Comparison of machine learning models to predict long-term outcomes after severe traumatic brain injury

*Dooman Arefan, PhD, Matthew Pease, MD, Shawn R. Eagle, PhD, David O. Okonkwo, MD, PhD, and Shandong Wu, PhD

OBJECTIVE An estimated 1.5 million people die every year worldwide from traumatic brain injury (TBI). Physicians are relatively poor at predicting long-term outcomes early in patients with severe TBI. Machine learning (ML) has shown promise at improving prediction models across a variety of neurological diseases. The authors sought to explore the following: 1) how various ML models performed compared to standard logistic regression techniques, and 2) if properly calibrated ML models could accurately predict outcomes up to 2 years posttrauma.

METHODS A secondary analysis of a prospectively collected database of patients with severe TBI treated at a single level 1 trauma center between November 2002 and December 2018 was performed. Neurological outcomes were assessed at 3, 6, 12, and 24 months postinjury with the Glasgow Outcome Scale. The authors used ML models including support vector machine, neural network, decision tree, and naive Bayes models to predict outcome across all 4 time points by using clinical information available on admission, and they compared performance to a logistic regression model. The authors attempted to predict unfavorable versus favorable outcomes (Glasgow Outcome Scale scores of 1–3 vs 4–5), as well as mortality. Models’ performance was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC) with 95% confidence interval and balanced accuracy.

RESULTS Of the 599 patients in the database, the authors included 501, 537, 469, and 395 at 3, 6, 12, and 24 months posttrauma. Across all time points, the AUCs ranged from 0.71 to 0.85 for mortality and from 0.62 to 0.82 for unfavorable outcomes with various modeling strategies. Decision tree models performed worse than all other modeling approaches for multiple time points regarding both unfavorable outcomes and mortality. There were no statistically significant differences between any other models. After proper calibration, the models had little variation (0.02–0.05) across various time points.

CONCLUSIONS The ML models tested herein performed with equivalent success compared with logistic regression techniques for prognostication in TBI. The TBI prognostication models could predict outcomes beyond 6 months, out to 2 years postinjury.

KEYWORDS severe traumatic brain injury; machine learning; predictive modeling; Glasgow Outcome Scale
challenge, researchers developed multivariate logistic regression models such as the International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT) and Corticosteroid Randomisation After Significant Head Injury (CRASH). These models are not in routine use in clinical care, in large measure because they were designed for research purposes and require tedious data entry. Additionally, both CRASH and IMPACT models make predictions at discharge from the hospital or at 6 months posttrauma, a time point that may be too early to capture the full recovery potential of patients with severe TBI. Recent work has shown that patients with severe TBI continue to experience gains in recovery at 2 years postinjury.

Machine learning (ML) is a field of computer science that trains computers to perform tasks by observing trends and rules in large data sets. ML-based models have improved prognostication compared to traditional techniques, especially in patients with TBI. To better understand the predictive abilities of ML in TBI, we compared the performance of a variety of ML models for tabular data (not deep learning) in a large, prospective cohort of patients with severe TBI who received systematic follow-up through 2 years posttrauma. We hypothesized that 1) advanced ML techniques could improve on basic ML techniques (i.e., logistic regression), and 2) these techniques could be extended past 6 months, the time limit at which CRASH and IMPACT models were validated.

Methods

This study received approval from the University of Pittsburgh Human Research Protection Office (Biomedical and genetic analysis of biospecimens following head injury; most recent approval June 29, 2022) with consent from subjects’ legal representatives. Research procedures followed the ethical standards of the responsible committee on human experimentation and complied with the Helsinki Declaration of 1975. We adhered to the Standards for Reporting of Diagnostic Accuracy Studies (STARD).

