Lab 1 - PECARN TBI Data, Spring 2025

Academic Integrity Pledge

Dear Bin Yu, I affirm that this work is entirely my own. I believe academic research honesty is necessary because it ensures the reliability of scientific knowledge, enables proper attribution of ideas, and allows for the meaningful advancement of our collective understanding. Through this work, I have strived to maintain the highest standards of academic integrity.

Introduction

Our problem focuses on identifying children at very low risk of clinically-important traumatic brain injuries (ciTBI) following head trauma, aiming to reduce unnecessary CT scans without compromising patient safety. Building upon the validated PECARN decision rules (Kuppermann et al., 2009), we specifically examine how Glasgow Coma Scale (GCS) scores, along with key clinical indicators like vomiting and seizures, can be optimized for risk stratification. This approach is particularly relevant given the increasing use of CT imaging in pediatric emergency care, despite the associated long-term risks of radiation-induced malignancies.

The significance of this work extends beyond individual patient outcomes to broader healthcare policy and clinical guidelines. Overuse of CT scans not only exposes children to unnecessary radiation but also increases healthcare costs and emergency department congestion. A well-validated prediction model could standardize assessment protocols, reducing variability in clinical practice and optimizing resource allocation. Additionally, by focusing on objective clinical indicators rather than subjective judgment alone, we can enhance diagnostic accuracy and physician confidence in identifying cases where imaging is unnecessary. This research has the potential to improve patient safety, streamline emergency care, and contribute to evidence-based guidelines for pediatric head trauma management.

Data

The dataset is structured with categorical groupings to standardize clinical variables and support predictive modeling. Injury mechanisms are classified by severity, while symptoms such as vomiting and headache are recorded based on frequency and intensity. Physical exam findings, including scalp hematomas and skull fractures, are categorized by size and location to ensure consistency in risk assessment. Additionally, age stratification can distinguish between preverbal (less than 2 years old) and verbal (greater than or equal to 2 years old) children, acknowledging differences in clinical presentation. By organizing data in this manner, the study enhances predictability across cases, facilitating the development of a robust and interpretable prediction model for evaluating ciTBI risk.

Domain Problem to Solve

The most relevant data for identifying low-risk ciTBI cases was selected based on both the PECARN clinical decision rules and current emergency care practices. Following domain expertise from Kuppermannet al.'s validated predictors and common emergency room practices, I focused particularly on GCS eye scores, vomiting, and seizures, which serve as a critical neurological indicator and proxy for pupillary response. While there were more indicators that emergency room nurses typically use to diagnose TBI, these three were the overlapping ones provided in our dataset.

Specific attention was paid to variables that emergency physicians can readily assess, ensuring the resulting model would be practical for real-time clinical decision-making. The focus on GCS scores, particularly the eye component, was motivated by its objective nature and universal use in emergency settings. This aligns with Kuppermann et al.'s emphasis on using standardized, reliable clinical indicators for risk assessment. Moreover, we supplement our chosen features with those found predictive of TBI in Kuppermann's paper.

Data Collection

The data was collected from pediatric patients presenting with head trauma at 25 emergency departments across North America. Inclusion criteria required patients to be under 18 years old, presenting within 24 hours of injury, and having a Glasgow Coma Scale (GCS) score of 14 or 15. The focus on children with mild head trauma was driven by the need to reduce unnecessary CT scans in this population, which is frequently imaged despite having a low risk of ciTBI.

Standardized clinical assessments were conducted by trained emergency physicians, who documented key findings using structured case report forms. To ensure reliability, a subset of patients underwent independent re-evaluations by additional clinicians within an hour of the initial assessment, allowing for inter-rater reliability checks and reducing subjective bias. This

methodology was specifically chosen to minimize variability in clinical assessments and ensure consistent data quality across all sites.

CT scans were performed at the discretion of emergency physicians using helical CT scanners with radiographic slices no greater than 10 mm apart. Faculty radiologists at each site interpreted the scans, while a designated pediatric radiologist independently reviewed inconclusive cases. Structured follow-up calls were conducted between 7 and 90 days post-visit for discharged patients to identify any missed injuries. If follow-up suggested a missed injury, medical records and imaging results were reviewed. Hospitalized patients had their records examined by research coordinators to confirm CT findings and clinical outcomes.

