AJNR

# Machine Learning in Differentiating Gliomas from Primary CNS Lymphomas: A Systematic Review, Reporting Quality, and Risk of Bias Assessment

G.I. Cassinelli Petersen, J. Shatalov, T. Verma, W.R. Brim, H. Subramanian, A. Brackett, R.C. Bahar, S. Merkaj, T. Zeevi, L.H. Staib, J. Cui, A. Omuro, R.A. Bronen, A. Malhotra and M.S. Aboian



*AJNR Am J Neuroradiol* 2022, 43 (4) 526-533
doi: https://doi.org/10.3174/ajnr.A7473
http://www.ajnr.org/content/43/4/526

# Machine Learning in Differentiating Gliomas from Primary CNS Lymphomas: A Systematic Review, Reporting Quality, and Risk of Bias Assessment

G.I. Cassinelli Petersen, J. Shatalov, T. Verma, W.R. Brim, H. Subramanian, A. Brackett, R.C. Bahar, S. Merkaj, T. Zeevi, L.H. Staib, J. Cui, A. Omuro, R.A. Bronen, A. Malhotra, and M.S. Aboian

## ABSTRACT

**BACKGROUND:** Differentiating gliomas and primary CNS lymphoma represents a diagnostic challenge with important therapeutic ramifications. Biopsy is the preferred method of diagnosis, while MR imaging in conjunction with machine learning has shown promising results in differentiating these tumors.

**PURPOSE:** Our aim was to evaluate the quality of reporting and risk of bias, assess data bases with which the machine learning classification algorithms were developed, the algorithms themselves, and their performance.

**DATA SOURCES:** Ovid EMBASE, Ovid MEDLINE, Cochrane Central Register of Controlled Trials, and the Web of Science Core Collection were searched according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines.

**STUDY SELECTION:** From 11,727 studies, 23 peer-reviewed studies used machine learning to differentiate primary CNS lymphoma from gliomas in 2276 patients.

**DATA ANALYSIS:** Characteristics of data sets and machine learning algorithms were extracted. A meta-analysis on a subset of studies was performed. Reporting quality and risk of bias were assessed using the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) and Prediction Model Study Risk Of Bias Assessment Tool.

**DATA SYNTHESIS:** The highest area under the receiver operating characteristic curve (0.961) and accuracy (91.2%) in external validation were achieved by logistic regression and support vector machines models using conventional radiomic features. Meta-analysis of machine learning classifiers using these features yielded a mean area under the receiver operating characteristic curve of 0.944 (95% CI, 0.898–0.99). The median TRIPOD score was 51.7%. The risk of bias was high for 16 studies.

**LIMITATIONS:** Exclusion of abstracts decreased the sensitivity in evaluating all published studies. Meta-analysis had high heterogeneity.

**CONCLUSIONS:** Machine learning–based methods of differentiating primary CNS lymphoma from gliomas have shown great potential, but most studies lack large, balanced data sets and external validation. Assessment of the studies identified multiple deficiencies in reporting quality and risk of bias. These factors reduce the generalizability and reproducibility of the findings.

**ABBREVIATIONS:** AI = artificial intelligence; AUC = area under the receiver operating characteristic curve; CNN = convolutional neural network; ML = machine learning; PCNSL = primary CNS lymphoma; PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROBAST = Prediction model study Risk Of Bias Assessment Tool; TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Gliomas are the most common primary malignancy of the CNS.[1] An important differential diagnosis for gliomas is primary CNS lymphoma (PCNSL), a more uncommon but highly malignant neoplasia.[2] Correct differentiation of these tumor entities is an important challenge for clinicians because

Received July 26, 2021; accepted after revision January 31, 2022.

From the Department of Radiology and Biomedical Imaging (G.I.C.P., T.V., H.S., R.C.B., S.M., T.Z., L.H.S., J.C., R.A.B., A.M., M.S.A.), Cushing/Whitney Medical Library (A.B.), and Department of Neurology (A.O.), Yale School of Medicine, New Haven, Connecticut; Universitätsmedizin Göttingen (G.I.C.P.), Göttingen, Germany; University of Richmond (J.S.), Richmond, Virginia; New York University (T.V.), New York, New York; and Whiting School of Engineering (W.R.B.), Johns Hopkins University, Baltimore, Maryland.

This work was supported by a American Society of Neuroradiology Fellow Award 2018 (M.S.A.). This publication was made possible by KL2 TR001862 from the National Center for Advancing Translational Science, components of the National Institutes of Health and National Institutes of Health Roadmap for Medical Research.

Its contents are solely the responsibility of the authors and do not necessarily represent the official view of National Institutes of Health.

Please address correspondence to Mariam Aboian, MD, PhD, Department of Radiology, Yale School of Medicine, 333 Cedar St, New Haven, 06510 CT; e-mail: mariam.aboian@yale.edu; @GabrielCassine1

Indicates open access to non-subscribers at www.ajnr.org

Indicates article with online supplemental data.

http://dx.doi.org/10.3174/ajnr.A7473

therapy differs vastly: High-grade gliomas are treated with surgery and adjuvant radiochemotherapy,[3] while standard PCNSL treatment consists of high-dose methotrexate chemotherapy.[4,5] Surgery, in the latter group, is mostly reserved for either biopsy and decompressive surgery in cases of increased intracranial pressure.[6] Currently, the standard diagnostic approach for suspected PCNSL consists of stereotactic biopsy and histopathologic analysis.[7] Nonetheless, this diagnostic method has morbidity and mortality rates of up to 6% and 3%, respectively.[8,9] Furthermore, while maximum surgical resection is the standard-of-care initial treatment for gliomas, its effectiveness in treating PCNSL has yet to be convincingly demonstrated.[4,10] Therefore, surgical biopsy poses important risks and yields no benefit besides histopathologic diagnosis. In this context, a noninvasive diagnostic procedure would be beneficial. An important candidate for this is artificial intelligence (AI)–assisted radiologic diagnosis.

