PECARN TBI Data - Exploratory Data Analysis

1 Introduction

This exploratory data analysis (EDA) leverages PECARN TBI data – encompassing patient history, injury details, and symptoms – to predict the risk of clinically important traumatic brain injury (ciTBI) in children following blunt trauma. The overarching goal is to contribute to a reduction in unnecessary CT scans, which expose children to radiation, while simultaneously ensuring the timely and accurate diagnosis of serious injuries. Studying this data has the potential to generate prediction rules for ciTBI risk, improve diagnostic accuracy, and inform targeted interventions. A better understanding of this data would lead to enhanced patient care, reduced healthcare costs, improved clinical decision-making, and the support of more effective public health strategies.

The EDA aims to deeply understand the data, uncover relationships between variables (particularly those predicting ciTBI), and discover interesting patterns. The report will proceed with data cleaning to understand the variables and handle missing data, followed by the EDA to visualize the data and explore relationships. We will then highlight three key findings related to ciTBI, implement two classification models to predict ciTBI risk, and conclude with a summary of the findings and a discussion of their implications.

2 Data

This analysis will focus on data collected from a prospective observational cohort study conducted across 25 PECARN emergency departments. The study enrolled children under 18 years of age presenting within 24 hours of minor head trauma (Glasgow Coma Scale scores of 14-15).

The data consists of case report forms capturing mechanism of injury, clinical variables (history, symptoms, and physical examination findings), and basic demographic information such as age and race. The primary outcomes include whether the patient experienced a clinically important traumatic brain injury (ciTBI) and the presence of traumatic brain injury as confirmed by CT scan.

This data is directly relevant to the problem of reducing unnecessary CT scans in children with minor head trauma. By analyzing the relationships between the recorded variables and the ciTBI outcome, we aim to identify potential predictors of ciTBI and contribute to the development or refinement of clinical prediction rules that can accurately identify children at very low risk, for whom CT scans can be safely avoided, aligning with the study's original objective.

2.1 Data Collection

Data was generated through a prospective study where trained medical personnel recorded patient history, injury details, symptoms, and signs on standardized forms before imaging results. Quality was ensured through inter-rater reliability checks, double/triple data entry, and site monitoring.

In the dataset, variable measurements varied: many clinical variables were categorical (present/absent, injury type), age was continuous, and outcomes (ciTBI, TBI on CT) were binary. CT scans were ordered at clinician discretion. CiTBI was a priori defined based on severe outcomes like death, surgery, or prolonged intubation related to TBI on CT.

2.2 Data Cleaning

First, we addressed special codes within the dataset. A review of the metadata revealed encoding rules indicating that certain variables used the value '92' to represent "not applicable", specifically when a related parent variable was answered as "no" or was missing. For instance, the 'VomitNbr' variable used '92' if 'vomit' was answered as "no" or was missing. Consequently, we handled these cases by replacing specific '92' codes with either 0 or NaN, depending on the value of the corresponding parent columns and the meanings of the variables. This ensures accurate representation of the data.

Next, we examined data consistency. We identified inconsistencies in the Glasgow Coma Scale (GCS) scores, specifically that 'GCSTotal' should equal the sum of 'GCSMotor', 'GCSVerbal', and 'GCSEye'. We corrected any inconsistent 'GCSTotal' values by replacing them with the correct sum of the three component scores, ensuring data integrity.

Following this, we assessed the extent of missing data (NA values). Figure 1 displays the top 20 columns with the most NA values. 'Dizzy' exhibits a high proportion of missing values, primarily because this variable was not recorded for children younger than 2 years, as per the study protocol. Additionally, 'Ethnicity' and 'Race' also show a significant number of NAs. Given that these variables are reported by the physician rather than the patient/guardian, the observed missingness may indicate inconsistencies in data capture. To avoid potential information loss at this early stage, we have chosen to leave these NA values as is for the time being, deferring imputation or removal to later stages of the analysis should it become necessary. This approach allows us to preserve all available data for initial exploration.

