Original Article

Artificial Intelligence Predicts Cost After Ambulatory Anterior Cruciate Ligament Reconstruction

Yining Lu, M.D., Kyle Kunze, M.D., Matthew R. Cohn, M.D., Ophelie Lavoie-Gagne, M.D., Evan Polce, B.S., Benedict U. Nwachukwu, M.D., M.B.A., and Brian Forsythe, M.D.

Purpose: To develop and internally validate a machine-learning algorithm to reliably predict cost after anterior cruciate ligament reconstruction (ACLR). **Methods:** A retrospective review of the New York State Ambulatory Surgery and Services database was performed to identify patients who underwent elective ACLR from 2015 to 2016. Features included in initial models consisted of patient characteristics (age, sex, insurance status, income, medical comorbidities as classified by the Clinical Classifications Software diagnosis code) as well as intraoperative variables (type of anesthesia and procedure-specific factors). Models were generated to predict total charges using 4 algorithms: random forest, extreme gradient boost, elastic net penalized regression, and support vector machines with radial kernels. Training was performed with 10-fold cross-validation followed by internal validation via 0.632 bootstrapping. Model discriminative performance was assessed by area under the receiver operating characteristic curve, calibration, and the Brier score. Decision curve analysis was performed to demonstrate the net benefit of using the final model in practice. **Results:** In total, 7,311 patients undergoing ambulatory ACLR were included. The random forest model demonstrated the best performance assessed via internal validation (area under the curve = 0.85), calibration, and the Brier score (0.208). Cost incurred was influenced by anesthesia type, operating room time, and number of chronic comorbidities. Decision curve analysis revealed a net benefit for use of the random forest model and the model was integrated into a web-based open-access application. **Conclusions:** The random forest model predicted cost after ambulatory ACLR using a large, statewide database with good performance. The top variables found to predict increased charges were general anesthesia, operating room time, meniscal repair, self-pay insurance, patient neighborhood characteristics, and number of chronic conditions. **Level of Evidence:** III, retrospective cohort study.

Anterior cruciate ligament reconstruction (ACLR) is a reproducible and effective procedure with low overall morbidity. Despite this consistency, patients have inherently variable risks for experiencing postoperative complications, readmission, and reoperation, which can translate to variabilities in total cost of the encounter episode. Focus on value-based health care necessitates methods to risk-stratify patients during preoperative evaluation to allow appropriate preoperative counseling, minimize the incidence of adverse events, and both minimize cost through changing modifiable risk factors as well as optimally allocate resources to high-risk cases. Previously identified risk factors for complications and readmissions in patients undergoing ACLR include age,[1,2] number and type of medical comorbidities,[2,3] operative time,[4] type of anesthesia,[5] body mass index,[5,6] and race.[5] However, economic analyses focused on the cost drivers of ambulatory ACLR is sparse, and the present study seeks to add to this body of evidence through the application of machine learning algorithms.

The development of a predictive model to accurately and rapidly quantify risk and excessive cost use after ACLR would be valuable for stakeholders and health

Department of Orthopaedic Surgery, Mayo Clinic, Rochester, Minnesota (Y.L., B.U.N.); Department of Orthopaedic Surgery, Hospital for Special Surgery, New York, New York (K.K.); and Department of Orthopaedic Surgery, Rush University Medical Center, Chicago, Illinois (M.R.C., O.L.-G., E.P., B.F.), U.S.A.

The authors report the following potential conflicts of interest or sources of funding: K.K. reports other from Arthroscopy, *outside the submitted work. B.U.N. reports personal fees from Remote Health, outside the submitted work. B.F. reports grants from Arthrex and Elsevier, other from Jace Medical, grants from Smith and Nephew, and grants and personal fees from Stryker, outside the submitted work. Full ICMJE author disclosure forms are available for this article online, as supplementary material.*

Yining Lu, M.D., and Kyle Kunze, M.D., contributed equally to this article.

Received February 3, 2021; accepted October 7, 2021.

Address correspondence to Brian Forsythe, M.D., Rush University Medical Center, 1611 W Harrison St., Ste 400, Chicago, IL 60612. E-mail: forsythe.research@rushortho.com

2666-061X/2193
https://doi.org/10.1016/j.asmr.2021.10.013

care institutions alike.[7] Bundled payments have become more common in various orthopaedic subspecialties, yet these payment models often do not account for patient-specific cost drivers and reimbursement may not accurately reflect the complexity of patients or cases. To this end, the application of artificial intelligence is an approach well-suited to establish a cost-predictive model, as it can identify how demographic and procedure-related risk factors can translate into excessive cost prior to deciding to perform ACLR.[8] Specifically, machine learning can be applied to and "learn" from complex datasets that contain this information to optimize predictive ability and identify which patients may be at risk for adverse events. Traditional statistics, such as regression models, are fixed constructs based on predefined relationships that are susceptible to collinearity and often fail to appropriately handle complex relationships and datasets with many inputs. In contrast, machine learning algorithms are capable of learning from these relationships and becoming more accurate when presented with additional data by refining the decisions made to come to a specific prediction.[9] Although machine learning has been successfully applied in orthopaedics to predict costs based on risk factors in total hip and knee arthroplasty,[10-12] total shoulder arthroplasty,[13] and spinal fusion,[14] it has not yet been applied to guide cost and resource use predictions after ACLR.

The ability to create a customized risk prediction model for cost use after ACLR would enable the future development of risk-adjusted, patient-specific payment models as well as inform preoperative prediction of the most impactful postoperative outcomes and metrics. The purpose of the current study was to develop and internally validate a machine learning algorithm to reliably predict cost after ACLR. The authors hypothesized that the best performing algorithm would accurately predict total charges following ACLR and allow for the development of a customized prediction tool to inform patient-centered decision-making.

Methods

Guidelines

The present analysis was performed adherent to The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis guidelines and the Guidelines for Developing and Reporting Machine Learning Models in Biomedical Research.[15,16]

Data Source

After institutional review board exemption, The New York State Ambulatory Surgery and Services Database (NYSASD) was queried for patients undergoing ACLR using the Current Procedural Terminology code 29888 for the years 2015 and2016. The NYSASD is a database developed by the Healthcare Cost and Utilization Project (HCUP) that documents data from both ambulatory surgery centers as well as outpatient services at hospital-owned facilities.[17] Patients undergoing revision ACLR as abstracted using *International Classification of Diseases*, 9th and 10th Revision diagnosis codes for mechanical complication due to graft failure (996.52/T84.490S), concomitant ligamentous procedures, cartilage restoration, or osteotomies, were excluded.

