Lab 1 - PECARN TBI Data, STAT 214, Sp25

Soohyun Kim

2025-02-19

1 Introduction

Traumatic Brain Injury (TBI) is a critical public health concern, particularly in pediatric populations, where timely and accurate diagnosis is essential to prevent severe health consequences. The PECARN (Pediatric Emergency Care Applied Research Network) study aimed to develop a clinical decision rule that could identify children at very low risk of clinically important TBI (ciTBI), thus reducing unnecessary CT scans and minimizing radiation exposure. Understanding the data collected in this study is crucial for assessing the validity of the clinical prediction rule and determining areas for improvement.

In this report, we conduct an exploratory data analysis (EDA) of the TBI Public Use Dataset (PUD) to investigate data quality, patterns, and relationships among variables. This process involves assessing missing data, identifying inconsistencies, and determining appropriate data cleaning techniques. Given the high stakes of TBI diagnosis, our analysis will also examine how different patient-level factors influence the likelihood of ciTBI and whether they align with the prediction model developed in the PECARN study.

2 Data

This study utilizes the PECARN TBI Public Use Dataset (PUD) to explore the effectiveness of clinical prediction rules for identifying clinically important traumatic brain injuries (ciTBI) in children. The dataset contains patient-level features collected from multiple pediatric emergency departments across North America, focusing on children under 18 years of age who presented with minor head trauma. The primary goal of analyzing this dataset is to assess data quality, identify key patterns, and evaluate the potential for improving clinical decision-making. The dataset provides rich information on patient demographics, clinical symptoms, injury mechanisms, and outcomes, including whether a child developed ciTBI and required further intervention. There are a total of 43399 observations(patients) with 125 columns.

2.1 Data Collection

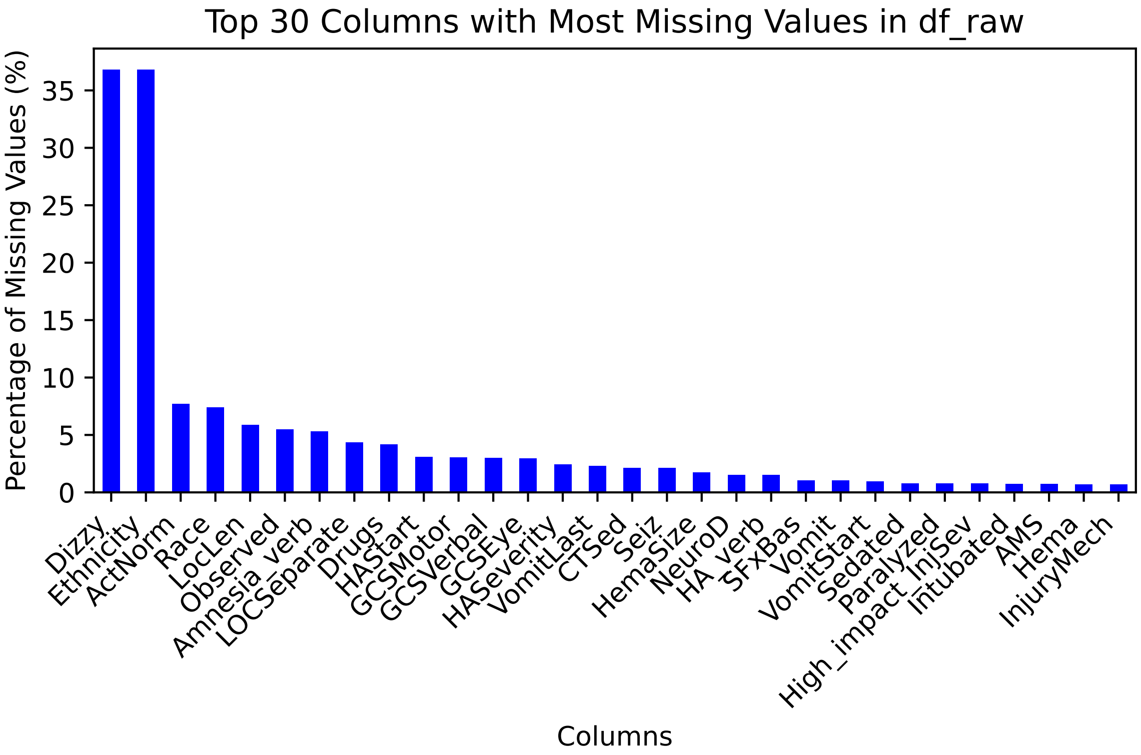
The data were collected as part of a prospective observational cohort study conducted across 25 PECARN emergency departments. The study enrolled children who presented to the emergency department within 24 hours of experiencing blunt head trauma. Patient history and clinical symptoms were recorded by trained site investigators and emergency department physicians using a standardized data form before knowing imaging results. The dataset includes Glasgow Coma Scale (GCS) scores, indicators of altered mental status, loss of consciousness, vomiting, and other clinical signs that help determine the severity of head trauma. Imaging data (CT scans) were obtained at the discretion of the emergency department physician rather than being required for every patient, meaning that not all cases have CT scan results.

