Recurrent Neural Network based Time-Series Modeling for Long-term Prognosis Following Acute Traumatic Brain Injury

Amin Nayebi[1], Sindhu Tipirneni[2], Brandon Foreman[3], Jonathan Ratcliff[4], Chandan K Reddy[2], Vignesh Subbian[1]

[1]The University of Arizona, AZ, USA; [2]Virginia Tech, VA, USA; [3]University of Cincinnati, OH, USA; [4]Emory University, GA, USA

Abstract

We developed a prognostic model for longer-term outcome prediction in traumatic brain injury (TBI) using an attention-based recurrent neural network (RNN). The model was trained on admission and time series data obtained from a multi-site, longitudinal, observational study of TBI patients. We included 110 clinical variables as model input and Glasgow Outcome Score Extended (GOSE) at six months after injury as the outcome variable. Designed to handle missing values in time series data, the RNN model was compared to an existing TBI prognostic model using 10-fold cross validation. The area under receiver operating characteristic curve (AUC) for the RNN model is 0.86 (95% CI 0.83-0.89) for binary outcomes, whereas the AUC of the comparison model is 0.69 (95% CI 0.67-0.71). We demonstrated that including time series data into prognostic models for TBI can boost the discriminative ability of prediction models with either binary or ordinal outcomes.

Introduction

Traumatic Brain Injury (TBI) is one of the leading causes of death and disability in the United States. There are nearly 2.8 million new TBI cases every year in the US (1). It is estimated that the general trend of worldwide TBI cases will continue to increase and be a growing significant health problem, particularly, for low- and middle-income countries (2). Accurate and early prediction of outcomes in TBI patients can help in clinical management as well as in optimizing resource allocation within the health system (3).

Even though not all TBI cases result in death, different forms of disability are common in complex TBI cases, and simple mortality predictions do not account for the long-term health consequences associated with TBI. Therefore, any practical TBI prediction model should consider outcomes other than mortality (2). The Extended Glasgow Outcome Scale (GOSE) is a functional TBI outcome measure of the severity of TBI and is often used in prognostic models (4). GOSE rates patients in eight categories, from death to upper good recovery, and has been commonly dichotomized into mortality (versus survival) or unfavorable (versus favorable). In this research, we develop a model to predict the GOSE outcome of TBI patients.

Prior studies have developed and evaluated models to predict mortality or severity of TBI patients. A systematic review shows that many prognostic studies suffer from methodological issues. For example, the sample size may not be sufficient: 75% of studies have less than 500 subjects (2,3,5). Two of the most widely used prediction studies for TBI patients are the International Mission on Prognosis and Analysis of Clinical trials in Traumatic Brain Injury (IMPACT) (6) and Corticosteroid Randomization After Significant Head injury (CRASH) (7). The covariates in these models are primarily based on the clinical, physiological, and lab data that are collected at the time of admission, which does not take into account the evolution of the primary injury, and the development of secondary brain injuries.

Regression models, such as logistic regression, to predict disease occurrence (diagnosis) or disease outcome (prognosis), are standard approaches to build a prediction model (8). However, machine learning (ML) algorithms are gaining acceptance for use in the clinical domain especially as the increasingly large and rich data sets such as Electronic Health Records (EHR) data are growingly available (9). On the other hand, recent data suggests that ML algorithms do not necessarily outperform regression models for prognosis of TBI cases, especially when the number of predictors is not high.

In our study, in addition to clinical and laboratory data collected at the time of admission to the Emergency Department (ED), we use time series data obtained from the first few days in the ICU. Recurrent neural networks (RNN) are well known to achieve strong results in many applications with time series and sequential data (10). Two common RNN structures, the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), can capture the long-term temporal dependencies in variable-length samples. GRUs are getting more attention since they can maintain the effect of LSTM units while they are simpler. A study showed that GRUs can outperform LSTM units both in terms of CPU time and generalization (11).

One of the limitations in working with clinical time series data is missingness. Approaches to handle missing values in time series data include, but are not limited to, deletion, mean imputation, and autoregression (12). A case study shows that when the missing rate is high, excluding incomplete data negatively impacts the performance of prediction models compared to alternative scenarios of imputation (13). In this work, imputing time series missingness and training the model are done simultaneously. GRU-D is a recently developed recurrent unit for managing the missingness that optimizes the model performance by imputing missing data while simultaneously learning the model parameters (14).

In this paper, we describe an attention-based RNN with new recurrent units known as GRU-D. Our study differentiates from others according to the following characteristics.

- The model combines the temporal features during the ICU stay (e.g., a sequence of vital signs) and non-temporal features such as age and sex.
- The model handles missing values in the time series data and learns from the missingness patterns. Training and imputation occur simultaneously. It uses GRU-D units that benefit from a decaying mechanism for imputation of missing values among time series data.
- The most important features for predicting the longer-term outcomes in acute TBI patients are identified.

Materials and methods

Source of data

This study was based on data from the prospective, multicenter Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) study (15). The TRACK-TBI study collected detailed clinical data on TBI patients from 18 different academic Level I trauma centers across the US. TRACK-TBI enrolled 2996 participants across the spectrum of TBI severity.