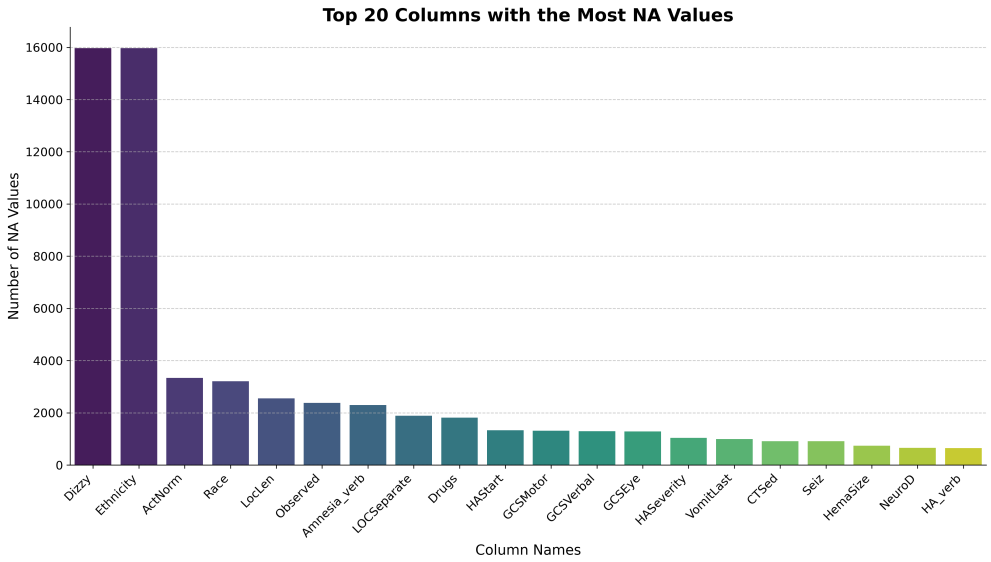

Figure 1: Top 20 columns with the most NA values

To examine potential outliers, we plotted the unique values of each variable (excluding identifiers such as 'patientNumber'; Figure 2). This visualization revealed points that appeared outside the main range of the data. However, upon closer inspection, these extreme values were identified as the special codes '91' and '92', which represent specific conditions such as "missing" or "not applicable" within the dataset. It's important to note that this plot was generated prior to replacing these special codes with more conventional missing value representations (NaN or 0, as described earlier). Therefore, these apparent outliers do not represent true data anomalies but rather the encoding scheme used in the original dataset.

In addition, we reviewed the column names to ensure they were reasonable, consistent, and accurately reflected the data contained within each column. This review confirmed that the existing column names aligned appropriately with the data description and metadata. Consequently, no changes to the column names were deemed necessary. This ensures clarity and ease of interpretation throughout the analysis.

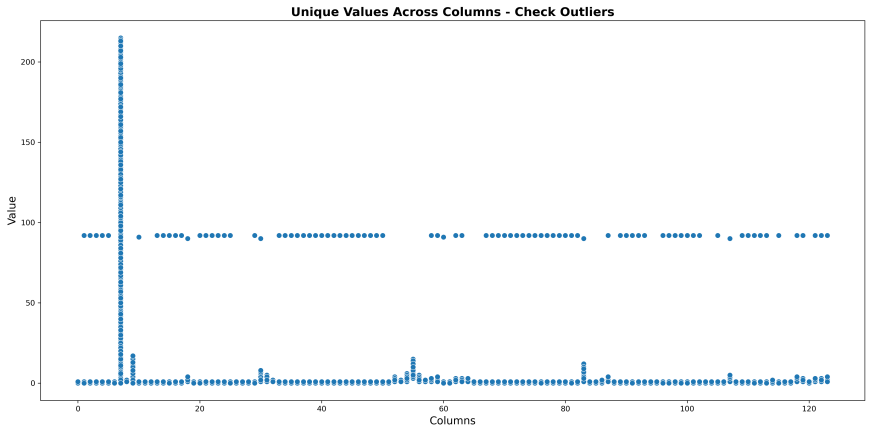

Figure 2: Unique values across columns - Check outliers

In summary, the following data cleaning steps were performed:

- Replaced special codes like '92' based on relationships between columns for more accurate information.
- Corrected inconsistent values, such as 'GCSTotal', based on related variables.
- Examined outliers, confirming they were special codes already addressed.
- Verified column names were clear and consistent.

2.3 Data Exploration

The dataset primarily comprises three categories of information: mechanism of injury, clinical variables (encompassing history, symptoms, and physical examination findings), and outcomes (ciTBI and TBI on CT). Our exploratory analysis begins by examining the distribution of injury mechanisms, followed by a focused investigation of the relationships between clinical variables and the defined outcomes. Finally, we will explore interesting patterns in outcomes such as factors influencing the decision to order a CT scan.

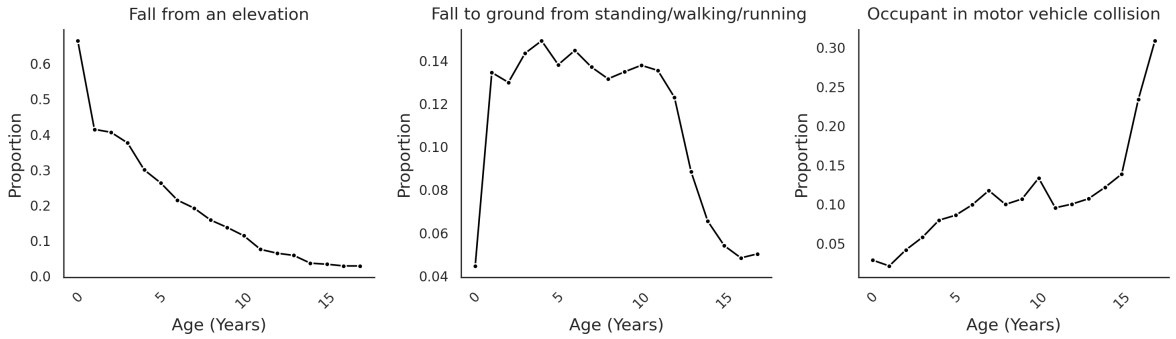

Figure 3: Injuries mechanisms vary by age

Figure 3 illustrates the relationship between age and the proportion of injuries resulting from different mechanisms. The proportion of injuries due to falls from an elevation decreases sharply with age, while falls to the ground from standing, walking, or running show a peak in early childhood followed by a decline. In contrast, the proportion of injuries sustained as an occupant in a motor vehicle collision generally increases with age, suggesting a shift in the dominant mechanisms of injury as children grow older. This aligns with our intuition that at younger ages, falls are a more common cause of head trauma, while as children age and become more mobile and independent, their risk of injuries related to motor vehicle collisions increases.