2.2 Data Cleaning

To prepare the dataset for analyzing factors influencing clinically important traumatic brain injury (ciTBI), we first removed columns deemed irrelevant to our predictive goal. This included patient identifiers (e.g., 'PatNum'), physician-related information, and redundant or overly specific variables. Additionally, detailed physical exam findings and CT scan ordering reasons were excluded, as our focus is solely on whether a CT scan was performed and the presence of ciTBI. Columns with excessive missing values (over 35%), such as 'Dizzy' and 'Ethnicity', were also removed, as their limited data would not provide meaningful insights.

Next, we addressed missing data based on three cases. First, rows with missing values for ciTBI, our primary outcome, were dropped to ensure the integrity of our analysis. For columns with less than 1% missing data, missing values were replaced with the mode to maintain consistency. Variables with 2% to 5% missingness were handled on a case-by-case basis, using appropriate imputations. Specific features such as LocSeparate (indicating loss of consciousness) were filled with a designated unknown category (92), while LocLen (duration of loss of consciousness) was imputed using the mode based on whether LocSeparate was 1 or 2. Similarly, for headache-related features (HA_verb, HAStart, HASeverity), missing values were either replaced with 92(inapplicable) or filled using the mode within their respective categories. This ensured that valuable information was not lost while maintaining the interpretability of patient symptoms.

For symptom-related columns such as vomiting (Vomit), hematoma (Hema), seizures (Seiz), and neurological deficits (NeuroD), missing values were also handled using mode imputation, following the same structured approach as the headache and loss of consciousness variables. Additionally, to simplify the Glasgow Coma Scale (GCS) features, we retained only GCSTotal, as it had no missing values, and removed individual GCS components (GCS Motor, GCS Verbal, GCS Eye). Finally, categorical variables were recoded for interpretability, such as replacing unclear responses (2) with "not applicable" (92) and simplifying injury severity into binary classifications (e.g., 'High_impact_InjSev' categorized as 0 for mild, 1 for moderate/high). These cleaning steps ensured the dataset was structured, complete, and ready for meaningful exploratory analysis and predictive modeling.



```
No NaN values for all columns.
Cleaned df shape is:  (43379, 55)
```

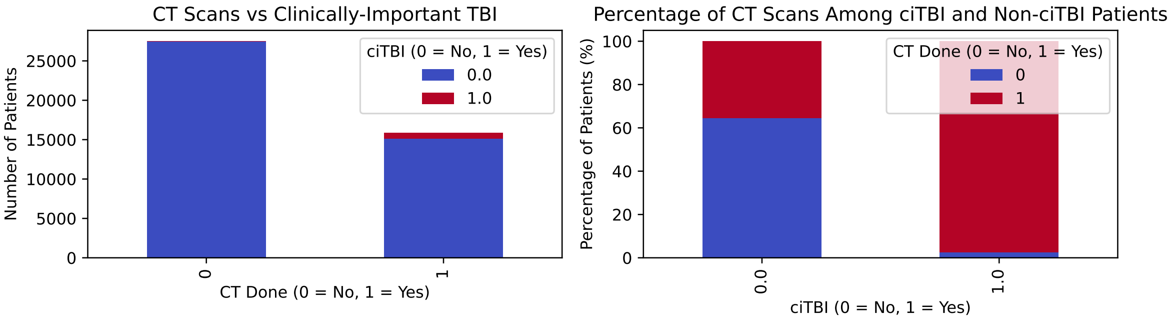
I have used the cleaned the data using the clean_data function in the clean.py file, and as we can see, there are no NA values in the cleaned dataframe, and the number of columns has decreased from 125 to 55, keeping only the essential columns that are relevant to our project.

2.3 Data Exploration

First, we examine the distribution of the final outcome column, which indicates whether a patient was diagnosed with clinically important traumatic brain injury (ciTBI). This step helps us understand the overall proportion of children with ciTBI in the dataset.

	Count	Percentage
PosIntFinal		
0.0	42616	98.241084
1.0	763	1.758916

From the frequency chart above, we observe that less than 2% of the children were diagnosed with ciTBI, highlighting a significant class imbalance in the data. Given this, our next step is to analyze the proportion of children who underwent CT scans and compare it with the actual ciTBI diagnoses. Specifically, we aim to determine how many children received a potentially unnecessary CT scan, as they ultimately did not have ciTBI. This analysis is crucial for evaluating the efficiency of clinical decision rules in reducing unnecessary radiation exposure while ensuring accurate diagnoses.