For this analysis, we included a total of 110 clinical variables that were collected at the time of arrival to the ED, discharge from ED, and during the ICU or hospital stay. During ICU stays, measurements were recorded as frequently as every hour, providing extensive time-series data for those variables. Of the total 110 variables, 59 were static variables (demographic or one-time-recorded measurements in ED), and 51 were time-series variables that were recorded during the patients hospitalization. The GOSE was measured at six months after injury and was used as our primary outcome variable for this analysis. Table 1 shows the summary statistics for clinical variables in the dataset.

Table 1. Summary statistics of clinical variables in TRACK-TBI dataset

Variable	Frequency/mean	Percentage	% Missing data
Demographic			
Age (mean +/- SD)	39.11 +/- 18.25		0%
Sex (female)	881	31.5%	0%
ED examination			
GCS (mean +/- SD)	12.95 +/- 3.93		4.9%
Pupil Reactivity			18.8%
Both	2124	93.4%	
Neither	118	5.2%	
One	32	1.4%	
Motor Score			4.9%
No response	225	8.4%	
Extension	25	0.9%	
Abnormal	22	0.8%	
Withdrawal	77	2.9%	
Localize	140	5.2%	
Obey	2134	80.1%	
Untestable	40	1.5%	
Diastolic Blood Pressure (mean +/- SD)	84.1 +/- 18.3		6.0%
Systolic Blood Pressure (mean +/- SD)	139.9 +/- 24.2		1.4%
Hemoglobin (mean +/- SD)	13.9 +/- 1.7		12.0%
Glucose (mean +/- SD)	134.7 +/- 53.0		13.1%
Complications and treatment			
Pre-hospital Hypotension (yes)	91	3.3%	0.6%
Pre-hospital Hypoxia (yes)	77	2.8%	0.5%
6-month Outcome			
GOSE			37.4%
1- Death	126	7.2%	
2- Vegetative state	6	0.3%	

3-	Lower severe disability	88	5.0%
4-	Upper severe disability	23	1.3%
5-	Lower moderate disability	160	9.1%
6-	Upper moderate disability	304	17.3%
7-	Lower good recovery	464	26.5%
8-	Upper good recovery	582	33.2%

Study participants

We included all adult participants with GOSE scores available and those that are admitted to the ICU. Non-adult participants and those who withdrew consent are excluded from the analysis. Out of 2996 participants, 902 met the mentioned criteria and are included in the analysis. Only the first five days of ICU data are used in this study. Figure 1 shows the exclusion criteria and the number of subjects that are left after applying each criterion.

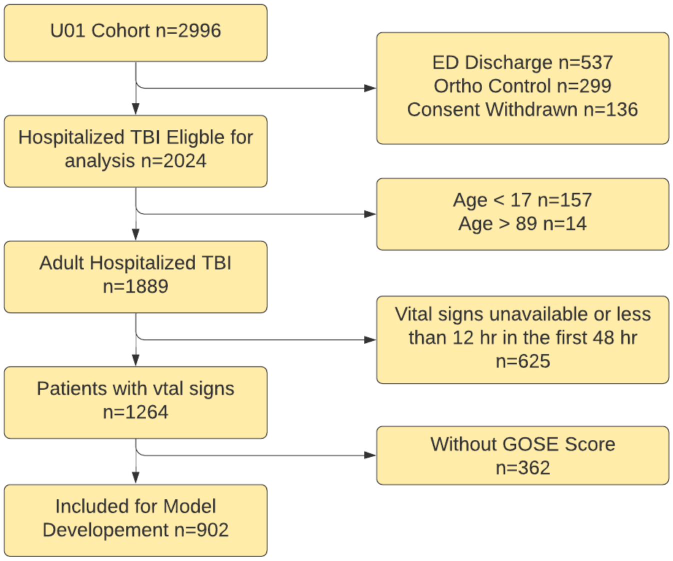

U01 Cohort n=2996
ED Discharge n=537
Ortho Control n=299
Consent Withdrawn n=136
Hospitalized TBI Eligble for analysis n=2024
Age < 17 n=157
Age > 89 n=14
Adult Hospitalized TBI n=1889
Vital signs unavailable or less than 12 hr in the first 48 hr n=625
Patients with vtal signs n=1264
Without GOSE Score n=362
Included for Model Developement n=902

Figure 1. Flowchart showing study participant selection

Missing data

Large amount of missing values in clinical time series data is common (16). To tackle sporadic measurements in the ICU, we discretized the observation time window into fixed-length time intervals. As shown in Figure 2, some variables might have missing values after discretizing the time window, while others have more than one value in a timestamp. We replaced the value of each variable in each timestamp with the average of recorded measurements in the corresponding interval. In order to handle the missingness, we utilized a modified version of Gated Recurrent Units which will be discussed later.

From all static variables, those with more than 20 percent missing data were excluded from the analysis. For imputing the missing static data, we used a Multivariate Imputation with Chained Equations (MICE) approach. In this approach, a series of predictions are used to impute the missing values of each variable. This is done iteratively until the imputed data does not change significantly (17).