Figure 4 presents two pie charts illustrating the proportion of various symptoms observed in patients with

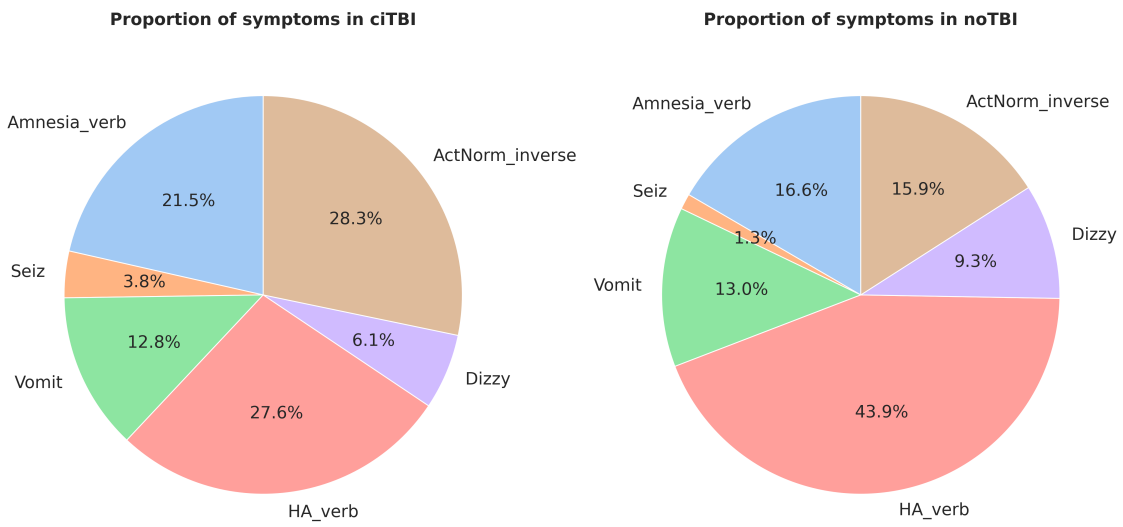

Figure 4: Proportion of Symptoms group by ciTBI

ciTBI (left) and patients without ciTBI (right). In the ciTBI group, "ActNorm_inverse" (abnormal behavior) and "HA_verb" (headache) have the highest proportions at 28.3% and 27.6%, respectively, followed by "Amnesia_verb" (amnesia) at 21.5%. In contrast, in the noTBI group, "HA_verb" dominates with 43.9%, with "ActNorm_inverse" and "Amnesi_verb" having considerably lower proportions (15.9% and 16.6%, respectively). Seizures ("Seiz") are a relatively rare symptom in both groups but appear more frequently in the ciTBI group (3.8%) compared to the noTBI group (1.3%). Dizziness ("Dizzy") is a higher proportion in the noTBI group compared to the ciTBI group. This may indicate that although many people experience headache after injury, the headache might not be as severe or indicative of significant injury in the absence of other symptoms (we will explore that later in Finding1). These differences in symptom profiles may provide valuable insights into differentiating patients at higher risk of ciTBI.

The correlation heatmap(Figure 5) provides a valuable overview of the relationships between variables, highlighting potential predictors of TBI on CT. The very strong positive correlation (0.77) between "PosCT" (positive CT scan) and "PosIntFinal" (positive interpretation, final) validates the consistency of outcome measures, as expected.

Among clinical variables, "ActNorm_inverse" (abnormal behavior) shows a moderate positive correlation with "AMS" (Altered Mental Status) (0.58), as expected. "AMS" shows a negative correlation with "GCS Total" (-0.26), suggesting that lower GCS scores are associated with altered behavior. The negative correlation between GCS Total and PosIntFinal/PosCT(-0.52/-0.38) further suggests that GCS is a good indicator for positive finding.

In general, individual symptoms such as "Amnesia_verb," "Seiz," "Vomit," "HA_verb," and "Dizzy" show weak correlations with each other and with the CT outcomes. This suggests that these individual symptoms alone may not be strong predictors of TBI on CT and emphasizes the need to consider symptom combinations and other factors in risk assessment.

