The Spine Journal 19 (2019) 1950–1959

Clinical Study

Development of machine learning algorithms for prediction of mortality in spinal epidural abscess

Aditya V. Karhade, BE[a], Akash A. Shah, MD[b], Christopher M. Bono, MD[a], Marco L. Ferrone, MD[c], Sandra B. Nelson, MD[d], Andrew J. Schoenfeld, MD, MSc[c], Mitchel B. Harris, MD[a], Joseph H. Schwab, MD, MS[a,*]

[a] *Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115, USA*
[b] *Department of Orthopaedic Surgery, University of California, Los Angeles, CA 90095, USA*
[c] *Department of Orthopaedic Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA*
[d] *Department of Infectious Diseases, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115, USA*
Received 30 March 2019; revised 23 June 2019; accepted 26 June 2019

Abstract

BACKGROUND CONTEXT: In-hospital and short-term mortality in patients with spinal epidural abscess (SEA) remains unacceptably high despite diagnostic and therapeutic advancements. Forecasting this potentially avoidable consequence at the time of admission could improve patient management and counseling. Few studies exist to meet this need, and none have explored methodologies such as machine learning.

PURPOSE: The purpose of this study was to develop machine learning algorithms for prediction of in-hospital and 90-day postdischarge mortality in SEA.

STUDY DESIGN/SETTING: Retrospective, case-control study at two academic medical centers and three community hospitals from 1993 to 2016.

PATIENTS SAMPLE: Adult patients with an inpatient admission for radiologically confirmed diagnosis of SEA.

OUTCOME MEASURES: In-hospital and 90-day postdischarge mortality.

METHODS: Five machine learning algorithms (elastic-net penalized logistic regression, random forest, stochastic gradient boosting, neural network, and support vector machine) were developed and assessed by discrimination, calibration, overall performance, and decision curve analysis.

RESULTS: Overall, 1,053 SEA patients were identified in the study, with 134 (12.7%) experiencing in-hospital or 90-day postdischarge mortality. The stochastic gradient boosting model achieved the best performance across discrimination, c-statistic=0.89, calibration, and decision curve analysis. The variables used for prediction of 90-day mortality, ranked by importance, were age, albumin, platelet count, neutrophil to lymphocyte ratio, hemodialysis, active malignancy, and diabetes. The final algorithm was incorporated into a web application available here: https://sorg-apps.shinyapps.io/seamortality/.

CONCLUSIONS: Machine learning algorithms show promise on internal validation for prediction of 90-day mortality in SEA. Future studies are needed to externally validate these algorithms in independent populations.

Keywords: Artificial intelligence; Healthcare; Machine learning; Mortality; Spinal epidural abscess; Spine surgery

FDA device/drug status: Not applicable.

Author disclosures: ***AVK:*** Nothing to disclose. ***AAS:*** Nothing to disclose. ***CMB:*** Royalties: Wolters Kluwer (A), Elsevier (B); Consulting: United Health Care (B); Other Office: The Spine Journal (D); Fellowship Support: OMEGA (D, Paid directly to institution/employer). ***MLF:*** Nothing to disclose. ***SBN:*** Nothing to disclose. ***AJS:*** Royalties: Wolters Kluwer (B), Springer (A); Scientific Advisory Board/Other Office: JBJS (C); Research Support (Investigator Salary, Staff/Materials): CMS-OMH (F); Grants: DoD (H), OREF (F), NIH-NIAMS (G). ***MBH:*** Nothing to disclose. ***JHS:*** Speaking and/or Teaching Arrangements: Stryker (B).

Ethics Statement: This retrospective study was approved by our institutional review board.

* Corresponding author. Department of Orthopedic Surgery, Massachusetts General Hospital, Associate Professor, Harvard Medical School, 55 Fruit St, Boston, MA 02114, USA. Tel.: 617-543-5227; fax: 617-726-7587.
E-mail address: jhschwab@mgh.harvard.edu (J.H. Schwab).

https://doi.org/10.1016/j.spinee.2019.06.024

Introduction

In-hospital and short-term mortality in spinal epidural abscess (SEA) remain unacceptably high despite diagnostic and therapeutic advancements in medical and surgical care [1–8] Identification of a subset of SEA patients at high risk of short-term mortality at the time of admission could help inform treatment decisions and aid in expectation management. Changes in the type and intensity of in-hospital management and alterations in postdischarge surveillance for high-risk patients may reduce the potential for short-term mortality.

Few studies exist to meet this need, and none have explored methodologies such as machine learning [4,9,10]. Machine learning is a subfield of computer science and statistics that has emerged as a growing tool in health care [11–13]. Previous studies in degenerative spine disease, spinal oncology, and spinal deformity have illustrated successful applications of this methodology within spine surgery [14–19]. Despite these trends, machine learning remains relatively underutilized in SEA.

The purpose of this study was to develop machine learning algorithms for prediction of in-hospital and 90-day postdischarge mortality in SEA. Additional aims of this study were to deploy these models as digital applications in order to increase the accessibility of machine learning for stakeholders in SEA. Finally, this study sought to increase the interpretability of the resulting machine learning models by providing global explanations as well as patient-centered explanations integrated into the digital application regarding survival after treatment of SEA.

Methods

Guidelines

The Multivariable Prediction Model for Individual Prognosis or Diagnosis guidelines and the Guidelines for Developing and Reporting Machine Learning Models in Biomedical Research were followed for this study [20,21].

Source of data

Our institutional review board approved retrospective review of electronic medical records. Chart review was used to identify patients admitted with a diagnosis of SEA at two academic centers and three community hospitals. The study was limited to chart review only and individual patient consent was waived by the institutional review board.