Variables and Outcome

Patient characteristics documented by the database and included in feature selection included both demographic and clinical variables. Feature selection for model input is critical as poorly chosen features can mislead the model predictions and irrelevant features can increase the variance of model prediction due to increased noise. The selection process involved both an automated recursive feature elimination process performed by the standardized modeling workflow to exclude irrelevant features, as well as post-hoc assessment of the selected features by the senior authors, and additional features were added or removed based on their clinical relevance. Demographic variables included: age, sex, race/ethnicity, insurance status, the quartile of annual income earned for the patient's zip code within that state, and community characteristics. Clinical variables included: number of medical comorbidities, quarter of discharge, time spent in the operating room (OR time), anesthesia type (for patients who were administered multiple anesthesia, this was hierarchically determined in the NYSASD database as general, regional, other, and local), Clinical Classifications Software diagnosis code, and concomitant procedures. Variables related to patient income level and surgery location were included as well. The full list of variables considered for feature selection is provided in Table 1.

The primary outcome of interest was total charges. Total charges were adjusted for inflation by converting all costs to 2015 US dollars using the medical care-specific consumer price index provided by the Bureau of Labor Statistics, which adjusts for inflation specific to medical care commodities and medical care services. Inflation-adjusted charges were then normalized to z-scores equal to one standard deviation above and below the mean charge, and continuous charges were binned into three categories: low, equal to one standard deviation below the mean cost, and high, greater than one standard deviation above the mean cost, and average, within one standard deviation of the mean cost. The decision boundaries for inflation-adjusted cost were <$1,660.57, between $1,660.57 and $16,707.9, and ≥$16,707.9.

Table 1. Baseline Characteristics of the Study Population, n = 7,311

Variable	n (%), Median (IQR)
Demographics and clinical history	
Age	31 (24-41)
Sex	
Female	2,808 (38.4)
Male	4,503 (61.6)
Race	
White	4,368 (59.7)
Black	604 (8.3)
Hispanic	616 (8.4)
Asian or Pacific Islander	317 (4.3)
Native American	34 (0.5)
Other	1,372 (18.8)
Hispanic	
Not Hispanic	6,691 (91.5)
Hispanic, White	186 (2.5)
Hispanic, Black	41 (0.6)
Hispanic, other race	393 (5.4)
Insurance status	
Medicare	46 (0.6)
Medicaid	954 (13.0)
Private insurance	5,094 (69.7)
Self-pay	424 (5.8)
No charge	4 (0.1)
Other	789 (10.8)
Discharge quarter	
1	1,896 (25.9)
2	1,948 (26.6)
3	1,668 (22.8)
4	1,799 (24.6)
Number of chronic conditions	1 (0-1)
Operative characteristics	
Anesthesia	
MAC/IV sedation	642 (8.8)
Local anesthesia	335 (4.6)
General anesthesia	4,695 (64.2)
Regional anesthesia	1,367 (18.7)
OR time	118 (89-150)
Concomitant procedures	
Meniscal repair	299 (4.1)
Menisectomy	1,349 (18.5)
Microfracture	90 (1.2)
Synovectomy	71 (1.0)
Graft from distance	15 (0.2)
Plica excision	50 (0.7)
Community characteristics	
Median household income state quartile	
1	1,257 (17.2)
2	1,613 (22.1)
3	1,908 (26.1)
4	2,533 (34.6)
Median household income for patient ZIP code	
1	1,257 (17.2)
2	1,613 (22.1)
3	1,908 (26.1)
4	2,533 (34.6)
Patient location: CBSA	
Non-CBSA	148 (2.0)
Micropolitan statistical area	354 (4.8)
Metropolitan statistical area	6,809 (93.1)
Total charges	
High	845 (11.6)
Average	6,278 (85.9)
Low	188 (2.6)

CBSA, core-based statistical area; IQR, interquartile range; IV, intravenous; MAC, monitored anesthesia care; OR, operating room.

Missing Data and Feature Selection

If a variable was considered important and missing in more than 30% of the study population, complete case analysis was performed after exclusion of incomplete cases. However, features with missing data were imputed to reduce bias and improve statistical power where possible.[18] Multiple imputation is a popular method for handling missing data. In this approach, missing value in the dataset is replaced with an imputed value based on a statistical estimation; this process is repeated randomly resulting in multiple "completed" datasets, each consisting of observed and imputed values. These are combined using a simple formulae known as Rubin's rule to give final estimates of target variables.[19]

The missForest multiple imputation method was used to impute remaining variables with less than 30% missing data.[20-22] Variables were assumed to be missing-at-random based on epidemiologic convention,[23,24] although multiple imputation is equipped to handle both missing completely at random and missing not at random data.[25]

Following imputation for missing data, feature selection was performed using recursive feature elimination using a random forest algorithm. Recursive feature elimination has been demonstrated to effectively select an optimal number of input variables with low collinearity within high dimensional data.[26,27]

Modeling

Following selection, modeling was performed using the selected features with each of the following candidate machine learning algorithms: 2 tree-based models: random forest and extreme gradient boosted machine (XGBoost); support vector machines (SVM) with radial kernel; and elastic net penalized regression. Candidate models were chosen to represent a diverse spectrum of modeling techniques, and these algorithms have been shown to develop robust predictive models for various orthopedic conditions.[6] Specifically, random forest and XGBoost are derived from the family of decision-tree based models, which have the advantages of improved flexibility and reduced bias over generalized linear models; however secondary to the same flexibility is an inherent tendency for random forest to overfit to the training data, with resultant increases in variance. Boosting is a separate ensemble technique

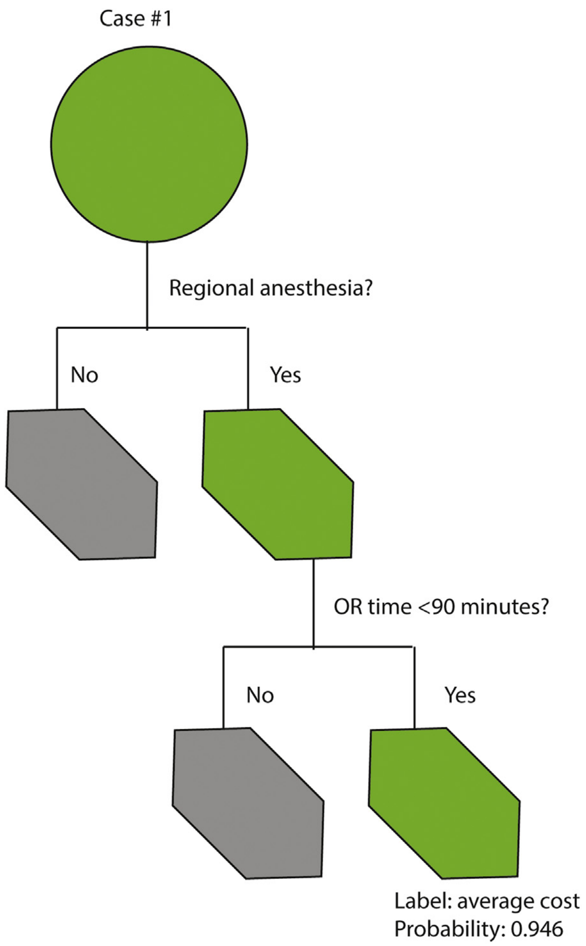

Fig 1. Simplified graphic demonstrating the basic decision-tree: from the root node (an example patient), the algorithm takes the case through several branching points based on the feature space until a leaf node is reached, where the patient falls into a cohort that cannot be further split, and the predicted probability and label are provided accordingly. (OR, operating room.)

