NEUROSURGICAL FOCUS

Neurosurg Focus 54(6):E5, 2023

Which supervised machine learning algorithm can best predict achievement of minimum clinically important difference in neck pain after surgery in patients with cervical myelopathy? A QOD study

Christine Park, MS,[1] Praveen V. Mummaneni, MD, MBA,[2] Oren N. Gottfried, MD,[1] Christopher I. Shaffrey, MD,[1] Anthony J. Tang, BBA, BSA,[3] Erica F. Bisson, MD, MPH,[4] Anthony L. Asher, MD,[5] Domagoj Coric, MD,[5] Eric A. Potts, MD,[6] Kevin T. Foley, MD,[7] Michael Y. Wang, MD,[8] Kai-Ming Fu, MD, PhD,[9] Michael S. Virk, MD, PhD,[9] John J. Knightly, MD,[10] Scott Meyer, MD,[10] Paul Park, MD,[11] Cheerag Upadhyaya, MD, MBA, MSc,[12] Mark E. Shaffrey, MD,[13] Avery L. Buchholz, MD, MPH,[13] Luis M. Tumialán, MD,[14] Jay D. Turner, MD, PhD,[14] Brandon A. Sherrod, MD,[4] Nitin Agarwal, MD,[15] Dean Chou, MD,[3] Regis W. Haid Jr., MBA, MD,[16] Mohamad Bydon, MD,[17] and Andrew K. Chan, MD[3]

[1]Department of Neurosurgery, Duke University, Durham, North Carolina; [2]Department of Neurosurgery, University of California, San Francisco, California; [3]Department of Neurological Surgery, Columbia University Vagelos College of Physicians and Surgeons, The Och Spine Hospital at NewYork-Presbyterian, New York, New York; [4]Department of Neurosurgery, University of Utah, Salt Lake City, Utah; [5]Neuroscience Institute, Carolinas Healthcare System and Carolina Neurosurgery & Spine Associates, Charlotte, North Carolina; [6]Goodman Campbell Brain and Spine, Indianapolis, Indiana; [7]Department of Neurosurgery, University of Tennessee, Semmes-Murphey Neurologic and Spine Institute, Memphis, Tennessee; [8]Department of Neurosurgery, University of Miami, Florida; [9]Department of Neurosurgery, Weill Cornell Medical Center, New York, New York; [10]Atlantic Neurosurgical Specialists, Morristown, New Jersey; [11]Department of Neurosurgery, University of Michigan, Ann Arbor, Michigan; [12]Marion Bloch Neuroscience Institute, Saint Luke's Health System, Kansas City, Missouri; [13]Department of Neurosurgery, University of Virginia, Charlottesville, Virginia; [14]Barrow Neurological Institute, Phoenix, Arizona; [15]Department of Neurosurgery, Washington University in St. Louis, Missouri; [16]Atlanta Brain and Spine Care, Atlanta, Georgia; and [17]Department of Neurologic Surgery, Mayo Clinic, Rochester, Minnesota

OBJECTIVE The purpose of this study was to evaluate the performance of different supervised machine learning algorithms to predict achievement of minimum clinically important difference (MCID) in neck pain after surgery in patients with cervical spondylotic myelopathy (CSM).

METHODS This was a retrospective analysis of the prospective Quality Outcomes Database CSM cohort. The data set was divided into an 80% training and a 20% test set. Various supervised learning algorithms (including logistic regression, support vector machine, decision tree, random forest, extra trees, gaussian naïve Bayes, k–nearest neighbors, multilayer perceptron, and extreme gradient boosted trees) were evaluated on their performance to predict achievement of MCID in neck pain at 3 and 24 months after surgery, given a set of predicting baseline features. Model performance was assessed with accuracy, F1 score, area under the receiver operating characteristic curve, precision, recall/sensitivity, and specificity.

RESULTS In total, 535 patients (46.9%) achieved MCID for neck pain at 3 months and 569 patients (49.9%) achieved it at 24 months. In each follow-up cohort, 501 patients (93.6%) were satisfied at 3 months after surgery and 569 patients (100%) were satisfied at 24 months after surgery. Of the supervised machine learning algorithms tested, logistic regression demonstrated the best accuracy (3 months: 0.76 ± 0.031, 24 months: 0.773 ± 0.044), followed by F1 score (3 months: 0.759 ± 0.019, 24 months: 0.777 ± 0.039) and area under the receiver operating characteristic curve (3 months: 0.762 ± 0.027, 24 months: 0.773 ± 0.043) at predicting achievement of MCID for neck pain at both follow-up

ABBREVIATIONS AUROC = area under the receiver operating characteristic curve; CSM = cervical spondylotic myelopathy; EQ-5D = EuroQol–5 Dimensions; EQ-VAS = EuroQol VAS; MCID = minimum clinically important difference; mJOA = modified Japanese Orthopaedic Association; NASS = North American Spine Society; NDI = Neck Disability Index; QOD = Quality Outcomes Database; SHAP = Shapley Additive Explanations; VAS = visual analog scale.

SUBMITTED January 31, 2023. **ACCEPTED** March 22, 2023.
INCLUDE WHEN CITING DOI: 10.3171/2023.3.FOCUS2372.

time points, with fair performance. The best precision was also demonstrated by logistic regression at 3 (0.724 ± 0.058) and 24 (0.780 ± 0.097) months. The best recall/sensitivity was demonstrated by multilayer perceptron at 3 months (0.841 ± 0.094) and by extra trees at 24 months (0.817 ± 0.115). Highest specificity was shown by support vector machine at 3 months (0.952 ± 0.013) and by logistic regression at 24 months (0.747 ± 0.18).

