ELSEVIER

The Spine Journal 19 (2019) 1657−1665

THE SPINE JOURNAL

Clinical Study

Development of a machine learning algorithm for prediction of failure of nonoperative management in spinal epidural abscess

Akash A. Shah, MD[a,*], Aditya V. Karhade, BS[b], Christopher M. Bono, MD[b], Mitchel B. Harris, MD[b], Sandra B. Nelson, MD[c], Joseph H. Schwab, MD, MD[b]

[a] *Department of Orthopaedic Surgery, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA*
[b] *Department of Orthopaedic Surgery, Massachusetts General Hospital, Boston, MA, USA*
[c] *Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, USA*

Received 24 February 2019; revised 15 April 2019; accepted 30 April 2019

Abstract

BACKGROUND CONTEXT: Data regarding risk of failure of nonoperative management in spinal epidural abscess (SEA) are limited. Given the potential for deterioration with treatment failure, a tool that predicts the probability of failure would be of great clinical utility.

PURPOSE: We primarily aim to build a machine learning model using independent predictors of nonoperative management failure. Secondarily, we aim to develop an open-access web-based application that provides a patient-specific probability of treatment failure.

STUDY DESIGN/SETTING: Retrospective, case-control study.

PATIENT SAMPLE: Patients 18 years or older diagnosed with SEA at 2 academic medical centers and 3 community hospitals.

OUTCOME MEASURES: Failure of nonoperative management.

METHODS: This is a retrospective cohort study of 367 patients with SEA initially managed nonoperatively between 1993 and 2016. The primary outcome was failure of nonoperative management defined as neurologic deterioration, worsened back and/or radicular pain, or persistent symptoms despite initiation of antibiotic therapy. Five machine learning algorithms were developed and assessed by discrimination, calibration, and overall performance.

RESULTS: Ninety-nine (27%) patients failed nonoperative management. Factors determined for prediction of nonoperative management were: motor deficit, diabetes, ventral component of abscess relative to thecal sac, history of compression or pathologic vertebral fracture, sensory deficit, active malignancy, and involvement of 3 or more vertebral levels. The elastic-net penalized logistic regression model was chosen as the final model given its superior discrimination, calibration, and overall model performance. This model was incorporated into an open access web application.

CONCLUSION: By building a discriminative and well-calibrated model in a user-friendly and open-access digital interface, we hope to provide a prognostic tool that can be used to inform clinical decision-making in real-time.

Keywords: Algorithms; Machine learning; Nonoperative management; Spinal epidural abscess; Infection; Outcomes; Application

FDA device/drug status: Not applicable.

Author disclosure: ***AS:*** Nothing to disclose; ***AK:*** Nothing to disclose; ***CB:*** Royalties: Wolters Kluwer (A), Elsevier (B); Consulting: United Health Care (B); Other Office: The Spine Journal (F); Fellowship Support: OMEGA (D, Paid directly to institution/employer).; ***MH:*** Nothing to disclose; ***SB:*** Nothing to disclose; ***JS:*** Scientific Advisory Board: Chordoma Foundation (None); Speaking and/or Teaching Arrangements: AO Spine (Travel Expense Reimbursement, Outside 12-Month Requirement), Stryker Spine (B, Outside 12-Month Requirement).

* Corresponding author. Department of Orthopaedic Surgery, David Geffen School of Medicine at UCLA, 10833 Le Conte Avenue, Los Angeles, CA 90095, USA. Tel.: +17143318707.

E-mail address: aashah@mednet.ucla.edu (A.A. Shah).

https://doi.org/10.1016/j.spinee.2019.04.022
1529-9430/© 2019 Elsevier Inc. All rights reserved.

Introduction

The variable clinical progression of spinal epidural abscess (SEA) and its associated risk of precipitous neurologic deterioration make it a challenging entity to manage.[1,2] Due likely to an aging population, increased number of spinal procedures, increased prevalence of diabetes mellitus, and intravenous drug use, as well as increased clinical suspicion with subsequent imaging, the incidence of SEA has increased in recent decades.[1−6] Although surgical decompression continues to be a common treatment of SEA, medical management has seen increased popularity.[3] Data regarding which patients will respond favorably to medical management remain limited; this is an important consideration given the potential for poor neurologic outcomes observed in those who fail nonoperative management.[7] Recent studies have identified many patient factors associated with treatment failure.[7−9] These studies have largely employed multivariable logistic regression and do not provide open access tools for healthcare professionals to predict outcomes.

Machine learning represents a set of techniques that allow machines to learn and make predictions by recognizing patterns in data. Unlike conventional regression methods, machine learning allows for detection of complex nonlinear relationships as well as multivariate effects.[10,11] This can be contrasted with traditional algorithms that are programmed with a desired behavior.[11] As Beam and Kohane explain, it may be useful to view a predictive algorithm as lying on a spectrum between fully human-guided and fully machine-guided, where an algorithm is considered more machine-guided as fewer human assumptions are placed on it.[12] By relying only on data to build nonlinear algorithms with the greatest predictive capability, machine learning methods have been shown to not only accurately predict outcomes or detect disease but also to outperform logistic regression in select cases.[10,13−15] Although machine learning methods are increasingly used in other medical disciplines, they have not yet been applied for outcomes in musculoskeletal infections.

In the current study, we primarily aim to build a machine-learning model using independent predictors of failure of nonoperative management. Secondarily, we aim to employ this predictive algorithm in an open-access web-based application that provides a patient-specific probability of management failure.

Methods

Study design and subjects

Our institutional review board approved a waiver of consent for this retrospective study. We included patients who were 18 years or older diagnosed with SEA by magnetic resonance imaging or computed tomography (CT) admitted to our health care system. We excluded patients who were initially treated operatively.

