Lab 1 - PECARN TBI Data, STAT 214, Spring 2025

§1 Introduction

Traumatic Brain Injury (TBI) is a significant cause of death and disability in children worldwide. In the United States, pediatric head trauma leads to approximately 7,400 deaths, over 60,000 hospital admissions, and more than 600,000 emergency department visits annually. Computed tomography (CT) scans are commonly used to diagnose TBI; however, their overuse poses risks due to radiation exposure, which can lead to malignancies, particularly in young children. Despite these risks, CT scans are often performed even in cases with a low likelihood of clinically-important TBI (ciTBI), leading to unnecessary healthcare costs and potential harm. [1]

The Pediatric Emergency Care Applied Research Network (PECARN) conducted a study to derive and validate clinical prediction rules for identifying children at very low risk of ciTBI after head trauma. This study analyzed data from over 42,412 patients, aiming to reduce the number of unnecessary CT scans without compromising patient safety. By understanding the key predictors associated with ciTBI and the patterns in this dataset, researchers can develop strategies to enhance clinical decision-making, minimize radiation exposure, and improve resource utilization in emergency settings.

Our exploratory data analysis (EDA) seeks to examine the dataset collected in this study, focusing on variables related to patient demographics, injury mechanisms, clinical signs, and imaging outcomes. The purpose of this EDA is to gain deeper insights into the data, identify areas for cleaning, and explore the relationships between key predictors and ciTBI. We aim to provide a clear foundation for future statistical modeling and validation efforts.

The rest of this report will outline our data cleaning approach, summarize key findings from the dataset, and highlight areas for further investigation. Through this analysis, we aim to better understand the patterns in pediatric TBI data, with the goal of supporting effective practices in emergency care.

§2 Data

§2.1 Data Collection

The study conducted by the authors [1] involved a prospective cohort of 43,499 pediatric patients under the age of 18 who presented to a hospital within 24 hours of experiencing

head trauma. Data collection took place across 25 pediatric emergency departments over a period of approximately two years, with the final months dedicated to gathering validation samples for the decision rules established in the primary study. The analysis focused on patients with Glasgow Coma Scale (GCS) scores of 14 or 15, while those with scores of 13 or lower were enrolled but analyzed separately. For each patient, a trained investigator or medical personnel systematically recorded key clinical details, including injury mechanism, medical history, and responses to standardized symptom assessments.

In the study, the outcome is ciTBI (Clinically-important TBI), which was defined as having at least one of the following: (1) neurosurgical procedure performed, (2) intubated $\geq$ 24 hours for head trauma, (3) death due to TBI or in the ED, (4) hospitalized for $\geq$ 2 nights due to head injury and having a TBI on CT.

There are 123 predictors in the dataset. Most of them are categorical or ordinal, except for age. It contains multiple binary indicator variables that track whether a child experienced specific symptoms or injuries following head trauma. These include loss of consciousness, seizure, headache, vomiting, altered mental state, palpable skull fracture, basilar skull fracture, hematoma, trauma above the clavicles, neurological deficits, and other significant non-head injuries. Additionally, many of these indicators are accompanied by detailed follow-up questions. Beyond these indicators, the dataset also includes key variables related to the injury mechanism, injury severity, and the child's clinical condition, such as whether they are acting normally, intubated, paralyzed, or sedated. We also have several demographic variables such as patient number, race, ethnicity, gender, and position of medical professional. These variables do not affect whether a patient will be positive for ciTBI.

One key variable is GCS score, which is used in emergency departments to determine a patient's level of consciousness by rating their ability for certain tests for eye and motor movement along with verbal ability. The lower the score a patient has, the lower the total GCS score, which means that the patient is in a worse situation.

§2.2 Data Cleaning

The outcome variable has some NA values. In fact, there seem to be 20 patients without the outcome variable PosIntFinal. However in the frequencies of categorical variables, the union of missing Intub24Head, Neurosurgery, HospHeadPosCT, and DeathTBI is 1 which means at most 1 of the 20 cannot be inferred from these. If all of the known values (we exclude any 'Unknown' values) of the following columns: Intub24Head, Neurosurgery, HospHeadPosCT, and DeathTBI agree then then known value is used to infer the outcome. Otherwise, the outcome is left as 'Unknown'. Using this method, we are able to infer values for all of the 'Unknown' PosIntFinal values. In the end, we figure out that there 3 values with 'Unknown', leading to 43396 known values.