Model Development

In this study, a deep RNN model was developed to predict TBI patients' functional outcome at six months, post injury. As the output of the model (GOSE) is an ordinal variable, we follow the same procedure presented by Chen et al. (18) to handle the ordinal output in a neural network. We transformed a classification problem with *K* ordinal categories to *K-1* binary classifications. To do so, we used a special type of output encoding. The i^{th} element of the encoded binary output shows whether the original output is larger than the i^{th} ordinal level. In other words, if a data point belongs to the i^{th} category, the first *i-1* binary variables of the encoded output vector are 1 and the rest are 0. As an example, if GOSE value is 3, the corresponding vector is *(1,1,0,0,0,0,0)*. We use the sigmoid function as the activation function of the output node, and a squared error loss function is used.

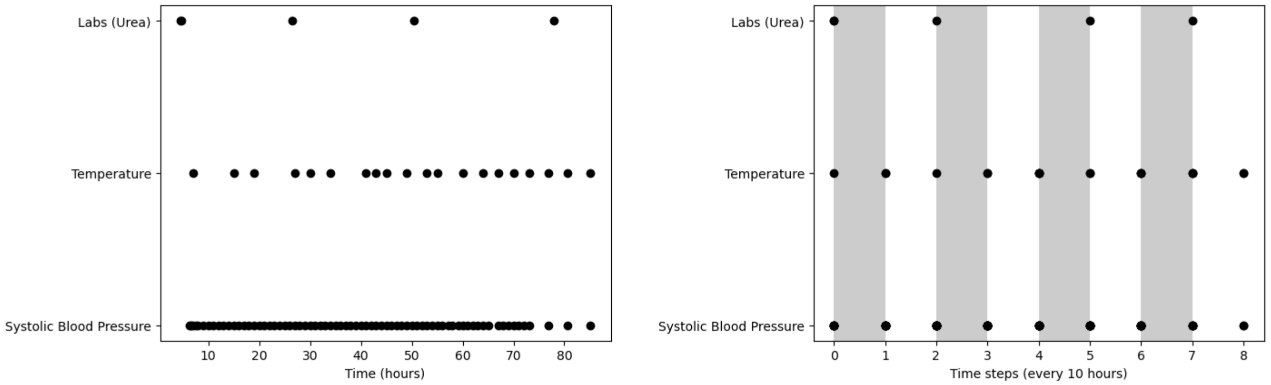


Figure 2. Distribution of recorded measurements across different variables over time for a single patient. The plot on the left shows that different variables have different frequencies of measurement. For example, blood pressure is taken hourly, but urea lab is taken at most once a day. The plot on the right illustrates how values are aggregated over 10-hour intervals.

The model consists of two parts. In the first part, time series data for the first five days of ICU is fed to an RNN which is elaborated on later. The output vectors of RNN at each time step go through an attention layer which takes a weighted sum of its input vectors. The attention layer output is concatenated with the static values and passed through a hidden dense layer before the output dense layer. Figure 3 shows the structure of the prediction model.

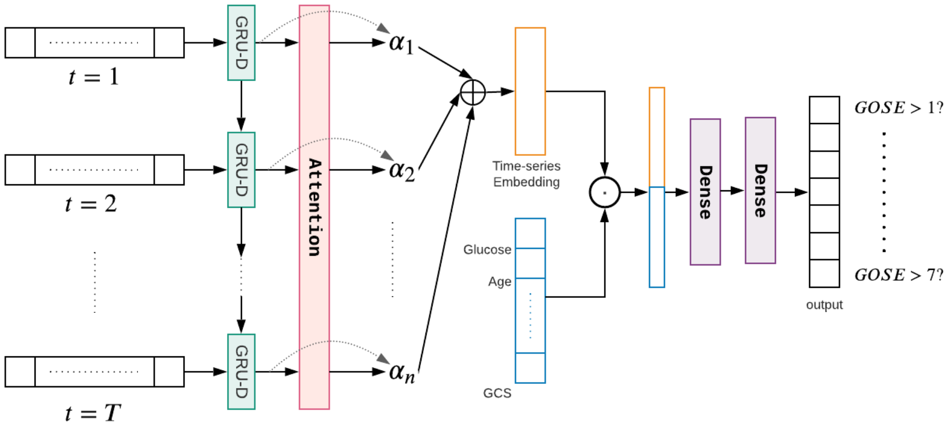


Figure 3. The deep RNN structure for predicting the output. Temporal features are fed to the GRU units and static features are connected to the output through a hidden layer. An attention layer is also used on top of the sequential part of the model.

In this model, we used a modified GRU unit presented by Che et al. (14) which is called GRU-D. They showed that the missingness pattern among variables in time series can provide useful information. They modified the GRU unit in such a way that it captures the missingness information and, at the same time, imputes the missing data. In this research, we used the GRU-D units for RNN cells in our model to predict the GOSE among TBI patients.

