Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation

Karhade, A.V.; Thio, Q.C.B.S.; Ogink, P.T.; Bono, C.M.; Ferrone, M.L.; Oh, K.S.; ... ; Schwab, J.H.

Predicting 90-Day and 1-Year Mortality in Spinal Metastatic Disease: Development and Internal Validation

Aditya V. Karhade, BE*
Quirina C.B.S. Thio, MD*
Paul T. Ogink, MD*
Christopher M. Bono, MD*
Marco L. Ferrone, MD‡
Kevin S. Oh, MD§
Philip J. Saylor, MD¶
Andrew J. Schoenfeld, MD‡
John H. Shin, MD‖
Mitchel B. Harris, MD*
Joseph H. Schwab, MD, MS*

*Department of Orthopedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts; ‡Department of Orthopedic Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; §Department of Radiation Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts; ¶Department of Hematology/Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts; ‖Department of Neurosurgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts

Correspondence:
Joseph H. Schwab, MD, MS,
Department of Orthopedic Surgery,
Massachusetts General Hospital,
Harvard Medical School,
55 Fruit Street,
Boston, MA 02114.
E-mail: jhschwab@mgh.harvard.edu

Received, September 14, 2018.
Accepted, February 12, 2019.
Published Online, March 14, 2019.

BACKGROUND: Increasing prevalence of metastatic disease has been accompanied by increasing rates of surgical intervention. Current tools have poor to fair predictive performance for intermediate (90-d) and long-term (1-yr) mortality.

OBJECTIVE: To develop predictive algorithms for spinal metastatic disease at these time points and to provide patient-specific explanations of the predictions generated by these algorithms.

METHODS: Retrospective review was conducted at 2 large academic medical centers to identify patients undergoing initial operative management for spinal metastatic disease between January 2000 and December 2016. Five models (penalized logistic regression, random forest, stochastic gradient boosting, neural network, and support vector machine) were developed to predict 90-d and 1-yr mortality.

RESULTS: Overall, 732 patients were identified with 90-d and 1-yr mortality rates of 181 (25.1%) and 385 (54.3%), respectively. The stochastic gradient boosting algorithm had the best performance for 90-d mortality and 1-yr mortality. On global variable importance assessment, albumin, primary tumor histology, and performance status were the 3 most important predictors of 90-d mortality. The final models were incorporated into an open access web application able to provide predictions as well as patient-specific explanations of the results generated by the algorithms. The application can be found at https://sorg-apps.shinyapps.io/spinemetssurvival/

CONCLUSION: Preoperative estimation of 90-d and 1-yr mortality was achieved with assessment of more flexible modeling techniques such as machine learning. Integration of these models into applications and patient-centered explanations of predictions represent opportunities for incorporation into healthcare systems as decision tools in the future.

KEY WORDS: Machine learning, 90-day, 1-year, Prognosis, Spine metastasis, Survival, Explanation

Neurosurgery 85:E671–E681, 2019 DOI:10.1093/neuros/nyz070 www.neurosurgery-online.com

The increasing prevalence of spinal metastatic disease has been accompanied by enhanced enthusiasm for surgical intervention.[1,2] However, accurate estimation of postoperative survival is a necessary prerequisite to justify surgical management and several prognostic aids have been developed for this purpose.[3-12] While several utilities demonstrate good discriminative capacity for short-term mortality (30 d) in spinal metastatic disease, they do not perform as well in terms of prognosticating intermediate (90-d) and long-term (1-yr) mortality following surgery.[3,8,13]

The primary purpose of this study was to develop models capable of prediction for intermediate and long-term postoperative mortality in spinal metastatic disease, including the application of flexible modeling techniques, such as machine learning. The drawbacks of machine

ABBREVIATIONS: ASIA, American Spinal Injury Association; **AUC,** area under the receiver operating curve; **ECOG,** Eastern Cooperative Oncology Group; **INR,** international normalized ratio; **PLR,** penalized logistic regression; **PT,** prothrombin time; **PTT,** partial thromboplastin time; **SGB,** stochastic gradient boosting; **TRIPOD,** Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

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

learning include an inability for clinicians to access fully developed models through simple tools such as risk scores or nomograms, and limited capacity to explain the predictions of these models at the global (for all patients in a given population) or local (for an individual patient) levels. The final aim of this study was to propose one method for overcoming these primary limitations of machine learning by providing an open access digital application for healthcare professionals that more effectively and transparently communicates the results generated by machine learning models for patients with spinal metastases.

METHODS

Guidelines

This retrospective, prognostic classification study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines and the Guidelines for Developing and Reporting Machine Learning Models in Biomedical Research.[14,15]

Source of Data

A retrospective chart review was conducted on patients surgically treated for spinal metastases at 2 large academic medical centers. Health records review was approved by our institutional review board. Individual patient consent was waived as the study was limited to retrospective chart review.

Participants

The following inclusion criteria were applied: (1) adult patients defined as age greater than 18 yr, (2) diagnosis of spinal metastatic disease (solid tumor metastases, multiple myeloma, lymphoma) with or without pathological fracture, and (3) initial surgical procedure performed between January 1, 2000 and December 31, 2016.

Outcome

Survival was assessed by cross-referencing the Social Security Death Index and through chart review. The date of last review was July 11, 2018. For this study, the primary outcomes consisted of 90-d and 1-yr mortality. Ninety-day and 1-yr mortality could be ascertained in 722 (98.6%) and 709 (96.9%) patients, respectively.

Predictors

The following preoperative variables were candidates based on the findings of previous research: age (years), sex, body mass index (kilograms per meter squared [kg/m^2]), preoperative presence of any Charlson comorbidity other than metastatic disease,[16] primary tumor histology (based on the classifications postulated by Katagiri et al[17]; Tables S1 and S2 in **Appendix, Supplemental Digital Content**), pathological fracture at presentation, pain at presentation, Eastern Cooperative Oncology Group (ECOG) performance status, American Spinal Injury Association (ASIA) Impairment Scale, spine tumor location, number of spinal metastases, other nonspine bone metastases, presence of visceral metastases (metastases in liver or lung), presence of brain metastases, history of local radiation to affected site, history of previous systemic therapy, white blood cell count ($\times 10^3$ per microliter [μL]), hemoglobin (grams per deciliter [g/dL]), platelet count ($\times 10^3/\mu$L), absolute lymphocyte count ($\times 10^3/\mu$L), absolute neutrophil count ($\times 10^3/\mu$L), platelet to absolute lymphocyte ratio, neutrophil-to-absolute lymphocyte ratio, albumin (g/dL), alkaline phosphatase (international units per liter [IU/L]), calcium (milligrams per deciliter [mg/dL]), creatinine (mg/dL), prothrombin time (PT), partial thromboplastin time (PTT), and international normalized ratio (INR).

