Natural and Artificial Intelligence in Neurosurgery: A Systematic Review

Joeky T. Senders, BS[*‡]
Omar Arnaout, MD[‡§]
Aditya V. Karhade, BS[‡]
Hormuzdiyar H. Dasenbrock, MD[‡]
William B. Gormley, MD, MPH, MBA[‡]
Marike L. Broekman, MD, PhD, JD[*‡]
Timothy R. Smith, MD, PhD, MPH[‡]

[*]Department of Neurosurgery, University Medical Center, Utrecht, the Netherlands; [‡]Cushing Neurosurgery Outcomes Center, Department of Neurosurgery, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; [§]Department of Neurological Surgery, Northwestern University School of Medicine, Chicago, Illinois

Correspondence:
Timothy R. Smith, MD, PhD, MPH,
Department of Neurosurgery,
Brigham and Women's Hospital,
15 Francis Street,
Boston, MA 02115.
E-mail: trsmith@partners.org

Received, February 9, 2017.
Accepted, August 11, 2017.
Published Online, September 7, 2017.

BACKGROUND: Machine learning (ML) is a domain of artificial intelligence that allows computer algorithms to learn from experience without being explicitly programmed.

OBJECTIVE: To summarize neurosurgical applications of ML where it has been compared to clinical expertise, here referred to as "natural intelligence."

METHODS: A systematic search was performed in the PubMed and Embase databases as of August 2016 to review all studies comparing the performance of various ML approaches with that of clinical experts in neurosurgical literature.

RESULTS: Twenty-three studies were identified that used ML algorithms for diagnosis, presurgical planning, or outcome prediction in neurosurgical patients. Compared to clinical experts, ML models demonstrated a median absolute improvement in accuracy and area under the receiver operating curve of 13% (interquartile range 4-21%) and 0.14 (interquartile range 0.07-0.21), respectively. In 29 (58%) of the 50 outcome measures for which a *P*-value was provided or calculated, ML models outperformed clinical experts ($P < .05$). In 18 of 50 (36%), no difference was seen between ML and expert performance ($P > .05$), while in 3 of 50 (6%) clinical experts outperformed ML models ($P < .05$). All 4 studies that compared clinicians assisted by ML models vs clinicians alone demonstrated a better performance in the first group.

CONCLUSION: We conclude that ML models have the potential to augment the decision-making capacity of clinicians in neurosurgical applications; however, significant hurdles remain associated with creating, validating, and deploying ML models in the clinical setting. Shifting from the preconceptions of a human-vs-machine to a human-and-machine paradigm could be essential to overcome these hurdles.

KEY WORDS: Artificial intelligence, Machine learning, Neurosurgery, Systematic review

Neurosurgery 83:181–192, 2018 DOI:10.1093/neuros/nyx384 www.neurosurgery-online.com

Artificial intelligence is the branch of computer science dealing with the simulation of intelligent behavior in computers.[1-3] Machine learning (ML) is a domain of artificial intelligence that allows computer algorithms to learn patterns by studying data directly without being explicitly programmed.[4,5] ML methods are already widely applied in multiple aspects of our daily lives, although this is not always obvious to the casual observer; common examples are email spam filters, search suggestions, online shopping suggestions, and speech recognition in smartphones.[3] Artificial intelligence by means of ML is entering the realm of medicine at an increasing pace and has been tested in a variety of clinical applications ranging from diagnosis to outcome prediction.[6,7]

ABBREVIATIONS: AUC, area under the curve; **CT,** computed tomography; **EHR,** electronic health records; **IQR,** interquartile range; **ML,** machine learning; **MRI,** magnetic resonance imaging; **NLP,** natural language processing; **PRISMA,** Preferred Reporting Items for Systematic Reviews and Meta-analysis; **TBI,** traumatic brain injury; **WHO,** World Health Organization

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

ML algorithms in general can be divided into supervised, unsupervised, and reinforcement learning algorithms, and each algorithm has its own mathematical structure (Table 1).[5-7] Reinforcement learning algorithms aim to determine the ideal behavior within a specific context based on simple reward feedback on their

TABLE 1. Frequently Used ML Algorithms Explained

ML algorithm	Mechanism	Advantages	Disadvantages
ANN—supervised learning	▷ Inspired on neural networks in the brain ▷ Organized in layers of interconnected nodes ▷ The nodes in the input layer and output layer represent the input features and target outputs, respectively ▷ The nodes in the "hidden" and output layers base the value of their output on the total input they receive	▷ Can model very complex relationships between input features and output ▷ No need to model the underlying data generating mechanism ▷ Robust to noise and incomplete data	▷ Difficult to interpret the explicit relationships between input features and outcome ▷ Prone to overfitting ▷ Long training times ▷ Requires significant memory and processing power for large data sets
SVM—supervised learning	▷ SVMs classify data points on their input features by calculating the ideal "separating hyperplane" ▷ SVMs select the hyperplane with the maximal distance to the nearest data point, "support vectors" ▷ A kernel function is mathematical trick that adds an extra dimension to the data ▷ Nonseparable 2-dimensional data, for example, could then be separated in a 3-dimensional space	▷ Can model very complex relationships between input features and output ▷ Effective in high-dimensional data ▷ Robust to noise and overfitting ▷ Fast fitting procedure	▷ Difficult to interpret the explicit relationships between input features and outcome ▷ Choosing a kernel function is a tuning parameter ▷ Requires significant memory and processing power
Decision tree—supervised learning	▷ Decision trees make predictions or classifications based on several input features with the use of bifurcating the feature space ▷ At each split of branches, training examples are divided based on class or value of the specific feature ▷ In ML, algorithms are used to find optimal features at which a split is made and the optimal value in case of a numeric feature	▷ Easy to interpret ▷ Fast ▷ Robust to noise and incomplete data ▷ Generally high accuracy performance among competitors	▷ Complex trees are hard to interpret ▷ Small variation in data can lead to different decision trees ▷ Does not work very well on a small training data set ▷ Prone to overfitting
KNN—supervised learning	▷ The algorithm compares a data point with unknown class to its K nearest neighbors, and determines its class as the most common class of its neighbors ▷ For the value K = 1, the algorithm compares the example with the single closest neighbor	▷ Easy to interpret ▷ Classes do not need to be linearly separable ▷ Naturally handles multiclass classification and regression ▷ Robust to noisy training data ▷ Computation only required at the time of evaluation	▷ Performs poorly on high dimensionality datasets ▷ Must define a meaningful distance function (distance function determines which neighbors will be considered closest) ▷ The value of K has a large effect on the behavior of this model.
Naïve Bayes—supervised learning	▷ Naïve Bayes calculates the most likely output based on the input features and the a priori chance. ▷ The probability of a certain outcome is the product of probabilities given by the individual features ▷ It assumes that the presence (or absence) of a feature is unrelated to the presence of any other feature	▷ Easy to interpret ▷ Easy to construct ▷ Fast ▷ Robust to overfitting	▷ Performs less in low-dimensional data ▷ It assumes that the presence of a feature is independent of the presence of other features, which rarely is the case in life

TABLE 1. Continued

ML algorithm	Mechanism	Advantages	Disadvantages
FCM—unsupervised learning	▷ FCM is an unsupervised learning algorithm that clusters data points based on their input features without having a desired output ▷ The 'fuzzy' aspect gives the algorithm the flexibility to classify a data point to each cluster to a certain degree relating to the likelihood of belonging to that cluster	▷ Allows data points to be in multiple clusters ▷ Could detect hidden patterns across large and complex datasets ▷ No training needed because there is no desired output	▷ The number of clusters (C) is a tuning parameter ▷ Need to determine membership cutoff value ▷ Sensitive to outliers

