Development of Machine Learning–Based Models to Predict Treatment Response to Spinal Cord Stimulation

Amir Hadanny, MD*
Tessa Harland, MD*
Olga Khazen, BS‡
Marisa DiMarzio, PhD‡
Anthony Marchese, BS‡
Ilknur Telkes, PhD‡
Vishad Sukul, MD*
Julie G. Pilitsis, MD, PhD *‡

*Department of Neurosurgery, Albany Medical College, Albany, New York, USA; ‡Department of Neuroscience and Experimental Therapeutics, Albany Medical College, Albany, New York, USA

Correspondence:
Julie G. Pilitsis, MD, PhD,
AMC Neurosurgery Group,
Physicians Pavilion, 1st Floor,
47 New Scotland Ave, MC 10,
Albany, NY 12208, USA.
Email: jpilitsis@yahoo.com

Received, June 17, 2021.
Accepted, November 3, 2021.
Published Online, February 22, 2022.

BACKGROUND: Despite spinal cord stimulation's (SCS) proven efficacy, failure rates are high with no clear understanding of which patients benefit long term. Currently, patient selection for SCS is based on the subjective experience of the implanting physician.
OBJECTIVE: To develop machine learning (ML)–based predictive models of long-term SCS response.
METHODS: A combined unsupervised (clustering) and supervised (classification) ML technique was applied on a prospectively collected cohort of 151 patients, which included 31 features. Clusters identified using unsupervised K-means clustering were fitted with individualized predictive models of logistic regression, random forest, and XGBoost.
RESULTS: Two distinct clusters were found, and patients in the cohorts significantly differed in age, duration of chronic pain, preoperative numeric rating scale, and preoperative pain catastrophizing scale scores. Using the 10 most influential features, logistic regression predictive models with a nested cross-validation demonstrated the highest overall performance with the area under the curve of 0.757 and 0.708 for each respective cluster.
CONCLUSION: This combined unsupervised–supervised learning approach yielded high predictive performance, suggesting that advanced ML-derived approaches have potential to be used as a functional clinical tool to improve long-term SCS outcomes. Further studies are needed for optimization and external validation of these models.

KEY WORDS: Spinal cord stimulation, Pain, Machine learning, Prediction

Neurosurgery 90:523–532, 2022 https://doi.org/10.1227/neu.0000000000001855

Spinal cord stimulation (SCS) is an Food and Drug Administration-approved treatment for managing chronic pain, most commonly for medically refractory back and neck pain and complex regional pain syndrome (CRPS). Devices have been increasingly used over the past 5 years at a growth rate of 20%, in part due to the opioid epidemic.[1] Despite patients undergoing psychological assessment and a trial of SCS before implant, suboptimal outcomes after SCS implant may occur in as many as 50% of patients at 2 years,[2] explant rates hover around 10%, and failure rates are estimated to be 25% to 30%.[2,3] There remains a lack of a clear understanding of which patients benefit long term. Thus, the ability to accurately predict patients who will not benefit from SCS would reduce the high financial burden of failed implants that plague the neuromodulation field. Currently, patient selection for SCS is based on the subjective experience of the implanting physician.[4]

Here, we address previous limitations by using the largest single-center database of prospectively collected longitudinal SCS outcomes[5-8] and applying a combination of unsupervised clustering and supervised classification to obtain individualized models for each subgroup/cluster of patients. Specifically, we use machine learning (ML) techniques for developing and internally validating predictive models for long-term SCS response.

ABBREVIATIONS: BDI, Beck's depression inventory; **CRPS,** complex regional pain syndrome; **CV,** cross-validation; **FBSS,** failed back surgery syndrome; **LR,** logistic regression; **ML,** machine learning; **MPQ,** McGill pain questionnaire; **NPV,** negative predictive value; **NRS,** numeric rating scale; **ODI,** Oswestry disability index; **PCS,** pain catastrophizing scale; **SCS,** spinal cord stimulation.

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

METHODS

Patients

The study protocol was approved by the Institutional Review Board. Data were collected prospectively and longitudinally except where otherwise noted. All patients who were consented to participate in the prospective outcomes database, underwent permanent SCS placement between November 1, 2012, and March 31, 2019, and had a 1-year follow-up (10-14 months) were included in our model (Figure 1). Of 261 patients who had baseline data included in the database, 108 patients were excluded. One patient requested and underwent explant before the 1-year follow-up, and 7 underwent revision surgeries (3 migration, 3 inadequate coverage, and 1 generator repositioning). The other 100 patients had their outcomes collected outside the designated time point (10-14 months). Both demographics and pain outcome data were gathered. Pain outcomes included the numeric rating scale score (NRS), pain catastrophizing scale (PCS), McGill pain questionnaire (MPQ), Oswestry disability index (ODI), and Beck's depression inventory (BDI). The NRS score documents pain intensity.[9] Anxiety and depression features were extracted from the neuropsychological evaluations that all patients underwent prospectively before surgery. All patients were approved as stable and fit for neuromodulation surgery by a trained pain psychologist.

Pain outcomes were collected in all patients pre-SCS placement and at 1-year postoperative follow-up. Patients were classified as responders if they had more than a 50% reduction of NRS (calculated as [baseline NRS – 1-year NRS/baseline NRS] × 100) and as high responders if they had more than a 70% NRS reduction.[10]

Features

Our database contained 49 features. We narrowed our focus to variables that could serve as preoperative predictors for training ML models, thus excluding 24 operative and postoperative factors. Features were collected from medical records. Pain diagnosis was divided into failed back surgery syndrome, CRPS, chronic neuropathic pain, or others (occipital neuralgia, plexitis, tethered cord, and combined diagnosis). Anxiety and depression features were processed using ordinal integer encoding (none/mild/moderate/severe). All other categorical features were processed using one-hot encoding (none/exists). Multicollinearity was evaluated, and highly correlated features (>0.7) were excluded (**Figure S1**, **Supplemental Digital Content**, http://links.lww.com/NEU/B419). After encoding of categorical features and exclusion of the 4 correlated features, a total of 31 factors were considered during model development.