Missing Data

Rates of missing data for covariates were as follows: body mass index = 90 (12.3%), ECOG performance status = 193 (26.4%), ASIA impairment scale = 11 (1.5%), white blood cell count = 89 (12.2%), hemoglobin = 89 (12.2%), platelet = 89 (12.2%), absolute lymphocyte = 218 (29.8%), absolute neutrophil = 215 (29.3%), albumin = 185 (25.3%), alkaline phosphatase = 194 (26.5%), calcium = 110 (15%), creatinine = 89 (12.2%), PT = 117 (15.9%), PTT = 143 (19.5%), INR = 134 (18.3%). Multiple imputation with the missForest methodology was used to impute variables with less than 30% missing data.[18]

Statistical Analysis and Methods

For 90-d mortality, the total population was divided into a training and testing cohort with 80:20 stratified split, ensuring equal proportions of the outcome in each set. The subsequent steps (variable selection, model building, predictive performance assessment) were conducted on the training set, and the final models developed on the training set were evaluated on the independent testing (holdout) set as described below. This process was repeated for 1-yr mortality. Feature selection was carried out by random forest algorithms with 10-fold cross-validation repeated 3 times.[19-22] Additional details are included in the supplemental methods in the **Appendix, Supplemental Digital Content**.

The following algorithms were chosen for modeling based on prior research: (1) random forest (2) stochastic gradient boosting (SGB) (3) neural network (4) support vector machine (5) penalized logistic regression.[23] Description of the methodology of these algorithms is included in the supplemental methods in the **Appendix, Supplemental Digital Content**.

Ten-fold cross-validation repeated 3 times was used to assess predictive performance on the training set. Discrimination was assessed by plotting the receiver operating curve and calculating the area under the curve.[24,25] The area under the receiver operating curve (AUC) represents the probability the model will correctly separate a given patient into an outcome category. Calibration was assessed by plotting calibration curves and examining calibration slope and calibration intercept.[24,25] Calibration intercept represents the overall tendency of the model to overestimate (greater than zero) or underestimate (less than zero) the true probability. Brier score was calculated for an overall assessment of discrimination and calibration:[24-26]

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N} Y_i - p_i^2,$$

where N is the sample size, Y is the true outcome, and p is the predicted probability. The null model Brier score was calculated (setting the predicted probability equal to the prevalence of the outcome of interest in the study population) to compare the gain of the model relative to this benchmark.

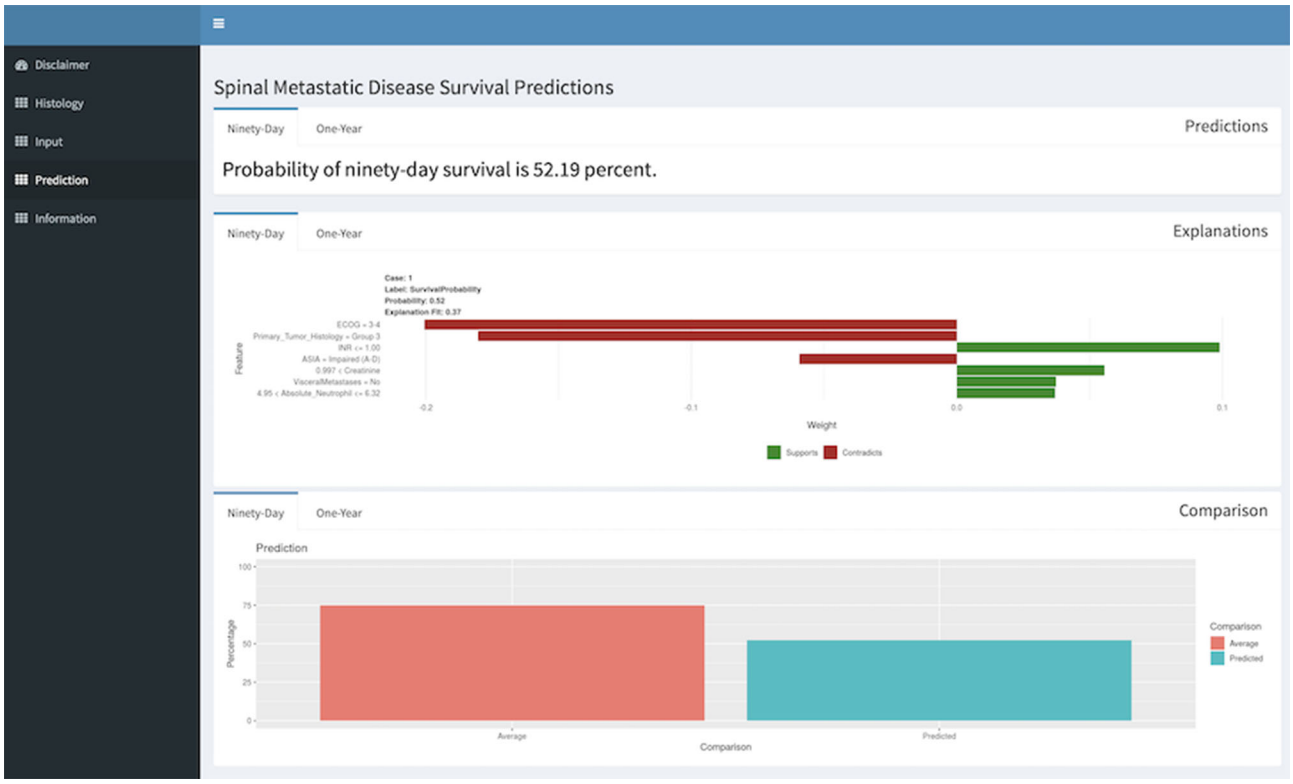

FIGURE 1. *Web application interface for final algorithm. Top, tabs for 90-d and 1-yr predictions. Middle, explanations for 90-d and 1-yr predictions. Bottom, comparison to average for 90-d and 1-yr mortality. Lower right, explanation of prediction. Abbreviations: ASIA: American Spinal Injury Association impairment scale; ECOG: Eastern Cooperative Oncology Group performance status; INR: International normalized ratio.*

Decision curves were created for prediction of 90-d mortality by plotting the net benefit over the range of predicted probabilities.[27,28] The net benefit is defined as:

$$\text{Net beneffit} = \frac{\text{True Positives} - \text{relativeWeight}\left(\text{False Positives}\right)}{\text{Number of Patients}}$$

Relative weight is determined by the threshold probability and is defined by the threshold:

$$\text{relativeWeight} = \frac{\text{threshold probability}}{1 - \text{threshold probability}}$$

For example, at a threshold probability of 20%, one true positive is weighted equal to 4 false positives and the relative weight is 0.25. Decision curves plot all possible threshold probabilities so the user can determine what threshold best suits individual needs and simultaneously assess the predicted net benefit of using the model at that particular threshold.

Predictions of the final models were explained globally and for individual assessments. Further description is included in the supplemental methods in the **Appendix, Supplemental Digital Content**.

