Clin Orthop Relat Res (2020) 478:322-333
DOI 10.1097/CORR.0000000000000997

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Clinical Research

Development and Internal Validation of Machine Learning Algorithms for Preoperative Survival Prediction of Extremity Metastatic Disease

Quirina C. B. S. Thio MD, Aditya V. Karhade BE, Bas Bindels BSc, Paul T. Ogink MD, Jos A. M. Bramer MD, PhD, Marco L. Ferrone MD, Santiago Lozano Calderón MD, PhD, Kevin A. Raskin MD, Joseph H. Schwab MD, MS

Received: 7 May 2019 / Accepted: 26 September 2019 / Published online: 18 October 2019
Copyright © 2019 by the Association of Bone and Joint Surgeons

Abstract

Background A preoperative estimation of survival is critical for deciding on the operative management of metastatic bone disease of the extremities. Several tools have been developed for this purpose, but there is room for improvement. Machine learning is an increasingly popular and flexible method of prediction model building based on a data set. It raises some skepticism, however, because of the complex structure of these models.

Questions/purposes The purposes of this study were (1) to develop machine learning algorithms for 90-day and 1-year survival in patients who received surgical treatment for a bone metastasis of the extremity, and (2) to use these algorithms to identify those clinical factors (demographic, treatment related, or surgical) that are most closely associated with survival after surgery in these patients.

Methods All 1090 patients who underwent surgical treatment for a long-bone metastasis at two institutions between 1999 and 2017 were included in this retrospective study. The median age of the patients in the cohort was 63 years (interquartile range [IQR] 54 to 72 years), 56% of patients (610 of 1090) were female, and the median BMI was 27 kg/m^2 (IQR 23 to 30 kg/m^2). The most affected

Each author certifies that neither he or she, nor any member of his or her immediate family, has funding or commercial associations (consultancies, stock ownership, equity interest, patent/licensing arrangements, etc.) that might pose a conflict of interest in connection with the submitted article.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research®* editors and board members are on file with the publication and can be viewed on request.
Clinical Orthopaedics and Related Research® neither advocates nor endorses the use of any treatment, drug, or device. Readers are encouraged to always seek additional information, including FDA approval status, of any drug or device before clinical use.
Each author certifies that his or her institution approved the human protocol for this investigation and that all investigations were conducted in conformity with ethical principles of research.
This work was performed at Massachusetts General Hospital, Boston, MA, USA.

Q. C. B. S. Thio, A. V. Karhade, B. Bindels, P. T. Ogink, S. L. Calderón, K. A. Raskin, J. H. Schwab, Department of Orthopedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

J. A. M. Bramer, Department of Orthopedic Surgery, Academic University Medical Center, University of Amsterdam, Amsterdam, the Netherlands

M. L. Ferrone, Department of Orthopedic Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Q. C. B. S. Thio (✉), Room 3.946, Yawkey Building, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114 USA, Email: quirina.thio@gmail.com

Wolters Kluwer

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

location was the femur (70%), followed by the humerus (22%). The most common primary tumors were breast (24%) and lung (23%). Intramedullary nailing was the most commonly performed type of surgery (58%), followed by endoprosthetic reconstruction (22%), and plate screw fixation (14%). Missing data were imputed using the missForest methods. Features were selected by random forest algorithms, and five different models were developed on the training set (80% of the data): stochastic gradient boosting, random forest, support vector machine, neural network, and penalized logistic regression. These models were chosen as a result of their classification capability in binary datasets. Model performance was assessed on both the training set and the validation set (20% of the data) by discrimination, calibration, and overall performance.

Results We found no differences among the five models for discrimination, with an area under the curve ranging from 0.86 to 0.87. All models were well calibrated, with intercepts ranging from -0.03 to 0.08 and slopes ranging from 1.03 to 1.12. Brier scores ranged from 0.13 to 0.14. The stochastic gradient boosting model was chosen to be deployed as freely available web-based application and explanations on both a global and an individual level were provided. For 90-day survival, the three most important factors associated with poorer survivorship were lower albumin level, higher neutrophil-to-lymphocyte ratio, and rapid growth primary tumor. For 1-year survival, the three most important factors associated with poorer survivorship were lower albumin level, rapid growth primary tumor, and lower hemoglobin level.

Conclusions Although the final models must be externally validated, the algorithms showed good performance on internal validation. The final models have been incorporated into a freely accessible web application that can be found at https://sorg-apps.shinyapps.io/extremitymetssurvival/. Pending external validation, clinicians may use this tool to predict survival for their individual patients to help in shared treatment decision making.

Level of Evidence Level III, therapeutic study.

Introduction

The incidence of cancer grows annually; approximately 14 million patients were diagnosed with the disease in 2012 [11], and approximately 18 million were diagnosed in 2018 [10]. Simultaneously, the survival rates of patients with cancer have increased because of improved treatment options, including those for metastatic cancer. Three cancers with the highest incidence, namely prostate, lung, and breast cancer, have a high propensity to metastasize to bone [6, 25]. It is therefore expected that both the incidence and prevalence of metastatic bone disease will increase. Bone metastases can lead to pathologic fractures and can dramatically decrease a patient's quality of life, causing pain and immobility [7]. Because metastatic cancer is generally deemed incurable, treatment is intended to treat the symptoms and maintain quality of life. Treatment options for these patients include systemic therapy, radiotherapy, and surgery. Determination of operative management is influenced by estimated survival [35]. In patients with bone metastases of the extremities, two survival thresholds are generally considered important: 90 days and 1 year [12]. Patients who are not expected to live beyond 90 days usually will not benefit from surgery, while patients who live beyond 1 year will benefit from more-invasive and enduring reconstruction procedures [35].

Several prognostic models have been developed in the past decades to help physicians in their estimations for patients with bone metastases of the extremities [1, 12, 18, 19, 22, 36]. These models range from simple scoring systems to more complex machine learning algorithms. Although these models perform well, it is necessary to keep modifying and optimizing them, particularly because new prognostic factors are continuously being investigated and more-advanced machine learning techniques now can be used for both prediction and explanation. Machine learning is a subset of computer science and statistics. It is capable of handling large amounts of data and recognizing complex combinations of predictors for a certain outcome by using modern computational and mathematic algorithms [9, 23]. So far, a limited number of studies have used machine learning to develop prognostic models for patients with a bone metastasis of the extremities [12, 18], and only a few machine learning algorithms have been explored for that purpose. We recently explored different machine learning algorithms for the survival prediction of patients suffering from chondrosarcoma [4, 31], using a similar methodology for model development and performance assessment as the current study. One of the main disadvantages of machine learning algorithms is the so-called "black box problem"; we are able to observe the data we enter into the computer as well as the output the algorithms give us, but what happens in between sometimes is unclear. In this study, we aimed to address these drawbacks while preserving the predictive performance of the resulting algorithms.