We identified our cohort by performing a computer query search of all patients admitted to our hospital system of 2 tertiary academic medical centers and 3 regional community hospitals between 1993 and 2016 for International Classification of Diseases, Ninth and Tenth Revisions (ICD-9, ICD-10) codes for SEA and synonyms (ICD-9 324.1 and ICD-10 G06.1). We also performed a search for Current Procedural Terminology codes for "laminectomy for excision or evacuation of intraspinal lesion other than neoplasm, extradural" for the cervical, thoracic, lumbar, and sacral spine (current procedural terminology 63275 to 63278). This initial search strategy yielded 2,756 unique patients, of which 1,053 patients were 18 years or older and began definitive treatment for SEA in our system. We also ensured that included patients had appropriate magnetic resonance imaging or CT reports that confirm presence of SEA.

Of these 1,053 potentially eligible patients, 472 were initially treated nonoperatively. The initial treatment modality was determined by the primary attending physician. Nonoperative management is defined as systemic antibiotic therapy with or without CT-assisted percutaneous drainage. Treatment groups were defined by the intention of the treating team; a patient was considered to have been treated nonoperatively if the primary spine service—or the consulting spine service if the primary team was not neurosurgery or orthopaedic surgery—initially elected for nonoperative management. We excluded patients who were treated nonoperatively for palliation or because they were too ill to undergo a surgical procedure.

Finally, we excluded patients without documented treatment failure if they had less than 60 days of follow-up since initiation of treatment. We did so to avoid definitively labeling patients as having been successfully treated nonoperatively without adequate follow-up. If patients had follow-up of <60 days but already had been identified as a treatment failure, they were included. This yielded 367 patients (Fig. 1).

Outcome and explanatory variables

The primary outcome measure was failure of nonoperative management. Failure was defined as neurologic deterioration, worsened back and/or radicular pain, or persistent symptoms despite initiation of antibiotic therapy that led to a change in management (eg, prolonged/altered antibiotic course, CT-guided drainage, or surgical management). If progression on serial imaging led to an alteration in treatment by the primary team, this was also considered failure.

Variables collected for the patients included demographics, signs and symptoms, duration of symptoms, laboratory values, microbiology, radiographic characteristics, as well as concurrent spinal and nonspinal infections (Table 1). Motor status was determined using the American Spinal Injury Association Scale.[16] We define sensory changes to include frank sensory deficit and subjective

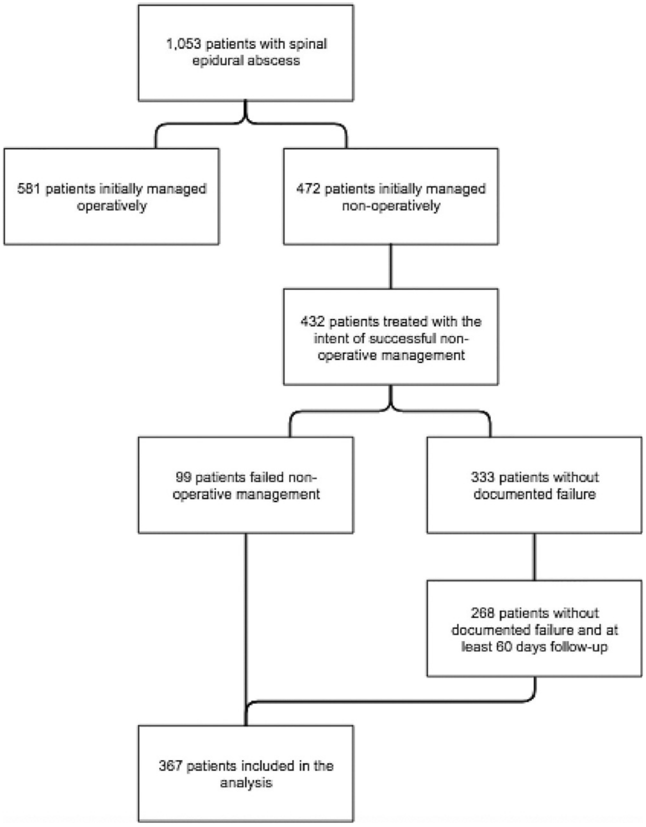
1,053 patients with spinal epidural abscess
581 patients initially managed operatively
472 patients initially managed non-operatively
432 patients treated with the intent of successful non-operative management
99 patients failed non-operative management
333 patients without documented failure
268 patients without documented failure and at least 60 days follow-up
367 patients included in the analysis

Fig. 1. Flowchart detailing the inclusion and exclusion criteria that led to the formation of our final study population. SEA = spinal epidural abscess.

paresthesias. The patient was considered to have motor or nonmotor neurologic dysfunction only if these were new symptoms at presentation.

Statistical analysis

The population was divided into a derivation and validation cohort by using a stratified 80:20 split. The derivation (training) cohort was used for recursive feature selection to determine the variables used for algorithm development. Random forest algorithms with 10-fold cross validation were used for the feature selection. Explanatory variables are listed in Table 1; factors that have more than 30% missing data were excluded from feature selection. Variables determined from the feature selection step were then used to develop 5 machine learning algorithms: Elastic-Net Penalized Logistic Regression, Stochastic Gradient Boosting, Random Forest, Support Vector Machine, and Neural Network. These models were chosen on the basis of prior machine learning studies.[17,18] The algorithms were trained on the derivation set and tested with 10-fold cross validation repeated 3 times.[19] In this method, the data was divided into 10 equally sized folds. The models were trained on 9 of the 10 folds (90% of the data) and tested on the remaining fold (10%). This sub-process was repeated 10 times to test on each fold while training on the remainder. Finally, the