PCNSL typically appears as a homogeneously contrast-enhancing parenchymal mass without necrosis,[11] while glioblastoma as an intra-axial tumor with irregular infiltrative margins and a central heterogeneously enhancing core, reflecting necrosis and hemorrhage.[12,13] While these qualitative features provide valuable clues for differentiation in typical cases, there are atypical presentations: PCNSL with ring-enhancing lesions and central necrosis can be observed in up to 13% of non-AIDS- and up to 75% of AIDS-related cases.[11]

An important tool that has recently emerged to improve the radiologic diagnosis is machine learning (ML). ML pipelines learn quantitative image features that are not visible to the human eye and correlate them to a clinical outcome.[14] In the past decades, considerable effort has been put into developing ML-based classification algorithms for differentiating gliomas and PCNSLs. This work has led to much data that should be identified, systematically evaluated, and synthesized. So far, 1 systematic review on this topic has been presented by Nguyen et al,[15] in 2018, but it was performed only on a single bibliographic data base. Prior studies have shown that single data base searches are insensitive and limit the scope of systematic reviews.[16] Therefore, we performed a more comprehensive search using 4 established data bases and wider-reaching keywords.

In this systematic review, we synthesized and evaluated the quality of reporting, risk of bias, data bases, algorithms, and their performance achieved thus far. We hope to provide an accurate picture of the current state of development, identifying shortcomings and providing recommendations to increase model performance, reproducibility, and generalizability to enable implementation into routine clinical practice.

## MATERIALS AND METHODS

### Search Strategy and Information Sources

This systematic review was performed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[17] The study was registered with the International Prospective Register of Systematic Reviews (PROSPERO, CRD42020209938). A data base search of Ovid EMBASE, Ovid MEDLINE, and the Cochrane Central Register of Controlled Trials (CENTRAL), and the Web of Science Core Collection was performed by a clinical librarian from anytime until February 2021. The search strategy included the following keywords and controlled vocabulary combining the terms for the following: "AI," "machine learning," "deep learning," "radiomics," "MR imaging," "glioma" as well as related terms (Online Supplemental Data). The search strategy was independently reviewed by a second institutional librarian. All publications were screened on Covidence (Veritas Health Innovation) software by a neuroradiology attending physician, a radiology resident, an AI graduate student, and a senior medical student.

### Selection Process and Eligibility Criteria

To select relevant studies, the 4 reviewers undertook the following steps independently: Initially, after duplicate removal, all study abstracts were screened to exclude studies not pertaining to neuro-oncology or not using ML methods. Next, full-text review was performed to exclude publications that met the following criteria: 1) were only abstracts; 2) were not original articles; 3) did not involve artificial intelligence or ML; 4) did not involve gliomas; 5) were not done on humans; 6) were not performed with either MR imaging, PET, or MR spectroscopy; and 7) were not in English. Lastly, only studies evaluating differentiation of gliomas versus PCNSL were included for data extraction. In an initial search, studies that used only logistic regression were excluded. These studies were, however, later included by filtering the excluded studies in Covidence by the terms "lymphoma" and "pcnsl." Here, studies that used logistic regression and differentiated gliomas from PCNSL were selected after abstract screening and full-text review. When disagreement between reviewers occurred, the neuroradiology attending physician made the final decision.

### Data-Collection Process and Data Items

Data was extracted independently by 2 reviewers using a custom-built data-extraction form (Online Supplemental Data). Disagreement was resolved by reaching a consensus through discussion. Data was collected on 1) the report (title, authors, year); 2) the patient characteristics (number of patients included, source of data, glioma/PCNSL case ratio, immune status of the patients with lymphoma, percentage of patients in training and testing, and use of an independent test cohort); 3) the tumor type studied and the definition of ground-truth (type of glioma, criterion standard for diagnosis); 4) the ML method used (classic ML or deep learning, algorithms studied, type and number of features used); 5) the imaging procedures performed (type of imaging studies used, magnetic field strength of MR imaging machine, MR imaging sequence studied); and 6) performance metrics as described in detail below.

### Reporting Quality and Risk of Bias Assessment

Reporting quality and risk of bias assessment was performed independently by 2 reviewers using the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) checklist [18] and the Prediction model study Risk Of Bias Assessment Tool (PROBAST), respectively.[19] TRIPOD is composed of 77 individual questions that address 30 different scorable domains, 29 of which are applicable to our study after excluding item 11 as listed in the Online Supplemental Data. The final TRIPOD score was calculated as described in the

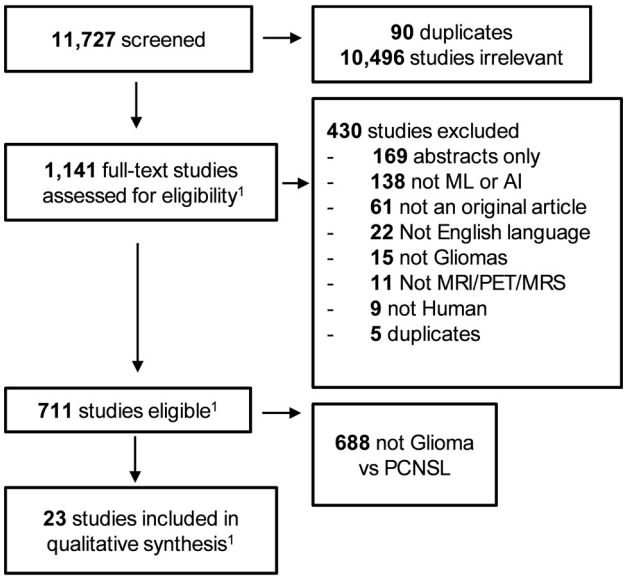


**FIG 1.** PRISMA flow diagram. This chart delineates the selection process that yielded the 23 studies included in this systematic review.[1] The initial search yielded 1135 studies for full-text review, and 704 were eligible, and 16 were included. A second literature search yielded 6 additional studies.