Therefore, the primary purpose of this study was to develop machine learning algorithms for 90-day and 1-year survival in patients who received surgical treatment for a bone metastasis of the extremity. Our secondary aim was to use these algorithms to identify those clinical factors (demographic, treatment related, or surgical) that are most closely associated with survival after surgery in these patients.

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

Patients and Methods

Study Design and Population

This study was performed according to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis [8] and the Strengthening the Reporting of Observational Studies in Epidemiology [33] guidelines.

All 1090 patients 18 years and older who underwent surgical treatment for a long-bone metastasis at the Massachusetts General Hospital and the Brigham and Women's Hospital between 1999 and 2017 were included in this retrospective study. Surgical treatment consisted of intramedullary nailing, endoprosthetic reconstruction, plate screw fixation or dynamic hip screw. Patients were excluded if the first surgical treatment of the long-bone metastasis was not performed at one of these institutions. If a patient received multiple surgical procedures for a long-bone metastasis, only the first procedure was included. In general, the decision to operate or not was based on the health of the patient and the patient's wishes. Patients who were expected to have a long survival were treated with more durable and invasive procedures than patients who were not expected to have a long survival.

Explanatory Variables and Outcome

We reviewed operative notes, medical records, radiology reports, and pathology reports to record the following variables: age; sex; BMI (kg/m^2); any Charlson comorbidity in addition to metastatic cancer; primary tumor type [19]; the Eastern Cooperative Oncology Group score; tumor location; the presence of a pathologic fracture, other bone metastases, spine metastases, visceral metastases, brain metastases; and previous systemic therapy and local radiation. Preoperative laboratory factors were the hemoglobin level (g/dL), white blood cell count (x 10^3/ uL), platelet count (x 10^3/uL), absolute lymphocyte count (x 10^3/uL), absolute neutrophil count (x 10^3/uL), neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte count, albumin level (g/dL), and alkaline phosphatase (IU/L), calcium (mg/dL), creatinine (mg/dL), and sodium (mg/dL) levels.

The primary outcome of interest was overall survival, and this was verified from the medical records or the Social Security Death Index [17].

Missing Data

Missing data were imputed using the missForest methods [27]. This was performed for variables with less than 30% of missing data: BMI: 237 (22%); hemoglobin level: 146 (13%); white blood cell count: 146 (13%); platelet count: 146 (13%); absolute lymphocyte count: 326 (30%); absolute neutrophil count: 322 (30%); albumin level: 320 (30%); alkaline phosphatase level: 316 (29%); calcium level: 200 (18%); creatinine level: 66 (15%); and sodium level: 199 (18%). Variables with more than 30% of missing data were dropped.

Baseline Characteristics

The median patient age in the cohort was 63.0 years (interquartile range [IQR] 54.0 to 72.0 years), and 56% of patients (610 of 1090) were female (Table 1). The median BMI was 26.6 kg/m^2 (IQR 23.2 to 30.3 kg/m^2). As for the primary tumor category, 43% of the patients (460 of 1090) were in the slow-growth group, 34% (367 of 1090) were in the rapid-growth group, and 24% (263 of 1090) were in the moderate-growth group (see Table 1, Supplemental Digital Content 1, http://links.lww.com/CORR/A244). The primary tumor categories were classified as per Katagiri et al. [19] (see Table 2, Supplemental Digital Content 2, http://links.lww.com/CORR/A245). In all, 55% of patients (595 of 1090) had pathologic fractures. Eastern Cooperative Oncology Group scores were available for 39% of patients (422 of 1090), of whom 85% (360 of 422) had an Eastern Cooperative Oncology Group score of 0 to 2. The lower extremity was affected in 77% of the patients. Fifty-eight percent of patients (637 of 1090) were treated with intramedullary nailing, followed by endoprosthetic reconstruction in 22% of patients (241 of 1090), and plate-screw fixation in 14% (155 of 1090). Overall, 78% of patients (845 of 1090) had other bone metastases, and 57% (626 of 1090) had spinal metastases. Visceral metastases were present in 45% of patients (487 of 1090) and brain metastases occurred in 16% (175 of 1090). Thirty-eight patients were lost to follow-up within 90 days, sixty-nine were lost to follow-up within 1 year. Twenty-nine percent of the 1052 patients (305) died within 90 days, and 62% (1031) died within 1 year.

Model Development

Variables were selected for the algorithms by 10-fold cross-validation of random forest algorithms [5, 21]. Random forest algorithms repeatedly select random features to build models [5]. Tenfold cross-validation entails that the data is split into 10 groups; each model is trained in nine groups and tested in the tenth [21]. This process is repeated until every group has been used for testing. The combination of random forest algorithms and 10-fold cross-validation enables us to find the optimal subset of features while keeping the variance of model performance low and avoiding overfitting.

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

Table 1. Baseline characteristics of the study population

Variable	Number (n = 1090)
Age (years)[a]	63 (54-72)
Female sex	56% (610)
BMI (kg/m^2)[a]	27 (23-30)
Other Charlson comorbidity	54% (584)
Primary tumor type	
Slow growth	42% (460)
Moderate growth	24% (263)
Rapid growth	34% (367)
Pathologic fracture	55% (594)
ECOG	
0-2	85% (360)
3-4	15% (62)
Tumor location	
Upper extremity	23% (255)
Lower extremity	77% (835)
Other bone metastases	78% (845)
Spine metastases	57% (626)
Visceral metastases	45% (487)
Brain metastases	16% (175)
Previous systemic therapy	62% (676)
Local radiation	18% (194)
Hemoglobin level (g/dL)[a]	11 (10-13)
White blood cell count (10^3/uL)[a]	7 (5-10)
Platelet count (10^3/uL) [a]	251 (184-332)
Absolute lymphocyte count (10^3/uL)[a]	1 (1-2)
Absolute neutrophil count (10^3/uL)[a]	5 (4-8)
Neutrophil-to-lymphocyte ratio[a]	5 (3-9
Platelet-to-lymphocyte ratio [a]	234 (158-374)
Albumin level (g/dL)[a]	4 (3-4)
Alkaline phosphatase level (IU/L)[a]	101 (74-146)
Calcium (mg/dL)[a]	9 (9-10)
Creatinine (mg/dL)[a]	0.8 (0.7-1.1)
Sodium (mg/dL)[a]	138 (136-140)
90-day mortality	29% (305)
1-year mortality	62% (639)

[a]Data are presented as median (range).
ECOG = Eastern Cooperative Oncology Group performance status score.