The bar chart (Figure 6) compares the proportion of different indications for ordering CT scans in all patients who received a CT ("all ordered") versus those specifically found to have ciTBI ("ciTBI"). Altered mental status (IndAMS) and Mechanism of Injury(IndMech) were the leading indications in both groups. However, compared to the "all ordered" group, Altered mental status (IndAMS) appears as a weaker indication in the ciTBI group as it takes a smaller portion. Loss of consciousness (IndLOC) also shows a lower proportion in the ciTBI group, while some indications, like clinical Significant Fracture (IndClinSFx) or Seizure(IndSeiz), only take small portions. These findings highlight the relative importance of various clinical factors in CT ordering decisions and their association with clinically important TBI, suggesting that while conditions like

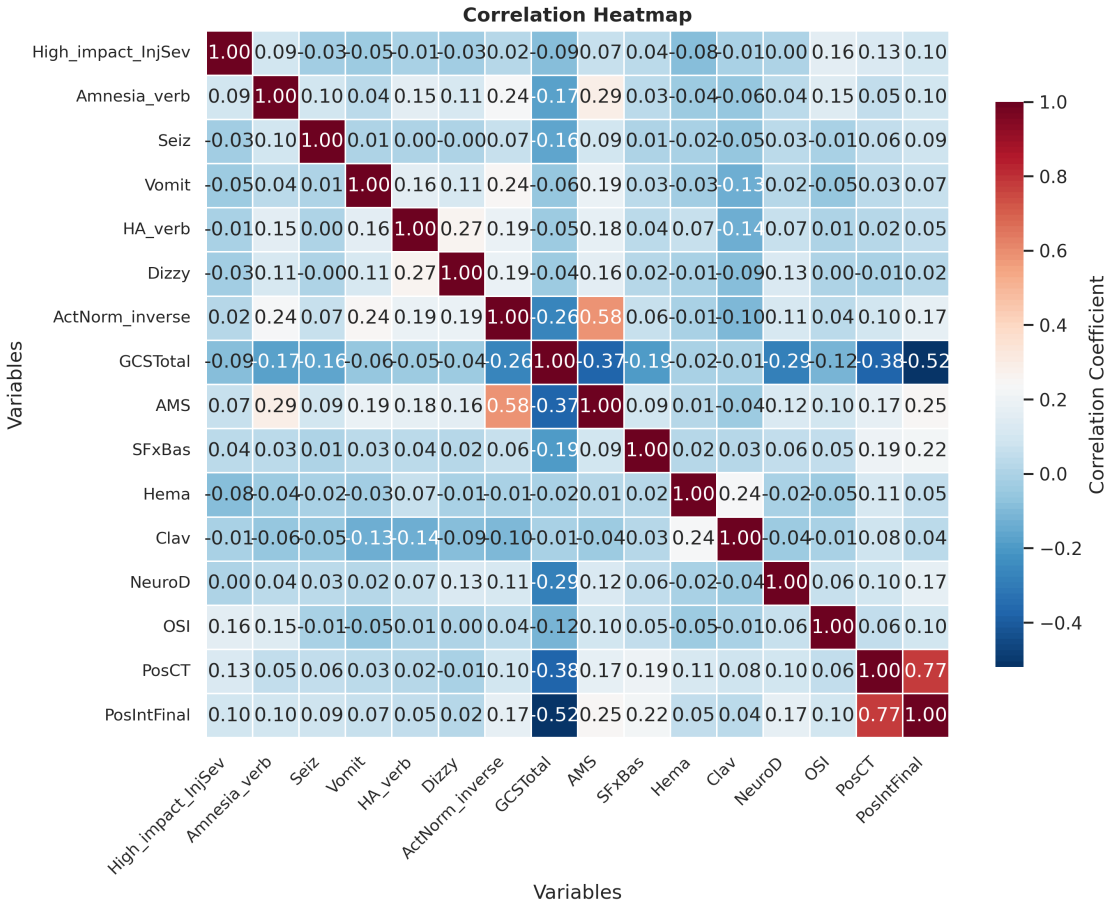

Figure 5: Correlation heatmap of important variables

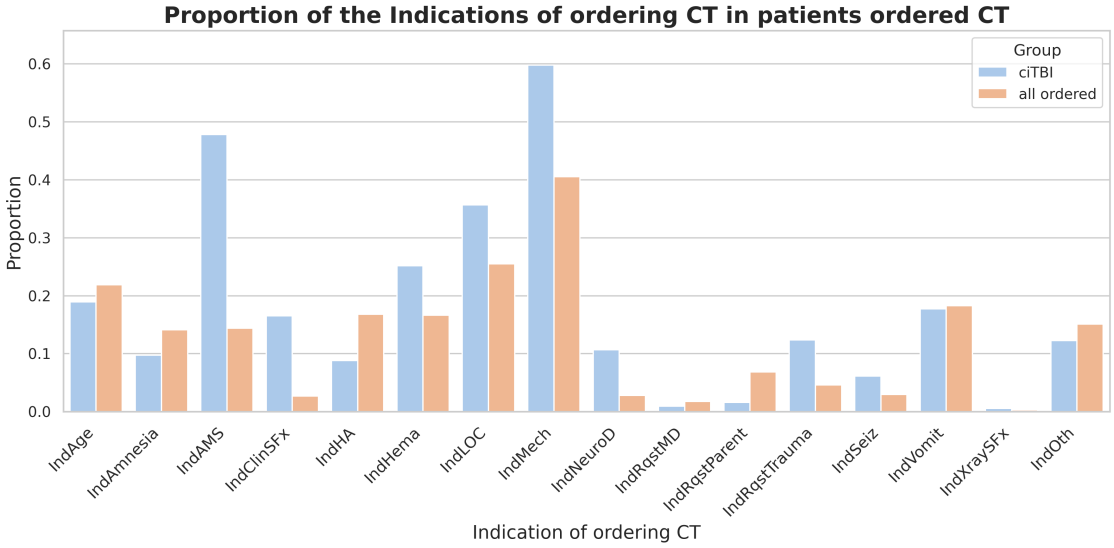

Figure 6: Proportion of the indications of ordering CT

altered mental status and concerning injury mechanisms drive overall CT utilization, they may not be as

strongly predictive of actually having a ciTBI.

