Predicting Pediatric Traumatic Brain Injury Mortalities

Abstract

Traumatic Brain Injury (TBI) is a widespread public health issue from which thousands of individuals suffer each year. These injuries are especially problematic for children whose brains are not yet fully developed, but despite concern there currently exist few TBI outcome prediction and triage methods that can be used with this demographic. Using pediatric patients entered into the National Trauma Data Bank with head injuries (n = 147,452), we construct a new TBI mortality predictive model via multiple logistic regression built specifically for this demographic with an emphasis on reducing high false negative rates brought on by an imbalanced dataset without excessively introducing false positives. Model performance was evaluated with a variety of measurements and results showed our proposed model to outperform the the gold standard for trauma injury outcome prediction. We also present a computer application of the model in an effort to increase accessibility.

Introduction

Traumatic Brain Injury (TBI) is a widespread public health issue from which thousands of individuals in the United States (US) and around the globe suffer each year. A summary of the frequency of these injuries over the course of 7 years in the US is presented in Table 1. Overall, TBI only accounts for 1-2% of all injury-related ED visits. However, even with this relatively low occurrence rate, these injuries has been previously measured as the cause of nearly 30% of all injury-related deaths [1, 2]. The severity of these injuries have broad range but are often categorized into one of three groups based on a measure of consciousness known as the Glasgow Coma Score (GCS) (a scale from 3-worst to 15-best) where mild injuries (GCS $>$ 13) are associated with short-term memory loss and severe (GCS $<$ 9) with comatose individuals [3, 4].

Table 1: **Emergency department injury visits in the United States (2008–2015)**

Year	Number of injury-related visits (in thousands)	Number of TBI-related visits (in thousands)	Percent of TBI-related visits
2015	38959	556	1.4
2014	40019	573	1.4
2013	37211	381	1.0
2012	37427	436	1.2
2011	40220	371	0.9
2010	37878	355	0.9
2009	45420	507	1.1
2008	42520	485	1.1

But even with an understanding of the possible long-term consequences that may arise with TBI, standardized approaches towards the treatment of these injuries are still well unestablished and it remains difficult to estimate outcomes [5]. Existing methods for such predictions are often criticized as outdated or suffer from a lack of usability in the pediatric cohort. For example, the Trauma and Injury Severity Score (TRISS) estimates an individual's probability of mortality after suffering any traumatic injury with a logistic regression model that originally formulated in the 1980s [6–12]. TRISS has been previously hailed as the gold standard for trauma injury outcome prediction (which ultimately lead to its widespread adoption in many clinics), but more than 30 years after its reception it has been failed to be updated with modern patient data [13, 14]. In addition, two relatively new models designed specifically for TBI, the Corticosteroid Randomization After Significant Head Injury (CRASH) and International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT) [15] models, were among the first TBI specific models to be built on relatively large data and externally validated by multiple countries, but both completely disregarded the pediatric cohort [16–19]. This is especially concerning for said younger individuals, as they can account for upwards of 75% of all TBI cases in certain countries and whose brains are still undergoing significant development [20]. But to worsen the situation, the rates of mortality due to TBI among children are much lower [21] than those seen in adults (Table 2), so any hope of attempting to use the adult models to accurately predict the future state of pediatric patients is unlikely. Furthermore, the rates of TBI related ED visits have shown to be increasing with the Center for Disease Control and Prevention (CDC) reporting a near doubling in the per 100,000 rate of TBI related ED visits from the period of 2001–2008 to 2009–2010 in the US [22].

The low mortality rate of TBI related injuries in children does more than potentially disable the use

Table 2: **Rates (per 100,000) of TBI-related ED visits, 2001–2010**

Age group (yrs)	ED Visits	Hospitalizations	Deaths
0–4	1112.6–2193.8	57.7–78.7	**4.3–5.2**
5–14	498.8–888.7	23.1–54.5	**1.9–3.2**
15–24	576.9–981.9	81.2–126.6	15.6–23.4
25–44	320.3–470.0	65.3–76.4	14.6–17.6
45–64	164.8–328.2	60.1–83.9	17.5–18.1
65+	293.3–603.3	191.5–294.0	41.2–45.2
all	420.6–715.7	82.7–98.7	**17.1–18.6**

of adult models as it also makes modeling such events with new models more difficult. As shown in Table 2 children of age 0–14 suffered from a mortality rate of 1.9–5.2 per 100,000 during the period of 2001–2010 in the US [21], and when there exist rare events in data (i.e., low mortality rate) many statistical models do not perform well. King and Zeng discussed this issue for logistic regression models in detail, and pointed out that the prevalence of rare outcomes can be severely underestimated [24].

Consider Table 3 as an example to describe our concern. Table 3 was calculated based on the values in St-Louis and colleagues in an evaluation of TRISS on their data [25]. In this result, sensitivity and specificity were 0.5% and 99.9%, respectively. The error rate was 12.17% which is acceptable in most studies. However, the low sensitivity indicates that this model severely underestimates trauma mortality. Therefore, using sensitivity, specificity, or error rate as sole performance measurements is not recommended when choosing the best prediction model. Optimizing both sensitivity and specificity might not be feasible, since there is a trade-off between these two measures, but in clinical situations and especially in the case of TBI where injuries are not always clearly visible and whose impacts might not be immediately felt, it is crucially important to predict non-survivors as non-survivors correctly when compared with survivors as survivors.