The algorithms chosen for this study, stochastic gradient boosting, random forest, support vector machine, neural network, and penalized logistic regression, were based on a previous study's method [34]. They are commonly used algorithms and are suitable for binary classification. All five algorithms have their separate way of classifying the data. Both stochastic gradient boosting and random forest algorithms are tree-based, in which outputs of individual decision trees are combined. For stochastic gradient boosting models, these individual trees are developed sequentially. Each tree "learns" from the previous tree, and redistributes the weight of the accurately and wrongly classified data. By giving less weight to the accurately classified data and more weight to the wrongly classified data, each tree will improve after reclassifying the data until further improvement is not possible. All trees combined will then give the final prediction. In random forest models each tree is independently developed with random feature subsets. All these different trees will then "vote" to form the final prediction model. The fundamental concept behind this is the wisdom of the crowds: all the uncorrelated models combined will outperform a single model. Support vector machines are kernel-based algorithms looking to define a hyperplane that best divides the dataset into two classes [16]. Simply said, a hyperplane can be imagined as a line that separates and classifies a set of data. The further the data points are from the hyperplane, the more certain it is they are correctly classified. Support vector machines classify data into higher dimensions, by which the line becomes three-dimensional and is no longer a line but a plane. Neural networks are modeled after the human brain and consist of input and output layers that are connected with a certain weight via one or multiple hidden layers. They are capable of recognizing patterns from the data and learn from it. Penalized logistic regression models impose a "penalty" to a logistic models for having too many variables. There are two ways to do that: With ridge regression all the predictors are kept but the coefficients of minor predictors are lowered close to zero. Lasso regression eliminates the minor predictors by setting their coefficients to 0. The elastic-net penalized logistic regression combines ridge regression and lasso regression to find a reduced set of variables for an optimal performing model.

The data were divided into a training set (80%) and a validation set (20%). The training set was used to develop the models, while the validation set was used to internally validate the models.

Assessment of Model Performance

After the model was developed, its performance was assessed on both the training set by means of 10-fold cross-validation and the validation set. Performance metrics included discrimination (area under the curve), calibration (intercept and slope), and Brier score. The area under the curve ranged from 0.50 to 1.0, with 0.50 indicating pure chance and 1.0 indicating the highest discriminating score. Graphically, discrimination is visualized with receiver operating characteristic curve plots. Calibration indicates agreement between the predicted outcome and the actual outcome, and perfect calibration has an intercept of 0 and a slope of 1 [28, 29]. The Brier score refers to overall performance, with 0 as a perfect Brier score. However, the prevalence of the outcome must be considered; therefore,

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

Table 2. Discrimination and calibration of algorithms on repeated cross-validation of the training set, n = 873, mean (95% CI)

Metric	Stochastic gradient boosting	Random forest	Support vector machine	Neural network	Penalized logistic regression
90-day mortality					
AUC	0.87 (0.86 to 0.88)	0.86 (0.85 to 0.88)	0.86 (0.84 to 0.87	0.87 (0.84 to 0.87)	0.86 (0.85 to 0.87)
Intercept	0.01 (-0.06 to 0.08)	0.01 (-0.06 to 0.07)	0.08 (-0.03 to 0.20)	-0.03 (-0.10 to 0.04)	0.04 (-0.05 to 0.13)
Slope	1.04 (0.96 to 1.12)	1.12 (1.01 to 1.23)	1.13 (1.00 to 1.27)	1.03 (0.94 to 1.11)	1.08 (0.97 to 1.20)
Brier[a]	0.13 (0.12 to 0.14)	0.13 (0.13 to 0.14)	0.14 (0.13 to 0.14)	0.14 (0.13 to 0.15)	0.14 (0.13 to 0.14)
1-year mortality					
AUC	0.85 (0.83 to 0.86)	0.85 (0.83 to 0.86)	0.85 (0.83 to 0.86)	0.85 (0.84 to 0.86)	0.85 (0.83 to 0.86)
Intercept	-0.04 (-0.12 to 0.03)	-0.12 (-0.19 to 0.04)	-0.03 (-0.10 to 0.05)	0.05 (-0.02 to 0.13)	0.02 (-0.05 to 0.09)
Slope	1.12 (1.02 to 1.21)	1.41 (1.29 to 1.53)	1.16 (1.03 to 1.28)	0.87 (0.81 to 0.94)	0.94 (0.84 to 1.05)
Brier[b]	0.16 (0.15 to 0.16)cccc	0.16 (0.15 to 0.16)	0.15 (0.15 to 0.16)	0.15 (0.15 to 0.16)	0.15 (0.15 to 0.16)

[a]90-day mortality null-model Brier score = 0.21.
[b]1-year mortality null-model Brier score = 0.24; AUC = area under the receiver operating characteristic curve.

the Brier score of the null model was also calculated by assigning a probability equal to the prevalence of the outcome to each patient [3, 29].

Decision curves were then plotted for the 90-day and 1-year prediction models. A decision curve analysis is a way of evaluating the net benefit of a model across a range of different threshold probabilities [32]. The user of the model can decide which threshold is important and determine if the model is valuable at that threshold and see what the predicted net benefit would be.

Model Explanations

The final prediction models were explained by visualizing the included features of the models with their weighted importance [15]. These plots give a global estimation of the models. Partial-dependence plots were created to reflect the association between continuous variables and the outcome [2]. Additionally, for individual patients, an explanation of the contribution of the different features to the outcome was given and shown with illustrative examples.

Internet Application

The stochastic gradient boosting model was chosen as the final model for both 90-day and 1-year survival prediction. These models were deployed as a freely accessible internet application and can be found at https://sorg-apps.shinyapps.io/extremitymetssurvival/. Anaconda Distribution (Continuum Analytics, Austin, TX, USA) with RStudio (Version 1.0.153, Boston, MA, USA), Python Version 3.6 (Python Software Foundation, Wilmington, DE, USA), and StataCorp 2013 (Stata Statistical Software: Release 13; StataCorp LP, College Station, TX, USA) were used for analyzing data, creating the model, and developing the internet application.

Results

Development and Performance of the Machine Learning Algorithms

The factors associated with 90-day survival were albumin level, neutrophil-to-lymphocyte ratio, primary tumor group, alkaline phosphatase level, hemoglobin level, calcium level, absolute neutrophil count, white blood cell count, age, and platelet count. The variables selected for 1-year survival were albumin level, primary tumor type, hemoglobin level, neutrophil-to-lymphocyte ratio, alkaline phosphatase level, absolute lymphocyte count, presence of visceral metastases, sodium level, platelet-to-lymphocyte ratio, and age.