The data was collected between 2004 and 2006, with separate derivation and validation cohorts. While changes in operators or instruments over time were not explicitly mentioned, variability in clinician decision-making, training differences, and evolving hospital protocols could have influenced data consistency. Additionally, the selective recording of certain variables (like IndSeiz) only for patients being considered for CT scans introduces potential sampling bias that must be considered in our analysis.

Meaning

The variables selected for the decision tree model were carefully chosen based on domain expertise and research evidence. Particular emphasis was placed on the GCS components, especially the eye score, given its common use in emergency settings in head trauma assessment. The GCS eye score provides a standardized measure of neurological function, though its interpretation can be challenging in very young children or those with pre-existing conditions.

The injury mechanism variable captures the nature of the trauma, differentiating between low-risk scenarios (such as ground-level falls) and high-risk events (such as motor vehicle crashes). The documentation provides clear categorical labels, but potential issues arise in how emergency physicians classify ambiguous cases, such as falls from moderate heights or sports-related injuries. The injury severity variable further refines the assessment by categorizing the impact force, but this is subject to physician judgment and might lack objective quantification, leading to variability.

Loss of consciousness (LOC) and LOC duration are key neurological indicators. LOC is typically reported by caregivers or inferred by physicians based on patient behavior, introducing possible recall bias or misclassification. LOC duration is even more problematic, as precise time measurements are often unavailable in emergency settings. Patients who are confused but not fully unconscious might be misclassified, affecting the reliability of this variable.

Post-traumatic seizures are another critical indicator, but the dataset has known issues regarding seizure classification. The IndSeiz variable serves as an indicator for seizure occurrence but was only recorded for patients being considered for a CT scan, as stated in the documentation. This systematic missingness creates potential bias in our analysis, as seizure data may

be underrepresented in cases deemed lower risk. The Seiz variable records the presence of a post-trauma seizure, but missing values introduce uncertainty—some of these may represent unreported seizures rather than true negatives. Similarly, SeizOccur and SeizLen suffer from similar documentation challenges.

Data Cleaning

For our data cleaning process, we focused on feature construction and cleaning for vomiting and seizure variables, guided by clinical expertise that identifies these as key symptoms of TBI along with pupillary response (measured through GCS Eye scores). Our primary cleaning involved the construction of a comprehensive seizure indicator (HadSeiz) that combined information from multiple variables to address the limitations of the original seizure data. We simply deduced wether we knew the patient has had a seizure, did not have a seizer, if we are unsure, or if there is a contradiction in the data (one feature says the patient never had a seizure when another says they have). IndSeiz, which indicates seizure occurrence, produced potential false negatives since it was only recorded for patients being considered for CT scans. Similarly, the Seiz variable capturing post-trauma seizures contained missing values that needed careful interpretation. The cleaning process also addressed the complexities of seizure timing (SeizOccur) and duration (SeizLen) variables, where "not applicable" entries could indicate either true absence or unrecorded data.

For vomiting data, we created evidence and lack-of-evidence counters to handle the complex missing values and documentation practices common in emergency settings. Moreover, we cleaned the outcome labels by identifying and handling missing values in PosIntFinal, paying particular attention to cases with positive CT findings to ensure no potential TBI cases were overlooked. We dropped 18 of the 20 unlabeled patients because we the number represented a small proportion of the data's total observations, allowing us to reasonably assume these patients would not significantly effect our final analysis. Our judgement call is especially justified considering we chose to use undersampling techniques to ensure our models do not overestimate the number of disproportionately non-TBI patients.

Lastly, we replaced each and every 92 variable left over after our data processing with NaN values to ensure consistent representation throughout our features.

Relevance

The focus on GCS scores, particularly the eye component, along with seizure and vomiting patterns, aligns with validated clinical decision rules and current emergency medicine practice. While these features provide strong discriminatory power, the systematic missingness in certain variables (particularly seizure data) may affect our model's generalizability. However, we discuessed how we constructed our own features and supplement them with the predictive features Kuppermann finds to alleviate these issues. Moreover, given our goal of reducing

unnecessary radiation exposure while maintaining safety, the current dataset provides sufficient clinical markers for developing an effective prediction model.

Comparability

The data units across variables are comparable, as they are based on standardized clinical assessments. Glasgow Coma Scale (GCS) scores, seizure presence, and injury severity are categorical, while LOC duration is recorded in time intervals. There is no evidence of incompatible data types being combined, ensuring that the model can treat them as exchangeable within their respective categories. The dataset appears to maintain independence between observations, as each patient is assessed separately. A check for redundant variables was performed to ensure that no duplicate columns were included, and each selected variable contributes unique predictive value to the decision tree model.