Abbreviations: ANN: artificial neural network; FCM: fuzzy C-means; KNN: K-nearest neighbors; ML: machine learning. SVM: support vector machine.

actions; the self-driving car is a typical example. Reinforcement learning is, however, beyond the scope of the current review. Supervised learning algorithms are trained on prelabeled data referred to as the training set.[7] This training process is an iterative process in which ML algorithms try to find the optimal combination of variables and weights given to the input variables (referred to as features) of the model with the goal of minimizing the training error as judged by the difference between predicted outcome and actual outcome.[6] A model with high error due to bias can fail to capture the regularities in the data, resulting in an inaccurate model underfitting the data. Increasing the complexity of the model, such as adding more parameters in the model, can reduce this bias. However, an excessively complex model, such as having too many parameters compared to the number of patients, can describe random error or noise instead of the meaningful relationships, referred to as overfitting of the data. This results in an increase in error due to variance and a reduced generalizability to previously unseen data. The complexity of a model should, therefore, be a tradeoff between bias and variance.[3] To avoid overfitting, predictive models must be validated on a test set that was not involved in the training process.

For unsupervised learning techniques on the other hand, no prelabeling is required. These algorithms cluster data points based on similarities in features and can be powerful tools for detecting previously unknown patters in multidimensional data.[6] The progress in this field of applied ML is continuously driven by the growing amount of available data and the increasing computational power.[8,9]

ML is increasingly tested in neurosurgical applications and even demonstrated to emulate the performance of clinical experts.[10-32] The complex diagnostic and therapeutic modalities employed in neurosurgery provide a rich assortment and quantity of data towards construction of ML models. The primary aim of this systematic review is to compare the performance of ML and clinical experts head-to-head in applications relevant to neurosurgery providing insights into the current state of advancement of ML and its potential to improve clinical decision making. The secondary aim is to discuss hurdles in creating, validating, and deploying ML models in the clinical setting. To our knowledge, this is the first systematic review comparing the performance of natural intelligence (clinical experts) and artificial intelligence (ML) in neurosurgery, but this review could offer valuable insights for other surgical and nonsurgical specialties too.

METHODS

A systematic search in the Pubmed and Embase databases has been performed according to the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) guidelines, to identify all potential relevant studies as of August 2016. The search syntax was built with the guidance of a professional librarian using search terms related to "Artificial intelligence" and "Neurosurgery." The exact search syntax for the Pubmed and Embase databases is provided in **Table, Supplemental Digital Content**.

Studies were included that compared ML models against clinical experts in applications relevant to the neurosurgical patient population, comparing them head-to-head to a third modality defined as the "ground truth." We defined the neurosurgical patient population as patients eligible for neurosurgical treatment throughout the course of the disease. No specific limitation was applied regarding the domain of application (including diagnosis, prognosis, treatment, and outcome prediction) to be as comprehensive as possible. The definition of "ground truth" depends on the specific task evaluated in the study and could include actual survival in a survival prediction task,[33-35] histological diagnosis in a radiological diagnostic classification task,[36-40] or consensus of a panel of experts when no "objective third modality" is available.[33,41,42] Exclusion criteria were lack of full text, languages other than English and Dutch, animal studies, and absence of a third modality as ground truth when comparing ML and experts. The study selection and data collection was done by 2 independent authors (JS, AK). Disagreements were solved by discussion, in which 2 other authors were involved (OA, HD).

Data extracted from each study were (a) year of publication, (b) disease condition, (c) specific application, (d) ML model used, (e) input features, (f) size of training set, (g) validation method, (h) size of test set, (i) (sub)specialization of clinical experts, (j) level of education of clinical experts, (k) ground truth, (l) statistical measure of performance, (m) performance of ML models, (n) performance of clinical experts, and (o) *P*-value of the difference in performance.

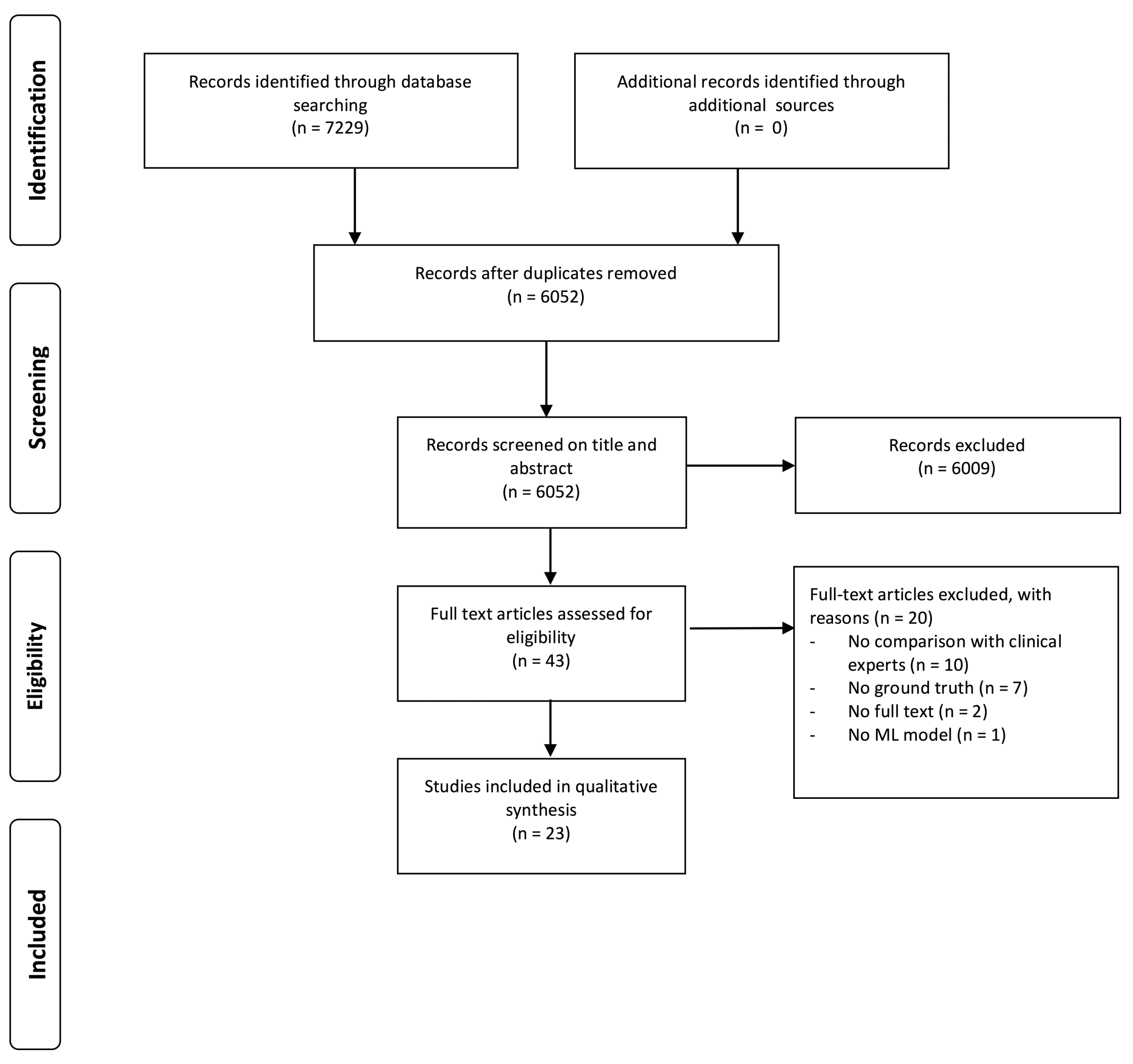
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 7229)
Additional records identified through additional sources (n = 0)
Records after duplicates removed (n = 6052)
Records screened on title and abstract (n = 6052)
Records excluded (n = 6009)
Full text articles assessed for eligibility (n = 43)
Full-text articles excluded, with reasons (n = 20)
- No comparison with clinical experts (n = 10)
- No ground truth (n = 7)
- No full text (n = 2)
- No ML model (n = 1)
Studies included in qualitative synthesis (n = 23)