Table 1
Baseline characteristics of study population n = 367

Variable	n (%) | median (IQR)
Age (years)	59.0 (49.0−70.5)
Female sex	130 (35.4)
History	
Intravenous drug use	72 (19.6)
Alcohol use	20 (5.4)
Smoking status*	
Never	185 (51.1)
Current	99 (27.3)
Quit >1 y	78 (21.5)
Diabetes mellitus	52 (14.2)
Active malignancy	82 (22.3)
Hemodialysis	25 (6.8)
HIV positive	12 (3.3)
Spinal instrumentation in place	20 (5.4)
Spinal procedure in past year	60 (16.3)
History of pathologic or compression fracture	17 (4.6)
Symptoms and signs	
Fever	83 (22.6)
Back pain	353 (96.2)
Motor deficit	54 (14.8)
Sensory changes	43 (11.7)
Symptom duration prior to presentation	
<=72 h	70 (19.1)
72 h - 2 wk	155 (42.2)
>2 wk	142 (38.7)
Laboratory values*	
Hemoglobin (g/dL)	10.7 (9.7−11.9)
WBC (10^3/uL)	10.4 (7.6−14.1)
Platelet (10^3/uL)	281.5 (196.0−389.2)
Absolute lymphocyte (10^3/uL)	1.26 (0.90−1.66)
Absolute neutrophil (10^3/uL)	6.54 (4.67−9.29)
Absolute eosinophils (10^3/uL)	0.10 (0.04−0.19)
Absolute basophils (10^3/uL)	0.020 (0.010−0.040)
Absolute monophils (10^3/uL)	0.59 (0.42−0.85)
Neutrophil to lymphocyte ratio	5.35 (3.24−8.42)
Platelet to lymphocyte ratio	217.8 (150.0−319.3)
Erythrocyte sedimentation rate (mm/h)	87.0 (53.0−106.0)
C-reactive protein (mg/dL)	100.4 (31.8−163.9)
Albumin (g/dL)	3.10 (2.60−3.50)
Alkaline phosphatase (IU/L)	101.0 (75.2−141.0)
AST (IU/L)	24.0 (16.8−41.0)
ALT (IU/L)	23.0 (14.0−37.0)
Bilirubin, total (mg/dL)	0.50 (0.40−0.80)
Creatinine (mg/dL)	0.87 (0.70−1.10)
Blood urea nitrogen (mg/dL)	16.0 (10.0−23.0)
Calcium (mg/dL)	8.70 (8.30−9.10)
Prothrombin time (s)	14.3 (13.5−15.6)
INR	1.10 (1.10−1.30)
Partial thromboplastin time (seconds)	32.2 (28.3−37.8)
Three or more affected levels	162 (44.3)
Spine location	
Cervical	43 (12.1)
Thoracic	105 (29.7)
Lumbosacral	241 (68.1)
Location of abscess relative to thecal sac*	
Anterior	243 (66.8)
Posterior	59 (16.2)
Circumferential	26 (7.1)
Multiple locations	36 (9.9)
Organism	
No growth	84 (22.9)
Methicillin-sensitive Staph aureus	124 (33.8)

Table 1 (**Continued**)

Variable	n (%) | median (IQR)
Methicillin-resistant Staph aureus	38 (10.4)
Streptococcus	39 (10.6)
Bacteremia	213 (58.0)
Local spinal infections	
Spondylodiscitis	212 (57.8)
Psoas/paraspinal abscesses	186 (50.7)
Vertebral osteomyelitis	50 (13.6)
Prevertebral abscess/retropharyngeal abscess	29 (7.9)
Discitis	20 (5.4)
Wound infection	15 (4.1)
Local nonspinal infections	
Endocarditis	23 (6.3)
Nonspinal abscess cellulitis	20 (5.4)
Septic arthritis	18 (4.9)
Pneumonia/empyema	13 (3.5)
Failure of nonoperative management	99 (27.0)

* Rates of missing data were: Smoking status = 5(1.6%), Hemoglobin = 9 (2.45%), Platelet = 11 (3.0%), Absolute lymphocyte = 57 (15.5%), Absolute neutrophil = 51 (13.9%), Absolute eosinophil = 59 (16.1%), Absolute basophil = 63 (17.2%), Absolute monophil = 56 (15.3%), Neutrophil to lymphocyte ratio = 57 (15.5%), Platelet to lymphocyte ratio = 58 (15.8%), Erythrocyte sedimentation rate = 50 (13.6%), C-reactive protein = 119 (32.4%), Albumin = 75 (20.4%), Alkaline phosphatase = 77 (21.0%), AST = 79 (21.5%), ALT = 95 (25.9%), Total bilirubin = 77 (21.0%), Creatinine = 9 (2.5%), Blood urea nitrogen = 8 (2.2%), Calcium = 22 (6.0%), Prothrombin time = 67 (18.3%), INR = 82 (22.3%), Partial thromboplastin time = 83 (22.6%), location of abscess = 3 (0.8%).

overall process was repeated 3 times. Model performance was examined by discrimination (receiver-operating curve, c-statistic), calibration (calibration plot, calibration slope, calibration intercept), overall model performance (Brier score), and decision curve analysis.

Discrimination refers to the model's ability to distinguish patients who failed nonoperative management from those who did not.[20−23] Discrimination was assessed graphically with the receiver-operating curve (ROC) and numerically with the area under the receiver-operating curve (AUC). Calibration measures how well the model's predicted probabilities correlate to the observed probabilities in the study population; it was assessed graphically with calibration plots and numerically with calibration slope and calibration intercept.[22,23] Overall model performance was assessed with the Brier score, the mean squared error between the observed values and the predicted probabilities.[23,24] The calculated Brier score was compared with the null model Brier score. The null model Brier score was calculated by assigning a predicted probability for all patients equivalent to the rate of nonoperative management failure in the study population. Decision curve analysis is a method based on the concept of net benefit. Net benefit is defined as a function of true positives, false positives, the relative weight assigned to true positives versus false positives, and the overall sample size. By plotting the net benefit for all probability thresholds, the result of any change in management on the basis of the model can be compared with the default results of not changing management for any patients or for changing management for all patients.[25] The final models were tested on the validation cohort and assessed again by discrimination, calibration, and overall performance.