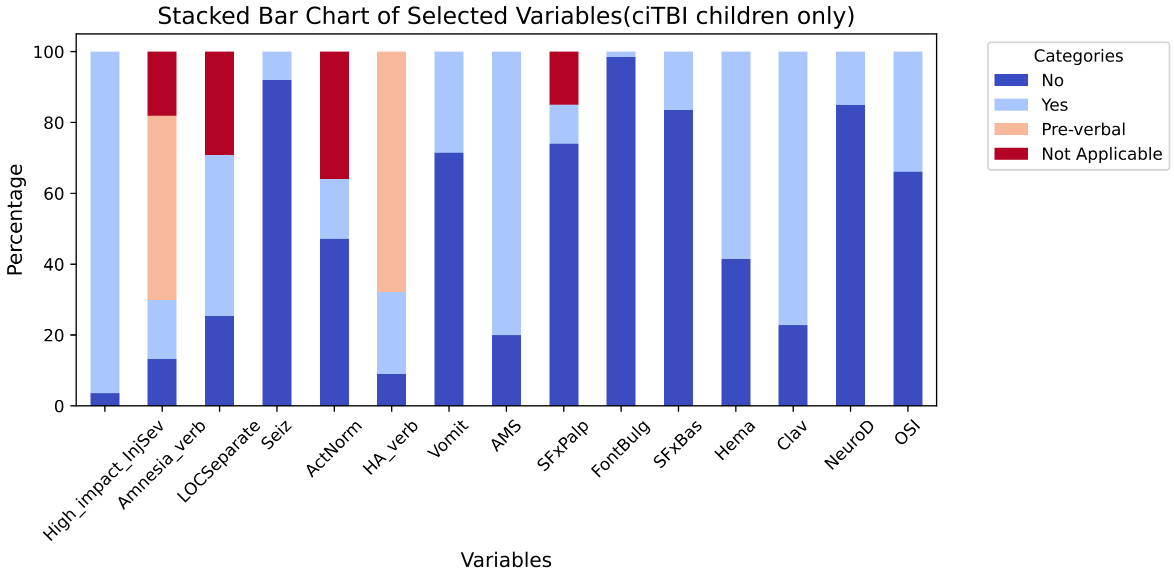
On the left chart we can see that most children did not receive a CT scan (left bar, "0" in CTDone). Among those who did receive a CT scan (right bar, "1" in CTDone), only a small fraction were diagnosed with clinically-important TBI (ciTBI). This suggests that many children underwent CT scans despite not having ciTBI, raising potential concerns about unnecessary radiation exposure. In the right chart, among children who did not have ciTBI (0 in PosIntFinal), a significant proportion still received CT scans (red section), meaning many CT scans were performed on children without severe injury. Among children who did have ciTBI (1 in PosIntFinal), nearly all of them received a CT scan, indicating high sensitivity in using CT scans for those who truly needed it. Additionally, the small blue section in the ciTBI group (1 in PosIntFinal) represents children who had ciTBI but did not receive a CT scan, suggesting a missed diagnosis risk. These findings suggest a need for improved decision rules to minimize unnecessary CT scans while ensuring that high-risk patients are properly diagnosed.

3 Findings

3.1 First Finding: Distribution of ciTBI Children Across Categories

To better understand the distribution of clinically important traumatic brain injury (ciTBI) cases across different categorical variables, we filtered the dataset to include only children diagnosed with ciTBI. We then calculated the percentage distribution for key variables, allowing us to assess which factors are most strongly associated with ciTBI. At this stage, we focus on high-level categories such as Altered Mental Status (AMS), Scalp Hematoma (Hema), Basilar

Skull Fracture (SFxBas), Neurological Deficits (NeuroD), and Other Significant Injury (OSI) while excluding overly specific features.