Participants

Inclusion criteria were as follows: (1) adult patients, 18 years or older, (2) initial presentation for a diagnosis of SEA confirmed by computed tomography or magnetic resonance imaging, and (3) inpatient admission between January 1, 1993 and December 31, 2016. Exclusion criteria for the study included (1) patients who presented to our institutions after beginning management for SEA elsewhere (2) patients who did not have SEA confirmed by imaging.

Outcome

The primary outcome of interest was mortality in-hospital or within 90 days of discharge. Follow-up was ascertained by cross-referencing manual chart review, structured clinical documentation, and the Social Security Death Index.

Predictors

Variables included in this study were age (years), sex, medical and social history (smoking, alcohol, intravenous drug use, diabetes, hemodialysis, active malignancy, HIV status, and previous spinal instrumentation, spinal procedures, or pathologic fractures), signs and symptoms at presentation [fever, back pain, motor deficit, sensory changes, urinary or fecal incontinence or retention, the American Spinal Injury Association scale (A–D)], symptom duration before presentation, laboratory values [hemoglobin (grams per deciliter, g/dL), white blood cell (thousand count per microliter, $10^3/\mu$L), platelet ($10^3/\mu$L), absolute lymphocyte ($10^3/\mu$L), absolute neutrophil ($10^3/\mu$L), absolute eosinophil ($10^3/\mu$L), absolute basophil ($10^3/\mu$L), absolute monocyte ($10^3/\mu$L), neutrophil to lymphocyte ratio, platelet to lymphocyte ratio, erythrocyte sediment rate (millimeters per hour, mm/h), c-reactive protein (g/dL), albumin (g/dL), alkaline phosphatase (international units per liter, IU/L), aspartate transaminase (IU/L), alanine transaminase (IU/L), total bilirubin (milligrams per deciliter, mg/dL), creatinine (mg/dL), blood urea nitrogen (mg/dL), calcium (mg/dL), prothrombin time (seconds), International Normalized Ratio, partial thromboplastin time (seconds)], imaging characteristics [affected levels, location of abscess relative to thecal sac], microbiologic data on growth of organisms from abscess culture, concurrent infections (bacteremia, other local spine infections, other localized nonspine infections), and initial management modality (operative, nonoperative). Laboratory values were included if they were collected on the day of admission or the day immediately before admission.

Missing data

Variables were included in this analysis if they had less than 30% missing data. Multiple imputation was undertaken with the missForest methodology. The rates of missing data for each variable were smoking status 24 (2.3%), American Spinal Injury Association Impairment Scale impairment scale=2 (0.2%), symptom duration before presentation=2 (0.2%), hemoglobin=25 (2.4%), platelet=29 (2.8%), absolute lymphocyte=182 (17.3%), absolute neutrophil=159 (15.1%), absolute eosinophil=199 (18.9%), absolute basophil=207 (19.7%), absolute monocytes=182 (17.3%), neutrophil to lymphocyte ratio=182 (17.3%), platelet to lymphocyte ratio=186 (17.7%), erythrocyte sediment rate=275 (26.1%), c-reactive protein=483 (45.9%),

albumin=230 (21.8%), alkaline phosphatase=231 (21.9%), aspartate transaminase=232 (22.0%), alanine transaminase=293 (27.8%), total bilirubin=233 (22.1%), creatinine=25 (2.4%), blood urea nitrogen=28 (2.7%), calcium=78 (7.4%), prothrombin time=151 (14.3%), international normalized ratio=206 (19.6%), and partial thromboplastin time=170 (16.1%).

Statistical analysis and methods

A stratified 80:20 split of the SEA patient population was undertaken to create the training and testing sets, respectively [22]. The training set was used for model development and the performance of final models was assessed on the testing set. Detailed technical explanation of the modeling strategies has been extensively described by previous studies [15,16].

In brief, the training set was used for variable selection, model development, and predictive performance assessment. Recursive feature elimination with random forest algorithms (10-fold cross-validation, repeated three times) was used to identify the subset of variables used as inputs to each machine learning model [22]. The following candidate algorithms were selected on the basis of prior work (1) random forest, (2) stochastic gradient boosting (SGB), (3) neural network, (4) support vector machine, and (5) penalized logistic regression (PLR) [23]. Further details of the algorithms and methodology have been provided in the Supplementary Methods. Discrimination, calibration, and decision curve analysis were used to assess each model in the training set [24]. During development, the algorithms were assessed by using the training set for 10-fold cross-validation repeated 3 times.

Discrimination was assessed by plotting the receiver operating curve and by calculating the c-statistic, also known as the area under the receiver operating curve (AUC) for binary classification [25]. Perfect models have c-statistic/AUC=1. Calibration was assessed by plotting calibration plots and calculating calibration slope and intercept. Perfect models have calibration slope=1 and calibration intercept=0. The Brier score was calculated as a combined measure of discrimination and calibration. Perfect models have Brier scores=0. However, the Brier score must be interpreted within the context of the null model Brier score. The null model Brier score is defined as the Brier score of the model that assigns a predicted probability to each patient equal to the observed prevalence of the outcome in the population. The null model Brier score was calculated to compare the relative gain of the algorithms to this benchmark. Decision curve analysis was undertaken to investigate the net benefit (weighted average of true positives and false positives) of the algorithms over the range of predicted probabilities [26].

To overcome the "black box" nature of machine learning models, explanations were provided for the algorithms at the global and local level (eg, patient-centered). Global explanations were defined as model insights applicable over the study population. For example, global explanations through variable importance plots averaged the relative importance of each variable to the algorithm's predictions over the total population. Similarly, partial dependence plots were provided to give greater insight into the relationship between the range of input continuous variables and the model predicted probability outputs [27,28]. Finally, local explanations are provided for individual patients to give providers greater insight into the specific factors influencing the probability of mortality [29].

