ELSEVIER

The SPINE JOURNAL

The Spine Journal 000 (2023) 1–14

Clinical Study

Precision medicine for traumatic cervical spinal cord injuries: accessible and interpretable machine learning models to predict individualized in-hospital outcomes

Mert Karabacak, MD, Konstantinos Margetis, MD, PhD*

Department of Neurosurgery, Mount Sinai Health System, 1468 Madison (Ave), New York, 10029 NY, USA

Received 7 May 2023; revised 28 June 2023; accepted 13 August 2023

Abstract

BACKGROUND CONTEXT: A traumatic spinal cord injury (SCI) can cause temporary or permanent motor and sensory impairment, leading to serious short and long-term consequences that can result in significant morbidity and mortality. The cervical spine is the most commonly affected area, accounting for about 60% of all traumatic SCI cases.

PURPOSE: This study aims to employ machine learning (ML) algorithms to predict various outcomes, such as in-hospital mortality, nonhome discharges, extended length of stay (LOS), extended length of intensive care unit stay (ICU-LOS), and major complications in patients diagnosed with cervical SCI (cSCI).

STUDY DESIGN: Our study was a retrospective machine learning classification study aiming to predict the outcomes of interest, which were binary categorical variables, in patients diagnosed with cSCI.

PATIENT SAMPLE: The data for this study were obtained from the American College of Surgeons (ACS) Trauma Quality Program (TQP) database, which was queried to identify patients who suffered from cSCI between 2019 and 2021.

OUTCOME MEASURES: The outcomes of interest of our study were in-hospital mortality, nonhome discharges, prolonged LOS, prolonged ICU-LOS, and major complications. The study evaluated the models' performance using both graphical and numerical methods. The receiver operating characteristic (ROC) and precision-recall curves (PRC) were used to assess model performance graphically. Numerical evaluation metrics included AUROC, balanced accuracy, weighted area under PRC (AUPRC), weighted precision, and weighted recall.

METHODS: The study employed data from the American College of Surgeons (ACS) Trauma Quality Program (TQP) database to identify patients with cSCI. Four ML algorithms, namely XGBoost, LightGBM, CatBoost, and Random Forest, were utilized to develop predictive models. The most effective models were then incorporated into a publicly available web application designed to forecast the outcomes of interest.

RESULTS: There were 71,661 patients included in the analysis for the outcome mortality, 67,331 for the outcome non-home discharges, 76,782 for the outcome prolonged LOS, 26,615 for the outcome prolonged ICU-LOS, and 72,132 for the outcome major complications. The algorithms exhibited an AUROC value range of 0.78 to 0.839 for in-hospital mortality, 0.806 to 0.815 for nonhome discharges, 0.679 to 0.742 for prolonged LOS, 0.666 to 0.682 for prolonged ICU-LOS, and 0.637 to 0.704 for major complications. An open access web application was developed as part of the study, which can generate predictions for individual patients based on their characteristics.

CONCLUSIONS: Our study suggests that ML models can be valuable in assessing risk for patients with cervical cSCI and may have considerable potential for predicting outcomes during hospitalization. ML models demonstrated good predictive ability for in-hospital mortality and nonhome discharges, fair predictive ability for prolonged LOS, but poor predictive ability for prolonged ICU-LOS and major complications. Along with these promising results, the development of

FDA device/drug status: Not applicable.

Author disclosures: ***MK:*** Nothing to disclose. ***KM:*** Consulting: Viseon Inc (B); Other Office: New York Neurotrauma Consortium (Nonfinancial); Trips/Travel: Accelus (A), Globus (B), Medtronic (A), Stryker (B).

*Corresponding author. Department of Neurosurgery, Mount Sinai Health System, 1468 Madison Ave, New York, 10029 NY, USA.

E-mail address:

https://doi.org/10.1016/j.spinee.2023.08.009

a user-friendly web application that facilitates the integration of these models into clinical practice is a significant contribution of this study. The product of this study may have significant implications in clinical settings to personalize care, anticipate outcomes, facilitate shared decision making and informed consent processes for cSCI patients.

Keywords: Artificial Intelligence; Machine learning; Neurotrauma; Outcome prediction; Spinal cord injury; Spinal trauma; Web application

Introduction

An acute traumatic spinal cord injury (SCI) can result in temporary or permanent motor and sensory impairment, leading to devastating short and long-term consequences that carry significant morbidity and mortality [1–3]. In North America, the estimated annual incidence of SCI is 40 cases per million [4]. While chronic SCI affects nearly 300,000 people in the United States alone, the impact of SCI on an individual's longevity, functional ability, psychological well-being, and socioeconomic stability is substantial [5–7]. Among all traumatic SCI cases, the cervical spine is the most commonly affected area, accounting for approximately 60% of cases [8,9]. Moreover, compared to thoracic or lumbosacral injuries, cervical SCI (cSCI) is associated with higher rates of adverse events and mortality [10–12].

Due to the critical role of timely and effective diagnosis and treatment in managing SCI, there is a promising opportunity for machine learning (ML) approaches to improve the quality of care and best practices [13]. Additionally, personalized or precision medicine can be beneficial in SCI patients by accounting for the inherent variability in outcomes, functional prognosis, and rehabilitation journey among this population, thus allowing for tailored expectations and management strategies. ML-based clinical predictive models offer several advantages over conventional models that typically utilize logistic regression or other linear regression techniques. One of the primary advantages is the ability to handle nonlinear relationships between predictor variables and outcomes. ML algorithms can capture complex patterns and interactions that may be missed by conventional models [14–16]. Additionally, ML algorithms can identify the most important features for prediction, which can be helpful for clinicians to identify which factors are most relevant for a particular outcome [17,18]. These algorithms can also handle missing data more effectively than conventional models, which can improve model accuracy [19]. Furthermore, ML algorithms have the potential to achieve higher accuracy than conventional models, especially when the data is complex, or the relationships between predictors and outcomes are nonlinear [18,20]. Finally, ML algorithms are better able to generalize to new data than conventional models, which can improve the generalizability of the model [18,21]. Overall, these advantages can lead to better clinical decision-making and patient outcomes.

Several studies have established the impressive predictive capabilities of ML models for SCI outcomes [22–25]. Despite this, there is currently no prognostic tool available that is specific to the traumatic cSCI patient population and can be easily used in a clinical setting. This gap in the literature and clinical practice highlights the need for further research to establish and validate ML models specifically tailored to predict cSCI outcomes and can be integrated into clinical practice. This study aims to develop a user-friendly web application that integrates ML algorithms to predict in-hospital outcomes for patients with cSCI. The tool identifies high-risk patients and provides visual explanations of the predictions to establish confidence and encourage adoption in clinical practice. It supports clinicians in guiding treatment strategies, prioritizing care, and planning for discharge needs. Additionally, the tool can facilitate shared decision-making and support quality assurance initiatives. Comparing the tool's predictions to clinical acumen could determine whether the analysis adds value to the existing decision-making process. The application has significant implications for clinical practice, improving patient care and outcomes in the management of cSCI through the integration of ML algorithms.

Methods

Data source

The data for this study were obtained from the American College of Surgeons (ACS) Trauma Quality Program (TQP) database, which was queried to identify patients who suffered from cSCI between 2019 and 2021.

Guidelines

We followed the Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis (TRIPOD) and Journal of Medical Internet Research (JMIR) Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [26,27]. Our study was a retrospective machine learning classification study aiming to predict the outcomes of interest, which were binary categorical variables, in patients diagnosed with cSCI.

Study population

Adult patients (aged 18 and over) with isolated cSCI were identified by the International Classification of Diseases, Tenth Revision (ICD-10) code S12X, S13X, and S14X. We excluded patients with the following criteria: (1)

patients with concurrent thoracolumbar SCI (S22X, S23X, S24X, S32X, S33X, S34X); (2) patients with severe injuries (Abbreviated Injury Scale [AIS] injury severity score≥3) to the head, face, thorax, abdomen, upper extremities, lower extremities, and unspecified body regions; and (3) patients with major polytrauma (Injury Severity Score [ISS] ≥27), (4) patients with minor cSCIs (AIS injury severity score=1), (5) patients with advanced directives limiting care, and (6) patients with prehospital cardiac arrests.

With the second exclusion criteria, the intention was to isolate cSCI to assess its impacts without the confounding effects of other severe injuries in other body regions. Regarding the exclusion of minor cSCIs (AIS injury severity score = 1), this choice was informed by the nature of the injuries that fall into this category. AIS injury severity score of one often encompasses injuries like spinal muscle strains. The inclusion of such injuries would introduce a significant level of heterogeneity to our patient group, which might compromise the specificity of our findings. By focusing on nonminor cSCIs, we aimed to provide a more homogenous patient group and hence more reliable model predictions.

Predictor variables

The predictor variables that were deemed to be known before the occurrence of the outcomes of interest were chosen from the TQP dataset. A list of the variables utilized in the analysis can be found in Supplementary Table 1.