Table 3: **TRISS performance**

		Truth	
		Non-survivor	Survivor
Prediction	Non-survivor	45	65
	Survivor	8868	64441

To tackle the aforementioned problems with existing methods and the pediatric cohort, in this paper, we construct the TBI Mortality Prediction model for Pediatric patients (TMPP) which is built on the largest trauma data bank available in North America with the following contributions: (1) TMPP is specifically designed to target pediatric patients with TBI, and (2) TMPP highlights the importance of predicting true non-survivors correctly.

Furthermore, a Java application of the model is presented to help with clinical accessibility.

Materials and methods

Software

All data analyses were performed with the statistical software platform R [26]. Java version 8 update 171 was used to build the user-friendly calculator application that will be described in the result section.

Data

Head injury patients aged 14 or less were extracted from the 2010-2015 National Trauma Data Bank Research Data Set (NTDB RDS). Head injuries were identified using a regional AIS (Abbreviated Injury Scale) score (e.g., severity score in the head region ≥ 1). Burn victims and those with unspecified severity were excluded. A total number of 147,452 pediatric trauma patients with head injuries were used for our study. The outcome of interest was the patient's 14-day mortality after the injury. Throughout the paper, we denoted patients who died within 14 days after the injury as cases and the rest as non-cases. The following variables were also examined and considered as potential predictor variables: patient demographics (age, race, gender), patient vitals at time of emergency department visit (systolic blood pressure, pulse rate, respiratory rate, oxygen saturation, body temperature), injury severity measurements (GCS, ISS, AIS severity rating), the presence of foreign substances in the body (narcotics/prescriptions, alcohol), the use of supplemental oxygen, and the type of injury (blunt, penetrating).

Descriptive statistics

Patient demographics and clinical characteristics were summarized by groups (case vs. non-cases) in Table 4. Counts and proportions were used as summary statistics for discrete variables. Median and median absolute deviation (MAD) were used for continuous variables in considering outliers [27,28]. The proportion of missing values for each variable was also calculated. Using simple logistic regressions, we measured and tested the associations between patients' mortality and their clinical/demographic characteristics. P-values and crude odds ratios with 95% confidence intervals were also estimated.

Model construction

Prior to model construction, we imputed missing values of potential predictor variables using a Random Forests based Chained Equation algorithm [29] (an approach previously shown to work well in trauma data [30]). Specifically, we used the MICE (Multiple Imputation By Chained Equation) R package [31] to build one single complete data set. Missing data mechanisms were assumed MAR (Missing At Random) [29] and single imputation was used in place of multiple to avoid issues with pooling models.

Our proposed model was then built with multiple logistic regression using backwards elimination variable selection and the Bayesian Information Criterion (BIC) to determine the final set of predictors [32, 33]. To test the performance of our final model, we split the data into training and testing

Table 4: **Summary statistics (NTDB 2010 - 2015)**

			Cases		Noncases		Univariate Analysis		
Categorical Variables		Missing %	n	%	n	%	Crude OR	95% CI	p-value
Gender	Female	0.04%	1337	38.79%	52989	36.79%	1.00 (ref.)	-	-
	Male		2209	61.20%	91017	63.20%	0.91	(0.85, 0.98)	0.016
Alcohol Use	No	0%	3415	99.10%	143041	99.32%	1.00 (ref.)	-	-
	Yes		31	0.89%	965	0.67%	1.34	(0.91, 1.89)	0.105
Drug Use	No	0%	3321	96.37%	141379	98.17%	1.00 (ref.)	-	-
	Yes		125	3.62%	2627	1.82%	2.02	(1.67, 2.42)	< 0.001
Supplemental Oxygen	No	22.05%	407	11.81%	119911	83.26%	1.00 (ref.)	-	-
	Yes		3039	88.18%	24095	16.73%	37.15	(33.52, 41.30)	< 0.001
Race	African American	6.24%	774	22.46%	23794	16.52%	1.00 (ref.)	-	-
	American Indian		56	1.62%	1502	1.04%	1.14	(0.86, 1.49)	0.333
	Asian		55	1.59%	2895	2.01%	0.58	(0.43, 0.76)	< 0.001
	Caucasian		2047	59.40%	94519	65.63%	0.66	(0.61, 0.72)	< 0.001
	Pacific Islander		19	0.55%	484	0.33%	1.20	(0.73, 1.86)	0.427
	Other Race		495	14.36%	20812	14.45%	0.73	(0.65, 0.81)	< 0.001
Type of Injury	Blunt	1.24%	1912	55.48%	129753	90.10%	1.00 (ref.)	-	-
	Penetrating		292	8.47%	743	0.51%	26.67	(23.09, 30.72)	< 0.001
	Other Trauma		1242	36.04%	13510	9.38%	6.28	(5.79, 6.71)	< 0.001
Continuous Variables		Missing %	Median	MAD	Median	MAD	Crude OR	95% CI	p-value
AIS Severity		0%	5	0.0	2	1.0	8.44	(7.97, 8.94)	< 0.001
GCS		5.99%	3	0.0	15	0.0	0.56	(0.55, 0.57)	< 0.001
Age		0%	3	3.0	5	5.0	0.96	(0.96, 0.97)	< 0.001
ISS		2.5%	29	6.0	9	5.0	1.15	(1.15, 1.15)	< 0.001
Systolic Blood Pressure (mmHg)		9.7%	99	21.0	114	11.0	0.96	(0.96, 0.96)	< 0.001
Pulse Rate (bpm)		2.7%	119	27.0	113	21.0	0.99	(0.99, 0.99)	< 0.001
Respiratory Rate		3.97%	20	7.0	24	4.0	0.91	(0.90, 0.91)	< 0.001
Blood Oxygen Saturation (%)		18.48%	100	0.0	100	0.0	0.97	(0.96, 0.97)	< 0.001
Body Temperature ($^{\circ}$C)		10.01%	35.40	1.2	36.70	0.3	0.68	(0.67, 0.70)	< 0.001

sets based on year of entry into the NTDB. 123,816 patients entered from years 2010-2014 were placed into a training set to build TMPP, and the remaining 23,636 individuals entered in 2015 were used as a testing set to validate performance.