Finalized models from the training set were then assessed using identical utilities when applied to the sample held out for independent testing. The best model from these series of validations was then deployed as a web application accessible on desktops, tablets, and smartphones (Fig. 1). Software used for analysis included the Anaconda Distribution (Anaconda, Inc, Austin, TX), R version 3.5.1 (The R Foundation, Vienna, Austria), RStudio version 1.0.153 (RStudio, Boston, MA), and Python version 3.6 (Python Software Foundation, Wilmington, DE).

Results

Participants

Overall, 1,053 patients met the inclusion criteria with in-hospital and 90-day mortality rates of 134 (12.7%). The median age was 59 (interquartile range=48–69) years and 408 (38.7%) patients were female. Other baseline characteristics are available for review in Table 1. Overall, 581 (55.2%) underwent initial operative management and ultimately 646 (61.3%) were managed operatively before discharge. By time period of admission, the 90-day mortality for patients admitted in 1999 or before (n=127) was 11 (8.7%). The 90-day mortality for patients admitted from 2000 to 2005 (n=235) was 41 (17.4%). The 90-day mortality for patients admitted from 2006 to 2010 (n=277) was 43 (15.5%). The 90-day mortality for patients admitted 2011 or later was 39 (9.4%; n=414).

Variables used for prediction of 90-day mortality were age, albumin, platelet count, neutrophil to lymphocyte ratio, hemodialysis, active malignancy, and diabetes. On cross-validation of the training set, n=844 (80%), the discrimination (AUC) of the algorithms ranged from 0.69 (support vector machine) to 0.83 (elastic-net PLR and neural network; Table 2. The SGB, random forest, and elastic-net PLR models had the best calibration ranging from calibration intercept=0.04 (SGB) to 0.09 (PLR) and calibration slope=0.94 (random forest) to 1.06 (PLR). The overall performance varied from Brier score=0.09 to 0.10, relative to the null model Brier score of 0.11.

On assessment in the independent testing set, n=209 (20%), the SGB algorithm had the best performance with AUC=0.89, calibration intercept=0.01, calibration slope=1.23, and Brier score=0.08 (Table 3). Variable importance plots for the SGB algorithm showed that the five most important variables were

Fig. 1. Web application interface for the SGB algorithm. SGB, stochastic gradient boosting.

age, albumin, platelet count, neutrophil to lymphocyte ratio, and hemodialysis (Fig. 2). Decision curve analysis of the SGB model showed that decision changes based on the model outperformed not only the default strategies of decision change for all patients or no patients as well as decision changes based on patient age alone (Fig. 3). At a threshold of 0.1, the SGB algorithm had a sensitivity (recall) of 1.00, specificity of 0.67, and positive predictive value (precision) of 0.30 in testing set. At a threshold of 0.2, the SGB algorithm had a sensitivity of 0.73, specificity of 0.80, and positive predictive value of 0.34 in testing set. At a threshold of 0.3, the SGB algorithm had a sensitivity of 0.62, specificity of 0.89, and positive predictive value of 0.44 in testing set. At a threshold of 0.4, the SGB algorithm had a sensitivity of 0.42, specificity of 0.94, and positive predictive value of 0.50 in testing set. At a threshold of 0.5, the SGB algorithm had a sensitivity of 0.15, specificity of 0.96, and positive predictive value of 0.50 in testing set.

Sensitivity analyses were undertaken by training and testing the algorithms on the subset of patients admitted only in the last 10 years of the available cohort (2007–2016). Overall, 635 patients were admitted from 2007 to 2016 with in-hospital or 90-day postdischarge mortality rate of 70 (11%). In the training set, the AUC ranged from 0.61 to 0.84, the calibration intercept ranged from −0.38 to 0.41, the calibration slope ranged from 0.73 to 1.52, and the Brier score ranged from 0.08 to 0.10 (Table 4). In the testing set, the elastic-net PLR achieved the highest AUC=0.89 but had less favorable calibration with intercept=0.86 and slope=1.50. The SGB algorithm resulted in AUC=0.86 and was well calibrated with intercept=0.07 and slope=1.00 (Table 5).

Model explanations

Partial dependence plots were created for the SGB, neural network, and PLR models for examining the relationship between the continuous variables of age, albumin, neutrophil to lymphocyte ratio, platelet count, and the model outputs ($\hat{y}$ or the predicted probability; Fig. 4). The predicted probability of mortality increased proportionately over the full range of albumin as albumin decreased for the PLR model. However, the predicted probability of mortality for the SGB model plateaued at albumin less than 2.1 g/dL, indicating that the algorithm did not increase the expectation of an adverse outcome once albumin decreased beyond this threshold.

Individual patient level explanations were provided for the SGB model (Fig. 5). For example, consider a 72-year-old patient with albumin=2.3, platelet=150, and neutrophil to lymphocyte ratio=14. The model predicted a probability of 0.63 of 90-day mortality for this patient. Factors that increased the likelihood of 90-day mortality were older age, hypoalbuminemia, thrombocytopenia, and systemic inflammation (elevated neutrophil to lymphocyte ratio). However,