TRIPOD Adherence Assessment Form. For each study, the percentage of successfully reported TRIPOD items applicable to the individual study was reported. Additionally, for every item in the assessment, we report an adherence index, which we calculated as the average achieved across all studies. PROBAST is a checklist composed of 4 domains and 20 signaling questions, useful for assessing the risk of bias in multivariate diagnostic prediction models.[19] The Cohen's $\kappa$ was used to calculate the interrater reliability of the assessment between the 2 independent reviewers, and interpreted as delineated by Altman.[20]

### Data Analysis and Synthesis

To assess the performance of the classifiers from each study, we extracted primarily the reported area under the receiver operating characteristic curve (AUC) and its corresponding 95% confidence interval if available. Other threshold-based performance metrics that were extracted were accuracy, sensitivity, and specificity. Different studies test the interaction of classifiers with different feature-selection methods, resulting in many permutations of the same classifier. Only the results of the best performing version of each studied classifier were reported because we deemed this information most relevant. We grouped the performance metrics according to whether they were calculated during training, internal or external validation. To plot graphs, we used the performance on validation. If a study reported both internal and external validation, only external validation was plotted. Some studies compared ML models with the performance of different neuroradiologists. In these cases, we reported only the results of the highest performing radiologist, unless stated otherwise.

We performed a meta-analysis on the AUC values of a subset of studies that used conventional radiomic features and conventional ML algorithms for model development. Studies were only included if they reported an AUC with a 95% CI in a validation set and if they used conventional radiomic features for model development. Studies that used a deep learning–based classifier were also excluded in the meta-analysis. These exclusion criteria were chosen to decrease the methodic diversity and increase the comparability of the studies included in the meta-analysis. If both internal and external validation were reported, we used the performance on external validation. The meta-analysis used a random-effects model, as described by Zhou et al,[21] and was performed on MedCalc (MedCalc Software). The calculated heterogeneity among studies is reported using Higgins $I^2$, which describes the percentage of total variation attributable to heterogeneity rather than chance alone.[22]

## RESULTS

### Study Selection

The study-selection process is presented in Fig 1. The literature search yielded 11,727 studies. After duplicate removal, 10,496 studies were excluded, 1141 studies underwent full-text review, and finally 23 articles were included in our systematic review as per our criteria.[23-45] Of note, the selection process was performed in two steps since 6 studies that were finally included, were initially excluded solely because only a Logisitc Regression model was developed. Data was extracted from these studies for qualitative synthesis. An outline of the data sets and the developed ML pipelines of the individual studies can be found in the Online Supplemental Data.

### Data Sets for Model Development

The data sets had a mean size of 99 patients per study (range: 17–259 patients) (Fig 2*A*), with a mean ratio of 1.9 glioma cases for every PCNSL case (range: 7.9–0.4 cases), with only 2 studies having a 1:1 ratio (Fig 2*B*); 56.5% ($n = 13$) of the studies used data from single-center hospital data bases, and 17.4% ($n = 4$) used private multicenter hospital data bases. The source of patients could not be determined in 26.1% ($n = 6$) of articles (Fig 2*C*). No study used public brain tumor data sets such as Brain Tumor Segmentation (BraTS) or The Cancer Imaging Archive (TCIA).

More than half of the studies did not use external validation, instead relying on k-fold cross-validation or randomly sampling subjects into 2 cohorts, training and validation. Five studies did not report any type of validation (Fig 2*D*). Among the 6 studies that externally validated their algorithm, 4 sampled the external data set from a different institution (geographic validation);[28,30,33,43] and 2, on a different timepoint (temporal validation)[31,37] than the training set.

### Tumor Entities

All studies used PCNSL and gliomas in their data sets. Among the gliomas, all studies included glioblastomas: 2 included World Health Organization grade III gliomas;[23,41] and 1, lower-grade gliomas.[41] 5 also included meningiomas[34] and/or metastatic lesions.[33,34,36,45] 3 studies specified that they incorporated atypical glioblastomas, defined as glioblastomas without central necrosis,[29,35,39] while 3 explicitly included atypical PCNSLs.[28,30,39] We also investigated whether the immune status of patients with lymphoma was reported. 5 studies included only immunocompetent patients,[28,29,31,38,44] whereas 2 included both immunocompetent and immunosuppressed patients.[23,45] The remaining studies did