The predictions of the algorithm with the best performance across these metrics was explained globally and locally. Averaging across all patients in the derivation cohort, the importance of each variable included in the model was demonstrated globally to show their relative importance.[26] Next, at the individual patient level, local explanations were provided to demonstrate which factors supported and contradicted the prediction of failure of nonoperative management for each patient.[27]

Application development

The best algorithm was incorporated into an interactive interface and deployed as an open-access web application able to provide both predictions and patient-specific explanations. R version 3.5.0 (The R Foundation, Vienna, Austria), RStudio version 1.0.153 (RStudio, Boston, MA, USA), and Python version 3.6 (Python Software Foundation, Wilmington, Delaware) were used for data analysis and model creation and deployment.

Results

Demographic characteristics

Of the 367 patients in the cohort, 99 patients (27%) failed nonoperative management. Fifty-four patients (15%) had a motor deficit at presentation, and 43 patients (12%) had sensory changes. With respect to medical comorbidities, 82 patients (22%) had diabetes mellitus and 25 (6.8%) had an active malignancy at the time of presentation. Seventeen patients (4.6%) had a pathologic/compression fracture at the affected levels. Two hundred and forty-three patients (67%) had abscesses located ventral to the thecal sac, 59 patients (16%) had abscesses located dorsally, and 26 patients (7.1%) had abscesses that circumferentially surrounded the thecal sac. Thirty-six patients' (9.9%) abscesses had components located in multiple locations relative to the thecal sac (Table 1). The median follow-up was 36 weeks.

Model performance and application

Using random forest algorithms with 10-fold cross validation in the derivation cohort, the following variables were identified to include in final algorithm development: motor deficit, diabetes mellitus, ventral component of abscess relative to thecal sac, history of compression or pathologic vertebral fracture, sensory dysfunction, active malignancy, and 3 or more spine levels (Fig. 2B).

Discrimination refers to the model's ability to distinguish between patients who failed and did not fail nonoperative management[20−23]; models with perfect discrimination have a c-statistic of 1. Calibration is a measure of how well

Table 2
Machine learning model performance on cross-validation of training set, mean (95% confidence interval), n = 295

Performance metric	Elastic-net penalized logistic regression	Neural network	Random forest	Stochastic gradient boosting	Support vector machine
C-statistic	0.80 (0.76, 0.83)	0.78 (0.75, 0.82)	0.75 (0.72, 0.79)	0.79 (0.75, 0.82)	0.72 (0.66, 0.76)
Calibration intercept	0.16 (−0.09, 0.42)	0.01 (−0.20, 0.22)	0.92 (0.23, 1.61)	0.17 (−0.09, 0.42)	1.10 (−0.58, 2.78)
Calibration slope	1.19 (0.99, 1.40)	1.05 (0.87, 1.23)	1.35 (0.00, 2.71)	1.15 (0.95, 1.35)	1.86 (0.69, 3.03)
Brier score	0.15 (0.14, 0.16)	0.16 (0.15, 0.17)	0.18 (0.17, 0.19)	0.15 (0.14, 0.17)	0.17 (0.16, 0.18)
Null model brier score	0.20				

the model's predicted probabilities compare to observed probabilities in the study population. Calibration slope measures the difference between predictor effects for the model in training and testing sets; a calibration slope of 1 indicates that the predictor effects for the model are equivalent in both sets. The Brier score is used to assess overall model performance.[23,24] The mean squared error between observed values and predicted probabilities, Brier scores closer to 0 indicate better models since this indicates a lower error between predicted and observed values.

The AUC for the 4 machine-learning models ranged from 0.56 for the Random Forest to 0.79 for the Elastic-Net Penalized Logistic Regression in the testing set. The calibration slope ranged from 0.08 for the Random Forest to 1.21 for the Elastic-Net Penalized Logistic Regression. The Brier score ranged from 0.14 for the Elastic-Net Penalized Logistic Regression to 0.18 for the Random Forest. The Null Model Brier score was 0.20 (Table 3). Assessed numerically by discrimination alone, the best model for predicting failure of nonoperative management was the Elastic-Net Penalized Logistic Regression. The receiver operating curve for the Elastic-Net Penalized Logistic Regression model is shown in Fig. 2A. The Elastic-Net Penalized Logistic Regression was best calibrated over the full range of predicted probabilities (Fig. 3A). Furthermore, the Elastic-Net Penalized Logistic Regression had the lowest Brier Score of the models. With superior discrimination, calibration, and overall model performance, the Elastic-Net Penalized Logistic Regression resulted in greater net benefit than the default strategies of changing management for all patients or for no patients for thresholds greater than 0.13 (Fig. 3B). We calculate positive predictive value, negative predictive value, and accuracy at 3 threshold values (Table 4).

The Elastic-Net Penalized Logistic Regression model was used to build a predictive algorithm for failure of nonoperative management. This was developed into a web application that is available as an open access tool for clinicians. The web application can be accessed at: https://sorg-apps.shinyapps.io/seanonop/ (Fig. 4).

Discussion

A better understanding of factors associated with nonoperative management failure in SEA would be of great utility. Nonoperative management emerged as a viable treatment strategy for SEA in recent decades, with several reports of successful medical management.[28−34] General characteristics of patients in whom nonoperative management have been proposed: a normal neurological exam, extensive panspinal infection, complete paralysis for >72 hours, poor surgical candidacy, or refusal of surgery.[1,4,28] Limited by relatively low case numbers in the SEA literature, independent predictors of unsuccessful medical management have been difficult to identify until recently.