Let's take a look at the rate of ciTBI based on GCS Class (3-13 vs 14-15). There are 40$ of patients with GCS scores of 3-13 being positive for ciTBI vs 0.8% of those with GCS scores of 14-15. Therefore we can safely suggest using CT scan for those patients with GCS score 3-13. Also, we will remove patients with GCS scores between 3-13.
To assess the completeness of the dataset, we examined the proportion of missing values across all variables. Notably, the variables ethnicity and dizzy exhibit the highest levels of

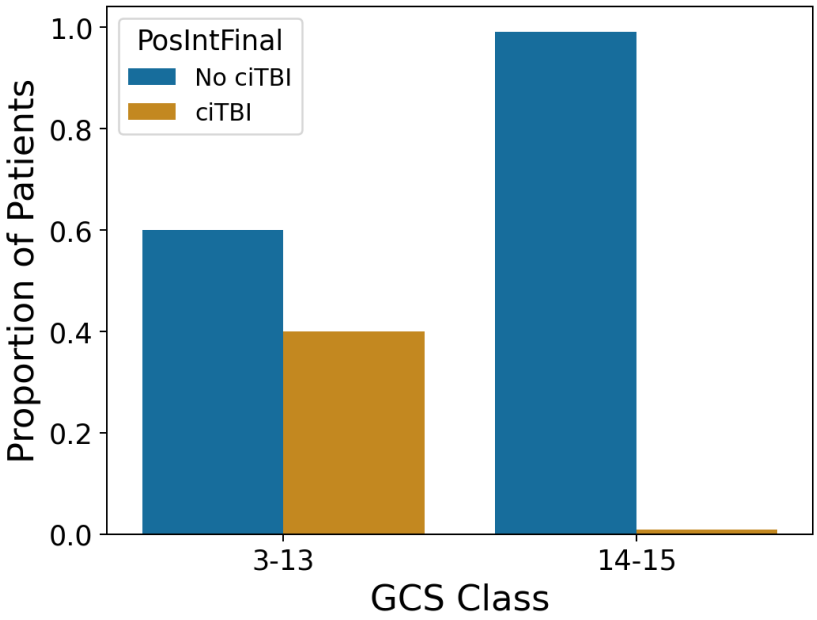

Figure 1: GCS Proportion v ciTBI

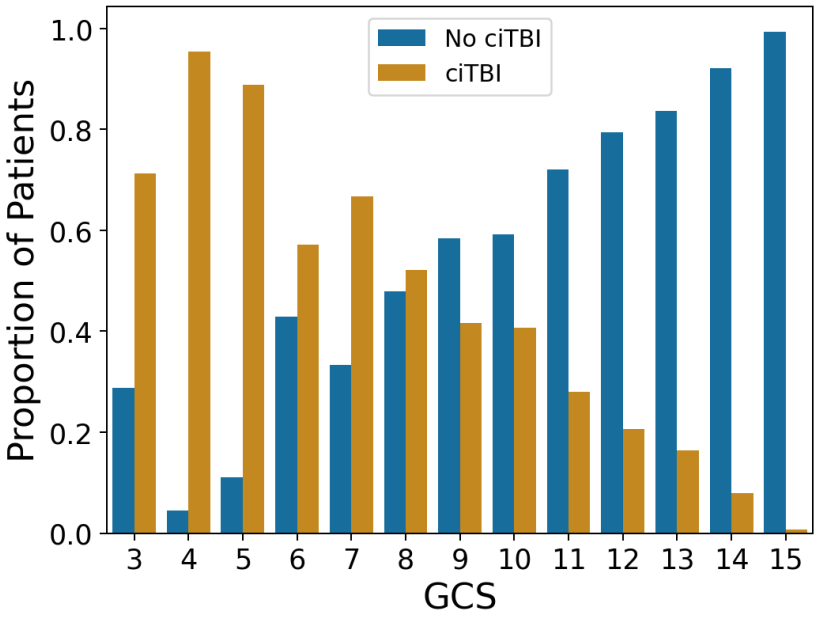

Figure 2: GCS Proportion v ciTBI by Age

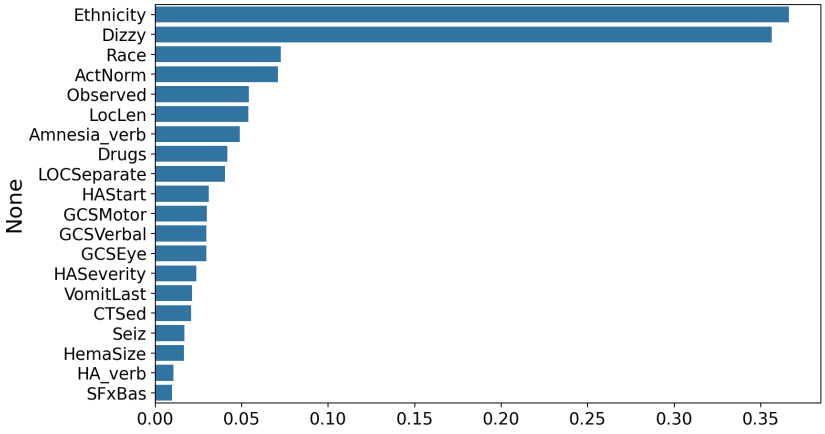

Figure 3: Missing Proportion

missingness, with more than 35% of their respective values absent. Given the substantial proportion of missing data, we determined that imputing these variables would not be methodologically justifiable. Additionally, we assessed their potential impact on the outcome variable, ciTBI, and concluded that their exclusion would not significantly affect the analysis. Consequently, we opted to remove these two variables from the dataset.