Existing scoring systems for spinal metastatic disease were evaluated by c-statistic on the overall study population to derive a reference point for the performance of the machine learning algorithms developed here. The original Tokuhashi, revised Tokuhashi, Tomita, modified Bauer, Van der Linden, Katagiri, New England Spine Metastasis Score, SORG Classic, SORG Nomogram, and Bollen scores were assessed for predictive performance at 90-d and 1-yr mortality.[5-7,10-12,29-31] The discrimination of the best performing existing scoring systems was compared using Delong's method to the highest performing algorithms developed in this study with a significance threshold set a priori to $P < .05$.

The final algorithms for 90-d and 1-yr mortality prediction were deployed as an open access web application (Figure 1). The Anaconda Distribution (Anaconda Inc, Austin, Texas), R version 3.4.3 (The R Foundation, Vienna, Austria), RStudio version 1.0.153 (RStudio, Boston, Massachusetts), and Python version 3.6 (Python Software Foundation, Wilmington, Delaware) were used for data analysis, model building, and application development.

RESULTS

Participants

Seven hundred thirty-two patients were included in this study with 90-d and 1-yr mortality rates of 181 (25.1%) and 385 (54.3%), respectively. Three hundred and six (41.8%) patients were female, and the median age was 61 (interquartile range = 53-69; Table 1).

TABLE 1. Baseline Characteristics of Study Population, n = 732

Variable	n (%) | median (IQR)
Age (years)	61 (53-69)
Female sex	306 (41.8)
BMI (kg/m^2)	26.3 (23.1-29.7)
Other Charlson comorbidity	441 (60.7)
Primary tumor histology	
Slow growth	219 (29.9)
Intermediate	254 (34.7)
Rapid	259 (35.4)
Pathological fracture	456 (62.3)
Pain	627 (85.7)
ECOG	
0-2	440 (81.6)
3-4	99 (18.4)
ASIA	
Normal (E)	379 (52.6)
Impaired (A-D)	342 (47.4)
Tumor location	
Cervical	104 (14.2)
Thoracic	425 (58.1)
Lumbar	164 (22.4)
Multiple	39 (5.3)
Spine metastases	
1	211 (28.8)
2	117 (16)
3 or more	404 (55.2)
Other bone metastases	388 (53)
Visceral metastases	252 (34.4)
Brain metastases	81 (11.1)
History of local radiation	252 (34.4)
Previous systemic therapy	418 (57.1)
White blood cell ($\times 10^3/\mu$L)	8.20 (6.06-10.9)
Hemoglobin (g/dL)	12.1 (10.7-13.3)
Platelet ($\times 10^3/\mu$L)	259 (196 - 337)
Absolute lymphocyte ($\times 10^3/\mu$L)	0.90 (0.58-1.43)
Absolute neutrophil ($\times 10^3/\mu$L)	6.32 (4.48-8.8)
Neutrophil to lymphocyte ratio	7.22 (3.64-12.82)
Platelet to lymphocyte ratio	280.5 (172.9-461.2)
Albumin (g/dL)	3.80 (3.4-4.2)
Alkaline phosphatase (IU/L)	94.5 (73-140)
Calcium (mg/dL)	9.10 (8.6-9.5)
Creatinine (mg/dL)	0.80 (0.69-1.0)
INR	1.10 (1.0-1.1)
PT	13.5 (12.8-14.3)
PTT	27.7 (25.2-31.3)
Number of levels operated	
1 or 2	531 (72.6)
Three or more	200 (27.4)
Anterior approach	105 (14.3)
Posterior approach	660 (90.2)
Combined approach	33 (4.5)
Decompression	699 (95.5)
Stabilization	640 (87.4)
Corpectomy	351 (48.0)

ASIA, American Spinal Injury Association Impairment Scale; BMI, body mass index; ECOG, Eastern Cooperative Oncology Group performance status; g/dL, grams per deciliter; IU/L, international units per liter; IQR, interquartile range; kg/m^2, kilograms per meter squared; mg/dL, milligrams per deciliter; μL, microliter.

Ninety-Day Mortality

Variables selected for prediction of 90-d mortality were as follows: primary tumor histology, ECOG, other Charlson comorbidity, ASIA scale, visceral metastasis, 3 or more spinal metastases, hemoglobin, platelet, absolute lymphocyte, absolute neutrophil, neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, creatinine, albumin, alkaline phosphatase, and INR.

Model Development and Performance

On cross-validation of the training set, n = 587 (80%), all 5 algorithms achieved good discriminative performance (AUC 0.83-0.85) but demonstrated varied calibration (Table 2). The final algorithms developed by cross-validation of the training set were also evaluated on the full training set to assess the propensity for bias and variance when compared to the testing set; for example, the random forest algorithm had high fluctuations from a c-statistic of 0.99 and Brier score of 0.02 on the full training set to a c-statistic of 0.80 and Brier score of 0.14 in the testing set (Table S3 in **Appendix**, **Supplemental Digital Content**). The SGB algorithm was chosen as the final model with AUC = 0.83, calibration intercept = −0.04, calibration slope = 0.95 and Brier score = 0.14 on cross-validation of the training set. In the independent sample not used for algorithm development, n = 145 (20%), SGB achieved the best performance with AUC = 0.83, calibration intercept = 0.04, calibration slope = 0.92, and Brier score = 0.13 (Table 3). In comparison, the null model Brier score was 0.19. On global variable importance assessment, preoperative albumin, primary tumor histology, and ECOG performance status were the 3 most important variables for prediction of 90-d mortality by the SGB model (Figure 2). On decision curve analysis, the SGB model showed greater standardized net benefit at all predicted probabilities relative to default strategies of decision change for all patients or no patients (Figure 3).

One-Year Mortality

Variables used for prediction of 1-yr mortality were primary tumor histology, ECOG, other Charlson comorbidity, brain metastases, previous systemic therapy, and body mass index, ASIA scale, visceral metastasis, hemoglobin, platelet, absolute lymphocyte, absolute neutrophil, neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, creatinine, albumin, alkaline phosphatase, and INR.