3 Findings

3.1 First finding

Expanding on the previous analysis focusing on individual symptoms, our first finding explores the role of symptom combinations through an examination of symptom counts in ciTBI and noTBI patients.

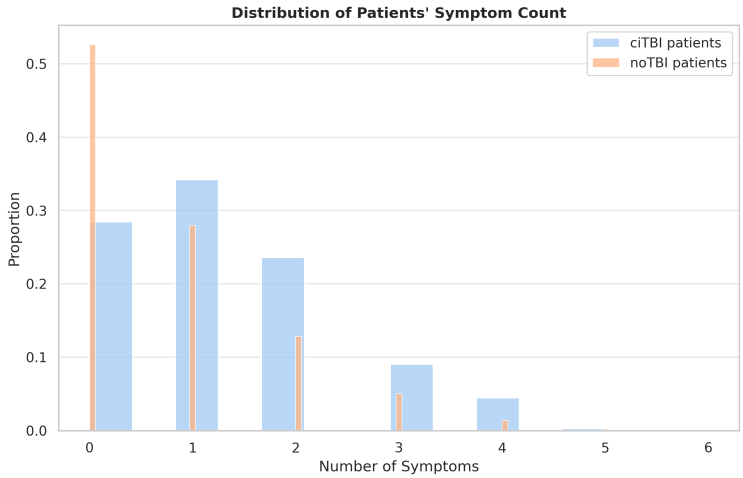

Figure 7: Distribution of Patients' Symptoms Count

Figure 7 illustrates the distribution of symptom counts in the two groups. A larger proportion of noTBI patients present with zero symptoms, whereas the proportion of ciTBI patients generally increases with symptom count up to a certain point, before declining. There's some overlap between the two groups, suggesting that although a higher symptom count correlates with an increased likelihood of ciTBI, a subset of ciTBI patients may present with few, or even no, recorded symptoms. Therefore, symptom count serves as a useful, but not definitive, factor in distinguishing between ciTBI and noTBI patients, highlighting the need to consider other clinical factors in addition to the total number of symptoms.

3.2 Second finding

Figure 8 provides a comparative analysis of positive exam result proportions, stratified by age group (under 2 years vs. over 2 years), across all observations and specifically within the ciTBI cohort. This allows us to assess how common certain findings are in the broader patient population versus specifically in confirmed cases of clinically important TBI.

Across all observations, the younger children (under 2 years) are more prone to having hematomas ("Hema"). The older age group has slightly higher proportions of AMS. However, the bottom chart, focusing specifically on confirmed ciTBI cases, shows a notable increase in the importance (proportion) of both "GCSGroup_01" and "AMS", especially as the age increases, which is a good indication to find out what results to consider.

In summary, by contrasting the overall trends with the ciTBI-specific trends, the findings imply the signals have a shift and become more important than the overall. While the proportions of positive exam findings are elevated in the ciTBI group, certain clinical variables may be more indicative of ciTBI in specific age groups.