In the clinical domain, variables tend to be close to a default number when they are unobserved for a long time (14). This would mean that missing values in time series would fade gradually to a default value (e.g., the empirical average of the variable). GRU-D enjoys a decay mechanism for input variables and hidden states to address the mentioned properties. Two types of decay variables are used in GRU-D, input decays (γ_x) and hidden state decays (γ_h). The general formulation of decay vectors is as follows:

$$\gamma_t = exp\{-max(0, W_\gamma \delta_t + b_\gamma)\} \qquad (1)$$

where δ_t shows how far the last observation of each variable is from time *t*. W_γ and $b_\gamma$are the parameters of the model which must be learned. Corresponding to each variable, a masking vector, *m*, is defined in such a way that is 1 if the variable is observed and otherwise is 0. To impute the input values of the time series, the following formulation is used.

$$\hat{x}_t^d = m_t^d x_t^d + (1 - m_t^d)\left(\gamma_{x_t}^d x_{t'}^d + \left(1 - \gamma_{x_t}^d\right)\tilde{x}^d\right) \qquad (2)$$

In Equation (2), indices d and t indicate d-th variable and t-th time slot, respectively. $\hat{x}_t^d$is the imputed value of the input variable and would be used in the traditional GRU equations. $x_t^d$is the input time series data. $\gamma_{x_t}^d$is the input decay calculated by Equation (1). $x_{t'}^d$is the last observed value of d-th variable at time t'. $\tilde{x}^d$is the empirical average of d-th variable over all time steps and observations.

To fully capture the missingness information, a decaying mechanism for the hidden states is utilized. Before modifying each hidden state using traditional GRU equations, a new decayed hidden state is calculated as follows:

$$\hat{h}_{t-1} = \gamma_{h_t} \odot h_{t-1} \qquad (3)$$

Instead of $x_t$and h_{t-1}, we use $\hat{x}_t$ and $\hat{h}_{t-1}$ in the update functions of the GRU-D.

To avoid our model overfitting on the training data, we used some tools to regularize the model. We utilized dropout after each layer of the model (both recurrent and feed-forward layers) to randomly drop some of the nodes. We also applied a L2-regularizer on the hidden layer weight parameters. L2-regulazier applies penalties on the layer weight parameters. These penalties are summed into the loss function and hence, avoids weight parameters taking large values.

For comparison purposes, we developed a regression model using static data. In this model, only the variables used in the IMPACT prediction model (6) are included, which are age, motor score, pupillary reactivity, hypoxia, hypotension, glucose, and hemoglobin. To build the model, we first converted the values of each variable to the IMPACT score, and then developed an ordinal logistic model on those scores. The IMPACT model originally was developed on patients with moderate and severe TBI (GCS ≤ 12), and a dichotomized GOS score based on favorable and unfavorable outcomes. To have a fair comparison with IMPACT, we also developed, trained, and validated our model under the same conditions. In order to dichotomize the GOSE score, an outcome is unfavorable if GOSE ≤ 4, otherwise it is considered favorable.

The output of the prediction is an ordinal variable, and generally three types of performance metrics are used to assess ordinal classifiers: accuracy, misclassification error, and rank association (19). All these three measurements are used in this study to compare the models. Accuracy (ACC) simply calculates the proportion of correct classifications. To take the misclassification into account, we also used Area Under the Curve (AUC), F1 score, and Mean Squared Error (MSE). MSE measures the degree of error between true and predicted labels. To calculate this error, we assigned a number to each class of GOSE score from one to eight. Since the class sizes are imbalanced in the output variable (Table 1), we used weighted average of MSE (AMSE) and weighted average accuracy (AACC) across all classes (20).

Another criterion measures the association between true (y) and predicted ($\hat{y}$) labels using a rank order correlation statistic called Kendall's correlation coefficient (τ_b) (21). Based on this criterion, values of $\tau_b$are in the interval of [-1,1]. Larger values of $\tau_b$indicate better association between two ranking vectors (predicted and true output values) and hence better prediction. However, this measurement has a drawback since it does not consider the predictions individually and only takes into account the ranking of the prediction and true values as two vectors.

To tune the model hyperparameters, we used a Bayesian Optimization (BO) method presented in (22) to find the best set of hyperparameters for the model. BO is a strategy for optimizing black-box objective functions that are expensive to calculate. To evaluate the performance of the models in each iteration of BO, a 10-fold cross validation is implemented which splits the data to training, testing, and validation sets. The proportion of different GOSE levels was preserved among all training, testing and validation sets. To assess the performance of models, the same 10-fold cross validation approach was utilized, and the metrics were evaluated on the testing data.

In order to interpret the proposed model, we utilized Shapley Additive Explanation (SHAP), a unified framework for interpreting predictions via feature importance (23). SHAP unifies methods like LIME (24) and DeepLIFT (25) under the additive feature attribution umbrella. An additive feature attribution method formulates the outcome of a prediction model in the form of $f(x) = \theta_0 + \sum_{i=0}^{m} \theta_i x'_i$ in which f is the prediction model, θ_i is the attribution assigned to each feature and x'_i is a simplified input showing whether the i^{th} feature is missing. To implement the method, we used the SHAP python package written by the authors of the original paper.

Results

A total of 902 participants and 110 variables met the inclusion criteria for this study. Our proposed prediction model was trained on the training data and its performance was measured on the test set based on different metrics. The results are shown in Table 2. The values represent the mean and the standard error of mean for a 10-fold cross validation.