Model Development and Performance

On cross-validation of the training set, all 5 algorithms achieved good discriminative performance (AUC 0.84-0.85) but again showed varied performance on calibration. Similar to the above results for 90-d mortality, the final random forest and neural network algorithms had high fluctuations from c-statistics of 0.96 to 0.99 and Brier score of 0.02 to 0.07 on evaluation in the full training set to c-statistics of 0.86 to 0.88 and Brier scores of 0.14 to 0.15 in the testing set (Table S3 in **Appendix**, **Supplemental Digital Content**). The SGB algorithm was chosen

TABLE 2. Discrimination and Calibration of Algorithms on Repeated Cross-Validation of Training Set, n = 587, mean (95% Confidence Interval)

Metric	Stochastic gradient boosting	Random forest	Support vector machine	Neural network	Penalized logistic regression
			Ninety-day mortality		
AUC	0.83 (0.81-0.85)	0.84 (0.83-0.86)	0.84 (0.83-0.86)	0.84 (0.82-0.85)	0.85 (0.84-0.87)
Intercept	− 0.04 (−0.17 to 0.09)	0.30 (0.17-0.43)	0.22 (0.02-0.43)	− 0.13 (−0.22 to 0.04)	0.39 (0.21-0.57)
Slope	0.95 (0.85-1.05)	1.42 (1.27-1.57)	1.20 (1.06-1.35)	0.77 (0.71-0.83)	1.44 (1.29-1.58)
Brier	0.14 (0.13-0.14)	0.14 (0.13-0.14)	0.13 (0.13-0.13)	0.14 (0.13-0.14)	0.13 (0.13-0.14)
			One-year mortality		
AUC	0.85 (0.83-0.87)	0.85 (0.83-0.87)	0.84 (0.83-0.86)	0.84 (0.82-0.86)	0.85 (0.83-0.87)
Intercept	0.01 (−0.08 to 0.11)	0.02 (−0.12 to 0.17)	0.04 (−0.05 to 0.12)	0.02 (−0.07 to 0.11)	0.01 (−0.09 to 0.08)
Slope	1.07 (0.93-1.21)	1.31 (1.10-1.51)	1.19 (1.03-1.35)	0.84 (0.74-0.94)	1.51 (1.33-1.68)
Brier	0.16 (0.15-0.17)	0.16 (0.15-0.17)	0.16 (0.15-0.17)	0.16 (0.15-0.17)	0.16 (0.15-0.17)

AUC, area under the receiver operating curve. Ninety-day mortality, null model Brier score = 0.19. One-year mortality null model Brier score = 0.25.

TABLE 3. Discrimination and Calibration of Algorithms in Holdout Set, n = 145

Metric	Stochastic gradient boosting	Random forest	Support vector machine	Neural network	Penalized logistic regression
			Ninety-day mortality		
AUC	0.83	0.80	0.77	0.77	0.79
Intercept	0.04	0.13	− 0.18	− 0.31	− 0.01
Slope	0.92	1.06	0.73	0.55	0.98
Brier	0.13	0.14	0.16	0.16	0.15
			One-year mortality		
AUC	0.89	0.88	0.89	0.86	0.90
Intercept	0.07	− 0.03	0.02	− 0.05	0.02
Slope	1.26	1.45	1.53	0.86	1.88
Brier	0.13	0.14	0.14	0.15	0.14

AUC, area under the receiver operating curve. Ninety-day mortality, null model Brier score = 0.19. One-year mortality, null model Brier score = 0.25.

as the final model with AUC = 0.85, calibration intercept = 0.01, calibration slope = 1.07, and Brier score = 0.16 (Table 2). In the independent sample not used for algorithm development, SGB also achieved excellent performance with AUC = 0.89, calibration intercept = 0.07, calibration slope = 1.26, and Brier score = 0.13 (Table 3). In comparison, the null model Brier score was 0.25. On global variable importance assessment, primary tumor histology, preoperative albumin, and preoperative hemoglobin were the 3 most important variables for prediction of 1-yr mortality by the SGB model (Figure 2). On decision curve analysis, the SGB model showed greater standardized net benefit at all predicted probabilities relative to management decision change based on all patients or no patients (Figure 3).

Model Explanations

Partial dependence plots for 90-d mortality are shown in Figure 4. The outcome ($\hat{y}$ or the predicted probability) of the penalized logistic regression (PLR) was linearly related to albumin, whereas the outcome of the SGB plateaued at albumin less than 3.3 g/dL and albumin greater than 4 g/dL. Similarly, the outcome of the PLR was linearly related to hemoglobin, whereas the outcome of the SGB plateaued at hemoglobin levels greater than 14.5 g/dL. These plots illustrate the differences in the transformation function that each algorithm applied to the input predictors to achieve the model estimations. The SGB was better able to capture nonlinearities in continuous predictors.

Local explanations for the SGB model for prediction of 90-d mortality are shown in Figure 5. For example, panel A shows a patient with impaired ASIA (A-D) neural function, visceral metastases, primary tumor histology group 3 (rapid growth), albumin 2.7, neutrophil-to-lymphocyte ratio 18.1, hemoglobin 9.7, alkaline phosphatase 237, absolute lymphocyte 0.85, platelet 369, platelet-to-lymphocyte ratio 434, creatinine 0.75, INR 1.1, absolute neutrophil 15.4, ECOG performance status 0-2, more