With a cohort of 51 patients who underwent nonoperative management, Patel et al. identified 4 risk factors predictive of treatment failure: diabetes mellitus, leukocytosis greater than 12.5, positive blood cultures, and C-reactive protein greater than 115.7 With a cohort of 142 nonoperatively managed patients, Kim et al. also identified 4 independent predictors of failure: age greater than 65 years, diabetes, methicillin-resistant *Staphylococcus aureus*, and pretreatment motor deficit. They also provide a simplified algorithm for probability of treatment failure.[8]

Table 3
Machine learning model performance on testing set, n = 72

Performance metric	Elastic-net penalized logistic regression	Neural network	Random forest	Stochastic gradient boosting	Support vector machine
C-statistic	0.79	0.78	0.56	0.78	0.72
Calibration intercept	0.07	−0.13	−0.01	0.31	−0.17
Calibration slope	1.21	0.96	0.08	1.16	0.87
Brier score	0.14	0.15	0.18	0.14	0.17
Null model brier score	0.20				

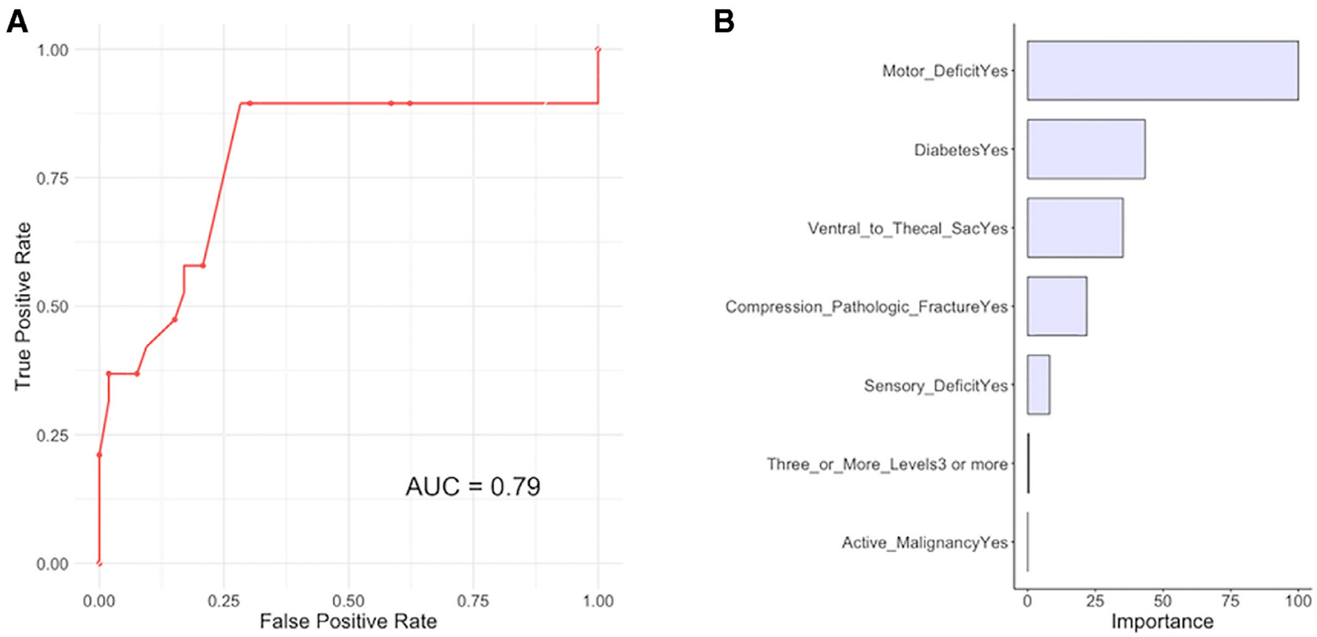

Fig. 2. (A) Receiver operating curve for elastic-net penalized logistic regression; (B) Variable importance plots for elastic-net penalized logistic regression.

