Lab 1 - PECARN TBI Data, Stat 215A, Fall 2024

1 Introduction

Traumatic Brain Injuries (TBI) remain one of the leading causes of death and severe disabilities in children under 18 years of age [1]. A key challenge in clinical settings is the timely identification of high-risk patients who may require immediate intervention, such as a computed tomography (CT) scan. However, the use of CT scans presents a trade-off, as the ionizing radiation can pose long-term health risks. To address this, clinical decision rules (CDRs) have been developed to identify children at very low risk of clinically important TBIs (ciTBIs), which allows clinicians to reduce unnecessary CT scans while maintaining patient safety [2, 3]. The dataset used in this lab, drawn from the Pediatric Emergency Care Applied Research Network (PECARN), is instrumental in validating and improving these CDRs. My goal is to perform an exploratory data analysis (EDA) on the PECARN dataset to identify patterns that can assist in better clinical decision-making, ultimately reducing the risks associated with both under- and over-diagnosis of TBI in pediatric patients.

This report will focus on two key aspects: data cleaning and exploratory data analysis. First, I will examine the structure of the dataset and clean up any inconsistencies or errors. Following this, I will explore the dataset to identify meaningful patterns that could contribute to the improvement of CDRs in pediatric emergency care.

2 Data

The PECARN dataset used in this analysis is derived from a prospective observational cohort study, conducted across 25 emergency departments, for identifying children at low risk for ciTBI. The study enrolled children younger than 18 years who presented within 24 hours of a blunt head trauma. The dataset includes a variety of patient-level features, including injury mechanisms, clinical assessments, and outcomes. Variables such as age, gender, and race were recorded along with clinical symptoms such as loss of consciousness, headache, and vomiting. Additionally, the data captures outcomes related to CT scans and hospitalizations, which are used to define whether a patient experienced a clinically important TBI.

2.1 Data Collection

The PECARN dataset includes a wide variety of features that provide detailed information on both patient demographics and clinical characteristics. The data were collected by trained site investigators, who completed standardized data forms for each patient. These forms captured detailed information regarding the mechanism of injury, clinical symptoms, and any treatments or interventions. Patient follow-ups were conducted via phone calls to assess delayed outcomes or missed diagnoses. This rigorous data collection process ensures the accuracy and completeness of the dataset for further analysis.

2.2 Data Cleaning

In order to ensure the integrity of the analysis and mitigate the effects of incomplete or inconsistent data, a systematic data cleaning process was employed. The Pediatric Emergency Care Applied Research Network (PECARN) dataset comprises 125 variables, several of which contain missing or invalid entries. Properly cleaning the data is essential for obtaining reliable and accurate conclusions.

Before initiating any data cleaning procedures, I conducted an initial assessment of the missing values in the dataset. This preliminary analysis allowed for a better understanding of the extent and distribution of missing data. Figure 1 illustrates the raw dataset prior to any cleaning or preprocessing steps, highlighting the top 30 variables where data is missing.