Model evaluation

Using a confusion matrix as in Table 5, we define some terminologies which will be used throughout this paper. The ideal prediction model would maximize both True Positive (TP) and True Negative (TN) totals subsequently minimizing its False Positives (FP) and False Negatives (FN).

Table 5: **Confusion matrices**

		Truth	
		Case	Non-case
Prediction	Case	True Positive (TP)	False Positive (FP)
	Non-case	False Negative (FN)	True Negative (TN)

However, it is unfeasible to optimize all quantities in a model, so instead it is common to choose a model minimizing the error rate, defined as $\frac{FP+FN}{TP+FP+FN+TN}$. However with a rare event outcome (e.g., low mortality rate), the error rate can be misleading. For example, consider an extreme situation where we predict all patients as non-cases. In Table 3, we would have TP = 0, FP = 0, FN = 8913, TN = 64506. Based on these values, we would have a sensitivity of 0%, a specificity of 100%, and an error rate of 12.14%. Even in this situation where no prediction model was applied, we are capable of achieving a low error rate (notably an error rate better than TRISS on this data). This is due to the large number of non-cases in the data set mixed with few instances of cases i.e. an imbalance across the outcome in our data. To avoid the imbalance entirely we could choose to maximize sensitivity, the accuracy of cases, to reduce the quantity of false negatives. This might be reasonable in some situations, however not in medical fields. Increases in sensitivity lead to decreases in specificity, and constantly predicting non-cases as cases and would cause an unnecessary and inappropriate spending of treatment costs. One of the most popular approaches for imbalanced data is down-sampling which is to select only some of the non-cases so that the total number of non-cases is nearly equal to the number of cases [34]. However, in this paper we choose to add weights to our logistic regression model as well as optimize the decision threshold used for classifying probabilities as cases or non-cases. This threshold was chosen by performing 5-fold cross validation on the training data and choosing the threshold which minimized a total cost we defined as in Eq. 1 where W was the weight placed on the cases in the regression model (i.e., the cost associated with false negatives). Multiple values for W were analyzed and the W which best classified the training data was chosen as the final weight.

$$Total\,Cost = FP + W * FN \tag{1}$$

All final performance evaluation was then done on the testing set. In addition to commonly used measurements derived from the confusion matrix (discussed below) and Receiver Operating Characteristic (ROC) curves, we also assessed the diagnostic odds ratio [35], F_β measure [36], and Youden's J Index [37] as alternative indicators of performance.

Finally, we also evaluated the performance of a naïve logistic regression model constructed without weights on our training data as well the aforementioned TRISS and compared both with TMPP.

Confusion matrix metrics

Performance measurements obtained directly from the confusion matrices included:

- Accuracy (ACC)= $\frac{TP+TN}{TP+FP+FN+TN}$
- Sensitivity or True Positive Rate (TPR) = $\frac{TP}{TP+FN}$
- Specificity or True Negative Rate (TNR) = $\frac{TN}{TN+FP}$
- Positive predictive value (PPV) = $\frac{TP}{TP+FP}$
- Negative predictive value (NPV) = $\frac{TN}{TN+FN}$
- False positive rate (FPR) = $\frac{FP}{FP+TN}$
- False negative rate (FNR) = $\frac{FN}{TP+FN}$

where TP, TN, FP, and FN are defined in Table 5.

ROC curves

The ROC curve along with the area under its curve (AUC) were both evaluated for model performance using the ROCR package in R. DeLong's test for the comparison of AUCs was used to test for significant differences in the AUC between each model [38].

Measures of effectiveness of the diagnostic test

The positive and negative likelihood ratios were used to measure how much the probability of suffering mortality changed given the output of our model. As a general rule, positive likelihood ratios greater than 10 suggest large increases in the probability of mortality after the model has evaluated a patient's status while negative likelihood ratios less than 0.1 indicate large decreases in said probability [39]. Both ratios can be computed directly from the confusion matrix (Eq. 2).

$$LR^{+} = \frac{TPR}{FPR} \qquad\qquad LR^{-} = \frac{FNR}{TNR} \tag{2}$$

In addition, the diagnostic odds ratio (DOR) which can be derived from both likelihood ratios (Eq.3) was also used as an indicator of discriminatory ability for our predictive model. Not only is this performance measurement independent of prevalence, but it can also be easily interpreted in a clinical setting as the ratio of the odds of a positive output of a model in cases relative to the odds of positive output in non-cases. [35].

$$DOR = \frac{LR^+}{LR^-} \tag{3}$$

The DOR takes values from 0 to infinity with larger values indicating better performance. A test for significant difference in odds ratio was used to compare the DOR between the models.