The five models showed no difference in discrimination, with an area under the curve of 0.86 for the random forest (95% CI 0.85 to 0.88), support vector machine (95% CI 0.84 to 0.87), and penalized logistic regression models (95% CI 0.85 to 0.87) and an area under the curve of 0.87 for stochastic gradient boosting (95% CI 0.86 to 0.88) and the neural network models (95% CI 0.84 to 0.87) (Table 2). No difference was found in calibration between the models, with intercepts ranging from -0.03 to 0.08 and slopes ranging from 1.03 to 1.12. Brier scores were 0.13 for the stochastic gradient boosting and random forest models and 0.14 for the support vector machine, neural network, and penalized logistic regression models. The null-model Brier score was 0.21. In the validation set, the discriminating

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

performance ranged from an area under the curve of 0.85 for support vector machine and neural network to an area under the curve of 0.87 for stochastic gradient boosting (Table 3). Calibration intercepts ranged from 0.01 to 0.13 and calibration slopes ranged from 1.02 to 1.09. Brier scores for all five models were 0.13 compared with the null-model Brier score of 0.21.

Clinical Factors Associated with Survival

The most important factors associated with a greater risk of 90-day mortality were lower albumin level, higher neutrophil-to-lymphocyte ratio, and rapid growth primary tumor (Fig. 1A-D). Scaled from 0 to 100, the relative importance of albumin level was 100, of neutrophil-to-lymphocyte ratio around 75 and of primary tumor category around 40. For all predicted probabilities, the model showed greater standardized net benefit relative to changes in management decision based on all patients or no patients (Fig. 2). Partial dependence plots for the continuous variables of albumin level, neutrophil-to-lymphocyte ratio, calcium level, and hemoglobin level show the relationships between the input variables and the algorithm outputs for the different models (Fig. 3). The variable input is shown on the x-axis while the algorithm output is shown on the y-axis. The ability of the stochastic gradient boosting model to display non-linear relationships between the variables and the predicted probability compared with the neural network and penalized logistic regression models was noticeable. In those two models, higher albumin and hemoglobin levels were linearly associated with a higher predicted probability, while a higher neutrophil-to-lymphocyte ratio and calcium level were linearly associated with lower predicted probabilities. The stochastic gradient boosting model showed that levels under or above certain thresholds did not affect the predicted probability. For instance, for the albumin level, between 3 g/dL and 4 g/dL, an increase in the predicted probability was seen, while above and below that level, there was a plateau. Similar relationships were observed for the hemoglobin level.

The most important factors associated with a greater risk of 1-year mortality for the stochastic gradient boosting model were lower albumin level, rapid growth primary tumor, and lower hemoglobin level (Fig. 1). The relative importance of albumin level was 100, for primary tumor category it was around 80 and for hemoglobin level it was around 70. For all predicted probabilities, the model showed greater standardized net benefit relative to change in management decision based on all patients or no patients (Fig. 2A-D).

An example of the 90-day mortality prediction shows which factors lead to a 90-day mortality probability of 49% (Fig. 4).

Discussion

In the past decades, different prognostic models ranging from classic scoring models to machine learning algorithms have been developed to predict mortality at different time points in patients who undergo surgical treatment of a bone metastasis of the extremity [1, 12, 18, 19, 22, 36]. Frequently updating and improving these models is important because new prognostic markers such as the neutrophil-to-lymphocyte ratio are continuously being identified [30]. More importantly, machine learning techniques are also improving, with recent advances in the ability to explain the transformation function that is applied to the inputs of the model to generate the outputs. This transparency allows for an increased understanding of the models while continuing to build models that can capture complex relationships between predictors. In this study we developed machine learning algorithms to estimate survival in patients with a metastasis of the extremity. The models

Table 3. Discrimination and calibration of algorithms in the holdout set (n = 217)

Metric	Stochastic gradient boosting	Random forest	Support vector machine	Neural network	Penalized logistic regression
90-day mortality					
AUC	0.87	0.86	0.85	0.85	0.86
Intercept	0.06	0.02	0.13	0.01	0.06
Slope	1.03	1.08	1.09	1.02	1.03
Brier[a]	0.13	0.13	0.13	0.13	0.13
1-year mortality					
AUC	0.81	0.81	0.80	0.80	0.79
Intercept	0.09	-0.01	0.05	0.09	0.08
Slope	0.85	1.10	0.81	0.69	0.73
Brier[b]	0.18	0.17	0.18	0.18	0.18

[a]90-day mortality null-model Brier score = 0.21.
[b]1-year mortality null-model Brier score = 0.24; AUC = area under the receiver operating characteristic curve.