FIGURE. *PRISMA flow diagram of systematic identification, screening, eligibility, and inclusion. After screening 6052 studies, 23 studies were included in the final analysis.*

We considered a quantitative synthesis to be inappropriate due to the heterogeneity in neurosurgical applications. A qualitative synthesis of results and assessment of risk of bias on outcome, study, and review level is provided by means of a narrative approach. However, to summarize the findings in some quantitative form, the median absolute improvement was calculated for the most commonly used statistical measures of performance, accuracy, and area under the receiver operating curve (AUC). Accuracy refers to the proportion of correct predictions among the total number of predictions, and the AUC corresponds to the probability that a binary classifier will rank a randomly chosen positive instance higher than a randomly chosen negative. The proportion of superior/equal/inferior performance was calculated in percentages. Superior performance is defined as a significantly better performance ($P < .05$) according to the statistical measure used for evaluation. Equal performance was defined as a nonsignificant difference in performance ($P > .05$). P-values were calculated manually if they were not reported in the papers. Fisher's exact test was used for outcome measures with a binomial distribution (accuracy, sensitivity, specificity, positive predictive value, negative predictive value). Student's t-test was used to compare continuous outcome measures. Bound continuous outcome scores, such as the F-measure or AUC (both range 0-1), were transformed to unbound scores (range 0-infinity) using a delta method to meet the assumptions of the Student's t-test. When multiple ML models and/or multiple clinicians were compared, the mean performances were used to calculate the P-value.

RESULTS

After removal of duplicates, a total of 6052 citations in the Pubmed and Embase databases were identified. Forty-four potential relevant studies were selected by title/abstract screening, of which 23 studies remained after full-text screening that evaluated a total of 34 ML models and compared their performance to that of clinical experts based on a total of 61 outcome measures (Figure).

Study Characteristics

The median size of the training set was 79 patients (range 9-7769; interquartile range [IQR] 36.5-123). The median size of the test set was 124.5 patients (range 29-1225; IQR 34.5-139.4). Ten studies used cross-validation methods only instead of a separate test set.[36,37,39-41,43-47] The median ratio between the size of the training set and size of the test set was 1:0.96 (range 1:0.01-1:1.63; IQR 1:0.50-1:1.40).

Twenty-one studies evaluated supervised learning algorithms,[34-54] and 2 studies evaluated unsupervised learning algorithms.[33,55] Of the various ML models used, artificial neural networks (n = 13)[35-40,42,45,48-50,52,54] and support-vector machines (n = 6) [34,38,41,44,49,51] were the most frequently used algorithms. Other learning algorithms used were decision tree (n = 2),[38,49] linear discriminant analysis (n = 2),[42,43] fuzzy C-means (unsupervised learning; n = 2),[33,55] deep learning (n = 1),[41] quadratic discriminant analysis (n = 1),[46] linear regression (n = 1),[52] naïve Bayes (n = 1),[47] and genetic algorithm (n = 1).[49] The clinical experts were neurosurgeons,[35,44,53] neurologists,[46,47] (neuro)radiologists,[33,34,36,37,39,40,44,46,52,55] neurophysiologists,[51] emergency medicine physicians,[48] and hospital hygienist physicians[53] with different levels of education and/or subspecialty.

Ground truth was established by another diagnostic modality,[33,36,37,39,40,44,46,48,52] clinical outcome,[33-35,47,49] expert agreement,[33,41,55] database information,[43] or a combination of these sources.[45,53]

Descriptive Summary of Results

The performance of ML models was compared to clinicians in 23 studies evaluating their utility for diagnosis, preoperative planning, and outcome prediction on a total of 61 outcome measures. The most frequently used statistical measures of performance were accuracy (48% of the studies and 22% of the outcome measures) and AUC (43% of the studies and 21% of the outcome measures). Compared to the clinical experts, the median absolute improvement of the prediction accuracy and AUC of the ML models was 13% (range −4% to 38%; IQR 4%-21%)[35,36,38,42-45,48,49,52,54]] and 0.14 (range 0.03-0.27; IQR 0.07-0.21),[33-35,38-40,44,48,50,54] respectively. A *P*-value was provided for 25 and manually calculated for another 25 of the 61 outcome measures. In 29 of 50 (58%), ML models outperformed clinical experts ($P < .05$). In 18 of 50 (36%), no significant difference was seen between ML and expert performance. In 3 of 50 (6%), clinical experts outperformed ML models ($P < .05$). All studies could be divided into 3 domains of applications: diagnosis, preoperative planning, and outcome prediction. Results on the performance of ML models and clinical experts are provided for each study in Table 2. Methodological details are provided for each study in Table 3. Statistical measures used for the evaluation of performance are explained in Table 4.

Diagnosis

Fourteen studies compared ML and experts on diagnostic performance.[33,36-40,43,44,48,50-54] Four studies focused on diagnostic classification of pediatric posterior fossa tumors,[36,37] intra-axial tumors,[40] or sellar–suprasellar masses.[39] All 14 studies used age and magnetic resonance imaging (MRI) scans as input features with or without additional clinical input (age, gender, medical history, symptoms, signs, and/or family history)[36,37,40] or radiological input (computed tomography (CT), magnetic resonance spectroscopy).[36,37] ML methods were compared against the performance of neuroradiologists[36,37,39] and/or general radiologists.[39,40] ML models performed significantly[36] and nonsignificantly[37] better in differentiating pediatric posterior fossa tumors. Two studies compared the performance of clinicians assisted by ML models against clinicians alone. ML models significantly increased the AUCs for the classification of suprasellar masses in general radiologists (0.88-0.98; $P = .008$) and neuroradiologists (0.95-0.99; $P = .04$).[39] Also, ML improved the diagnostic classification of intracerebral tumors by radiologists (AUC 0.90-0.95; $P < .001$).[40]

Five studies attempted to predict World Health Organization (WHO) grade in gliomas based on MRI features alone[33,38,52,54] or in combination with age.[44] ML models were compared against neuroradiologists,[33,44,54] general radiologists,[38,52] and/or neurosurgeons.[44] ML models performed significantly[38,44,54] and nonsignificantly[33,38, 44,52] better in predicting WHO grade in gliomas.