CONCLUSIONS Appropriate selection of models for studies should be based on the strengths of each model and the aims of the studies. For maximally predicting true achievement of MCID in neck pain, of all the predictions in this balanced data set the appropriate metric for the authors' study was precision. For both short- and long-term follow-ups, logistic regression demonstrated the highest precision of all models tested. Logistic regression performed consistently the best of all models tested and remains a powerful model for clinical classification tasks.

https://thejns.org/doi/abs/10.3171/2023.3.FOCUS2372

KEYWORDS cervical spondylotic myelopathy; machine learning; neck pain; patient satisfaction; patient-reported outcomes; Quality Outcomes Database

Cervical spondylotic myelopathy (CSM) is a common condition affecting older adults, and its incidence is growing along with the aging population.[1] Many patients with CSM suffer from neck pain, with more than 40% of them presenting with severe neck pain.[2] The neck pain can be debilitating, preventing one from performing daily tasks and affecting one's quality of life. Although surgery accomplishes the main goal of neural decompression in CSM, neck pain is not uniformly improved.[3,4] With this in mind, it would be helpful to better predict postoperative neck pain improvement to help guide preoperative counseling in this patient population.

Machine learning algorithms are often used to build prediction models that learn to recognize patterns from past data to predict future outcomes. In healthcare, the supervised learning variant is commonly used to aid in making classification decisions based on labeled data (whereas unsupervised learning aims to uncover hidden patterns in unlabeled data sets).[5] Supervised machine learning is trained on past input-output pairs to predict outputs for new input data. During training, the algorithm searches for patterns in the data that can be used to predict the output. The model can then be applied to new inputs to forecast the desired outcome. Due to its predictive power, supervised machine learning may potentially be used to evaluate and select patients for surgical intervention based on the estimated likelihood of achieving clinical improvement. The purpose of this study was to compare the performance of well-established supervised machine learning algorithms in classifying patients with CSM based on their likelihood to achieve a minimum clinically important difference (MCID) in visual analog scale (VAS) neck pain score after surgery.

Methods

Study Design and Selection of Patients

This was a retrospective study performed using the prospective Quality Outcomes Database (QOD) CSM data set. Institutional review board approval (Columbia University) was obtained and patient consent waived due to the study design. This specific data set consists of adult patients diagnosed with CSM at 14 hospital sites that combined their prospective QOD registry data for patients who met the following inclusion criteria: 1) underwent elective cervical spine surgery for an indication of CSM between January 2016 and December 2018, 2) had a modified Japanese Orthopaedic Association (mJOA) score < 17, and 3) had a predominant symptom of myelopathy. Patients were excluded if they had a spinal infection, tumor, fracture, traumatic dislocation, deformity, or neurological paralysis due to preexisting spine disease or injury.

Study Variables

Demographic information included the following: age; sex; self-reported race and ethnicity; socioeconomic status index;[6] insurance coverage; smoking status; medical comorbidities; American Society of Anesthesiologists grade; employment status; baseline symptoms and symptom duration; underlying pathology; approach (anterior vs posterior); levels treated; and patient-reported outcomes such as mJOA, Neck Disability Index (NDI), arm and neck pain according to the VAS, EuroQol VAS (EQ-VAS), and EuroQol–5 Dimensions (EQ-5D [measured in quality-adjusted life-years—QALY]) scores. Myelopathy, as per the mJOA score, was classified as mild (15–17), moderate (12–14), and severe (< 12).[7]

Outcome of Interest

Achievement of MCID at 3- and 24-month follow-ups for VAS neck pain was defined as a reduction of 2.6 points from baseline.[8–10] We compared the performance metrics of the models based on their successful prediction of achieving MCID for neck pain. Patients were defined as satisfied if they reported a North American Spine Society (NASS) score of 1 or 2 based on the 4-point NASS scale.

Statistical Analysis

Descriptive categorical and continuous variables were summarized using frequency counts (percentages) and means (SDs), respectively. For missing values, we used the scikit-learn SimpleImputer function,[11] which substitutes the mean value for continuous variables and the median for categorical variables. A summary of the missing values for the predictors is reported in Supplementary Table 1. Feature importance was conducted via logistic regression.

Selection of Algorithms and Performance Metrics

Several supervised machine learning algorithms can be used to predict achievement of MCID for neck pain. In

TABLE 1. Strengths and weaknesses of different supervised machine learning models

Supervised ML Algorithm	Main Advantages	Main Limitations
Logistic regression	Basic model usually used as a baseline for classification problems when comparing prediction performance; fast	Does not perform well when input variables have complex relationships; potential to overfit
Support vector machine	More robust than simple logistic regression or gaussian naïve Bayes, yet fast; can handle multiple data types	Can be computationally expensive for large data sets; generic support vector machine cannot classify >2 classes unless extended
Decision tree	Multiple data types are supported; results are easy to interpret	Requires classes to be mutually exclusive; depends on the order of features
Extra trees	Uses entire input sample, which reduces bias; works well w/ large data sets; fast	Randomly chooses split of each node, which reduces variance
Random forest	Subsamples input data w/ replacement, which increases variance & diversity of data; works well w/ large data sets	Complex & can be computationally expensive
k–nearest neighbors	Another simple algorithm based on clustering; fast	Features are given equal importance so can lead to poor classification performance
Gaussian naïve Bayes	Simple yet useful for large data sets; assigns probabilities to each feature for making prediction; fast	Requires classes to be mutually exclusive; assumes normal distribution of numeric features
Multilayer perceptron	Useful when there are complex relationships among variables	Computationally expensive; can be hard to understand results
Extreme gradient boosted trees	Can handle multiple data types; works well w/ large data sets	Prone to overfitting; complex & can be computationally expensive

ML = machine learning.

this work, we focused on the set of supervised machine learning algorithms that were commonly used in clinical studies, namely logistic regression, support vector machine, decision tree, random forest, extra trees, gaussian naïve Bayes, k–nearest neighbors, and multilayer perceptron from the scikit-learn[11] package, and extreme gradient boosted trees from the XGBoost[12] package. In applicable cases, a class-weighted variant of certain algorithms was used for improved accuracy. Strengths and weaknesses of each model are summarized in Table 1. The model hyperparameters are outlined in Supplementary Table 2.

Model performance was assessed with accuracy, F1 score, precision, sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC). Specifically, we performed stratified 5-fold cross-validation to evaluate our models, meaning that the training set was divided into 5 folds, with 1 fold serving as the test set and the other 4 as training for each cross-validation process, giving us 5 evaluations of 80% of the data set as the training set and 20% as the test set. Thus, the model performance was reported with the mean and SD of each metric over the 5 validation runs. Additionally, we calculated the Shapley Additive Explanations (SHAP) value for each of the algorithms tested to evaluate the contribution of each feature of the algorithms to making the final predictions.