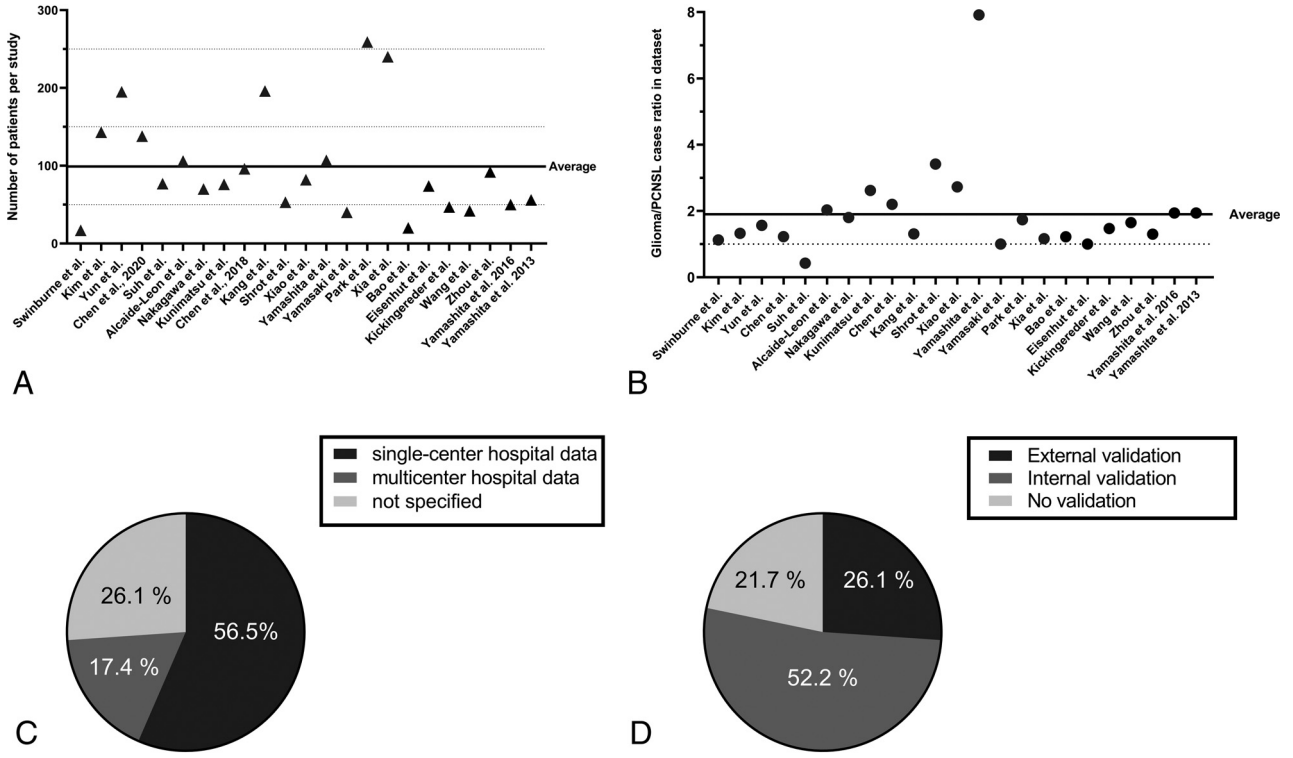


**FIG 2.** *A*, Scatterplot displaying the number of patients included in the data sets. *B*, Scatterplot of the glioma/PCNSL case ratio in the data sets. A ratio of >1 means more gliomas, a ratio of <1 means more PCNSL, and a ratio = 1 means equal number. *C*, Source of patients in the data sets. *D*, Type of validation performed in the studies.

not specify immunologic status. Importantly, all except 2 studies solely used images of tumors whose final diagnosis had been histopathologically confirmed. The other 2 combined histopathologic and clinicoradiologic criteria for diagnosis.[40,42]

### Image Features and Classification Algorithms

Nineteen studies used classic ML: 2 solely deep learning methods;[33,41] and 2 a combination of both.[36,43] Ten studies used combinations of shape and conventional radiomic features (first order, texture matrices, and wavelet-transformed images). Among these, the mean number of features used for model development was 29 (range, 3–80). A combination of diffusion and perfusion features was used in 8 studies,[24,27,29,34,36,40,42,45] while 1 also included SWI-derived features.[29] Other types of image features were used such as scale-invariant feature transform features,[26,46] luminance histogram range,[39] temporal patterns of time-signal intensity curves from DSC perfusion imaging extracted with the help of an autoencoder neural network,[33] and [$^{18}$F] PET–derived metrics.[40,42,44] After feature selection, the number of features ranged from 1 to 496.[26,36]

For classification, 10 different classic ML and 3 different deep learning algorithm types were used. The most common classic ML methods were support vector machines and logistic regression (each $n = 11$), a multilayer perceptron network ($n = 3$), and a convolutional neural network (CNN) ($n = 2$) for deep learning. Other algorithms were random forests ($n = 4$), decision tree ($n = 3$), Naïve Bayes ($n = 2$), linear discriminant analysis ($n = 2$), generalized linear model ($n = 2$), XGBoost ($n = 1$), AdaBoost ($n = 1$), and k-nearest neighbor ($n = 1$).

### Imaging

All studies except for 1 were performed on MR images. MR imaging sequences used were contrast-enhanced T1 (100% of studies performing MR imaging, $n = 22$), noncontrast T1 (50%, $n = 11$), T2 (59.1%, $n = 13$), FLAIR (50%, $n = 11$), DWI (68.2%, $n = 15$), intravoxel incoherent motion (4.6%, $n = 1$), and perfusion images (45.5%, $n = 10$). Three studies implemented [$^{18}$F] FDG PET/CT imaging.

### Model Performance and Meta-analysis

The reported metrics varied among different studies. AUC, accuracy, sensitivity, and specificity were reported in 91.3%, 65.2%, 73.9%, and 69.6% of the studies, respectively. The highest validation AUC of every study and respective 95% CI, if reported, are shown in the Online Supplemental Data. For a summary of the performance of every classifier by study, please refer to the Online Supplemental Data.

The classifiers that reached the highest AUC and accuracy in external validation were logistic regression[30] (AUC = 0.961) and a support vector machine[30] and logistic regression model[37] (both accuracy = 91.2%), respectively. All were trained on conventional radiomic features extracted from routine and DWI sequences. An XGBoost classifier[32] and a support vector machine classifier trained on scale-invariant feature transform features[26] were the only models that reached an AUC of >0.98 in internal validation but were not explored further in external validation.

Some studies compared the classification performance of ML models with that of radiologists tasked with comparing the same set of images,[23,28,32,35,43] and 2 studies examined the effect of