ML showed a significantly higher AUC in differentiating single-photon emission computer tomography images with brain lesion from images without brain lesions.[50,56]

Three studies evaluated the diagnostic application of ML models without the use of radiological input features.[43,48,53] ML showed a significantly higher accuracy in differentiating single-cell and multiunit spike clusters based on intracranial electroencephalography in epilepsy patients.[43] ML using clinical input features had a significantly higher accuracy in predicting the presence of CT abnormalities in pediatric traumatic brain injury (TBI) patients compared to pediatric emergency medicine fellows and residents.[48] In a study that used natural language processing (NLP) to identify surgical site infections from free text of electronic health records (EHR), a significantly higher sensitivity and F-measure and lower positive predictive value was found for the NLP method compared to a neurosurgeon and/or hospital hygienist physician.[53] NLP is a technique used to process written text such that it may be used to generate ML-based predictive models.[57,58] This tool is especially important when analyzing the large amount of free-text data inherit to the EHR.[47,53]

TABLE 2. Performance of ML Models and Clinical Experts

First author, year of publication	Output	Input features	Outcome measures	ML models	Clinical experts	*P*-value
Diagnosis						
Diagnostic tumor classification						
Kitajima, 2009[39]	Differentiate pituitary adenoma, craniopharyngioma, Rathke's Cleft[a]	Age, MRI	AUC	0.990	0.910	NA[d]
Yamashita, 2008[40]	Differentiate brain metastases, glioma grade II-V, malignant lymphoma[a]	Age, history of brain tumor, MRI	AUC	0.95	0.90	NA[d]
Bidiwala, 2004[37]	Differentiate pediatric posterior fossa tumors: medulloblastoma, cerebellar astrocytoma, ependymoma	Age, gender, symptoms, signs, CT, MRI	Sensitivity	73%-86%	57%-59%	.074[c]
			Specificity	86%-93%	82%-83%	77[c]
			PPV	73%-86%	62%-63%	17[c]
Arle, 1997[36]	Differentiate pediatric posterior fossa tumors: astrocytoma, PNET, ependymoma/other	Age, gender, MRI, MRS	Accuracy	95%	73%	**<.001**[c]
Tumor grading						
Juntu, 2010[38]	Differentiate between benign and malignant soft-tissue tumors including neural tumors	MRI	Accuracy	93%	90%	.61[c]
			Sensitivity	94%	81%	**.009**[c]
			Specificity	91%	92%	1.00[c]
			AUC	0.92	0.85	NA[d]
Zhao, 2010[44]	Classify glioma into grade I-IV	Age, MRI	Accuracy overall	82%	65%	**.001**
			Accuracy LGG	82%	62%	09
			Accuracy HGG	85%	66%	**.008**
			Kappa value	0.68	0.47	NA[d]
			AUC	0.870	0.71	**.004**
Emblem, 2009[33]	Classify glioma into grade I-IV	MRI	AUC	NA	NA	.56-.97
Abdolmaleki, 1997[54]	Differentiate between low and high-grade astrocytomas[a]	MRI	Accuracy	89%	80%	**.003**
			AUC	0.91	0.84	**<.001**[c]
			r	0.87	0.56	NA[d]
Christy, 1995[52]	Classify glioma into grade I-IV	MRI	Accuracy	61%	57%	.84[c]
			Sensitivity	64%	53%	.43[c]
			Specificity	56%	55%	1.00[c]
Other applications						
Campillo, 2013[53]	Detection of surgical site infection	Free text of EHR	Sensitivity	92%	23%	**<.001**[c]
			PPV	40%	100%	**<.001**
			F-measure	0.56	0.38	
Duun-Henriksen, 2012[51]	Automated seizure detection in epilepsy patients	iEEG	Sensitivity	96%	96%	1.00[c]
			FDR	0.14	0.18	1.00[c]
Tankus, 2009[43]	Classify spike clusters in epilepsy	iEEG	Accuracy	91%-92%	38%-69%	**<.001**
Sinha, 2001[48]	Predict the presence of CT abnormalities and DSF in pediatric TBI patients	Age, gender, symptoms, signs, history of trauma	Accuracy	94%	92%	**<.05**[c]
			Sensitivity	82%	62%	**<.001**[c]
			Specificity	96%	96%	1.00[c]
			PPV	75%	72%	.73[c]
			NPV	97%	95%	0.25[c]
			LR+	21	17.3	NA[d]
			LR-	0.18	0.39	NA[d]
			AUC	0.93	0.90	NA[d]
			DP	2.75	2.10	NA[d]
			Sensitivity DSF	79%	54%	**<.001**
Floyd, 1992[50]	Lesion detection in undefined patients	SPECT	AUC	0.86	0.68	**<.001**[d]
Preoperative planning						
Surgical candidate selection						
Cohen, 2016[47]	Identify surgical candidates among pediatric epilepsy patients by means of NLP	Free text of EHR	F-measure	0.77-0.82[b]	0.71	**<.001**[c]

TABLE 2. Continued

First author, year of publication	Output	Input features	Outcome measures	ML models	Clinical experts	*P*-value
Segmentation						
Dolz, 2016[41]	Segmentation of brain stem in trigeminal neuralgia, brain metastases, brainstem cavernoma patients	MRI	DSC pVD Speed	0.88-0.92 3.1%-7.2% 36-40 s	0.84-0.90 19%-39% 20.2 min	**<.001** **<.001** **<.001**[c]
Emblem, 2009[33]	Tumor segmentation in glioma patients grade I-IV	MRI	Sensitivity LGG Sensitivity HGG PPV LGG PPV HGG	83% 69% 66% 73%	59% 57% 89% 87%	**<.001** **.005<.001** **004**
Clarke, 1998[55]	Tumor segmentation in glioma patients grade III-IV	MRI		0.96	0.28-0.49	NA[d]
Localizing epileptogenic zone						
Chiang, 2015[46]	Differentiate L-TLE and R-TLE	fMRI	Accuracy	96%	67%	**.05**
Kassahun, 2014[49]	Differentiate TLE and E-TLE	Symptoms, genetics	Accuracy	66%-78%	56%-78%	.41
Kerr, 2013[45]	Differentiate L-TLE, R-TLE and NES[a]	FDG-PET	Accuracy	76%	80%	.53[c]
Lee, 2000[42]	Localization of epileptogenic zone	FDG-PET	Accuracy	85%	81%	.43[c]
Outcome prediction						
Emblem, 2015[34]	Predict survival in glioma patients grade II-IV	MRI	AUC 6 mo AUC 1 yr AUC 2 yr AUC 3 yr	0.794 0.762 0.806 0.851	0.50-0.66 0.50-0.66 0.50-0.66 0.50-0.66	**<.01** **<.01** **<.01** **<.01**
Rughani, 2010[35]	Predict in-hospital survival in TBI patients	Age, gender, vital parameters	Accuracy Sensitivity Specificity AUC	88% 97% 74% 0.86	72% 73% 75% 0.74	**<.001** **<.001** 40 **<.001**
Emblem, 2009[33]	Predict survival in glioma patients grade I-IV	MRI	Log-rank value	14.4	10.6-12.8	NA[d]

Abbreviations: AUC: area under the curve; CT: computer tomography; DSC: dice similarity coefficient; DSF; depressed skull fracture; E-TLE: extra temporal lobe epilepsy; EHR: electronic health record; iEEG: intracranial encephalography; FDG-PET: positron emission tomography with fludeoxyglucose; FDR: false detection rate; HGG: high-grade glioma; L-TLE: left-sided TLE; LGG: low-grade glioma; MRI: magnetic resonance imaging; MRS: magnetic resonance spectroscopy; NA: not available; NES: nonepileptic seizure; NLP; natural language processing; NPV: negative predicative value; PFT: posterior fossa tumor; PNET: primitive neuroectodermal tumor; PPV: positive predictive value; pVD: percentage volume difference; r: correlation coefficient; R-TLE: right-sided TLE; SPECT: single-photon emission computer tomography; TBI: traumatic brain injury; TLE: temporal lobe epilepsy

[a] This study also measured performance of clinical experts that used ML as second opinion.