For the remaining variables, the proportion of missing values is below 10%. Given this relatively low level of missingness, we assume that the missing data mechanism is missing at random (MAR), making standard imputation techniques more appropriate for handling these instances. We opted to impute the missing values with mode, as we believe that imputing a small proportion of missing data has a negligible impact on the results. Additionally, many variables include follow-up questions. For example, the variable Vomit has a follow-up question regarding the number of vomiting episodes, which is only asked if the respondent initially reported vomiting. If the response to Vomit is 'No', the follow-up question is typically marked as 'Not Applicable'.
To streamline the analysis and maintain consistency, we converted all 'Not Applicable' responses to 'No' where appropriate. This transformation ensures that categorical variables remain analyzable without introducing additional complexity. Furthermore, we decided not to drop any observations in the dataset.

Next, we addressed the issue of categorical variables being incorrectly represented as numeric values in the dataset. Variables such as '1.0', '2.0', and '3.0' are intended to be categorical but are assigned numeric values, which may lead to misinterpretation in modeling. Then we mapped these numeric representations to their corresponding categorical labels using the data dictionary.

§2.3 Data Exploration

The proportions looking at injury severity in Figure 4 are similar across age category
From Figure 5, falls from elevation ('FallElev') appear to be the most common injury mechanism in children under 2 years old, making up more than 50% of cases. In contrast, older children show a more even distribution across different injury types, with falls, motor vehicle crashes (MVC), and sports-related injuries being more prevalent.

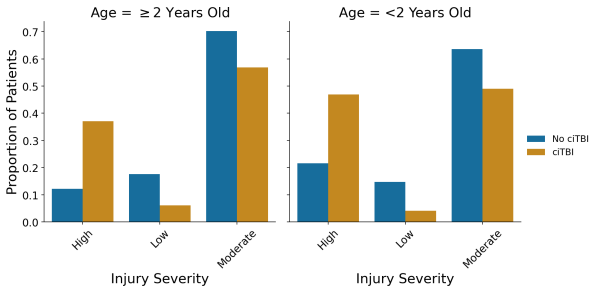

Figure 4: Proportion of patients by Age

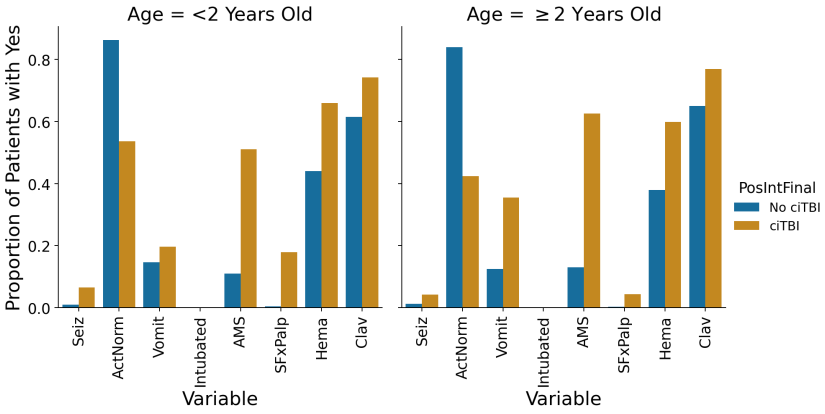

Figure 5: ciTBI by Age

In both age groups, ciTBI cases tend to follow similar proportions as non-ciTBI cases, meaning the likelihood of ciTBI appears relatively consistent across injury types.
In children under 2 years, falls from elevation and other injuries have a higher proportion of ciTBI. In children 2 years and older, motor vehicle crashes (MVC), falls from elevation, and PedesMV appear to be leading contributors to ciTBI.
Next, we look more closely at the distribution of binary variables (Fig 6) that could contribute to predicting ciTBI for each age group. We see that the proportion of ciTBI is each age group is quite the same. For each variable and age group, we see the respective proportion with patients with 'Yes'. We see that the proportion is different among two age groups for 'Vomit'.
Next, we examine if any features are with the outcome by calculating the Spearman coefficient ρ categorical variables and the binary outcome. None seem to have a really strong correlation to the outcome. The maximum correlation is around 0.13 for variable AMS (GCS $\leq$ 15 or other signs of altered mental status (agitated, sleepy, slow to respond, repetitive questions in the ED, other) and AMSOth. Some variables such as LocLen (Duration of loss of consciousness) and VomitNbr (number of vomiting episodes) show very weak correlation with ciTBI.