Table 2. Performance metrics for each model

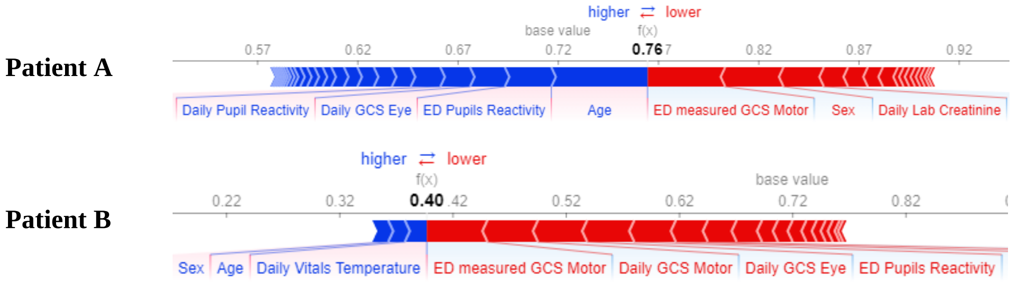
GCS range	Type of outcome	Models	AMSE	AACC	Kendall	AUC	F1
All GCS scores	8-level GOSE	RNN	1.63 ± 0.11	0.24 ± 0.02	0.55 ± 0.04	0.59 ± 0.01	0.24 ± 0.02
		IMPACT	2.02 ± 0.09	0.16 ± 0.04	0.38 ± 0.04	0.51 ± 0.03	0.16 ± 0.03
	Binary outcome	RNN	0.35 ± 0.04	0.86 ± 0.03	0.75 ± 0.05	0.86 ± 0.03	0.91 ± 0.01
		IMPACT	0.55 ± 0.02	0.69 ± 0.02	0.45 ± 0.05	0.69 ± 0.02	0.81 ± 0.01
GCS ≤ 12	8-level GOSE	RNN	1.63 ± 0.13	0.21 ± 0.03	0.60 ± 0.04	0.58 ± 0.02	0.23 ± 0.03
		IMPACT	2.08 ± 0.13	0.15 ± 0.03	0.38 ± 0.06	0.51 ± 0.03	0.14 ± 0.03
	Binary outcome	RNN	0.42 ± 0.03	0.81 ± 0.02	0.64 ± 0.05	0.81 ± 0.02	0.82 ± 0.02
		IMPACT	0.57 ± 0.04	0.66 ± 0.05	0.33 ± 0.10	0.66 ± 0.05	0.66 ± 0.05

We also analyzed the importance of time series data in the prediction task. To do so, we first only added the static data to the model and eliminated the recurrent part of the model (i.e., GRU-D units). We then trained another model which only incorporated the time series data. This model lacked the static inputs of the main model, so it only had the GRU-D units fed by time series data. The results of this analysis are shown in Table 3.

Table 3. Performance metrics for different models with only time series and static data

Data used	AMSE	AACC	Kendall	AUC	F1
All data	1.63 ± 0.11	0.24 ± 0.02	0.55 ± 0.04	0.59 ± 0.01	0.24 ± 0.02
Time series data	1.65 ± 0.13	0.25 ± 0.02	0.55 ± 0.03	0.56 ± 0.01	0.23 ± 0.01
Static data	2.66 ± 0.09	0.13 ± 0.01	0.15 ± 0.02	0.50 ± 0.01	0.07 ± 0.01

Figure 4 illustrates the important features that contributed to the outcome prediction of two sample patients with different outcomes. The important features were derived based on the magnitude of the SHAP values. Figure 5 shows the top important features derived using SHAP values. To assign a single value to each temporal feature, we took an average over all data for the first 120 hours of each patient. Figure 6 illustrates the time series values of three numerical variables among top features for favorable and unfavorable outcomes over the first five days of ICU stay. The missing values in the time series data were imputed using the linear interpolation.



Patient A

Patient B

Figure 4. Important features contributing to the outcome prediction of two patients. Patient A had a favorable outcome with GOSE eight, and patient B died after six months (i.e., GOSE equals one). Features indicated with red and blue colors contribute to the unfavorable and favorable outcomes, respectively. Patient A is a 24 years old male with both reactive pupils at ED, while patient B is a 42 years old female with only one reactive pupil at ED.