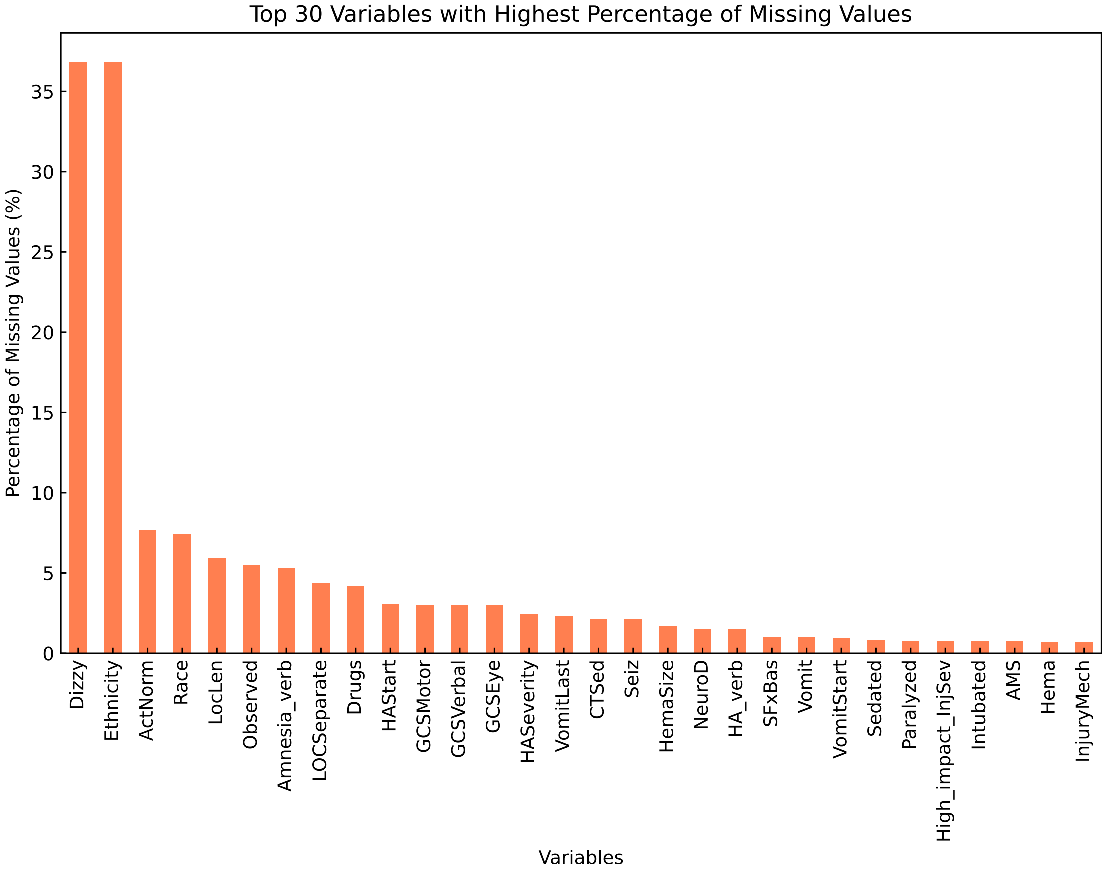

Figure 1: Missing value plot

From the initial analysis of the plot, it was observed that the variables `Dizzy` and `Ethnicity` have more than 30% of their values missing. This significant amount of missing data in these variables suggests that careful consideration is required before proceeding with any imputation or removal strategies.

The following steps outline the cleaning procedure applied to the dataset:

- **Replacing Missing Value Placeholders**: The dataset contained various placeholders representing missing values, such as `n/a`, `na`, `--`, `-`, and numerical codes like `92` indicating "not applicable." These placeholders were replaced with `NaN` (Not a Number) to standardize the missing data across the dataset. Additionally, extreme values such as `999`, `-999`, and out-of-range values like `inf` and `-inf` were also replaced by `NaN`. This step ensured consistency for the subsequent handling of missing values.
- **Removing Rows with Excessive Missing Data**: Rows with excessive missing data were removed based on a pre-defined threshold. Any row containing more than the allowed number of missing values was discarded. This strategy prevented incomplete rows from adversely affecting the analysis, ensuring that the remaining data was sufficiently complete for further analysis.
- **Categorizing Variables**: The dataset comprised both numerical and categorical variables. Numerical variables included `AgeinYears`, `AgeInMonth`, `GCSTotal`, and `PatNum`. Other columns, such as clinical symptoms, demographic information, and injury mechanisms, were categorized as categorical variables.
- **Imputation of Missing Values**: Missing values in both numerical and categorical columns were handled using different approaches:
  - **Numerical Variables**: Missing values in numerical columns were imputed using the mean of

each respective column. This simple method helps preserve the distribution of the data while filling in the missing values without introducing biases.

- **Categorical Variables**: For categorical columns, missing values were imputed using the mode (i.e., the most frequent value) for each respective variable. In cases where no mode could be identified, the missing values were replaced with a default value of `unknown`, ensuring that the dataset remained as complete as possible without introducing unnecessary bias.

After following these steps, the dataset was cleaned and prepared for further exploratory analysis. This methodical cleaning process minimized the risk of bias from missing values or outliers, thereby ensuring the validity and reliability of subsequent findings.

2.3 Data Exploration

In the initial phase of data exploration, two key plots were analyzed to understand the distribution of age and injury mechanisms within the dataset. Figure 2 shows the age distribution, ranging from 0 to 18 years.