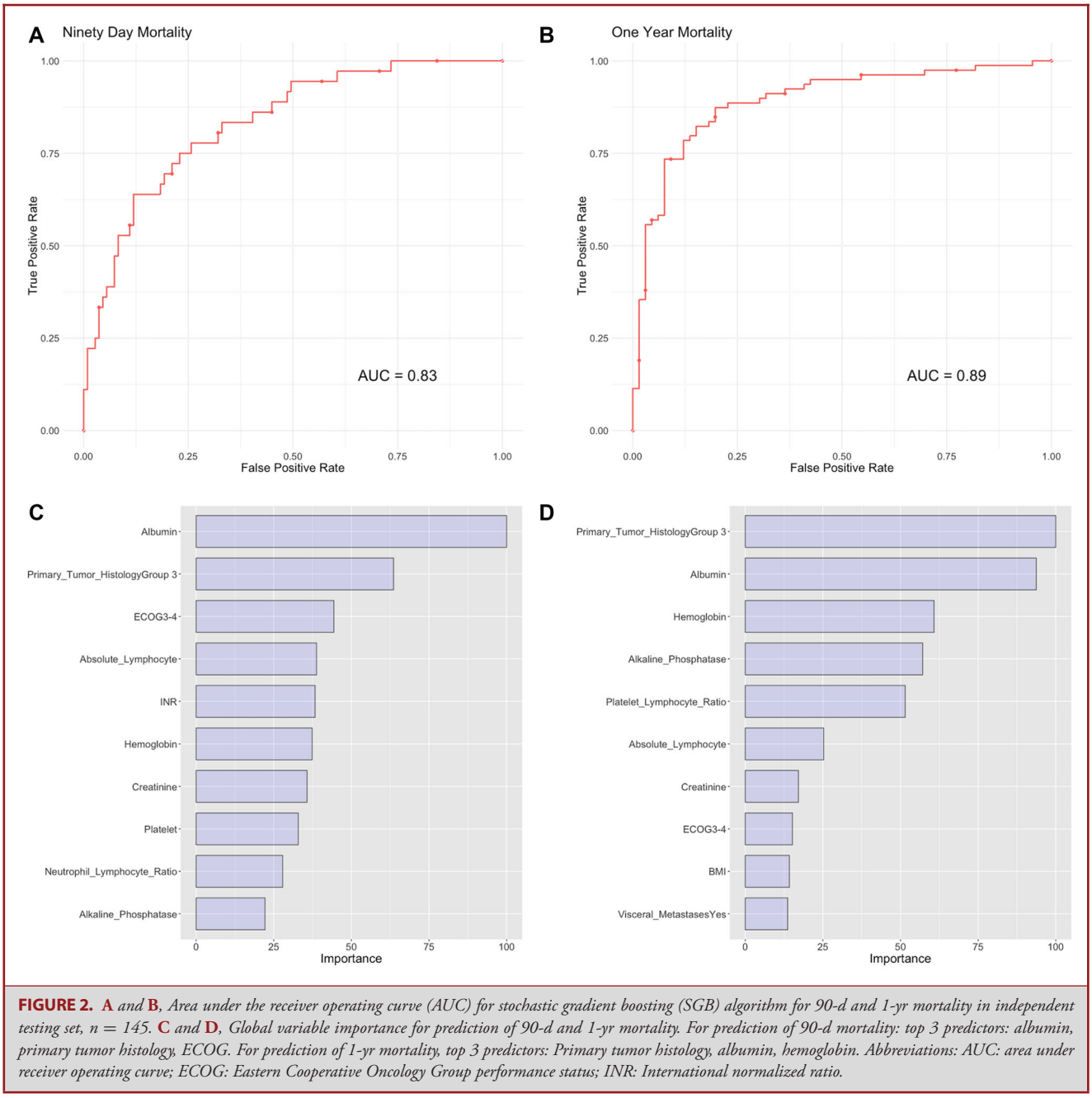

FIGURE 2. **A** *and* **B**, *Area under the receiver operating curve (AUC) for stochastic gradient boosting (SGB) algorithm for 90-d and 1-yr mortality in independent testing set, n = 145.* **C** *and* **D**, *Global variable importance for prediction of 90-d and 1-yr mortality. For prediction of 90-d mortality: top 3 predictors: albumin, primary tumor histology, ECOG. For prediction of 1-yr mortality, top 3 predictors: Primary tumor histology, albumin, hemoglobin. Abbreviations: AUC: area under receiver operating curve; ECOG: Eastern Cooperative Oncology Group performance status; INR: International normalized ratio.*

than 2 spine metastases, and other Charlson comorbidity. The model's 90-d mortality prediction for this patient was 0.835 and the following factors resulted in an adjustment that increased the probability of mortality: albumin < 3.5, primary tumor histology group 3, hemoglobin < 10.8, absolute neutrophil > 8.5, impaired ASIA, and neutrophil-to-lymphocyte ratio greater than 12.6. In contrast, ECOG performance 0 to 2, resulted in an adjustment that favored survival in the 90 d following surgery.

Model Specification

The final model is available at https://sorg-apps.shinyapps.io/spinemetssurvival/

The default values included in the web application are placeholders. Clinicians can modify the inputs as per the clinical characteristics of individual patients to examine the impact on survival prediction in real time. Clinicians should be aware that the algorithms require complete information for the factors included in the interface.

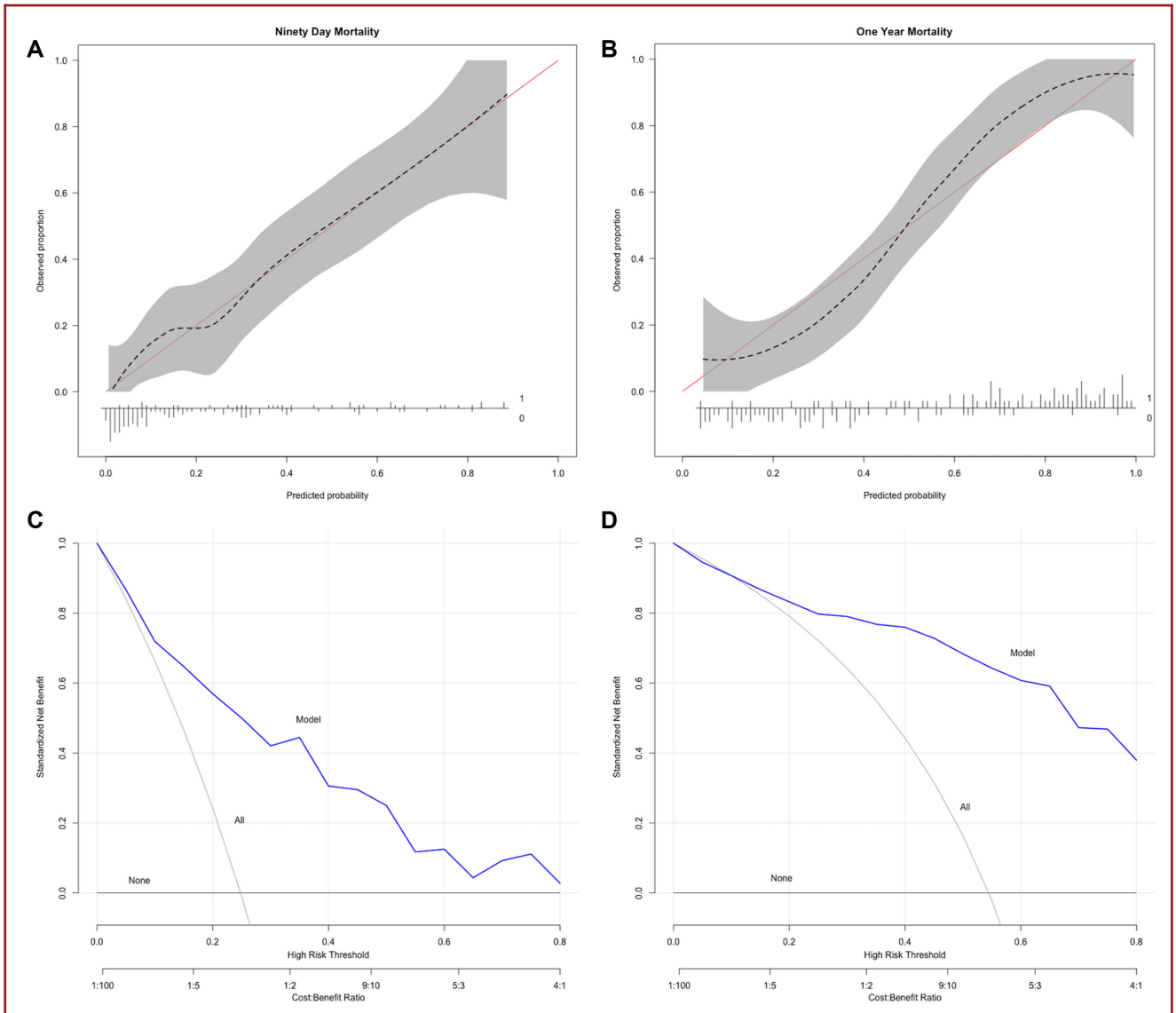

FIGURE 3. **A** *and* **B**, *Calibration plots for stochastic gradient boosting (SGB) algorithm for 90-d mortality and 1-yr mortality in independent testing set, n = 145.* **C** *and* **D**, *Decision curve analysis with expected standardized net benefit by threshold probability for 90-d mortality and 1-yr mortality.*

Evaluation of Existing Scoring System Performance

Discrimination of existing scoring systems and the final SBG algorithms was assessed by the c-statistic on the overall study population (n = 732). For 90-d mortality, the c-statistics were: original Tokuhashi = 0.72, revised Tokuhashi = 0.74, Tomita = 0.70, modified Bauer = 0.71, Van der Linden = 0.68, Katagiri = 0.74, New England Spine Metastasis Score = 0.75, SORG Classic = 0.73, SORG nomogram = 0.74, Bollen = 0.67. For 90-d mortality, discrimination of the SGB model (c = 0.88) was greater than the New England Spine Metastasis Score (c = 0.75), $P < .001$.