Results

Predictors Used for Models

All available baseline characteristic variables were iterated through the models to find the best combination of predictors for achievement of MCID for VAS neck pain (Tables 2 and 3). The final set of predictors used for the models included the following: age; sex; BMI; approach; number of levels treated; and baseline neck pain, arm pain, NDI, EQ-VAS, EQ-5D, and mJOA scores. The SHAP value plots are shown in Supplementary Fig. 1A–I for each of the models. Baseline VAS neck pain was the highest positive contributing factor for making the prediction for achieving MCID in neck pain at respective follow-up for logistic regression, decision tree, extra trees, random forest, gaussian naïve Bayes, and multilayer perceptron. Baseline EQ-VAS score was the highest positive contributor for support vector machine, and baseline NDI was highest for k–nearest neighbors.

Model Predictions for Neck Pain at 3 Months

The follow-up rate was 83.7% (955 of 1141 patients) at 3 months. The results for predicting MCID for VAS neck pain after surgery at 3 months are shown in Table 4. In total, 535 patients (46.9%) achieved MCID for neck pain at 3 months after surgery; of these, 501 patients (93.6%) were satisfied. Logistic regression demonstrated the best accuracy (0.76 ± 0.031), followed by F1 score (0.759 ± 0.019) and AUROC (0.762 ± 0.027). The best precision was demonstrated by logistic regression with 0.724 ± 0.058, whereas the best recall/sensitivity was demonstrated by multilayer perceptron with 0.841 ± 0.094. The highest specificity was shown by support vector machine with 0.952 ± 0.013. The AUROC and precision/recall curves are shown in Fig. 1A and B, respectively.

Model Predictions for Neck Pain at 24 Months

The follow-up rate was 80.3% (916 of 1141 patients) at 24 months. In total, 569 patients (49.9%) achieved MCID in neck pain at 24 months after surgery; of these, all 569 patients (100%) were satisfied. The model metric results

TABLE 2. Baseline characteristics of patients who did versus did not achieve MCID in neck pain after surgery for CSM at 3 months

	Achieved MCID, n = 535	Did Not Achieve MCID, n = 606	p Value
Age in yrs, mean (SD)	59.7 (11.5)	61.3 (12.0)	0.02*
Female sex, no. (%)	279 (52.1)	262 (43.2)	0.003*
BMI, mean (SD)	30.3 (6.43)	30.1 (6.46)	0.61*
Insurance, no. (%)			0.08
Medicare	196 (36.6)	244 (40.3)	
Medicaid	40 (7.5)	39 (6.4)	
VA/government	16 (3.0)	12 (2.0)	
Private	281 (52.5)	298 (49.2)	
College or higher, no. (%)	165 (30.8)	239 (39.4)	0.009
Caucasian race, no. (%)	398 (74.4)	472 (77.9)	0.16
Smoking, no. (%)	98 (18.3)	106 (17.5)	
Comorbidities, no. (%)			
Diabetes	125 (23.4)	120 (19.8)	0.15
Depression	124 (23.2)	127 (21.0)	0.37
Anxiety	98 (18.3)	114 (18.8)	0.83
CAD	49 (9.2)	59 (9.7)	0.74
PVD	21 (3.9)	21 (3.5)	0.68
Arthritis	152 (28.4)	174 (28.7)	0.91
CKD	20 (3.7)	28 (4.6)	0.46
COPD	38 (7.1)	43 (7.1)	0.20
MS	8 (1.5)	10 (1.7)	0.81
PD	1 (0.2)	4 (0.7)	0.20
ASA grade, no. (%)			0.88
1	12 (2.2)	10 (1.7)	
2	275 (51.4)	304 (50.2)	
3	238 (44.5)	285 (47.0)	
4	11 (2.1)	7 (1.2)	
Currently employed, no. (%)	291 (54.4)	280 (46.2)	0.83
Procedure breakdown, no. (%)			0.004*
ACDF	317 (59.3)	338 (55.8)	
ACCF	50 (9.3)	49 (8.1)	
CDR	26 (4.9)	11 (1.8)	
Laminectomy w/ fusion	106 (19.8)	132 (21.8)	
Laminectomy w/o fusion	20 (3.7)	48 (7.9)	
Laminoplasty	16 (3.0)	28 (4.6)	
Radicular deficit, no. (%)	183 (34.2)	172 (28.4)	0.04
Radicular arm pain, no. (%)	280 (52.3)	244 (40.3)	<0.001
Numbness, no. (%)	324 (60.6)	352 (58.1)	0.40
Neck pain, no. (%)	397 (74.2)	332 (54.8)	<0.001
Motor deficit, no. (%)	339 (63.4)	356 (58.7)	0.11
Duration of symptoms, no. (%)			0.88
<3 mos	75 (14.0)	71 (11.7)	
3–12 mos	155 (29.0)	210 (34.7)	
>12 mos	253 (47.3)	264 (43.6)	
Independent ambulation, no. (%)	444 (83.0)	488 (80.5)	0.52
Underlying pathology, no. (%)			
Intervertebral disc herniation	156 (29.2)	159 (26.2)	0.27
Foraminal stenosis	225 (42.1)	263 (43.4)	0.65
Central stenosis	398 (74.4)	463 (76.4)	0.43
Dynamic instability at level of surgery	12 (22.6)	20 (3.3)	0.27
Pseudarthrosis	0 (0)	2 (0.3)	0.16
Adjacent-segment disease	0 (0)	5 (0.8)	0.12
Baseline neck pain on VAS, mean (SD)	7.31 (1.91)	3.40 (3.09)	<0.001*
Baseline arm pain on VAS, mean (SD)	6.13 (3.12)	3.79 (3.34)	<0.001*
Baseline NDI score, mean (SD)	46.3 (17.7)	31.6 (20.9)	<0.001*
Baseline mJOA score, mean (SD)	11.8 (2.83)	12.3 (2.81)	0.004*
Baseline EQ-VAS score, mean (SD)	57.2 (21.7)	61.1 (21.3)	0.002*
Baseline EQ-5D score, mean (SD)	0.52 (0.21)	0.60 (0.21)	<0.001*
Levels treated, mean (SD)	2.51 (1.48)	2.70 (1.52)	0.03*

ACCF = anterior cervical corpectomy and fusion; ACDF = anterior cervical discectomy and fusion; ASA = American Society of Anesthesiologists; CAD = coronary artery disease; CDR = cervical disc replacement; CKD = chronic kidney disease; COPD = chronic obstructive pulmonary disease; MS = multiple sclerosis; PD = Parkinson's disease; PVD = peripheral vascular disease; VA = Veterans Affairs.