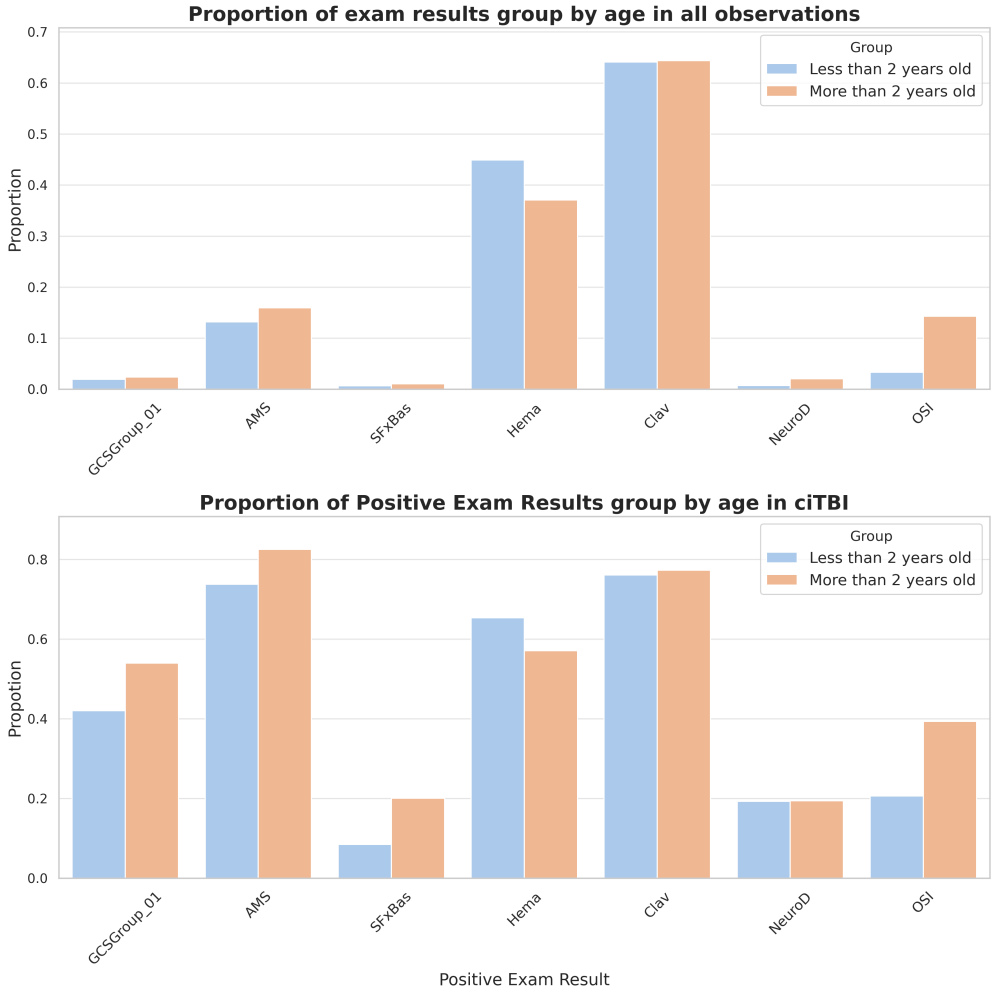

Figure 8: Proportion of Positive Exam Results

3.3 Third finding

The third finding, while perhaps less directly related to the prediction of ciTBI, represents an interesting observation regarding the distribution of specific CT findings across different pediatric age groups.

Figure 9 presents age-related variations in the types of CT findings observed in patients with ciTBI, offering potential insights into how mechanisms of injury and brain vulnerability change with age. Cerebral contusions, characterized by bruising of brain tissue, show fluctuating proportions across age, with peaks around 7.5, 10, and a general increase towards older ages. This trend could be related to older children participating in higher-risk activities that may lead to direct head impacts. Cerebral edema, representing brain swelling often triggered by trauma, peaks in proportion around age 12. Diastasis of the skull, involving a separation of cranial sutures, is more prevalent among younger children (0-5 years). This is plausibly explained by the incomplete fusion of skull sutures at that age, increasing their susceptibility to separation under traumatic forces. The proportion of extra-axial hematoma, which is bleeding outside the brain tissue itself, decreases after age 5.

Overall, the observed fluctuations in these CT findings suggest that the mechanisms and patterns of brain injury differ across pediatric age groups. A potential example of such a difference involves how younger children are more likely to injure their sutures due to the nature of the skull. However, whether these interpretations truly make sense, are supported by existing research, and accurately reflect the underlying

physiology requires validation with more specialized domain knowledge. It is crucial to remember that this data only reflects confirmed ciTBI cases, so the reported age-specific patterns might not generalize to all head injuries.