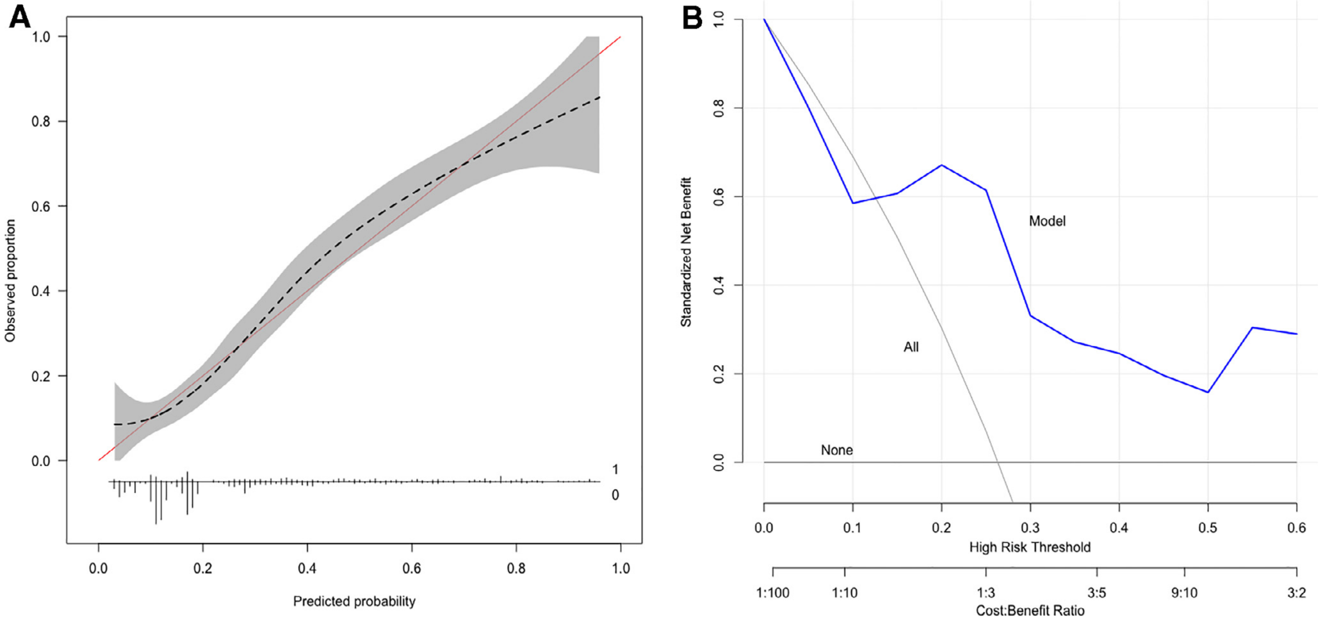

Fig. 3. The Elastic-Net Penalized Logistic Regression model was A) well calibrated over the full range of predicted probabilities and B) resulted in greater net benefit than default strategies of changing management for all patients or no patients for thresholds greater than 0.13.

Shah et al. built on these studies with a cohort of 367 nonoperatively managed patients, identifying 6 independent predictors of failure.[9] Motor deficit at presentation, sensory changes, diabetes, active malignancy, and pathologic/compression fracture in affected levels were positive predictors, whereas dorsal location of abscess relative to the thecal sac was a negative predictor. These factors encapsulate neurologic status at the time of presentation, medical comorbidities, as well as local abscess anatomy. Furthermore, they developed an algorithm that provides a patient-specific risk of failure based on the presence or absence of these 6 risk factors.[9]

Machine learning technology enables machines to learn and make predictions through recognition of patterns in large datasets.[11] By capturing complex, nonlinear relationships in data sets, machine learning represents an opportunity for improving the accuracy of predicting clinical outcomes.[10] Machine learning algorithms have been employed to aid decision-making in dermatology, ophthalmology, and oncology.[13−15] Recently, machine learning

Table 4
Positive predictive value, negative predictive value, and accuracy at three different threshold values

Threshold	Positive predictive value	Negative predictive value	Accuracy
0.25	0.53	0.93	0.76
0.50	0.70	0.81	0.79
0.75	1.00	0.78	0.79

algorithms have been developed to predict surgical site infections as well as survival in patients with metastatic bone disease.[35−37] Machine learning techniques have not been applied to SEA thus far. Previously performed studies identifying independent predictors have all employed regression modeling.[7−9] Accurately predicting failure of nonoperative management would be of great utility given the risk of clinical deterioration and protracted time of antibiosis incurred through treatment failure.[7]

With a cohort of 367 patients with SEA managed nonoperatively, we have developed a machine learning algorithm for prediction of failure of nonoperative management. Assessment of discrimination, calibration, and overall model performance led to the selection of the Elastic-Net Penalized Logistic Regression model. By including the largest cohort of nonoperatively managed patients with SEA over a 26-year period across multiple tertiary care and community hospitals, we are hopeful that our findings are generalizable to a larger population.

Developing a discriminative and calibrated model for predicting failure of nonoperative management is of