Patient-cohort identity was based on imputation for those missing 3-month VAS neck pain follow-up.

* Variables included in the machine learning models after logistic regression analysis.

are shown in Table 5. Logistic regression demonstrated the best accuracy (0.773 ± 0.044), followed by F1 score (0.777 ± 0.039) and AUROC (0.773 ± 0.043). The best precision was demonstrated by logistic regression with 0.780 ± 0.097, whereas the best recall/sensitivity was demonstrated by extra trees with 0.817 ± 0.115. The highest specificity was shown by logistic regression with 0.747 ± 0.18. The AUROC and precision/recall curves are shown in Fig. 2A and B, respectively.

Discussion

In total, 535 patients (46.9%) achieved MCID for neck pain at 3 months and 569 patients (49.9%) achieved it at 24 months. In each follow-up cohort, 501 patients (93.6%) were satisfied at 3 months after surgery and 569 patients (100%) were satisfied at 24 months after surgery. Of the supervised machine learning algorithms tested, logistic regression demonstrated the best accuracy—followed by F1 score and AUROC—at predicting achievement of

TABLE 3. Baseline characteristics of patients who did versus did not achieve MCID in neck pain after surgery for CSM at 24 months

	Achieved MCID, n = 569	Did Not Achieve MCID, n = 572	p Value
Age in yrs, mean (SD)	60.3 (11.5)	60.8 (12.1)	0.47*
Female sex, no. (%)	286 (50.3)	255 (44.6)	0.06*
BMI, mean (SD)	30.3 (6.44)	30.0 (6.44)	0.43*
Insurance, no. (%)			0.42
Medicare	209 (36.7)	231 (40.4)	
Medicaid	53 (9.3)	26 (4.5)	
VA/government	12 (2.1)	16 (2.8)	
Private	291 (51.1)	288 (50.3)	
College or higher, no. (%)	174 (30.6)	230 (40.2)	<0.001
Caucasian race, no. (%)	434 (76.3)	436 (76.2)	0.99
Smoking, no. (%)	114 (20.0)	90 (15.7)	0.04
Comorbidities, no. (%)			
Diabetes	128 (22.5)	117 (20.5)	0.40
Depression	131 (23.0)	120 (21.0)	0.41
Anxiety	108 (19.0)	104 (18.2)	0.73
CAD	59 (10.4)	49 (8.6)	0.30
PVD	25 (4.4)	17 (3.0)	0.20
Arthritis	154 (27.1)	172 (30.1)	0.26
CKD	20 (3.5)	28 (4.9)	0.25
COPD	43 (7.6)	38 (6.6)	0.55
MS	11 (1.9)	7 (1.2)	0.35
PD	1 (0.2)	4 (0.7)	0.17
ASA grade, no. (%)			0.19
1	11 (1.9)	11 (1.9)	
2	279 (49.0)		
3	269 (47.3)	254 (44.4)	
4	10 (1.8)	8 (1.4)	
Currently employed, no. (%)	257 (45.2)	270 (47.2)	0.45
Procedure breakdown, no. (%)			0.14*
ACDF	333 (58.5)	322 (56.3)	
ACCF	52 (9.1)	47 (8.2)	
CDR	21 (3.7)	16 (2.8)	
Laminectomy w/ fusion	116 (20.4)	122 (21.3)	
Laminectomy w/o fusion	24 (4.2)	44 (7.7)	
Laminoplasty	23 (4.0)	21 (3.7)	
Radicular deficit, no. (%)	185 (32.5)	170 (29.7)	0.31
Radicular arm pain, no. (%)	285 (50.1)	239 (41.8)	0.005
Numbness, no. (%)	336 (59.1)	340 (59.4)	0.89
Neck pain, no. (%)	414 (72.8)	315 (55.1)	<0.001
Motor deficit, no. (%)	356 (62.6)	339 (59.3)	0.25
Duration of symptoms, no. (%)			0.39
<3 mos	69 (12.1)	77 (13.5)	
3–12 mos	173 (30.4)	192 (33.6)	
>12 mos	275 (48.3)	242 (42.3)	

CONTINUED IN NEXT COLUMN »

» CONTINUED FROM PREVIOUS COLUMN

TABLE 3. Baseline characteristics of patients who did versus did not achieve MCID in neck pain after surgery for CSM at 24 months

	Achieved MCID, n = 569	Did Not Achieve MCID, n = 572	p Value
Independent ambulation, no. (%)	465 (81.7)	467 (81.6)	0.86
Underlying pathology, no. (%)			
Intervertebral disc herniation	150 (26.4)	165 (28.8)	0.35
Foraminal stenosis	254 (44.6)	234 (40.9)	0.20
Central stenosis	429 (75.4)	432 (75.5)	0.96
Dynamic instability at level of surgery	17 (3.0)	15 (2.6)	0.71
Pseudarthrosis	1 (0.2)	1 (0.2)	0.99
Adjacent-segment disease	2 (0.4)	4 (0.7)	0.42
Baseline neck pain on VAS, mean (SD)	7.23 (1.95)	3.25 (3.08)	<0.001*
Baseline arm pain on VAS, mean (SD)	6.01 (3.18)	3.77 (3.33)	<0.001*
Baseline NDI score, mean (SD)	45.8 (17.3)	31.2 (21.3)	<0.001*
Baseline mJOA score, mean (SD)	11.8 (2.83)	12.2 (2.82)	0.02*
Baseline EQ-VAS score, mean (SD)	57.5 (21.5)	61.0 (21.5)	0.005*
Baseline EQ-5D score, mean (SD)	0.52 (0.21)	0.60 (0.21)	<0.001*
Levels treated, mean (SD)	2.61 (1.47)	2.61 (1.53)	0.99*

Patient-cohort identity was based on imputation for those missing 24-month VAS neck pain follow-up.
* Variables included in the machine learning models after logistic regression analysis.