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

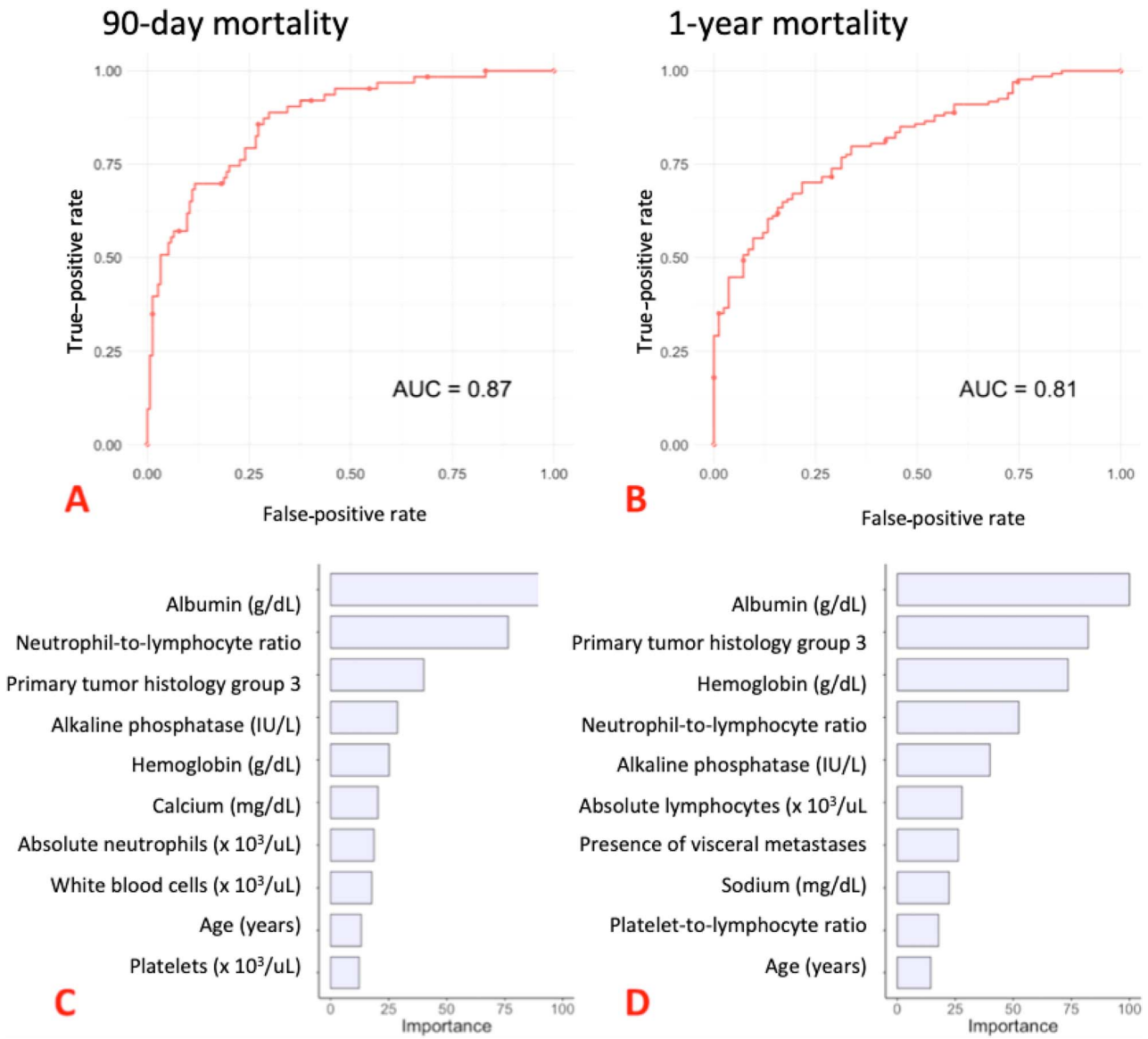

Fig. 1 A-D This image shows receiver operating characteristic curves for stochastic gradient boosting for (**A**) 90-day and (**B**) 1-year survival and overall feature importance for (**C**) 90-day and (**D**) 1-year survival. It represents the capability of the model of distinguishing between classes; AUC = area under the curve.

showed good performance and can be accessed through: https://sorg-apps.shinyapps.io/extremitymetssurvival/. Global and individual explanations are provided there.

Limitations

This study has several limitations. First, the developed models have not been externally validated. External validation is a crucial step in using the models in daily practice and this is an avenue for future research. However, we did validate our models internally with a validation set containing data that was not used for the model development. Second, this is a retrospective study and only patients who were surgically treated were included. Prospective validation is needed to assess the validity of the models. We were unable to include the Eastern Cooperative Oncology Group or Karnofsky scores because most patients did not have these scores recorded preoperatively. These performance scores have been used in many previous models [19, 20, 22, 24, 26, 36, 37]. Future studies should aim to include these factors and determine if they improve algorithm performance. Fourth, the patients in this study were from one geographic area in the United States. The algorithms may therefore apply mainly to patients in urban areas in the United States and perhaps Western Europe, where decisions to surgically treat these patients are

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

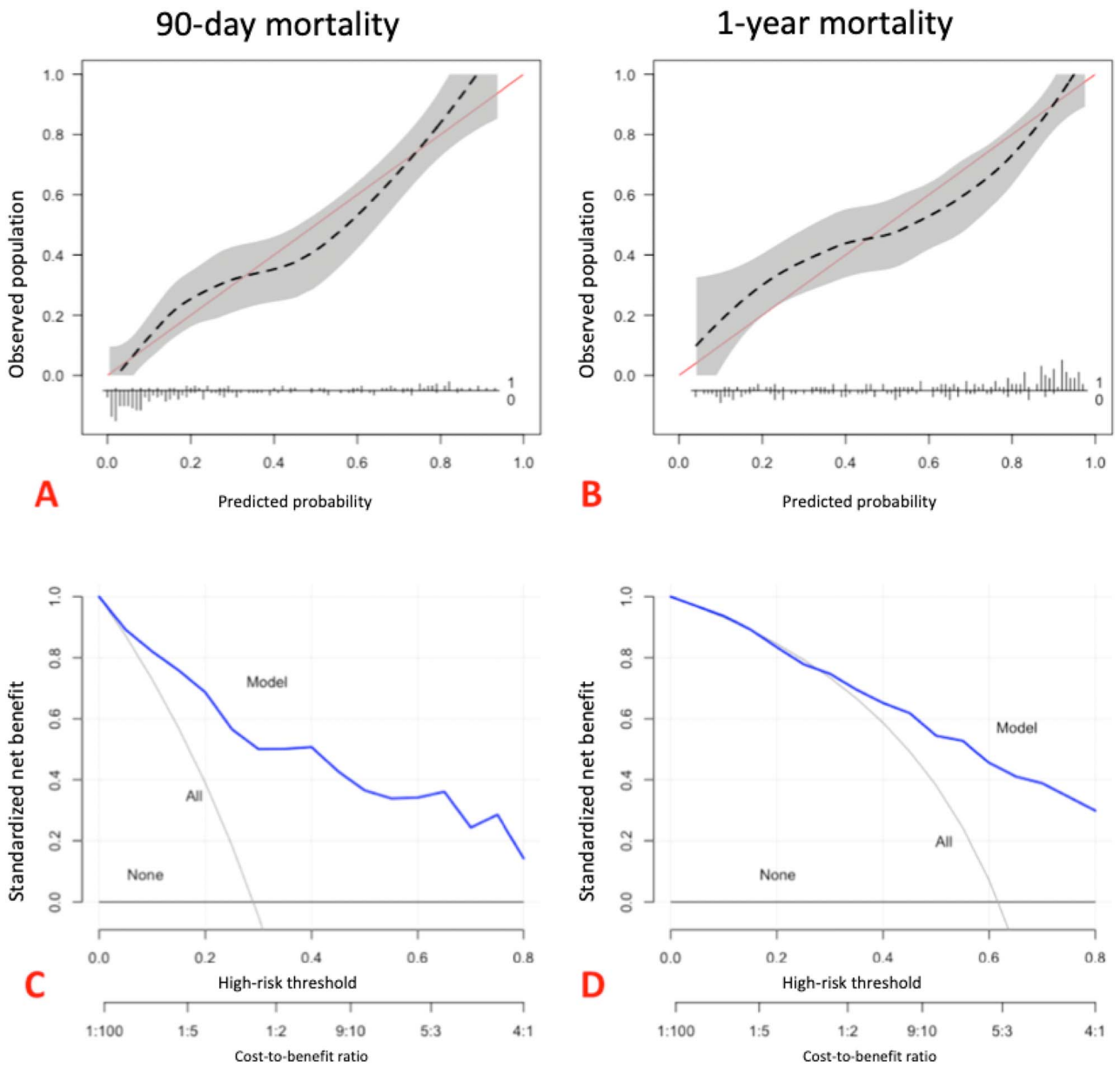

Fig. 2 Calibration plots for stochastic gradient boosting with 95% confidence intervals are shown for (**A**) 90-day and (**B**) 1-year survival and decision curve analysis plots are shown for (**C**) 90-day and (**D**) 1-year survival. The calibration plot visualizes how accurate the predictions are over different probabilities. The diagonal red line represents the optimal calibration; the closer the line of the model, the more accurate the prediction.

approached in a similar fashion and with similar healthcare support systems. It remains to be determined if the models have similar performance in different populations.