Study Cohort

We performed an analysis of a prospectively collected database of consecutive patients with severe TBI treated at a single level 1 trauma center between November 2002 and December 2018. This database has previously been described and includes patients aged 16–80 years with severe TBI from blunt trauma (defined as a postresuscitation Glasgow Coma Scale [GCS] score of 8 or lower); it excludes patients with imminent brain death, pregnancy, and penetrating TBI. Our database includes patients regardless of cranial injury pattern (i.e., epidural, subdural, contusion) or whether the patient required surgery. Neurological outcomes were assessed at 3, 6, 12, and 24 months postinjury based on a structured interview conducted by trained neuropsychologists using the Glasgow Outcome Scale (GOS), in which 1 = death; 2 = vegetative state; 3 = severe disability; 4 = moderate disability; and 5 = low/no disability.

Despite our database spanning nearly 2 decades, we had similar care patterns throughout. All patients were promptly evaluated in the emergency room and offered craniotomy/craniectomy if deemed appropriate by the on-call neurosurgeon. After admission, a subspecialized neurotrauma neurosurgeon assumes care for all patients with severe TBI and guides management with a multidisciplinary team of critical care physicians and general trauma surgeons. This role for the neurosurgeon at the University of Pittsburgh Medical Center was filled by one person for approximately the first quarter of the study period; this person subsequently left and was replaced by another person (D.O.O.) for the remainder of the study period. Across the entire study period, the practice at our institution was to place both an intraparenchymal intracranial pressure (ICP) and brain tissue oxygenation monitor, as well as an external ventricular drain. Both an ICP monitor and external ventricular drain are placed to allow for continuous pressure monitoring and CSF diversion, which is associated with improved control of ICP. Our multidisciplinary teams used guided medical and surgical therapies to attempt to maintain normal brain tissue oxygenation and ICP.

Clinical and qualitative radiographic covariates used in both IMPACT and CRASH models, as well as some demographic information, were collected from the prospectively assembled database. Missing information, including all radiographic descriptors, was retrospectively collected through chart review if available. Our variable list includes age, race, sex, mechanism of injury, pupil reactivity, GCS score, GCS motor score, hypoxia, hypotension, Marshall CT score, presence of traumatic subarachnoid hemorrhage, epidural hematoma, glucose, and hemoglobin. Patients with any missing variables were excluded from our study cohort.

We included the same variable set as IMPACT and CRASH analyses, with sex and race added, because IMPACT and CRASH are the two most widely used and validated prognostic models in TBI. Both of these models use admission information to prognosticate, which does limit the variable set. Potentially important variables, such as measurements of ICP, brain tissue oxygenation, or hematoma expansion, are not available on admission. We have previously validated CRASH and IMPACT models on our data set, and our validation had similar performance to that in other groups, highlighting the utility of this variable set in our population.

Model Building

We built and tested multiple ML models at various time points posttrauma. Similarly to IMPACT and CRASH models, we attempted to predict favorable (GOS score 4–5) versus unfavorable (GOS score 1–3) outcomes, as well as mortality (GOS score 1) versus survival (GOS score 2–5). Based on our repeated-measure design, we made independent predictions at 3, 6, 12, and 24 months postinjury. Our goal was to test a variety of different modeling structures at various time points against standard logistic regression models. Of note, although logistic regression is an ML modeling approach, we are considering this the standard technique. Any patients with missing outcome data, as well as those with any missing model input data, were excluded from model building.