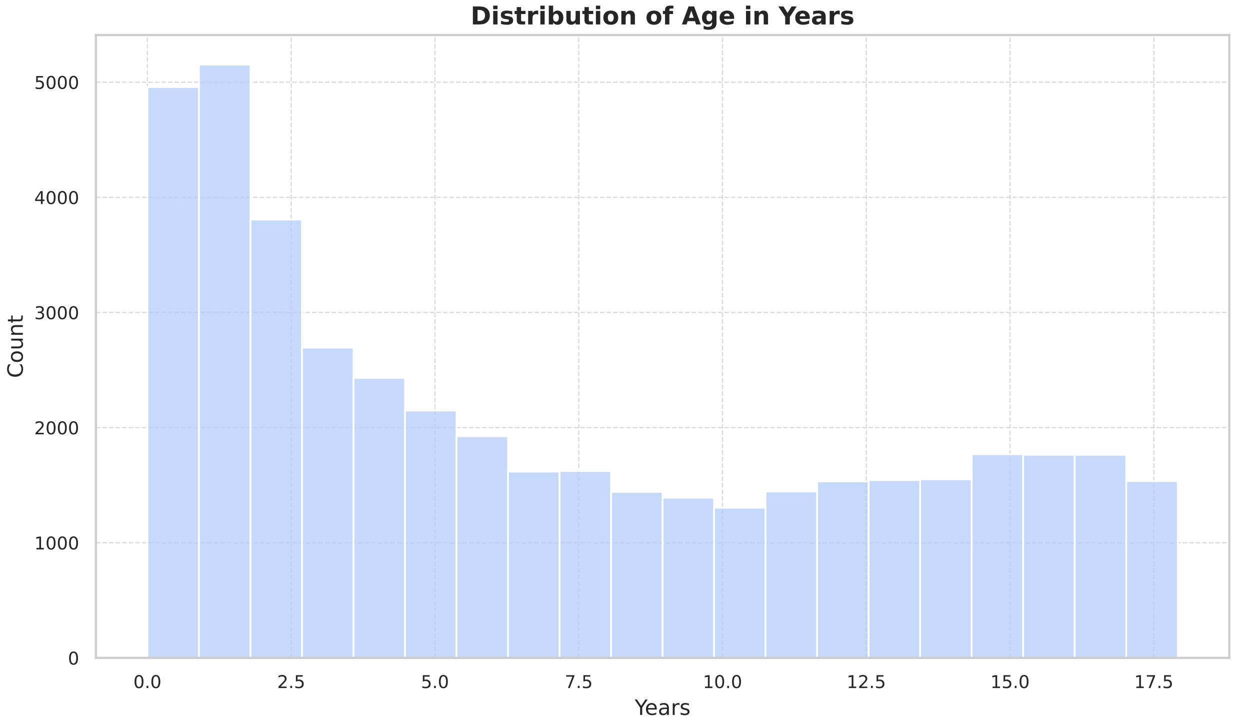

Figure 2: Distribution of Age (in years)

The majority of individuals fall into the younger age groups, with a noticeable peak in the early years and a gradual decline in frequency as age increases. This suggests that the dataset is heavily weighted toward younger individuals, which may have implications for further analysis, particularly if age-related injury severity or outcomes are of interest. The relatively low representation of older children or teenagers implies that conclusions drawn from this data may be more applicable to younger demographics.

Figure 3 reveals the most common causes of injury in the dataset.

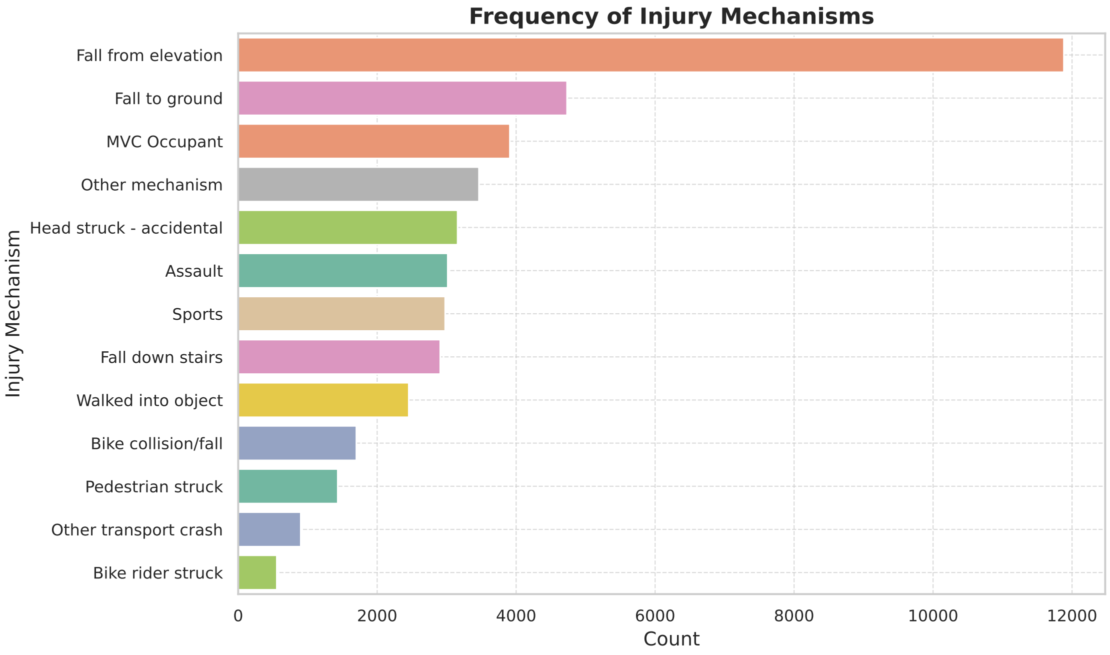

Figure 3: Frequency of Injury Mechanisms

The most frequent mechanisms include *Fall from elevation* and *Fall to ground*, followed closely by *MVC Occupant (Motor Vehicle Crash)*. These injury types significantly outnumber other mechanisms, such as *Bike rider struck* or *Other transport crash*. This concentration suggests that falls are a predominant source of injury, highlighting an area where preventive measures may be most effective. Understanding the dominant mechanisms provides critical context for further analysis, such as evaluating how these mechanisms correlate with injury severity or medical interventions.

Figure 4 illustrates the distribution of hospitalization due to head injuries and deaths resulting from traumatic brain injuries (TBI).