Table 1
Baseline characteristics of study population, n=1,053

Variable	n (%)|median (IQR)
Age	59.0 (48.0–69.0)
Female sex	408 (38.7)
History	
Smoking status	
Current	286 (27.8)
Quit >1 year	208 (20.2)
Alcohol use	183 (17.4)
Intravenous drug use	191 (18.1)
Diabetes mellitus	241 (22.9)
Hemodialysis	44 (4.2)
Active malignancy	79 (7.5)
HIV positive	21 (2.0)
Spinal instrumentation in place	70 (6.6)
Spinal procedure in past year	207 (19.7)
Pathologic or compression fracture	56 (5.3)
Signs and symptoms	
Fever	283 (26.9)
Back pain	1,014 (96.3)
Motor deficit	362 (34.4)
Sensory changes	250 (23.7)
Urinary incontinence or retention	113 (10.7)
Fecal incontinence or retention	41 (3.9)
ASIA	
Normal (E)	656 (62.3)
Incomplete injury (B–D)	331 (31.4)
Complete (A)	31 (2.9)
Sedated/existing deficit	33 (3.1)
Symptom duration before presentation	
≤72 h	209 (19.9)
72 h–2 wk	345 (32.8)
>2 wk	497 (47.3)
Laboratory values	
Hemoglobin (g/dL)	10.4 (9.4–11.6)
White blood cell ($10^3/\mu$L)	11.4 (8.2–15.8)
Platelet ($10^3/\mu$L)	285.5 (195.0–390.0)
Absolute lymphocyte ($10^3/\mu$L)	1.21 (0.87–1.65)
Absolute neutrophil ($10^3/\mu$L)	7.90 (5.50–11.81)
Absolute eosinophil (10^3/uL)	0.080 (0.020–0.180)
Absolute basophil ($10^3/\mu$L)	0.020 (0.010–0.040)
Absolute monocyte ($10^3/\mu$L)	0.58 (0.40–0.86)
Neutrophil lymphocyte Ratio	6.73 (3.91–11.16)
Platelet lymphocyte ratio	230.3 (157.1–340.0)
Erythrocyte sediment rate (mm/h)	88.0 (59.0–107.0)
C-Reactive protein (mg/dL)	112.7 (33.4–183.3)
Albumin (g/dL)	2.80 (2.30–3.30)
Alkaline phosphatase (IU/L)	102.0 (76.0–143.8)
Aspartate transaminase (IU/L)	27.0 (17.0–44.0)
Alanine transaminase (IU/L)	25.0 (14.0–42.0)
Total bilirubin (mg/dL)	0.50 (0.40–0.90)
Creatinine (mg/dL)	0.82 (0.66–1.10)
Blood urea nitrogen (mg/dL)	15.0 (10.0–23.0)
Calcium (mg/dL)	8.60 (8.00–9.00)
Prothrombin time (seconds)	14.4 (13.5–15.6)
International normalized ratio	1.20 (1.10–1.30)
Partial thromboplastin time (seconds)	30.5 (27.2–36.1)
Three or more affected levels	587 (55.8)
Spine location*	
Cervical	190 (19.3)
Thoracic	351 (35.6)
Lumbosacral	571 (57.9)
Location of abscess relative to thecal sac*	
Dorsal	464 (44.1)
Ventral	783 (74.4)
Circumferential	80 (7.6)
Organism	
No growth	176 (16.7)
Methicillin-sensitive Staph aureus	421 (40.0)
Methicillin-resistant Staph aureus	151 (14.3)
Streptococcus	104 (9.9)
Coagulase-negative staphylococcal species	68 (6.5)
Escherichia coli	27 (2.6)
Other	106 (10.1)
Bacteremia	614 (58.3)
Local spinal infections	
Spondylodiscitis	483 (45.9)
Psoas/paraspinal abscesses	476 (45.2)
Prevertebral abscess/retropharyngeal abscess	88 (8.4)
Wound infection	67 (6.4)
Local nonspinal infections	
Endocarditis	64 (6.1)
Nonspinal abscess cellulitis	55 (5.2)
Septic arthritis	56 (5.3)
Pneumonia/empyema	57 (5.4)
Meningitis	20 (1.9)
Nonvertebral osteomyelitis	14 (1.3)
Initial treatment modality	
Operative	581 (55.2)
Nonoperative	472 (44.8)
Death in hospital or within 90 days of discharge	134 (12.7)

Table 1 (***Continued***)

* Percent greater than 100 as abscess in certain patients was in multiple locations.ASIA, American Spinal Injury Association Impairment Scale; BMI, body mass index; CRP, c-reactive protein; ESR, erythrocyte sediment rate; (g/dL), grams per deciliter; h, hours; IQR, interquartile range; kg/m^2, kilogram per meter squared; mg/dL, milligrams per deciliter; μL: microliter; WBC, white blood cell.

lack of active malignancy, diabetes, and hemodialysis dependency in this patient reduced the likelihood of 90-day mortality.

Model specification

The SGB model was incorporated into a digital application and made available here: https://sorg-apps.shinyapps.io/seamortality/.

The digital application includes default values. These default values are placeholders that users can modify as per the individual characteristics of the patient under consideration. The SGB model requires complete data for these seven factors in order to generate predictions and explanations for 90-day mortality in SEA.

Discussion

In this study, the SGB model emerged as the best performing algorithm from an analysis of five machine learning models for prediction of 90-day mortality in patients with a diagnosis of SEA. The algorithm performed well on discrimination, calibration, and decision curve analysis. Internal validation of this algorithm and immediate

Table 2
Discrimination and calibration of algorithms on repeated cross-validation of training set, n=844, mean (95% confidence interval)

Metric	Stochastic gradient boosting	Random forest	Support vector machine	Neural network	Elastic-net penalized logistic regression
AUC	0.82 (0.80, 0.84)	0.82 (0.80, 0.85)	0.69 (0.66, 0.72)	0.83 (0.81, 0.85)	0.83 (0.81, 0.85)
Intercept	0.04 (−0.14, 0.23)	0.05 (−0.17, 0.26)	0.66 (−0.37, 1.69)	2.32 (1.23, 3.40)	0.09 (−0.12, 0.29)
Slope	1.04 (0.93, 1.15)	0.94 (0.81, 1.08)	1.34 (0.83, 1.85)	1.46 (1.25, 1.66)	1.06 (0.94, 1.18)
Brier	0.09 (0.09, 0.10)	0.09 (0.09, 0.10)	0.10 (0.10, 0.11)	0.10 (0.10, 0.11)	0.09 (0.09, 0.10)

AUC, area under the receiver operating curve. Null model Brier score=0.11.