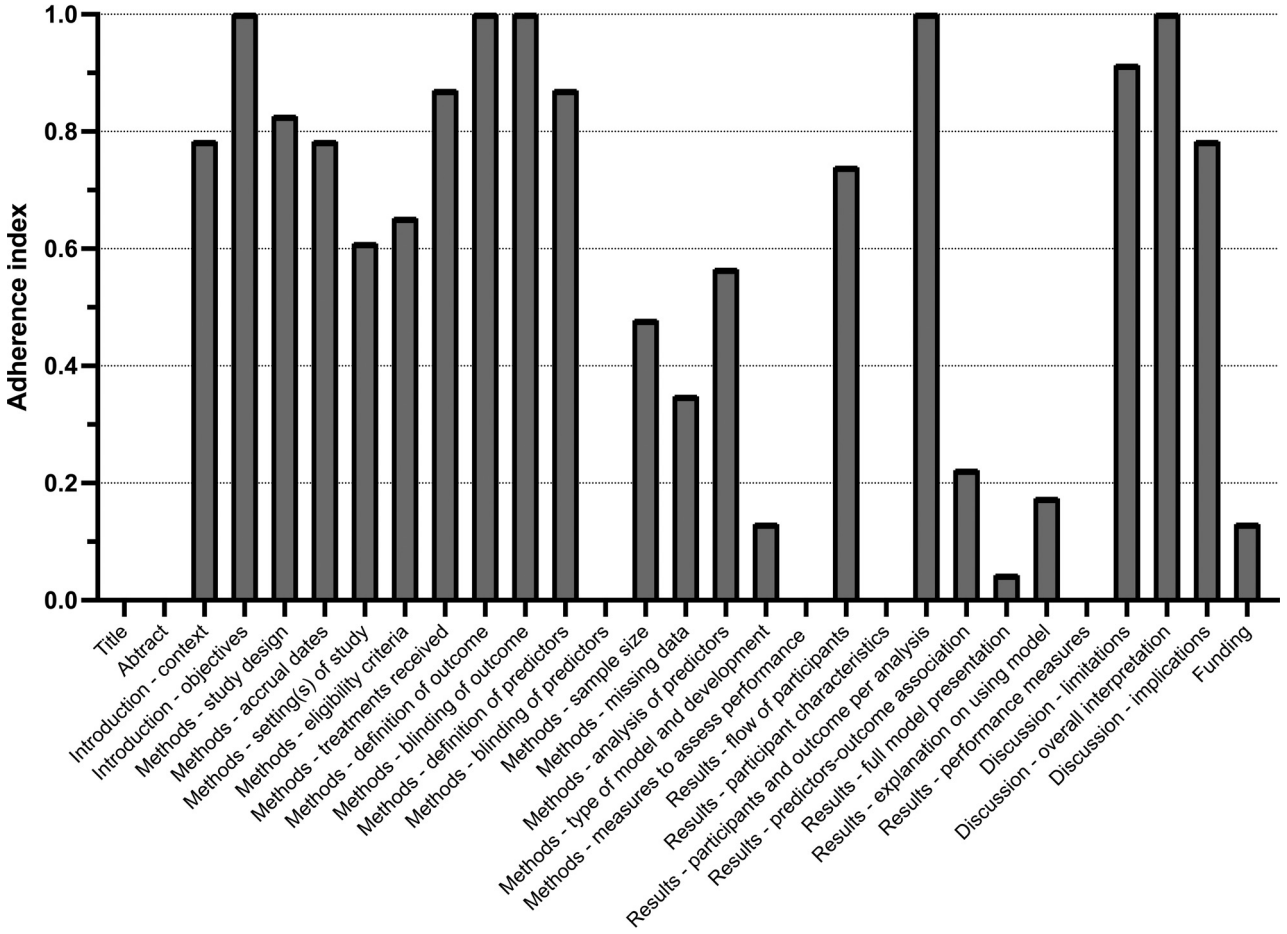


**FIG 3.** TRIPOD adherence index. The adherence index for a particular item was calculated as the average points achieved across all studies.

integrating the results of an ML algorithm in the radiologists' decision process (Online Supplemental Data).[37,41] Two publications found the ML algorithm superior,[32,35] while one found it significantly noninferior.[23] Both Yamashita et al[41] and Xia et al[37] found that incorporating ML models into the classification of novice radiologists significantly improved the AUC to levels comparable with their more experienced counterparts. Among experienced neuroradiologists, the effect was smaller-but-significant in 1 study.

Because conventional radiomics was the most used type of feature, we decided to conduct a random-effects AUC meta-analysis on a subset of studies that used these features in classic ML classifiers. We identified 6 studies that reported AUCs with confidence intervals in a validation test.[23,28,31,33,35,43] We excluded one because the radiomic features it used were not conventional[33] and one because its best classifier was a deep learning model.[43] In total, 4 studies were included in the meta-analysis.[23,28,31,35] The pooled AUC was calculated as 0.944 (95% CI, 0.918–0.980; $I^2$ = 74.3%). A forest plot of the meta-analysis can be seen in the Online Supplemental Data.

### *Adherence to Reporting Standards and Risk of Bias Assessment*

We performed a reporting quality assessment according to the TRIPOD checklist. Thirteen studies had an adherence index of <50%. Overall, the median TRIPOD score among all studies was 51.7% (interquartile range, 41.4%–62.1%). The individual adherence index for every item is shown in Fig 3 and the Online Supplemental Data. We performed a risk of bias assessment using the PROBAST tool. The overall risk of bias was deemed high in 69.6% ($n$ = 16) of studies and unclear in the rest. The risk of bias per PROBAST domain is further specified in the Online Supplemental Data. The interrater reliability between the 2 independent reviewers was very good in both the reporting quality ($\kappa$ = 0.965; 95% CI, 0.945–0.985) and risk of bias assessment ($\kappa$ = 0.851; 95% CI, 0.809–0.892).

## DISCUSSION

Our systematic review identified and analyzed 23 articles that published ML-based classification algorithms for noninvasive differentiation of gliomas and PCNSL.[23-45]

Analysis of study data sets revealed them to be predominantly small and unbalanced because glioma cases were overrepresented compared with PCNSL. This finding likely reflects the difficulty in sampling lymphoma cases due to their low prevalence. Moreover, a minority of studies validated their algorithm externally.[28,30,31,33,37,43] These factors decreased the generalizability of the findings and increased the risk of overlooking overfitted classifiers. Thus, we encourage multicenter collaborations to create larger, more balanced data sets. Additionally, cross-center

collaborations would facilitate the construction of geographically distinct external validation data sets on which to test these models.

We were also interested in the specific tumor entities that researchers used for model development. Strikingly, only a few articles specified the inclusion of atypical glioblastomas and lymphomas.[28-30,35,39] Considering that it is the atypical variants of the tumors that appear most similar, only including typical-appearing tumors might make classification easier without reflecting the everyday challenges faced by diagnosticians. Similarly, only 7 studies reported the immune status of the included patients with lymphoma.[23,28,29,31,38,44,45] Overall, we recommend inclusion of atypical cases in future data sets and clear reporting of their fraction and patients' immune status.

Classic ML classifiers trained on conventional radiomic features of routine sequences and DWI reached AUCs of >0.95 and the highest accuracies in external validation.[30,37] These findings, along with the high mean AUC in the meta-analysis, suggest that radiomic features extracted from conventional sequences are powerful in differentiating gliomas from PCNSL. This finding should make clinical implementation faster, considering that open-source packages for conventional radiomic feature extraction, like PyRadiomics,[47] are readily available. XGBoost, a decision tree–based algorithm popular among data scientists, performed very well in internal validation but was not tested on external validation.[32] Considering that random forest models (also decision-tree based) performed well in external validation, it would be reasonable to also expect good performance with XGBoost and hence encourage further research using this algorithm. These results are in line with other systematic reviews on ML in neuro-oncology. Our research group has also performed systematic reviews on the role of ML in predicting glioma grade and differentiating gliomas from brain metastases.[48,49] Both studies found, similar to our findings, a high mean accuracy despite small data sets. Overall, these findings are encouraging because they show that even though PCNSL is a rarer disease than other brain neoplasms, the development of ML applications for its diagnosis is on a par with that for other tumor entities.