Outcome of interest

The outcomes of interest of our study were in-hospital mortality, nonhome discharges, prolonged LOS, prolonged ICU-LOS, and major complications.

Nonhome discharges included discharges to various healthcare facilities such as inpatient rehab or designated units, home under the care of organized home health services, skilled nursing facilities, long-term care hospitals, short-term general hospitals for inpatient care, and intermediate care facilities. Patients with discharges to court/law enforcement, psychiatric hospital or psychiatric unit of a hospital, another type of institution not defined elsewhere, and patients who left against medical advice or discontinued care were excluded from the nonhome discharges analysis since these outcomes were not considered as a proxy for functional status at discharge. Patients who died during hospitalization or were discharged to hospice care were also excluded from the analysis for nonhome discharges.

LOS and ICU-LOS were assessed by excluding patients who died during hospitalization, left against medical advice, or discontinued care since these patients may have artificially lowered LOS. Prolonged LOS was defined as total LOS greater than 80% of the included patient population (>9 days), and prolonged ICU-LOS was defined as total ICU-LOS greater than 80% of the included patient population (>7 days).

Major complications included severe in-hospital complications such as cardiac arrest with resuscitation, central line-associated bloodstream infection, catheter-related bloodstream infection, deep surgical site infection, deep vein thrombosis, pulmonary embolism, unplanned intubation, acute kidney injury, myocardial infarction, acute respiratory distress syndrome, unplanned return to the operating room, severe sepsis, stroke or cerebrovascular event, unplanned admission to the ICU, and ventilator-associated pneumonia.

Data preprocessing

Imputation was employed to avoid introducing bias by excluding patients with missing values. After removing variables with missing values for more than 25% of the patient population, the k-nearest neighbor (kNN) imputation algorithm was used to fill in the missing values in the remaining continuous variables. To ensure that all feature values were weighed equally, a Min-Max Scaler was used to place each continuous variable in the (0, 1) range. After missing values of categorical variables were imputed with “Unknown” or “Unknown/Other,” all were label-encoded.

Training, validation, and test sets

The data from 2019 to 2021 was split into training, validation, and test sets in a 60:20:20 ratio. The training set was used to train the models, the validation set to fine-tune the hyperparameters and calibrate the models, and the test set to evaluate the models’ performance.

To account for the class imbalance for the positive outcomes of interest the Synthetic Minority Over-sampling Technique (SMOTE) was used to artificially generate cases of positive outcomes of interest based on the training sets [28].

Modeling

This study employed four machine learning algorithms, namely XGBoost, LightGBM, CatBoost, and Random Forest, to build the prediction models for each outcome. The optimization of these algorithms was performed using the Optuna library, with the objective of maximizing the area under the receiver operating characteristic curve (AUROC) metric [29]. To establish a benchmark for optimization, the TPESampler algorithm was used. The final models were developed using the training sets and the optimized hyperparameters. Platt scaling, also known as isotonic regression, was used to calibrate the models [30,31]. All machine learning analyses were conducted using Python version 3.7.15.

Performance evaluation

The study evaluated the models’ performance using both graphical and numerical methods. The receiver operating characteristic (ROC) and precision-recall curves (PRC) were used to assess model performance graphically.

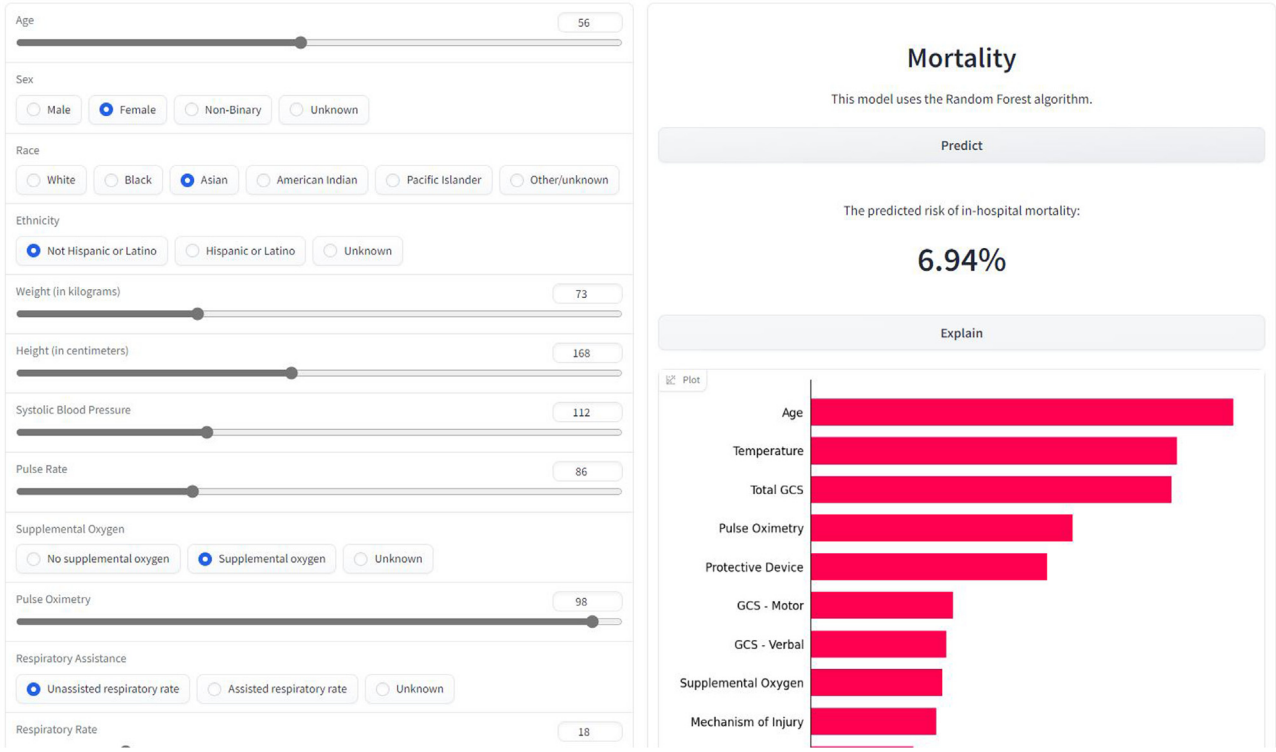

Fig. 1. A screenshot of the online web application.

Numerical evaluation metrics included AUROC, balanced accuracy, weighted area under PRC (AUPRC), weighted precision, and weighted recall. Calibration was assessed using the Brier score. We also employed SHapley Additive exPlanations (SHAP) framework to determine the relative importance of predictor variables [32].

Online prediction tool

An online prediction tool was developed to generate patient-level predictions (Fig. 1). This tool is built upon the models discussed in the present study. Both the tool and its source code are publicly available through Hugging Face, a platform that facilitates sharing of machine learning models. The following link will take readers to the online prediction tool: https://huggingface.co/spaces/MSHS-Neurosurgery-Research/TQP-cSCI.

Statistical analysis

For continuous variables with a normal distribution, means (± standard deviations) were reported, whereas medians (interquartile ranges) were presented for non-normally distributed continuous variables. For categorical variables, the number of patients was reported with percentages. Statistical differences between groups were determined using various statistical tests. Specifically, the independent t-test was applied for normally distributed continuous variables with equal variances, Welch's t-test was used for normally distributed continuous variables with unequal variances, Mann-Whitney U test was used for non-normally distributed continuous variables, and Pearson's chi-squared test was used for categorical variables. The normality of the continuous variables was assessed using the Shapiro-Wilk test, and the equality of variances was evaluated using Levene's test. Statistical significance was considered when $p<.05$. All statistical analyses were performed using Python version 3.7.15.

Results

Initially, 391,960 patients were identified with the ICD-10 codes S12X, S13X, and S14X. After excluding 166,058 patients with concurrent thoracolumbar SCI, other exclusion criteria were applied sequentially (Fig. 2). There were 71,661 patients included in the analysis for the outcome mortality (n=2,280 [3.18%] mortality), 67,331 for the outcome non-home discharges (n=31,300 [46.49%] nonhome discharges), 76,782 for the outcome prolonged LOS (n=13,404 [17.46%] prolonged LOS), 26,615 for the outcome prolonged ICU-LOS (n=4,445 [16.7%] prolonged ICU-LOS), and 72,132 for the outcome major complications (n=3,891 [5.39%] major complications). Characteristics of the patient population are presented in Table 1. Differences among the patients belonging to different outcome groups are presented in Supplementary Tables 2, 3, 4, 5, and 6.