We developed our own logistic regression models, rather than comparing our models to CRASH and IMPACT, for two reasons. First, IMPACT and CRASH models were
developed in large, multiinstitutional data sets. We wanted
to ensure that the performance of our ML models was not
artificially improved through overfitting single-institution
data when comparing to our base logistic regression model.
Thus, by creating an in-house logistic regression model
built on available data, we created a fairer comparison.
Second, we wanted to use the same set of predictors for
both our logistic regression models and our ML models.
The CRASH and IMPACT models do not include all of
the covariates we included—such as total GCS score, race,
sex, and mechanism of injury. We have previously vali-
dated CRASH and IMPACT models on our data and have
reported the results. 17,25,26
We built prediction models using the following ML
techniques: logistic regression (the base model), support
vector machine (SVM), neural network (NN), decision
tree (DT), and naïve Bayes (NB). 1) For logistic regres-
sion, we used a generalized linear model that contained an
intercept and linear term for each independent variable. 2)
We used a linear kernel with automatic hyperparameter
optimization in the SVM technique. 3) An NN with 10
hidden layers, 1000 maximum epochs, Levenberg-Mar-
quardt training function, and early stopping regularization
technique was used to improve generalization. 27 4) We
used a standard classification and regression tree (CART)
algorithm for DT, and the minimum leaf size, maximum
number of splits, and split criteria were selected using the
automatic hyperparameter optimization technique in the
modeling. 5) A normal kernel smoother (gaussian) was
used in the NB technique.
The least absolute shrinkage and selection operator
(LASSO) technique was used to select a subset of inde-
dependent variables. A subset with minimum mean square
error on the training set was used as input for the pre-
diction models. All coding was done using the statistics
and ML toolbox in MATLAB (R2020b, The MathWorks,
Inc.).

Statistical Analysis
For each model, its performance was evaluated using
the 5-fold cross-validation technique to increase the ro-
 bustness of the results. A randomized stratified method
was used to split data at the patient level into training and
test sets in each fold of the 5-fold cross-validation. We re-
ported the area under the receiver operating characteristic
(ROC) curve (AUC) with the 95% confidence interval and
balanced accuracy. Balanced accuracy was calculated us-
ing arithmetic averaging of sensitivity and specificity. Ac-
curacy is not an appropriate metric to evaluate a model's
performance when we use an imbalanced data set in a bi-
nary classification task. 28
The p values were calculated to compare AUCs to the
base AUC (logistic regression model) by using a 5-fold
cross-validated paired t-test. We controlled the false dis-
covery rate by using the Benjamini-Hochberg procedure. 29
In addition, the contribution of each clinical variable to
the prediction of outcomes at 6 and 24 months for both
mortality and unfavorable outcome based on SVM mod-
els was evaluated using the Shapley Additive Explan-
tations (SHAP) approach. 30 SHAP is a common, post hoc,
explainable ML technique that identifies the relative im-
portance of various model features (i.e., input variables).
A higher Shapley value indicates a higher contribution to
prediction of the outcome.

Results
We identified 599 patients in our prospective database
who were eligible for our study. We excluded patients from
model building if the following information was missing:
demographics, clinical covariates, qualitative radiograph-
ic descriptors, and GOS scores. A total of 501 (84%), 537
(90%), 469 (78%), and 395 (66%) patients remained who
had 3-, 6-, 12-, and 24-month GOS scores and complete
model input data, respectively (Fig. 1). Table 1 displays

FIG. 1. CONSORT diagram for our cohort in this study.
the demographic, clinical, and radiographic information, as well as the GOS scores for patients used in model building at all time points. Specifically, this includes all patients with complete outcomes and model input data.

Overall, we tested 5 models, including the base logistic regression model, at 4 different time points (Tables 2 and 3). Our AUCs ranged from 0.71 to 0.85 for mortality and from 0.62 to 0.82 for unfavorable outcomes with various modeling strategies across all 4 time points. Across all time points, the logistic regression model either performed best, albeit nonsignificantly, or had very similar performance to the best model. The DT model was significantly worse than the other models at most time points (p < 0.05 for both mortality and unfavorable outcome prediction at 6- and 24-month follow-ups, as well as for mortality prediction at 3-month follow-up). Our models had little variation between time points, with most models differing by 0.02–0.05 points over time. Figures 2 and 3 display the ROC curves for our various models and for mortality and unfavorable outcome predictions, respectively.