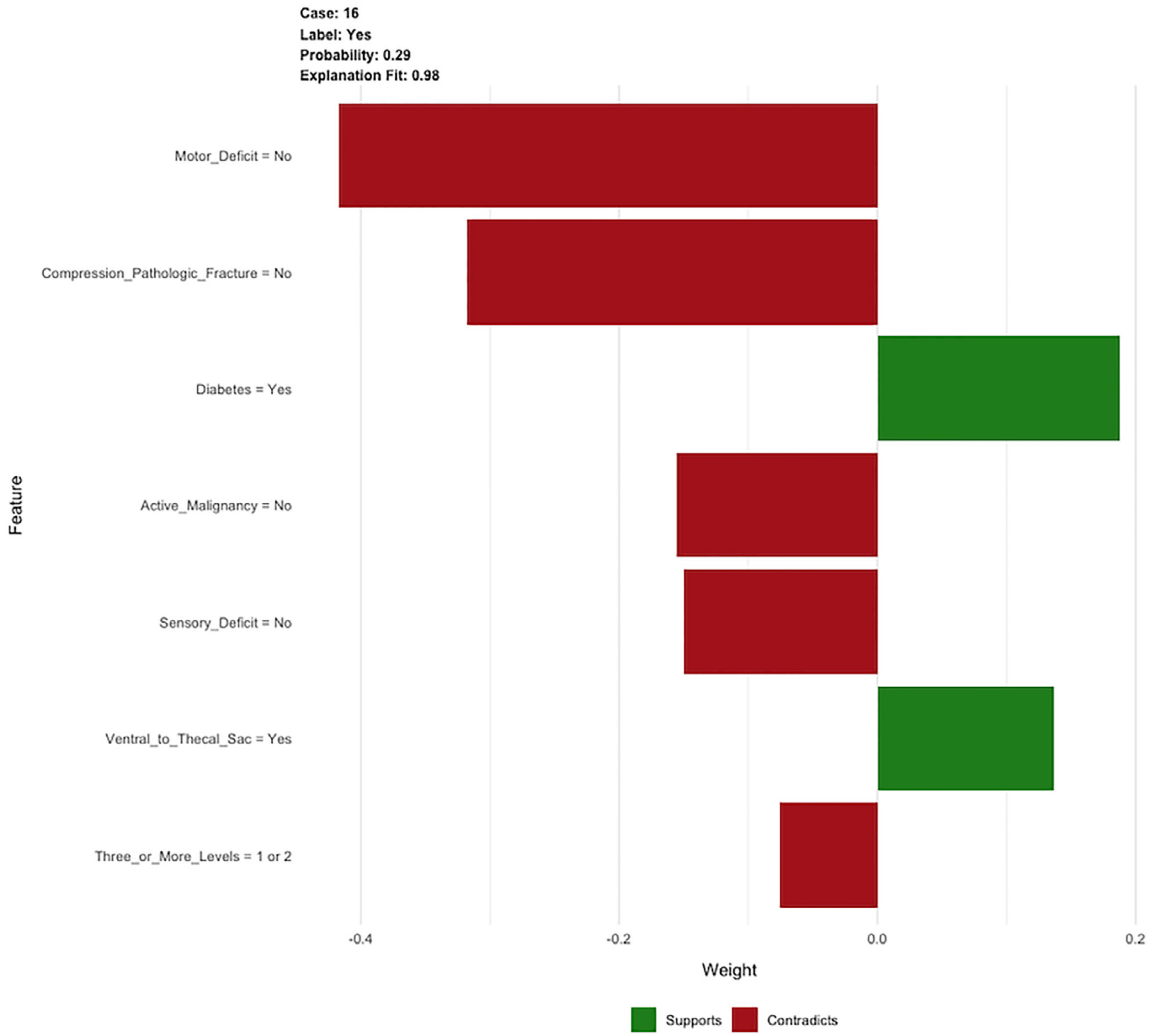

Fig. 4. Screenshot of the web-based application for a sample patient with diabetes and SEA spanning three or more levels that is ventral to the thecal sac. This patient has a 29% probability of failing nonoperative management.

prognostic value and can be used to inform decision-making in SEA. The clinical utility of the algorithm hinges not just on the accuracy of the model but also on the ease of use for clinicians. Previous studies reporting predictive algorithms for nonoperative management failure have provided charts of the outcome's probability listed for all conceivable scenarios or the algorithm itself.[8,9] Yet, charts or equations are not as straightforward to apply in real-time as digital applications are. To facilitate direct use of this algorithm by healthcare providers, we incorporated the Elastic-Net Penalized Logistic Regression model into an open access digital interface. It must be emphasized that this algorithm is not meant to supplant the role of clinical judgment. The decision to manage SEA operatively versus nonoperatively is a complex one and must take into consideration the broader clinical and psychosocial context of the patient. Although, an algorithm cannot make the determination of whether the dangers of failed nonoperative management outweigh the benefits of avoiding surgery, an algorithm that provides an accurate risk of failure can provide an important data point for the clinician.

This study has limitations, first of which is its retrospective design. Second, radiologic images were not always available for review in our electronic medical record before 2007; thus location relative to thecal sac in these cases was determined and classified solely from radiology reports. Furthermore, the decision for nonoperative management was not made with any clearly defined criteria; rather, it was a decision made by the attending spine surgeon. This may represent a source of selection bias. In any predictive model, there is a concern for model overfitting. Including more variables with a small sample size can lead to overfitting. In overfitting, the apparent performance of the algorithm on the development set improves but the generalizability of the algorithm to new samples decreases because the algorithm focuses on the idiosyncrasies of the development dataset rather than learning general principles. It is thus important to perform future studies where this algorithm is externally validated.

The utility of machine-learning models lies in their predictive nature, not in an explanatory capacity. Unlike in logistic regression, the relative contributions of different risk factors toward failure of nonoperative management are not offered as odds ratios. Nonetheless, improving the accuracy of clinical predictions is of great utility so long as it is used thoughtfully to translate into better clinical care. It should be noted that the patient population used in this study overlaps with the cohort reported in Shah et al.[9]; however, we employ a novel methodology to predict treatment failure and provide a web-application interface for increased accessibility.

Finally, it is important to recognize that an algorithm − whether it is machine learning or regression − is only as trustworthy as the data it is built on. Systematic biases in data collection and clinical decisions impact the patterns that are detected by machine learning; this can adversely affect underrepresented groups such as women, ethnic minorities, and patients of lower socioeconomic status.[11] Clinicians must be cognizant that algorithms can reflect past biases. External validation of machine learning models is also important to guard against institutional biases in a single-system study. Future studies can seek to validate or refute the models created in this analysis by using data from multiple institutions or with prospective study designs.

Conclusion

Using the largest cohort of nonoperatively managed patients with SEA, we have built a robust machine learning algorithm to predict failure of nonoperative management. An accurate model is necessary but not sufficient to achieve true clinical utility, however. We also report an open-access web-based application that simplifies use of the algorithm for clinicians, the first such tool for prediction of nonoperative management failure in SEA.

References

[1] Rigamonti D, Liem L, Sampath P, Knoller N, Numaguchi Y, Schreibman DL, et al. Spinal epidural abscess: contemporary trends in etiology, evaluation, and management. Surg Neurol 1999;52(2):189–97 http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed4&NEWS=N&AN=1999268148.

[2] Darouiche R. Spinal epidural abscess. N Engl J Med 2006;355(19):2012–20. https://doi.org/10.1056/NEJMra055111.

[3] Arko L, Quach E, Nguyen V, Chang D, Sukul V, Kim B-S. Medical and surgical management of spinal epidural abscess: a systematic review. Neurosurg Focus 2014;37(2):E4. https://doi.org/10.3171/2014.6.FOCUS14127.

[4] Reihsaus E, Waldbaur H, Seeling W. Spinal epidural abscess: a meta-analysis of 915 patients. Neurosurg Rev 2000;23(4):175–204 discussion 205. https://doi.org/10.1007/PL00011954.