§3 Findings

§3.1 First Finding

After completing the exploratory analysis, we proceeded with a more detailed statistical analysis. First, we applied Principal Component Analysis (PCA) to reduce the dimensionality of the dataset while retaining as much variance as possible. Prior to implementing

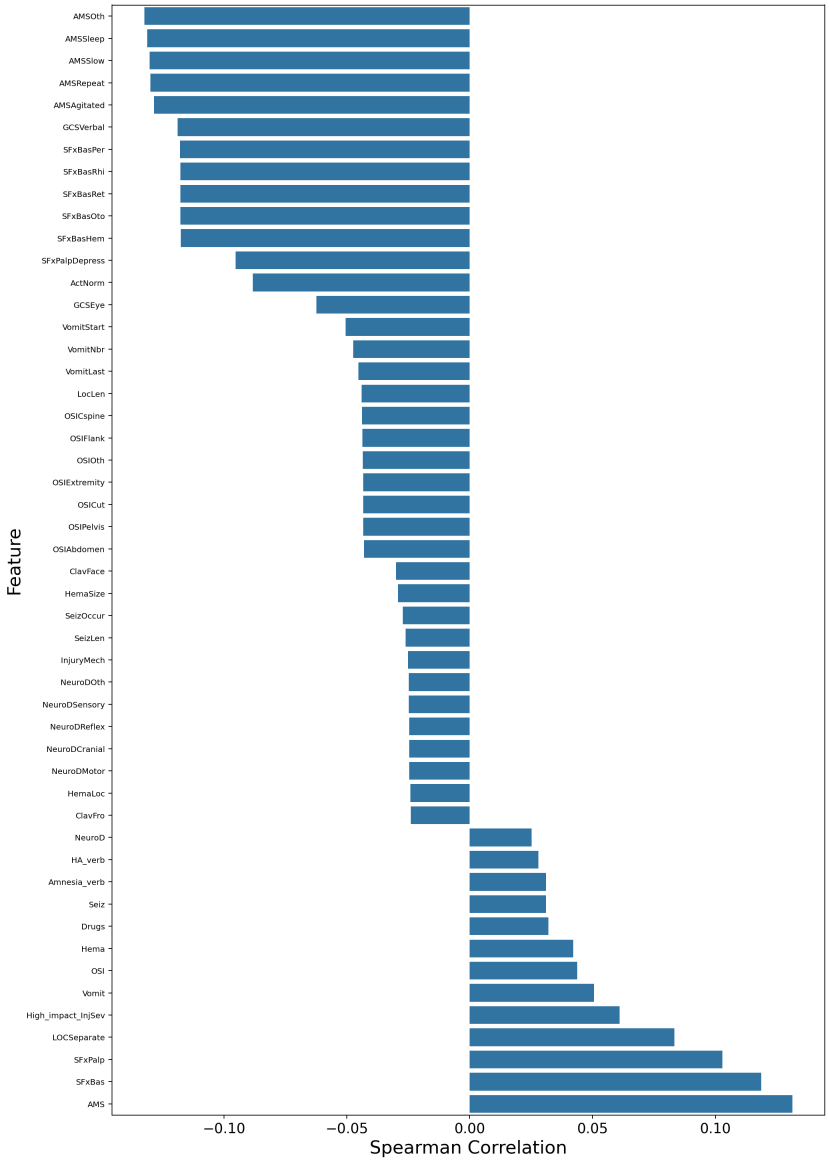

Figure 6: Spearman's Correlation between Features and Outcome

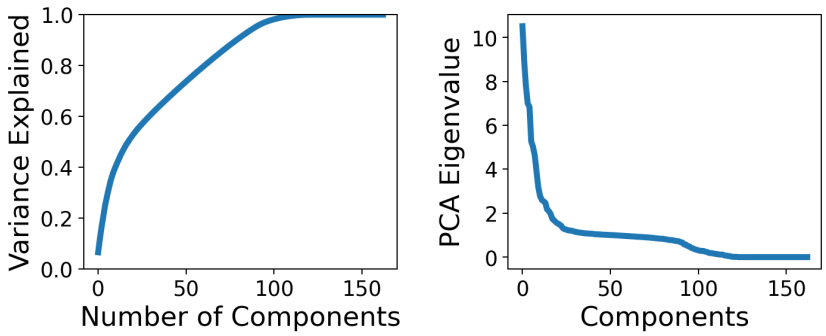

Figure 7: PCA

PCA, we excluded variables that were not relevant for predicting the outcome variable ciTBI. These included:
'Finding' variables and CT Form variables (e.g., IndAge, IndAmnesia, IndAMS), Variables directly generating the outcome variable (HospHeadPosCT, DeathTBI, HospHead, Intub24Head, Neurosurgery), Demographic variables (e.g., Ethnicity, Race, Gender, Dizzy).
After filtering out these variables, we reduced the number of predictor variables to 64, ensuring that only the most relevant features were retained for predictive modeling.
Next, we aimed to visualize the dataset using PCA with two principal components to assess its structure and investigate potential relationships with the outcome variable ciTBI and AgeGroup. This visualization allows us to explore potential clustering patterns and understand the distribution of data points across different outcome categories.
Followed by the cleaning process, after handling NA values by mode, we decide to do one-hot encoding on all categorical variables. Then we performed PCA on one-hot encoded data. Wee see that all the variance is explained by the first 100 components over 165 components totally. The first five components explains around 30$ of the total variance and the first 50 variables explains around 50% of the total variance. Given the attributes of the principal components, we acknowledge that this dataset is low-rank.