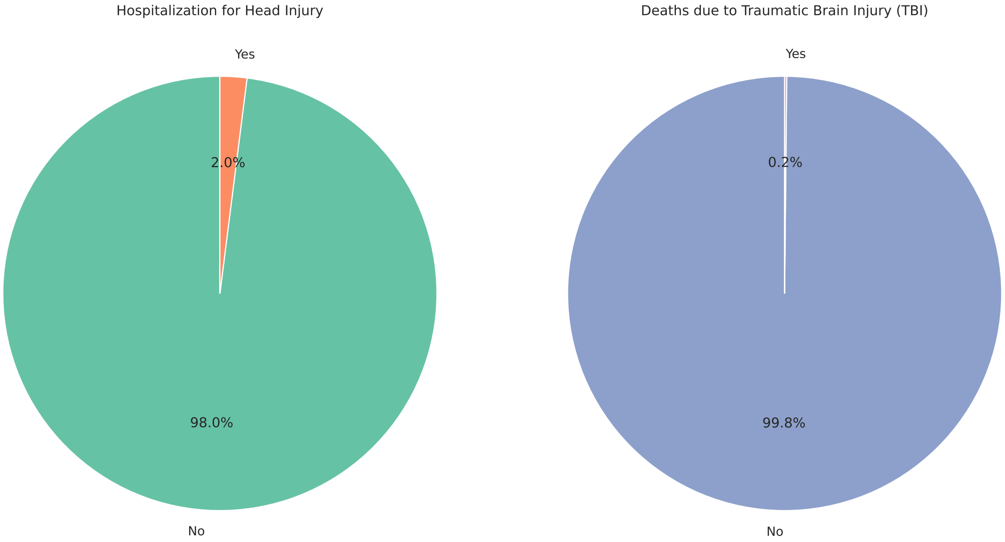

Figure 4: Distribution of Hospitalization for Head Injury and Deaths Due to Traumatic Brain Injury

The left pie chart represents the proportion of cases where patients were hospitalized for head injuries, while the right pie chart shows the proportion of deaths attributed to TBI. From the left chart, we observe that the vast majority of patients, 98.0%, did not require hospitalization for head injuries, with only 2.0% of the cases involving hospitalization. Similarly, the right chart demonstrates that 99.8% of patients did not experience a fatal TBI, while only a small fraction (0.2%) of cases resulted in death due to TBI. These insights suggest that both hospitalization and death are relatively rare outcomes in this dataset, which should be considered when interpreting the results.

3 Findings

In this section, the key findings from the exploratory data analysis of the dataset is presented. The analysis reveals important patterns related to injury mechanisms, severity levels, and outcomes, which can be used to inform future safety measures and preventive strategies.

3.1 First Finding

Figure 5 illustrates the distribution of injury mechanisms across three levels of injury severity: Low, Moderate, and High.