The F_β score

The F_β score was used to evaluate the tradeoff the models were making between sensitivity and positive predictive value. This metric can also be weighted to emphasize the importance of false negatives and in this study weights of $\beta = 1$, $\beta = 2$ (the most common weights) and $\beta = W$ were all used to analyze how the models perform as the importance of reducing the number of false negatives increased.

The score takes on values from 0 (worst performance) to 1 (best performance) and can be computed as below (Eq.4).

$$F_\beta = (1 + \beta^2) * \frac{TPR * PPV}{(\beta^2 * PPV) + TPR} \tag{4}$$

Youden's J Index

The final performance measurement used to evaluate all three models was Youden's J Index [37]. This measurement is often used to optimize the decision threshold in conjunction with ROC curves, but it also can be used to estimate the probability of a model giving an informed decision as opposed to a random guess. It can be computed with the sensitivity and specificity derived from the confusion matrix (Eq. 5) and takes values from 0 (worst performance) to 1 (best performance).

$$J = TPR + TNR - 1 \tag{5}$$

Results

Naïve model

Variables selected in the final naïve model included the AIS severity rating, ISS, GCS, type of injury, systolic blood pressure, pulse rate, blood oxygen saturation, and body temperature. A summary of this model is presented in Table 6.

Table 6: **Summary for Naïve model**

Variable	Adj. OR	95% OR CI	p-value

Intercept	5.14	-	-
Blunt Injury	1.00 (ref.)	-	-
Penetrating Injury	7.14	(5.66, 9.03)	$<$0.001
Other Injury Type	2.71	(2.41, 3.05)	$<$0.001
AIS Severity	1.78	(1.66, 1.91)	$<$0.001
Injury Severity Score	1.05	(1.05, 1.06)	$<$0.001
Glasgow Coma Score	0.65	(0.63, 0.66)	$<$0.001
Systolic Blood Pressure	0.98	(0.97, 0.98)	$<$0.001
Pulse Rate	0.99	(0.99, 0.99)	$<$0.001
Body Temp.	0.86	(0.85, 0.88)	$<$0.001

TMPP

Variables selected for the proposed model included the AIS severity rating, age, gender, drug use, ISS, systolic blood pressure, pulse rate, blood oxygen saturation, body temperature, supplemental oxygen use, GCS, race and type of injury. The weight which resulted in the best discrimination on the training data was $W = 100$ and minimizing Eq. 1 during cross-validation resulted in a decision threshold of 0.44. A summary of this model is presented in Table 7.

Table 7: **Summary for TMPP**

Variable	Adjusted OR	95% CI	p-value
Intercept	12.74	-	-
Female	1.00 (ref.)	-	-
Male	1.20	(1.12, 1.28)	$<$0.001
No Drug Use	1.00 (ref.)	-	-
Drug Use	1.74	(1.46, 2.09)	$<$0.001
No Supplemental Oxygen	1.00 (ref.)	-	-
Supplemental Oxygen	1.62	(1.49, 1.75)	$<$0.001
African American	1.00 (ref.)	-	-
American Indian	1.53	(1.15, 2.04)	0.003
Asian	0.72	(0.56, 0.94)	0.016
Caucasian	1.44	(1.32, 1.57)	$<$0.001
Pacific Islander	1.22	(0.81, 1.89)	0.355
Other Race	1.21	(1.08, 1.36)	$<$0.001
Blunt Injury	1.00 (ref.)	-	-
Penetrating Injury	10.99	(8.87, 13.72)	$<$0.001
Other Injury	2.91	(2.68, 3.15)	$<$0.001
AIS Severity	1.49	(1.44, 1.54)	$<$0.001
Injury Severity Score	1.10	(1.10, 1.11)	$<$0.001
Glasgow Coma Score	0.69	(0.68, 0.69)	$<$0.001
Age	0.95	(0.95, 0.96)	$<$0.001
Systolic Blood Pressure	0.98	(0.98, 0.98)	$<$0.001
Pulse Rate	0.99	(0.99, 0.99)	$<$0.001
Oxygen Saturation	0.97	(0.96, 0.97)	$<$0.001
Body Temperature	0.77	(0.75, 0.79)	$<$0.001

Comparison between models

The naïve model showed better accuracy (0.983 vs 0.920), specificity (0.995 vs 0.918), positive predictive value (0.722 vs 0.222), false positive rate (0.004 vs 0.081), positive likelihood ratio (111.344 vs 12.203), and F_1 score (0.567 vs 0.362) when compared with TMPP. However, TMPP had improved sensitivity (0.990 vs 0.467), false negative rate (0.009 vs 0.532), negative predictive value (0.999 vs 0.987), negative likelihood ratio (0.010 vs 0.534), diagnostic odds ratio (1213.218 vs 208.279), F_2 score (0.585 vs 0.502), F_{100} score (0.989 vs 0.467) and Youden's J Index (0.908 vs 0.462). The performance of TRISS was comparable with the naïve model in nearly every metric with the exception of the positive predictive value, positive likelihood ratio, and diagnostic odds ratio each of which were better in the naïve model. A summary of all measurements is given in Table 8. The best value for each measure is given in bold.