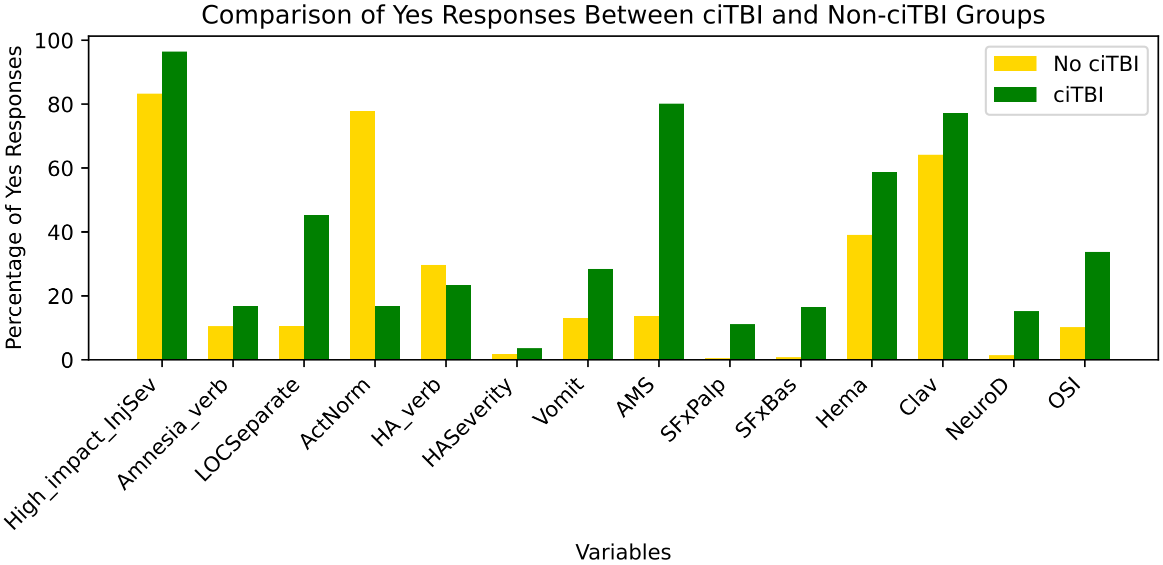


The distribution analysis of children with clinically important traumatic brain injury (ciTBI) reveals several key insights. Certain features, including High-Impact Injuries, Loss of Consciousness (LOC), Altered Mental Status (AMS), Scalp Hematoma (Hema), Clavicle Injury (Clav), and Other Significant Injury (OSI), show a high proportion of "Yes" responses, suggesting they are strong predictors of ciTBI. Conversely, Acting Normally (ActNorm) has a high proportion of "No" responses, indicating that a large share of ciTBI cases involve abnormal behavior at the time of evaluation. In contrast, Seizures (Seiz) and Anterior Fontanelle Bulging (FontBulg) exhibit very low associations with ciTBI, suggesting they may not serve as reliable standalone predictors. Some variables, such as Amnesia, Headache, Vomiting, Basilar Skull Fracture (SFxBas), and Neurological Deficit (NeuroD), display a mixed distribution of responses, indicating a more context-dependent relationship with ciTBI. Notably, the PECARN study identified AMS, LOC, Mechanism of Injury, Acting Normally, and Scalp Hematoma as critical predictors, aligning well with our findings. However, Headache, Vomiting, and Basilar Skull Fracture, which were considered significant in the study, exhibit a more ambiguous distribution in our dataset. This suggests that further statistical testing, such as logistic regression or chi-square tests, may be necessary to determine their predictive significance.

3.2 Second Finding: Comparison of ciTBI and Non-ciTBI Cases Across Categories

The bar chart below compares the percentage of "Yes" responses for selected variables between children with and without clinically important traumatic brain injury. The selected features include strong and mixed indicators from Finding 1, such as High-Impact Injuries, Loss of

Consciousness (LOC), Altered Mental Status (AMS), Scalp Hematoma (Hema), Clavicle Injury (Clav), Other Significant Injury (OSI), Amnesia, Acting Normal, Headache (HA), Vomiting (Vomit), Basilar Skull Fracture (SFxBas), and Neurological Deficit (NeuroD). We exclude Seizures (Seiz) and Anterior Fontanelle Bulging (FontBulg) due to their weak association with ciTBI. Additionally, we introduce Headache Severity (HASeverity) to validate the findings from the PECARN prediction rule in the original study. A feature is considered a strong predictor if its "Yes" percentage is significantly higher in ciTBI cases compared to non-ciTBI cases. Conversely, if the "Yes" percentages are similar in both groups, the feature is likely less useful for prediction.