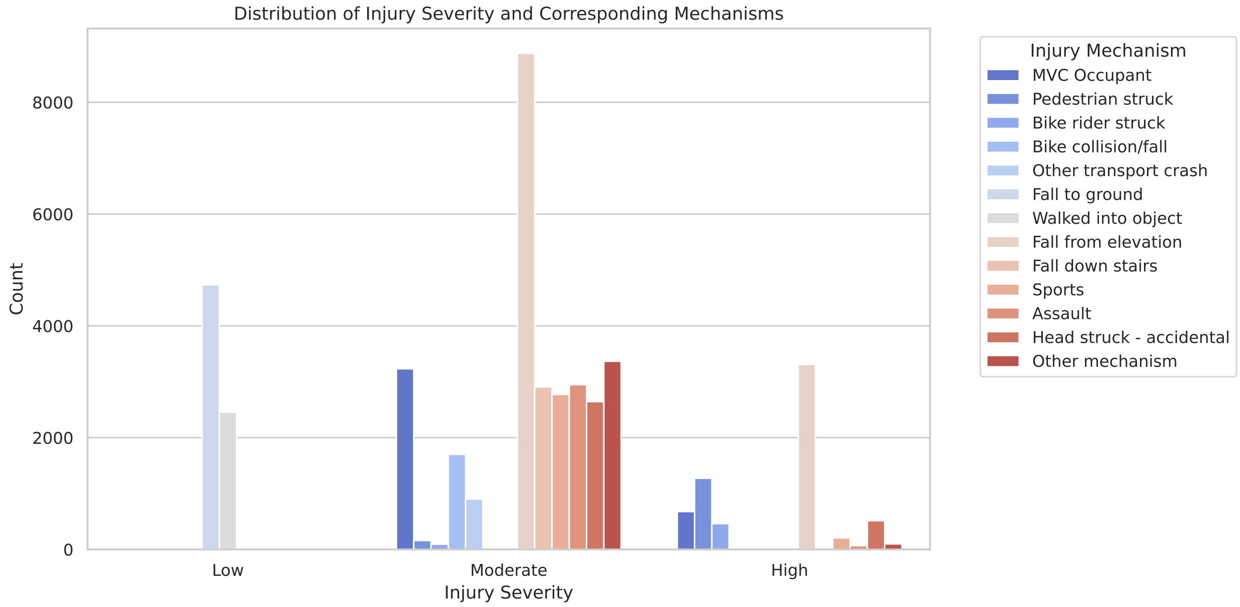

Figure 5: Distribution of Injury Severity and Corresponding Mechanisms

Several interesting trends are evident. Motor Vehicle Crash (MVC) Occupants account for a significant portion of injuries in both the Low and Moderate severity categories, suggesting that while motor vehicle incidents are frequent, they often do not result in high-severity injuries. Falls to the ground are the most common mechanism in the Moderate severity category, exhibiting the highest count of injuries overall. Interestingly, there are fewer cases of high-severity injuries associated with falls, indicating that this mechanism frequently results in moderate but not critical injuries. Additionally, accidental head strikes contribute significantly to moderate severity injuries, highlighting the need for improved safety measures to prevent accidental head trauma, particularly in environments where such incidents are common. Other mechanisms, such as bike collisions/falls and assaults, are less frequent across all severity levels, with very few cases resulting in high-severity injuries. These findings underscore the importance of targeting specific injury mechanisms, such as motor vehicle crashes and falls, with interventions aimed at reducing their frequency and severity. The distribution of injuries across severity levels provides valuable insights for prioritizing safety strategies.

3.2 Second Finding

Figure 6 shows the predominant causes of hospitalization. The bars show the count of "Yes" responses for each cause, ordered from highest to lowest, providing insight into the leading factors contributing to hospitalization.

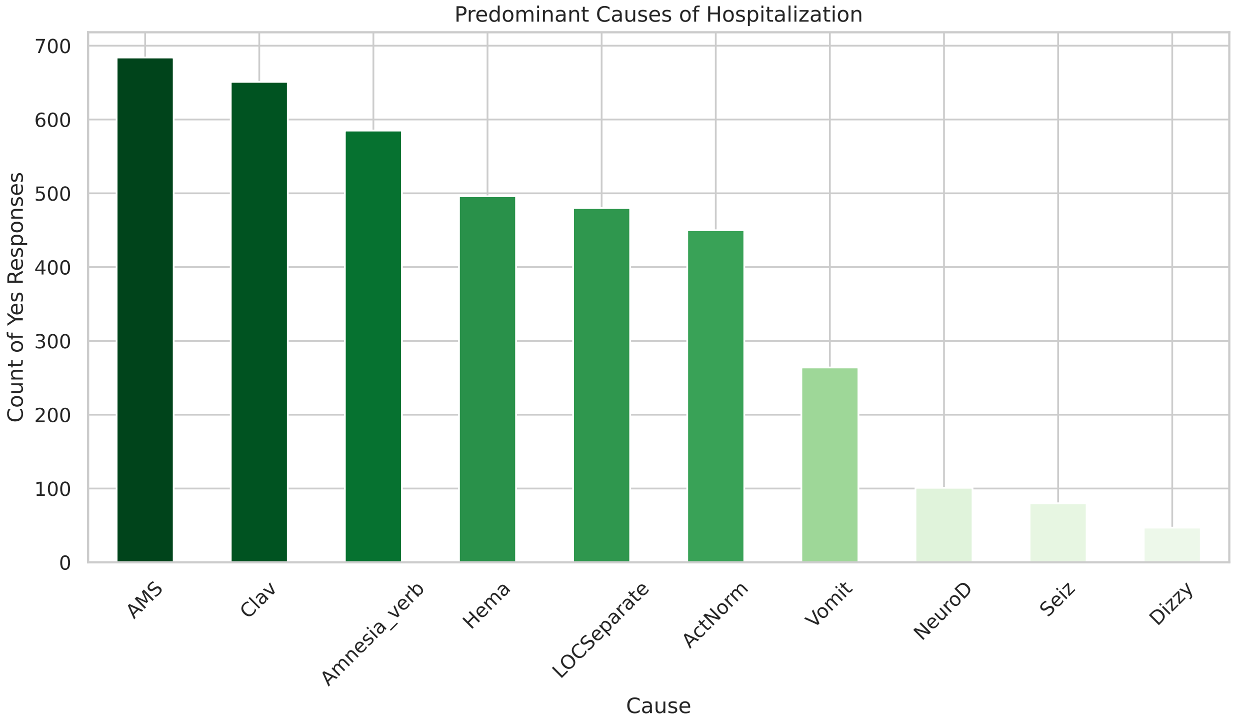

Figure 6: Predominant Causes of Hospitalization

The key finding from the analysis is that *AMS (Altered Mental Status)* is the most frequent cause of hospitalization, with over 700 cases. This suggests that a significant portion of patients admitted for hospitalization exhibited signs of mental disorientation or confusion. The second most common cause is *Clavicle Injuries (Clav)*, with nearly the same frequency as *AMS*, indicating that clavicle fractures or related injuries are a major reason for hospital admission.

Amnesia Verbally Reported (Amnesia_verb) follows closely, highlighting that memory loss or inability to recall events is another prevalent cause for hospitalization. *Hematoma (Hema)* and *Loss of Consciousness (LOCSeparate)* also feature prominently, suggesting that head trauma and its associated symptoms are significant factors in hospital admissions. *Activity Normalization (ActNorm)*, representing the patient's return to normal activities, is slightly lower but still a frequent reason for hospitalization. It might indicate a subset of patients needing observation or care before they can return to regular activities. *Vomiting (Vomit)*, although present in fewer cases, remains a notable symptom, often associated with trauma or other injuries.