### Table 1. Model input and outcome data at various time points posttrauma for all patients in cohort

<table>
<thead>
<tr>
<th>Variable</th>
<th>Time Post-TBI</th>
<th>3 Mos</th>
<th>6 Mos</th>
<th>12 Mos</th>
<th>24 Mos</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of patients*</td>
<td></td>
<td>501</td>
<td>537</td>
<td>469</td>
<td>395</td>
</tr>
<tr>
<td>Demographic information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (yrs) ± SD</td>
<td></td>
<td>40 ± 17</td>
<td>40 ± 17</td>
<td>41 ± 17</td>
<td>42 ± 17</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td>7% (37)</td>
<td>7% (38)</td>
<td>7% (32)</td>
<td>6% (23)</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>91% (455)</td>
<td>91% (490)</td>
<td>91% (429)</td>
<td>92% (364)</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>2% (9)</td>
<td>2% (9)</td>
<td>2% (8)</td>
<td>2% (8)</td>
</tr>
<tr>
<td>Male sex</td>
<td></td>
<td>78% (391)</td>
<td>78% (422)</td>
<td>78% (364)</td>
<td>77% (306)</td>
</tr>
<tr>
<td>Clinical covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCS score ± SD</td>
<td></td>
<td>5.5 ± 1.7</td>
<td>5.5 ± 1.7</td>
<td>5.4 ± 1.8</td>
<td>5.3 ± 1.8</td>
</tr>
<tr>
<td>GMS ± SD</td>
<td></td>
<td>3.1 ± 1.7</td>
<td>3.1 ± 1.7</td>
<td>3.0 ± 1.7</td>
<td>3.0 ± 1.7</td>
</tr>
<tr>
<td>Glucose (mg/dL) ± SD</td>
<td></td>
<td>163 ± 63</td>
<td>162 ± 62</td>
<td>164 ± 64</td>
<td>166 ± 63</td>
</tr>
<tr>
<td>Hb (g/dL) ± SD</td>
<td></td>
<td>13 ± 2.1</td>
<td>13 ± 2.2</td>
<td>13 ± 2.2</td>
<td>13 ± 2.1</td>
</tr>
<tr>
<td>Pupil reactivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both</td>
<td></td>
<td>65% (328)</td>
<td>65% (351)</td>
<td>64% (299)</td>
<td>60% (239)</td>
</tr>
<tr>
<td>One</td>
<td></td>
<td>9% (48)</td>
<td>9% (51)</td>
<td>10% (48)</td>
<td>11% (42)</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>25% (125)</td>
<td>25% (135)</td>
<td>26% (122)</td>
<td>29% (114)</td>
</tr>
<tr>
<td>Radiographic descriptors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypoxia</td>
<td></td>
<td>17% (86)</td>
<td>17% (90)</td>
<td>18% (86)</td>
<td>18% (71)</td>
</tr>
<tr>
<td>Hypotension</td>
<td></td>
<td>25% (128)</td>
<td>26% (139)</td>
<td>26% (124)</td>
<td>26% (104)</td>
</tr>
<tr>
<td>Marshall CT score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>4% (21)</td>
<td>4% (21)</td>
<td>3% (14)</td>
<td>3% (12)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>49% (245)</td>
<td>50% (270)</td>
<td>48% (223)</td>
<td>46% (180)</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>12% (59)</td>
<td>11% (62)</td>
<td>13% (60)</td>
<td>13% (50)</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>7% (37)</td>
<td>7% (40)</td>
<td>7% (35)</td>
<td>9% (35)</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>22% (113)</td>
<td>22% (118)</td>
<td>24% (111)</td>
<td>23% (93)</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>5% (26)</td>
<td>5% (26)</td>
<td>5% (26)</td>
<td>6% (25)</td>
</tr>
<tr>
<td>tSAH</td>
<td></td>
<td>82% (410)</td>
<td>82% (439)</td>
<td>83% (388)</td>
<td>84% (331)</td>
</tr>
<tr>
<td>Epidural mass</td>
<td></td>
<td>11% (55)</td>
<td>11% (58)</td>
<td>11% (54)</td>
<td>12% (47)</td>
</tr>
<tr>
<td>Outcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOS score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>40% (199)</td>
<td>39% (208)</td>
<td>45% (213)</td>
<td>55% (216)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>5% (26)</td>
<td>2% (11)</td>
<td>2% (8)</td>
<td>0.5% (2)</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>36% (180)</td>
<td>30% (163)</td>
<td>21% (100)</td>
<td>16% (62)</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>16% (78)</td>
<td>18% (99)</td>
<td>16% (76)</td>
<td>15% (58)</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>4% (18)</td>
<td>10% (56)</td>
<td>15% (72)</td>
<td>14% (57)</td>
</tr>
</tbody>
</table>