Table 8: **Performance measurements**

Performance Measurement	Naïve	TMPP	TRISS
Accuracy	**0.983**	0.920	0.979
Sensitivity	0.467	**0.990**	0.491
Specificity	**0.995**	0.918	0.991
False Positive Rate	**0.004**	0.081	0.008
False Negative Rate	0.532	**0.009**	0.508
Positive Predictive Value	**0.722**	0.222	0.563
Negative Predictive Value	0.987	**0.999**	0.988
Area Under ROC Curve	**0.987**	**0.987**	0.983
Positive Likelihood Ratio	**111.344**	12.203	55.123
Negative Likelihood Ratio	0.534	**0.010**	0.512
Diagnostic Odds Ratio	208.279	**1213.218**	107.475
F_1 Score	**0.567**	0.362	0.524
F_2 Score	0.502	**0.585**	0.503
F_{100} Score	0.467	**0.989**	0.491
Youden's Index	0.462	**0.908**	0.482

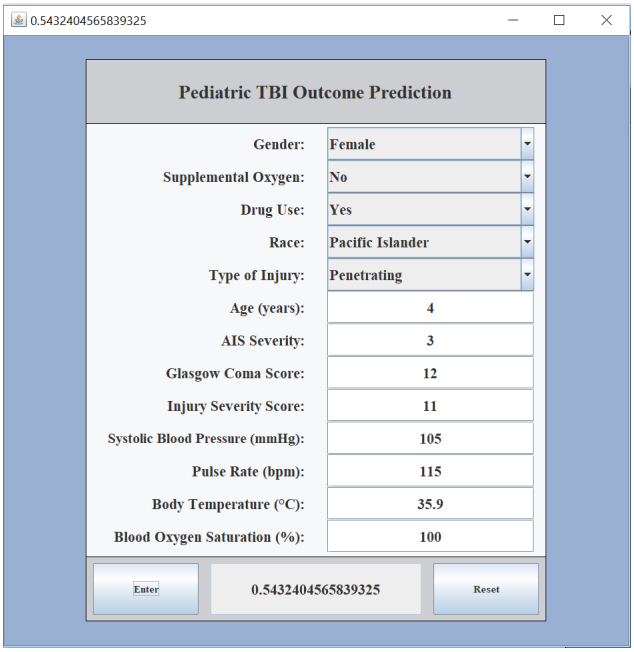
The difference in AUC between TMPP and the naïve model was not significant according to DeLong's test ($p > 0.5$). However, the difference in AUC was significant between TMPP and TRISS ($p < 0.001$) (this was also true of the difference in the AUC between the naïve model and TRISS ($p < 0.001$)). A comparison of the DORs showed a significant difference of DOR between TMPP and TRISS ($p < 0.001$), and TMPP and the naïve model ($p < 0.001$). This was not true for the difference of DOR between the naïve model and TRISS ($p > 0.4$).

Java application

A java application was created to help with the accessibility of TMPP. It works offline and has been successfully tested on Mac, Windows, and Linux machines supporting an appropriate Java version. An overview of the Graphical User Interface (GUI) is presented in Figure 1.

To ensure the validity of an input, the application will throw error messages to the user if any of the conditions are violated:

Figure 1: **Calulator GUI**



- Invalid AIS severity. Value should be integer between 1 and 6.
- Invalid ISS. Value should be integer between 1 and 75.
- Invalid GCS. Value should be integer between 3 and 15.
- Invalid blood oxygen saturation %. Value should be between 0 and 100.
- Invalid age. Value should be between 0 and 14
- Negative vital measurements.
- Empty field. Every field should be filled.
- Incorrect input type. All values should be integers or doubles (depending on the variable).

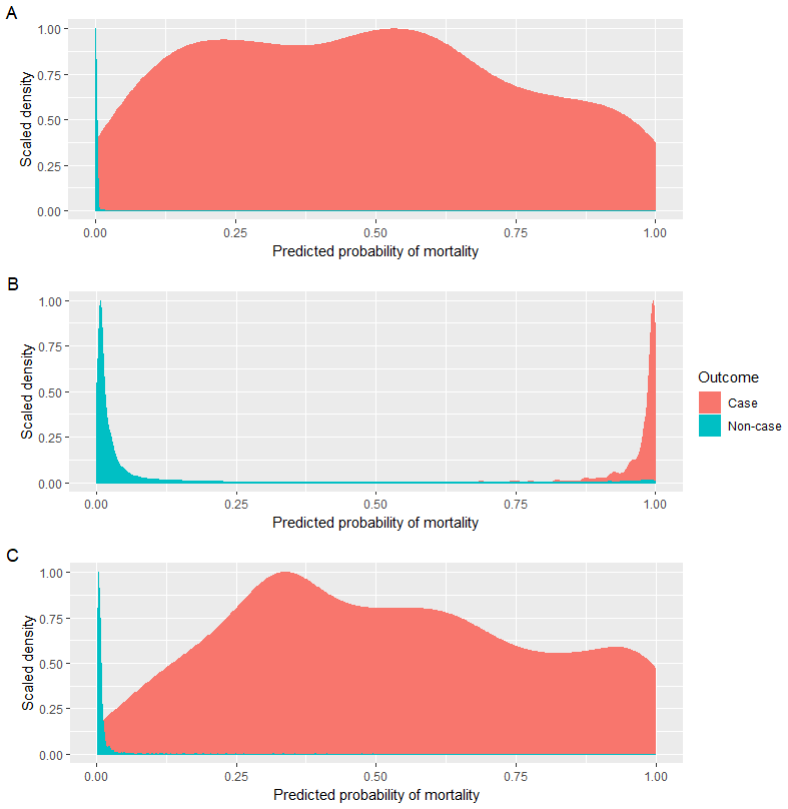
Discussion

Results from the performance measurements suggest TMPP to be an effective diagnostic tool for evaluating incoming pediatric patients' risks of TBI mortality. Nearly every variable was included in this model after backwards elimination, but the Java application allows the model to be used by anyone running the appropriate software with ease. Furthermore all variables considered are routinely measured and do not require any advanced medical instrumentation. The simplicity of these variables opens the possibility for the application of this model to see use in low-middle income countries (LMIC) where more advanced diagnostic methods such as brain imaging methods may be less frequently available.