[b] F-measure higher (0.74) even 12 mo before surgical referral.

[c] Manually calculated *P*-values.

[d] *P*-value not calculated due to absence of a measure of distribution.

Bold values indicate difference between the performance of clinicians and machine learning models was statistically significant.

Preoperative planning

Seven studies compared ML and experts for performance in preoperative planning.[33,41,42,45-47,55] Tumor segmentation is used for neurosurgical planning to extract the 3-dimensional shape of the tumor from an MRI scan and its relationship with the surrounding anatomy. ML-based MRI segmentation was compared against manual segmentation by (neuro)radiologists in 3 studies, of which 2 evaluated glioma segmentation[33,55] and 1 brainstem segmentation.[41] Comparison consisted of assessing the similarity in the classification of individual voxels by both manual and automated methods. Consensus of multiple experts was considered as the ground truth. ML showed a significantly higher dice similarity coefficient with ground truth and a lower percentage of volume difference for brainstem segmentation.[41] Furthermore, median segmentation speed was 36 to 40 s instead of 20.2 min. ML models had a significantly higher sensitivity for the segmentation of low-grade gliomas and high-grade gliomas at the cost of a lower PPV.[33] In another study evaluating glioma segmentation, the correlation coefficient with ground truth was significantly higher for ML compared to 2 out of 3 experts.[55]

Four studies evaluated epileptogenic zone localization tasks.[42,45,46,49] ML demonstrated a significantly higher accuracy in differentiating left-sided temporal lobe epilepsy (TLE) and right-sided TLE based on functional MRI.[46] No significant difference was found in differentiating TLE from extratemporal lobe epilepsy based on symptoms and genetic information.[49] ML models performed similarly to human experts in localizing the epileptogenic zone by means of positron emission tomography

TABLE 3. Details on Clinical Experts, ML Models, Size Training/Test Set, Validation Method, and Ground Truth

First author, Year of publication	Experts	ML models	Size training set	Validation method	Size test set	Ground truth
Diagnosis						
Diagnostic tumor classification						
Kitajima, 2009[39]	5 general radiologists + 4 neuroradiologists[a]	ANN	43	LOOCV	–	Histological diagnosis
Yamashita, 2008[40]	9 radiologists[a]	ANN	126	LOOCV	–	Histological diagnosis
Bidiwala, 2004[37]	1 neuroradiologist	ANN	33	CV (NOS)	–	Histological diagnosis
Arle, 1997[36]	1 neuroradiologist	ANN	80	5-FCV	–	Histological diagnosis
Tumor grading						
Juntu, 2010[38]	2 radiologists	SVM, ANN, DT(C4.5)	60-100	10-FCV	–	Histological diagnosis
Zhao, 201 0[44]	1 neurosurgeon + 1 neuroradiologist	SVM	106	5-FCV	–	Histological grading
Emblem, 2009[33]	4 neuroradiologists	FCM	–	–	50	Histological grading
Abdolmaleki, 1997[54]	3 neuroradiologists	ANN	43	–	36	Histological grading
Christy, 1995[52]	1 radiologist	ANN, LR	52	–	29	Histological grading
Other applications						
Campillo, 2013[53]	1 neurosurgeon + 1 hospital hygienist physicians	NA	3785	–	1225	Patients identified by expert, NLP or ICD-10 code database
Duun-Henriksen, 2012[51]	1 neurophysiologist	SVM	10	–	10	NA
Tankus, 2009[43]	1 human observer (NOS)	LDA	12	LOOCV	–	Synthetic database with known ground truth
Sinha, 2001[48]	9 pediatric EM attendees + 6 pediatric EM fellows	ANN	382	–	351	CT imaging
Floyd, 1992[50]	6 human observers (NOS)	ANN	40	–	120	NA
Preoperative planning						
Surgical candidate selection						
Cohen, 2016[47]	4 pediatric epileptologists	SVM, NB	523	10-FCV	–	Clinical outcome
Segmentation						
Dolz, 2016[41]	4 experts (NOS)	SVM, DL	9	LOOCV	–	≥75% expert agreement
Emblem, 2009[33]	4 neuroradiologists	FCM	–	–	50	≥75% radiologist agreement
Clarke, 1998[55]	2 radiologists + 2 residents	FCM	–	–	6	Segmentation by multiple radiologists
Localizing epileptogenic zone						
Chiang, 2015[46]	1 neuroradiologist	QDA	24	LOOCV	–	VEEG
Kassahun, 2014[49]	7 clinicians (NOS)	GA, DT, ANN, SVM	79	CV (NOS)	129	Disease free after operation
Kerr, 2013[45]	1 neurologist[a]	ANN	105	LOOCV	–	Consensus of history, physical and neurological exam, neuropsychiatric testing, VEEG, interictal and ictal FDG-PET, MRI and CT
Lee, 2000[42]	1 human expert (NOS)	ANN, LDA	120	–	141	Consensus of 2 expert physicians
Outcome prediction						
Emblem, 2015[34]	1 neuroradiologists	SVM	101	10-FCV	134	True survival
Rughani, 2010[35]	4 neurosurgeons + 5 neurosurgical residents	ANN	7769	–	100	True in-hospital survival
Emblem, 2009[33]	4 neuroradiologists	FCM	–	–	50	True survival

Abbreviations: ANN: artificial neural networks; CT: computer tomography; DL: deep learning; DT(C4.5): decision tree (C4.5 is type of DT); GA: genetic algorithm; ICD: international classification of disease; LDA: linear discriminant analysis; LOOCV: leave-one out cross validation; LR: linear regression; FCM: fuzzy c-means; FCV: fold cross validation; FDG-PET: positron emission tomography using fludeoxyglucose; NA: not available; NB; naïve bayes; NLP: natural language processing; NOS: not otherwise specified; QDA: quadratic discriminant analysis; SVM: support vector machines; VEEG: video electroencephalography monitoring test; -: not applicable

[a] This study also measured performance of clinical experts that used ML as second opinion.

