether or not there was an appropriate interval between the IT & the ref standard; all of the pts received a ref standard, but it was unclear if they received the same ref standard; all pts were included in the study, so this domain received an unclear risk of bias (U,Y,U,Y)
Yamashita et al., 2008	It was unclear as to whether a random sample of pts was enrolled, or whether or not the study avoided inappropriate exclusions, & a case-control design was not avoided, thus a high risk of bias in this domain (U,N,U)	While it was unclear as to whether or not the IT results were interpreted w/o prior knowledge of the ref standard, a prespecified threshold was used, giving this domain an overall unclear risk of bias (U,Y)	The ref standard likely correctly classified the target condition & the ref standard results were interpreted w/o knowledge of the results of the IT, resulting in low risk of bias (Y,Y)	There was an appropriate interval between the IT & the ref standard, all of the pts received the same ref standard & all of the pts were included in the analysis, giving this domain a low risk of bias (Y,Y,Y,Y)

IT = index test; pts = patients; ref = reference.

N = no, U = unclear, Y = yes, per the QUADAS-2 questions for assessing risk of bias in diagnostic accuracy studies.

of either PCNSL or GBM based on features that could be extracted or inferred from imaging. These articles all trained their models using MR images, an imaging modality with relatively high 3D spatial resolution that is used in examination of soft tissue pathology.

At first glance, the preliminary results are promising. ML algorithms performed exceptionally well, with the lowest AUC being 0.878. Yamashita et al. even demonstrated a statistically significant improvement in radiologist performance when utilizing ML-trained model results. Furthermore, our meta-analysis of ML models demonstrated that they were more consistent than radiologists, as indicated by a smaller 95% confidence region (Fig. 2). The results also indicated a trend toward improved performance, but these were not statistically significant, as indicated by overlapping 95% confidence regions.

The noninferiority of ML-trained models in differentiating between GBM and PCNSL on radiological imaging then raises the question of why we do not immediately incorporate their use in clinical practice. However, these positive results must be interpreted with caution. A number of factors, including patient selection for inclusion into the training set and the use of data from single institutions, may have contributed to the development of model overfitting. Overfitting is a common issue in ML that decreases its utility in problem-solving and is the equivalent of confounding or selection bias in ML.^{8,10,17,32} As a consequence of overfitting, the ML algorithm becomes highly proficient in handling a specific set of situations, but unless it is trained using a highly heterogeneous population, it lacks the ability to successfully solve slightly differing cases. The ML algorithm may be invisibly biased by subtle differences between hospital sites, such as MRI hardware or

FIG. 2. Bivariate model displaying average optimal FP rate (1 – specificity) and TP rate (sensitivity) of 3 of 4 studies that prospectively evaluated ML versus radiologists in distinguishing GBM from PCNSL. The 95% confidence regions are shown for the ML models collectively (dotted line region) and for the radiologists (dashed line region).

imaging parameters, differences in histological technique, or underlying patient characteristics. Indeed, in the only study that applied their model to a data set from an external hospital, the algorithm performed worse on the external data set in terms of AUC, sensitivity, and specificity, indicating that the model was overfit to the training data from its home institution. The radiologists, on the other hand, performed roughly equivalently in both the internal and external cases.

To avoid overfitting and to adequately assess ML performance, multiple steps must be undertaken. Proper training should involve large sample sizes, k-fold cross validation, and multiple imaging modalities that the model could consider.^{7,13,34} Furthermore, the model should be trained to make nondichotomous outputs, as there is always the possibility, in prospective situations, that the radiological image is neither PCNSL nor GBM. Evaluation should include prospective radiologists' interpretation as a control comparison so that researchers can ensure that ML algorithms are an improvement upon current diagnostic modalities. Although this might appear to increase index test, reference standard, and flow bias, both "diagnostic tests" in this case would be evaluating the same already-collected information. Furthermore, ML focuses on pattern recognition and subsequent classification based on constellations of features. External validation is also a necessity, and only one of the reviewed papers externally validated its model. The above steps should absolutely be undertaken prior to considering integration of ML approaches into clinical practice.