Deep learning classifiers were explored by only 4 different studies.[33,36,41,43] Yun et al[43] developed a CNN-based model that showed good performance in internal validation (AUC = 0.879), but performance decreased drastically when externally validated (AUC = 0.486). CNNs, if not regularized properly, are prone to overfitting and benefit from large multisite data sets.[50] Using multiple sites facilitates larger data sets and incorporates valuable heterogeneity for training. Park et al[33] also developed a CNN-based model which achieved a higher AUC (0.89) in external validation. Overall, further evaluation of applications of CNN in the classification of gliomas from lymphomas in larger data sets is needed.

In recent years, the utility of ML algorithms as computer-aided diagnosis systems in oncologic practice has been repeatedly postulated.[51,52] By showing that ML can achieve a performance similar to that of radiologists (and sometimes even surpass them), the studies included in this systematic review support this notion.[37,41] Furthermore, Xia et al[37] and Yamashita et al[41] highlight the special utility of ML algorithms in helping radiologists in training achieve diagnostic performance comparable with that of their more experienced colleagues.

We performed a reporting quality assessment using the TRIPOD checklist.[18] TRIPOD addresses topics similar to those on the Checklist for Artificial Intelligence in Medical Imaging but is structured in 77 clearly defined questions and is, to our knowledge, the most comprehensive checklist for reporting quality assessment.[53] Adherence to reporting standards was generally low. Important shortcomings were found in reporting the full model to enable individual predictions, methods for measuring performance, the performance measures themselves, and incomplete disclosure of funding. Moreover, no study provided the programming code that was used to create the model, severely hindering reproducibility. Furthermore, no study reported calibration measurements, and only <50% reported confidence intervals of performance metrics, limiting the reader's ability to assess the achieved performance. These results are in line with a previously published systematic review that showed similar TRIPOD adherence indices in studies regarding radiomics in oncologic studies.[54] TRIPOD assessments were also performed in the above-mentioned systematic reviews from our group. Both studies found very similar TRIPOD adherence indices (44% and 48%) as well as similar deficiencies in the individual items.[55,56] Our results suggest that deficiencies in transparent reporting are a broader issue in the field of neuro-oncologic imaging.

We also performed a risk of bias assessment using the PROBAST tool.[19] PROBAST uses 20 signaling questions organized in 4 domains to assess the risk of bias related to the selection of participants, definition and measurement of predictors, definition and determination of outcomes, and quality of analysis methods in studies developing predictive diagnostic models.[19] While all studies included in this systematic review had a low risk of bias in the domains concerned with defining and measuring predictors and outcomes, a high proportion of high or unclear risk of bias was determined for most studies in participant selection and analysis. Regarding PROBAST Domain 1, the main concern rose from a selection of patients that did not represent the intended target population: Three studies excluded immunosuppressed patients;[31,38,44] and 1, hemorrhagic tumors,[24] likely skewing the participant population in the direction of typical patients and making discrimination easier for classifiers. The main concerns raised in Domain 4 were the low patient-to-feature ratio and the exclusion of participants with missing data in several studies. These factors have the potential of introducing bias because the former can lead to overfitting and thus to overestimation of performance metrics, while the latter is risky in small data sets because it can skew the patient population and render it not representative.[19] The risk of bias of several studies remained, nonetheless, unclear because of the several reporting deficiencies discussed above.

This systematic review had several limitations. First, by excluding studies that were presented only as abstracts, we reduced the sensitivity of our systematic review. We, nonetheless, accepted this loss of information because the inherent brevity of abstracts impedes a comprehensive appraisal of the study design, methods, and results.[57,58] Moreover, the developed pipelines and data sets are different and hence not always comparable. Using public brain tumor data sets, such as BraTS, could make comparisons between classifiers easier, though images in these data sets

are highly curated and might not reflect variable quality of images encountered in clinical practice. The meta-analysis was performed on a small subset of studies because most publications did not report sufficient data for statistical synthesis. Interestingly, the studies included in the meta-analysis showed high heterogeneity, reflecting the diversity of the ML model pipelines used. This level of heterogeneity is lower but comparable to one calculated in another published meta-analysis on ML in neuroradiological diagnosis.[59] The TRIPOD and PROBAST checklists are applicable to ML-based prediction models but were developed with conventional multivariate regression–based models in mind.[18,19,60] Due to the use of slightly different terminology and the lack of ML-based examples in both PROBAST's and TRIPOD's Elaboration and Examples document, the reporting quality assessment was burdensome at times. The TRIPOD and PROBAST creators have, however, acknowledged these shortcomings in a communication released in 2019 and announced the development of TRIPOD-AI and of PROBAST-AI.[60] We welcome and encourage this development to help improve transparent reporting and risk of bias assessment of ML-based prediction models.

## CONCLUSIONS

ML models for the differentiation of gliomas from PCNSL have great potential and have demonstrated high-level performance, sometimes even comparable with that of senior subspecialty-trained radiologists. ML models have also been shown to be powerful computer-aided diagnosis tools that can improve diagnostic performance, especially among junior radiologists. However, to be able to implement these into clinical practice, it is still necessary to perform further model development in larger, more balanced, and heterogeneous data sets that include other disease entities as well as test the robustness of models in external data sets. This more extensive development should increase the generalizability and reliability of the developed model. In addition, transparent reporting of model development should always be a priority, and we recommend adherence to the TRIPOD statement in future publications. This reporting will increase reproducibility, potentially enabling incorporation of these techniques into routine clinical practice.


## ACKNOWLEDGMENTS

We would like to acknowledge Thomas Mead, Mary Hughes, and Vermetha Polite from the Harvey Cushing/John Hay Whitney Medical Library for their support in doing this research.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.