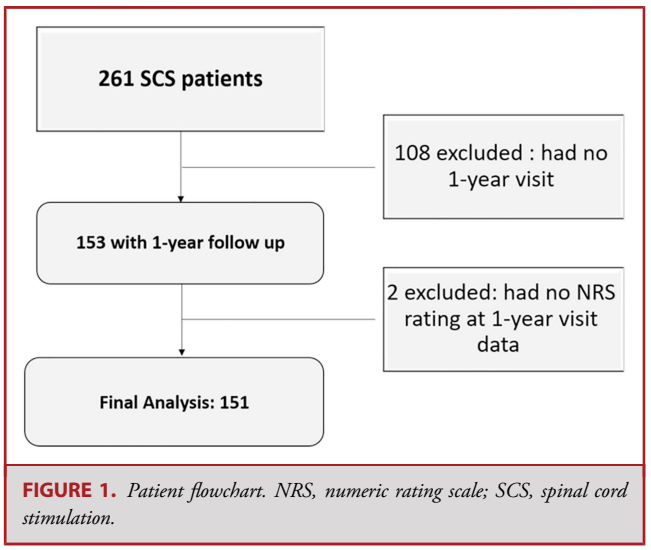

FIGURE 1. *Patient flowchart. NRS, numeric rating scale; SCS, spinal cord stimulation.*

Approach

A combined unsupervised and supervised ML approach was used over 2 stages. First, the study aimed to explore the potential presence of coherent patient clusters/phenotypes.[11] The K-means algorithm was used to discover patient subgroups from a mere data-driven perspective.[12] The patients were clustered based on the following numeric features: age, pain duration (in months), baseline NRS score, and baseline PCS total score. The elbow method was used to determine the number of clusters (K) (**Figure S2**, **Supplemental Digital Content**, http://links.lww.com/NEU/B419). The second stage focused on the development of ML models for each cluster separately to predict SCS outcomes (**Figure S3**, **Supplemental Digital Content**, http://links.lww.com/NEU/B419).

Models

Models were developed, and prediction performance was evaluated using a nested cross-validation (CV) scheme with K = 10-fold for both inner (hyperparameter tuning and feature selection) and outer (testing) loops. This approach reduces the overfitting of data and the optimistic bias in error estimation in small sample sizes.[13] Missing values imputation using the mean/mode method and numeric features normalization were performed on each iteration of the outer loop. Because of the significant imbalance of nonhigh responders to high responders, the synthetic minority oversampling technique was additionally applied to each loop iteration in the high-responder models.

Feature selection was performed using the 10 most influential features based on importance weights per model. Prediction performance was averaged across all outer loop folds. Models tested included logistic regression (LR), Random Forest, and XGBoost. Hyperparameter tuning details are given in **Table S1**, **Supplemental Digital Content**, http://links.lww.com/NEU/B419. Predictive performance was assessed by the area (AUC) under the receiver operating characteristic curve, specificity, sensitivity, positive predictive value, and negative predictive value.[14]

Statistical Analysis

Continuous data were expressed as mean ± standard deviation. The normal distribution for all variables was tested using the Kolmogorov–Smirnov test. Categorical data were expressed in numbers and percentages. Univariate analysis between responders and nonresponders was performed using unpaired *t*-tests and χ^2/Fisher's exact test to identify significant variables. The statistical significance threshold was set to 0.05. Data were statistically analyzed, and models were developed and tested using Python (Python Software Foundation; http://python.org).

Data Availability

The data that support the findings of this study are available on request from the corresponding author (J.P.).

RESULTS

A total of 151 SCS participants with a mean age of 54.8 ± 12.0 years were included in this study (Figure 1). Seventy-eight (51.7%) were treated for failed back surgery syndrome, 24 (15.9%) were treated for

CRPS, and 18 (11.9%) were treated for neuropathic pain. The majority of participants suffered from back (60.3%) and leg pain (70.2%), whereas 34.4% suffered from pelvic pain and 22.5% had arm pain.

Sixty-two participants demonstrated at least a 50% NRS reduction at 1 year (responders), and of those, 31 demonstrated at least a 70% NRS reduction at 1 year (high responders). The average age was

TABLE 1. Patients' Characteristics Divided by Nonresponders and Responders

	Missing	Total	Nonresponders	Responders	*P*-value
N		151	89	62	
Age, y, mean (SD)	0	54.8 (12.0)	53.3 (11.7)	57.0 (12.3)	.065
Sex, n (%)					
Male	0	67 (44.4)	36 (40.4)	31 (50.0)	.319
Female		84 (55.6)	53 (59.6)	31 (50.0)	
BMI, mean (SD)	0	32.2 (7.6)	32.7 (7.7)	31.4 (7.4)	.327
Pain duration, mean (SD)	3	47.8 (17.5)	48.3 (16.8)	47.2 (18.5)	.718
Pelvis pain, n (%)		52 (34.4)	36 (40.4)	16 (25.8)	.091
Back pain, n (%)		91 (60.3)	58 (65.2)	33 (53.2)	.191
Neck pain, n (%)		28 (18.5)	20 (22.5)	8 (12.9)	.202
Leg pain, n (%)		106 (70.2)	66 (74.2)	40 (64.5)	.274
Arm pain, n (%)		34 (22.5)	28 (31.5)	6 (9.7)	**.003**
FBSS	0	78 (51.7)	44 (49.4)	34 (54.8)	.732
CRPS		24 (15.9)	14 (15.7)	10 (16.1)	
Diagnosis, n (%)					
Neuropathy		18 (11.9)	10 (11.2)	8 (12.9)	
Others		31 (20.5)	21 (23.6)	10 (16.1)	
Psychiatric family history, n (%)					
None	0	142 (94.0)	81 (91.0)	61 (98.4)	.082
Yes		9 (6.0)	8 (9.0)	1 (1.6)	
Anxiety, n (%)					
None	0	115 (76.2)	67 (75.3)	48 (77.4)	.956
Mild		32 (21.2)	20 (22.5)	12 (19.4)	
Moderate		2 (1.3)	1 (1.1)	1 (1.6)	
Severe		2 (1.3)	1 (1.1)	1 (1.6)	
Depression, n (%)					
None	0	97 (64.2)	54 (60.7)	43 (69.4)	.434
Mild		49 (32.5)	33 (37.1)	16 (25.8)	
Moderate		3 (2.0)	1 (1.1)	2 (3.2)	
Severe		2 (1.3)	1 (1.1)	1 (1.6)	
Smoking, n (%)					
Current	0	67 (44.4)	46 (51.7)	21 (33.9)	**.020**
Never		43 (28.5)	18 (20.2)	25 (40.3)	
Former		41 (27.2)	25 (28.1)	16 (25.8)	
Insurance, n (%)					
Commercial	0	110 (72.8)	57 (64.0)	53 (85.5)	**.027**
Medicare		3 (2.0)	3 (3.4)		
No-fault (auto)		5 (3.3)	4 (4.5)	1 (1.6)	
Workers' compensation		33 (21.9)	25 (28.1)	8 (12.9)	
Previous spinal surgeries (SD)	0	1.3 (1.6)	1.2 (1.2)	1.3 (1.9)	.815
Months from previous surgery[a] (SD)	0	80.6 (101.0)	78.4 (97.3)	85.1 (109.8)	.79
NRS baseline, mean (SD)	0	6.9 (1.7)	7.1 (1.9)	6.7 (1.5)	.084
ODI_Baseline, mean (SD)	10	25.2 (7.2)	26.4 (6.8)	23.3 (7.5)	**.014**
BDI_Baseline, mean (SD)	21	13.3 (9.0)	14.5 (9.6)	11.6 (7.8)	.056
PCSTotal_Baseline, mean (SD)	17	23.2 (12.9)	24.2 (13.0)	21.7 (12.6)	.269
MPQTotal_Baseline, mean (SD)	0	5.2 (2.8)	5.6 (2.7)	4.6 (2.9)	**.029**
MPQAffective_Baseline, mean (SD)	0	0.7 (0.9)	0.8 (0.9)	0.6 (1.0)	.330