TABLE 4. Statistical Measures Explained

Accuracy	Proportion of correct predictions among the total no. of predictions (TP + TN)/total population; range 0% to 100%
Sensitivity	Proportion of positively classified cases among the total no. of positive cases; TP/(TP + FN); range 0% to 100%
Specificity	Proportion of negatively classified cases among the total no. of negative cases; TN/(TN + FP); range 0% to 100%
Positive predictive value	Proportion of positive cases among the total no. of positively classified cases; TP/(TP + FP); range 0% to 100%
Negative predictive value	Proportion negative cases among the total no. of negatively classified cases; TN/(TN + FN); range 0% to 100%
Positive likelihood ratio	Refers to the increase in probability if the condition of the test is positive; Sens/(1 – Spec); range 1 to ∞
Negative likelihood ratio	Refers to the decrease in probability if the condition of the test is positive (1 – Sens)/Spec; range 0 to 1
F-measure	Measure of accuracy that uses the sensitivity and positive predictive value (2 × TP)/(2 × TP + FP + FN); range 0 to 1
Area under the ROC curve	ROC is a graphical plot illustrating the sensitivity as a function of "1 – specificity" in a binary classifier with a varying discrimination threshold. The area under the curve corresponds to the probability that a binary classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one; range 0 to 1
Kappa value	Statistic to calculate the degree of agreement considering the agreement occurring by chance only; range 0 to 1
Dice similarity coefficient	Statistic for similarity; range 0 to 1
Percentage volume difference	Statistic for volume similarity; range 0% to 100%
Correlation coefficient	Statistic for strength and direction of linear relationship; range 0 to 1
Log-rank value	Statistic to compare survival distributions from 2 samples; range 0 to ∞

Abbreviations: FN: false negative; FP: false positive; no.: number; ROC: receiver operating characteristic; Sens: sensitivity; Spec: specificity; TN: true negative; TP: true positive

with fludeoxyglucose; however, combining the 2 improved the classification accuracy significantly.[45]

One study showed that ML-based NLP could identify surgical candidates among pediatric epilepsy patients. NLP methods used data of free-text clinical notes extracted from the EHR as input feature. Actual surgery as reported in the EHR was considered as ground truth. The F-measure was higher than that of pediatric epileptologists at time of referral (0.78-0.82 vs 0.71) and even 12 mo before referral (0.74).[47]

Outcome Prediction

Three studies evaluated outcome prediction.[33-35] Two studies evaluated survival prediction in glioma patients based on MRI features.[33,34] One study provided *P*-values, showing significantly higher AUCs for 6-mo, 1-, 2-, and 3-yr survival compared to a neuroradiologist.[34]

One study evaluated in-hospital survival prediction in TBI patients based on clinical features. Compared to the neurosurgeons and/or neurosurgical residents, the ML models performed superiorly on accuracy, sensitivity, and AUC and performed equally on specificity.[35]

DISCUSSION

The aim of this systematic review was to compare the performance of ML and clinical experts head-to-head in applications relevant to neurosurgery, providing insights into the current state of advancement of ML and its potential to improve clinical decision making. Compared to the clinical experts, the median absolute improvement of the prediction accuracy and AUC of the ML models was 13% and 0.14, respectively. In 29 of 50 (58%) studies that provided a *P*-value or sufficient details to calculate a *P*-value, ML models outperformed clinical experts significantly ($P < .05$). Our results show that ML models can emulate the performance of clinicians in some neurosurgical applications within the domains of diagnosis, preoperative planning, and outcome prediction.

Implications of Implementing ML in Neurosurgical Care

Although the output consisted of a wide range of neurosurgical applications, ML was most frequently used for analyzing radiological data by means of artificial neural networks. Since every voxel can be used as an individual input feature, the amount of information that can be extracted by ML is extremely high, making it faster and more accurate than is humanly possible. Automated analysis of radiological data for diagnosis, segmentation, or outcome prediction could, therefore, be one of the first ML applications that finds its way to actual clinical practice.[2]

Due to the vast amount of radiological data, the total amount of data remains high even in smaller sample sizes. The median size of training sets across all studies was relatively low (79 patients); however, many studies were still able to construct high-performing ML models using radiological data. The training of ML models is, therefore, dependent on both the sample size and the volume of data collected per patient.

An important preconception regarding the role of ML models in the clinical realm is that it would result in displacement of clinicians, the so-called human-vs-machine paradigm. In reality, even if ML models are able to perform a given analysis with very high accuracy, clinicians still must consider the implications of this analysis in the global sense. The results of an ML model are better utilized to augment the clinician in medical decision making, which we refer to as human-and-machine. As an example, 4 studies assessed the combined performance of clinicians and ML models for radiological diagnosis or segmentation.[39,40,45,54] In all studies, ML in conjunction with clinical decision making was superior to clinical decision making or ML models alone. This highlights the fact that both ML models and human experts contribute to the classification task and are, therefore, complementary to each other.

Although ML can be a very powerful tool for clinicians to make sense of and use big data and even emulate clinical expertise, it remains ambiguous how these methods should be implemented in actual clinical practice. In 2 studies that used a human-and-machine approach, radiologists used ML models as a second opinion for radiological diagnosis,[39,40] whereas the radiologist's impression was used as an ML input feature in the other 2 studies.[45,54] In both cases, ML is used to increase diagnostic accuracy. On the other hand, ML can also be used to increase efficiency of clinicians by automating clinical decision making and select cases that need secondary evaluation. Lastly, ML can produce error detection systems for clinicians (in training), operating on the background only.

Challenges Regarding the Implementation of ML Models

A barrier with the introduction of ML models into clinical practice is that the mechanisms driving the algorithms can be very complex, or even impossible, to interpret; as such, these algorithms are sometimes referred to as "black box" techniques. This contrasts with most of the conventional statistical methods, such as regression coefficients, odds ratios, or hazard ratios that are familiar and can be easily interpreted for clinical meaning. This creates a conflict where on one hand we have powerful predictive algorithms, and on the other hand our inability to peer inside the "black box" can result in adoption hesitancy. We suspect that as more ML models are produced and validated and as such become more familiar, clinicians will be more comfortable deploying them in daily practice.

Another hurdle is that the generation of ML model requires a large amount of complete and adequately categorized data. Due to higher quality data, the performance of ML models could be overestimated in research setting compared to their true performance in the clinical setting. Lastly, access to patient data, especially across institutions, is rightfully restricted given privacy considerations, which can cause difficulties in obtaining high-volume training data sets.[6]

To overcome these challenges, more studies should take a human-and-machine approach to explore how clinicians can benefit most from the powerful analyses that ML offers. This means, for example, assessing the diagnostic performance and efficiency of radiologists assisted by an ML automated diagnosis system or assessing surgical outcomes after operations that have been planned with the use of ML segmentation models. Furthermore, future studies should focus on creating an ethical and legal framework that supports collection of training data, validation of ML models on heterogeneous test sets prior to deployment, and regulation of the ML performance after deployment in clinical care.

Limitations

This review has some inherent limitations. It is possible that the ML models reviewed here are being compared to clinical experts in situations where ML models are more likely to perform superiorly. Due to positive publication bias, the performance of ML could be overestimated even more. This review is limited to situations where human decision making and ML methods are comparable, ignoring the potential of ML methods on tasks that lie beyond the capacity of human experts. Furthermore, most studies evaluated performance in terms of precision and accuracy, and only 1 study evaluated clinicians and ML models based on the speed of performing a classification/prediction task.[41] This underexposes the potential in which computers perform at their best: taking over very simple and repetitive tasks at a much higher speed. Additionally, the performance of clinical experts could be overestimated when ground truth is defined as a consensus of multiple experts; the individually tested experts could be influencing that ground truth.