Because the clinical and radiological presentations of GBM and PCNSL overlap considerably, in principle ML algorithms can be a tool to assist radiologists in approaching cases that may have features common to both PCNSL and GBM. The hope is that the clinical utility of ML will continue to increase as algorithms become more complex and as we continue to learn more about disease pathophysiology. Although currently imperfect, ML algorithms are likely to help increase the quality of neurosurgical care by decreasing the need for stereotactic and open biopsy in the future, thereby reducing the incidence of complications that compromise patient quality of life and life expectancy while expediting initiation of intervention. Although a couple of recent studies support craniotomy and resection as an alternative to biopsy of PCNSL in specifically selected patients, the results also indicated that patients with high risk scores did not benefit from resection.^{23,39} It is important to note that this study did not evaluate outcomes of patients who had definite PCNSL on radiological imaging and did not undergo subsequent biopsy, as biopsy and histological examination is still currently the gold standard. However, if ML models could output "confidence of diagnosis" their usability would increase and the need for biopsy and surgery may decrease. This would certainly help in cases in which cytoreductive surgery is not indicated, which should be reserved for carefully selected patients.²⁸ Radiologists, pathologists, and neurosurgeons could decide whether histological confirmation is necessary. In many cases, the question of "PCNSL or GBM" may become, "resection or nonsurgical treatment," bypassing the steps of biopsy, histological evaluation, and postsurgical patient recovery. This would undoubtedly help expert neurosurgeons to op-

timize both patient outcomes and the cost-effectiveness of neurosurgical care.

Although we did not focus on assessing the inputs given to the ML algorithms, we noticed much heterogeneity between studies in the steps leading up to classifier training. Studies varied between their form of MRI (i.e., T1-weighted, T2-weighted, diffusion-weighted, etc.), machine settings, tumor segmentation, and feature selection. This fact certainly decreases comparability and generalizability of the studies as a whole. Institutions vary in terms of their default settings and hardware capabilities. Future studies should consider allowing ML algorithms to classify image sets that may not necessarily include all the different forms of MRI. Another possibility would be for studies to standardize what forms are used, either using the most basic modality or the singly most applicable. Furthermore, investigators should consider having experts segment and contour tumors or utilize automatic segmentation algorithms such as "seed and grow." This would help to standardize the ML training and evaluation process. Another observation of interest is the fact that patient demographics and clinical history are available to radiologists but not ML algorithms. Future studies could consider incorporating these important data points with radiological features to improve the classification abilities of ML models.

Limitations of the Study

Application of ML in neuroradiology for solving the dilemma of whether an image depicts GBM or PCNSL is relatively new. There is currently a limited number of publications that address this scientific inquiry, although this may have been a result of our methods. One important limitation is that we only searched 1 database and attempted to expand our review by including the citations of each paper we initially included in our review. Our meta-analysis lacked the power necessary to detect any statistically significant differences between approaches. We were also unable to assess ML performance in external validation data sets, as only 1 study had externally validated their model. Being able to compare ML to radiologists in internal versus external validation processes would have produced the clearest results. However, we maintain that based on the principle of ML and previous studies, overfitting currently precludes the widespread use of ML in medicine. Future studies that address distinguishing GBM from PCNSL should prospectively evaluate the performance of their model against several radiologists in cases that have not been seen by either entity. Future studies should also consider the utility of newer MRI techniques such as MR spectroscopy that may improve differentiation of these two pathologies.

Additionally, our assessment of bias revealed inherent issues with applying the QUADAS-2 to ML studies. For example, the assessment of Yamashita et al. on how ML results could impact radiologists' diagnosis could have been interpreted as bias that affects domains 2 and 3. Furthermore, while domain 4 examines whether there is adequate time between the index test and reference standard, comparison of ML versus radiologists does not require a time interval as both groups utilize the same collected images. Despite these limitations, we maintain that assessment of bias is an

absolute necessity. We contend that the field would benefit from the development of a bias assessment tool tailored to studies evaluating diagnostic ML algorithms.