BMI, body mass index; BDI, Beck's depression inventory; CRPS, complex regional pain syndrome; FBSS, failed back surgery syndrome; MPQ, McGill pain questionnaire; NRS, numeric rating scale; ODI, Oswestry disability index; PCS, pain catastrophizing scale; SD, standard deviation.

[a]Only including patients with at least 1 previous surgery.

N = Sample size.

The statistically significant differences are highlighted in bold.

53.3 ± 11.7 years in nonresponders compared with 57.0 ± 12.3 years in responders (P = .065). The statistical analysis demonstrated that nonresponders more frequently reported arm pain (P = .003), smoked (P = .02), and had noncommercial insurance including workers' compensation (P = .027). Nonresponders also had a statistically higher baseline ODI (P = .014) and MPQ score (P = .029).

TABLE 2. Patients' Characteristics Divided by High Responders vs Nonhigh Responders

	Missing	Total	Nonhigh responders	High responders	*P*-value
N		151	120	31	
Age, y, mean (SD)	0	54.8 (12.0)	53.9 (11.5)	58.2 (13.5)	.111
Sex, n (%)					
Male	0	67 (44.4)	57 (47.5)	10 (32.3)	.187
Female		84 (55.6)	63 (52.5)	21 (67.7)	
BMI, mean (SD)	0	32.2 (7.6)	32.9 (7.8)	29.2 (6.3)	**.008**
Pain duration, mean (SD)	3	47.8 (17.5)	47.8 (17.2)	47.8 (18.8)	.999
Pelvis pain, n (%)		52 (34.4)	47 (39.2)	5 (16.1)	**.028**
Back pain, n (%)		91 (60.3)	78 (65.0)	13 (41.9)	**.033**
Neck pain, n (%)		28 (18.5)	25 (20.8)	3 (9.7)	.244
Leg pain, n (%)		106 (70.2)	89 (74.2)	17 (54.8)	.06
Arm pain, n (%)		34 (22.5)	32 (26.7)	2 (6.5)	**.031**
Diagnosis, n (%)					
FBSS	0	78 (51.7)	62 (51.7)	16 (51.6)	.169
CRPS		24 (15.9)	20 (16.7)	4 (12.9)	
Neuropathy		18 (11.9)	11 (9.2)	7 (22.6)	
Others		31 (20.5)	27 (22.5)	4 (12.9)	
Psychiatric family history, n (%)					
None	0	142 (94.0)	111 (92.5)	31 (100.0)	.205
Yes		9 (6.0)	9 (7.5)		
Anxiety, n (%)					
None	0	115 (76.2)	89 (74.2)	26 (83.9)	.613
Mild		32 (21.2)	27 (22.5)	5 (16.1)	
Moderate		2 (1.3)	2 (1.7)		
Severe		2 (1.3)	2 (1.7)		
Depression, n (%)					
None	0	97 (64.2)	77 (64.2)	20 (64.5)	.138
Mild		49 (32.5)	41 (34.2)	8 (25.8)	
Moderate		3 (2.0)	1 (0.8)	2 (6.5)	
Severe		2 (1.3)	1 (0.8)	1 (3.2)	
Smoking, n (%)					
Current	0	67 (44.4)	56 (46.7)	11 (35.5)	.068
Never		43 (28.5)	29 (24.2)	14 (45.2)	
Former		41 (27.2)	35 (29.2)	6 (19.4)	
Insurance, n (%)					
Commercial	0	110 (72.8)	85 (70.8)	25 (80.6)	.640
Medicare		3 (2.0)	3 (2.5)		
No-fault (auto)		5 (3.3)	4 (3.3)	1 (3.2)	
Workers' compensation		33 (21.9)	28 (23.3)	5 (16.1)	
Previous spinal surgeries (SD)	0	1.3 (1.6)	1.3 (1.5)	1.2 (1.7)	.923
Months from previous surgery (SD)[a]	0	80.6 (101.0)	81.1 (103.8)	78.3 (88.2)	.921
AverageNRS_Baseline, mean (SD)	0	6.9 (1.7)	7.0 (1.8)	6.8 (1.6)	.529
ODI_Baseline, mean (SD)	10	25.2 (7.2)	26.1 (7.1)	21.7 (6.9)	**.004**
BDI_Baseline, mean (SD)	21	13.3 (9.0)	14.1 (9.2)	10.3 (7.3)	**.034**
PCSTotal_Baseline, mean (SD)	17	23.2 (12.9)	23.6 (13.0)	21.5 (12.4)	.448
MPQTotal_Baseline, mean (SD)	0	5.2 (2.8)	5.6 (2.7)	3.6 (2.4)	**<.001**
MPQAffective_Baseline, mean (SD)	0	0.7 (0.9)	0.8 (1.0)	0.4 (0.7)	**.014**

BDI, Beck's depression inventory; BMI, body mass index; CRPS, complex regional pain syndrome; FBSS, failed back surgery syndrome; MPQ, McGill pain questionnaire; NRS, numeric rating scale; ODI, Oswestry disability index; PCS, pain catastrophizing scale; SD, standard deviation.

[a]Only including patients with at least 1 previous surgery.

N = Sample size.

The statistically significant differences are highlighted in bold.