The most accurately predicted outcome in terms of AUROC was the in-hospital mortality, with a mean AUROC of 0.816 (95% confidence interval [CI] 0.790–0.826). Among the four different algorithms, the Random

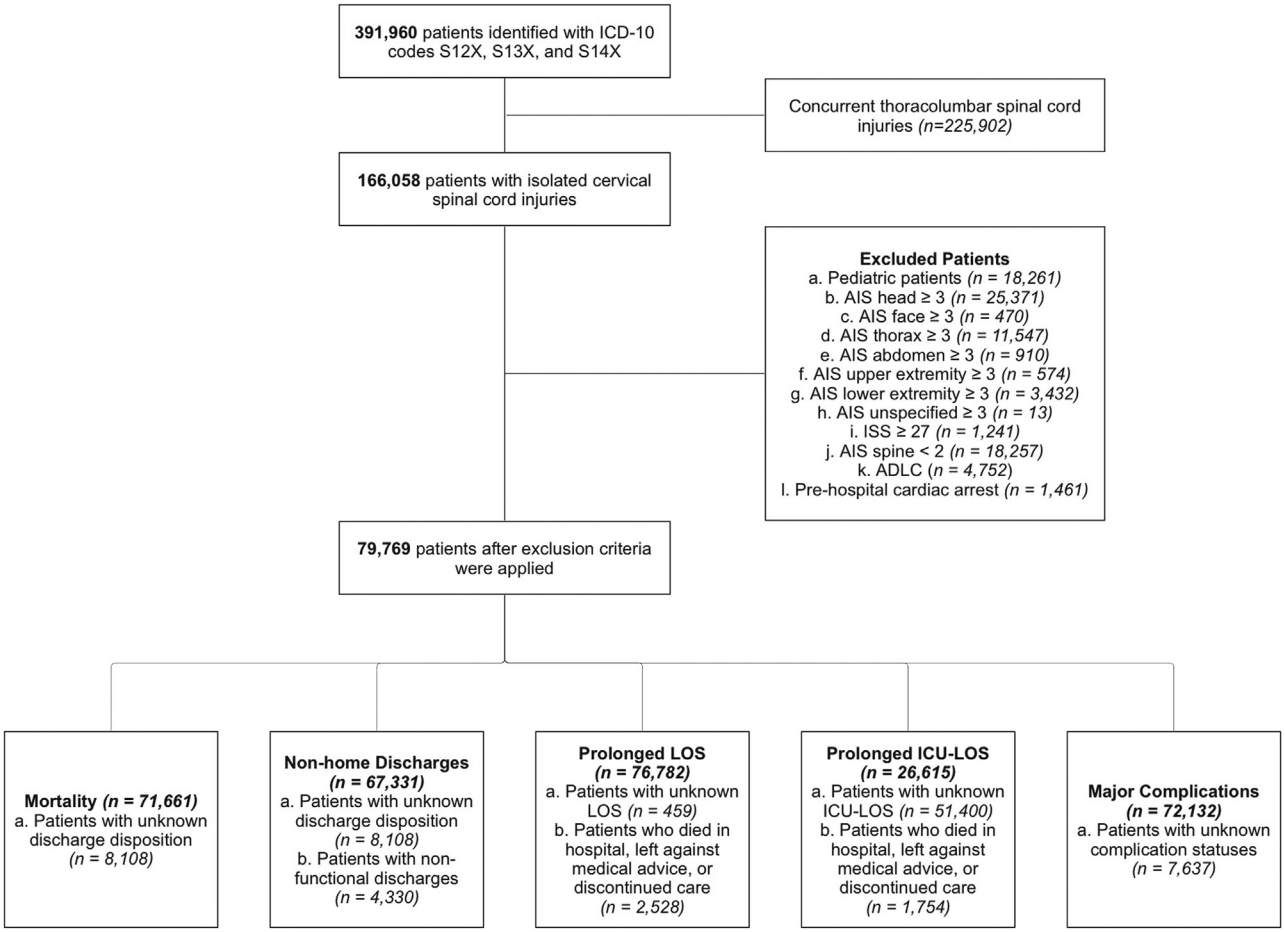

Fig. 2. Patient selection flowchart.

Forest algorithm demonstrated the best discriminative ability across all outcomes, with a mean AUROC of 0.752 (95% CI 0.731–0.760). The performance metrics for the algorithms are provided in Table 2.

The ROC and PRC for each of the five outcomes are illustrated in Figs. 3 and 4, respectively. Meanwhile, Fig. 5 presents the SHAP bar plots for the top-performing algorithm for each outcome. Supplementary Figures 1 to 5 contain the SHAP bar plots for the remaining algorithms for every outcome. SHAP bar plots offer an overview of the significance of features in a model. Within these plots, the value of each feature is signified by a bar, its length relating to the average absolute SHAP value from all occurrences. This importance gauge signifies the average strength of a feature's input to the model's prediction. The features are arranged according to their significance, with the most influential feature at the top. For instance, Fig. 5D shows the SHAP bar plot for the XGBoost model's prediction of prolonged ICU-LOS, where "Systolic Blood Pressure" possesses the longest bar. This signifies that systolic blood pressure is the most critical predictor of prolonged ICU-LOS in the model, meaning that systolic blood pressure has globally the greatest impact on the model's prediction of prolonged ICU-LOS compared to the other features.

Discussion

The purpose of this study was to explore the potential of ML models in enhancing the prediction of adverse in-hospital outcomes following cSCI by utilizing an expanded set of clinical variables. Our ML models were developed to forecast in-hospital mortality, nonhome discharges, prolonged LOS, prolonged ICU-LOS, and major complications. The findings of the study suggest that ML algorithms can facilitate the risk stratification of cSCI patients and offer valuable insights into predicting adverse in-hospital outcomes. The results regarding the discriminatory performances of our models indicate good classification performance for predicting in-hospital mortality and nonhome discharges, fair performance for predicting prolonged LOS, and poor performance for predicting prolonged ICU-LOS and major complications [33]. Additionally, we developed an open-access web application that provides probabilistic predictions for the outcomes investigated in this study for cSCI patients. The application incorporates SHAP plots to offer visual explanations of the predictions, aiming to establish confidence in the predictions and encourage their adoption in clinical practice. To the best of our knowledge, this

Table 1
Patient characteristics.

Variables		Total Mean (±SD), Median (IQR), or n (%)
Age		62.0 (±33.0)
Sex	Male	50263 (63.0%)
	Female	29199 (36.6%)
	Non-binary	8 (0.0%)
	Unknown	299 (0.4%)
Race	White	59494 (74.6%)
	Black	12218 (15.3%)
	Asian	1536 (1.9%)
	Other/unknown	6521 (8.2%)
Ethnicity	Not Hispanic or Latino	70480 (88.4%)
	Hispanic or Latino	6409 (8.0%)
	Unknown	2880 (3.6%)
Weight		79.4 (±25.0)
Height		172.7 (±15.0)
Systolic Blood Pressure		142.0 (±34.0)
Pulse Rate		82.0 (±22.0)
Supplemental Oxygen	No supplemental oxygen	66663 (83.6%)
	Supplemental oxygen	7854 (9.8%)
	Unknown	5252 (6.6%)
Pulse Oximetry		98.0 (±3.0)
Respiratory Assistance	Unassisted respiratory rate	74428 (93.3%)
	Assisted respiratory rate	1593 (2.0%)
	Unknown	3748 (4.7%)
Respiratory Rate		18.0 (±4.0)
Temperature		36.7 (±0.5)
GCS - Eye		4.0 (±0.0)
GCS - Verbal		5.0 (±0.0)
GCS - Motor		6.0 (±0.0)
Total GCS		15.0 (±0.0)
Fracture of C1 Vertebra	No	76584 (96.0%)
	Yes	3185 (4.0%)
Fracture of C2 Vertebra	No	70710 (88.6%)
	Yes	9059 (11.4%)
Fracture of C3 Vertebra	No	78541 (98.5%)
	Yes	1228 (1.5%)
Fracture of C4 Vertebra	No	78209 (98.0%)
	Yes	1560 (2.0%)
Fracture of C5 Vertebra	No	77472 (97.1%)
	Yes	2297 (2.9%)
Fracture of C6 Vertebra	No	76286 (95.6%)
	Yes	3483 (4.4%)
Fracture of C7 Vertebra	No	75685 (94.9%)
	Yes	4084 (5.1%)
Rupture of Cervical Intervertebral Disc	No	79549 (99.7%)
	Yes	220 (0.3%)
Subluxation and Dislocation of C0/C1 Vertebrae	No	79677 (99.9%)
	Yes	92 (0.1%)
Subluxation and Dislocation of C1/C2 Vertebrae	No	79397 (99.5%)
	Yes	372 (0.5%)
Subluxation and Dislocation of C2/C3 Vertebrae	No	79660 (99.9%)
	Yes	109 (0.1%)
Subluxation and Dislocation of C3/C4 Vertebrae	No	79579 (99.8%)
	Yes	190 (0.2%)
Subluxation and Dislocation of C4/C5 Vertebrae	No	79512 (99.7%)
	Yes	257 (0.3%)
Subluxation and Dislocation of C5/C6 Vertebrae	No	79457 (99.6%)
	Yes	312 (0.4%)
Subluxation and Dislocation of C6/C7 Vertebrae	No	79549 (99.7%)
	Yes	220 (0.3%)
Subluxation and Dislocation of C7/T1 Vertebrae	No	79733 (100.0%)
	Yes	36 (0.0%)
Concussion and Edema of cSC	No	78654 (98.6%)
	Yes	1115 (1.4%)
Complete Lesion of cSC	No	75477 (94.6%)
	Yes	4292 (5.4%)
Anterior Cord Syndrome of cSC	No	79677 (99.9%)
	Yes	92 (0.1%)
Brown-Sequard Syndrome of cSC	No	79510 (99.7%)
	Yes	259 (0.3%)
Other Incomplete Lesions of cSC	No	78474 (98.4%)
	Yes	1295 (1.6%)
Current Smoker	No	62076 (77.8%)
	Yes	17667 (22.2%)
	Unknown	26 (0.0%)
Alcohol Use Disorder	No	73286 (91.9%)
	Yes	6435 (8.1%)
	Unknown	48 (0.1%)
Substance Abuse Disorder	No	74620 (93.6%)
	Yes	5100 (6.4%)
	Unknown	49 (0.1%)
Diabetes Mellitus	No	65628 (82.3%)
	Yes	14118 (17.7%)
	Unknown	23 (0.0%)
Hypertension	No	44189 (55.4%)
	Yes	35546 (44.6%)
	Unknown	34 (0.0%)
Congestive Heart Failure	No	75761 (95.0%)
	Yes	3998 (5.0%)
	Unknown	10 (0.0%)
History of Myocardial Infarction	No	78986 (99.0%)
	Yes	640 (0.8%)
	Unknown	143 (0.2%)
Angina Pectoris	No	79568 (99.8%)
	Yes	193 (0.2%)
	Unknown	8 (0.0%)
History of Cerebrovascular Accident	No	77521 (97.2%)
	Yes	2222 (2.8%)
	Unknown	26 (0.0%)
Peripheral Arterial Disease	No	78877 (98.9%)
	Yes	860 (1.1%)
	Unknown	32 (0.0%)
Chronic Obstructive Pulmonary Disease	No	73689 (92.4%)
	Yes	6069 (7.6%)
	Unknown	11 (0.0%)