For 1-yr mortality, the c-statistics were: original Tokuhashi = 0.73, revised Tokuhashi = 0.74, Tomita = 0.71, modified Bauer = 0.73, Van der Linden = 0.65, Katagiri = 0.76, New England Spine Metastasis Score = 0.75, SORG Classic = 0.76, SORG nomogram = 0.78, Bollen = 0.65. For 1-yr mortality, discrimination of the SGB model (c = 0.89) was greater than the SORG nomogram (c = 0.78), $P < .001$.

DISCUSSION

Five models were developed for prediction of 90-d and 1-yr mortality after surgery for spinal metastatic disease. The SGB model achieved superior performance on both cross-validation of the training set and testing in the independent holdout set. This

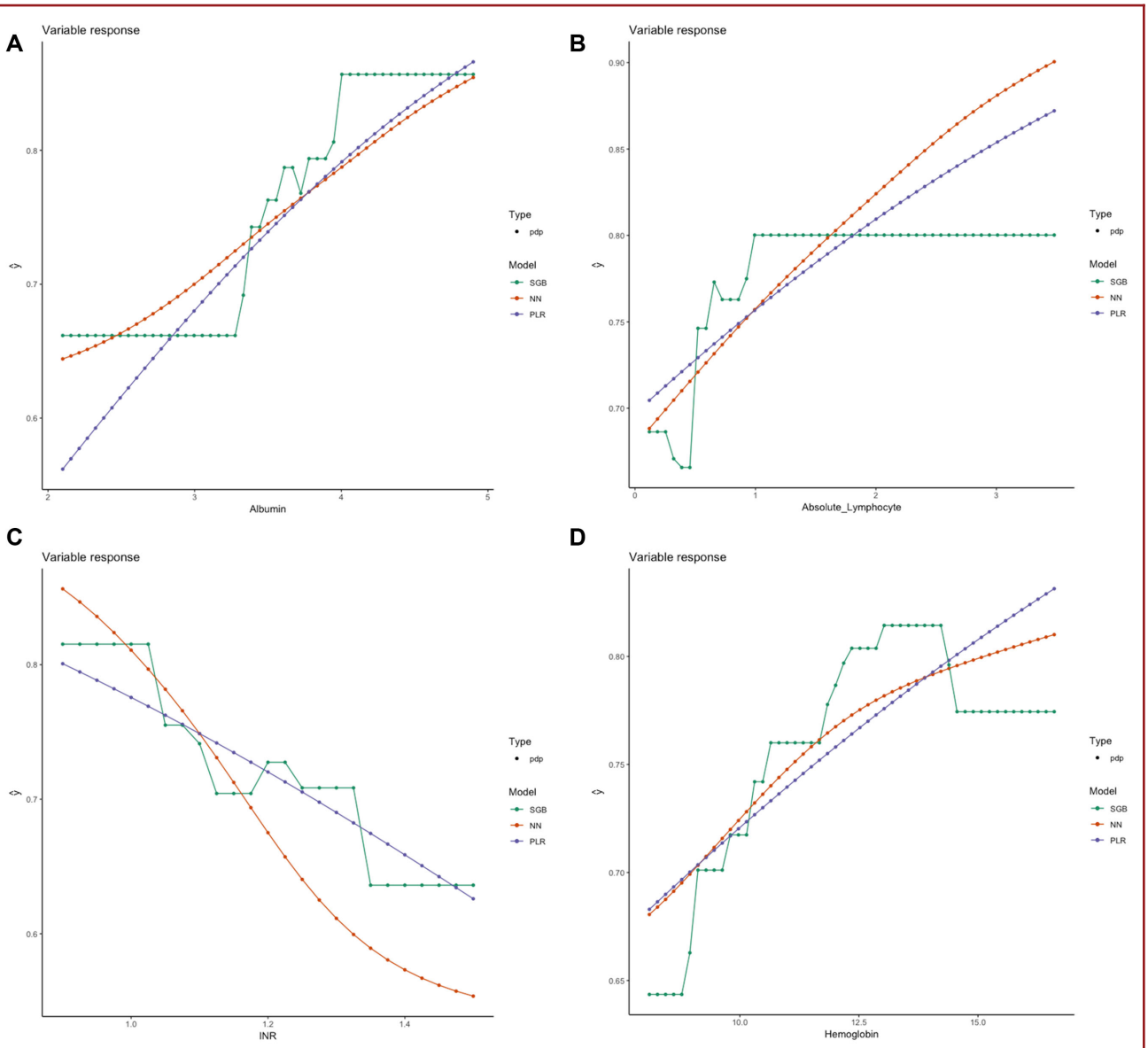

FIGURE 4. **A–D**, *Partial dependence plots for* **A**, *albumin,* **B**, *absolute lymphocyte,* **C**, *international normalized ratio (INR), and* **D**, *hemoglobin. These plots show the relationship between the input variables (albumin, absolute lymphocyte, INR, and hemoglobin) and the outputs of the machine learning algorithms for 90-d mortality over the range of the input variables. Abbreviations: NN: neural network; PDP: partial dependence plot; PLR: penalized logistic regression; SGM: stochastic gradient boosting.*

model was incorporated into an open access application with the ability to provide both predictions and explanations of results for individual patients.

Interpretation

Ahmed et al[3] recently compared 9 scoring for survival prediction in spinal metastatic disease (SORG Classic Scoring Algorithm, SORG nomogram, Tokuhashi, revised Tokuhashi, Tomita, Bauer, modified Bauer, Katagiri, and van der Linden). The highest discrimination for 30-d mortality was encountered for the SORG nomogram. In comparison, the highest c-statistic achieved for 90-d mortality and 1-yr mortality were the SORG nomogram and Tokuhashi score, respectively. Our analysis of the baseline performance of existing scoring systems on our study population yielded similar results. The SORG nomogram includes age, ECOG performance status, primary tumor histology (good and poor prognosis groupings), presence of 2 or more spine metastases, visceral metastases, brain