TABLE 3. Patients' Characteristics Divided by Two Distinct Clusters

	Missing	Total	Cluster 1	Cluster 2	*P*-value
M		151	79	72	
Responder, n (%)		62	29 (36.7%)	33 (45.8%)	.331
High responder, n (%)		31 (20.5)	14 (17.7)	17 (23.6)	.488
Age, y, mean (SD)	0	54.8 (12.0)	51.5 (11.8)	58.5 (11.2)	**<.001**
Sex, n (%)					
Female	0	84 (55.6)	48 (60.8)	36 (50.0)	.244
Male		67 (44.4)	31 (39.2)	36 (50.0)	
BMI, mean (SD)	0	32.2 (7.6)	32.3 (7.7)	32.0 (7.6)	.838
Pain duration, mean (SD)	3	47.8 (17.5)	43.8 (19.4)	52.3 (13.8)	**.002**
Pelvis_Baseline, n (%)		52 (34.4)	29 (36.7)	23 (31.9)	.657
Back_Baseline, n (%)		91 (60.3)	46 (58.2)	45 (62.5)	.712
Neck_Baseline, n (%)		28 (18.5)	12 (15.2)	16 (22.2)	.368
Legs_Baseline, n (%)		106 (70.2)	57 (72.2)	49 (68.1)	.710
Arms_Baseline, n (%)		34 (22.5)	18 (22.8)	16 (22.2)	.911
Diagnosis (%)					
FBSS		78 (51.7)	34 (43.0)	44 (61.1)	**.04**
CRPS		24 (15.9)	19 (24.1)	5 (6.9)	**.008**
Neuropathy		18 (11.9)	8 (10.1)	10 (13.9)	.645
Others		31 (20.5)	18 (22.8)	13 (18.1)	.605
Psychiatric family history, n (%)					
None	0	142 (94.0)	74 (93.7)	68 (94.4)	1.000
Yes		9 (6.0)	5 (6.3)	4 (5.6)	
Anxiety, n (%)					
None	0	115 (76.2)	58 (73.4)	57 (79.2)	.534
Mild		32 (21.2)	18 (22.8)	14 (19.4)	
Moderate		2 (1.3)	2 (2.5)		
Severe		2 (1.3)	1 (1.3)	1 (1.4)	
Depression, n (%)					
None	0	97 (64.2)	48 (60.8)	49 (68.1)	.361
Mild		49 (32.5)	27 (34.2)	22 (30.6)	
Moderate		3 (2.0)	3 (3.8)		
Severe		2 (1.3)	1 (1.3)	1 (1.4)	
Smoking (%)					
Never smoking		67 (44.4)	38 (48.1)	29 (40.3)	
Former smoking		43 (28.5)	16 (20.3)	27 (37.5)	**.030**
Current smoking		41 (27.2)	25 (31.6)	16 (22.2)	
Insurance (%)					
Medicare		3 (2.0)		3 (4.2)	.106
Commercial		110 (72.8)	56 (70.9)	54 (75.0)	.701
No-fault		5 (3.3)	3 (3.8)	2 (2.8)	1
Workers' compensation		33 (21.9)	20 (25.3)	13 (18.1)	.378
Previous spinal surgeries (SD)	0	1.3 (1.6)	0.9 (1.2)	1.6 (1.8)	**.005**
Months from previous surgery[a] (SD)	0	80.6 (101)	66.1 (75.8)	92.9 (117.5)	.22
NRS Baseline, mean (SD)	0	6.9 (1.7)	7.9 (1.3)	5.8 (1.4)	**<.001**
ODI_Baseline, mean (SD)	10	25.2 (7.2)	27.5 (6.3)	22.6 (7.4)	**<.001**
BDI_Baseline, mean (SD)	21	13.3 (9.0)	17.0 (9.6)	9.6 (6.6)	**<.001**
PCSTotal_Baseline, mean (SD)	17	23.2 (12.9)	32.0 (9.7)	14.1 (8.7)	**<.001**
MPQTotal_Baseline, mean (SD)	0	5.2 (2.8)	5.6 (2.9)	4.8 (2.6)	**.048**
MPQAffective_Baseline, mean (SD)	0	0.7 (0.9)	0.8 (1.0)	0.6 (0.9)	.167

BDI, Beck's depression inventory; BMI, body mass index; CRPS, complex regional pain syndrome; FBSS, failed back surgery syndrome; MPQ, McGill pain questionnaire; NRS, numeric rating scale; ODI, Oswestry disability index; PCS, pain catastrophizing scale; SD, standard deviation.

[a]Only including patients with at least 1 previous surgery.

N = Sample size.

The statistically significant differences are highlighted in bold.

High responders had a lower body mass index (P = .008) and were less likely to have pelvic (P = .028), back (P = .033), or arm pain (P = .031) than nonresponders. They also had lower preoperative ODI (P = .004), BDI (P = .034), MPQ total (P < .001), and MPQ affective subscore (P = .014) compared with nonhigh responders. Additional patient characteristics can be found in Tables 1 and 2.

Clustering

After K-means clustering optimization, 2 distinct clusters (cluster 1: n = 79; cluster 2: n = 72) were found (Table 3). As expected, there were significant differences between the clusters. Cluster 1 included patients who were younger (51.5 ± 11.8 vs 58.5 ± 11.2, P < .001), had shorter pain duration (43.8 ± 17.5 vs 52.3 ± 13.8, P = .002), and had higher baseline NRS (7.9 ± 1.3 vs 5.8 ± 1.4, P < .001) and higher PCS total scores (32.0 ± 9.7 vs 14.1 ± 18.7, P < .001) compared with cluster 2. In addition, patients in cluster 1 had higher BDI scores (17.0 ± 9.6 vs 9.6 ± 6.6, P < .001), higher ODI scores (27.5 ± 6.3 vs 22.6 ± 7.4, P < .001), and higher rates of CRPS (24.1% vs 6.9%, P = .008). Notably, both clusters had similar rates of responders (36.7% in cluster 1 and 45.8% in cluster 2) and high responders (17.7% in cluster 1 and 23.6% in cluster 2) (Table 3).