Table 3
Discrimination and calibration of algorithms in holdout set, n=209

Metric	Stochastic gradient boosting	Random forest	Support vector machine	Neural network	Elastic-net penalized logistic regression
AUC	0.89	0.83	0.70	0.89	0.89
Intercept	0.01	−0.01	−0.62	2.74	0.16
Slope	1.23	0.97	0.68	1.78	1.30
Brier	0.08	0.09	0.11	0.10	0.08

AUC, area under the receiver operating curve. Null model Brier score=0.11.

availability of a digital platform offer direct avenues for clinical application.

Variables used for prediction of 90-day mortality in this analysis concurred with previous studies. Shah et al. identified age greater than 65 years, diabetes, active malignancy, hemodialysis, pretreatment motor deficit, endocarditis, symptom duration before presentation, and leukocytosis as independent risk factors for prediction of 90-day mortality [4]. Overlapping factors with this study included age and medical comorbidities (active malignancy, hemodialysis, and diabetes). Schoenfeld et al. previously determined that age and renal failure were risk factors for in-hospital mortality for SEA patients in the National Inpatient Sample [1], and like findings along with diabetes and disseminated cancer were highlighted by Du et al. using 30-day mortality figures from the National Surgical Quality Improvement Program [10]. The remaining factors featured in the SGB model (eg, albumin, neutrophil to lymphocyte ratio, and platelet count) are aligned with more recent work that has established hypoalbuminemia, elevated neutrophil to lymphocyte ratio, and thrombocytopenia as independent risk factors for short-term mortality in SEA [10,30,31]. Spinal epidural abscess most commonly occurs in patients with impaired immune function who are otherwise also prone to transient bacteremia or septicemia. This includes elderly patients, individuals with active malignancy, recipients of hemodialysis, and intravenous drug users. The most fulminant cases are likely to present in individuals with impaired physiologic reserve, manifested by hypoalbuminemia and thrombocytopenia. Markers of increased inflammation, such as the neutrophil to lymphocyte ration, may be indicative of the size of the abscess, extent of neurologic

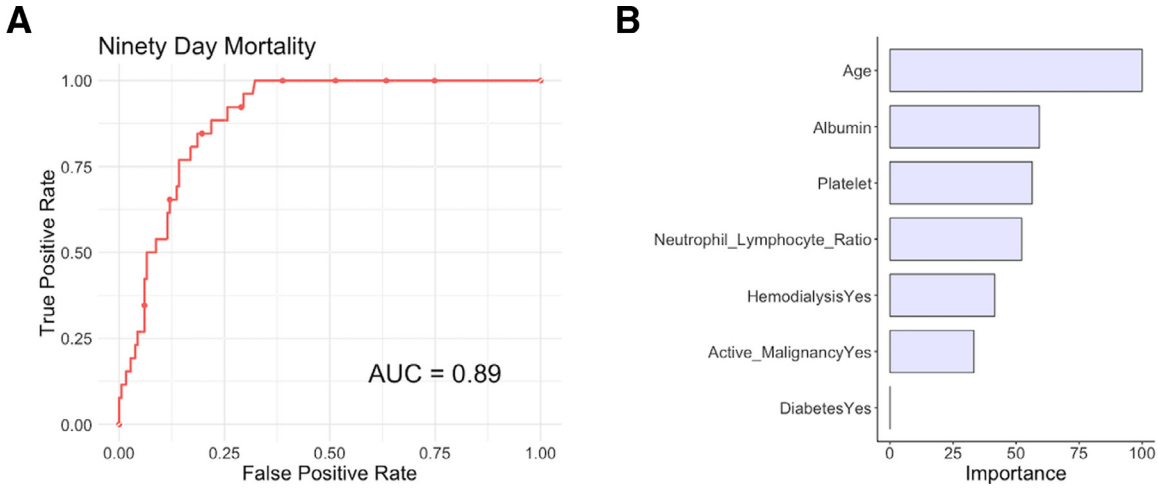

Fig. 2. (A) Receiver operating curve for the SGB algorithm, AUC=0.89, in the testing set, n=209. (B) Variable importance plot for the SGB model. For prediction of 90-day mortality: top 5 predictors: age, albumin, platelet, neutrophil to lymphocyte ratio, and hemodialysis. AUC, area under the receiver operating curve; SGB, stochastic gradient boosting.