MCID for neck pain at both follow-up time points, with fair performance. The best precision was demonstrated by logistic regression at 3 and 24 months. The best recall/sensitivity was demonstrated by multilayer perceptron at 3 months and by extra trees at 24 months. Highest specificity was shown by support vector machine at 3 months and logistic regression at 24 months.

In medical research, logistic regression is one of the most frequently used statistical analysis methods to explore the relationship between one or more independent variables (the predictors or features) and a binary dependent variable (the outcome). There are several reasons for its popularity in medical research applications.[13,14] First, logistic regression can be used to estimate the probability/risk of a particular outcome given the independent variables.[14] Second, confounding variables can be easily accounted for by holding the other independent variables constant while studying the relationship between each variable and the outcome.[15] Third, the process is fast, and the results are intuitive to interpret.[14] Particularly, the ex-

TABLE 4. Prediction of achieving MCID for neck pain at 3 months

Model	Accuracy	F1	AUROC	Precision	Recall/Sensitivity	Specificity
Logistic regression	0.76 (0.031)	0.759 (0.019)	0.762 (0.027)	0.724 (0.058)	0.804 (0.052)	0.721 (0.094)
Support vector machine	0.537 (0.012)	0.12 (0.024)	0.51 (0.012)	0.555 (0.109)	0.067 (0.014)	0.952 (0.013)
Decision tree	0.697 (0.042)	0.671 (0.046)	0.695 (0.041)	0.687 (0.058)	0.662 (0.074)	0.728 (0.084)
Extra trees	0.736 (0.029)	0.729 (0.039)	0.738 (0.028)	0.708 (0.056)	0.766 (0.115)	0.709 (0.118)
Random forest	0.726 (0.038)	0.719 (0.032)	0.728 (0.035)	0.702 (0.066)	0.746 (0.077)	0.709 (0.112)
Gaussian naïve Bayes	0.718 (0.041)	0.707 (0.033)	0.718 (0.037)	0.7 (0.079)	0.727 (0.081)	0.709 (0.121)
k–nearest neighbors	0.674 (0.037)	0.678 (0.033)	0.677 (0.035)	0.636 (0.049)	0.733 (0.083)	0.622 (0.106)
Multilayer perceptron	0.75 (0.034)	0.759 (0.034)	0.755 (0.032)	0.698 (0.052)	0.841 (0.094)	0.67 (0.106)
Extreme gradient boosted trees	0.691 (0.021)	0.666 (0.049)	0.689 (0.023)	0.676 (0.038)	0.669 (0.118)	0.709 (0.093)

Values are expressed as the mean (SD).

ponentiated logistic regression slope coefficient can be translated into an odds ratio that indicates how much the odds of a particular outcome change for a 1-unit increase in the independent variable (for continuous variables) or against a reference category (for categorical variables).[16] However, the use of logistic regression may become limited when the independent input variables have a complex relationship (e.g., nonlinear) with the dependent outcome variable.[17] This is because logistic regression models assume that the relationship between the predictor variables and the dependent variable is constant and unchanging in direction over the entire range of values. Furthermore, independent variables that are highly correlated with other independent variables cannot be used because the effect of these variables will produce imprecise logistic regression results.[17]

In an effort to expand on the current statistical methodology to discover new insights and improve the current prognostic and diagnostic accuracy in medical research, there has been an increasing amount of interest in exploring and assessing different types of machine learning models (Table 1).[18] When selecting the most appropriate model, it is important to consider which performance metric (i.e., accuracy, F1 score, AUROC, precision, recall/sensitivity, and specificity) is relevant in addressing a given clinical research question. Definitions and examples of different performance metrics are summarized in Table 6. In our study we applied several well-known supervised machine learning models to predict achievement of MCID in patients with neck pain. Because neck pain significantly impacts quality of life in this patient population, it would be useful to select a model that can correctly predict true achievement of MCID in neck pain out of all the predictions. In this context, precision would be the metric we would want to optimize. If the purpose of the study were to minimize the number of false negatives, the model that demonstrates high recall/sensitivity could be used to avoid incidences in which the patients remain undiagnosed and fail to receive proper treatment. Finally, if the goal were to minimize false positives, specificity would be useful.

For both short- and long-term follow-ups, logistic regression demonstrated the highest precision of all the models tested. The observation that logistic regression performs similarly (and slightly superiorly) to other com-