Internally validated performances of the ML predictive models for responders are summarized in Table 4 and Figure 2. When all 31 features were used to predict the responders in cluster 1, best performance was obtained with the LR model with an AUC of 0.757, a sensitivity of 61.7%, a specificity of 80%, and an accuracy of 73.4%. When the features were downsized to the 10 most important features (see **Table S2**, **Supplemental Digital Content**, http://links.lww.com/NEU/B419), overall performance remained high with an AUC of 0.757, whereas the sensitivity decreased to 50%. Responders in cluster 2 were best predicted by the LR model using the 10 most important features (**Table S2**, **Supplemental Digital Content**, http://links.lww.com/NEU/B419) with an AUC of 0.708, a sensitivity of 63.3%, a specificity 61.7%, and an accuracy of 62%. The combination of the separate performances of the LR models based on the 10 most important features in the 2 clusters showed higher performance than that of the model based on the entire cohort (AUC: 0.732 vs 0.653, respectively). The performance of both random forest and XGBoost models on the entire cohort was higher than that performed on individual clusters or the combination of the clusters' separate performances (AUC: 0.706 and 0.655, respectively) (see Table 4).

Internally validated performances of the ML predictive models for high responders are summarized in Table 5. Similarly, best model performance to predict high responders in each cluster was obtained with the LR model using the 10 most important features (AUC 0.729 in cluster 1 and AUC 0.647 in cluster 2). LR using

TABLE 4. Performance Comparison of Predictive Models: Responders

Algorithms	Clusters	AUC	Sensitivity	Specificity	PPV	NPV	Accuracy
Logistic regression							
All features	1	0.757 (0.213)	0.617 (0.193)	0.800 (0.133)	0.658 (0.202)	0.790 (0.095)	0.734 (0.110)
	2	0.608 (0.297)	0.592 (0.382)	0.642 (0.171)	0.517 (0.285)	0.725 (0.224)	0.621 (0.171)
	Combination	0.682 (0.262)	0.604 (0.294)	0.721 (0.169)	0.587 (0.251)	0.757 (0.170)	0.677 (0.151)
	Whole cohort	0.638 (0.144)	0.400 (0.154)	0.760 (0.177)	0.570 (0.209)	0.642 (0.110)	0.615 (0.131)
10 features	1	0.757 (0.232)	0.500 (0.360)	0.900 (0.141)	0.658 (0.390)	0.783 (0.125)	0.759 (0.110)
	2	0.708 (0.233)	0.633 (0.343)	0.617 (0.172)	0.550 (0.270)	0.710 (0.236)	0.620 (0.207)
	Combination	0.732 (0.227)	0.566 (0.349)	0.758 (0.211)	0.604 (0.331)	0.746 (0.187)	0.689 (0.176)
	Whole cohort	0.653 (0.146)	0.371 (0.276)	0.778 (0.128)	0.422 (0.243)	0.654 (0.085)	0.609 (0.076)
Random forest							
All features	1	0.710 (0.233)	0.367 (0.331)	0.900 (0.141)	0.583 (0.466)	0.723 (0.131)	0.707 (0.171)
	2	0.550 (0.220)	0.433 (0.288)	0.750 (0.204)	0.517 (0.309)	0.620 (0.167)	0.600 (0.197)
	Combination	0.630 (0.235)	0.400 (0.303)	0.825 (0.187)	0.550 (0.386)	0.671 (0.155)	0.683 (0.181)
	Whole cohort	0.706 (0.192)	0.431 (0.267)	0.794 (0.139)	0.571 (0.230)	0.677 (0.126)	0.648 (0.142)
10 features	1	0.550 (0.291)	0.233 (0.274)	0.760 (0.310)	0.370 (0.462)	0.607 (0.190)	0.570 (0.244)
	2	0.500 (0.297)	0.183 (0.254)	0.750 (0.264)	0.333 (0.471)	0.515 (0.150)	0.487 (0.217)
	Combination	0.525 (0.287)	0.208 (0.258)	0.755 (0.280)	0.351 (0.454)	0.561 (0.173)	0.528 (0.228)
	Whole cohort	0.697 (0.098)	0.443 (0.249)	0.796 (0.142)	0.548 (0.214)	0.686 (0.062)	0.655 (0.062)
XGBoost							
All features	1	0.657 (0.185)	0.550 (0.273)	0.760 (0.207)	0.607 (0.330)	0.746 (0.137)	0.684 (0.118)
	2	0.477 (0.272)	0.433 (0.235)	0.550 (0.307)	0.492 (0.287)	0.540 (0.220)	0.407 (0.195)
	Combination	0.567 (0.244)	0.491 (0.255)	0.655 (0.276)	0.549 (0.306)	0.643 (0.207)	0.545 (0.211)
	Whole cohort	0.655 (0.150)	0.476 (0.223)	0.662 (0.139)	0.488 (0.152)	0.650 (0.138)	0.583 (0.136)
10 features	1	0.607 (0.343)	0.433 (0.316)	0.760 (0.207)	0.525 (0.389)	0.704 (0.148)	0.643 (0.201)
	2	0.423 (0.230)	0.392 (0.125)	0.550 (0.258)	0.457 (0.216)	0.497 (0.134)	0.475 (0.145)
	Combination	0.515 (0.299)	0.412 (0.234)	0.655 (0.251)	0.491 (0.308)	0.600 (0.173)	0.559 (0.191)
	Whole cohort	0.652 (0.065)	0.486 (0.206)	0.708 (0.106)	0.530 (0.049)	0.675 (0.086	0.616 (0.058)

AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

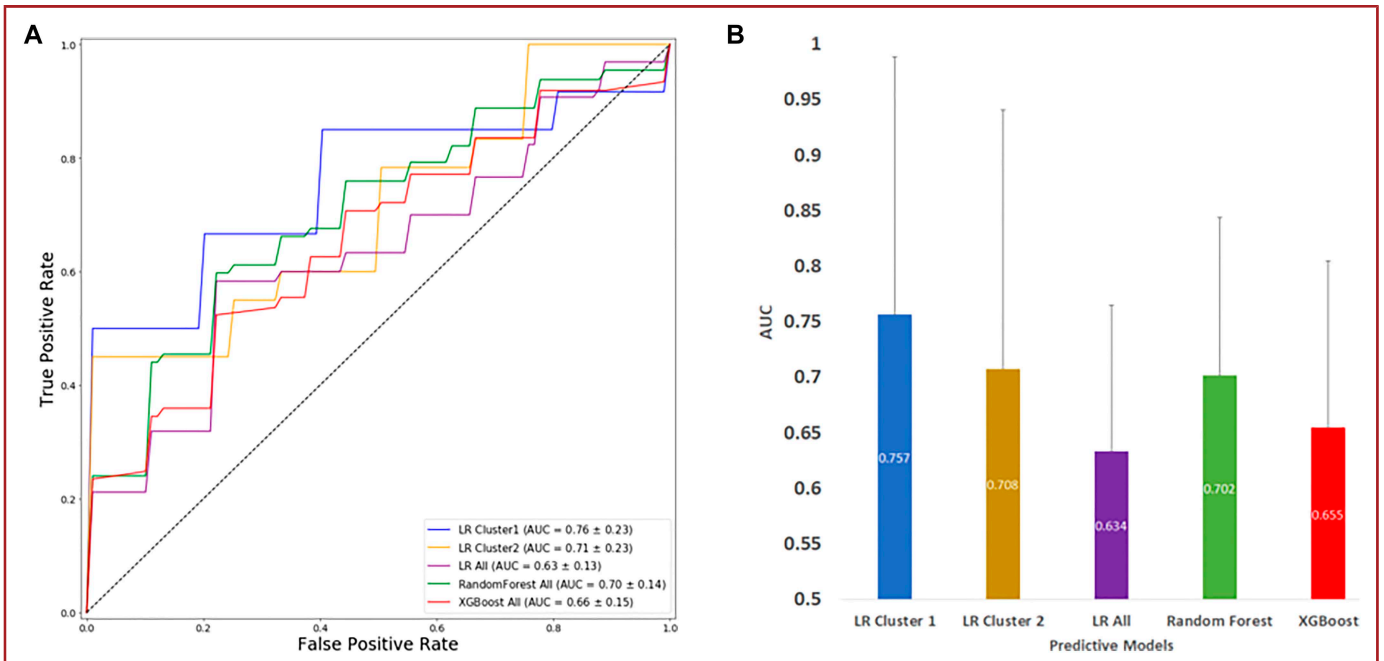

FIGURE 2. *AUC comparison for responders' predictive models:* **A**, *receiver operating curves comparison for LR on cluster 1; LR on cluster 2; and LR, random forest, XGBoost on the entire cohort and* **B**, *Comparison of models' mean AUCs. AUC, area under the curve; LR, logistic regression.*

their combined performance with 10 features showed an AUC of 0.688, a sensitivity of 57.5%, a specificity of 79.6%, and an accuracy of 74.2%. These were higher than those of the LR model when used on the entire cohort.

DISCUSSION

Key Results

This study demonstrated, for the first time, the ability of ML-derived algorithms to predict long-term patient response to SCS placement with relatively high performance (0.708–0.757 AUC for prediction of responders; 0.647–0.729 AUC for prediction of high responders). In addition, the study demonstrated how the combination of unsupervised and supervised learning can develop patient-individualized models based on predicted clusters to increase overall predictive performance (0.757 and 0.708 for the clusters and 0.706 for the entire cohort).[15] To the best of our knowledge, this is the first time that the combined approach was applied in the neuromodulation field.

Although there are rare examples of ML in the SCS literature, our methodology addressed challenges encountered in other studies. De Jaeger et al[16] developed a prediction model to identify participants with low back pain and leg pain who would respond to high-dose SCS using a classification and regression trees (CART) model. Although the predictive performance was higher (AUC of 0.79 for low back pain and 0.80 for leg pain), our models included all types of implanted SCS and waveforms and were internally validated using the nested CV method. Internal validation is critical to avoid overfitting of the data and low performance when applied on an external cohort. Similarly, Goudman et al[17] applied several ML algorithms to predict high frequency-SCS long-term SCS response, yet none achieved high accuracy or overall predictive performance above 58.33%. In addition, by splitting their 119 patients into 80% and 20% segments 1 time (of many possible 20% splits), the model was prone to both over- and underfitting. Moreover, Goudman et al[17] used retrospective outcome data with a wider range of time points (10 months up to 3 years postsurgery) that may detract from the external validity.

Interpretation

Although we provided sensitivity, specificity, and accuracy statistics (similarly to previous studies), these measures can be problematic because they depend on a diagnostic criterion for positivity, which is often chosen arbitrarily. In this study, we used 0.5 as the standard threshold with accuracy results at 62% to 73.4%, which are significantly higher than the clinical accuracy of 40.5% response rate. The area under the curve (AUC) of a receiver operating characteristic curve circumvented this arbitrary threshold and provided a more effective method to evaluate predictive performance between different models.[18] Thus, our models provided a relatively high overall performance of 0.64 to 0.76. Moreover, the models reported the probabilities for a responder/high responder and, ultimately, would allow the clinician to decide on the threshold. Our ultimate goal is to