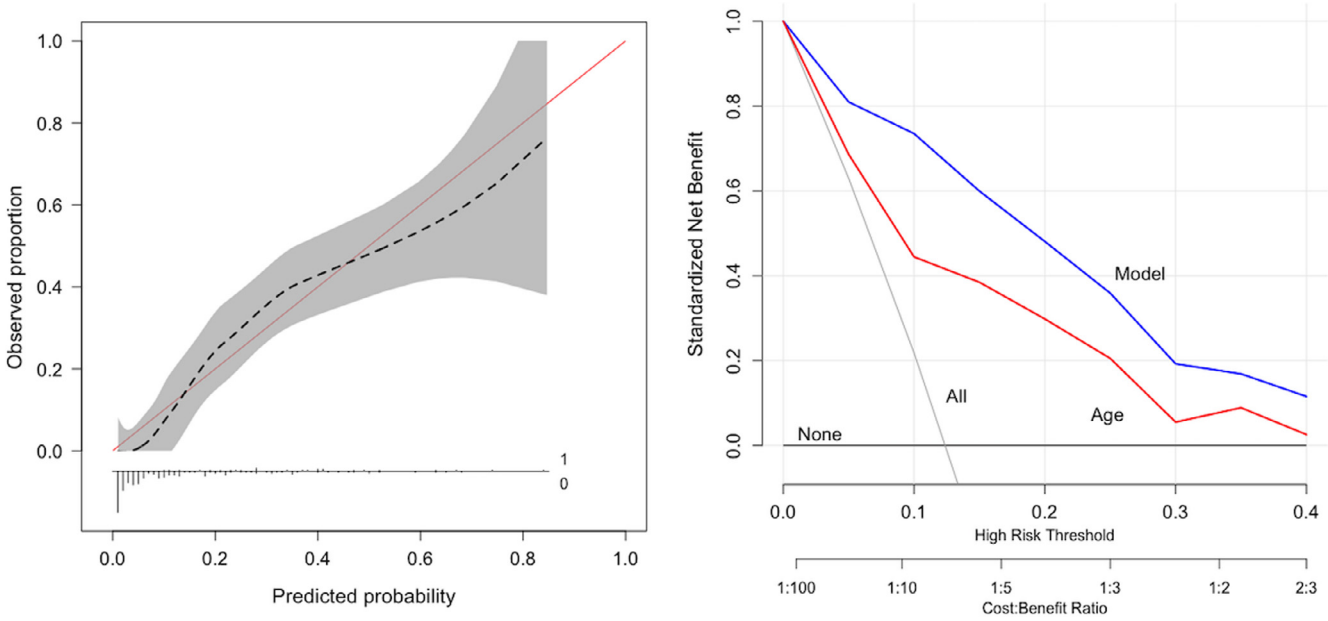

Fig. 3. (A) Calibration plot for the SGB algorithm, calibration intercept=0.01, calibration intercept=1.23, in the testing set, n=209. (B) Decision curve analysis with net benefit achieved by management changes based on the SGB algorithm relative to the default strategies of changing management for no patients and for all patients as well as those based solely on patient age. SGB, stochastic gradient boosting.

Table 4
Sensitivity analysis of patients admitted only in the last 10 years of the available cohort (2007–2016). Discrimination and calibration of algorithms on repeated cross-validation of training set, n=508, mean (95% confidence interval)

Metric	Stochastic gradient boosting	Random forest	Support vector machine	Neural network	Elastic-net penalized logistic regression
AUC	0.81 (0.79, 0.84)	0.78 (0.75, 0.81)	0.61 (0.56, 0.67)	0.84 (0.81, 0.86)	0.84 (0.82, 0.87)
Intercept	−0.08 (−0.47, 0.31)	−0.38 (−0.63, −0.12)	1.03 (−1.64, 3.70)	0.10 (−0.32, 0.33)	0.41 (0.08, 0.74)
Slope	0.95 (0.72, 1.19)	0.73 (0.58, 0.88)	1.52 (0.24, 2.79)	1.12 (0.87, 1.39)	1.29 (1.07, 1.52)
Brier	0.09 (0.08, 0.09)	0.09 (0.09, 0.10)	0.10 (0.10, 0.10)	0.08 (0.08, 0.09)	0.08 (0.08, 0.08)

AUC, area under the receiver operating curve. Null model Brier score=0.10.

Table 5
Sensitivity analysis of patients admitted only in the last 10 years of the available cohort (2007–2016). Discrimination and calibration of algorithms in holdout set, n=127

Metric	Stochastic gradient boosting	Random forest	Support vector machine	Neural network	Elastic-net penalized logistic regression
AUC	0.86	0.83	0.76	0.87	0.89
Intercept	0.07	−0.17	5.51	0.06	0.86
Slope	1.00	0.88	3.74	1.58	1.50
Brier	0.08	0.08	0.11	0.08	0.08

AUC, area under the receiver operating curve. Null model Brier score=0.11.

compromise, and the nature of the infecting organism(s). These facets of SEA, when considered holistically, help to explain why they figure so prominently in the prognostic model developed here.

Few other studies have developed predictive algorithms for short-term mortality in SEA. Shah et al. previously developed a nomogram for predicting 90-day mortality in patients admitted with a diagnosis of SEA; the current study extended this prior work by assessing the utility of machine learning algorithms [4]. Du et al. also developed a risk score for 90-day postoperative mortality using the National Surgical Quality Improvement Program. Neither study examined calibration plots, overall performance (Brier score), or decision curve analysis [25]. The present study overcame these previous limitations by rigorously examining these standardized metrics of model performance assessment.

Nonetheless, this study has several limitations. This remains a retrospective assessment of SEA patients admitted