[5] Adogwa O, Karikari IO, Car KR, Krucoff M, Ajay D, Fatemi P, et al. Spontaneous spinal epidural abscess in patients 50 years of age and older: a 15-year institutional perspective and review of the literature. J Neurosurg Spine 2014;20(3):344–9. https://doi.org/10.3171/2013.11.SPINE13527.

[6] Vakili M, Crum-Cianflone NF. Spinal epidural abscess: a series of 101 cases. Am J Med 2017;130(12):1458–63. https://doi.org/10.1016/j.amjmed.2017.07.017.

[7] Patel A, Alton T, Bransford R, Lee M, Bellabarba C, Chapman J. Spinal epidural abscesses: risk factors, medical versus surgical management, a retrospective review of 128 cases. Spine J 2014;14(2):326–30. https://doi.org/10.1016/j.spinee.2013.10.046.

[8] Kim S, Melikian R, Ju KL, Zurakowski D, Wood KB, Bono CM, et al. Independent predictors of failure of nonoperative management of spinal epidural abscesses. Spine J 2014;14(8):1673–9. https://doi.org/10.1016/j.spinee.2013.10.011.

[9] Shah AA, Ogink PT, Nelson SB, Harris MB, Schwab JH. Nonoperative management of spinal epidural abscess: development of a predictive algorithm for failure. J Bone Jt Surgery Am Vol 2018;100(7):546–55.

[10] Chen J, Asch S. Machine learning and prediction in medicine - beyond the peak of inflated expectations. N Engl J Med 2017;376(26):2507–9.

[11] Hashimoto D, Rosman G, Rus D, Meireles O. Artifical intelligence in surgery: promises and perils. Ann Surg 2018;268(1):70–6.

[12] Beam A, Kohane I. Big data and machine learning in health care. JAMA 2018;319(13):1317–8.

[13] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7660):686.

[14] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc 2016;316(22):2402–10.

[15] Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS One 2013;8 (4):e61318.

[16] Kirshblum SC, Burns SP, Biering-Sorensen F, Donovan W, Graves DE, Jha A, et al. International standards for neurological classification of spinal cord injury (Revised 2011). J Spinal Cord Med 2011;34 (6):535–46. https://doi.org/10.1179/107902611X13186000420242.

[17] Wainer J. Comparison of 14 different families of classification algorithms on 115 binary datasets. arXiv 2016:1606. 00930.

[18] Karhade AV, Thio QC, Ogink PT, Shah AA, Bono CM, Oh KS, et al. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. Neurosurgery 2018;0(0). https://doi.org/10.1016/j.wneu.2018.07.276. Epub ahead of print.

[19] Kim JH. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Anal 2009;53(11):3735–45. https://doi.org/10.1016/j.csda.2009.04.009.

[20] Cook N. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115(7):928–35.

[21] Hanley J, McNeil B. The meaning and use of the area under a receiving operating characteristic (ROC) curve. Radiology 1982;143(1):29–36.

[22] Steyerberg E, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35(29):1925–31.

[23] Steyerberg E, Vickers A, Cook N, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 2010;21(1):128–38.

[24] Brier G. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78(1):1–3.

[25] Vickers A, Elkin E. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak 2006;26(6):565–74.

[26] Greenwell B, Boehmke B, McCarthy A. A simple and effective model-based variable importance measure. arXiv 2018:1805. 04755.

[27] Ribeiro M, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv 2016:1606. 05386.

[28] Leys D, Lesoin F, Viaud C, Pasquier F, Rousseaux M, Jomin M, Petit H. Decreased morbidity from acute bacterial spinal epidural abscesses using computed tomography and nonsurgical treatment in selected patients. Ann Neurol 1985;17(4):350–5. https://doi.org/10.1002/ana.410170408.

[29] Wheeler D, Keiser P, Rigamonti D, Keay S. Medical management of spinal epidural abscesses: case report and review. Clin Infect Dis 1992;15(1):22–7.

[30] Mampalam T, Rosegay H, Andrews B, Rosenblum M, Pitts L. Nonoperative treatment of spinal epidural infections. J Neurosurg 1989; 71:208–10.

[31] Tang H, Lin H, Liu Y, Li C. Spinal epidural abscess - experience with 46 patients and evaluation of prognostic factors. J Infect 2002;45 (2):76–81. https://doi.org/10.1053/jinf.2002.1013.

[32] Siddiq F, Chowfin A, Tight R, Sahmoun A, Smego R. Medical vs surgical management of spinal epidural abscess. Arch Intern Med 2004;164(22):2409–12. https://doi.org/10.1001/archinte.164.22.2409.

[33] Curry WT, Hoh BL, Amin-Hanjani S, Eskandar EN. Spinal epidural abscess: clinical presentation, management, and outcome. Surg Neurol 2005;63(4):364–71. https://doi.org/10.1016/j.surneu.2004.08.081.

[34] Savage K, Holtom P, Zalavras C. Spinal epidural abscess: early clinical outcome in patients treated medically. Clin Orthop Relat Res 2005; 439:56–60. https://doi.org/10.1097/01.blo.0000183089.37768.2d.

[35] Paulino Pereira NR, Janssen SJ, van Dijk E, Harris MB, Hornicek FJ, Ferrone ML, et al. Development of a prognostic survival algorithm for patients with metastatic spine disease. J Bone Jt Surgery Am Vol 2016;98(21):1767–76.

[36] Forsberg J, Wedin R, Boland P, Healey J. Can we estimate short- and intermediate-term survival in patients undergoing surgery for metastatic bone disease? Clin Orthop Relat Res 2017;475 (4):1252–61.

[37] Soguero-Ruiz C, Wang F, Jenssen R, et al. Data-driven temporal prediction of surgical site infection. AMIA Annu Symp Proc 1989: 1164–73.