TABLE 5. Performance Comparison of Predictive Models: High Responders

Algorithms	Clusters	AUC	Sensitivity	Specificity	PPV	NPV	Accuracy
Logistic regression							
All features	1	0.737 (0.218)	0.450 (0.438)	0.826 (0.157)	0.350 (0.388)	0.867 (0.108)	0.761 (0.170)
	2	0.593 (0.320)	0.450 (0.438)	0.737 (0.247)	0.400 (0.394)	0.787 (0.178)	0.655 (0.252)
	Combination	0.665 (0.276)	0.450 (0.426)	0.781 (0.206)	0.375 (0.381)	0.827 (0.149)	0.708 (0.468)
	Whole cohort	0.683 (0.212)	0.467 (0.358)	0.783 (0.181)	0.363 (0.303)	0.854 (0.095)	0.716 (0.154
10 features	1	0.729 (0.212)	0.550 (0.438)	0.805 (0.139)	0.358 (0.314)	0.907 (0.081)	0.761 (0.090)
	2	0.647 (0.351)	0.600 (0.394)	0.787 (0.129)	0.442 (0.299)	0.857 (0.135)	0.723 (0.139)
	Combination	0.688 (0.285)	0.575 (0.406)	0.796 (0.130)	0.400 (0.301)	0.882 (0.111)	0.742 (0.115)
	Whole cohort	0.653 (0.155)	0.467 (0.322)	0.733 (0.146)	0.336 (0.293)	0.845 (0.089)	0.676 (0.124)
Random forest							
All features	1	0.727 (0.212)	0.300 (0.422)	0.907 (0.081)	0.300 (0.422)	0.858 (0.091)	0.796 (0.108)
	2	0.537 (0.336)	0.200 (0.350)	0.843 (0.125)	0.133 (0.219)	0.780 (0.091)	0.682 (0.054)
	Combination	0.632 (0.290)	0.250 (0.380)	0.875 (0.107)	0.216 (0.338)	0.819 (0.097)	0.739 (0.101)
	Whole cohort	0.622 (0.203)	0.300 (0.246)	0.892 (0.097)	0.350 (0.326)	0.832 (0.056)	0.768 (0.090)
10 features	1	0.683 (0.218)	0.400 (0.459)	0.893 (0.101)	0.350 (0.412)	0.868 (0.107)	0.798 (0.133)
	2	0.547 (0.249)	0.250 (0.354)	0.747 (0.161)	0.158 (0.217)	0.773 (0.097)	0.623 (0.100)
	Combination	0.615 (0.238)	0.325 (0.406)	0.820 (0.151)	0.254 (0.335)	0.820 (0.110)	0.710 (0.145)
	Whole cohort	0.568 (0.141)	0.292 (0.246)	0.900 (0.102)	0.382 (0.371)	0.833 (0.050)	0.775 (0.082)
XGBoost							
All features	1	0.717 (0.251)	0.350 (0.412)	0.829 (0.119)	0.267 (0.335)	0.857 (0.088)	0.746 (0.118)
	2	0.475 (0.202)	0.250 (0.354)	0.750 (0.107)	0.183 (0.242)	0.770 (0.103)	0.627 (0.105)
	Combination	0.596 (0.254)	0.300 (0.374)	0.789 (0.117)	0.225 (0.287)	0.813 (0.103)	0.686 (0.124)
	Whole cohort	0.613 (0.234)	0.333 (0.314)	0.850 (0.135)	0.382 (0.379)	0.871 (0.079)	0.742 (0.134)
10 features	1	0.687 (0.324)	0.500 (0.471)	0.800 (0.143)	0.308 (0.329)	0.870 (0.125)	0.736 (0.159)
	2	0.475 (0.202)	0.250 (0.354)	0.750 (0.107)	0.183 (0.242)	0.770 (0.103)	0.627 (0.105
	Combination	0.581 (0.284)	0.375 (0.425)	0.775 (0.125)	0.245 (0.288)	0.820 (0.122)	0.681 (0.142)
	Whole cohort	0.628 (0.175)	0.267 (0.211)	0.858 (0.131)	0.337 (0.373)	0.818 (0.058)	0.736 (0.117)

AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

create a clinical decision support tool, which could augment clinical decision-making.

Using the unsupervised approach as a first stage, we found 2 distinct clusters based on patients' age, pain duration, baseline NRS, and baseline PCS total scores. All of these scores have been previously associated with SCS outcomes, yet they were never clustered using the ML techniques described here.[19] We suggest that these clusters represent 2 distinct SCS populations as 2 different phenotypes: younger patients with higher pain scores who have been suffering for a shorter duration and an older population with a longer chronic pain duration with lower pain scores.

Through hyperparameter fine-tuning and supervised intrinsic feature selection, we identified the 10 most influential features that contribute the most to the model performance. Importantly, these top features are expected to vary between methods because of different algorithms used. Several of these features, including the presence of depression, number of previous spinal surgeries, body mass index, insurance type, and smoking status, have been documented as predictors of poor response in the literature.[20-26] The congruency of these selected features with characteristics identified in previous studies substantiates the validity of our ML-derived models. Moreover, identification of these features in our model can help guide preoperative optimization by addressing these modifiable patient factors to increase the chance of clinical success. Ultimately, these factors likely represent confounders that complicate the underlying pathophysiology, and processing of chronic pain through mechanism research is yet to be fully elucidated. Notably, the study's objective was to enable long-term SCS prediction rather than identifying new presurgical parameters.

The responders and high-responder rates were 40.5% and 20.5%, respectively, which are considerably lower than company-sponsored clinical trial results.[27-30] These findings may be explained by several factors. First, our data represent real-world data in clinical practice, which is significantly different from a study setting. Second, although we used the insurance-based definition of response, NRS usefulness is limited and may create significant results bias.[31]

Limitations

This study had several limitations. First, although we used 1 of the largest prospectively collected data sets, a cohort of 151 patients is considered a small sample size for predictive models. Because of the small sample size, splitting the data into train, test, and validation sets would be highly biased because any split can

create different results (eg, choosing a random 20% test set, compared with another 20% random set). Therefore, we have used the nested CV method. Although a nested CV technique was applied in an effort to overcome this limitation and enable generalization of the results, both prospective and external validations of our models remain a necessary next step. In addition, we aim to create a multicenter registry because currently, multicenter registries in SCS do not usually contain most of the parameters that we have collected. Second, mainly demographics and subjective outcome metrics scores were used as model features. The incorporation of additional objective features (eg, imaging factors, activity level, pedometric data, chronic medications, and stimulation waveforms) may improve overall predictive performance. Third, our data had missing values, which had to be imputed and thus might have attributed to overfitting of data. However, these imputed values were rare (Table 1) and were ultimately processed within the nested CV loop to avoid data leaking and overfitting.

CONCLUSION

The combined unsupervised–supervised ML approach yielded relatively high predictive performance for long-term SCS outcomes in chronic pain patients. ML models of SCS response may be integrated to clinical routine and used to augment, not replace, clinical judgement. Our study suggests that the advanced ML-derived approaches have potential to be used as a functional clinical tool to improve SCS outcomes. Further studies are needed for optimization and external validation of these models.

Funding

This study did not receive any funding or financial support.

Disclosures

The authors have no personal, financial, or institutional interest in any of the drugs, materials, or devices described in this article. Dr Pilitsis is a consultant for Boston Scientific, Nevro, TerSera, and Abbott and receives grant support from Medtronic, Boston Scientific, Abbott, Nevro, TerSera, NIH 2R01CA166379-06, and NIH U44NS115111. She is a medical advisor for Aim Medical Robotics and Karuna and has stock equity. Dr Hadanny has stock equity in Aviv Scientific and EEG Sense. Dr Telkes has grant support from NIH/NINDS K99NS119672 and NIH U44NS115111. Dr Sukul is a consultant for Boston Scientific.