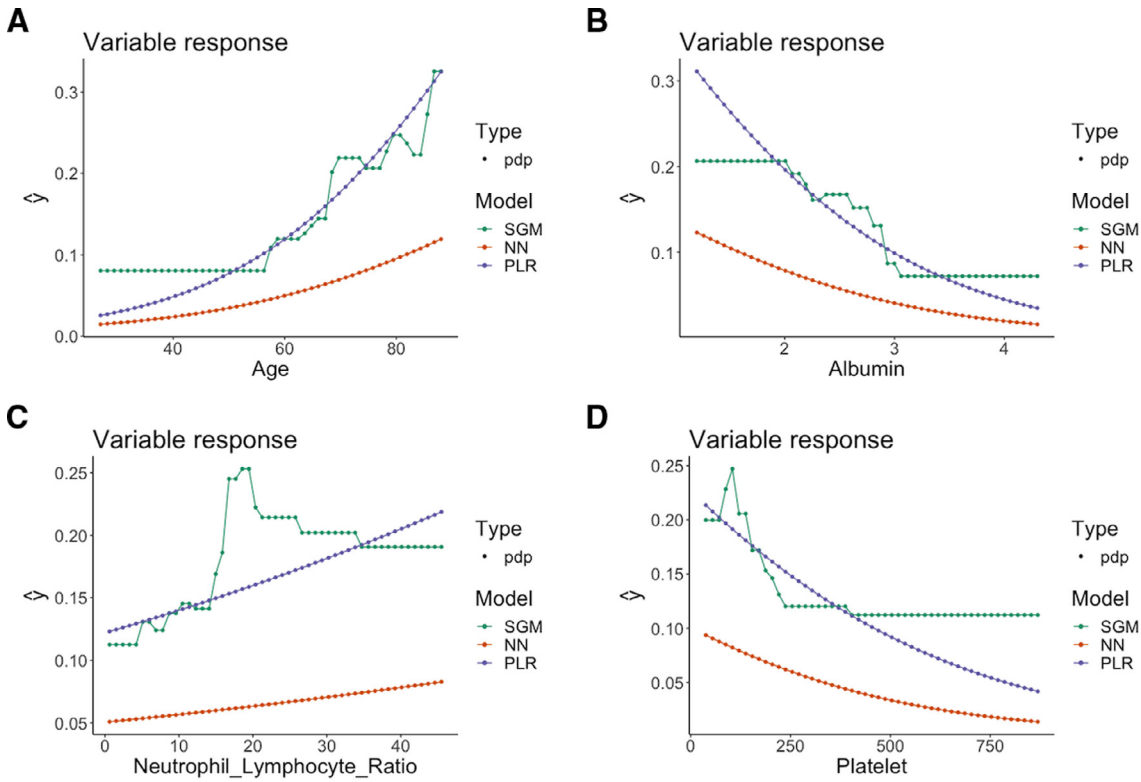

Fig. 4. (A–D) Partial dependence plots for age (years) (A), albumin (g/dL) (B), neutrophil to lymphocyte ratio (NLR) (C), and platelet count ($10^3/\mu$L) (D). These plots show the relationship between the model outputs for the elastic-net penalized logistic regression (PLR), neural network (NN), and stochastic gradient boosting (SGB) models and the input continuous variables of age, albumin, NLR and platelet. g/dL, grams per deciliter; NN, neural network; PLR, penalized logistic regression; SGB, stochastic gradient boosting; μL, microliter.

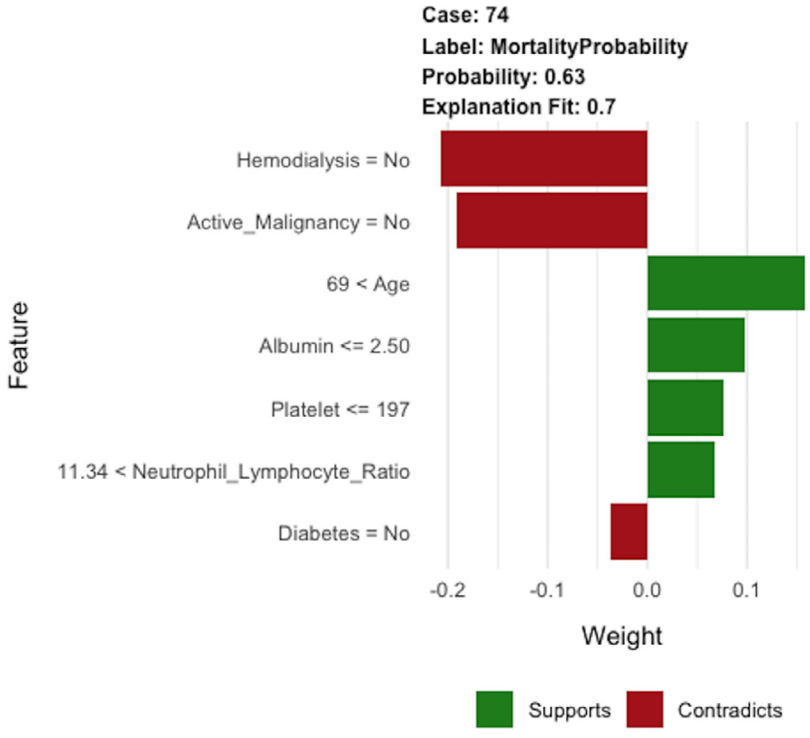

Fig. 5. Illustrative example of case-specific local explanation of the SGB model predictions. SGB, stochastic gradient boosting.

to hospitals in one health-care system, with the potential for many shared institutional practices. This may have led to clustering and influenced the direction of the study findings, including decisions for surgery, extent of nonoperative care and consequent survival. The generalization potential of these models in multi-institutional samples remains to be determined but is clearly an avenue for future research. The patients included in this study spanned two decades of SEA admission at our institution. Accrual of a sufficient volume of patients to enable meaningful evaluation of variation in prognostic factors is a contributing factor to this limitation but future studies should nonetheless consider study designs that span narrower time periods, if practicable. Certain variables such as c-reactive protein had levels of missing data greater than 30% and were not included in the analysis; future studies should seek to further characterize the utility of these markers for mortality in SEA patients. This study included all patients with a confirmed diagnosis of SEA; future studies should be undertaken to examine mortality for patients managed only operatively or nonoperatively and to study factors such as the influence of time to intervention on subsequent outcomes.