## REFERENCES


1. Ostrom QT, Gittleman H, Liao P, et al. **CBTRUS Statistical Report: primary brain and other central nervous system tumors diagnosed in the United States in 2010-2014.** *Neuro Oncol* 2017;19:v1–88 CrossRef Medline
2. Villano JL, Koshy M, Shaikh H, et al. **Age, gender, and racial differences in incidence and survival in primary CNS lymphoma.** *Br J Cancer* 2011;105:1414–18 CrossRef Medline
3. Tan AC, Ashley DM, Lopez GY, et al. **Management of glioblastoma: state of the art and future directions.** *CA Cancer J Clin* 2020;70:299–312 CrossRef Medline
4. Hoang-Xuan K, Bessell E, Bromberg J, et al; European Association for Neuro-Oncology Task Force on Primary CNS Lymphoma, **Diagnosis and treatment of primary CNS lymphoma in immunocompetent patients: guidelines from the European Association for Neuro-Oncology.** *Lancet Oncol* 2015;16:e322–32 CrossRef Medline
5. Batchelor TT. **Primary central nervous system lymphoma: a curable disease.** *Hematol Oncol* 2019;37(Suppl 1):15–18 CrossRef Medline
6. Elder JB, Chen TC. **Surgical interventions for primary central nervous system lymphoma.** *Neurosurg Focus* 2006;21:E13 CrossRef Medline
7. Yang H, Xun Y, Yang A, et al. **Advances and challenges in the treatment of primary central nervous system lymphoma.** *J Cell Physiol* 2020;235:9143–65 CrossRef Medline
8. Malikova H, Liscak R, Latnerova I, et al. **Complications of MRI-guided stereotactic biopsy of brain lymphoma.** *Neuro Endocrinol Lett* 2014;35:613–18 Medline
9. Malone H, Yang J, Hershman DL, et al. **Complications following stereotactic needle biopsy of intracranial tumors.** *World Neurosurg* 2015;84:1084–89 CrossRef Medline
10. Weller M, Martus P, Roth P, et al; German PCNSL Study Group. **Surgery for primary CNS lymphoma? Challenging a paradigm.** *Neuro Oncol* 2012;14:1481–84 CrossRef Medline
11. Haldorsen IS, Espeland A, Larsson EM. **Central nervous system lymphoma: characteristic findings on traditional and advanced imaging.** *AJNR Am J Neuroradiol* 2011;32:984–92 CrossRef Medline
12. Yuguang L, Meng L, Shugan Z, et al. **Intracranial tumoural haemorrhage: a report of 58 cases.** *J Clin Neurosci* 2002;9:637–39 CrossRef Medline
13. Villanueva-Meyer JE, Mabray MC, Cha S. **Current clinical brain tumor imaging.** *Neurosurgery* 2017;81:397–415 CrossRef Medline
14. Wang S, Summers RM. **Machine learning and radiology.** *Med Image Anal* 2012;16:933–51 CrossRef Medline
15. Nguyen AV, Blears EE, Ross E, et al. **Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: a systematic review and meta-analysis.** *Neurosurg Focus* 2018;45:E5 CrossRef Medline
16. Whiting P, Westwood M, Burke M, et al. **Systematic reviews of test accuracy should search a range of databases to identify primary studies.** *J Clin Epidemiol* 2008;61:357–64 Medline
17. Page MJ, McKenzie JE, Bossuyt PM, et al. **The PRISMA 2020 statement: an updated guideline for reporting systematic reviews.** *BMJ* 2021;372:n71 CrossRef Medline
18. Collins GS, Reitsma JB, Altman DG, et al. **Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.** *BMJ* 2015;350:g7594 CrossRef Medline
19. Moons KG, Wolff RF, Riley RD, et al. **PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration.** *Ann Intern Med* 2019;170:W1–33 CrossRef Medline
20. Altman DG. *Practical Statistics for Medical Research*. Chapman and Hall; 1991
21. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley; 2011
22. Higgins JP, Thompson SG, Deeks JJ, et al. **Measuring inconsistency in meta-analyses.** *BMJ* 2003;327:557–60 CrossRef Medline
23. Alcaide-Leon P, Dufort P, Geraldo AF, et al. **Differentiation of enhancing glioma and primary central nervous system lymphoma by texture-based machine learning.** *AJNR Am J Neuroradiol* 2017;38:1145–50 CrossRef Medline
24. Bao S, Watanabe Y, Takahashi H, et al. **Differentiating between glioblastoma and primary CNS lymphoma using combined whole-tumor histogram analysis of the normalized cerebral blood volume and the apparent diffusion coefficient.** *Magn Reson Med Sci* 2019;18:53–61 CrossRef Medline