REFERENCES

1. Leung N, Tsourmas NF, Yuspeh L, et al. Increased spinal cord stimulator use and continued opioid treatment among injured workers: a regional pilot study. *J Occup Environ Med.* 2020;62(8):e436.
2. Brinzeu A, Cuny E, Fontaine D, et al. Spinal cord stimulation for chronic refractory pain: long-term effectiveness and safety data from a multicentre registry. *Eur J Pain.* 2019;23(5):1031-1044.
3. Nissen M, Ikäheimo TM, Huttunen J, Leinonen V, von Und Zu Fraunberg M. Long-term outcome of spinal cord stimulation in failed back surgery syndrome: 20 years of experience with 224 consecutive patients. *Neurosurgery.* 2019;84(5):1011-1018.
4. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *J Clin Epidemiol.* 2016;71:76-85.
5. Prabhala T, Sabourin S, DiMarzio M, Gillogly M, Prusik J, Pilitsis JG. Duloxetine improves spinal cord stimulation outcomes for chronic pain. *Neuromodulation.* 2019;22(2):215-218.
6. Slyer J, Scott S, Sheldon B, Hancu M, Bridger C, Pilitsis JG. Less pain relief, more depression, and female sex correlate with spinal cord stimulation explants. *Neuromodulation.* 2020;23(5):673-679.
7. Sheldon BL, Khazen O, Feustel PJ, et al. Correlations between family history of psychiatric illnesses and outcomes of spinal cord stimulation. *Neuromodulation.* 2020;23(5):667-672.
8. Khan H, Pilitsis JG, Prusik J, Smith H, McCallum SE. Pain remission at one-year follow-up with spinal cord stimulation. *Neuromodulation.* 2018;21(1):101-105.
9. Williamson A, Hoggart B. Pain: a review of three commonly used pain rating scales. *J Clin Nurs.* 2005;14(7):798-804.
10. Moore RA, Moore OA, Derry S, Peloso PM, Gammaitoni AR, Wang H. Responder analysis for pain relief and numbers needed to treat in a meta-analysis of etoricoxib osteoarthritis trials: bridging a gap between clinical trials an d clinical practice. *Ann Rheum Dis.* 2010;69(2):374-379.
11. Levitt J, Edhi MM, Thorpe RV, et al. Pain phenotypes classified by machine learning using electroencephalography features. *NeuroImage.* 2020;223:117256.
12. Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recognition Lett.* 2010;31(8):651-666.
13. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS One* 2019;14(11):e0224365.
14. Higgins JP, Thomos J, Chandler J, et al., editors. *Cochrane Handbook for Systematic Reviews of Interventions Version 6.1 (Updated September 2020)*. Cochrane; 2020. Available at www.training.cochrane.org/handbook.
15. Elbattah M, Molloy O. Clustering-aided approach for predicting patient outcomes with application to elderly healthcare in Ireland. In: Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, 2017.
16. De Jaeger M, Goudman L, Brouns R, et al. The long-term response to high-dose spinal cord stimulation in patients with failed back surgery syndrome after conversion from standard spinal cord stimulation: an effectiveness and prediction study. *Neuromodulation.* 2020;24(3):546-555.
17. Goudman L, Van Buyten JP, De Smedt A, et al. Predicting the response of high frequency spinal cord stimulation in patients with failed back surgery syndrome: a retrospective study with machine learning techniques. *J Clin Med.* 2020;9(12):4131.
18. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med.* 2013;4(2):627-635.
19. Pilitsis JG, Fahey M, Custozzo A, Chakravarthy K, Capobianco R. Composite score is a better reflection of patient response to chronic pain therapy compared with pain intensity alone. *Neuromodulation.* 2020;24(1):68-75.
20. Marola O, Cherala R, Prusik J, et al. BMI as a predictor of spinal cord stimulation success in chronic pain patients. *Neuromodulation.* 2017;20(3):269-273.
21. Sparkes E, Duarte RV, Mann S, Lawrence TR, Raphael JH. Analysis of psychological characteristics impacting spinal cord stimulation treatment outcomes: a prospective assessment. *Pain Phys.* 2015;18(3):E369-E377.
22. Patel SK, Gozal YM, Saleh MS, Gibson JL, Karsy M, Mandybur GT. Spinal cord stimulation failure: evaluation of factors underlying hardware explantation. *J Neurosurg Spine.* Published online ahead of print October 4, 2019. doi: 10.3171/2019.6.SPINE181099.
23. Sparkes E, Raphael JH, Duarte RV, LeMarchand K, Jackson C, Ashford RL. A systematic literature review of psychological characteristics as determinants of outcome for spinal cord stimulation therapy. *Pain.* 2010;150(2):284-289.
24. Campbell CM, Jamison RN, Edwards RR. Psychological screening/phenotyping as predictors for spinal cord stimulation. *Curr Pain Headache Rep.* 2013;17(1):307.
25. Mekhail N, Azer G, Saweris Y, Mehanny DS, Costandi S, Mao G. The impact of tobacco cigarette smoking on spinal cord stimulation effectiveness in chronic spine-related pain patients. *Reg Anesth Pain Med.* 2018;43(7):768-775.
26. Turner JA, Hollingworth W, Comstock BA, Deyo RA. Spinal cord stimulation for failed back surgery syndrome: outcomes in a workers' compensation setting. *Pain.* 2010;148(1):14-25.
27. Amirdelfan K, Yu C, Doust MW, et al. Long-term quality of life improvement for chronic intractable back and leg pain patients using spinal cord stimulation: 12-month results from the SENZA-RCT. *Qual Life Res.* 2018;27(8):2035-2044.

28. Levy R, Deer TR, Poree L, et al. Multicenter, randomized, double-blind study protocol using human spinal cord recording comparing safety, efficacy, and neurophysiological responses between patients being treated with evoked compound action potential-controlled closed-loop spinal cord stimulation or open-loop spinal cord stimulation (the evoke study). *Neuromodulation.* 2019;22(3):317-326.
29. Deer T, Slavin KV, Amirdelfan K, et al. Success using neuromodulation with BURST (SUNBURST) study: results from a prospective, randomized controlled trial using a novel burst waveform. *Neuromodulation.* 2018;21(1):56-66.
30. Fishman MA, Calodney A, Kim P, et al. Prospective, multicenter feasibility study to evaluate differential target multiplexed spinal cord stimulation programming in subjects with chronic intractable back pain with or without leg pain. *Pain Pract.* 2020;20(7):761-768.
31. Safikhani S, Gries KS, Trudeau JJ, et al. Response scale selection in adult pain measures: results from a literature review. *J Patient Rep Outcomes.* 2017;2:40.

Acknowledgments

We would like to thank Dr Matthew Shapiro and his laboratory team for providing computational resources.

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

Supplemental Digital Content. Methods, Figure, Table. The Supplemental Digital Content expands on the Methods provided. **Figure S1.** Features correlation heatmap. **Figure S2.** Unsupervised clustering with K-means results. **Figure S3.** Combined unsupervised and supervised approach scheme. **Table S1.** Model hyperparameters. **Table S2.** Selected features for responders' models.