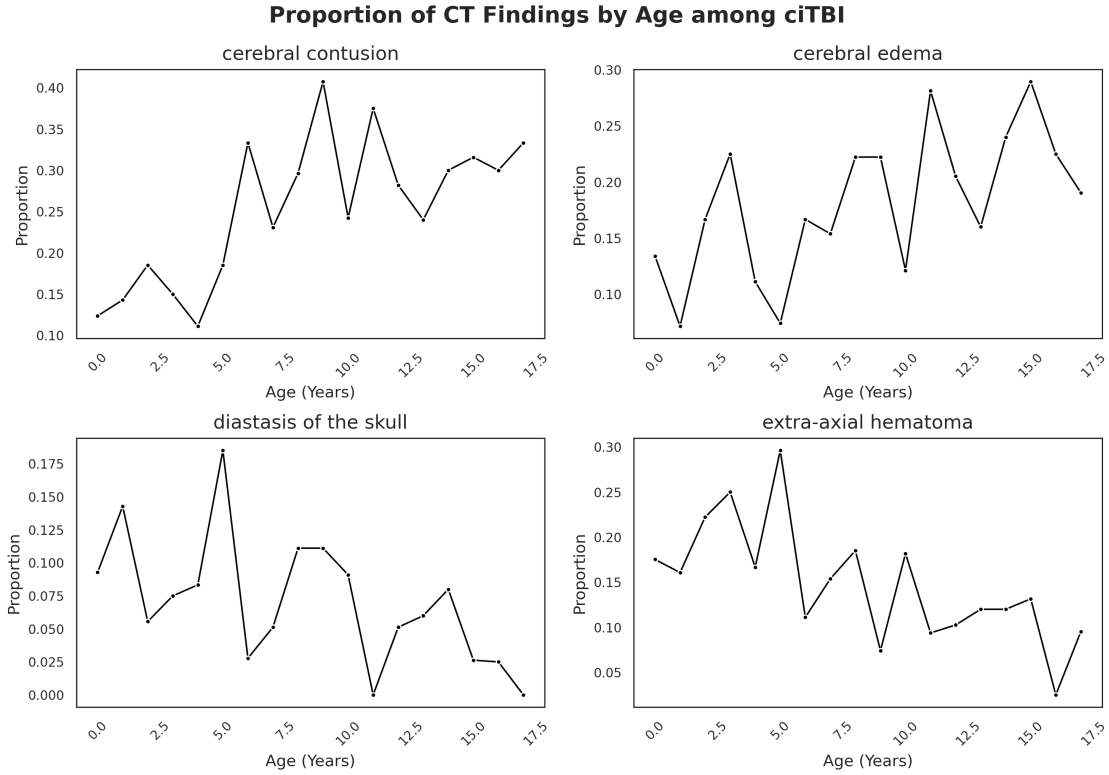

Figure 9: Proportion of CT Findings by Age among ciTBI

3.4 Reality Check

A reality check for our data involves comparing our findings with established research on exam findings in pediatric TBI, stratified by age. Medical literature suggests that findings such as AMS and signs of a concussion are typically found more often in those presenting TBI than in a normal setting, and GCS score has an inverse relationship with results from TBI. Our data, presented in Figure 8 and Figure 5, suggests the same trend which is very promising. This consistency suggests that our data is reliable in reflecting real-world patterns for this population. However, it's important to acknowledge that the proportions alone do not guarantee predictive power, and factors such as sample size biases will need more tests. Despite this, the presence of expected and confirmed real world effects is shown, giving the cleaned data a pass for now.

3.5 Stability Check

To evaluate stability, I utilized a dataset without performing inconsistency checks. Figure 10 reveals that omitting these checks resulted in minimal differences in finding2.

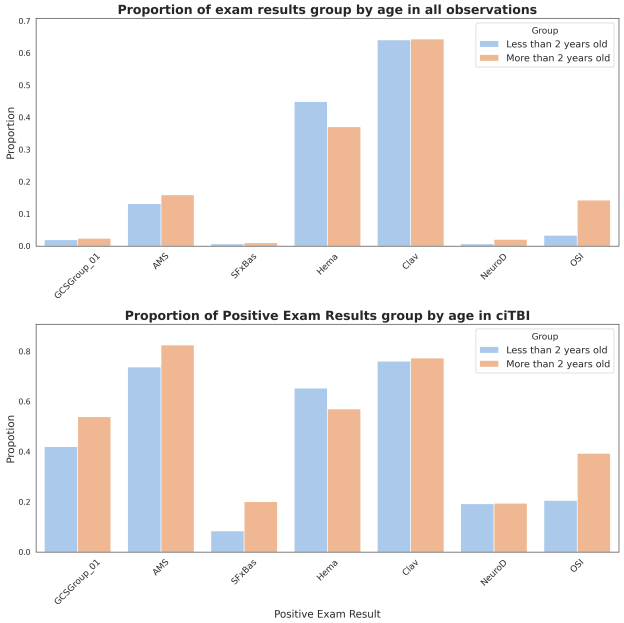

Figure 10: Perturbed version of finding2

4 Modeling

4.1 Implementation

To predict the need for a CT scan in pediatric head trauma cases, we implemented and compared two classification models: Logistic Regression and Random Forest. These choices were guided by the imbalanced nature of our dataset, where ciTBI cases were significantly less frequent, and by the paramount clinical goal of minimizing false negatives (i.e., avoiding missed diagnoses of ciTBI).

Logistic Regression was selected for its inherent interpretability and computational efficiency, allowing us to understand the individual contributions of predictors in a straightforward manner. Random Forest, a more complex and non-linear ensemble method, was chosen to potentially capture more intricate relationships within the data and achieve higher overall predictive accuracy.

To address the class imbalance, we set the class_weight hyperparameter to "balanced" for both models, effectively penalizing misclassification of ciTBI cases more heavily. Additionally, recognizing the greater clinical risk associated with false negatives, we deliberately lowered the default prediction threshold from 0.5 to 0.3. This adjustment aimed to increase the models' sensitivity, reducing the likelihood of missing a true ciTBI case, albeit with a potential increase in the number of unnecessary CT scans.

4.2 Interpretability

Logistic Regression offers a relatively interpretable model, where coefficients associated with each predictor show the direction and magnitude of their influence on the probability of needing a CT scan. While Random Forest is less inherently interpretable, its feature importance scores provide insight into the relative influence of each variable in making predictions. These scores allow us to understand which factors are most important in the model's decision-making process, even if we cannot see the exact decision rules.

4.3 Stability

Figure 11 indicates that the system exhibits little sensitivity to perturbation.