GMS = GCS motor score; Hb = hemoglobin; tSAH = presence of traumatic subarachnoid hemorrhage.

* See Fig. 1 for explanation of total patient numbers for each time point.
Table 4 shows selected clinical variables assessed using the 5-fold cross-validated LASSO technique for 3-, 6-, 12-, and 24-month follow-ups and for both mortality and unfavorable outcome predictions.

We implemented SHAP on the best-performing model—i.e., SVM—for both mortality and unfavorable outcome predictions at 6 and 24 months, to understand the relative importance of various model features. As shown in Fig. 4, age, GCS score, and Marshall CT score had the greatest contribution for both predictions at 6 and 24 months.

### Discussion
Analyzing a prospective cohort of nearly 600 patients with severe TBI from a single institution, we demonstrated that multiple ML models (SVM, NN, NB) can obtain favorable performance, with AUCs > 0.80 for predicting both mortality and unfavorable outcomes.
mortality and unfavorable outcomes up to 2 years post-trauma. Over the past several years, ML techniques have allowed for the superior ability to diagnose and prognosticate neurological disease. IMPACT and CRASH are the two most widely used prognostication models in moderate and severe TBI, although they are rarely used clinically due to tedious data entry requirements and distrust of models designed for research purposes. We attempted to show that ML models could improve performance and provide a viable alternative to IMPACT and CRASH, but found that standard logistic regression techniques, as used in IMPACT and CRASH, performed equally well or better than our alternative ML models. These findings complement those of Gravesteijn et al., who similarly found that logistic regression models performed well compared to ML approaches. Our results stand in contrast to other works that found high performance for ML models. Recognizing the difficulties of early prognostication, including with modeling, many societies advocate for a wait and see approach for prognostication of outcome in comatose patients.

Similarly to IMPACT and CRASH, our ML models may have limited abilities to predict long-term outcomes accurately by only using information available on admission. Whereas predicting outcomes on admission may be beneficial in research studies, this time point may be too early for accurate prognostication. Physicians typically lack the ability to accurately predict outcomes early, across a variety of diseases, for patients in a comatose state. Modeling may face similar challenges for severe TBI, in which the average patient with a favorable outcome spends nearly 2 weeks in a coma. The IMPACT and CRASH models, built using logistic regression techniques similar to the ones that we used, account for only 20%–30% of the variability in mortality and have poor discrimination for most cases. Recognizing the difficulties of early prognostication, including with modeling, many societies advocate for a wait and see approach for prognostication of outcome in comatose patients.

FIG. 2. ROC curves for mortality prediction. This figure displays the ROC curves at each time point for the 4 models: 3-month (A), 6-month (B), 12-month (C), and 24-month (D) follow-ups. LR = logistic regression.
FIG. 3. ROC curves for unfavorable outcome prediction. This figure displays the ROC curves at each time point for the 4 models: 3-month (A), 6-month (B), 12-month (C), and 24-month (D) follow-ups.