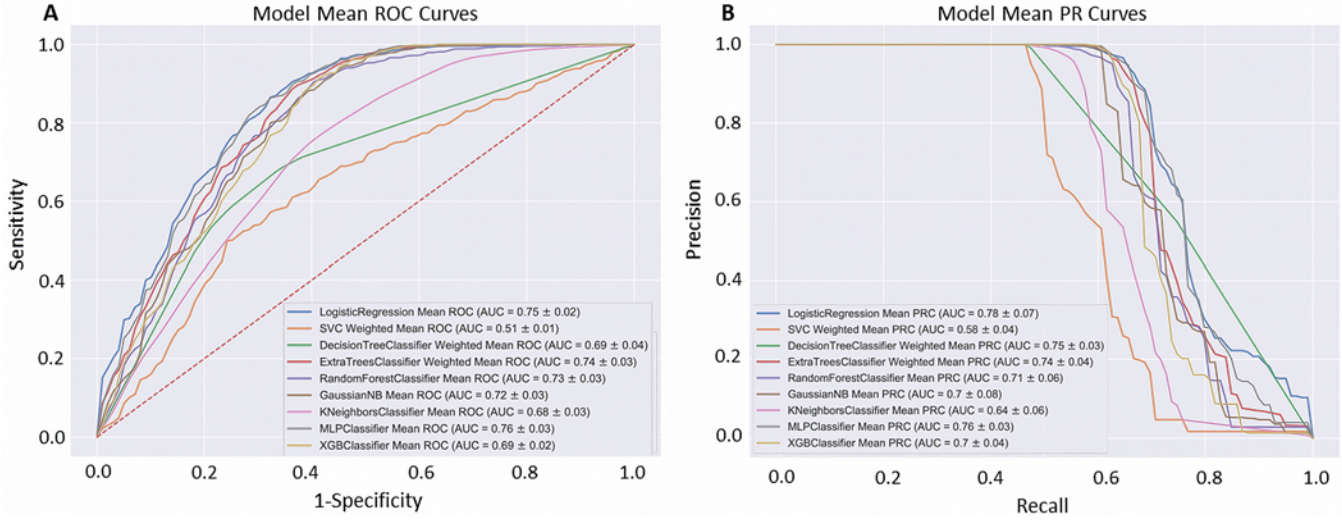

FIG. 1. AUROC (**A**) and precision/recall (**B**) curves at 3 months for prediction of MCID for neck pain. MLP = multilayer perceptron; NB = naïve Bayes; PR = precision/recall; SVC = support vector machine; XGB = extreme gradient boosted trees.

TABLE 5. Prediction of achieving MCID for neck pain at 24 months

Model	Accuracy	F1	AUROC	Precision	Recall/Sensitivity	Specificity
Logistic regression	0.773 (0.044)	0.777 (0.039)	0.773 (0.043)	0.78 (0.097)	0.8 (0.136)	0.747 (0.18)
Support vector machine	0.593 (0.033)	0.505 (0.128)	0.594 (0.032)	0.663 (0.078)	0.463 (0.253)	0.724 (0.262)
Decision tree	0.696 (0.033)	0.69 (0.026)	0.696 (0.032)	0.714 (0.078)	0.682 (0.092)	0.71 (0.136)
Extra trees	0.741 (0.043)	0.758 (0.023)	0.741 (0.042)	0.725 (0.088)	0.817 (0.115)	0.664 (0.184)
Random forest	0.728 (0.046)	0.746 (0.033)	0.729 (0.045)	0.715 (0.086)	0.802 (0.117)	0.656 (0.18)
Gaussian naïve Bayes	0.729 (0.045)	0.736 (0.036)	0.729 (0.044)	0.734 (0.093)	0.761 (0.122)	0.698 (0.18)
k–nearest neighbors	0.649 (0.046)	0.68 (0.034)	0.65 (0.046)	0.63 (0.055)	0.747 (0.081)	0.552 (0.134)
Multilayer perceptron	0.757 (0.026)	0.762 (0.013)	0.758 (0.025)	0.761 (0.088)	0.786 (0.124)	0.729 (0.169)
Extreme gradient boosted trees	0.72 (0.031)	0.728 (0.019)	0.72 (0.031)	0.72 (0.082)	0.758 (0.119)	0.682 (0.17)

Values are expressed as the mean (SD).

plex models is important because logistic regression–based machine learning remains a useful initial technique for addressing simple binary classification tasks such as those done in our study, without adding in the complexity, computational cost, and reduced interpretability that come with the more complex models.[19] Ultimately, model selection should depend on the characteristics of data sets and the clinical research question being addressed. If the database is imbalanced, the F1 score should be used. If the data set is well balanced (such as in our study), accuracy and AUROC may be more useful. In many clinical circumstances, the use of simple logistic regression models is likely to provide adequate performance. However, in circumstances that involve large data sets, unstructured data (i.e., imaging, genomics, texts), multiclass classification outcomes, and/or great value for small increase in performance, application of advanced machine learning algorithms may be indicated. For large data sets, extra trees, random forest, gaussian naïve Bayes, multilayer perceptron, and extreme gradient boosted trees can be considered. For unstructured data with complex feature relationships, multilayer perceptron and extreme gradient boosted trees may be more appropriate. For multiclass classification, models other than logistic regression would work well.

The medical literature has increasingly focused on clinical prediction and decision tools that use patient history, examination, and/or diagnostic tests to generate outcome-based risk stratification and/or intervention recommendations. These tools are often in the form of a simple clinical score, which can be easily calculated by clinicians and provide actionable information, clinical decision support, and standardization of care. Medical calculators that operationalize these tools are widely available. Grading criteria for these tools have been proposed, but are not commonly used, leading to a need for appropriate use and clinical validity demonstration, particularly for calculator tools that present recommended therapeutic or management steps. The decision to perform surgery is complex and is influenced by contextual details related to the patient, physician, and overall social support availability. Thus, it is essential to understand not only how clinical prediction and decision tools are used but also their appropriateness for a given environment and situation. As clinical risk tools become more prevalent, ensuring their quality and appropriate application of their results is crucial, particularly with