We also project the one-hot encoded data to two dimensions to study if both the age and ciTBI outcome are visible separable. Some patients with more than 2 years old and without ciTBI dominate the space, showing that most individuals belong to this category. Patients less than 2 years old without ciTBI form another large but somewhat separate cluster.
ciTBI Present cases are more sparsely distributed, implying they are less common but still cover different areas. Therefore, the overlapping regions suggest that classification of ciTBI vs. No ciTBI based on PCA alone may be challenging. However there are two distinct clusters in the data. This encourages us to keep diving into how these clusters mean. ((Fig. 8)

§3.2 Second Finding

We performed K-Means clustering with 2 clusters on the first two principal components (PC1 and PC2) derived from our dataset.
Then we calculate the difference in feature means between Cluster 1 and Cluster 0. While negative values mean The feature is more common in Cluster 0 than Cluster 1, positive values mean The feature is more common in Cluster 1 than Cluster 0. What we found is that almost all individuals in Cluster 0 did not have a recorded organ system injury

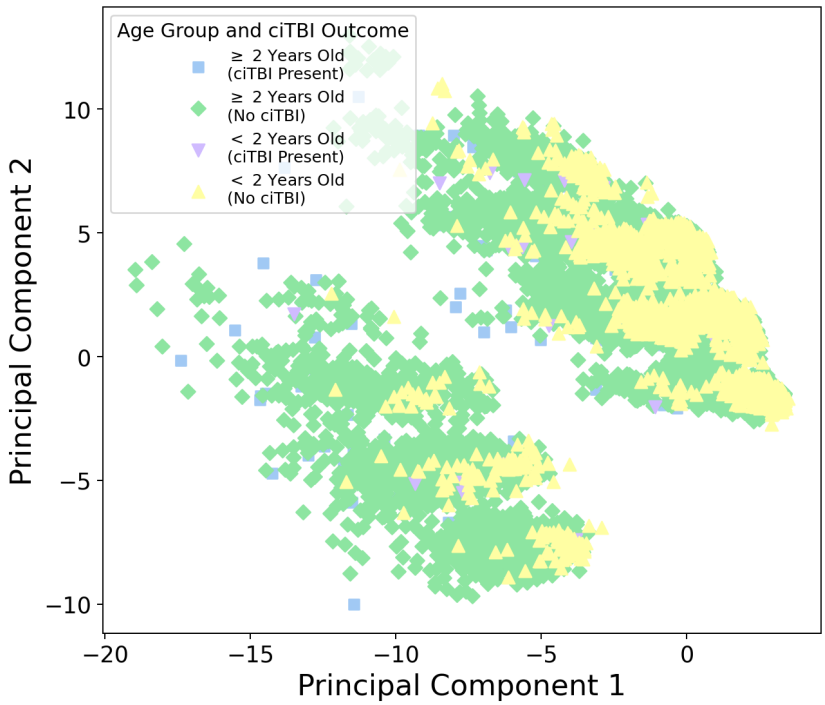

Figure 8: PCA

(OSI). Also, Cluster 0 contains younger children, possibly under 2 years old, who cannot describe their symptoms well. (Fig. 9)

Feature	**Mean Difference**
OSIPelvis_Not applicable	-0.995342
OSIExtremity_Not applicable	-0.995342
OSICspine_Not applicable	-0.995342
OSICut_Not applicable	-0.995342
OSIAbdomen_Not applicable	-0.995342
OSIOth_Not applicable	-0.995342
OSIFlank_Not applicable	-0.995342
Amnesia_verb_Pre/Non-verbal	-0.236737
HA_verb_Pre/Non-verbal	-0.229759

Table 1: Mean Differences for Selected Features

§3.3 Third Finding

We propose relative risk (RR), Table 2, which is a measure that compares the probability of an event (ciTBI in this case) occurring in one group to the probability of it occurring in another (baseline) group. If RR ≥ 1 then the group is at higher risk of ciTBI compared to reference group, otherwise, it is at lower risk.