Table 1 (***Continued***)

Table 1 (***Continued***)

Variables		Total Mean (±SD), Median (IQR), or n (%)
Chronic Renal Failure	No	78208 (98.0%)
	Yes	1541 (1.9%)
	Unknown	20 (0.0%)
Cirrhosis	No	78874 (98.9%)
	Yes	871 (1.1%)
	Unknown	24 (0.0%)
Bleeding Disorder	No	79016 (99.1%)
	Yes	746 (0.9%)
	Unknown	7 (0.0%)
Disseminated Cancer	No	79247 (99.4%)
	Yes	514 (0.6%)
	Unknown	8 (0.0%)
Currently Receiving Chemotherapy for Cancer	No	79330 (99.4%)
	Yes	429 (0.5%)
	Unknown	10 (0.0%)
Dementia	No	75436 (94.6%)
	Yes	4327 (5.4%)
	Unknown	6 (0.0%)
Attention Deficit Disorder or Attention Deficit Hyperactivity Disorder	No	78903 (98.9%)
	Yes	861 (1.1%)
	Unknown	5 (0.0%)
Mental or Personality Disorder	No	70540 (88.4%)
	Yes	9163 (11.5%)
	Unknown	66 (0.1%)
Ability to Complete Age-Appropriate ADL	No	71467 (89.6%)
	Yes	8288 (10.4%)
	Unknown	14 (0.0%)
Pregnancy	Not applicable (male patient)	50263 (63.0%)
	No	29398 (36.8%)
	Yes	74 (0.1%)
	Unknown	34 (0.0%)
Anticoagulant Therapy	No	69992 (87.7%)
	Yes	9766 (12.2%)
	Unknown	11 (0.0%)
Steroid Use	No	78825 (98.8%)
	Yes	938 (1.2%)
	Unknown	6 (0.0%)
Days from Incident to ED or Hospital Arrival		1.0 (±0.0)
Transport Mode	Ground ambulance	65057 (81.6%)
	Private/public vehicle/walk-in	7628 (9.6%)
	Air ambulance	6399 (8.0%)
	Other/unknown	685 (0.9%)
Inter-Facility Transfer	No	49348 (61.9%)
	Yes	30411 (38.1%)
	Unknown	10 (0.0%)
Trauma Type	Blunt	77651 (97.3%)
	Penetrating	1422 (1.8%)
	Other/unknown	696 (0.9%)
Injury Intent	Unintentional	76594 (96.0%)
	Assault	2436 (3.0%)
	Other/unknown	739 (0.9%)
Mechanism of Injury	Fall	43530 (54.6%)
	MVT occupant	20907 (26.2%)
	Struck by or against	2863 (3.6%)
	Other MVT	2548 (3.2%)
	MVT motorcyclist	2094 (2.6%)
	Other transport	1919 (2.4%)
	Other pedal cyclist	1494 (1.9%)
	MVT pedestrian	962 (1.2%)
	Firearm	1068 (1.3%)
	Other/unknown	2384 (3.0%)
Protective Device	None	56196 (70.4%)
	Belt	13939 (17.5%)
	Airbag present	4759 (6.0%)
	Helmet	2761 (3.5%)
	Other/unknown	2114 (2.6%)
Work-Related	No/unknown	77762 (97.5%)
	Yes	2007 (2.5%)
Blood Transfusion		0.0 (±0.0)
Surgical Intervention	None	77843 (97.6%)
	Fusion	1253 (1.6%)
	Decompression	563 (0.7%)
	Other	110 (0.1%)
Alcohol Screen	Yes	41339 (51.8%)
	No	38277 (48.0%)
	Unknown	153 (0.2%)
Alcohol Screen Result		0.0 (±0.0)
Drug Screen - Amphetamine	No	26407 (33.1%)
	Yes	2488 (3.1%)
	Not tested	50874 (63.8%)
Drug Screen - Barbiturate	No	28612 (35.9%)
	Yes	283 (0.4%)
	Not tested	50874 (63.8%)
Drug Screen - Benzodiazepines	No	27451 (34.4%)
	Yes	1444 (1.8%)
	Not tested	50874 (63.8%)
Drug Screen - Cannabinoid	No	22285 (27.9%)
	Yes	6610 (8.3%)
	Not tested	50874 (63.8%)
Drug Screen - Cocaine	No	26636 (33.4%)
	Yes	2259 (2.8%)
	Not tested	50874 (63.8%)
Drug Screen - MDMA or Ecstasy	No	28744 (36.0%)
	Yes	151 (0.2%)
	Not tested	50874 (63.8%)
Drug Screen - Methadone	No	28750 (36.0%)
	Yes	145 (0.2%)
	Not tested	50874 (63.8%)
Drug Screen - Methamphetamine	No	28120 (35.2%)
	Yes	775 (1.0%)
	Not tested	50874 (63.8%)
Drug Screen - Opioid	No	26882 (33.7%)
	Yes	2013 (2.5%)
	Not tested	50874 (63.8%)
Drug Screen - Oxycodone	No	28465 (35.7%)
	Yes	430 (0.5%)
	Not tested	50874 (63.8%)
Drug Screen - Phencyclidine	No	28701 (36.0%)
	Yes	194 (0.2%)
	Not tested	50874 (63.8%)
Drug Screen - Tricyclic Antidepressant	No	28783 (36.1%)
	Yes	112 (0.1%)
	Not tested	50874 (63.8%)
ACS Verification Level	Level I Trauma Center	35220 (44.2%)
	Level II Trauma Center	19979 (25.0%)

Table 1 (***Continued***)

Variables		Total Mean (±SD), Median (IQR), or n (%)
	Level III Trauma Center	3900 (4.9%)
	Unknown	20670 (25.9%)
Hospital Type	Non-profit	69392 (87.0%)
	For profit	9896 (12.4%)
	Government	449 (0.6%)
	Unknown	32 (0.0%)
Facility Bed Size	More than 600	28994 (36.4%)
	401 to 600	22663 (28.4%)
	201 to 400	22735 (28.5%)
	200 or fewer	5377 (6.7%)
Primary Method of Payment	Medicare	31343 (39.3%)
	Private/commercial insurance	26534 (33.3%)
	Medicaid	9621 (12.1%)
	Self-pay	6571 (8.2%)
	Other/unknown	5700 (7.2%)

SD, standard deviation; IQR, interquartile range; GCS, Glasgow Coma Scale; cSC, cervical spinal cord, ADL, activities of daily living; ED, emergency department; ACS, American College of Surgeons

is the first web application of its kind to provide predictions with additional interpretability for cSCI outcomes using ML.