Although this study showed promising results on internal validation, much work remains before the SGB model can be applied with confidence to clinical care by spine care providers. As mentioned above, rigorous assessment of the SGB model must be carried out in an independent population of SEA patients, ideally from multiple institutions. Overfitting occurs when algorithms memorize patterns that are idiosyncratic of the developmental population rather than extracting patterns that generalize to independent populations. With overfitting, algorithm performance may appear to be excellent in the developmental population but may vary dramatically when applied to external cohorts. Clinicians cannot rely on the developmental study alone and must look to external validation studies to substantiate the final algorithm. Ultimately, for algorithms that are found to be externally valid, the next steps are to derive interventions that are capable of influencing survival for patients with SEA and to measure the impact of these proposed interventions. A series of future studies are required to address the needs highlighted above.

The primary benefit of this analysis has been the establishment of the first digital clinical prediction tool for mortality in SEA. This tool not only applies machine learning but also provides individual patient level explanations to overcome the conventional drawbacks of other statistical models. Continued advancements in the application of data science may result in improved outcomes for patients with SEA.

Conclusions

Machine learning algorithms show promise on internal validation for prediction of 90-day mortality in SEA. Future studies are needed to externally validate these algorithms in independent populations.

Supplementary materials

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.spinee.2019.06.024.

References

[1] Schoenfeld AJ, Wahlquist TC. Mortality, complication risk, and total charges after the treatment of epidural abscess. Spine J 2015;15:249–55.
[2] Baker AS, Ojemann RG, Swartz MN, Richardson EP. Spinal epidural abscess. N Engl J Med 1975;293:463–8.
[3] Darouiche RO. Spinal epidural abscess. N Engl J Med 2006;355:2012–20.
[4] Shah AA, Ogink PT, Harris MB, Schwab JH. Development of predictive algorithms for pre-treatment motor deficit and 90-day mortality in spinal epidural abscess. JBJS 2018;100:1030–8.
[5] Reihsaus E, Waldbaur H, Seeling W. Spinal epidural abscess: a meta-analysis of 915 patients. Neurosurg Rev 2000;23:175–204.
[6] Hlavin ML, Kaminski HJ, Ross JS, Ganz E. Spinal epidural abscess: a ten-year perspective. Neurosurgery 1990;27:177–84.
[7] Curry Jr WT, Hoh BL, Amin-Hanjani S, Eskandar EN. Spinal epidural abscess: clinical presentation, management, and outcome. Surg Neurol 2005;63:364–71.
[8] Soehle M, Wallenfang T. Spinal epidural abscesses: clinical manifestations, prognostic factors, and outcomes. Neurosurgery 2002;51:79–87.
[9] Chaker AN, Bhimani AD, Esfahani DR, Rosinski CL, Geever BW, Patel AS. Epidural abscess: a Propensity analysis of surgical treatment strategies. Spine 2018;43:E1479–E85.
[10] Du JY, Schell AJ, Kim CY, Trivedi NN, Ahn UM, Ahn NU. 30-day mortality following surgery for spinal epidural abscess: incidence, risk factors, predictive algorithm, and associated complications. Spine 2018;44:E500–9.
[11] Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719.
[12] Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML. Natural and artificial intelligence in neurosurgery: a systematic review. Neurosurgery 2017;83:18–92.
[13] Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. Front Bioeng Biotechnol 2018;6.
[14] Thio QCBS, Karhade AV, Ogink PT, Raskin KA, De Amorim Bernstein K, Lozano Calderon SA. Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma? Clin Orthop Related Res 2018;476:2040–8.
[15] Karhade AV, Ogink P, Thio Q, Broekman M, Cha T, Gormley WB. Development of machine learning algorithms for prediction of discharge disposition after elective inpatient surgery for lumbar degenerative disc disorders. Neurosurg Focus 2018;45:E6.
[16] Karhade AV, Thio QCBS, Ogink PT, Shah AA, Bono CM, Oh KS. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. Neurosurgery 2018;85:E83–91.
[17] Karhade AV, Ogink PT, Thio QCBS, Broekman MLD, Cha TD, Hershman SH. Machine Learning for Prediction of Sustained Opioid Prescription After Anterior Cervical Discectomy and Fusion. Spine J 2019.
[18] Kim JS, Arvind V, Oermann EK, Kaji D, Ranson W, Ukogu C. Predicting surgical complications in patients undergoing elective adult spinal deformity procedures using machine learning. Spine Deform 2018;6:762–70.
[19] Staartjes VE, Marlies P, Vandertop WP, Schröder ML. Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar diskectomy: feasibility of center-specific modeling. Spine J 2018;19:853–61.
[20] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or

diagnosis (TRIPOD): the TRIPOD statement. BMJ (Clinical research ed) 2015;350:g7594.
[21] Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18:e323.
[22] Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013.
[23] Wainer J. Comparison of 14 different families of classification algorithms on 115 binary datasets. arXiv preprint arXiv:160600930. 2016.
[24] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer Science & Business Media; 2008.
[25] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31.
[26] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74.
[27] Greenwell BM, Boehmke BC, McCarthy AJ. A Simple and effective model-based variable importance measure. arXiv preprint arXiv:180504755. 2018.
[28] Biecek P. DALEX: explainers for complex predictive models in R. J Mach Learn Res 2018;19:3245–9.
[29] Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:160605386. 2016.
[30] Karhade AV, Shah AA, Lin KY, Ogink PT, Shah KC, Nelson SB. Albumin and spinal epidural abscess: derivation and validation in two independent data sets. World Neurosurg 2019;123:e416–e26.
[31] Karhade AV, Shah KC, Shah AA, Ogink PT, Nelson SB, Schwab JH. Neutrophil to lymphocyte ratio and mortality in spinal epidural abscess. Spine J 2019;19:1180–5.