that combines multiple weak classifiers to improve both model bias and variance. A simplified sample decision tree is provided in Figure 1. A dictionary of frequently encountered terms and concepts in machine learning is provided in Appendix Table 1, available at www.arthroscopyjournal.org.[18-22,28-36]

Models were trained and internally validated via 0.632 bootstrapping with 1,000 resampled datasets. In brief, model evaluation consists of reiterative partitions of the complete dataset into train and test sets. For each combination of train and test set, the model is trained on the train set using 10-fold cross validation repeated 3 times. The performance of this model is then evaluated on the respective test set, and no data points from the training set was included in the test set. This sequence of steps is then repeated for 999 more data partitions. The model is thus trained and tested on all datapoints available and evaluation metrics are summarized with standard distributions of values. Bootstrapping has been found to optimize evaluation of both model bias and variance as well as improve overall performance compared with internal validation through splitting the data into training and holdout sets.[28] The optimal model was chosen based on area under the receiver operating characteristics curve (AUROC). Models were compared by discrimination, calibration, and Brier score values.

Discriminative power was assessed via the AUROC. Calibration of the model's predicted probabilities as a function of observed frequencies within the test population are summarized in a calibration plot. An ideal model is a straight line with intercept 0 and slope of 1 (i.e., perfect concordance of model predictions to observed frequencies within the retrospective data). Based on the works of Hosmer and Lemeshow, an AUROC of 0.70 to 0.80 was considered acceptable and an AUROC of 0.80 to 0.90 was considered excellent.[29] Finally, the mean squared difference between predicted probabilities of models and observed outcomes, known as the Brier score, was calculated for each candidate model. The Brier score of candidate algorithm is then assessed by comparison to the Brier score of the null model, which is a model that assigns a class probability equal to the sample prevalence of the outcome for every prediction.

Decision curve analysis was used to determine the benefit of implementing the predictive algorithm in practice. The curve plots net benefit against the predicted probabilities of each outcome and provides the cost—benefit ratio for every value of the predicted probability. These ratios provide useful guidance for individualized decision-making and accounts for variability in clinician and patients thresholds for what is considered high-risk. In addition, decision curves for the default strategies of changing management for no patients or all patients are plotted for comparison purposes. To further highlight the utility of the final machine learning model over traditionally reported logistic regression, decision—curve analysis also was performed comparing a learned multivariate logistic regression model using the same parameters and inputs.

Both global and local model interpretability and explanations were provided. The global model variable importance plot demonstrates variable importance normalized against the input considered most contributory to the model predictive power. Local explanations for model behavior were provided for transparency into each individual output using local-interpretable model-agnostic explanations. The explanation algorithm generates optimized fits based on an established distance measure for the predicted probabilities of each outcome label based on the values of both categorical and continuous input, which can be visualized.[30,31]

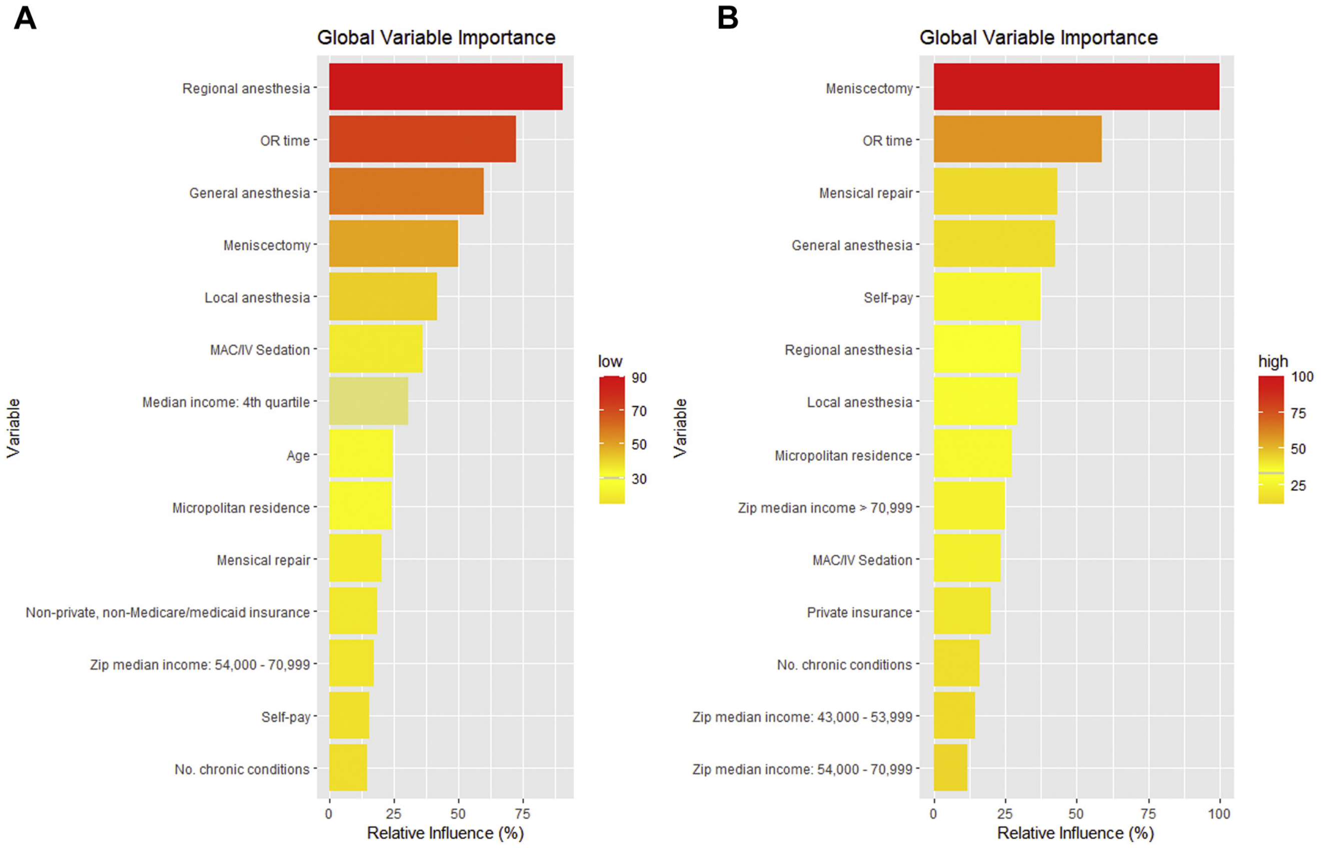

Fig 2. (A) Variable importance plot of the random forest for patients with predicted charges <$1,660.57 and (B) those with predicted charges >$1,6707.9. The variable importance plot demonstrates a global ranking of variables that were the most contributory to improved model performance, importance is relative and provided as a dimensionless quantity. (IV, intravenous care; MAC, monitored anesthesia care; OR, operating room.)