Lastly, *Neurological Deficit (NeuroD)*, *Seizures (Seiz)*, and *Dizziness (Dizzy)* are the least frequent causes of hospitalization in this dataset, yet they still represent critical conditions requiring medical attention.

This analysis reveals that *AMS, Clavicle Injuries, and Amnesia* are the predominant causes for hospitalization among the cases studied. Medical professionals might prioritize monitoring and addressing these conditions, as they represent the majority of cases requiring hospital care.

3.3 Third Finding

Figure 7 illustrates the rate of clinically important traumatic brain injuries (TBI) across three age groups (0-5, 5-12, and 12-18 years) for both males (Gender 1) and females (Gender 2).

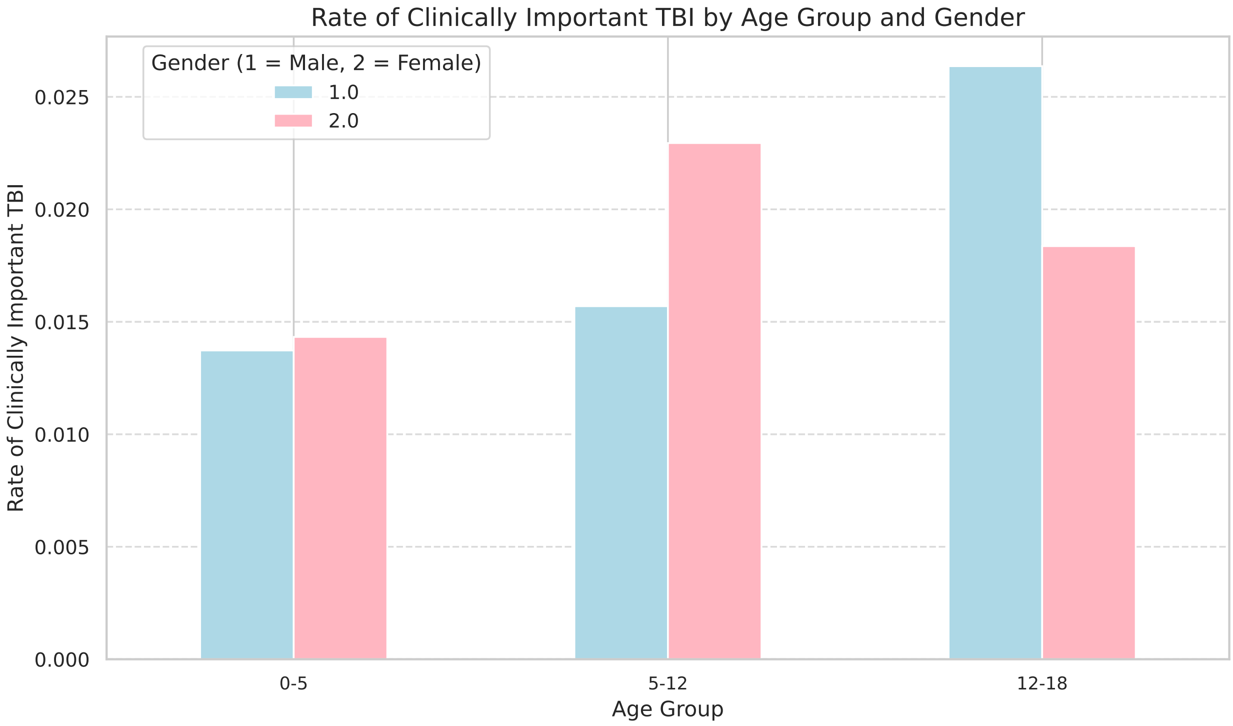

Figure 7: Predominant Causes of Hospitalization

The vertical axis represents the rate of clinically important TBIs, while the horizontal axis displays the different age groups. Notably, the rate of TBI shows variations across age and gender. In the 0-5 age group, males and females exhibit similar rates, with females having a slightly higher rate. Both genders in this age group demonstrate relatively low TBI rates compared to older age groups, with values around 0.01-0.015. However, in the 5-12 age group, females experience a higher rate of TBI, nearing 0.02, compared to males, whose rate remains below 0.015. The most significant gender difference is observed in the 12-18 age group, where males display the highest TBI rate across all categories, exceeding 0.025, while females maintain a lower but still considerable rate close to 0.02.

These results suggest distinct trends in TBI incidence based on both age and gender. Males in the 12-18 age group are at a particularly high risk, which could be attributed to factors such as increased engagement in contact sports, driving, and other high-risk activities. The higher TBI rates for females in the 5-12 age group may also warrant further investigation. Overall, the figure highlights that as children age, the likelihood of experiencing a clinically important TBI increases, with gender playing a crucial role in the risk, especially during adolescence. These findings emphasize the need for targeted prevention strategies, such as promoting safety in sports, educating about helmet use, and reducing risky behaviors among adolescents, particularly boys in the 12-18 age range.