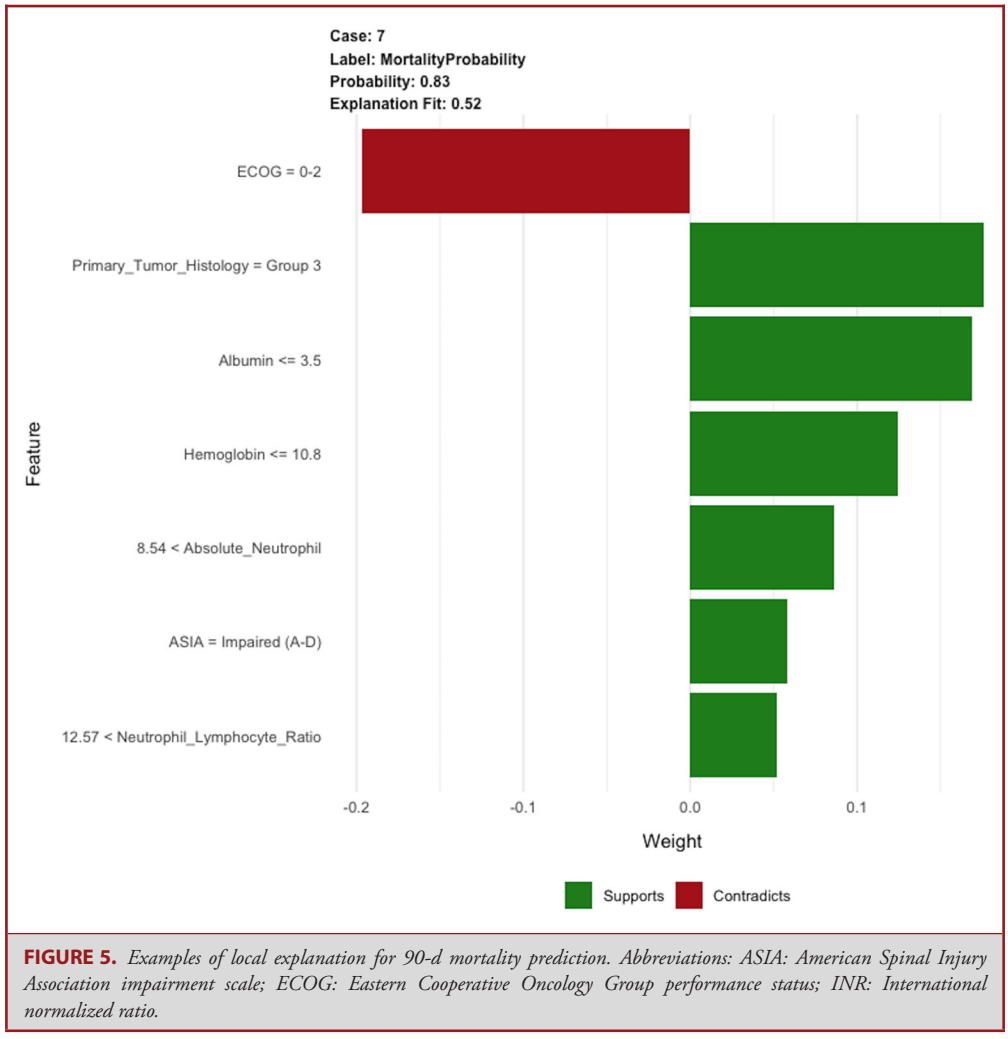

FIGURE 5. *Examples of local explanation for 90-d mortality prediction. Abbreviations: ASIA: American Spinal Injury Association impairment scale; ECOG: Eastern Cooperative Oncology Group performance status; INR: International normalized ratio.*

metastases, previous systemic therapy, white blood cell count, and hemoglobin.[9] The Tokuhashi score includes Karnofsky performance status, number of extraspinal bone metastatic foci, number of metastases in vertebral body, presence and operability of metastases to internal organs, primary tumor histology, and neurological deficit (Frankel score).[5,6] Ghori et al[30] studied 1-yr mortality after surgical intervention for spinal metastatic disease and determined that ambulatory status and preoperative albumin improved prediction relative to the modified Bauer score alone.[32] This study built on these prior scoring systems by incorporation of additional factors (alkaline phosphatase, neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio) identified as important prognostic factors in metastatic disease.[13,17,30,33,34] Previous studies have also neglected to assess the calibration of their models, in contrast to this effort which intensively assessed and demonstrated good performance of the SGB model in terms of both calibration and overall performance (Brier score).[25]

Limitations

There are several limitations to the work presented here. Firstly, this was a retrospective study and prospective validation of the models presented here remains to be evaluated. The patients included in this analysis were from 1 geographic region and treated at 2 academic medical centers that share a parent healthcare corporation, university affiliation, and many residency and fellowship programs. There is clear potential for homogeneity in terms of clinical decision making, surgical practice, and clustering along with selection and indication bias. The predictive performance of existing scoring systems was calculated as a reference point but the potential for the biases highlighted above limits direct comparison of the machine learning algorithms developed here to these previous scoring systems. As a result, the model findings presented here require further validation in an independent, geographically distinct and ideally, multi-institutional, sample of patients. To the extent that practice

in other clinical contexts differs substantively from that of our centers, the performance of the SGB model may be altered. This is clearly an important avenue for future research. As all patients included in this analysis were operatively managed, the findings are not applicable to survival in patients who receive only palliative therapy, radiation, or other nonoperative treatment regimens. Furthermore, there are additional factors such as novel biologic and targeted therapies that have been developed over the time period of this study. Future multicenter studies should seek to curate registries of patients by individual primary tumor histology with both sufficient volume and granular data in order to further evaluate the potential for preoperative prognostication on the basis of these therapies. The prospect for survival is but one factor considered in the determination of operative intervention, and future studies should look to examine other outcomes including postoperative ambulatory status, complications, reoperations, and unplanned readmissions.

Implications

Nonetheless, we believe that the results remain useful in light of the potential for enhanced predictive capacity for intermediate and long-term mortality in spinal metastatic disease within the SGB model and the fact that we have made this utility available for clinical application. After external validation of these algorithms, preoperative prediction of survival could be used to inform timing of surgery, invasiveness of intervention, expected postoperative course, consideration of radiotherapy, medical management, palliative therapy, and, most important, multidisciplinary shared decision making with patients around expectations and preferences. There are previous examples of studies in the literature that have explored machine learning models for survival prediction in oncology.[35,36] However, none of the existing efforts in spinal oncology have overcome the challenges of explaining machine learning models while simultaneously making them openly available to healthcare professionals. One milestone achieved by this study was to provide global and local explanations of the output of these complex algorithms. Explanations of predictions are crucial for practicing clinicians and patients to comprehend the determinations and necessary for application on a case-specific level in clinical practice.

This study relied on structured data collected from electronic health records. It is also possible to embed algorithms that automate the extraction of this structured data from patient health records and automate the generation of predictions and explanations.[37] Alternatively, algorithms may be developed that are able to predict meaningful outcomes for patients with spinal metastasis on the basis of unstructured data (free text, imaging, digital phenotyping).[38,39] Regardless of the approach, explanations of the predictions and the evolution of data science are likely to enhance patient care and shared decision making for spinal metastatic disease in the years to come.