Digital Application

The final model is incorporated into a web-based digital application to illustrate possible future model integration into clinical practice. It should be noted that this digital application remains exclusively for research and educational purposes until rigorous external validation is conducted. In the digital application, preoperative clinical data are entered to generate outcome predictions with accompanying explanations. All data analysis was performed in R 4.0.2 using RStudio, version 1.2.5001 (RStudio, Boston, MA).

Results

Variable Breakdown

A total of 7,311 patients undergoing ALCR between 2015 and 2016 were included in the study. Demographic characteristics of the cohort are as follows: the median age was 31 (interquartile range [IQR] 24-41) years and 2808 (38.4%) patients were female. Variables with missing data considered for modeling are as follows: anesthesia type (1,004, 13.7%), discharge quarter (10, 0.14%), sex (2, 0.03%), OR time (5, 0.07%), median household income state quartile (76, 1.04%), number of chronic comorbidities (1,695, 23.18%), Hispanic ethnicity (580, 7.93%), insurance status (5, 0.07%), median household income for patient zip code (76, 1.04%), and core-based statistical area (18, 0.25%). The full breakdown of variables available for feature selection are provided in Table 1. The median number of chronic comorbidities based on Clinical Classifications Software definitions was 1 (IQR 0-1). Median charges incurred were $6,830.7 (IQR $4,525.7-$1,1814.1). Following discretization, there were 845 (11.6%) patients with total charges >1 standard deviation above the median, 188 (2.6%) with total charges <1 standard deviation below the median, and 6,278 (85.9%) within one standard deviation of the median cost.

Recursive Feature Elimination

Following recursive feature elimination with the random forest algorithm, the following variables were important contributors to increased costs: usen of regional and general anesthesia, concomitant meniscectomy or meniscal repair, increased OR time, self-pay, micropolitan residence, increased zip code median income, and increased number of chronic

Table 2. Model Performances on Internal Validation via 0.632 Bootstrap

Metric	Accuracy	AUROC	Multinomial Brier Score
Elastic net	0.8614 (0.8613-0.8616)	0.799 (0.798-0.801)	0.224 (0.223-0.225)
Random forest	0.8783 (0.8782-0.8784)	0.848 (0.847-0.849)	0.208 (0.207-0.209)
XGBoost	0.8742 (0.8741-0.8742)	0.849 (0.847-0.850)	0.208 (0.207-0.209)
SVM	0.8688 (0.8687-0.8689)	0.783 (0.782-0.783)	0.231 (0.230-0.232)

AUROC, area under receiver operator curve SVM, support vector machines; XGBoost, extreme gradient boosted machine.
Null Brier: 1.27

conditions. Full plots of global importance of the input variables used for training are provided in Figure 2.

Model Performance

Following model tuning, the candidate model performances on internal validation were compared. Accuracy ranged from 0.861 (elastic net) to 0.878 (random forest) and discrimination as measured by AUROC ranged from 0.783 (SVM) to 0.849 (XGBoost) (Table 2). The multinomial Brier score ranged from 0.208 (XGBoost and random forest) to 0.231 (SVM) (Table 2). The null model Brier score was 1.27. Overall, the random forest demonstrated the best performance on discrimination, accuracy, and overall performance (Table 2). All models were appropriately calibrated, and the calibration line for the random forest algorithm had an intercept of −0.056 (95% confidence interval −0.059 to 0.052) and a slope of 1.068 (95% confidence interval 1.064-1.072) (Fig 3).

Decision Curve Analysis

Decision curve analysis were used to compare the net benefit derived from random forest algorithm against the default practices of changing management for all patients or no patients. A decision curve also was plotted for a trained multivariate logistic regression model trained using the same parameters and inputs for comparison to traditional methods used in the orthopaedic literature. The random forest model trained on the complete feature set demonstrated greater net benefit compared to all 3 alternatives (Fig 4).

Explanations

An example of a patient-level explanation accompanying predicted probability of the outcome of interest generated by the digital application is provided in Figure 5. This patient was assigned a probability of 0.967 for incurring cost within 1 standard deviation of the mean charges of the cohort. Features that supported this prediction included regional anesthesia, 0 chronic conditions, OR time >105 minutes, and private insurance. Features that did not support this prediction included zip code median income within the first state quartile, no concurrent meniscectomy or meniscal repair, age <28 years, and zip code median income.

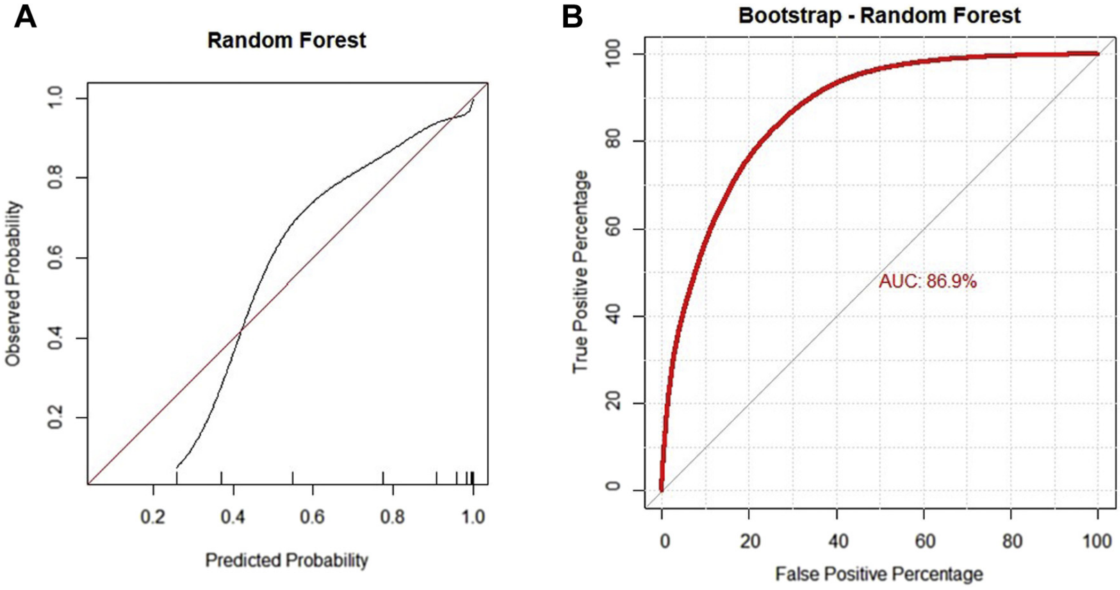

Fig 3. (A) Calibration and (B) discrimination as illustrated by the area under the receiver operating characteristics curve (AUROC) between low-cost and high-cost patients of the random forest algorithm. The ideal calibration curve should have an intercept of 0 and a slope of 1.