3.4 Reality Check

The bar plot in Figure 8 compares the number of consistent and inconsistent cases of clinically important traumatic brain injury (TBI) in relation to trauma variables such as death, hospitalization, neurosurgery, and intubation.

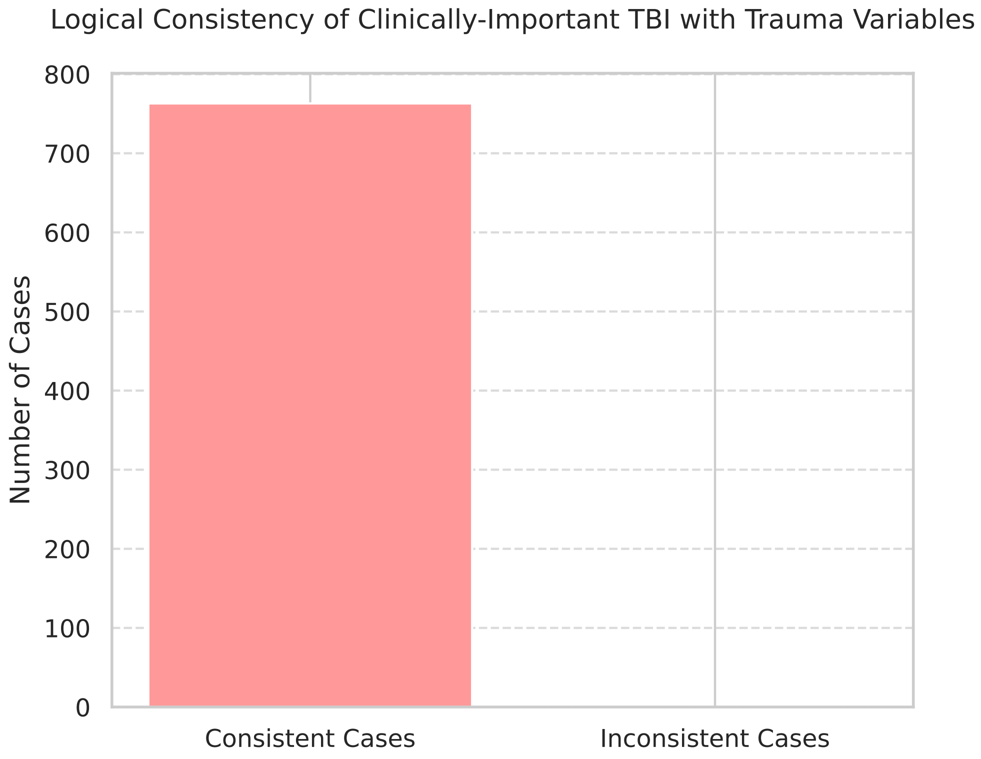

Figure 8: Logical Consistency of Clinically-Important TBI with Trauma Variables

The reality check here compares the cleaned data against real-world expectations that clinically important TBIs should be associated with medical interventions. In this case, the dataset passes the reality check since all cases where PosIntFinal = 1 (clinically important TBI) have associated trauma variables (e.g., hospitalization, surgery). This logical consistency indicates that the data is reliable, with no severe TBI cases missing critical trauma interventions. Therefore, no issues are found with the dataset's alignment with reality.

3.5 Stability Check

In the stability check, I applied small perturbations to the dataset to assess the sensitivity of the findings on the predominant causes of hospitalization. The plot (Figure 9) compares the original and perturbed data, showing minor variations in the counts of different causes.