Development and Performance of the Machine Learning Algorithms

In this study, we developed prediction models for 90-day and 1-year survival in patients, using five different machine learning techniques. For 90-day survival, the following factors were included as predictors associated with a greater risk of death: lower albumin level, higher neutrophil-to-lymphocyte ratio, higher alkaline phosphatase level, lower hemoglobin level, higher calcium level, higher absolute neutrophil count, higher white blood cell count, higher age, and lower platelet count. For 1-year survival, those factors were lower albumin level, lower hemoglobin level, higher neutrophil-to-lymphocyte ratio, higher alkaline phosphatase level, lower absolute lymphocyte count, lower sodium level,

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

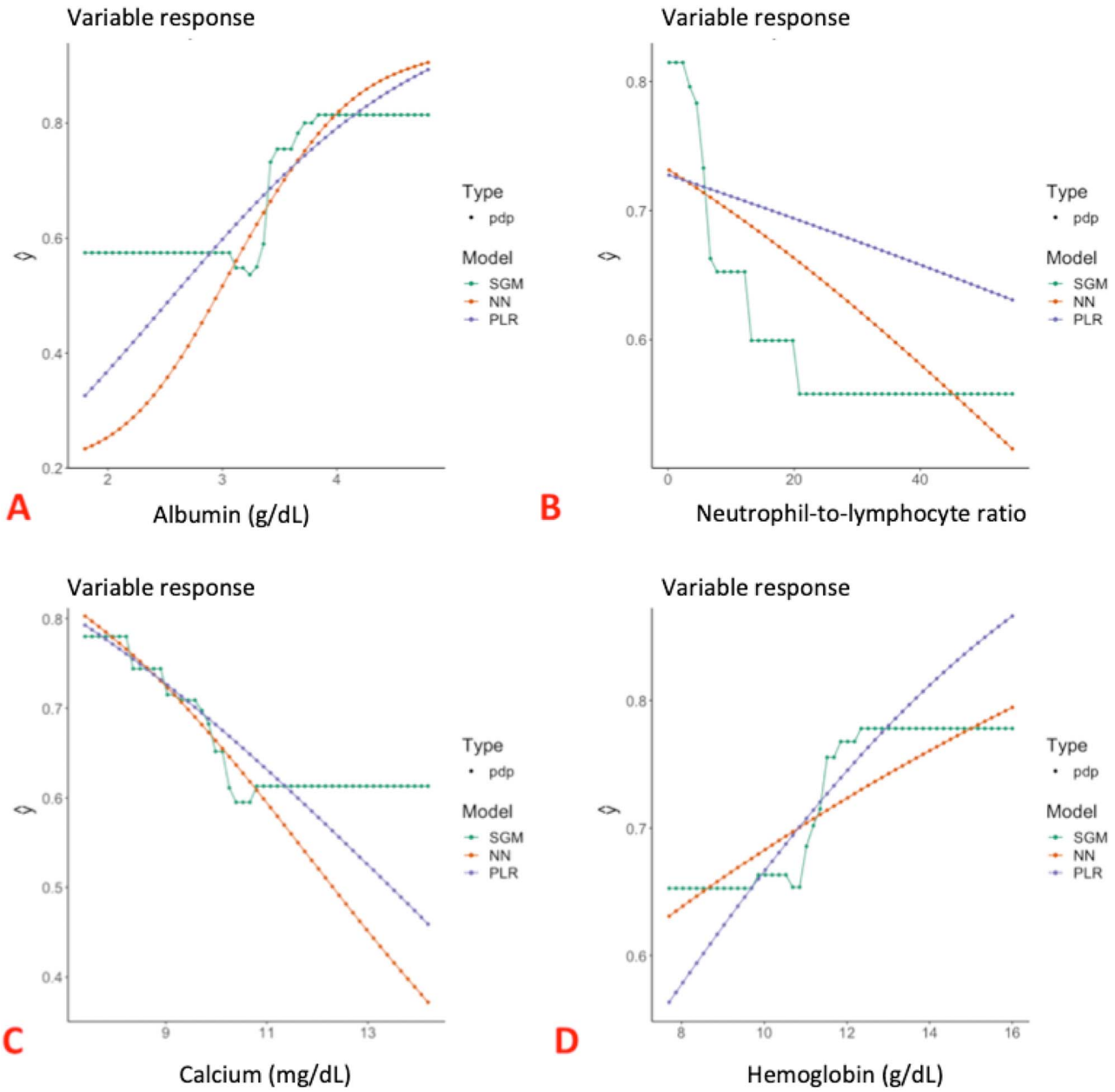

Fig. 3 A-D This figure shows partial dependence plots for 90-day mortality by model for (**A**) albumin level (g/dL), (**B**) neutrophil-to-lymphocyte ratio, (**C**) calcium level (mg/dL), and (**D**) hemoglobin level (g/dL). These plots show the relationship between input variables and the outputs of the different machine learning algorithms over the range of the input variables. For example, the albumin plot shows that the stochastic gradient boosting model has a constant output with albumin levels below 3 g/dL and levels above 4 g/dL. Between these levels there is a turning point in which the output rapidly increases. The neural networks model and the penalized logistic regression model show a more linear association between the input and the output; SGM = stochastic gradient boosting model, NN = neural networks model, PLR = penalized logistic regression model.

and higher platelet-to-lymphocyte ratio. We assessed predictive performance with discrimination, calibration, overall performance with the Brier score and a decision curve analysis. Assessing model performance is an important step in developing prediction models to determine the quality of the model [28]. Most previous studies [1, 19, 22, 24, 26, 37] did not report discrimination with c-statistics or areas under the curve, which makes it difficult to assess the discriminating capabilities of the models. An easy-to-use prognostic model was developed by means of a flow chart for 1520 patients with symptomatic bone metastases [36]. The authors reported a c-statistic of 0.70 but did not use no other