Our study introduces ML models, coupled with a web application, that could offer personalized and quantitative risk assessments for specific undesired outcomes following cSCI. This advancement stands to significantly enhance traditional methods that rely on generic risks based on population averages or subjective assessments by physicians. While we acknowledge that our study does not directly demonstrate the impact of these models on shared decision-making or the informed consent process, we posit that the potential clinical applications of our models are deserving of further investigation. The integration of these models into clinical practice could support clinical decision-making throughout a patient's hospital stay by forecasting the risk of functional impairment, thus assisting in prioritizing care and planning for discharge needs. This information can guide informed consent processes and shared decision-making, giving patients and caregivers insights into potential needs for nursing assistance postdischarge, enabling appropriate arrangements. Additionally, the implementation of our ML models and web application may contribute to quality assurance initiatives. They can serve as tools to identify unexpected patterns of undesired outcomes, particularly in patients predicted to be at low risk. This discrepancy between predicted and actual outcomes may reflect gaps in processes, hospital policies, or specific care gaps that vary across different populations. The findings from this comparison can then drive policy changes or resource optimization strategies to improve patient outcomes. For instance, if we observe undesired outcomes in low-risk patients, this might indicate a systemic process gap or a population-specific resource barrier that needs to be addressed. In summary, our approach could potentially improve patient care and bolster clinical decision-making, leading to improved outcomes for individuals with cSCI. We envisage the web application being used by clinicians to validate their clinical decisions based on the predictions generated by the ML models. By providing additional quantitative risk assessments, the tool may aid clinicians in making more informed decisions that consider a patient's specific circumstances and medical history. Such an approach could enhance the integration and workflow of multimodal and multidisciplinary care for cSCI patients. Nonetheless, future studies are needed to ascertain if the analysis outperforms or supplements clinical acumen, as well as to explore its integration into daily care and management.

We acknowledge that the study's main limitations are the clinical accuracy and relevance of the predictions, given the many confounders associated with the outcomes of interest that have not been accounted for due to the limited granularity within this database. While the study is mainly academic at this point, we believe that the potential clinical applications of the tool warrant further investigation. Customization of ML models for specific hospitals is possible by training them with data from that particular hospital, and updates can be made as new data becomes available to enhance their predictive capabilities. In response to concerns raised regarding the clinical relevance and accuracy of our models in real-world scenarios, future research should focus on addressing these limitations by incorporating more granular and comprehensive data sources that account for these confounders. By doing so, the predictive capabilities of the ML models could be significantly improved, further enhancing their potential clinical applications and utility.

When dealing with imbalanced datasets in ML classification tasks, it is crucial to exercise caution and understand the metrics used to evaluate model performance. Class distribution refers to the proportion of instances in each category in a classification problem. In our context, the "majority class" signifies the category with more instances, while the "minority class" denotes the category with fewer instances. For instance, in predicting in-hospital mortality following cSCI, the majority of patients would likely fall into the "patients without in-hospital mortality" category, making it the majority class. In contrast, the "patients with in-hospital mortality" category, having fewer instances, would constitute the minority class. In our study, we used metrics such as balanced accuracy, weighted precision, weighted recall, and weighted AUPRC to assess the performance of our ML models for predicting outcomes after cSCI. These metrics take into account the class distribution of the data, giving more weight to the minority class [34–36]. This allows for the fair evaluation of a model's performance in both classes and a more comprehensive view of the model's performance, considering the class distribution in the data. In contrast, unweighted versions of these metrics may not be reliable in

Table 2
Model performances

Outcome	Algorithm	Weighted precision (95% CI)	Weighted recall (95% CI)	Weighted AUPRC (95% CI)	Balanced accuracy (95% CI)	AUROC (95% CI)	Brier score (95% CI)
Mortality	XGBoost	0.958 (0.955−0.961)	0.789 (0.782−0.796)	0.141 (0.135−0.147)	0.713 (0.706−0.72)	0.821 (0.791−0.827)	0.028 (0.025−0.031)
	LightGBM	0.958 (0.955−0.961)	0.788 (0.781−0.795)	0.142 (0.136−0.148)	0.715 (0.708−0.722)	0.822 (0.793−0.831)	0.028 (0.025−0.031)
	CatBoost	0.949 (0.945−0.953)	0.949 (0.945−0.953)	0.055 (0.051−0.059)	0.573 (0.565−0.581)	0.78 (0.758−0.798)	0.029 (0.026−0.032)
	Random Forest	0.951 (0.947−0.955)	0.961 (0.958−0.964)	0.145 (0.139−0.151)	0.564 (0.556−0.572)	0.839 (0.816−0.848)	0.028 (0.025−0.031)
	Mean	*0.954 (0.95−0.958)*	*0.872 (0.866−0.877)*	*0.121 (0.115−0.126)*	*0.641 (0.634−0.649)*	*0.816 (0.79−0.826)*	*0.028 (0.025−0.031)*
Nonhome discharges	XGBoost	0.735 (0.728−0.742)	0.733 (0.726−0.74)	0.757 (0.75−0.764)	0.734 (0.727−0.741)	0.807 (0.799−0.813)	0.18 (0.174−0.186)
	LightGBM	0.737 (0.73−0.744)	0.735 (0.728−0.742)	0.759 (0.752−0.766)	0.735 (0.728−0.742)	0.806 (0.801−0.815)	0.179 (0.173−0.185)
	CatBoost	0.739 (0.732−0.746)	0.737 (0.73−0.744)	0.641 (0.633−0.649)	0.737 (0.73−0.744)	0.815 (0.803−0.818)	0.177 (0.171−0.183)
	Random forest	0.733 (0.726−0.74)	0.73 (0.723−0.737)	0.745 (0.738−0.752)	0.732 (0.725−0.739)	0.811 (0.796−0.81)	0.18 (0.174−0.186)
	Mean	*0.736 (0.729−0.743)*	*0.734 (0.727−0.741)*	*0.726 (0.718−0.733)*	*0.734 (0.728−0.742)*	*0.81 (0.8−0.814)*	*0.179 (0.173−0.185)*
Prolonged LOS	XGBoost	0.796 (0.79−0.802)	0.83 (0.824−0.836)	0.407 (0.399−0.415)	0.586 (0.578−0.594)	0.736 (0.737−0.757)	0.126 (0.121−0.131)
	LightGBM	0.775 (0.768−0.782)	0.698 (0.691−0.705)	0.308 (0.301−0.315)	0.624 (0.616−0.632)	0.679 (0.668−0.69)	0.175 (0.169−0.181)
	CatBoost	0.788 (0.782−0.794)	0.772 (0.765−0.779)	0.259 (0.252−0.266)	0.64 (0.632−0.648)	0.718 (0.712−0.732)	0.129 (0.124−0.134)
	Random forest	0.786 (0.78−0.792)	0.816 (0.81−0.822)	0.372 (0.364−0.38)	0.596 (0.588−0.604)	0.742 (0.721−0.742)	0.128 (0.123−0.133)
	Mean	*0.786 (0.78−0.792)*	*0.779 (0.772−0.786)*	*0.336 (0.329−0.344)*	*0.612 (0.604−0.62)*	*0.719 (0.71−0.73)*	*0.14 (0.134−0.145)*
Prolonged ICU-LOS	XGBoost	0.784 (0.773−0.795)	0.776 (0.765−0.787)	0.327 (0.314−0.34)	0.615 (0.602−0.628)	0.674 (0.66−0.701)	0.129 (0.12−0.138)
	LightGBM	0.779 (0.768−0.79)	0.699 (0.687−0.711)	0.327 (0.314−0.34)	0.617 (0.604−0.63)	0.666 (0.651−0.69)	0.132 (0.123−0.141)
	CatBoost	0.775 (0.764−0.786)	0.765 (0.754−0.776)	0.219 (0.208−0.23)	0.599 (0.586−0.612)	0.682 (0.657−0.696)	0.13 (0.121−0.139)
	Random forest	0.779 (0.768−0.79)	0.727 (0.715−0.739)	0.325 (0.312−0.338)	0.616 (0.603−0.629)	0.675 (0.651−0.692)	0.13 (0.121−0.139)
	Mean	*0.779 (0.768−0.79)*	*0.742 (0.73−0.753)*	*0.3 (0.287−0.312)*	*0.612 (0.599−0.625)*	*0.674 (0.655−0.695)*	*0.13 (0.121−0.139)*
Major complications	XGBoost	0.909 (0.904−0.914)	0.943 (0.939−0.947)	0.121 (0.116−0.126)	0.51 (0.502−0.518)	0.704 (0.683−0.72)	0.05 (0.046−0.054)
	LightGBM	0.911 (0.906−0.916)	0.726 (0.719−0.733)	0.098 (0.093−0.103)	0.586 (0.578−0.594)	0.637 (0.618−0.658)	0.103 (0.098−0.108)
	CatBoost	0.904 (0.899−0.909)	0.896 (0.891−0.901)	0.06 (0.056−0.064)	0.531 (0.523−0.539)	0.645 (0.619−0.658)	0.051 (0.047−0.055)
	Random forest	0.904 (0.899−0.909)	0.936 (0.932−0.94)	0.102 (0.097−0.107)	0.514 (0.506−0.522)	0.691 (0.672−0.707)	0.05 (0.046−0.054)
	Mean	*0.907 (0.902−0.912)*	*0.875 (0.87−0.88)*	*0.095 (0.09−0.1)*	*0.535 (0.527−0.543)*	*0.669 (0.648−0.686)*	*0.064 (0.059−0.068)*

AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; CI, confidence interval.