TABLE 4. Selected clinical variables assessed using 5-fold cross-validated LASSO technique for 3-, 6-, 12-, and 24-month follow-ups and both mortality and unfavorable outcome predictions

<table>
<thead>
<tr>
<th>Feature</th>
<th>3 Mos Mortality</th>
<th>3 Mos Unfavorable</th>
<th>6 Mos Mortality</th>
<th>6 Mos Unfavorable</th>
<th>12 Mos Mortality</th>
<th>12 Mos Unfavorable</th>
<th>24 Mos Mortality</th>
<th>24 Mos Unfavorable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pupil</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GCS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GMS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Hypoxia</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Hypotension</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Marshall CT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>tSAH</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Epidural</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Glucose</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Hb</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Age</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Race</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Sex</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MoI</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

Epidural = epidural hematoma; MoI = mechanism of injury; pupil = pupil reactivity; ✓ = selected; ✗ = not selected.
One major contribution of our work is demonstrating that, when properly calibrated, models can successfully predict outcomes past the 2-week or 6-month time point typically used in prognostic models of TBI. Recent work has demonstrated that patients with severe TBI have a remarkable ability to improve functional outcomes, and this recovery continues for up to 2 years postinjury. By predicting outcomes at only 6 months, many modeling approaches are underestimating the potential for patients to achieve favorable recovery. Here we have shown that advanced modeling techniques, albeit with similar performance to logistic regression models, can capture the full recovery from TBI in patients with recorded long-term outcomes without significant reduction in performance.

Our work has several limitations. First, we performed a retrospective, single-center study. Despite this, we used a large, prospective data set that we previously validated using the IMPACT model. Second, we did not use every ML modeling approach. However, we used the most common supervised algorithms. Deep learning was not used in this study because the input is a small set of tabular variables. Additionally, we used a popular feature selection technique, LASSO, to automatically select more important features to reduce feature dimension and overfitting in the modeling. The LASSO technique has few constraints; for example, it is less stable when trained on a data set with correlated features. Given that the number of features (14 clinical variables) in our study was much lower than in some radiomics studies, in which hundreds of computer-extracted features are used, the constraints of LASSO were of less concern in our study. Missing data were observed for each follow-up time point, with more missing data further away in time from injury. This is typical of research on severe TBI, and our follow-up rates were consistent with overall trends in the broader literature. Last, we chose to develop a model using only information available on admission, similar to CRASH and IMPACT. This time point may be too early to accurately prognosticate in patients with TBI, given that CRASH and IMPACT only account for one-third of the variation in outcomes. Previous work has demonstrated that additional information obtained after admission may improve prediction performance.

![FIG. 4. Average 5-fold cross-validated Shapley values of clinical variables assessed using the SVM model for 6-month (left) and 24-month (right) outcome predictions. Clinical variables are sorted from largest to smallest absolute Shapley value. Upper Row: Mortality prediction. Lower Row: Unfavorable outcome prediction. Epidural = epidural hematoma; GMS = GCS motor score; Hb = hemoglobin; Mol = mechanism of injury; Pupil = pupil reactivity; tSAH = presence of traumatic subarachnoid hemorrhage.](image-url)
Conclusions

We compared the performance of multiple ML approaches to a standard logistic regression model in a prospective cohort of patients with severe TBI and found equivalence among ML models. When properly calibrated, TBI prognostication models can predict outcomes out to 2 years postinjury.

Acknowledgments

Dr. Wu reports grants from the NIH/National Institute of Biomedical Imaging and Bioengineering (grant no. 1R01EB032896) and from Amazon Machine Learning Research Award outside the submitted work. Dr. Wu also received funding from the University of Pittsburgh. Dr. Pease reports funding from the Congress of Neurological Surgeons (CNS).

References

33. Muhlestein WE, Akagi DS, McManus AR, Chambless LB. Machine learning ensemble models predict total charges and


**Disclosures**

The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

**Author Contributions**

Conception and design: Wu, Arefan, Pease, Okonkwo. Acquisition of data: Pease, Okonkwo. Analysis and interpretation of data: Wu, Arefan, Pease, Eagle. Drafting the article: Arefan, Pease. Critically revising the article: all authors. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Wu. Statistical analysis: Arefan. Administrative/technical/material support: Wu, Okonkwo. Study supervision: Wu, Okonkwo.

**Supplemental Information**

**Data Availability**

Data are available to qualified researchers upon request.

**Correspondence**

Shandong Wu: University of Pittsburgh, PA. wus3@upmc.edu.