Another weakness to a systemic analysis is that most outcome measures did not provide a *P*-value; however, we have calculated *P*-values manually for 25 outcome measures, and these outcome measures were also represented in the median absolute difference in accuracy and AUC, demonstrating similar results in favor of the ML models (Table 2). Summarizing the results based on outcome level does not provide the adequate weight that should be given to each study based on the size and quality of the study. Many studies only used cross-validation without a distinct test set when creating the ML model.[35-37,39-41,43-47] Cross-validation divides the original data in training and test data multiple times, and the validation results are averaged over the number of rounds. This could overestimate the performance of ML models as the data is used for both training and selection of the model. Some studies assess multiple outcome measures that are interrelated, resulting in an overrepresentation of the findings in these studies.[34,44]

Nevertheless, we deem the limitations proportionate to the strength of this systematic review. This review provides a thorough overview of ML models in neurosurgical applications. To our knowledge, this is the first systematic review comparing the performance of ML models with that of clinical experts in neurosurgery. By using the performance of clinical experts as a benchmark, this review provides valuable insights into the current state of advancement of artificial intelligence in neurosurgery and its future role in patient care. Lastly, it identifies hurdles to overcome in creating, validating, and deploying ML models in the clinical setting.

CONCLUSION

This is the first systematic review comparing the performance of ML models with that of clinical experts in neurosurgery. It shows that artificial intelligence has the potential to augment the decision-making capacity of clinicians within the realms diagnosis, preoperative planning, and outcome prediction in neurosurgery; however, it is of paramount importance to address hurdles associated with creating, validating, and deploying ML models in the clinical setting. Shifting from a human-vs-machine to a human-and-machine paradigm could be essential for this. Future studies should, therefore, focus not only on the technical aspects in constructing these models but also on methods to validate ML models prior to deployment, increasing the reproducibility and interpretability of these "black-box" algorithms, improving the accessibility of clinical data, and investigating the combined performance of ML models and clinicians.

Disclosures

Dr Gormley is a proctor and consultant for Codman and Coviden (Medtronic). The authors have no personal, financial, or institutional interest in any of the drugs, materials, or devices described in this article.

REFERENCES

1. Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature*. 2015;521(7553):452-459.
2. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216-1219.
3. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255-260.
4. Mitchell, TM. *Machine Learning*. Vol. 1. New York: McGraw-Hill Science/Engineering/Math; 1997.
5. Noble WS. What is a support vector machine? *Nat Biotechnol*. 2006;24(12):1565-1567.
6. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920-1930.
7. Azimi P, Mohammadi HR, Benzel EC, Shahzadi S, Azhari S, Montazeri A. Artificial neural networks in neurosurgery. *J Neurol Neurosurg Psychiatry*. 2015;86(3):251-256.
8. Rodrigues JF, Jr, Paulovich FV, de Oliveira MC, de Oliveira ON, Jr. On the convergence of nanotechnology and Big Data analysis for computer-aided diagnosis. *Nanomedicine (Lond)*. 2016;11(8):959-982.
9. Coveney PV, Dougherty ER, Highfield RR. Big data need big theory too. *Philos Trans A Math Phys Eng Sci*. 2016;374(2080).
10. Mariak Z, Swiercz M, Krejza J, Lewko J, Lyson T. Intracranial pressure processing with artificial neural networks: classification of signal properties. *Acta Neurochir (Wien)*. 2000;142(4):407-411; discussion 411-402.
11. Nucci CG, De Bonis P, Mangiola A, et al. Intracranial pressure wave morphological classification: automated analysis and clinical validation. *Acta Neurochir (Wien)*. 2016;158(3):581-588; discussion 588.
12. Sieben G, Praet M, Roels H, Otte G, Boullart L, Calliauw L. The development of a decision support system for the pathological diagnosis of human cerebral tumours based on a neural network classifier. *Acta Neurochir (Wien)*. 1994;129(3-4):193-197.
13. Mathew B, Norris D, Mackintosh I, Waddell G. Artificial intelligence in the prediction of operative findings in low back surgery. *Brit J Neurosurg*. 1989;3(2):161-170.
14. Arle JE, Perrine K, Devinsky O, Doyle WK. Neural network analysis of preoperative variables and outcome in epilepsy surgery. *J Neurosurg*. 1999;90(6):998-1004.
15. Gazit T, Andelman F, Glikmann-Johnston Y, et al. Probabilistic machine learning for the evaluation of presurgical language dominance. *J Neurosurg*. 2016;125(2):1-13.
16. Shi HY, Hwang SL, Lee KT, Lin CL. In-hospital mortality after traumatic brain injury surgery: a nationwide population-based comparison of mortality predictors used in artificial neural network and logistic regression models. *J Neurosurg*. 2013;118(4):746-752.
17. Azimi P, Mohammadi HR. Predicting endoscopic third ventriculostomy success in childhood hydrocephalus: an artificial neural network analysis. *J Neurosurg Pediatr*. 2014;13(4):426-432.
18. Azimi P, Mohammadi H. Prediction of successful ETV outcome in childhood hydrocephalus: an artificial neural networks analysis. *J Neurosurg*. 2015;122(6):426-432.
19. Chang K, Zhang B, Guo X, et al. Multimodal imaging patterns predict survival in recurrent glioblastoma patients treated with bevacizumab. *Neuro-oncology*. 2016;18(12):1680-1687.
20. Jones TL, Byrnes TJ, Yang G, Howe FA, Bell BA, Barrick TR. Brain tumor classification using the diffusion tensor image segmentation (D-SEG) technique. *Neuro-oncology*. 2015;17(3):466-476.
21. Macyszyn L, Akbari H, Pisapia JM, et al. Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-oncology*. 2016;18(3):417-425.
22. Teplyuk NM, Mollenhauer B, Gabriely G, et al. MicroRNAs in cerebrospinal fluid identify glioblastoma and metastatic brain cancers and reflect disease activity. *Neuro-oncology*. 2012;14(6):689-700.
23. Zhang B, Chang K, Ramkissoon S, et al. Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas. *Neuro-oncology*. 2017;19(1):109-117.
24. Fouke SJ, Weinberger K, Kelsey M, et al. A machine-learning-based classifier for predicting a multi-parametric probability map of active tumor extent within glioblastoma multiforme. *Neuro-oncology*. 2012;14:vi124-vi125.
25. Kim LM, Commean P, Boyd A, et al. Predicting the location and probability of viable tumor within glioblastoma multiforme with multiparametric magnetic resonance imaging. *Neuro-oncology*. 2012;14:vi120-vi128.
26. Orphanidou-Vlachou E, Vlachos N, Davies N, Arvanitis T, Grundy R, Peet A. Texture analysis of T1-and t2-weighted magnetic resonance images to discriminate posterior fossa tumors in children. *Neuro-oncology*. 2014;16:i123-i126.
27. Rayfield C, Swanson K. Predicting the response to treatment in GBM: Machine learning on clinical images. *Neuro-oncology*. 2015;17:v167.
28. Akbari H, Macyszyn L, Da X, et al. Imaging surrogates of infiltration obtained via multiparametric imaging pattern analysis predict subsequent location of recurrence of glioblastoma. *Neurosurgery*. 2016;78(4):572-580.
29. Mitchell TJ, Hacker CD, Breshears JD, et al. A novel data-driven approach to preoperative mapping of functional cortex using resting-state functional magnetic resonance imaging. *Neurosurgery*. 2013;73(6):969-982; discussion 982-963.
30. Oermann EK, Kress MA, Collins BT, et al. Predicting survival in patients with brain metastases treated with radiosurgery using artificial neural networks. *Neurosurgery*. 2013;72(6):944-951; discussion 952.
31. Taghva A. An automated navigation system for deep brain stimulator placement using hidden Markov models. *Neurosurgery*. 2010;66(3 Suppl Operative):108-117; discussion 117.
32. Dumont TM, Rughani AI, Tranmer BI. Prediction of symptomatic cerebral vasospasm after aneurysmal subarachnoid hemorrhage with an artificial neural network: feasibility and comparison with logistic regression models. *World Neurosurg*. 2011;75(1):57-63; discussion 25-58.
33. Emblem KE, Nedregaard B, Hald JK, Nome T, Due-Tonnessen P, Bjornerud A. Automatic glioma characterization from dynamic susceptibility contrast imaging: brain tumor segmentation using knowledge-based fuzzy clustering. *J Magn Reson Imaging*. 2009;30(1):1-10.
34. Emblem KE, Pinho MC, Zollner FG, et al. A generic support vector machine model for preoperative glioma survival associations. *Radiology*. 2015;275(1):228-234.
35. Rughani AI, Dumont TM, Lu Z, et al. Use of an artificial neural network to predict head injury outcome. *J Neurosurg*. 2010;113(3):585-590.
36. Arle JE, Morriss C, Wang ZJ, Zimmerman RA, Phillips PG, Sutton LN. Prediction of posterior fossa tumor type in children by means of magnetic resonance image properties, spectroscopy, and neural networks. *J Neurosurg*. 1997;86(5):755-761.
37. Bidiwala S, Pittman T. Neural network classification of pediatric posterior fossa tumors using clinical and imaging data. *Pediatr Neurosurg*. 2004;40(1):8-15.
38. Juntu J, Sijbers J, De Backer S, Rajan J, Van Dyck D. Machine learning study of several classifiers trained with texture analysis features to differentiate benign