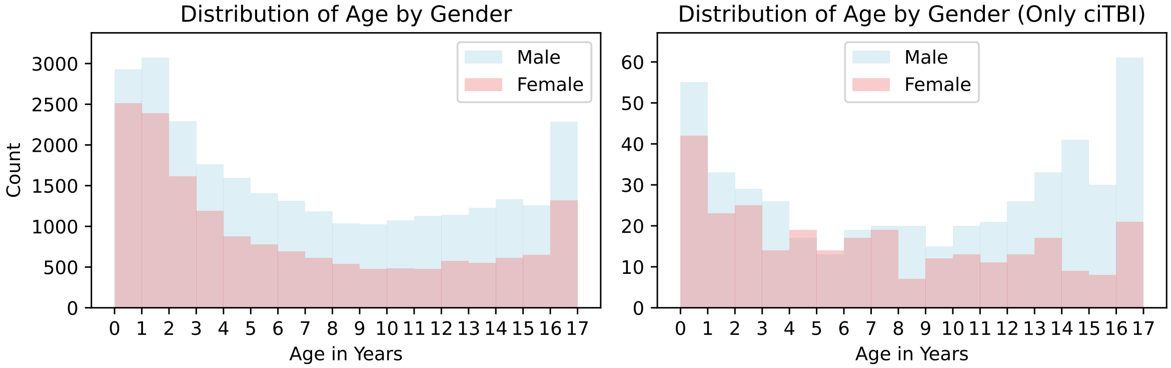
Building on Finding 1, the chart clearly shows that for almost all variables, the percentage of "Yes" responses is higher in ciTBI cases than in non-ciTBI cases. Notably, the largest differences between ciTBI and non-ciTBI groups occur in the following variables:Loss of Consciousness (LOC), Acting Normal (ActNorm) (where over 80% of ciTBI cases responded "No"), Altered Mental Status (AMS), Basilar Skull Fracture (SFxBas), Palpable Skull Fracture (SFx-Palp), Neurological Deficits (NeuroD). Most of these variables, except for NeuroD, were identified in the PECARN prediction rule as key indicators for ciTBI, reinforcing the consistency between the dataset and the findings in the paper.

Interestingly, while NeuroD and OSI also exhibit a high percentage difference between ciTBI and non-ciTBI cases, they were not included in the PECARN prediction rule. However, this does not contradict the study. The prediction rule was designed as a concise screening tool, meaning that while NeuroD and OSI may still be associated with ciTBI, they might not have been as impactful as other variables in reducing unnecessary CT scans.

Finally, HA_verb (self-reported headache) stands out as an exception, with a lower "Yes" percentage in ciTBI cases compared to non-ciTBI cases. This suggests that headache alone may not be a strong predictor of ciTBI. Overall, the bar chart aligns well with the PECARN study's key findings, demonstrating that variables like AMS, LOC, Mechanism of Injury, Acting Normal, and Scalp Hematoma remain strong indicators of ciTBI. However, additional exploratory analysis is needed to fully understand the role of variables like NeuroD and OSI, which show notable differences but were not included in the final prediction model.

3.3 Third Finding: Distribution of Age by Gender (All Children vs. ciTBI Cases)

Understanding the distribution of age by gender is crucial in assessing which age groups are most likely to be tested for TBI and which are at higher risk of clinically important traumatic brain injury (ciTBI). By comparing the full dataset (left histogram) with the subset of children diagnosed with ciTBI (right histogram), we can identify potential age-related patterns and gender differences in TBI risk.



The histograms reveal distinct age-related trends in TBI evaluation and diagnosis. In both the full dataset and ciTBI subset, younger children (ages 0-3) are more likely to be evaluated for TBI, but the number of cases gradually declines as age increases. This likely reflects the higher clinical caution exercised for young children, as they may have difficulty verbally expressing symptoms and are at greater risk for severe injury consequences. Interestingly, we observe a secondary peak at ages 16-17, suggesting that older adolescents experience a higher incidence of head trauma requiring evaluation. This could be attributed to increased participation in high-risk activities, such as contact sports, driving-related accidents, or riskier physical behaviors.

When focusing specifically on ciTBI cases (right histogram), we see a notable gender difference. While female cases decline more steadily with age, male cases show a sharp increase in ciTBI diagnoses at ages 16-17. This aligns with existing research suggesting that older male adolescents are at greater risk for severe head trauma due to higher engagement in contact sports and risk-taking behaviors. Additionally, in the full dataset, males are more frequently tested for TBI, and in the ciTBI subset, the number of cases in older ages (16-17) is nearly twice as high for males compared to females. This suggests that males are not only more frequently evaluated for TBI but also more likely to be diagnosed with ciTBI in later years. Overall, this analysis highlights the importance of considering both age and gender when assessing TBI risk and diagnostic patterns.