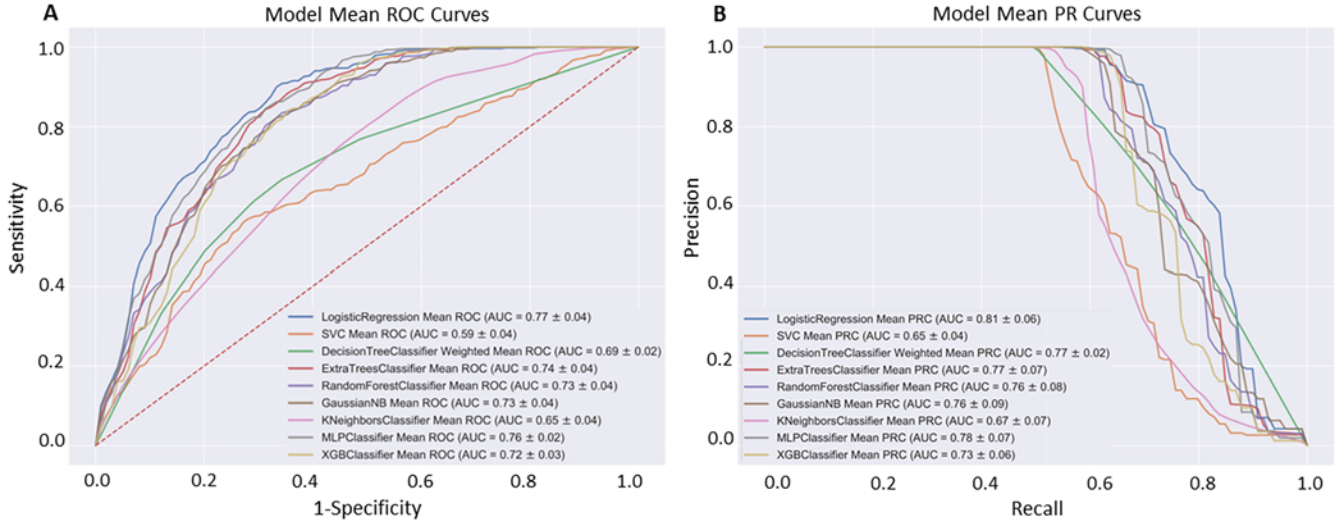

FIG. 2. AUROC (**A**) and precision/recall (**B**) curves at 24 months for prediction of MCID for neck pain.

TABLE 6. Descriptions of performance metrics with clinical context

Metric	Definition	Formula	When to Use
Accuracy	No. of observations (both positive & negative) that were correctly classified	TP + TN/TP + FP + TN + FN	Works well for balanced data sets
F1	Combination (harmonic mean) of precision & recall	$(1 + \beta^2)$ Precision * Recall/β^2 * Precision + Recall (higher beta when more weight should be put on recall over precision)	Can be used for imbalanced data sets
AUROC	Measures the trade-off btwn TPR & FPR	TPR = TP/TP + FN; FPR = FP/FP + TN	Works well for balanced data sets
Precision	No. of actual positive cases out of all positive predictions	TP/TP + FP	Used when occurrence of FP is undesirable (i.e., want to be confident of the predicted positives)
Recall/ sensitivity	No. of correct positive predictions out of all actual positive cases	TP/TP + FN	Used when occurrence of FN is undesirable (i.e., identification of positives is crucial); indicates how well the model can screen for a disease (negative test better rules out disease)
Specificity	No. of correct negative predictions out of all actual negative cases	TN/TN + FP	Used when occurrence of FP is undesirable (i.e., negative categorization is top priority); indicates how well the model can confirm diagnosis (positive test result better rules in the condition)

FN = false negative; FP = false positive; FPR = FP rate; TN = true negative; TP = true positive; TPR = TP rate.

the growing interest in embedding clinical scores within the electronic medical record and integrating them into clinical decision support mechanisms. Prospective validation studies are often necessary to assess the effectiveness of these tools when implemented in clinical care, given that initial derivation of a risk score may not include such validation.

Limitations

Limitations associated with retrospective studies, including selection bias and unaccounted confounding variables, are applicable here due to the nature of this study. The number of samples available for training the machine learning models was also limited—a bigger data set would have allowed for improved performance and would have enhanced the generalizability of the results to a broader population. Also, our study focused on well-established supervised machine learning algorithms for our simple binary classification task of predicting whether patients achieved MCID in neck pain or not, but we could have explored semisupervised and unsupervised methods if we had wanted to explore a more complex relationship among different variables and had a larger data set to work with. Additionally, assessment of VAS neck pain is subjective and may be more variable than other patient-reported outcomes such as NDI, mJOA, and EQ-5D—which may have contributed to the fair performance the models demonstrated at both short- and long-term follow-ups. Furthermore, our study showed that achievement of MCID in neck pain and patient satisfaction are highly correlated. However, satisfaction has many contributing factors (i.e., doctor/patient rapport, preoperative counseling, expectation setting), given that the NASS survey primarily assesses expectations. To this end, it would be interesting to explore the models' performance for other patient-reported outcomes.

Conclusions

Model selection depends on the characteristics of data sets and the clinical research question being addressed. For maximally predicting true achievement of MCID in neck pain out of all the predictions in our balanced data set, the appropriate metric for our study was precision. For both short- and long-term follow-ups, logistic regression demonstrated the highest precision of all the models tested. Logistic regression performed consistently the best of all models tested and remains a powerful model for clinical classification tasks.

Acknowledgments

This research was supported by the NeuroPoint Alliance (NPA), the Neurosurgery Research and Education Foundation (NREF), and the Spine Section. The NPA is a 501(c)(6) affiliate nonprofit organization of the AANS dedicated to the improvement of the quality of care in neurosurgical practice via the institution of national quality registries, such as the one used for this study. The NREF is the philanthropic arm of the AANS and has financially supported the creation and maintenance of the QOD. The Spine Section is a neurosurgical community formed in collaboration between the AANS and the CNS to advance spine and peripheral nerve patient care through education, research, and advocacy.

References

1. Nouri A, Cheng JS, Davies B, Kotter M, Schaller K, Tessitore E. Degenerative cervical myelopathy: a brief review of past perspectives, present developments, and future directions. *J Clin Med*. 2020;9(2):535.
2. Chan AK, Shaffrey CI, Gottfried ON, et al. Cervical spondylotic myelopathy with severe axial neck pain: is anterior or posterior approach better? *J Neurosurg Spine*. 2022;38(1):42-55.
3. van Middelkoop M, Rubinstein SM, Ostelo R, et al. Surgery versus conservative care for neck pain: a systematic review. *Eur Spine J*. 2013;22(1):87-95.