We used 3 features (altered mental status(AMS), history of loss of consciousness (LOC-Separate), and severity of injury mechanism (high Impact InjSev) and combined them to calculate the relative risk. The reference group is when AMS, LOCSeparate, and high impact injsev is (No, No, Low).

We can observe that any combination including AMS has a much higher RR than those

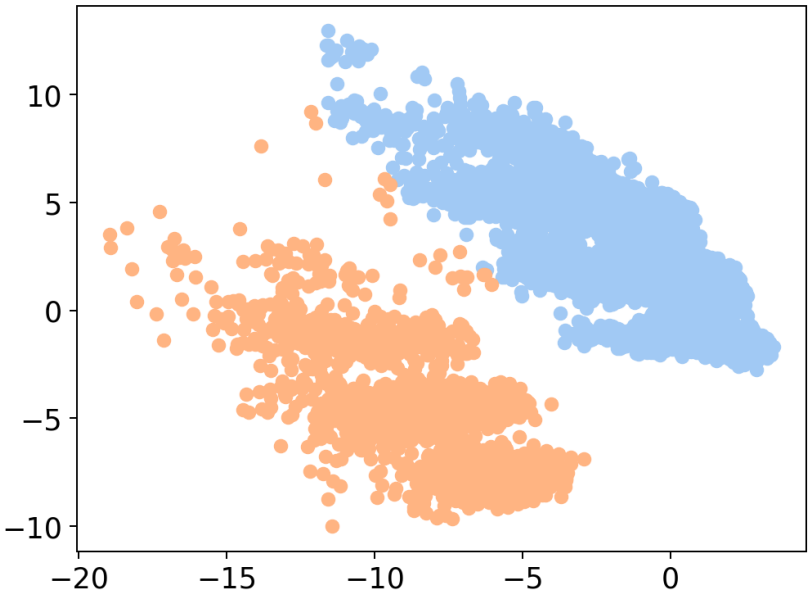

Figure 9: PCA

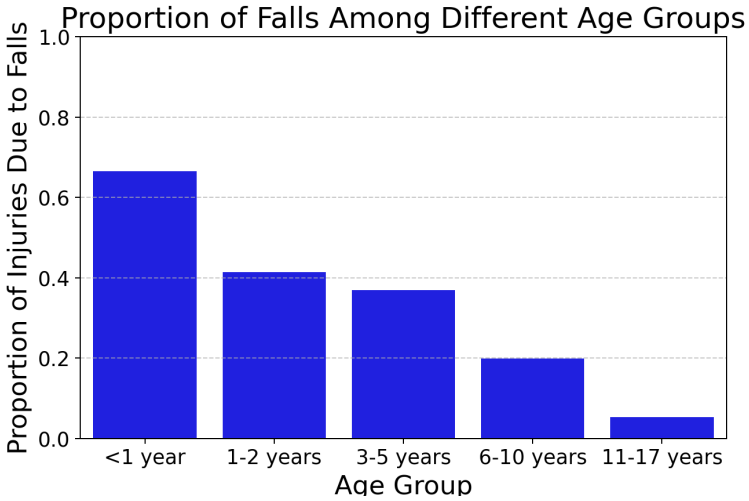

Figure 10: Proportion of Falls Among Different Age Groups

without it. Low-Impact Injuries can be dangerous if AMS and LOC are present, which means that even in mirror trauma, AMS and LOC are red flags for ciTBI.

§3.4 Reality Check

From Fig. 10, infants (less than 1 year old) have the highest proportion of fall-related injuries, with more than 65% of injuries in this age group being attributed to falls. On ther other hand, iadolescents (11-17 years old), falls contribute to less than 10% of injuries. The proportion of fall-related injuries gradually decreases as age increases.
Falls are the leading injury mechanism in young children (especially infants and toddlers). This aligns with developmental factors, as younger children are still developing motor coordination. For older children and teenagers, other injury mechanisms (e.g. motor vehicle crashes, or assaults) become more common.