3.4 Reality Check

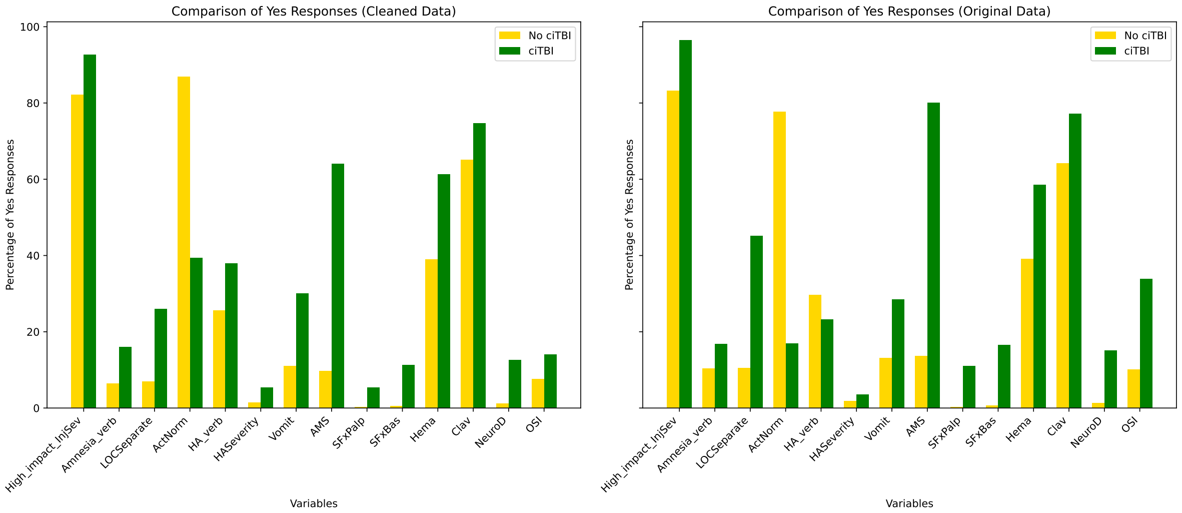
After extensive data cleaning, we verify whether our dataset aligns with real-world epidemiological trends. Our findings show that males, particularly in adolescence (16-17 years old), have higher ciTBI rates than females, which aligns with established research from the CDC, JAMA Pediatrics, and WHO. Studies confirm that males are at greater risk for TBIs due to higher participation in contact sports, increased risk-taking behavior, and anatomical differences. Additionally, the observed higher birth rate of male infants (105 males per 100 females) supports the slight male dominance in our dataset. However, we must consider the possibility of medical testing biases, where males may have been more frequently tested for TBI than females, potentially inflating their ciTBI representation. Further investigation is needed to determine whether gender-based differences in testing frequency impact the dataset's findings.

3.5 Stability Check

In this section, we assess the impact of different data cleaning strategies on our findings. Initially, we handled missing values by imputing them with the mode (most frequent value) or assigning 92 ("Inapplicable") for variables with less than 8% missingness. However, in this analysis, we take an alternative approach by removing all rows that contained any missing values in these columns. This allows us to evaluate whether excluding ambiguous data points leads to significant differences in our findings. After applying this stricter cleaning method, we observe that the ciTBI rate decreases from 1.75% in the original dataset to 0.5% in the cleaned dataset, meaning that a greater proportion of ciTBI cases were removed when we excluded missing data.

```
new df for check has no NaN values for all columns.
new df for stability check has dimension of  (27745, 58)
```

	Count	Percentage
PosIntFinal		
0.0	27595	99.459362
1.0	150	0.540638



The graphs above compare the Yes response percentages for key variables in children with and without ciTBI before and after the new cleaning method. Despite minor fluctuations in the bar heights, a consistent trend is maintained across both datasets: All columns except HA_verb (headache presence) show higher Yes percentages in ciTBI children compared to non-ciTBI children. Key indicators from the PECARN study—such as High-Impact Injury, Acting Normal, and Altered Mental Status (AMS)—exhibit an even stronger contrast between ciTBI and non-ciTBI children after stricter data cleaning. This finding is particularly interesting, as it suggests that HA_verb (whether a child reported a headache) is not a reliable predictor of ciTBI, aligning with the conclusions from the PECARN study. Additionally, the increased separation between ciTBI and non-ciTBI groups in critical predictor variables may indicate that rather than imputing missing values, a more accurate approach to refining a clinical prediction rule might be to eliminate ambiguous data points altogether. This stricter method of handling missing data could enhance the model's ability to differentiate between ciTBI and non-ciTBI cases.