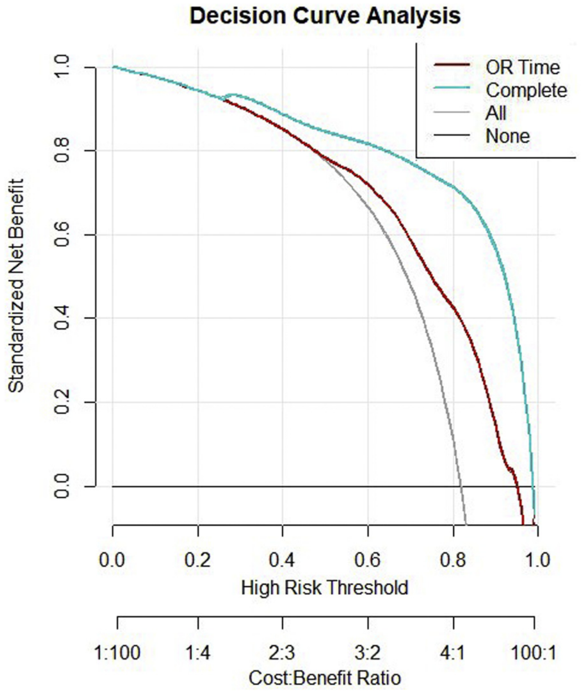

Fig 4. Decision curve analysis of comparing the complete model with model predictions using only OR time. The downsloping line marked by "all" plots the net benefit from the default strategy of changing management for all patients, while the horizontal line marked "none" represents the strategy of changing management for none of the patients (net benefit is zero at all thresholds). The "all" line slopes down because at a threshold of zero, false positives are given no weight relative to true positives; as the threshold increases, false positives gain increased weight relative to true positives and the net benefit for the default strategy of changing management for all patients decreases. (OR, operating room.)

The final model is incorporated into a web-based digital application accessible on desktops, tablets, and smartphones, and can be found at https://sportsmed.shinyapps.io/ACLR_cost/. Default values are provided as placeholders in the interface and the model requires complete cases to generate predictions and explanations.

Discussion

The principle finding of the current study was that the best machine-learning algorithm developed and internally validated in a large population of primary ACLR patients predicted cost and resource use with excellent performance. The most important features determined to influence postoperative cost and resource utilization after ambulatory ACLR included (1) type of anesthesia used, specifically general anesthesia, (2) OR time, and (3) total number of medical comorbidities. The possible integration of this algorithm into the clinical workflow to predict total charges, pending rigorous prospective external validation using geographically or temporally distinct cohorts, was demonstrated using an open access application. However, as noted, the application should not be deployed to patients until the performance of these algorithms is confirmed, and this application in its current form remains only for demonstration and education.

The model identified variables predictive of increased charges following ACLR, including general anesthesia, increased OR time, increased medical comorbidities, self-pay or private insurance, concomitant procedures, top and bottom quartile of income, and micropolitan residence. General anesthesia has been associated with a greater anesthetic cost in spine and hand surgery when compared with regional anesthesia, likely due to increased room time allotted for induction.[37,38] In addition, the increased number of comorbidities is a modifiable risk factor for increased postoperative charges. Isolated ACLR under regional anesthesia is the procedure incurring minimal cost.

The current study developed and internally validated 4 unique machine-learning algorithms on 7,311 patients who underwent ACLR, the best of which was the random forest algorithm. This algorithm has a c-statistic of 0.85, calibration intercept of −0.056, calibration slope of 1.068, and Brier score of 0.208. These values correspond to good discrimination, excellent calibration, and excellent performance for predicting cost use. Previous machine-learning studies in orthopaedics have developed algorithms with similar performance. Karnuta et al.[13] used artificial neural networks to predict inpatient charges for total shoulder arthroplasty and reported that the c-statistics of their models ranged between 0.75 and 0.89; however, this group did not evaluate their algorithms with calibration and Brier score, which are important metrics for demonstrating the precision, estimations, and overall performance of algorithm predictions. Navarro et al.[11] applied machine learning to a large cohort of total knee arthroplasty patients from a national database and found that a Bayesian model incorporating age, race, sex, and comorbidities had acceptable performance with c-statistics that ranged between 0.738 and 0.782. The quantitative metrics describing the performance of the algorithm in the current study are comparable to those predicting cost after other orthopaedic procedures and contribute to its validity as a patient risk stratification model for future use.

Value-based care is dependent on risk-adjustment and appropriate resource allocation for patients undergoing elective procedures with varying risk profiles.[39] The economic risk inherent in these models encourages surgeons to risk-stratify patients to optimize metrics considered in reimbursement. Minimizing excessive cost associated with an episode of care, such as that incurred as a function of unplanned readmissions or procedure-related complications, rewards surgeons who devote effort to doing so.[7] Therefore, it is