Group	Relative Risk (RR)	Interpretation
(Yes, Yes, High)	157.27	Highest risk group – 157x more likely to have ciTBI compared to the reference. These patients are in critical danger and must be prioritized for imaging/treatment.
(Yes, Yes, Moderate)	60.64	Very high risk – The combination of AMS and LOC in moderate-impact injuries still increases risk 60x over baseline.
(Yes, No, High)	50.08	High-impact injuries with AMS (but no LOC) still have a 50x increased risk. This shows AMS alone is a major predictor of ciTBI.
(Yes, Yes, Low)	34.98	Even low-impact injuries with both AMS and LOC increase ciTBI risk significantly. This suggests mechanism of injury alone is not enough—neurological symptoms matter.
(No, Yes, Low/Moderate)	∼8.4	LOC in low/moderate injuries increases ciTBI risk significantly, but less than AMS.
(No, No, High)	7.76	High-impact injuries alone increase ciTBI risk 7.7x, even without AMS or LOC.
(No, No, Moderate)	2.15	Moderate-impact injuries have only a small increase in risk (2.1x) if no AMS or LOC is present.

Table 2: Relative Risk (RR) of ciTBI for Different Groups

§3.5 Stability Check

We split the data into training and validation set randomly, and then we repeat the plot from Figure 6. What we obtained is Figure 12, which we see the consistent patterns in proportion of patients with 'Yes' in each age group when we randomly split the dataset.

§4 Modeling

§4.1 Implementation

In this section, we perform fitting models to predict whether patients have a ciTBI from their predictors. We perform logistic regression for the ease of interpretability. The metric we want to focus on is the recall score because we want to accept decision rules with extremely low missed positives (ciTBI cases) and we want to minimize unnecessary CT scans.
Next, we split the dataset into 2 groups by age group (less than 2 years old and more than 2 years) and fit models for each group. For each, we split the dataset into 70% for training and 30% for validation.

§4.2 Interpretability

From table 3, we have Parietal/Temporal Hematoma with coefficient (1.756), showing that hematoma in the parietal or temporal region is the strongest predictor of ciTBI, suggesting that trauma to these regions is highly associated with serious brain injuries in infants.

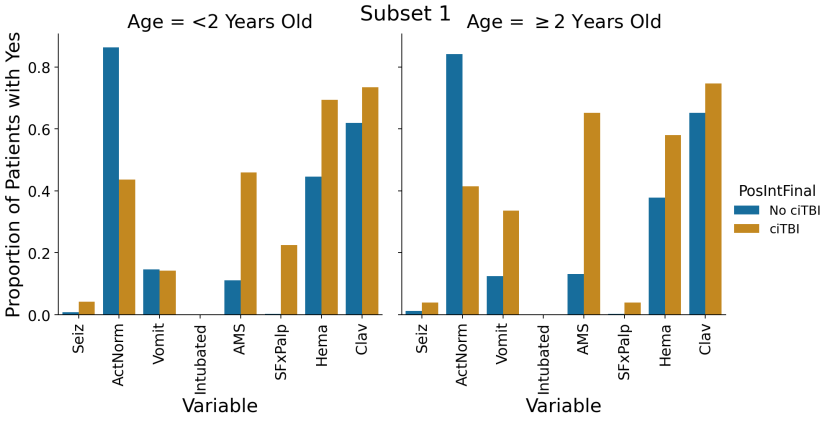

Figure 11: Subset 1 Stability Check

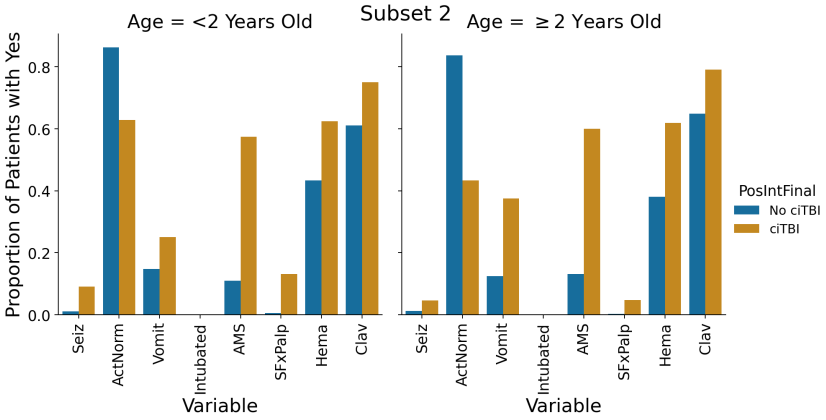

Figure 12: Subset 2 Stability Check

Feature	Coefficient
HemaLoc_Parietal/Temporal	1.755616
InjuryMech_MVC	1.247799
High_impact_InjSev_Low	-1.069603
SFxPalpDepress_Yes	1.066085
HemaLoc_Occipital	1.060413
⋮	⋮
Vomit_Yes	-0.012179
InjuryMech_Sports	-0.009567
HASeverity_Moderate	-0.009280
InjuryMech_BikeMV	-0.006922
GCSMotor_Pain withdraws	-0.004161

Table 3: Feature Coefficients for Age Group ≤ 2

Also, infants involved in motor vehicle crashes have higher odds of sustaining a ciTBI. This aligns with expectations that such crashes increase the likelihood of severe head trauma.
The presence of a depressed skull fracture is associated with ciTBI. This is expected because a depressed skull fracture suggests significant force, making intracranial injuries more likely.