Conclusions

ML algorithms can be trained to perform a number of tasks and excel at pattern recognition, leading to the hypothesis that ML can be incorporated into diagnostic fields such as radiology and pathology. There has been recent interest in ML applications on differentiating between PCNSL and GBM on radiological imaging as evidenced by the growing body of research on the issue within the past 5 years. This review supports the idea that ML algorithms are a promising avenue for optimizing healthcare and improving patient outcomes. Formal predictive analytics, of which ML is the newest and possibly most complicated step, have the ability to improve clinician performance by supplementing human expertise and experience with computational power. However, developers must be careful to avoid the pitfall of overfitting: to maximize the generalizability and utility of ML models, they must be trained on large, heterogeneous data sets that account for the many possible differences encountered in real-world practice.

References

- Alcaide-Leon P, Dufort P, Geraldo AF, Alshafai L, Maralani PJ, Spears J, et al: Differentiation of enhancing glioma and primary central nervous system lymphoma by texture-based machine learning. *AJNR Am J Neuroradiol* **38**:1145–1150, 2017
- Bloch O, Han SJ, Cha S, Sun MZ, Aghi MK, McDermott MW, et al: Impact of extent of resection for recurrent glioblastoma on overall survival: clinical article. *J Neurosurg* **117**:1032–1038, 2012
- Bühning U, Herrlinger U, Krings T, Thies R, Weller M, Küker W: MRI features of primary central nervous system lymphomas at presentation. *Neurology* **57**:393–396, 2001
- Chen M, Hao Y, Hwang K, Wang L: Disease prediction by machine learning over big data from healthcare communities. *IEEE Access* **5**:8869–8879, 2017
- Chen Y, Li Z, Wu G, Yu J, Wang Y, Lv X, et al: Primary central nervous system lymphoma and glioblastoma differentiation based on conventional magnetic resonance imaging by high-throughput SIFT features. *Int J Neurosci* **128**:608–618, 2018
- Deo RC: Machine learning in medicine. *Circulation* **132**:1920–1930, 2015
- Erickson BJ, Korfiatis P, Kline TL, Akkus Z, Philbrick K, Weston AD: Deep learning in radiology: does one size fit all? *J Am Coll Radiol* **15** (3 Pt B):521–526, 2018
- Foster KR, Koprowski R, Skufca JD: Machine learning, medical diagnosis, and biomedical engineering research—commentary. *Biomed Eng Online* **13**:94, 2014
- Ghahramani Z: Probabilistic machine learning and artificial intelligence. *Nature* **521**:452–459, 2015
- Hawkins DM: The problem of overfitting. *J Chem Inf Comput Sci* **44**:1–12, 2004
- Jordan MI, Mitchell TM: Machine learning: trends, perspectives, and prospects. *Science* **349**:255–260, 2015
- Kang D, Park JE, Kim YH, Kim JH, Oh JY, Kim J, et al: Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: development and multicenter external validation. *Neuro Oncol* [pub ahead of print], 2018
- Kohli M, Prevedello LM, Filice RW, Geis JR: Implementing machine learning in radiology practice and research. *AJR Am J Roentgenol* **208**:754–760, 2017
- Kononenko I: Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med* **23**:89–109, 2001
- Küker W, Nägele T, Korfel A, Heckl S, Thiel E, Bamberg M, et al: Primary central nervous system lymphomas (PCNSL): MRI features at presentation in 100 patients. *J Neurooncol* **72**:169–177, 2005
- Leefflang MMG: Systematic reviews and meta-analyses of diagnostic test accuracy. *Clin Microbiol Infect* **20**:105–113, 2014
- Lemm S, Blankertz B, Dickhaus T, Müller KR: Introduction to machine learning for brain imaging. *Neuroimage* **56**:387–399, 2011
- Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* **6**:e1000100, 2009
- Liu YH, Muftah M, Das T, Bai L, Robson K, Auer D: Classification of MR tumor images based on Gabor wavelet analysis. *J Med Biol Eng* **32**:22–28, 2012
- Lu Y, Yeung C, Radmanesh A, Wiemann R, Black PM, Golby AJ: Comparative effectiveness of frame-based, frameless, and intraoperative magnetic resonance imaging-guided brain biopsy techniques. *World Neurosurg* **83**:261–268, 2015
- Manlhiot C: Machine learning for predictive analytics in medicine: real opportunity or overblown hype? *Eur Heart J Cardiovasc Imaging* **19**:727–728, 2018
- Noble WS: What is a support vector machine? *Nat Biotechnol* **24**:1565–1567, 2006
- Rae AI, Mehta A, Cloney M, Kinslow CJ, Wang TJC, Bhagat G, et al: Craniotomy and survival for primary central nervous system lymphoma. *Neurosurgery* [pub ahead of print], 2018
- Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ: Assessing methodological quality, in Deeks JJ, Bossuyt PM, Gatsonis C (eds): *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0*. Oxford: The Cochrane Collaboration, 2009
- Ricard D, Idhah A, Ducray F, Lahutte M, Hoang-Xuan K, Delattre JY: Primary brain tumours in adults. *Lancet* **379**:1984–1996, 2012
- Rubenstein JL, Gupta NK, Mannis GN, Lamarre AK, Treseler P: How I treat CNS lymphomas. *Blood* **122**:2318–2330, 2013
- Santaguida PL, Riley CM, Matchar DB: Assessing risk of bias as a domain of quality in medical test studies, in Chang SM, Matchar DB, Smetana GW, et al (eds): *Methods Guide for Medical Test Reviews [Internet]*. Rockville, MD: Agency for Healthcare Research and Quality (US), 2012
- Sarpong Y, Litofsky NS: When less is more—the value of stereotactic biopsy for diagnosis in the era of cytoreductive neuro-oncology. *J Tumor* **4**:374–377, 2016
- Senders JT, Arnaut O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, et al: Natural and artificial intelligence in neurosurgery: a systematic review. *Neurosurgery* **83**:181–192, 2018
- Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, et al: Machine learning and neurosurgical outcome prediction: a systematic review. *World Neurosurg* **109**:476–486, 2018
- Senders JT, Zaki MM, Karhade AV, Chang B, Gormley WB, Broekman ML, et al: An introduction and overview of machine learning in neurosurgical care. *Acta Neurochir (Wien)* **160**:29–38, 2018
- Subramanian J, Simon R: Overfitting in prediction models—is it a problem only in high dimensions? *Contemp Clin Trials* **36**:636–641, 2013