4 Modeling

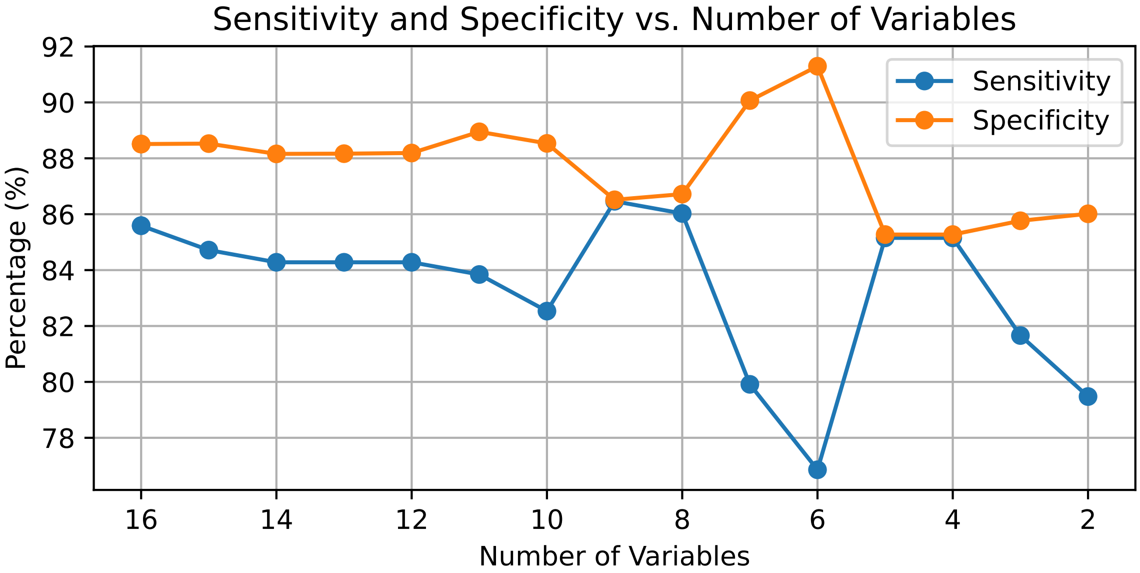
4.1 Implementation

Logistic Regression Model

To build our logistic regression model, we first refine the dataset by addressing missing and inapplicable values. We drop columns where more than 40% of values are 92, as these variables provide limited predictive value. For the remaining columns, we manually select features based on Findings 1, 2, and 3, ensuring we retain variables that showed a strong association with ciTBI. The selected features include key clinical and mechanism-based predictors (High-ImpactInj, LOCSeparate, Seiz, ActNorm, Vomit, GCSGroup, AMS), physical examination findings (SFxPalp, FontBulg, SFxBas, Hema, Clav, NeuroD, OSI), and demographic variables (AgeInYrs, Gender). We exclude features such as Amnesia and HA_verb due to their inconsistent predictive value in previous findings, as well as specific clavicle fracture details and race. Additionally, for variables LOCSeparate, ActNorm, and SFxPalp, where the proportion of 92 values is less than 0.1%, we replace these values with the mode, which is 0.

```
Initial Model Performance (All Variables):
Sensitivity (Recall): 85.59%
Specificity: 88.51%
Positive Predictive Value (PPV): 11.77%
Negative Predictive Value (NPV): 99.71%
Accuracy: 88.46%

Best Model Performance (After Backward Elimination):
Features: ['LOCSeparate', 'ActNorm', 'GCSGroup', 'AMS', 'SFxPalp', 'SFxBas', 'Hema', 'NeuroD
Sensitivity (Recall): 86.46%
Specificity: 86.52%
Positive Predictive Value (PPV): 10.30%
Negative Predictive Value (NPV): 99.72%
Accuracy: 86.51%
```



To identify the optimal set of variables for predicting ciTBI in children, I used a backward elimination approach in the logistic regression model, starting with all 16 predictor variables (High_impact_InjSev, LOCSeparate, Seiz, ActNorm, Vomit, GCSGroup, AMS, SFxPalp, FontBulg, SFxBas, Hema, Clav, NeuroD, OSI, AgeTwoPlus, Gender). The process involved fitting an initial model with class_weight='balanced' to address the dataset's class imbalance (98.24% no ciTBI vs. 1.76% ciTBI), then iteratively removing the least significant variable based on p-values from statsmodels logistic regression. I evaluated each model on the test set, and continued until only one variable remained or no significant variables could be removed. The "best model" was selected by maximizing sensitivity, minimizing false negatives critical in this clinical context.

The graph illustrating sensitivity and specificity versus the number of variables in the logistic regression model for predicting ciTBI in children reveals a dynamic trade-off as variables are removed through backward elimination. Starting with all 16 variables, the initial model achieved a sensitivity of 85.59% and specificity of 88.51%, reflecting a balanced performance. As variables were iteratively excluded based on p-values, sensitivity fluctuated, peaking at 86.46% with a reduced set of 9 variables (LOCSeparate, ActNorm, GCSGroup, AMS, SFxPalp, SFxBas, Hema, NeuroD, OSI), while specificity slightly decreased to 86.52%. This indicates that removing less significant variables improved sensitivity marginally but maintained a reasonable specificity, optimizing ciTBI detection while managing false positives, though PPV remained low (10.30%) due to the rarity of ciTBI, and NPV stayed high (99.72%) for ruling out ciTBI.