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

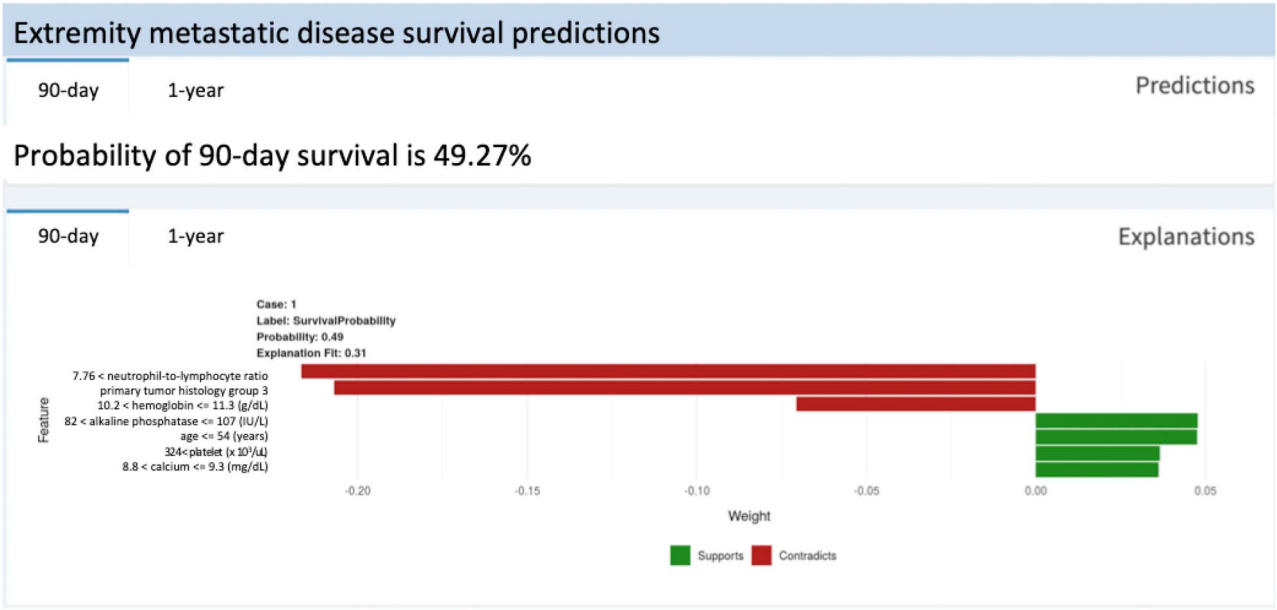

Fig. 4 An example of the 90-day survival prediction of a selected patient is shown here. This patient is a 54-year-old man with a rapid-growth primary tumor without visceral or brain metastases. He previously received systemic therapy. His laboratory values were as follows: hemoglobin level of 11.0 g/dL, platelet count of 375 x 103/uL, absolute lymphocyte count of 1.16 10^3/uL, absolute neutrophil count of 10.8 10^3/uL, creatinine of level of 2 mg/dL, white blood cell count of 8 10^3/uL, albumin level of 3.5 g/dL, alkaline phosphatase level of 89 IU/L, sodium level of 135 mg/dL, and calcium level of 9 mg/dL. Factors that support survival are visualized by the green bar. These are his alkaline phosphatase level, his age, his platelet count, and his calcium level. Factors that contradict survival are visualized by the red bars, which represent his neutrophil-to-lymphocyte ratio, his primary tumor histology, and his hemoglobin level. The prediction model shows a 90-day survival probability of 49%.

performance metrics such as calibration. Others developed a classic scoring algorithm, nomogram, and boosting algorithm for 927 patients surgically treated for a bone metastasis of the extremities and achieved areas under the curve on the testing set of 0.70, 0.75, and 0.75, respectively, at 90 days and 0.68, 0.73, and 0.72, respectively, at 1 year [18]. A c-statistic says something about the capability of a model to discriminate between the two outcomes (death or survival). It ranges from 0.5 (no discrimination, equal to a coin toss) to 1.0. The closer to 1.0, the better the discrimination. They also did not report using any other performance metrics. A Bayesian belief network model was developed by another research group for 189 patients who were treated for a bone metastasis of the extremities at 90 days and 1 year [12]. On external validation of a set of 815 patients, areas under the curve of 0.79 and 0.76 were achieved for 90 days and 1 year, respectively [13]. Again, no other metrics were described to assess performance.

Clinical Factors Associated with Survival

For the stochastic gradient boosting algorithm, which we used as the web-based application, the stochastic gradient boosting algorithm, the most important factors associated with a greater risk of 90-day mortality were lower albumin level, higher neutrophil-to-lymphocyte ratio, and rapid growth primary tumor. For 1-year mortality, the most important factors were lower albumin level, rapid growth primary tumor, and lower hemoglobin level. While primary tumor histology is incorporated in all previous models as important predictor [1, 12, 18, 19, 22, 36], most of them did not fully assess the laboratory factors identified as important predictors in the current study. Some studies have included the hemoglobin level [12, 14, 18, 22, 26], absolute lymphocyte count [12, 14], and a combination of C-reactive protein, lactate dehydrogenase, albumin level, platelet count, calcium level, and bilirubin level [19]. Ideally, prospective studies should seek to confirm the importance of laboratory factors, possibly in a non-surgical cohort.

Previous studies that sought to evaluate factors associated with overall survival of patients with metastatic bone disease in the extremities were extremely important in introducing machine learning and demonstrating the external validation of machine learning techniques in independent samples [12, 18]. Our work extends these previous studies by incorporating new factors recently identified to be associated with survival in metastatic bone disease in the extremities and integrating explanations of machine learning algorithms into accessible interfaces for clinicians.

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

We provide an accessible tool for clinicians to help them in their daily practice when they deal with surgical decision making for a patient with a metastasis of the extremity. The tool can be found here: https://sorg-apps.shinyapps.io/extremitymetssurvival/. The model explanations both on a global level and an individual level give clinicians more insight than predicted probabilities alone. Having some understanding about which factors are associated with outcome and how they are linked, both in general and for a specific patient, may help clinicians trust the models and help them better inform patients.

It is important to realize that the decision to operate or not is a difficult one and should not solely be based on the outcome of the prediction models. The patients' surgeon should discuss the options with the patient (and family), explaining the pros and cons of proceeding with surgery/deciding not to have surgery. Survival time is only one of the aspects that should be considered and our prediction models help the surgeon to estimate that. Unlike most previous models [1, 12, 18, 19, 22, 36], we used multiple important performance metrics which showed that our models performed well.