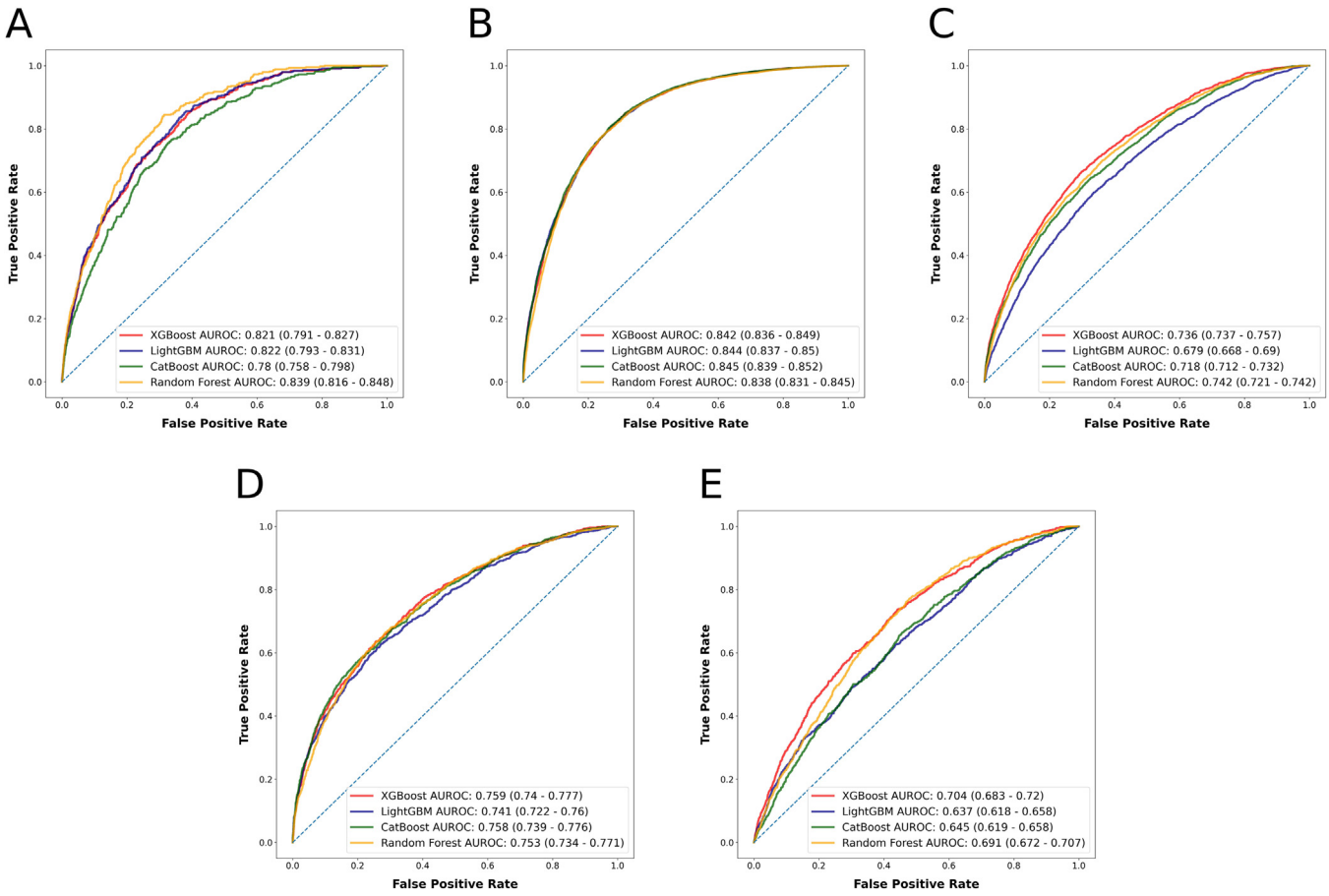

Fig. 3. (A). Algorithms' receiver operator curves for the outcome in-hospital mortality. (B). Algorithms' receiver operator curves for the outcome nonhome discharges. (C). Algorithms' receiver operator curves for the outcome prolonged length of stay. (D). Algorithms' receiver operator curves for the outcome prolonged length of intensive care unit stay. (E). Algorithms' receiver operator curves for the outcome major complications.

scenarios with imbalanced datasets since they do not consider the class distribution and may give a false sense of good performance by ignoring the minority class. Moreover, interpreting AUPRC can be challenging since its baseline is equal to the fraction of positive examples in the dataset, which can lead to significantly lower values than the AUROC, particularly for datasets with a low fraction of positive examples [37]. However, AUPRC may be more meaningful for a specific classification task. Despite this, it is often reported less frequently than AUROC due to its lower absolute values. In our study, the mean weighted AUPRC for predicting in-hospital mortality was 0.121, while the in-hospital mortality ratio was 0.032, representing the baseline. Finally, to assess the models' calibration, we used the Brier score, which measures the average squared difference between predicted and actual probabilities [31,38]. A well-calibrated model will have a Brier score close to zero, indicating that the predicted probabilities are very close to the actual probabilities.

We did not find any studies that presented ML models for predicting all adverse in-hospital outcomes we investigated after cSCI. However, some studies employed ML techniques to predict various outcomes following SCI. For example, Inoue et al. [25] evaluated the efficacy of ML algorithms in predicting neurological outcomes in patients with cSCI. The authors analyzed data from 165 patients with cSCI and used commonly utilized predictors such as demographics, magnetic resonance variables, and treatment strategies. The predictive tools used were XGBoost, logistic regression, and decision tree. The results showed that XGBoost had the highest accuracy (81.1%) and the second highest AUROC (0.867), followed by logistic regression and decision tree. Similarly, Fallah et al. [24] aimed to develop and validate a prognostic tool that could predict mortality following traumatic SCI. They developed the Spinal Cord Injury Risk Score (SCIRS) using ML techniques on patient-level data from 849 participants. The validation cohort consisted of 396 participants. The performance of SCIRS was compared with the ISS, a measure used to predict mortality following general trauma. The SCIRS was found to be more accurate than ISS in predicting both in-hospital and 1-year mortality following traumatic SCI. The AUROC for the SCIRS was 0.84 and 0.86 for 1-year mortality prediction in the development and validation cohorts, respectively. For in-hospital mortality, AUROC values were 0.87 and 0.85 for the development and validation