While the naïve model outperformed TMPP in several measurements, TMPP still demonstrated excellent performance in many of these areas. TMPP's accuracy, specificity, positive likelihood

ratio, and false positive rate were all still indicative of a good diagnostic tool. In contrast, in the instances where TMPP was suggested to perform better than the naïve model, the difference in performance was much more dramatic. The false negative rate and negative likelihood ratio decreased by more than 0.50 in TMPP. And a similar gain was seen in Youden's J index where the estimated probability of an informed decision increased from 0.462 to 0.908 as well as the F_β scores as more cost was associated with false negative errors. Furthermore, the median prediction on cases made by the naïve model was only 0.47 as opposed to TMPP's 0.99 and when looking at the density plot of the predictions made by all three models (Fig. 2) it is clear that both the naïve model and TRISS have difficulty classifying cases as opposed to TMPP's strong discrimination.

Figure 2: **Density plot of model predictions**
A–Naïve , B–TMPP, C–TRISS



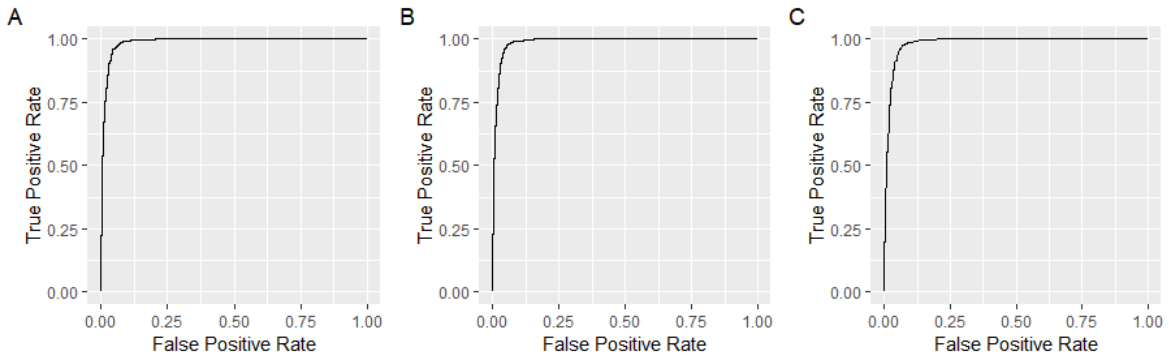
The greatest drop in performance observed in TMPP was the loss of positive predictive value. This metric has been regarded as one of the most important performance measurements for clinical diagnostic tools, however emphasis in this study was placed on ensuring that those predicted to survive were actual survivors. To accomplish this, the model needed to be less conservative with its mortality predictions which ultimately increased its false positive rate and lowered its positive predictive value. Surveys have shown that, depending on the type of injury/illness, patients and medical personnel have been willing to accept more than 2000 false positives in exchange for a single reduction in the number of false negatives [40, 41] given by a diagnostic test, and so when looking at the confusion matrices of each model (Table 9) we do not believe this drop in positive predictive value to be particularly concerning. As when compared with the naïve model or TRISS, TMPP exchanges roughly 2000 false positives for a drop of more than 200 false negatives.

Table 9: **Confusion matrices**

	Naïve		TMPP		TRISS	
	Cond. Pos.	Cond. Neg.	Cond. Pos.	Cond. Neg.	Cond. Pos.	Cond. Neg.
Pred. Pos.	253	97	536	1875	266	206
Pred. Neg.	288	22998	5	21220	275	22889

An interesting result of this study was the lack of use of the ROC curve in determining the model with the best performance. The difference between the AUC of the naïve model and TMPP was insignificant, and despite TRISS' AUC having a significant difference from both naïve and TMPP models, visually all three ROC curves were identical (Fig 3). Several other machine learning models were also constructed, including rule-based learners (OneR, RIPPER algorithms [42, 43]) and tree methods (decision trees, random forests) and compared with TMPP, but their AUC still indicated similar performance across models despite the information given by their confusion matrices (data not shown). Thus as machine learning methods continue to gain more popularity [44], we feel this highlights the necessity for a thorough use of multiple performance measurements during model evaluation for those models with intentions of real-world use.

Figure 3: **Receiver operating characteristic curves**
A–Naïve , B–TMPP, C–TRISS



We acknowledge that there were limitations to this study. Between 2010 and 2015, several commonly known predictors of TBI outcome were not recorded into the NTDB and therefore could not be included in model construction. This also meant that we could not compare our proposed model's performance with existing adult TBI models such as CRASH and IMPACT. In addition, during this time period, drug and alcohol indicator variables were defined vaguely and only specified the presence of foreign substances as opposed to specific drugs or blood alcohol contents thus limiting the information gained from these variables. Injury specification was yet another limitation imposed on this study by the NTDB. Head injuries were specified with the AIS predot code, but this code only asserts injury to the appropriate region and does not guarantee TBI. Other studies have used specific definitions given by the CDC derived from International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes to specify TBI when working with similar data sets, but even these definitions have shown to be inaccurate for TBI specification [45]. TBI is a broad classification of injuries and until a formal definition can be accurately used to specify these injuries from general trauma databases, this issue will continue to be present. Though it may have introduced multiple limitations, we still feel that the NTDB remains and invaluable resource for trauma data. New variables are continuously added (including those well-suited for TBI prediction)

and existing variables are redefined to increase the information gained from analyzing them.