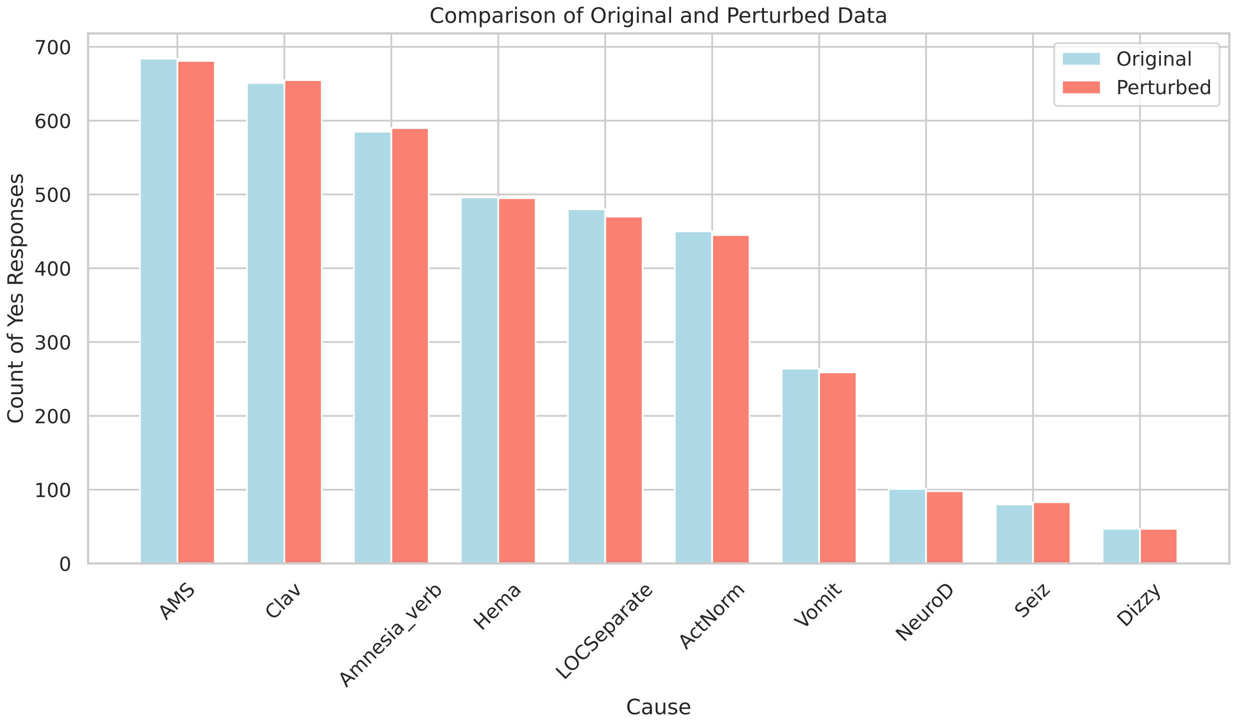

Figure 9: Comparison of Original and Perturbed Data for Hospitalization Causes

Key differences between the original and perturbed data include a small decrease in *AMS* cases (-3) and a slight increase in *Clavicle Injuries (Clav)* (+4). *Amnesia Verbally Reported (Amnesia_verb)* shows a moderate increase (+5), while *Hematoma (Hema)* and *Loss of Consciousness (LOCSeparate)* exhibit a decrease, with *LOCSeparate* showing the largest drop (-10). Other categories like *Vomit*, *Neurological Deficit (NeuroD)*, and *Activity Normalization (ActNorm)* also decrease, while *Seizures (Seiz)* shows a slight increase (+3).

Overall, the results indicate that while some categories, such as *LOCSeparate* and *ActNorm*, are more sensitive to perturbations, key causes like *AMS* and *Clav* remain stable, suggesting robustness in the main findings despite minor data fluctuations.

4 Discussion

The findings presented in this report provide important insights into the patterns of traumatic brain injury (TBI) mechanisms, severity levels, and hospitalization causes among pediatric patients. The exploratory data analysis (EDA) revealed that motor vehicle crashes and falls are the most frequent injury mechanisms, particularly contributing to moderate injury severity levels. Additionally, the analysis of hospitalization causes underscored the prominence of *Altered Mental Status (AMS)* and *Clavicle Injuries* as the leading reasons for hospital admission, suggesting these conditions are critical for clinical attention. The reality check confirmed the logical consistency of the dataset, demonstrating that all clinically important TBI cases were associated with appropriate medical interventions, such as neurosurgery or hospitalization, ensuring the dataset's reliability for clinical research.

In the stability check, small perturbations introduced to the dataset showed minimal impact on the overall findings. While certain categories, like *Loss of Consciousness* and *Activity Normalization*, were more sensitive to data perturbations, the key causes of hospitalization, such as *AMS* and *Clavicle Injuries*, remained stable, indicating that the findings are robust and reliable. Data size did not impose major restrictions on the analysis, although handling missing data required careful attention. This report demonstrates that there is a strong correspondence between the data and reality, and the use of data visualization techniques helped simplify and clarify these complex findings for clinical application.

5 Conclusion

The report provides valuable insights into the patterns of traumatic brain injuries (TBI) in pediatric patients, particularly in relation to injury mechanisms, severity levels, and causes of hospitalization. The analysis highlighted that motor vehicle crashes and falls are the most frequent mechanisms of injury, while *Altered Mental Status (AMS)* and *Clavicle Injuries* were the predominant causes of hospitalization. Through the reality check, we ensured the dataset's logical consistency, confirming that severe TBI cases were associated with appropriate medical interventions. The stability check demonstrated that the findings remained robust even when subjected to minor data perturbations. These insights can assist in refining clinical decision rules (CDRs) to reduce unnecessary CT scans while maintaining patient safety, ultimately supporting improved care for pediatric patients at risk of TBI.

6 Academic honesty statement

I personally designed and conducted all the data analysis procedures presented in this report. I also wrote all the text and created all the figures in this report. All procedures are fully documented, and the results can be completely reproduced. Wherever I have incorporated work from others, I have properly cited the sources.

References

[1] Takashi Araki, Hiroyuki Yokota, and Akio Morita. Pediatric traumatic brain injury: characteristic features, diagnosis, and management. *Neurologia medico-chirurgica*, 57(2):82–93, 2017.

[2] Nathan Kuppermann et al. Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. *The Lancet*, 374:1160–1170, 2009.

[3] Diana L. Miglioretti, Eric Johnson, Andrew Williams, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. *JAMA Pediatrics*, 167(8):700–707, 2013.