from malignant soft-tissue tumors in T1-MRI images. *J Magn Reson Imaging*. 2010;31(3):680-689.
39. Kitajima M, Hirai T, Katsuragawa S, et al. Differentiation of common large sellar-suprasellar masses effect of artificial neural network on radiologists' diagnosis performance. *Acad Radiol*. 2009;16(3):313-320.
40. 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(6):1153-1158.
41. Dolz J, Betrouni N, Quidet M, et al. Stacking denoising auto-encoders in a deep network to segment the brainstem on MRI in brain cancer patients: a clinical study. *Comput Med Imaging Graph*. 2016;52:8-18.
42. Lee JS, Lee DS, Kim SK, et al. Localization of epileptogenic zones in F-18 FDG brain PET of patients with temporal lobe epilepsy using artificial neural network. *IEEE Trans Med Imaging*. 2000;19(4):347-355.
43. Tankus A, Yeshurun Y, Fried I. An automatic measure for classifying clusters of suspected spikes into single cells versus multiunits. *J Neural Eng*. 2009;6(5):056001.
44. Zhao ZX, Lan K, Xiao JH, et al. A new method to classify pathologic grades of astrocytomas based on magnetic resonance imaging appearances. *Neurol India*. 2010;58(5):685-690.
45. Kerr WT, Nguyen ST, Cho AY, et al. Computer-Aided Diagnosis and Localization of Lateralized Temporal Lobe Epilepsy Using Interictal FDG-PET. *Front Neurol*. 2013;4:1-14.
46. Chiang S, Levin HS, Haneef Z. Computer-automated focus lateralization of temporal lobe epilepsy using fMRI. *J Magn Reson Imaging*. 2015;41(6):1689-1694.
47. Cohen KB, Glass B, Greiner HM, et al. Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning. *Biomed Inform Insights*. 2016;8(8):11-18.
48. Sinha M, Kennedy CS, Ramundo ML. Artificial neural network predicts CT scan abnormalities in pediatric patients with closed head injury. *J Trauma*. 2001;50(2):308-312.
49. Kassahun Y, Perrone R, De Momi E, et al. Automatic classification of epilepsy types using ontology-based and genetics-based machine learning. *Artif Intell Med*. 2014;61(2):79-88.
50. Floyd CE, Jr, Tourassi GD. An artificial neural network for lesion detection on single-photon emission computed tomographic images. *Invest Radiol*. 1992;27(9):667-672.
51. Duun-Henriksen J, Kjaer TW, Madsen RE, Remvig LS, Thomsen CE, Sorensen HB. Channel selection for automatic seizure detection. *Clin Neurophysiol*. 2012;123(1):84-92.
52. Christy PS, Tervonen O, Scheithauer BW, Forbes GS. Use of a neural network and a multiple regression model to predict histologic grade of astrocytoma from MRI appearances. *Neuroradiology*. 1995;37(2):89-93.
53. Campillo-Gimenez B, Garcelon N, Jarno P, Chapplain JM, Cuggia M. Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France. *Stud Health Technol Inform*. 2013;192:572-575.
54. Abdolmaleki P, Mihara F, Masuda K, Buadu LD. Neural networks analysis of astrocytic gliomas from MRI appearances. *Cancer Lett*. 1997;118(1):69-78.
55. Clarke LP, Velthuizen RP, Clark M, et al. MRI measurement of brain tumor response: comparison of visual metric and automatic segmentation. *Magn Reson Imaging*. 1998;16(3):271-279.
56. Chan T, Huang HK. Effect of a computer-aided diagnosis system on clinicians' performance in detection of small acute intracranial hemorrhage on computed tomography. *Acad Radiol*. 2008;15(3):290-299.
57. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544-551.
58. Liu K, Hogan WR, Crowley RS. Natural Language Processing methods and systems for biomedical ontology learning. *J Biomed Inform*. 2011;44(1):163-179.

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

Acknowledgment

The authors would like to thank P. Staples for his help on manually calculating *P*-values.

COMMENT

This article delves into the relatively new field of medically applicable machine learning, a process in which computers are programed to exceed the initial parameters defined within them as new information is input. It is the artificial corollary of clinical experience used by advanced practitioners to better provide patient care. The authors review 23 published studies that evaluate machine learning related to neurosurgical care. These studies were performed with the goal to compare machine learning based programs to clinician judgement in aspects of diagnosis, planning, and outcome prediction where a "gold-standard" was available as the defined correct choice. Over half of these measures demonstrated a greater performance from the machine learning programs when compared to clinician judgement by a median of 13%. Machines tended to be more dominant in the prevention of false negatives (higher sensitivity) whereas human observers were found to be able to detect more minute differences manifested by a higher specificity and positive predictive value. As typical for machine based processing, decisions made by computers were more accurate and many times faster than their human counterparts.

Perhaps the most impressive demonstration of utility from machine learning models was in the domain of radiographic assessment. Human assessment of radiology images is based on the construct provided after processing, but this in fact represents a crude averaging of the actual data collected to allow for something that can be comprehended visually. The data are vastly more detailed than what can be perceived visually. Analysis of differences in intensity on a voxel by voxel basis provides a distinctive advantage in identifying minute differences that may in fact represent diagnostically significant differences in tumor appearance that cannot be visually processed by human practitioners. In a sense, the machine programs are "cheating".

Another important topic explored by the authors is the concept of "machine and human" rather than simply "machine versus human". The application of machine learning may serve as a supplement to clinical judgement that allows for performance frequently greater than human-alone or machine-alone performance. This is important given the natural advantages demonstrated that allow for rapid and accurate evaluation of results by machines with the ability to avoid false positives due to small nuances that are currently better able to be detected by humans. We suggest that the future of medicine will likely involve an embracement of technological supplements rather than a complete replacement of clinicians by machine practitioners.

Michael M. McDowell
Peter C. Gerszten
Pittsburgh, Pennsylvania