Conclusions

We successfully developed machine learning models to predict 90-day and 1-year survival in patients with bone metastases of the extremities. The final models must be externally validated and future studies must assess the performance of these algorithms in other populations. The final models have been incorporated into a freely accessible web application that can be found at https://sorg-apps.shinyapps.io/extremitymetssurvival/. The values entered in the digital application are placeholders that clinicians can modify based on the individual characteristics of the patient. After inputting values, clinicians have access to the predicted probabilities and can further examine the explanations for these predicted probabilities. Pending external validation, clinicians may use this tool to aid preoperative shared decision making for patients with extremity metastatic bone disease.

Acknowledgments None.

References

1. Bauer HCF, Wedin R. Survival after surgery for spinal and extremity metastases: Prognostication in 241 patients. *Acta Orthop Scand.* 1995;66:143–146.
2. Biecek P. DALEX: explainers for complex predictive models. 2018. Available at: https://arxiv.org/abs/1806.08915. Accessed October 31, 2018.
3. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, Cohen ME. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg.* 2013;217:833-42.e1–3.
4. Bongers MER, Thio QCBS, Karhade A V, Stor ML, Raskin KA, Lozano Calderon SA, DeLaney TF, Ferrone ML, Schwab JH. Does the SORG algorithm predict 5-year aurvival in patients with xhondrosarcoma? An external validation. *Clin Orthop Relat Res.* [Published online ahead of print April 27, 2019]. DOI: 10.1097/CORR.0000000000000748.
5. Breiman L. Random forests. *Mach Learn.* 2001;45:5-32.
6. Coleman RE. Metastatic bone disease: clinical features, pathophysiology and treatment strategies. *Cancer Treat Rev.* 2001;27:165-176.
7. Coleman RE. Clinical features of metastatic bone disease and risk of skeletal morbidity. *Clin Cancer Res.* 2006;12:6243s-6249s.
8. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med* 2015;162:735.
9. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132:1920-1930.
10. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, Znaor A, Bray F. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer*. 2018;144:ijc.31937.
11. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136:E359–E386.
12. 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:e19956.
13. Forsberg JA, Wedin R, Bauer HCF, Hansen BH, Laitinen M, Trovik CS, Keller JØ, Boland PJ, Healey JH. External validation of the Bayesian Estimated Tools for Survival (BETS) models in patients with surgically treated skeletal metastases. *BMC Cancer*. 2012;12:493.
14. 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;475:1252-1261.
15. Greenwell BM, Boehmke BC, McCarthy AJ. A simple and effective model-based variable importance measure. 2018. Available at: https://arxiv.org/abs/1805.04755. Accessed October 31, 2018.
16. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning–Data mining, inference, and prediction. Available at: https://web.stanford.edu/~hastie/Papers/ESLII.pdf. Accessed June 17, 2018.
17. Huntington JT, Butterfield M, Fisher J, Torrent D, Bloomston M. The Social Security Death Index (SSDI) most accurately reflects true survival for older oncology patients. *Am J Cancer Res.* 2013;3:518-522.
18. Janssen SJ, van der Heijden AS, van Dijke M, Ready JE, Raskin KA, Ferrone ML, Hornicek FJ, Schwab JH. 2015 Marshall Urist Young Investigator Award: Prognostication in patients with long bone metastases: Does a boosting algorithm improve survival estimates? *Clin Orthop Relat Res.* 2015;473:3112-3121.
19. Katagiri H, Okada R, Takagi T, Takahashi M, Murata H, Harada H, Nishimura T, Asakura H, Ogawa H. New prognostic factors and scoring system for patients with skeletal metastasis. *Cancer Med.* 2014; 3:1359-1367.

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.

20. 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;87:698-703.
21. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial I.* 1995;2:1137-1143.
22. Nathan SS, Healey JH, Mellano D, Hoang B, Lewis I, Morris CD, Athanasian EA, Boland PJ. Survival in patients operated on for pathologic fracture: implications for end-of-life orthopedic xare. *J Clin Oncol.* 2005;23:6072-6082.
23. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med.* 2016;375: 1216-1219.
24. Ratasvuori M, Wedin R, Keller J, Nottrott M, Zaikova O, Bergh P, Kalen A, Nilsson J, Jonsson H, Laitinen M. Insight opinion to surgically treated metastatic bone disease: Scandinavian Sarcoma Group Skeletal Metastasis Registry report of 1195 operated skeletal metastasis. *Surg Oncol.* 2013;22:132-138.
25. Roodman GD. Mechanisms of bone metastasis. *N Engl J Med.* 2004;350:1655-1664.
26. Sorensen MS, Gerds TA, Hindso K, Petersen MM. Prediction of survival after surgery due to skeletal metastases in the extremities. *Bone Joint J.* 2016;98-B:271-277.
27. Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28: 112-118.
28. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35:1925-1931.
29. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models. *Epidemiology*. 2010;21:128-138.
30. Thio QCBS, Goudriaan WA, Janssen SJ, Paulino Pereira NR, Sciubba DM, Rosovksy RP, Schwab JH. Prognostic role of neutrophil-to-lymphocyte ratio and platelet-to-lymphocyte ratio in patients with bone metastases. *Br J Cancer*. 2018;119: 737-743.
31. Thio QCBS, Karhade A V, Ogink PT, Raskin KA, De Amorim Bernstein K, Lozano Calderon SA, Schwab JH. Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma? *Clin Orthop Relat Res.* 2018:476: 2040-2048.
32. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Med Decis Making.* 2006;26: 565-574.
33. VonElm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. *Int J Surg.* 2014;12:1495-1499.
34. Wainer J. Comparison of 14 different families of classification algorithms on 115 binary datasets. 2016. Available at: http://arxiv.org/abs/1606.00930. Accessed July 13, 2018.
35. Wedin R. Surgical treatment for pathologic fracture. *Acta Orthop Scand Suppl.* 2001;72:1-29.
36. Willeumier JJ, van der Linden YM, van der Wal CWPG, Jutte PC, van der Velden JM, Smolle MA, van der Zwaal P, Koper P, Bakri L, de Pree I, Leithner A, Fiocco M, Dijkstra PDS. An easy-to-use prognostic model for survival estimation for patients with symptomatic long bone metastases. *J Bone Joint Surg Am.* 2018; 100:196-204.
37. Zhang W-Y, Li H-F, Su M, Lin R-F, Chen X-X, Zhang P, Zou C-L. A simple scoring system predicting the survival time of patients with bone metastases after RT. *PLoS One*. 2016;11:e0159506.

Copyright © 2019 by the Association of Bone and Joint Surgeons. Unauthorized reproduction of this article is prohibited.