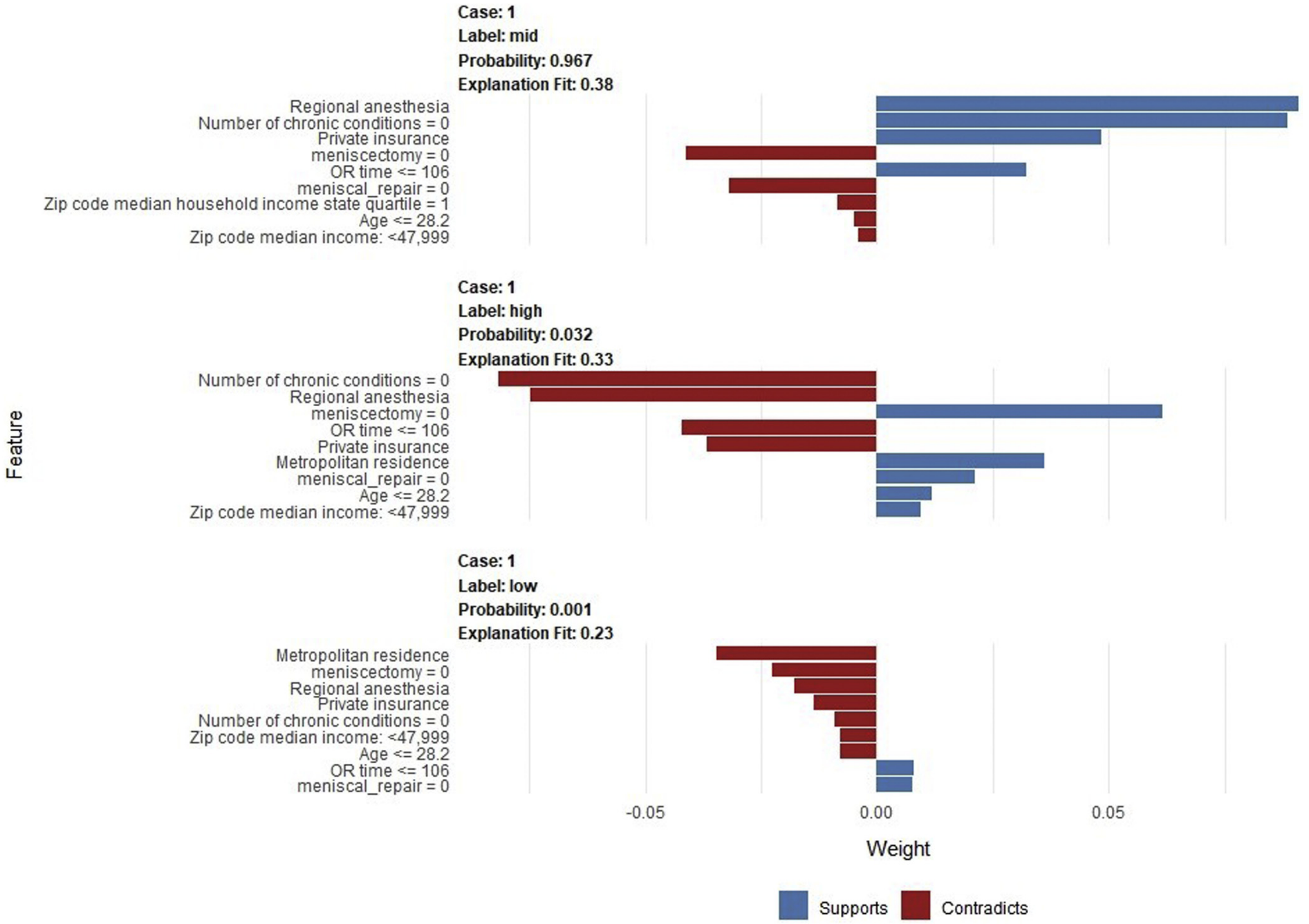

Fig 5. Example of individual patient-level explanation for random forest algorithm predictions. This patient had a predicted probability of 96.7% of incurring cost within one standard deviation of the median cost, and features that supported this prediction included the use of regional anesthesia, no chronic conditions, and OR time <106. (OR, operating room.)

essential that predictive modeling used in this context is dynamic and able to capture the nuances of individual patient's medical context, currently a limitation to traditional models that rely on conforming to predetermined mathematical relationships. The current study incorporated patient-specific data and used local explanations to demonstrate the potential clinical utility of machine learning. Specifically, our models explored age, sex, type of admission, insurance status, procedural factors, and number of chronic comorbidities, all of which may influence value-based outcomes. If integrated into clinical workflow, predictive models such as the one developed in the present study can inform preoperative planning and produce improvements in outcomes and cost-savings. These efforts will be important considerations for reimbursement arbitration of preauthorization with insurance firms and adjusted payment models for episodes of care.

We cannot currently recommend the use of the clinical decision-making tool developed as an open-source application; however, it is important to demonstrate the power and potential clinical impact that machine learning can confer in the current health care environment. External validation of these algorithms is imperative before the introduction of this predictive model into clinical practice. Despite this, transforming patient-specific risk factors for customized prediction into an open-source interface capable of being used in office-based settings holds numerous benefits. The use of these tools will allow for providers to graphically depict individualized predictions and explanations to patients in real time. These applications may also continue to improve and become more accurate with additional data, further strengthening their performance and value.

Limitations

This study is not without limitations. The current machine learning algorithm was developed on a single, large cohort of ambulatory ACLR patients. Although robust about sample size, it represents a single cohort of patients and may be influenced by both selection and algorithm bias. Although the current study demonstrated good predictive performance and responsiveness, the predictive ability may improve with additional data, which included specific breakdown of anesthesia for patients who received multiple types. Another limitation of the current study is the inherent black box phenomenon of machine learning, which does not currently allow for quantitative labels of the predictive strength of each variable in the model. However, by using local agnostic model explanation methodology, we can demonstrate the relative strength of how each variable contributes to a patient's overall risk, as demonstrated by the open-source application. In addition, the database consisted solely of patients undergoing ACLR in an ambulatory setting in the state of New York, which may limit generalizability of identified cost-drivers to other practice settings. Although the vast majority of ACLRs are now performed as outpatient procedures, the costs associated with ACLRs that require postoperative admission are not reflected in the current study's findings. External validation of this model will provide further data regarding the generalizability of the presently developed algorithm. Lastly, although a variety of patient and procedure-related factors were included in the machine learning algorithms, certain nuances in other factors such as personnel, facilities, implants, and graft type, were not available to be incorporated. Pending a robust external validation effort where these limitations are addressed, this algorithm could effectively help predict patient charges after ACLR and optimize resource allocation based on preoperative demographics and comorbidities.

Conclusions

The random forest model predicted cost after ambulatory ACLR using a large, statewide database with good performance. The top variables found to predict increased charges were general anesthesia, OR time, meniscal repair, self-pay insurance, patient neighborhood characteristics, and number of chronic conditions.