4.2 Interpretability

The logistic regression model predicts ciTBI in children, with its best feature set of 9 variables. Below is a graph that summarizes what the coefficients mean in terms of log odds. As we can see, palpable skull fracture (SFxPalp) and basilar skull fracture (SFxBas) are the strongest predictors, while acting normally significantly reduces risk. Since logistic regression provides probability estimates for ciTBI (i.e. P(Y=1|X) = 1/(e^-betaX)), for real-world use, a probability threshold or a clinical risk score can make the model actionable. For instance, a clinician could set a probability threshold (e.g., 10% probability → recommend a CT scan). This allows a risk-based approach instead of binary "Yes/No" predictions.

Converted Odds Ratios for Clinical Interpretation

Variable	Log-Odds ( )	Odds Ratio (e^)	Interpretation
LOCSeparate	0.9087	2.48	LOC increases ciTBI likelihood **2.48×**
ActNorm	-1.0951	0.33	Acting normal **reduces ciTBI risk by 67%**
GCSGroup	-2.0216	0.13	Higher GCS score **reduces ciTBI risk by 87%**

Variable	Log-Odds ( )	Odds Ratio (e^)	Interpretation
AMS	1.6208	5.06	AMS increases ciTBI likelihood **5.06**×
SFxPalp	3.1911	24.30	Palpable skull fracture increases risk **24**×
SFxBas	2.9986	20.07	Basilar skull fracture increases risk **20**×
Hema	0.8941	2.44	Scalp hematoma increases risk **2.44**×
NeuroD	0.9360	2.55	Neurological deficit increases risk **2.55**×
OSI	0.9522	2.59	Other significant injury increases risk **2.59**×

4.3 Stability Check

Similary with 3.5, we use cleaned dataframe which was made removing all rows that contained any missing values in the columns during the initial data cleaning process. Now, we will compare which variables are selected as the best model.

```
Best Model Performance (After Backward Elimination) for cleaned data:
number of Features:  16
Sensitivity: 73.33%
Specificity: 89.02%
Positive Predictive Value (PPV): 3.50%
Negative Predictive Value (NPV): 99.84%
Accuracy: 88.94%
```

We can see that using the clean data, the best model chosen with the same algorithm contains even more variables than with imputed data, with sensitivity and PPV much lower. This indicates that the cleaning process of eliminating all rows with missing values might not be the best option.

5 Discussion

The dataset was significantly larger than expected. While the sample size (number of patients) was reasonable given the extensive nature of this research—considering that ciTBI is a leading cause of death in children—the number of columns (125 variables) was unexpectedly high. One might assume that medical experts would have pre-selected the most relevant variables for predicting ciTBI, rather than including such a wide range of features. Additionally, more than 30% of patients had at least one missing (NaN) or inapplicable (92) value, making it challenging for non-experts to determine whether to impute or remove data based on limited judgment. This increased the complexity of data cleaning and decision-making, as different imputation strategies could lead to slightly different results. Regarding the three realms (data/reality, algorithms/models, and future data/reality), this lab primarily fit into data/reality and algorithms/models. The dataset represents real-world clinical data, but preprocessing decisions—such as imputing or removing missing values—introduce uncertainties and assumptions that may slightly distort its direct connection to reality. Because some variables were removed or modified, there is not a perfect one-to-one correspondence between the dataset and reality. However, given that Findings 1, 2, and 3 closely aligned with the prediction rule proposed in the original research paper, the data still captured meaningful patterns that reflect real-world clinical decision-making.

6 Conclusion

The PECARN prediction rule outlined in the original paper closely aligned with our findings, despite the challenges faced during the data cleaning process. While our preprocessing methods were not perfect, the fact that key variables—such as Altered Mental Status (AMS), Acting Normal, Vomiting, and Palpable Skull Fracture (SFxPalp)—emerged as strong predictors reinforces the reliability of the clinical rule. However, data cleaning and model selection are best handled by experts or engineers who work closely with medical professionals, rather than statistics students unfamiliar with the domain. Additionally, while our logistic regression model identified relevant predictors, further research and more rigorous data preprocessing are necessary. The sensitivity and specificity of our model were lower than those reported in the paper, highlighting the need for more refined feature selection, better handling of missing values, and potentially alternative modeling approaches to improve prediction accuracy.

7 Academic Honesty

I affirm that this work is my own and adheres to the academic integrity policies of this course. Any external sources, discussions, or collaborations have been properly cited. I have not engaged in unauthorized assistance or plagiarism in completing this assignment.

8 Collaborators

I used large language models (LLMs) to refine code for data visualization and improve the clarity of summaries. All analytical work, interpretations, and conclusions were my own. The LLM assistance was limited to debugging, code optimization, and enhancing readability.