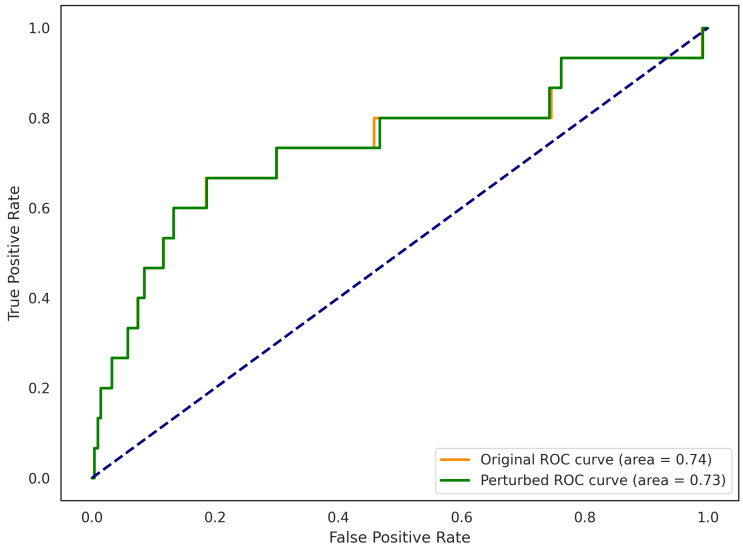

Figure 11: Perturbed Version vs Original

5 Discussion

A significant challenge arose from the imbalanced dataset, specifically the scarcity of ciTBI cases compared to non-ciTBI cases. This imbalance limited our ability to robustly identify patterns and restricted the complexity of models we could reliably train without overfitting. For instance, as observed in our third finding, the fluctuating patterns in certain plots might be attributable to the small sample sizes within specific age categories, hindering our ability to discern clear trends.

Our analysis spans three interconnected realms: the realm of data and reality, the algorithmic/modeling realm, and the realm of future data and reality. The initial data collection and subsequent cleaning stages represented our attempt to capture a portion of the real-world clinical experience of pediatric head trauma cases. However, this process inherently involves simplification and abstraction, with clinical nuances and individual patient characteristics being compressed into a finite set of variables. The model selection and implementation phase was entirely within the algorithmic realm. These models, like Logistic Regression and Random Forest, attempted to distill patterns and relationships from the data, thereby creating a simplified mathematical representation of the underlying clinical processes. Ultimately, the goal is to bridge the algorithmic realm with future reality by applying our models to new, unseen patients to inform clinical decisions and potentially reduce unnecessary CT scans.

It is crucial to recognize that there is no perfect one-to-one correspondence between our data and the complex clinical reality it seeks to represent. The data is, by necessity, a simplified and imperfect representation, influenced by factors such as measurement error, selection bias, and missing information. Similarly, data visualizations, while powerful tools for exploring and communicating insights, are also abstractions of reality. The choice of chart type, color scales, and other design elements can influence how patterns are perceived and may emphasize certain aspects of the data while downplaying others. Therefore, it is essential to approach both the data and its visualizations with a critical mindset, acknowledging their inherent limitations and potential for bias.

6 Conclusion

In conclusion, our analysis of the PECARN dataset provides several insights into the complex challenge of identifying children at risk for ciTBI following minor head trauma, all with the overarching goal of refining clinical decision-making regarding CT scans.

First, we observed that while a higher symptom count generally correlates with an increased likelihood of

ciTBI, a subset of ciTBI patients may present with few or even no recorded symptoms. This underscores the need to consider multiple clinical factors beyond the total number of symptoms when assessing risk.

Second, our age-stratified analysis of positive exam findings revealed important age-related differences in the predictors of ciTBI, while these patterns are shifted and changed across the age, but two key things have become more important: AMS and GCS score, this points to the important of neuro status, and suggest neuro results are more correlated to TBI outcomes.

Finally, our exploration of age-related variations in CT findings among ciTBI patients, while perhaps less directly predictive, offers a window into how the patterns and mechanisms of injury might differ across the pediatric age spectrum.

Overall, our work on the PECARN data reinforces the complexity of pediatric TBI and supports the need for nuanced, age-appropriate clinical decision rules that incorporate a range of clinical factors beyond any single symptom or sign. Future research, incorporating larger datasets and more advanced modeling techniques, is warranted to refine these prediction rules and ultimately improve the safety and well-being of children presenting with minor head trauma.

7 Academic honesty statement

I guarantee the following is true: I designed and carried out all the data analysis processes in this report myself. I also wrote all the text and put together all the figures myself. I've made sure to document all the steps, so the results can be fully reproduced. Whenever I used someone else's work, I cited them.

8 Collaborators

The main contents of this report are all my own thoughts and all the contents are completed by myself without any collaborators.

9 Bibliography

1. Kuppermann N, Holmes JF, Dayan PS, Hoyle JD Jr, Atabaki SM, Holubkov R, Nadel FM, Monroe D, Stanley RM, Borgialli DA, Badawy MK, Schunk JE, Quayle KS, Mahajan P, Lichenstein R, Lillis KA, Tunik MG, Jacobs ES, Callahan JM, Gorelick MH, Glass TF, Lee LK, Bachman MC, Cooper A, Powell EC, Gerardi MJ, Melville KA, Muizelaar JP, Wisner DH, Zuspan SJ, Dean JM, Wootton-Gorges SL; Pediatric Emergency Care Applied Research Network (PECARN). Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. Lancet. 2009 Oct 3;374(9696):1160-70. doi: 10.1016/S0140-6736(09)61558-0. Epub 2009 Sep 14. Erratum in: Lancet. 2014 Jan 25;383(9914):308. PMID: 19758692.

2. Nevo A, Cheney SM, Callegari M, Moore JP, Stern KL, Zell MA, Abdul-Muhsin H, Humphreys MR. Median lobe vs. complete gland holmium laser enucleation of the prostate: A propensity score matching. Can Urol Assoc J. 2023 Jan;17(1):E39-E43. doi: 10.5489/cuaj.7890. PMID: 36121884; PMCID: PMC9872827.