25. Chen C, Zheng A, Ou X, et al. **Comparison of radiomics-based machine-learning classifiers in diagnosis of glioblastoma from primary central nervous system lymphoma.** *Front Oncol* 2020;10:1151CrossRef Medline
26. Chen Y, Li Z, Wu G, et al. **Primary central nervous system lymphoma and glioblastoma differentiation based on conventional magnetic resonance imaging by high-throughput SIFT features.** *Int J Neurosci* 2018;128:608–18 CrossRef Medline
27. Eisenhut F, Schmidt MA, Putz F, et al. **Classification of primary cerebral lymphoma and glioblastoma featuring dynamic susceptibility contrast and apparent diffusion coefficient.** *Brain Sci* 2020;10:886 CrossRef Medline
28. Kang D, Park JE, Kim YH, et al. **Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: development and multicenter external validation.** *Neuro Oncol* 2018;20:1251–61 CrossRef Medline
29. Kickingereder P, Wiestler B, Sahm F, et al. **Primary central nervous system lymphoma and atypical glioblastoma: multiparametric differentiation by using diffusion-, perfusion-, and susceptibility-weighted MR imaging.** *Radiology* 2014;272:843–50 CrossRef Medline
30. Kim Y, Cho HH, Kim ST, et al. **Radiomics features to distinguish glioblastoma from primary central nervous system lymphoma on multi-parametric MRI.** *Neuroradiology* 2018;60:1297–1305 CrossRef Medline
31. Kunimatsu A, Kunimatsu N, Yasaka K, et al. **Machine learning-based texture analysis of contrast-enhanced MR imaging to differentiate between glioblastoma and primary central nervous system lymphoma.** *Magn Reson Med Sci* 2019;18:44–52 CrossRef Medline
32. Nakagawa M, Nakaura T, Namimoto T, et al. **Machine learning based on multi-parametric magnetic resonance imaging to differentiate glioblastoma multiforme from primary cerebral nervous system lymphoma.** *Eur J Radiol* 2018;108:147–54 CrossRef Medline
33. Park JE, Kim HS, Lee J, et al. **Deep-learned time-signal intensity pattern analysis using an autoencoder captures magnetic resonance perfusion heterogeneity for brain tumor differentiation.** *Sci Rep* 2020;10:21485 CrossRef Medline
34. Shrot S, Salhov M, Dvorski N, et al. **Application of MR morphologic, diffusion tensor, and perfusion imaging in the classification of brain tumors using machine learning scheme.** *Neuroradiology* 2019;61:757–65 CrossRef Medline
35. Suh HB, Choi YS, Bae S, et al. **Primary central nervous system lymphoma and atypical glioblastoma: differentiation using radiomics approach.** *Eur Radiol* 2018;28:3832–39 CrossRef Medline
36. Swinburne NC, Schefflein J, Sakai Y, et al. **Machine learning for semi-automated classification of glioblastoma, brain metastasis and central nervous system lymphoma using magnetic resonance advanced imaging.** *Ann Transl Med* 2019;7:232 CrossRef Medline
37. Xia W, Hu B, Li H, et al. **Multiparametric-MRI-based radiomics model for differentiating primary central nervous system lymphoma from glioblastoma: development and cross-vendor validation.** *J Magn Reson Imaging* 2021;53:242–50 CrossRef Medline
38. Xiao DD, Yan PF, Wang YX, et al. **Glioblastoma and primary central nervous system lymphoma: preoperative differentiation by using MRI-based 3D texture analysis.** *Clin Neurol Neurosurg* 2018;173:84–90 CrossRef Medline
39. Yamasaki T, Chen T, Hirai T, et al. **Classification of cerebral lymphomas and glioblastomas featuring luminance distribution analysis.** *Comput Math Methods Med* 2013;2013:619658 CrossRef Medline
40. Yamashita K, Hiwatashi A, Togao O, et al. **Diagnostic utility of intravoxel incoherent motion MR imaging in differentiating primary central nervous system lymphoma from glioblastoma multiforme.** *J Magn Reson Imaging* 2016;44:1256–61 CrossRef Medline
41. Yamashita K, Yoshiura T, Arimura H, 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* 2008;29:1153–58 CrossRef Medline
42. Yamashita K, Yoshiura T, Hiwatashi A, et al. **Differentiating primary CNS lymphoma from glioblastoma multiforme: assessment using arterial spin labeling, diffusion-weighted imaging, and (1)(8)F-fluorodeoxyglucose positron emission tomography.** *Neuroradiology* 2013;55:135–43 CrossRef Medline
43. Yun J, Park JE, Lee H, et al. **Radiomic features and multilayer perceptron network classifier: a robust MRI classification strategy for distinguishing glioblastoma from primary central nervous system lymphoma.** *Sci Rep* 2019;9:5746 CrossRef Medline
44. Zhou W, Wen J, Hua F, et al. **(18)F-FDG PET/CT in immunocompetent patients with primary central nervous system lymphoma: differentiation from glioblastoma and correlation with DWI.** *Eur J Radiol* 2018;104:26–32 CrossRef
45. Wang S, Kim S, Chawla S, et al. **Differentiation between glioblastomas, solitary brain metastases, and primary cerebral lymphomas using diffusion tensor and dynamic susceptibility contrast-enhanced MR imaging.** *AJNR Am J Neuroradiol* 2011;32:507–14 CrossRef Medline
46. Lowe DG. **Distinctive image features from scale-invariant keypoints.** *Int J Comput Vis* 2004;60:91–110 CrossRef
47. van Griethuysen JJM, Fedorov A, Parmar C, et al. **Computational radiomics system to decode the radiographic phenotype.** *Cancer Res* 2017;77:e104–07 CrossRef Medline
48. Bahar R, Merkaj S, Brim WR, et al. **NIMG-23: machine learning methods in glioma grade prediction: a systematic review.** *Neuro-Oncology* 2021;23:vi133 CrossRef
49. Brim WR, Jekel L, Petersen GC, et al. **OTHR-12. The development of machine learning algorithms for the differentiation of glioma and brain metastases: a systematic review.** *Neuro-Oncology Advances* 2021;3:iii17 CrossRef
50. Lee JG, Jun S, Cho YW, et al. **Deep learning in medical imaging: general overview.** *Korean J Radiology* 2017;18:570–84 CrossRef Medline
51. Takahashi R, Kajikawa Y. **Computer-aided diagnosis: a survey with bibliometric analysis.** *Int J Med Inform* 2017;101:58–67 CrossRef Medline
52. Booth TC, Williams M, Luis A, et al. **Machine learning and glioma imaging biomarkers.** *Clin Radiol* 2020;75:20–32 CrossRef Medline
53. Mongan J, Moy L, Kahn CE Jr. **Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers.** *Radiol Artif Intell* 2020;2:e200029 CrossRef Medline
54. Park JE, Kim D, Kim HS, et al. **Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement.** *Eur Radiol* 2020;30:523–36 CrossRef Medline
55. Jekel L, Brim WR, Petersen GC, et al. **OTHR-15. Assessment of TRIPOD adherence in articles developing machine learning models for differentiation of glioma from brain metastasis.** *Neuro-oncology Advances* 2021;3:17–18 CrossRef
56. Merkaj S, Bahar R, Brim W, et al. **NIMG-35: machine learning glioma grade prediction literature: a TRIPOD analysis of reporting quality.** *Neuro-Oncology* 2021;23:vi136 CrossRef
57. Scherer RW, Saldanha IJ. **How should systematic reviewers handle conference abstracts? A view from the trenches.** *Syst Rev* 2019;8:264 CrossRef Medline
58. Scherer RW, Sieving PC, Ervin AM, et al. **Can we depend on investigators to identify and register randomized controlled trials?** *PLoS One* 2012;7:e44183 CrossRef Medline
59. Bhandari AP, Liong R, Koppen J, et al. **Noninvasive determination of IDH and 1p19q status of lower-grade gliomas using MRI radiomics: a systematic review.** *AJNR Am J Neuroradiol* 2021;42:94–101 CrossRef Medline
60. Collins GS, Moons KG. **Reporting of artificial intelligence prediction models.** *Lancet* 2019;393:1577–79 CrossRef Medline