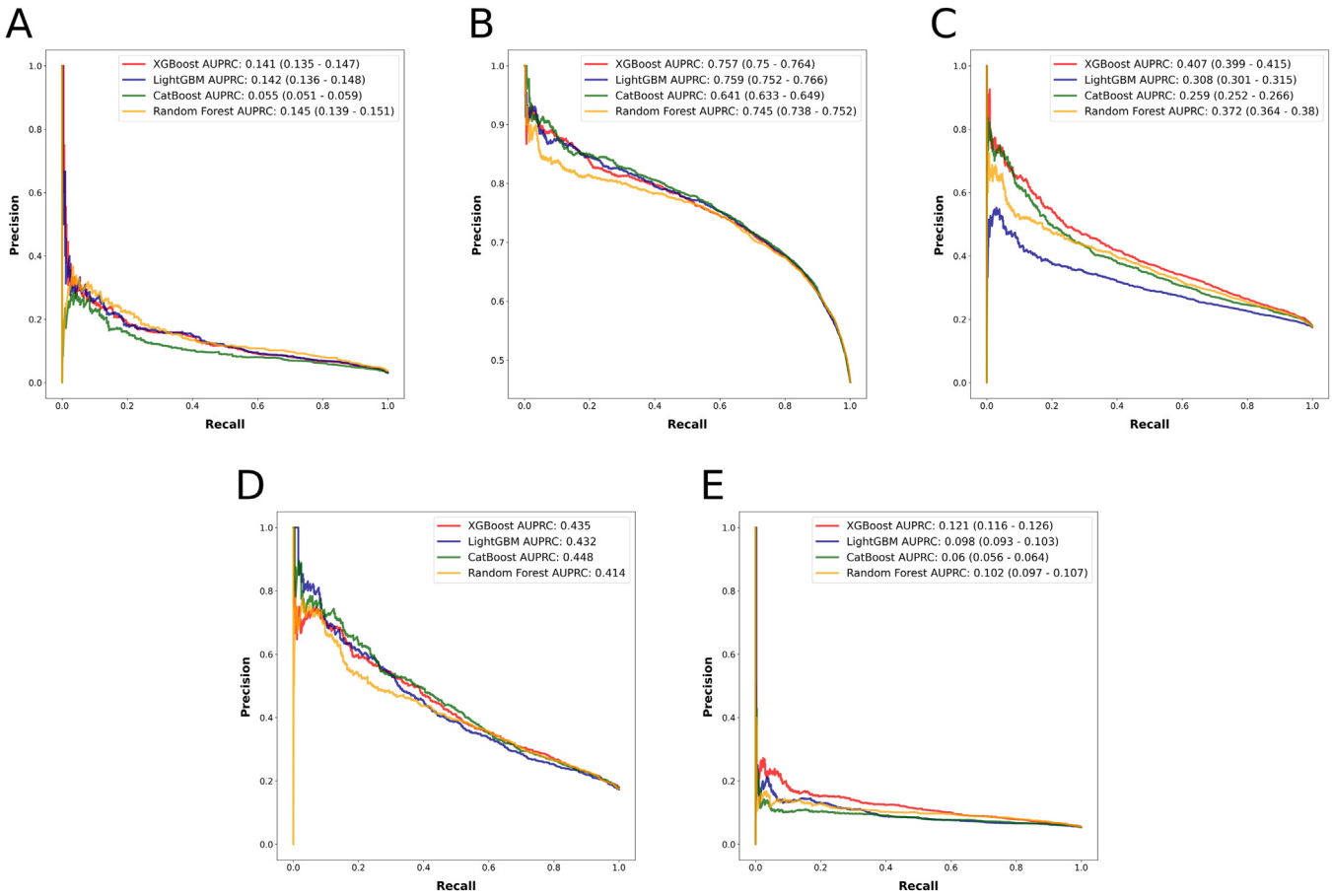

Fig. 4. (A). Algorithms' precision-recall curves for the outcome in-hospital mortality. (B). Algorithms' precision-recall curves for the outcome nonhome discharges. (C). Algorithms' precision-recall curves for the outcome prolonged length of stay. (D). Algorithms' precision-recall curves for the outcome prolonged length of intensive care unit stay. (E). Algorithms' precision-recall curves for the outcome major complications.

cohorts, respectively. Furthermore, Fan et al. aimed to develop ML classifiers to predict prolonged ICU-LOS and prolonged LOS in critical patients with SCI [23]. A total of 1,599 critical patients were included in the study, and data were extracted from two databases. The authors developed 91 initial ML classifiers, and the top three initial classifiers with the best performance were stacked into an ensemble classifier with a logistic regressor. The ensemble classifiers successfully predicted prolonged ICU-LOS and prolonged LOS, with AUROCs of 0.864 and 0.815, in the three-time five-fold cross-validation and 0.802 and 0.799, respectively, in independent testing.

Although the reported performance metrics were comparable with our study, these studies have some serious drawbacks. First, compared to our study, these models were developed using very small sample sizes. Developing ML-based clinical predictive models using small sample sizes can have several disadvantages. Firstly, small sample sizes can result in overfitting, where the model is optimized to perform well on the training data but does not generalize well to new data. This can lead to poor performance when the model is applied to real-world clinical settings. Secondly, small sample sizes can result in biased or incomplete data, which can affect the accuracy and generalizability of the model. For example, if the data is biased toward a particular demographic group, the model may not perform well on other groups. Lastly, small sample sizes can limit the complexity of the model that can be developed, as more complex models require larger sample sizes to learn and generalize well. Therefore, it is important to carefully consider the sample size when developing ML-based clinical predictive models to ensure that they are accurate, unbiased, and generalizable.

While our study provides a comprehensive approach to the application of ML in the context of traumatic cSCIs, we recognize several limitations that necessitate further research and refinement. Primarily, our study's population may not be entirely representative of all patients with traumatic cSCI. The data leveraged for our analysis was extracted from the ACS-TQP dataset, which primarily represents patients from hospitals equipped to meet the ACS-TQP reporting requirements, potentially leading to an overrepresentation of these specific hospitals. Thus, it is possible our dataset may inherently hold biases, which should be taken into account when interpreting the results. Moreover, our study's geographical limitation to the United States also narrows the applicability

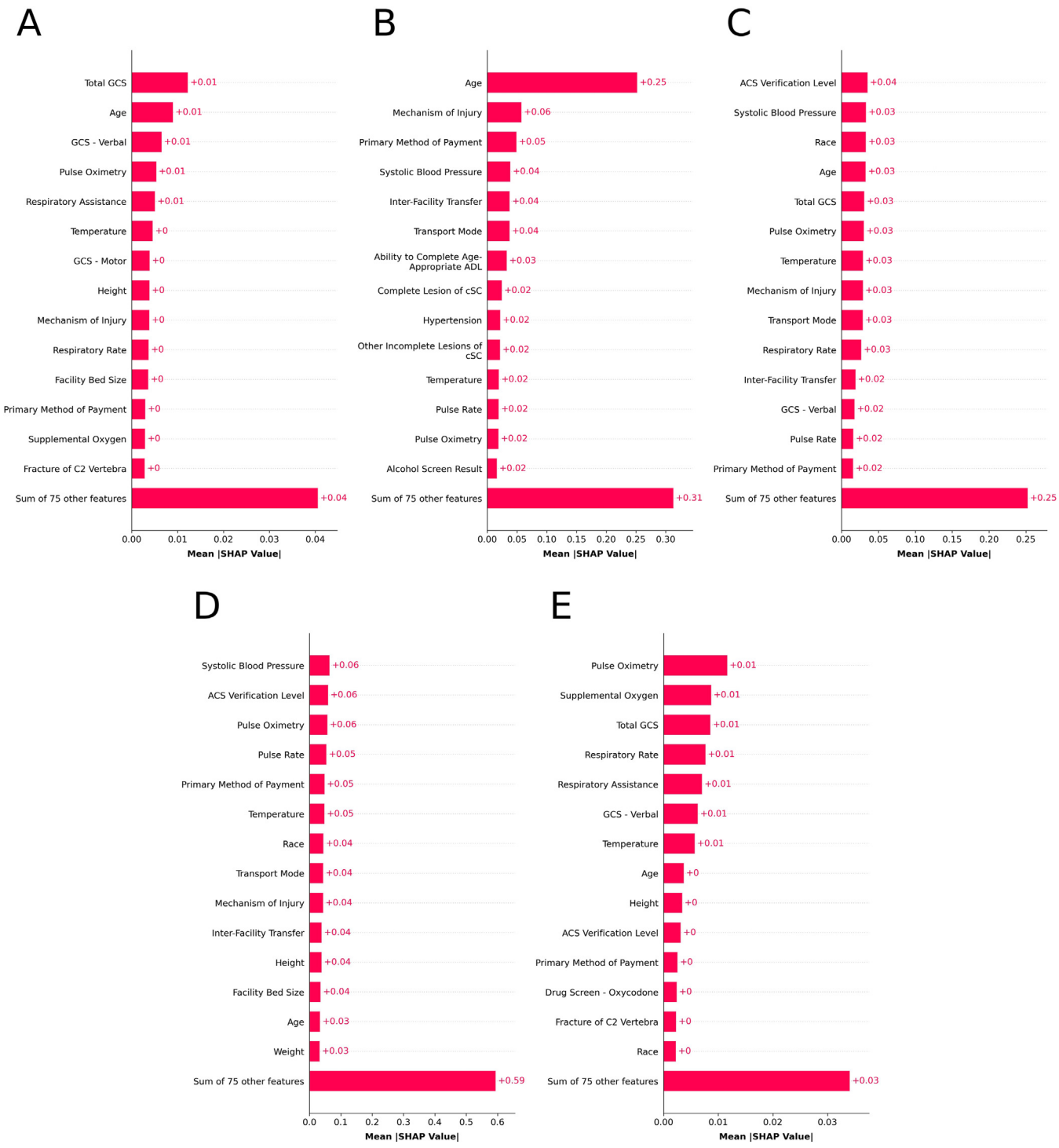

Fig. 5. (A). The fifteen most important features and their mean SHAP values for the model predicting the outcome in-hospital mortality with the Random Forest algorithm. (B). The fifteen most important features and their mean SHAP values for the model predicting the outcome nonhome discharges with the CatBoost algorithm. (C). The fifteen most important features and their mean SHAP values for the model predicting the outcome prolonged length of stay with the Random Forest algorithm. (D). The fifteen most important features and their mean SHAP values for the model predicting the outcome prolonged length of intensive care unit stay with the CatBoost algorithm. (E). The fifteen most important features and their mean SHAP values for the model predicting the outcome major complications with the XGBoost algorithm.

of our findings. In light of these limitations, it is essential to acknowledge that the outcomes may not be universally applicable or generalizable to different clinical environments across the globe. The scope of the data sources is another important limitation to consider. While the ACS-TQP dataset provides exhaustive information, the reliance on a singular

dataset potentially limits the broader applicability of our models. Therefore, external validation using independent datasets from diverse sources and geographical locations would further reinforce our models' robustness and generalizability. Additionally, the potential presence of coding errors and inaccuracies in large clinical databases, like the one used in our study, should be taken into consideration. For instance, the possibility of inaccuracies in the recording of patients' comorbidity information within the ACS-TQP database may impact the overall performance of our models. Finally, although our study offers a promising start, additional relevant variables such as detailed imaging parameters might improve the performance of our ML models. The inclusion of these specific variables could offer more nuanced and accurate predictions for individual patient outcomes, pushing the boundaries of precision medicine in the context of traumatic cSCIs. To summarize, while our study demonstrates the potential of ML in enhancing precision medicine for traumatic cSCIs, it is incumbent to conduct further research, including more diverse data sources and external validation, before our models can be fully integrated into clinical practice.