Lastly, while it should be expected that a model built on a specific injury type will outperform a general trauma outcome prediction tool such as TRISS when evaluated on data based around said injury type, the degree that our proposed model outperforms TRISS in several key measurements that may have significant clinical interpretation is alarming and serves, as a minimum, as additional evidence for the need to retire or update TRISS. Especially in the context of predictive modeling for medical applications, data is bound to change overtime and predictive models built on 30-year-old data can't be expected to have the same performance they would've had 30 years prior without any modifications to account for any sort of medical advances.

Conclusion

TMPP has shown to be a powerful diagnostic tool in predicting pediatric mortality in TBI. In situ validation is needed to verify the performance observed, but if successful our model fills a void created from outdated and ill-applicable existing methods and has the potential save many young lives.

References

[1] Taylor CA, Bell JM, Breiding MJ, Xu L. Traumatic Brain Injury–Related Emergency Department Visits, Hospitalizations, and Deaths — United States, 2007 and 2013. MMWR Surveillance Summaries. 2017;doi:10.15585/mmwr.ss6609a1.

[2] Faul M, Xu L, Wald MM, Coronado VG. Traumatic brain injury in the United States: emergency department visits, hospitalizations, and deaths. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. 2010;doi:10.1016/B978-0-444-52910-7.00011-8.

[3] Teasdale G, Jennett B. Assessment of Coma and Impaired Consciousness. A Practical Scale. The Lancet. 1974;doi:10.1016/S0140-6736(74)91639-0.

[4] Guluma K, Zink B. Traumatic brain injury. Semin Respir Crit Care Med. 2002;23:37–45. doi:10.1055/s-2002-20587.

[5] Silverberg N, Gardner A, Iverson GL, Brubacher JR. Systematic Review of Prognostic Models for Mild Traumatic Brain Injury. Archives of Physical Medicine and Rehabilitation. 2014;95(10):e7. doi:10.1016/j.apmr.2014.07.377.

[6] Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: The TRISS method. Journal of Trauma - Injury, Infection and Critical Care. 1987;27(4):370–378. doi:10.1097/00005373-198704000-00005.

[7] Champion HR, Copes WS, Sacco WJ, Lawnick MM, Keast SL, Bain LW, et al. The major trauma outcome study: Establishing national norms for trauma care. Journal of Trauma - Injury, Infection and Critical Care. 1990;doi:10.1097/00005373-199011000-00008.

[8] Champion HR, Sacco WJ, Carnazzo AJ, Copes W, Fouty WJ. Trauma score; 1981.

[9] Valderrama-Molina CO, Giraldo N, Constain A, Puerta A, Restrepo C, León A, et al. Validation of trauma scales: ISS, NISS, RTS and TRISS for predicting mortality in a Colombian population. European Journal of Orthopaedic Surgery and Traumatology. 2017;27(2):213–220. doi:10.1007/s00590-016-1892-6.

[10] Champion HR, Sacco WJ, Copes WS, Gann DS, Gennarelli TA, Flanagan ME. A revision of the trauma score. Journal of Trauma - Injury, Infection and Critical Care. 1989;29(5):623–629. doi:10.1097/00005373-198905000-00017.

[11] Champion H, Copes W, Sacco W, Frey C, Holcroft J, Hoyt D, et al. Improved prediction from a severity characterization of trauma (ascot) over trauma and injury severity score (triss): Results of an independent evaluation. Journal of Trauma Nursing. 1996;doi:10.1097/00043860-199604000-00008.

[12] Hannan EL, Mendeloff J, Farrell LS, Cayten CG, Murphy JG. Validation of TRISS and ASCOT using a non-MTOS trauma registry. J Trauma. 1995;doi:10.1097/00005373-199501000-00022.

[13] Schluter PJ, Nathens A, Neal ML, Goble S, Cameron CM, Davey TM, et al. Trauma and Injury Severity Score (TRISS) coefficients 2009 revision. Journal of Trauma - Injury, Infection and Critical Care. 2010;68(4):761–770. doi:10.1097/TA.0b013e3181d3223b.

[14] Schluter PJ. The Trauma and Injury Severity Score (TRISS) revised. Injury. 2011;doi:10.1016/j.injury.2010.08.040.

[15] Perel P, Arango M, Clayton T, Edwards P, Komolafe E, Poccock S, et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. British Medical Journal. 2008;doi:10.1136/bmj.39461.643438.25.

[16] Roozenbeek B, Lingsma HF, Lecky FE, Lu J, Weir J, Butcher I, et al. Prediction of Outcome after Moderate and Severe Traumatic Brain Injury: External Validation of the IMPACT and CRASH Prognostic Models. Critical Care Medicine. 2012;40(5):1609–1617. doi:10.1097/CCM.0b013e31824519ce.Prediction.

[17] Majdan M, Lingsma HF, Nieboer D, Mauritz W, Rusnak M, Steyerberg EW. Performance of IMPACT, CRASH and Nijmegen models in predicting six months outcome of patients with severe or moderate TBI: an external validation study. Scand J Trauma Resusc Emerg Med. 2014;22(1):68. doi:10.1186/s13049-014-0068-9.