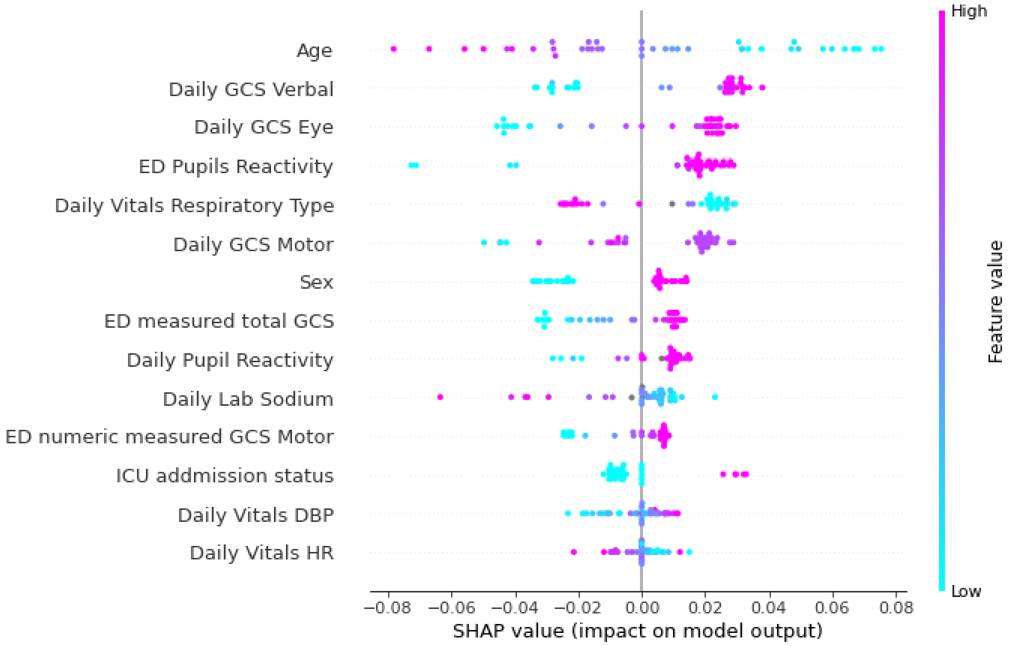


Figure 5. Top important features based on the absolute value of SHAP. Each point represents a patient, and the horizontal axis indicates the SHAP value of each feature for a patient. Negative and positive SHAP values imply the contribution to an unfavorable and favorable outcome, respectively. The color of each point represents the value of a feature for a patient. Feature names that start with “Daily” corresponds to time series variables.

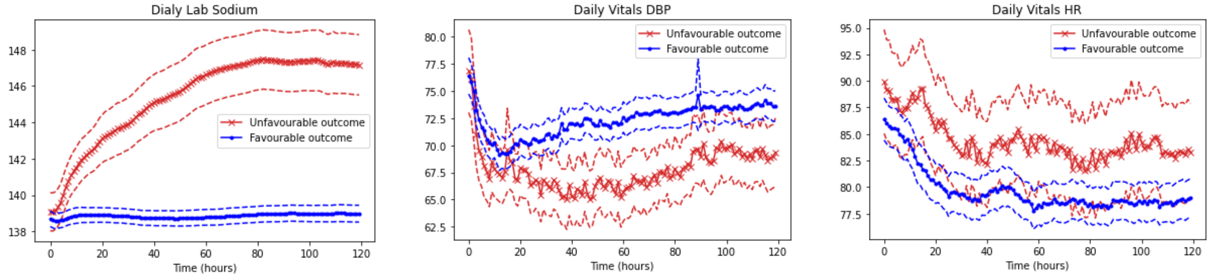


Figure 6. Time series values for different lab measurements over the first five days of ICU stay. Solid lines show the average and dashed lines are three standard errors. Favorable and unfavorable outcomes are separable for all three measurements.

Discussion

The main objective of this study is to predict 6-month outcome for the TBI patients admitted into the ICU. Prior studies using the TRACK-TBI dataset developed a clustering approach for identifying TBI cohorts (26) and implemented a linear regression model to predict the post-concussive symptoms among mild TBI patients (27).

Comparison of model performance and relevance

The comparison between the RNN model and IMPACT shows that the RNN model with time series data performs better based on all metrics. When all population is included and the outcome consists of all eight levels of GOSE, the AUC for both models is low (less than 0.6). One of the possible reasons for not achieving a high AUC is the implementation of a multiclass classification with eight classes. After dichotomizing the GOSE and refitting the models, the AUC for both IMPACT and RNN increases (0.69 and 0.86, respectively). However, RNN still outperforms the IMPACT on all metrics. Even though we did not include the CRASH model in our analysis, one study validated CRASH on TRACK-TBI dataset and demonstrated a poor discriminative ability (AUC of 0.49-0.50) for mild TBI patients (28).

Since the IMPACT model was originally developed on severe and moderate TBI patients, we expected a better performance for IMPACT on this population. However, compared to their performance on all populations, the RNN and IMPACT have a weaker performance. The RNN model with the binary outcome has an AUC of 0.86 on the whole cohort whereas it shows a lower AUC (0.81) on the severe and moderate TBI population. One potential reason is that TBI patients with lower acuity represented over half the studied cohort (555 subjects), thus by excluding these patients, training data shrinks by a significant amount.

Results in Table 3 illustrate that the time series data play an important role in the RNN model's performance. The model trained with only time series data significantly outperforms the model trained with static data. For example, the Kendall coefficient for the model with time series data is 0.55 while this metric for the model with static data is only 0.15. The model trained on only time series data is even superior to the IMPACT model as all the metrics show a better performance. This improvement implies that training our model only with time series data demonstrates a better predictive ability in comparison with the IMPACT model.

Interpretability of deep learning models

To support the interpretability of the RNN model used in the work, we used SHAP to rank the features based on their contribution to the outcome prediction. The important features identified by our model highlight the validity of our model development strategy. The model found that age was one of the most important variables, which is a well-recognized observation (29) and indeed is a core component of the IMPACT prediction model. Other variables from the IMPACT model, namely motor GCS and pupillary reactivity, are also featured. Interestingly, the time series of the GCS are also featured in our model which is novel. This likely reflects the evolution of the patients' neurological exam over time and therefore contributes to prognosis.