Conclusions

This study has demonstrated that ML algorithms can effectively predict in-hospital outcomes for patients with cSCI, and the development of a user-friendly web application makes the integration of these algorithms into clinical practice feasible. The results of this study show that ML algorithms can assist in risk stratification for cSCI patients, specifically for predicting in-hospital mortality and non-home discharges with good discriminatory power and prolonged LOS with fair discriminatory ability. While their performance in predicting prolonged ICU-LOS and major complications was relatively poor, the potential benefits of using ML algorithms to personalize care and predict outcomes for cSCI patients are significant. By providing visual explanations of the predictions, this tool can help establish confidence in the predictions and encourage their adoption in clinical practice. The incorporation of this tool can support quality assurance initiatives, guide treatment strategies, prioritize care, plan for discharge needs, facilitate shared decision-making, and ultimately improve patient outcomes.

Ethical approval

This study was deemed exempt from approval by the Icahn School of Medicine at Mount Sinai institutional review board because it involved analysis of deidentified patient data.

Data availability

Restrictions apply to the availability of these data. Data were obtained from the American College of Surgeons Trauma Quality Program and are available (https://www.facs.org/quality-programs/trauma/quality/tqp-participant-hub/) with the permission of the American College of Surgeons.

Code availability

The source code for preprocessing and analyzing the data is available on GitHub (https://github.com/mertkarabacak/TQP-cSCI).

Declarations of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research did not receive any external funding.

Supplementary materials

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.spinee.2023.08.009.

References

[1] Ahuja CS, Wilson JR, Nori S, Kotter MRN, Druschel C, Curt A, et al. Traumatic spinal cord injury. Nat Rev Dis Primers 2017;3:1–21. https://doi.org/10.1038/nrdp.2017.18.

[2] Middleton JW, Lim K, Taylor L, Soden R, Rutkowski S. Patterns of morbidity and rehospitalisation following spinal cord injury. Spinal Cord 2004;42:359–67. https://doi.org/10.1038/sj.sc.3101601.

[3] Silva V, Costa P, Pereira I, Faria R, Salgueira AP, Costa MJ, et al. Depression in medical students: insights from a longitudinal study. BMC Med Educ 2017;17:184. https://doi.org/10.1186/s12909-017-1006-0.

[4] Lee BB, Cripps RA, Fitzharris M, Wing PC. The global map for traumatic spinal cord injury epidemiology: update 2011, global incidence rate. Spinal Cord 2014;52:110–6. https://doi.org/10.1038/sc.2012.158.

[5] Migliorini C, Tonge B, Taleporos G. Spinal cord injury and mental health. Aust N Z J Psychiatry 2008;42:309–14. https://doi.org/10.1080/00048670801886080.

[6] Ugiliweneza B, Guest J, Herrity A, Nuno M, Sharma M, Beswick J, et al. A two-decade assessment of changing practice for surgical decompression and fixation after traumatic spinal cord injury−impact on healthcare utilization and cost. Cureus 2019;11:e6156. https://doi.org/10.7759/cureus.6156.

[7] National Spinal Cord InjuryStatistical Center. Spinal cord injury facts and figures at a glance. J Spinal Cord Med 2014;37:355–6. https://doi.org/10.1179/1079026814Z.000000000260.

[8] Kang Y, Ding H, Zhou H, Wei Z, Liu L, Pan D, et al. Epidemiology of worldwide spinal cord injury: a literature review. J Neurorestoratol 2020;6:1–9. https://doi.org/10.2147/JN.S143236.

[9] Chen Y, He Y, DeVivo MJ. Changing demographics and injury profile of new traumatic spinal cord injuries in the United States, 1972−2014. Arch Phys Med Rehab 2016;97:1610–9. https://doi.org/10.1016/j.apmr.2016.03.017.

[10] Strauss DJ, DeVivo MJ, Paculdo DR, Shavelle RM. Trends in life expectancy after spinal cord injury. Arch Phys Med Rehab 2006;87:1079–85. https://doi.org/10.1016/j.apmr.2006.04.022.

[11] Middleton JW, Dayton A, Walsh J, Rutkowski SB, Leong G, Duong S. Life expectancy after spinal cord injury: a 50-year study. Spinal Cord 2012;50:803–11. https://doi.org/10.1038/sc.2012.55.

[12] Hagen EM, Lie SA, Rekand T, Gilhus NE, Gronning M. Mortality after traumatic spinal cord injury: 50 years of follow-up. J Neurol, Neurosurg Psychiatry 2010;81:368–73. https://doi.org/10.1136/jnnp.2009.178798.

[13] Raju B, Jumah F, Ashraf O, Narayan V, Gupta G, Sun H, et al. Big data, machine learning, and artificial intelligence: a field guide for neurosurgeons. J Neurosurg 2020;1:1–11. https://doi.org/10.3171/2020.5.JNS201288.

[14] Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. N Engl J Med 2017;376:2507–9. https://doi.org/10.1056/NEJMp1702071.

[15] Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health 2019. https://doi.org/10.48550/arXiv.1806.00388.

[16] Bzdok D, Krzywinski M, Altman N. Machine learning: a primer. Nat Methods 2017;14:1119–20. https://doi.org/10.1038/nmeth.4526.

[17] Musolf AM, Holzinger ER, Malley JD, Bailey-Wilson JE. What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. Hum Genet 2022;141:1515–28. https://doi.org/10.1007/s00439-021-02402-z.

[18] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347–58. https://doi.org/10.1056/NEJMra1814259.

[19] Alabadla M, Sidi F, Ishak I, Ibrahim H, Affendey LS, Che Ani Z, et al. Systematic review of using machine learning in imputing missing values. IEEE Access 2022;10:44483–502. https://doi.org/10.1109/ACCESS.2022.3160841.

[20] He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30–6. https://doi.org/10.1038/s41591-018-0307-0.

[21] Buddhiraju A, Chen TL-W, Subih MA, Seo HH, Esposito JG, Kwon Y-M. Validation and generalizability of machine learning models for the prediction of discharge disposition following revision total knee arthroplasty. J Arthroplasty 2023;38(6S):S253–8. https://doi.org/10.1016/j.arth.2023.02.054.

[22] Dietz N, Jaganathan V, Alkin V, Mettille J, Boakye M, Drazin D. Machine learning in clinical diagnosis, prognostication, and management of acute traumatic spinal cord injury (SCI): a systematic review. J Clin Orthop Trauma 2022;35:102046. https://doi.org/10.1016/j.jcot.2022.102046.

[23] Fan G, Yang S, Liu H, Xu N, Chen Y, He J, et al. Machine learning-based prediction of prolonged intensive care unit stay for critical patients with spinal cord injury. Spine (Phila Pa 1976) 2022;47:E390–8. https://doi.org/10.1097/BRS.0000000000004267.

[24] Fallah N, Noonan VK, Waheed Z, Rivers CS, Plashkes T, Bedi M, et al. Development of a machine learning algorithm for predicting in-hospital and 1-year mortality after traumatic spinal cord injury. Spine J 2022;22:329–36. https://doi.org/10.1016/j.spinee.2021.08.003.

[25] Inoue T, Ichikawa D, Ueno T, Cheong M, Inoue T, Whetstone WD, et al. XGBoost, a machine learning method, predicts neurological recovery in patients with cervical spinal cord injury. Neurotrauma Reports 2020;1:8–16. https://doi.org/10.1089/neur.2020.0009.

[26] Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13:1. https://doi.org/10.1186/s12916-014-0241-z.

[27] Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A Multidisciplinary View. J Med Internet Res 2016;18:e323. https://doi.org/10.2196/jmir.5870.

[28] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Jair 2002;16:321–57. https://doi.org/10.1613/jair.953.

[29] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework 2019. https://doi.org/10.48550/arXiv.1907.10902.

[30] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classifiers 1999;10:61–74.

[31] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning - ICML '05. Bonn, Germany: ACM Press; 2005. p. 625–32. https://doi.org/10.1145/1102351.1102430.

[32] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017. p. 4765–74.

[33] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36. https://doi.org/10.1148/radiology.143.1.7063747.

[34] Feng Y, Zhou M, Tong X. Imbalanced classification: a paradigm-based review. Stat Anal Data Min 2021;14:383–406. https://doi.org/10.1002/sam.11538.

[35] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst, Man, Cybern C 2012;42:463–84. https://doi.org/10.1109/TSMCC.2011.2161285.

[36] Mullick SS, Datta S, Dhekane SG, Das S. Appropriateness of performance indices for imbalanced data classification: an analysis 2020. https://doi.org/10.48550/ARXIV.2008.11752.

[37] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 2015;10:e0118432. https://doi.org/10.1371/journal.pone.0118432.

[38] On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative, Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.