33. Suh HB, Choi YS, Bae S, Ahn SS, Chang JH, Kang SG, et al: Primary central nervous system lymphoma and atypical glioblastoma: differentiation using radiomics approach. **Eur Radiol** **28**:3832–3839, 2018
34. Waljee AK, Higgins PDR, Singal AG: A primer on predictive models. **Clin Transl Gastroenterol** **5**:e44, 2014
35. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al: QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. **Ann Intern Med** **155**:529–536, 2011
36. Yamasaki T, Chen T, Hirai T, Murakami R. Classification of cerebral lymphomas and glioblastomas featuring luminance distribution analysis. **Comput Math Methods Med** **2013**:619658, 2013
37. Yamashita K, Yoshiura T, Arimura H, Mihara F, Noguchi T, Hiwatashi A, et al: Performance evaluation of radiologists with artificial neural network for differential diagnosis of intra-axial cerebral tumors on MR images. **AJNR Am J Neuroradiol** **29**:1153–1158, 2008
38. Yang Z, Feng P, Wen T, Wan M, Hong X: Differentiation of glioblastoma and lymphoma using feature extraction and support vector machine. **CNS Neurol Disord Drug Targets** **16**:160–168, 2017
39. Yun J, Yang J, Cloney M, Mehta A, Singh S, Iwamoto FM, et al: Assessing the safety of craniotomy for resection of primary central nervous system lymphoma: a Nationwide Inpatient Sample analysis. **Front Neurol** **8**:478, 2017
40. Zusman EE, Benzil DL: The continuum of neurosurgical care: increasing the neurosurgeon's role and responsibility. **Neurosurgery** **80** (4 Suppl):S34–S41, 2017

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: all authors. Acquisition of data: Nguyen, Blears, Ross. Analysis and interpretation of data: Nguyen, Blears, Ross. Drafting the article: all authors. Critically revising the article: all authors. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Lall. Statistical analysis: Nguyen. Administrative/technical/material support: Nguyen, Ortega-Barnett. Study supervision: Lall, Nguyen, Ortega-Barnett.

Correspondence

Rishi Lall: University of Texas Medical Branch, Galveston, TX. rilall@utmb.edu.