References

1. Kraus Schmitz J, Lindgren V, Janarv PM, Forssblad M, Stalman A. Deep venous thrombosis and pulmonary embolism after anterior cruciate ligament reconstruction: Incidence, outcome, and risk factors. *Bone Joint J* 2019;101-B:34-40.
2. Abram SGF, Judge A, Beard DJ, Price AJ. Rates of adverse outcomes and revision surgery after anterior cruciate ligament reconstruction: A study of 104,255 procedures using the national hospital episode statistics database for England, UK. *Am J Sports Med* 2019;47:2533-2542.
3. Brophy RH, Wright RW, Huston LJ, Nwosu SK, Group MK, Spindler KP. Factors associated with infection following anterior cruciate ligament reconstruction. *J Bone Joint Surg Am* 2015;97:450-454.
4. Agarwalla A, Gowd AK, Liu JN, et al. Effect of operative time on short-term adverse events after isolated anterior cruciate ligament reconstruction. *Orthop J Sports Med* 2019;7:2325967118825453.
5. Bokshan SL, DeFroda SF, Owens BD. Risk factors for hospital admission after anterior cruciate ligament reconstruction. *Arthroscopy* 2017;33:1405-1411.
6. Cvetanovich GL, Chalmers PN, Verma NN, Cole BJ, Bach BR Jr. Risk factors for short-term complications of anterior cruciate ligament reconstruction in the United States. *Am J Sports Med* 2016;44:618-624.
7. Jayakumar P, Moore MLG, Bozic KJ. Value-based healthcare: Can artificial intelligence provide value in orthopaedic surgery? *Clin Orthop Relat Res* 2019;477:1777-1780.
8. Ramkumar PN, Haeberle HS, Bloomfield MR, et al. Artificial intelligence and arthroplasty at a single institution: Real-world applications of machine learning to big data, value-based care, mobile health, and remote patient monitoring. *J Arthroplasty* 2019;34:2204-2209.
9. Helm JM, Swiergosz AM, Haeberle HS, et al. Machine learning and artificial intelligence: Definitions, applications, and future directions. *Curr Rev Musculoskelet Med* 2020;13:69-76.
10. Ramkumar PN, Navarro SM, Haeberle HS, et al. Development and validation of a machine learning algorithm after primary total hip arthroplasty: Applications to length of stay and payment models. *J Arthroplasty* 2019;34:632-637.
11. Navarro SM, Wang EY, Haeberle HS, et al. Machine learning and primary total knee arthroplasty: Patient forecasting for a patient-specific payment model. *J Arthroplasty* 2018;33:3617-3623.
12. Karnuta JM, Navarro SM, Haeberle HS, et al. Predicting inpatient payments prior to lower extremity arthroplasty using deep learning: Which model architecture is best? *J Arthroplasty* 2019;34:2235-2241 e2231.
13. Karnuta JM, Churchill JL, Haeberle HS, et al. The value of artificial neural networks for predicting length of stay, discharge disposition, and inpatient costs after anatomic and reverse shoulder arthroplasty. *J Shoulder Elbow Surg* 2020;29:2385-2394.
14. Goyal A, Ngufor C, Kerezoudis P, McCutcheon B, Storlie C, Bydon M. Can machine learning algorithms accurately predict discharge to nonhome facility and early unplanned readmissions following spinal fusion? Analysis of a national surgical registry [published online June 7, 2019]. *J Neurosurg Spine*. doi: 10.3171/2019.3.SPINE181367.
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:e323.
16. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Br J Surg* 2015;102:148-158.
17. (HCUP). HCaUP. Overview of the State Ambulatory Surgery and Services Databases (SASD). Rockville, MD: Agency for Healthcare Research and Quality, 2019.
18. Hughes JD, Hughes JL, Bartley JH, Hamilton WP, Brennan KL. Infection rates in arthroscopic versus open rotator cuff repair. *Orthop J Sports Med* 2017;5:2325967117715416.

19. Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: An overview and case study. *Emerging Themes Epidemiol* 2017;14:8.
20. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012;28:112-118.
21. Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Med Res Methodol* 2018;18:168-168.
22. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* 2009;338. b2393-b2393.
23. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006;59:1092-1101.
24. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol* 2017;9:157-166.
25. Van Buuren S. *Flexible imputation of missing data*. Boca Raton, FL: CRC Press, 2018.
26. Nwachukwu BU, Beck EC, Lee EK, et al. Application of machine learning for predicting clinically meaningful outcome after arthroscopic femoroacetabular impingement surgery. *Am J Sports Med* 2020;48:415-423.
27. Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet* 2018;19:65.
28. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic model research. *PLoS Med* 2013;10.
29. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. *3. Aufl.*, ed. New York: Wiley, 2013.
30. Greenwell BM, Boehmke BC, McCarthy AJ. *A simple and effective model-based variable importance measure* 2018. arXiv preprint arXiv:1805.04755.
31. Ribeiro MT, Singh S, Guestrin C. *Model-agnostic interpretability of machine learning* 2016. *arXiv preprint arXiv:1606.05386*.
32. Karhade AV, Ogink PT, Thio Q, et al. Machine learning for prediction of sustained opioid prescription after anterior cervical discectomy and fusion. *Spine J* 2019;19:976-983.
33. Raschka S. *Model evaluation, model selection, and algorithm selection in machine learning* 2018. arXiv preprint arXiv:1811.12808.
34. Dietterich TG. *Ensemble methods in machine learning. International workshop on multiple classifier systems*. Berlin, Heidelberg: Springer, 2000;1-15.
35. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;3:18.
36. Kuhn M, Johnson K. *Applied predictive modeling*. Berlin: Springer Science & Business Media, 2013.
37. Kahveci K, Doger C, Ornek D, Gokcinar D, Aydemir S, Ozay R. Perioperative outcome and cost-effectiveness of spinal versus general anesthesia for lumbar spine surgery. *Neurol Neurochir Pol* 2014;48:167-173.
38. Chan VW, Peng PW, Kaszas Z, et al. A comparative study of general anesthesia, intravenous regional anesthesia, and axillary block for outpatient hand surgery: Clinical outcome and cost analysis. *Anesth Analg* 2001;93:1181-1184.
39. Novikov D, Cizmic Z, Feng JE, Iorio R, Meftah M. The historical development of value-based care: How we got here. *J Bone Joint Surg Am* 2018;100:e144.

Appendix

Appendix 1. Detailed Machine-Learning Modeling Workflow

Missing Data

Features with missing data were imputed using the missForest multiple imputation method to reduce bias and improve statistical robustness.[18] If a variable was considered important and missing in more than 30% of the study population, complete case analysis was performed after exclusion of incomplete cases. The miss Forest multiple imputation method was used to impute remaining variables with less than 30% missing data,[20-22] variables were assumed to be missing-at-random based on epidemiological convention.[23,24]

Modeling

Following imputation for missing data, highly collinear variables (defined as Spearman's correlation coefficients >0.5 or those considered clinically confounding) were identified and excluded. Notably, we did not explicitly exclude outcome variables of one model as input features in other models, therefore, recurrence was considered as an input feature in the model for progression to surgery, and whether patients underwent surgical treatment was considered an input feature in the model for development of symptomatic osteoarthritis.