Feature	Coefficient
InjuryMech_OtherWheelCrash	1.151628
SFxBasHem_Yes	1.062050
InjuryMech_BikeMV	1.014491
AMSAgitated_Yes	0.907694
SFxPalpDepress_Yes	0.899336
⋮	⋮
OSICut_Not applicable	0.002731
OSIFlank_Not applicable	0.002731
OSIAbdomen_Not applicable	0.002731
OSIPelvis_Not applicable	0.002731
OSIOth_Not applicable	0.002731

Table 4: Feature Coefficients for Age Group ≥ 2

Followed by table 4, children involved in non-bicycle wheeled crashes (e.g., scooters, skateboards, ATVs) are at a significantly higher risk of ciTBI. These vehicles provide less protection than cars and less stability than bicycles, increasing the severity of head injuries.
Moreover, a basilar skull fracture with associated hemorrhage is strongly associated with ciTBI. This aligns with clinical expectations, as such fractures indicate high-force trauma to the skull base, increasing the risk of intracranial bleeding.
A bicycle crash involving a motor vehicle increase the likelihood of ciTBI, suggesting that collisions with moving vehicles lead to more severe head trauma in older children.
Agitation as a sign of altered mental status (AMS) is also strong predictor of ciTBI. Agitation could indicate neurological impairment, possibly due to intracranial pressure or hemorrhage.

Similar to younger children, a palpable depressed skull fracture is a strong indicator of ciTBI. This suggests direct skull trauma with potential underlying brain injury, necessitating urgent evaluation

§4.3 Stability

To perform stability, we propose two approaches: bootrapping the training dataset and evaluating the bootstrapped logistic regression models on the validationset and doing k-fold cross validation.
In this report, we would implement the first approach by doing 10 bootstrapped training dataset and fit the logistic regression model on that dataset before we calculate the log-loss on the validation set. Ideally, we expect a small standard error among the log-loss on the validation set.

Log Loss	**Age ≥ 2**	**Age < 2**
Mean	0.0387	0.0407
Standard Deviation	0.0008	0.0013

Table 5: Bootstrap Log-Loss Summary for Different Age Groups

We present mean and standard deviation of bootstrapped log-loss. There is a tight concentration close to the various log-loss we could obtain from the validation set.

§5 Discussion

The logistic regression model exhibited poor sensitivity in both age groups, with a sensitivity score of approximately 0.1 on the validation set. Despite achieving high overall accuracy, this low sensitivity indicates that the model fails to correctly identify a significant proportion of ciTBI cases. This issue stems from the imbalance in the dataset, where ciTBI cases are much less frequent compared to non-ciTBI cases. As a result, the model is biased toward predicting the majority class, leading to high accuracy but poor recall for the minority class.
To improve the model's ability to detect ciTBI cases, several techniques can be considered:
- Further feature engineering: Further imputation method or feature engineering techniques can be explored to enhance the model performance
- Oversampling the minority class (ciTBI) using SMOTE (Synthetic Minority Oversampling Technique) or simple random oversampling. Undersampling the majority class (non-ciTBI) to create a more balanced training set.
- Instead of using the default 0.5 threshold, selecting a lower probability threshold can help increase sensitivity while maintaining reasonable precision.
- Tree-based models such as Random Forest or Gradient Boosting Machines (GBM) may perform better in handling class imbalances.

§6 Conclusion

In this report, we have looked at patient data from a prospective cohort study wherein patients between the ages of 0-17 who visited one of a series of hospitals presenting with a potential TBI were enrolled. We were given data from patient questionnaires that was converted to numerical features and fit models to predict the need for a CT scan. We also

performed data cleaning and feature engineering. Also, we did feature selection using domain knowledge to remove unnecessary variables, decreasing from 123 to 64. Next, we worked on some EDA to discover insights about ciTBI and features such as injury mechanism, symptoms and recorded patient history. Finally, we implemented logistic regression on each age group to predict ciTBI cases. Although we did not achieve the expected metric, we believe that a larger dataset with more positive samples (more ciTBI cases) would be helpful in this task.

§7 Academic honesty statement

References

[1] Nathan Kuppermann et al. "Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study". In: *Lancet (London, England)* 374.9696 (2009), pp. 1160–1170. DOI: 10.1016/S0140-6736(09)61558-0.