4. Cohen SP. Epidemiology, diagnosis, and treatment of neck pain. *Mayo Clin Proc*. 2015;90(2):284-299.
5. Elfanagely O, Toyoda Y, Othman S, et al. Machine learning and surgical outcomes prediction: a systematic review. *J Surg Res*. 2021;264:346-361.
6. Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas—the Public Health Disparities Geocoding Project. *Am J Public Health*. 2002;92(7):1100-1102.
7. Tetreault L, Kopjar B, Nouri A, et al. The modified Japanese Orthopaedic Association scale: establishing criteria for mild, moderate and severe impairment in patients with degenerative cervical myelopathy. *Eur Spine J*. 2017;26(1):78-84.
8. Jenkins NW, Parrish JM, Lynch CP, et al. The association of preoperative duration of symptoms with clinical outcomes and minimal clinically important difference following anterior cervical discectomy and fusion. *Clin Spine Surg*. 2020; 33(9):378-381.
9. Parker SL, Godil SS, Shau DN, Mendenhall SK, McGirt MJ. Assessment of the minimum clinically important difference in pain, disability, and quality of life after anterior cervical discectomy and fusion: clinical article. *J Neurosurg Spine*. 2013;18(2):154-160.
10. Youssef JA, Heiner AD, Montgomery JR, et al. Outcomes of posterior cervical fusion and decompression: a systematic review and meta-analysis. *Spine J*. 2019;19(10):1714-1729.
11. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011; 12(2011):2825–2830.
12. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016;785-794.
13. Schober P, Vetter TR. Logistic regression in medical research. *Anesth Analg*. 2021;132(2):365-366.
14. Nick TG, Campbell KM. Logistic regression. *Methods Mol Biol*. 2007;404:273-301.
15. Müllner M, Matthews H, Altman DG. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. *Ann Intern Med*. 2002;136(2):122-126.
16. Boateng EY, Abaye D. A review of the logistic regression model with emphasis on medical research. *J Data Anal Info Proc*. 2019;7:190-207.
17. Zabor EC, Reddy CA, Tendulkar RD, Patil S. Logistic regression in clinical studies. *Int J Radiat Oncol Biol Phys*. 2022; 112(2):271-277.
18. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*. 2019;19(1):64.
19. Issitt RW, Cortina-Borja M, Bryant W, Bowyer S, Taylor AM, Sebire N. Classification performance of neural networks versus logistic regression models: evidence from healthcare practice. *Cureus*. 2022;14(2):e22443.

Disclosures

Dr. Mummaneni reported grants from NREF during the conduct of the study; personal fees from DePuy Synthes, Globus, NuVasive, Stryker, SI Bone, BK Medical, and Brainlab; stock ownership with Spinicity/ISD; and grants from AO Spine outside the submitted work. Dr. C. Shaffrey reported personal fees from NuVasive, Medtronic, SI Bone, and Proprio outside the submitted work. Dr. Bisson reported personal fees from Stryker, Medtronic, and MiRus; and stock ownership with Proprio outside the submitted work. Dr. Asher reported personal fees from Globus outside the submitted work. Dr. Coric reported personal fees from Spine Wave, Medtronic, and Globus Medical outside the submitted work. Dr. Potts reported royalties/consulting fees from Medtronic outside the submitted work. Dr. Foley reported royalties, consulting, and stock ownership from Medtronic; and stock ownership with Discgenics, Accelus, DuraStat, RevBio, NuVasive, and Spine Wave outside the submitted work. In addition, Dr. Foley has multiple patents with royalties paid from Medtronic. Dr. Wang reported personal fees from DePuy Synthes, Stryker, Spineology, Pacira, Surgalign, and NuVasive; and purchased stock from Kinesiometrics, Medical Device Partners, and ISD outside the submitted work. In addition, Dr. Wang has a patent with DePuy Synthes with royalties paid. Dr. Fu reported personal fees from Johnson & Johnson outside the submitted work. Dr. Virk reported personal fees as a consultant for DePuy Synthes and Brainlab, and stock ownership with OnPoint Surgical outside the submitted work. Dr. Knightly reported being chair of the NPA outside the submitted work. Dr. P. Park reported personal fees from Globus, NuVasive, DePuy Synthes, Accelus, and LifeNet; grants from Cerapedics, SI Bone, ISSG, and DePuy Synthes; royalties from Globus; and lodging and travel for meeting from Medtronic outside the submitted work. Dr. Turner reported grants and personal fees from ATEC, SeaSpine, and NuVasive outside the submitted work. Dr. Sherrod reported grants from AO Spine and Cervical Spine Research Society outside the submitted work; in addition, Dr. Sherrod has a patent for a motorized skeletal traction device pending. Dr. Agarwal reported personal fees from Thieme Medical Publishers and Springer International Publishing outside the submitted work. Dr. Chou reported personal fees from Globus and Orthofix outside the submitted work.

Author Contributions

Conception and design: Chan, C Park, Mummaneni, CI Shaffrey, Bisson, Virk, Knightly, Haid, Bydon. Acquisition of data: Chan, Mummaneni, CI Shaffrey, Bisson, Asher, Coric, Potts, Foley, Wang, Fu, Virk, Knightly, P Park, Upadhyaya, ME Shaffrey, Buchholz, Tumialán, Sherrod, Chou. Analysis and interpretation of data: Chan, C Park, Mummaneni, Tang, Coric, Virk, ME Shaffrey, Turner, Agarwal, Haid. Drafting the article: Chan, C Park, Tang, ME Shaffrey, Agarwal. Critically revising the article: Chan, C Park, Mummaneni, CI Shaffrey, Tang, Bisson, Potts, Foley, Wang, Meyer, Upadhyaya, ME Shaffrey, Agarwal, Haid, Bydon. Reviewed submitted version of manuscript: Chan, C Park, Gottfried, CI Shaffrey, Tang, Bisson, Asher, Coric, Potts, Foley, Wang, Virk, Knightly, Meyer, P Park, Upadhyaya, Turner, Sherrod, Agarwal, Chou, Haid, Bydon. Approved the final version of the manuscript on behalf of all authors: Chan. Statistical analysis: C Park. Administrative/technical/material support: Mummaneni, Bydon. Study supervision: Mummaneni, CI Shaffrey, Bisson, Potts.

Supplemental Information

Online-Only Content

Supplemental material is available online.

Supplementary Tables and Figure. https://thejns.org/doi/suppl/10.3171/2023.3.FOCUS2372.

Correspondence

Andrew K. Chan: Columbia University Vagelos College of Physicians and Surgeons, The Och Spine Hospital at NewYork-Presbyterian, New York, NY. akc2136@columbia.edu.