[18] Honeybul S, Ho KM. Predicting long-term neurological outcomes after severe traumatic brain injury requiring decompressive craniectomy: A comparison of the CRASH and IMPACT prognostic models. Injury. 2016;47(9):1886–1892. doi:10.1016/j.injury.2016.04.017.

[19] Wong GKC, Teoh J, Yeung J, Chan E, Siu E, Woo P, et al. Outcomes of traumatic brain injury in Hong Kong: Validation with the TRISS, CRASH, and IMPACT models. Journal of Clinical Neuroscience. 2013;20(12):1693–1696. doi:10.1016/j.jocn.2012.12.032.

[20] Dewan MC, Mummareddy N, Wellons JC, Bonfield CM. The epidemiology of global pediatric traumatic brain injury: a qualitative review. World neurosurgery. 2016;doi:10.1016/j.wneu.2016.03.045.

[21] Rates of TBI-related Deaths by Age Group — United States, 2001–2010. Centers for Disease Control and Prevention; 2016. Available from:https://www.cdc.gov/traumaticbraininjury/data/rates_deaths_byage.html

[22] Rates of TBI-related Hospitalizations by Age Group — United States, 2001–2010 Centers for Disease Control and Prevention; 2016. Available from:https://www.cdc.gov/traumaticbraininjury/data/rates_hosp_byage.html

[23] Rates of TBI-related Emergency Department Visits by Age Group — United States, 2001–2010 Centers for Disease Control and Prevention; 2016. Available from:https://www.cdc.gov/traumaticbraininjury/data/rates_ed_byage.html

[24] King G, Zeng L. Logistic Regression in Rare Events Data. Political Analysis. 2001;doi:10.1093/oxfordjournals.pan.a004868.

[25] St-Louis E, Bracco D, Hanley J, Razek T, Baird R. Development and validation of a new pediatric resuscitation and trauma outcome (PRESTO) model using the U.S. National Trauma Data Bank. Journal of Pediatric Surgery. 2017;53(1):136–140. doi:10.1016/j.jpedsurg.2017.10.039.

[26] R Core Team. R: A Language and Environment for Statistical Computing; 2018. Available from: https://www.R-project.org/.

[27] Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology. 2013;49(4):764–766. doi:10.1016/j.jesp.2013.03.013.

[28] Fleming S, Thompson M, Stevens R, Heneghan C, Plüddemann A, MacOnochie I, et al. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: A systematic review of observational studies. The Lancet. 2011;377(9770):1011–1018. doi:10.1016/S0140-6736(10)62226-X.

[29] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple Imputation by Chained Equations: What is it and how does it work? International Journal of Methods in Psychiatric Research. 2011;20(1):40–49. doi:10.1002/mpr.329.Multiple.

[30] Oyetunji TA, Crompton JG, Ehanire ID, Stevens KA, Efron DT, Haut ER, et al. Multiple imputation in trauma disparity research. Journal of Surgical Research. 2011;165(1). doi:10.1016/j.jss.2010.09.025.

[31] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1–67.

[32] Akaike H. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika. 1973;doi:10.1093/biomet/60.2.255.

[33] Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978;doi:10.1214/aos/1176344136.

[34] He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2009;21(9):1263–1284. doi:10.1109/TKDE.2008.239.

[35] Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: A single indicator of test performance. Journal of Clinical Epidemiology. 2003;56(11):1129–1135. doi:10.1016/S0895-4356(03)00177-X.

[36] Rijsbergen CJ. Information Retrieval. 1979.

[37] Youden W. Index for rating diagnostic tests. Cancer. 1950

[38] Delong ER, Carolina N. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves : A Nonparametric Approach Author (s): Elizabeth R . DeLong , David M . DeLong and Daniel L . Clarke-Pearson Published by : International Biometric Society Stable . Biometrics. 2016;44(3):837–845.

[39] Deeks JJ. Diagnostic tests 4: likelihood ratios. Bmj. 2004;329(7458):168–169. doi:10.1136/bmj.329.7458.168.

[40] Boone D, Mallett S, Zhu S, Yao G L, Bell N, Ghanouni A, Wagner C, Taylor S A, Altman D G, Lilford R, Halligan S. Patients' and Healthcare Professionals' Values Regarding True- and False-Positive Diagnosis when Colorectal Cancer Screening by CT Colonography: Discrete Choice Experiment. Plos One. 2013; 8(12). doi:10.1371/journal.pone.0080767

[41] Schwartz LM, Woloshin S, Sox HC, Fischoff B, Welch HG. US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. BMJ. 2000;320(7250):1635–1640

[42] Holte RC. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning. 1993;11(1):63–90 doi:10.1023/A:10226331118932

[43] Cohen WW. Fast Effective Rule Induction. Machine Learning: Proceedings of the Twelfth International Conference. 1995.

[44] Koohy H. The rise and fall of machine learning methods in biomedical research. F1000Research. 2017;6(0):2012. doi:10.12688/f1000research.13016.1.

[45] Bazarian JJ, Veazie P, Mookerjee S, Lerner EB. Accuracy of mild traumatic brain injury case ascertainment using ICD-9 codes. Academic Emergency Medicine. 2006;13(1):31–38. doi:10.1197/j.aem.2005.07.038.