The inclusion of vital sign data is noteworthy. The diastolic blood pressure is a known risk factor for long-term cardiovascular health but has not been described in traumatic brain injury populations. Heart rate on the other hand may reflect shock or dysautonomia which impacts patients after TBI and may contribute to prognosis as well. The inclusion of sodium levels in our model suggests that the therapeutic intensity of treatment for elevations in intracranial pressure, which includes administration of hypertonic saline solution, is reflective of prognosis – the sicker patient is treated more aggressively. Therefore, elevated sodium as a negative prognostic marker likely represents the reality that the patient was exhibiting signs of more substantial injury requiring more aggressive therapy. However, this hypothesis would need to be formally tested as patients with the most severe injuries may develop diabetes insipidus (DI), which results in increased sodium, but DI remains relatively uncommon. Despite the limitations, it is worth highlighting that the model uncovered new and/or hypothesis-generating findings by leveraging both static and time-series data within this cohort.

The SHAP contributions of the top features cohere with their measured values, which is clearly illustrated in Figure 5. Most of the time, a consistent trend exists between SHAP contribution and measured values of each variable. Age is one of the best examples of this consistency between feature values and their SHAP contribution. As shown in Figure 5, the SHAP values decrease when the patients' age increases, which means a patient's age contributes more to an unfavorable outcome for older people. Figure 4 shows the SHAP contributions for patients A and B with favorable and unfavorable outcomes, respectively. Based on the SHAP scores, ED pupil reactivity contributes to the favorable and unfavorable outcomes for patients A and B, respectively, which matches the value of this feature since both pupils are reactive for patient A while only one of the pupils of patient B is reactive.

Limitations

One of the challenges in working with clinical time series data is that variables are not measured consistently over time. For example, vitals are recorded regularly, but mostly when patients are unstable. On the other hand, lab results are recorded only when physicians or nurses order them. As a result, clinical variables are recorded irregularly, and the measurement frequency varies between patients and is dependent on the place where the variables are taken. This frequency might be different across variables and even over time (16). However, our deep learning algorithm requires time series data to have regular time intervals. To make it possible, we discretize the observation time window into fixed-length time intervals (one-hour intervals) and aggregate all data within each interval. This method is a tradeoff between losing some information and increasing missing data. By increasing the length of time intervals, some information is lost since all available data in each interval is aggregated. Comparably, by decreasing the length of time intervals, as much data as possible is retained, but the missing data would increase since some of the intervals have no data.

Another limitation of this work is related to the amount of available data as well as external data for validation. The two notable prognostic models that are developed on large cohorts are IMPACT and CRASH. While both models are trained on data with more than 9,000 subjects, our model is trained only on 902 patients. Since the available data for this study is significantly lower than other major studies in the literature, providing more data would be beneficial for this prediction model. Furthermore, due to the lack of external data with enough time series data, we did not validate our model on an external dataset. We also acknowledge biases that may exist in the underlying data. For instance, if elevations in intracranial pressure were treated by a new medication that did not iatrogenically elevate the serum sodium concentrations, it might be expected that this laboratory value would no longer contain important prognostic information. However, if the model we present were not revised to reflect this change in practice, it could result in erroneous predictions. Methods for the reproducibility and validation of models have yet to be standardized for clinical data science but are crucial in prognostic prediction applications.

Conclusion

We propose a deep RNN based model for long-term prognosis of TBI, which is trained on both static and time series data and predicts the GOSE after six months of injury. The model handles the time series missing values and utilizes the information from missingness patterns in temporal features. In summary, our results show that training the model on time series data for TBI patients can be informative and boost the performance of the predictions. Even the model that is solely trained on time series data outperforms the well-known IMPACT prediction model. Top important features are derived from the RNN model, and their values show a separable trend for favorable versus unfavorable outcomes. This study shows the magnitude of information that can be derived from time series data to prognose TBI more accurately.

Acknowledgment

This material is based upon work supported by the National Science Foundation under grants #1838730 and #1838745. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors acknowledge all TRACK-TBI Study Investigators for providing access to data used in this work.

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 Surveill Summ [Internet]. 2017;66(9):1–16. Available from: http://www.cdc.gov/mmwr/volumes/66/ss/ss6609a1.htm

2. Menon DK, Zahed C. Prediction of outcome in severe traumatic brain injury. Curr Opin Crit Care. 2009;15(5):437–41.

3. Hukkelhoven CWPM, Rampen AJJ, Maas AIR, Farace E, Habbema JDF, Marmarou A, et al. Some prognostic models for traumatic brain injury were not valid. J Clin Epidemiol. 2006;59(2):132–43.

4. Jennett B, Snoek J, Bond MR, Brooks N. Disability after severe head injury: observations on the use of the Glasgow Outcome Scale. Neurosurgery, and Psychiatry [Internet]. 1981 [cited 2020 Dec 4];44:285–93. Available from: http://jnnp.bmj.com/

5. Mushkudiani NA, Hukkelhoven CWPM, Hernández A V., Murray GD, Choi SC, Maas AIR, et al. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. J Clin Epidemiol. 2008;61(4):331–43.

6. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting outcome after traumatic brain injury: Development and international validation of prognostic scores based on admission characteristics. PLoS Med. 2008;5(8):1251–61.

7. Perel PA, Olldashi F, Muzha I, Filipi N, Lede R, Copertari P, et al. Predicting outcome after traumatic brain injury: Practical prognostic models based on large cohort of international patients. Bmj. 2008;336(7641):425–9.

8. Steyerberg E. Clinical prediction models [Internet]. 2019 [cited 2020 Dec 3]. Available from: https://link.springer.com/content/pdf/10.1007/978-3-030-16399-0.pdf

9. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Vol. 110, Journal of Clinical Epidemiology. Elsevier USA; 2019. p. 12–22.

10. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. 2014. p. 3104–12.

11. Chung J, Gulcehre C, Cho K. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

12. Little R, Rubin D. Statistical analysis with missing data [Internet]. 2019 [cited 2021 Mar 10].

13. Stiglic G, Kocbek P, Fijacko N, Sheikh A, Pajnkihar M. Challenges associated with missing data in electronic health records: A case study of a risk prediction model for diabetes using data from Slovenian primary care. Health Informatics J [Internet]. 2019 Sep 1 [cited 2020 Dec 10];25(3):951–9. Available from: http://journals.sagepub.com/doi/10.1177/1460458217733288

14. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Sci Rep [Internet]. 2018;8(1):1–12. Available from: http://dx.doi.org/10.1038/s41598-018-24271-9

15. Yue JK, Vassar MJ, Lingsma HF, Cooper SR, Okonkwo DO, Valadka AB, et al. Transforming Research and Clinical Knowledge in Traumatic Brain Injury Pilot: Multicenter Implementation of the Common Data Elements for Traumatic Brain Injury. https://home.liebertpub.com/neu [Internet]. 2013 Oct 30 [cited 2021 Jul 6];30(22):1831–44. Available from: https://www.liebertpub.com/doi/abs/10.1089/neu.2013.2970

16. Lipton ZC, Kale DC, Wetzel R. Modeling Missing Data in Clinical Time Series with RNNs. 2016;56. Available from: http://arxiv.org/abs/1606.04130

17. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw. 2010;1–68.

18. Cheng J, Wang Z, Pollastri G. A neural network approach to ordinal regression. Proc Int Jt Conf Neural Networks. 2008;1279–84.

19. George NI, Lu T-P, Chang C-W. Cost-sensitive performance metric for comparing multiple ordinal classifiers. Artif Intell Res. 2016;5(1):135–43.

20. Baccianella S, Esuli A, Sebastiani F. Evaluation measures for ordinal regression. In: 2009 Ninth international conference on intelligent systems design and applications. IEEE; 2009. p. 283–7.

21. Kendall MG. Rank correlation methods. 1948;

22. Munteanu A, Nayebi A, Poloczek M. A framework for Bayesian optimization in embedded subspaces. In: 36th International Conference on Machine Learning, ICML 2019. 2019.

23. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems [Internet]. Neural information processing systems foundation; 2017 [cited 2021 Mar 25]. p. 4766–75. Available from: http://arxiv.org/abs/1705.07874

24. Ribeiro MT, Singh S, Guestrin C. “Why should i trust you?” Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. Association for Computing Machinery; 2016 [cited 2021 Mar 26]. p. 1135–44. Available from: http://dx.doi.org/10.1145/2939672.2939778

25. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: International Conference on Machine Learning. PMLR; 2017. p. 3145–53.

26. Folweiler KA, Sandsmark DK, Diaz-Arrastia R, Cohen AS, Masino AJ. Unsupervised Machine Learning Reveals Novel Traumatic Brain Injury Patient Phenotypes with Distinct Acute Injury Profiles and Long-Term Outcomes. J Neurotrauma [Internet]. 2020 Jun 15 [cited 2020 Nov 18];37(12):1431–44. Available from: https://www.liebertpub.com/doi/10.1089/neu.2019.6705

27. Cnossen MC, Winkler EA, Yue JK, Okonkwo DO, Valadka AB, Steyerberg EW, et al. Development of a Prediction Model for Post-Concussive Symptoms following Mild Traumatic Brain Injury: A TRACK-TBI Pilot Study. J Neurotrauma [Internet]. 2017 Aug 15 [cited 2020 Nov 18];34(16):2396–408. Available from: http://www.liebertpub.com/doi/10.1089/neu.2016.4819

28. Lingsma HF, Yue JK, Maas AIR, Steyerberg EW, Manley GT, Cooper SR, et al. Outcome prediction after mild and complicated mild traumatic brain injury: External validation of existing models and identification of new predictors using the TRACK-TBI pilot study. J Neurotrauma [Internet]. 2015 Jan 15 [cited 2021 Jan 12];32(2):83–94. Available from: /pmc/articles/PMC4291219/?report=abstract

29. Teasdale G, Skene A, Parker L, Jennett B. Age and outcome of severe head injury. Acta Neurochir Suppl (Wien). 1979;28(1):140–3.