CONCLUSION

Preoperative estimation of 90-d and 1-yr mortality was achieved with assessment of more flexible modeling techniques such as machine learning. Integration of these models into applications that can be easily incorporated into clinical practice represents the next step in this process and one which has the potential to transform treatment recommendations and shared decision making for patients with spinal metastases.

Disclosures

The authors have no personal, financial, or institutional interest in any of the drugs, materials, or devices described in this article.

REFERENCES

1. Kelly ML, Kshettry VR, Rosenbaum BP, Seicean A, Weil RJ. Effect of a randomized controlled trial on the surgical treatment of spinal metastasis, 2000 through 2010: a population-based cohort study. *Cancer*. 2014;120(6):901-908.
2. Patchell RA, Tibbs PA, Regine WF, et al. Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: a randomised trial. *Lancet (London, England)*. 2005;366(9486):643-648.
3. Ahmed AK, Goodwin CR, Heravi A, et al. Predicting survival for metastatic spine disease: a comparison of nine scoring systems. *Spine J*. 2018;18(10):1804-1814.
4. Sciubba DM, Petteys RJ, Dekutoski MB, et al. Diagnosis and management of metastatic spine disease. *J Neurosurg Spine*. 2010;13(1):94-108.
5. Tokuhashi Y, Matsuzaki H, Oda H, Oshima M, Ryu J. A revised scoring system for preoperative evaluation of metastatic spine tumor prognosis. *Spine*. 2005;30(19):2186-2191.
6. Tokuhashi Y, Matsuzaki H, Toriyama S, Kawano H, Ohsaka S. Scoring system for the preoperative evaluation of metastatic spine tumor prognosis. *Spine*. 1990;15(11):1110-1113.
7. Tomita K, Kawahara N, Kobayashi T, Yoshida A, Murakami H, Akamaru T. Surgical strategy for spinal metastases. *Spine*. 2001;26(3):298-306.
8. Ghori AK, Leonard DA, Schoenfeld AJ, et al. Modeling 1-year survival after surgery on the metastatic spine. *Spine J*. 2015;15(11):2345-2350.
9. Paulino Pereira NR, McLaughlin L, Janssen SJ, et al. The SORG nomogram accurately predicts 3- and 12-months survival for operable spine metastatic disease: external validation. *J Surg Oncol*. 2017;115(8):1019-1027.
10. van der Linden YM, Dijkstra SP, Vonk EJ, Marijnen CA, Leer JWH. Dutch Bone Metastasis Study Group. Prediction of survival in patients with metastases in the spinal column. *Cancer*. 2005;103(2):320-328.
11. Bauer HC, Wedin R. Survival after surgery for spinal and extremity metastases: prognostication in 241 patients. *Acta Orthop Scand*. 1995;66(2):143-146.
12. Pereira NRP, Janssen SJ, van Dijk E, et al. Development of a prognostic survival algorithm for patients with metastatic spine disease. *J Bone Joint Surg Am*. 2016;98(21):1767-1776.
13. Schoenfeld AJ, Leonard DA, Saadat E, Bono CM, Harris MB, Ferrone ML. Predictors of 30- and 90-day survival following surgical intervention for spinal metastases. *Spine (Phila Pa 1976)*. 2016;41(8):E503-E509.
14. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med*. 2015;13(1):1.
15. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18(12).
16. Quan H, Li B, Couris CM, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol*. 2011;173(6):676-682.
17. Katagiri H, Okada R, Takagi T, et al. New prognostic factors and scoring system for patients with skeletal metastasis. *Cancer Med*. 2014;3(5):1359-1367.
18. Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-118.

19. Kim JH. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computat Stat Data Analysis*. 2009;53(11):3735-3745.
20. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32.
21. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*. Vol. 1: Springer Series in Statistics. Springer: New York; 2001.
22. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Vol 112. New York: Springer; 2013.
23. Wainer J. Comparison of 14 different families of classification algorithms on 115 binary datasets. arXiv:1606.00930. 2016.
24. Steyerberg EW, Van Calster B, Pencina MJ. Performance measures for prediction models and markers: evaluation of predictions and classifications. *Rev Esp Cardiol*. 2011;64(9):788-794.
25. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925-1931.
26. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Wea Rev*. 1950;78(1):1-3.
27. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574.
28. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.
29. Bollen L, van der Linden YM, Pondaag W, et al. Prognostic factors associated with survival in patients with symptomatic spinal bone metastases: a retrospective cohort study of 1 043 patients. *Neuro Oncol*. 2014;16(7):991-998.
30. Ghori AK, Leonard DA, Schoenfeld AJ, et al. Modeling 1-year survival after surgery on the metastatic spine. *Spine J*. 2015;15(11):2345-2350.
31. Katagiri H, Takahashi M, Wakai K, Sugiura H, Kataoka T, Nakanishi K. Prognostic factors and a scoring system for patients with skeletal metastasis. *J Bone Joint Surg Br*. 2005;87(5):698-703.
32. Schoenfeld AJ, Le HV, Marjoua Y, et al. Assessing the utility of a clinical prediction score regarding 30-day morbidity and mortality following metastatic spinal surgery: the New England Spinal Metastasis Score (NESMS). *Spine J*. 2016;16(4):482-490.
33. Thio Q, Goudriaan WA, Janssen SJ, et al. Prognostic role of neutrophil-to-lymphocyte ratio and platelet-to-lymphocyte ratio in patients with bone metastases. *Br J Cancer*. 2018;119(6):737-743.
34. Karhade AV, Thio Q, Ogink PT, Schwab JH. Serum alkaline phosphatase and 30-day mortality after surgery for spinal metastatic disease. *J Neurooncol*. 2018;140(1):165-171.
35. Forsberg JA, Eberhardt J, Boland PJ, Wedin R, Healey JH. Estimating survival in patients with operable skeletal metastases: an application of a bayesian belief network. *PLoS ONE*. 2011;6(5):e19956.
36. Forsberg JA, Wedin R, Boland PJ, Healey JH. Can we estimate short- and intermediate-term survival in patients undergoing surgery for metastatic bone disease? *Clin Orthop Relat Res*. 2017;475(4):1252-1261.
37. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk. *Ann Surg*. 2018.
38. Elfiky AA, Pany MJ, Parikh RB, Obermeyer Z. Development and application of a machine learning approach to assess short-term mortality risk among patients with cancer starting chemotherapy. *JAMA Network Open*. 2018;1(3):e180926.
39. Karhade AV, Gormley WB, Smith TR. Improving outcomes: big data and predictive analytics. In: Guillaume DJ, Hunt MA, ed. *Quality and Safety in Neurosurgery*. San Diego: Elsevier; 2018:205-212.

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

Supplemental Digital Content. Appendix. 3 Tables and Methods. The Supplemental Digital Content expands on the Results & Methods provided. Table S1: Primary tumor histology, n = 732. Table S2: Histology groupings. Table S3: Discrimination and calibration of algorithms in training set, n = 587. Additional detail of methodology.