The following 5 algorithms were developed on the training data set: (1) Support vector machines, (2) elastic net penalized logistic regression, (3) random forest, (4) neural network, and (5) extreme gradient boosting. These algorithms have been shown to develop robust predictive models for various orthopedic conditions.[32] Each model was trained and validated via 0.632 bootstrapping with 1,000 resampled datasets, also known as Monte Carlo cross-validation. In brief, model evaluation consists of reiterative partitions of the complete dataset into train and test sets. For each combination of train and test set, the model is trained on the train set using 10-fold cross validation repeated 3 times.[33] The performance of this model is then evaluated on the respective test set, and no data points from the training set was included in the test set. This sequence of steps is then repeated for 999 more data partitions. The model is thus trained and tested on all datapoints available and evaluation metrics summarized with standard distributions of values. Bootstrapping has been found to optimize both model bias and variance and improve overall performance compared to internal validation through splitting the data into a partition of training and holdout sets.[28] In addition, a gradient-boosted ensemble model of the 5 candidate models was constructed and trained, similarly through 0.623 bootstrapping. Advantages of ensemble modeling include decreasing variance and bias as well as improving predictions, whereas disadvantages include increased memory requirements and reduced speed of implementation.[34]

Model Assessment

Model performance for each algorithm was assessed for (1) discrimination by comparing area under the receiver operating curve, with >0.80 defined as excellent concordance based on the works of Hosmer and Lemeshow[29]; (2) calibration by calibration curve plots, intercept, and slope; (3) decision curve analysis; and (4) Brier score, which is the mean squared difference between predicted probabilities of models and observed outcomes. The Brier score for each algorithm was compared with the null Brier score, which is calculated by assigning each patient a probability equivalent to the population prevalence of the predicted outcome.

Decision curve analysis was used to compare the benefit of implementing the best-performing algorithm to the logistic regression in practice.[35] The curve plots net benefit against the predicted probabilities of each outcome and provide the cost—benefit ratio for every value of the predicted probability. These ratios provide useful guidance for individualized decision-making and accounts for variability in clinician and patients thresholds for what is considered high-risk. In addition, decision curve for the default strategies of changing management for no patients or all patients are plotted for comparison purposes. Equations for the calculation of the cost-benefit ratio and net benefit are as follows:

$$Cost : Benefit\ Ratio = \frac{risk\ threshold\ probability}{1 - risk\ threshold\ probabilty}$$

$$Net\ benefit = \frac{True\ positives - Cost : Benefit\ Ratio\ (False\ Positives)}{Total\ number\ of\ Patients}$$

Both global and local model interpretability and explanations were provided. The global model variable importance plot demonstrates variable importance normalized to the input considered most contributory to the model predictive power. Local explanations for model behavior were provided for transparency into each individual output using local-interpretable model-agnostic explanations.[30,31] The explanation algorithm generates optimized fits based on an established distance measure for the predicted probabilities of each outcome label based on the values of both categorical and continuous input, which can be plotted for visualization.[30,31]

Appendix Table 1. Definition of Machine Learning Concepts and Methods Used

Term	Definition
Multiple imputation	A popular method for handling missing data, which is often a source of bias and error in model output. In this approach, missing value in the dataset is replaced with an imputed value based on a statistical estimation; this process is repeated randomly resulting in multiple "completed" datasets, each consisting of observed and imputed values. These are combined utilizing a simple formulae known as Rubin's rule to give final estimates of target variables.[19]
Recursive feature elimination	A feature selection algorithm that searches for an optimal subset of features by fitting a given machine learning algorithm (random forest and naïve Bayes in our case) to the predicted outcome, ranking the features by importance, and removing the least important features, this is done repeatedly, in a "recursive" manner until a specified number of features remain or a threshold value of a designated performance metric has been reached. The features can then be entered as inputs into the candidate models for prediction of the desired outcome.[36]
0.632 Bootstrapping	The method for training an algorithm based on the input features selected from recursive feature elimination. Briefly, model evaluation consists of reiterative partitions of the complete dataset into train and test sets. For each combination of train and test set, the model is trained on the train set using 10-fold cross validation repeated 3 times. The performance of this model is then evaluated on the respective test set, and no data points from the training set was included in the test set. This sequence of steps is then repeated for 999 more data partitions.[33] The model is thus trained and tested on all datapoints available and evaluation metrics summarized with standard distributions of values.[33] Bootstrapping has been found to optimize both model bias and variance and improve overall performance compared to internal validation through splitting the data into training and holdout sets
Extreme gradient boosting	Algorithm of choice among stochastic gradient boosting machines, a family in which multiple weak classifiers (a classifier that predicts marginally better than random) are combined (in a process known as boosting) to produce an ensemble classifier with a superior generalized misclassification error rate.[36]
Random forest	Algorithm of choice among tree-based algorithms, an ensemble of independent trees, each generating predictions for a new sample chosen from the training data, and whose predictions are averaged to give the forest's prediction. The ensembling process is distinct in principle from gradient boosting.[36]
Neural network	A nonlinear regression technique based on one or more hidden layers consisting of linear combinations of some or all predictor variables, through which the outcome is modeled, these hidden layers are not estimated in a hierarchical fashion. The structure of the network mimic neurons in a brain.[36]
Elastic-net penalized logistic regression	A penalized linear regression based on a function to minimize the squared errors of the outputs, belongs to the family of penalized linear models including ridge regression and the lasso.[36]
Support vector machines	A supervised learning algorithm that performs classification problems by representation of each data point as a point in abstract space and defines a plane known as a hyperplane that separates the points into distinct binary classes, with maximal margin. Hyperplanes can be linear or nonlinear, as we have implemented in the presented analysis, using a circular kernel.[36]
Area under the receiver operating characteristic curve	A common metric to model performance, utilizing the receiver operating characteristics curve, which plots calculated sensitivity and specificity given the class probability of an event occurring (instead of using a 50:50 probability). The area under the ROC curve classically ranges from 0.5 to 1, with 0.5 being a model that is no better than random and 1 being a model that is completely accurate in assigning class labels.[36]
Calibration	The ability of a model to output probability estimates that reflect the true event rate in repeat sampling from the population. An ideal model is a straight line with intercept 0 and slope of 1 (i.e., perfect concordance of model predictions to observed frequencies within the data). A model can correctly assign a label, as reflected by the area under the receiver operating characteristic, yet it can output class probabilities of a binary outcome that is dramatically different from its true event rate in the population, such a model is not well calibrated.[36]
Brier's Score	The mean squared difference between predicted probabilities of models and observed outcomes in the testing data. The Brier score can generally range from 0 for a perfect model to 0.25 for a noninformative model.[36]
Decision curve analysis	A measure of clinical utility whereby a clinical "net benefit" for one or more prediction models or diagnostic tests is calculated in comparison to default strategies of treating all or no patients. This value is calculated based on a set threshold, defined as the minimum probability of disease at which further intervention would be warranted. The decision curve is constructed by plotting the ranges of threshold values against the net benefit yielded by the model at each value; as such, a model curve that is farther from the bottom left corner yields more net benefit than one that is closer.[35]