Presenting Findings

Our analysis of the PECARN TBI dataset reveals several key patterns relevant to clinical decision-making. We present four visualizations that highlight critical relationships in the data.
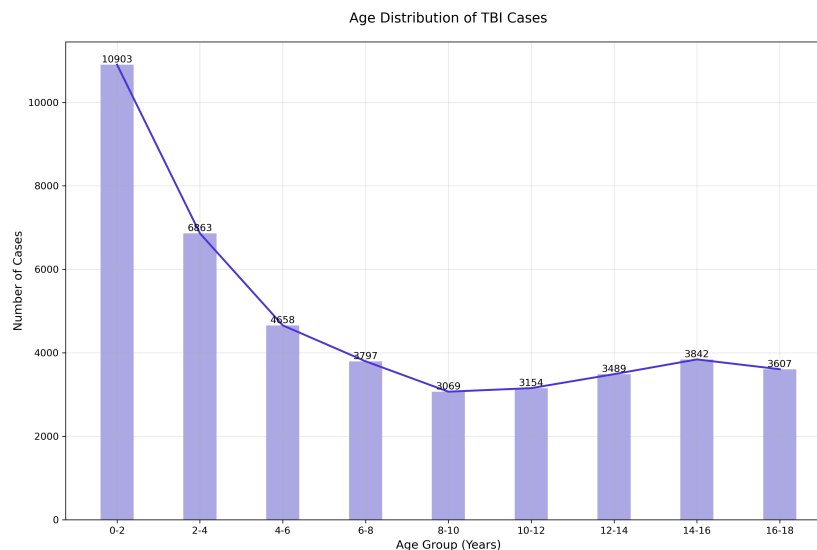

Figure 1: Age Distribution of TBI Cases in Pediatric Patients

Figure 1 demonstrates the age distribution of TBI cases, showing a clear predominance in very young children (0-2 years) with 10,904 cases. This early peak likely reflects the vulnerability of toddlers developing mobility without full coordination or risk awareness. A second,

smaller peak in youth (14-16 years) could suggest increased risk from sports and recreational activities.
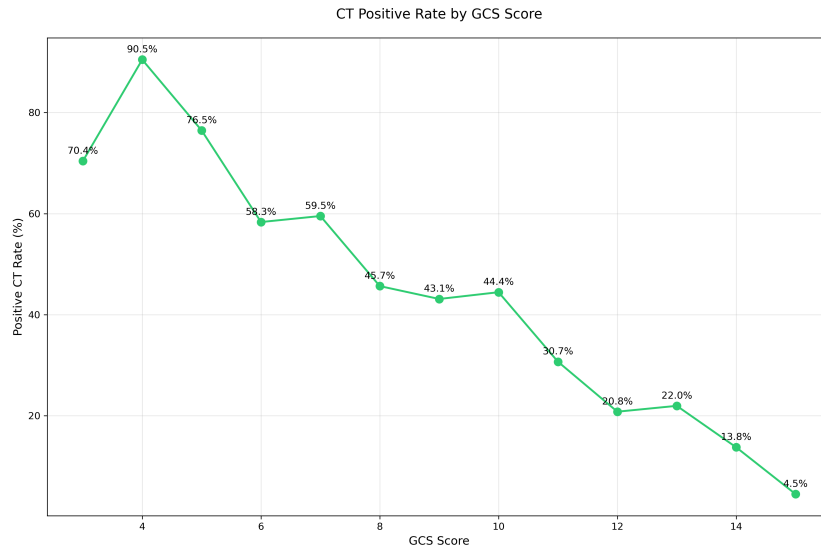

Figure 2: Relationship Between GCS Score and Positive CT Findings

Figure 2 illustrates the strong inverse relationship between Glasgow Coma Scale scores and positive CT findings. The dramatic increase in positive CT rates at lower GCS scores (reaching 90.5% at GCS 4) provides strong evidence for the scale's predictive value. This relationship supports current clinical protocols that prioritize CT imaging for patients with lower GCS scores.

Figure 3 examines the specific predictive power of the GCS eye component. The parallel decline in both CT positive rates and neurosurgical intervention rates as eye scores improve suggests this component alone could serve as a valuable rapid screening tool. Notably, patients with no eye opening (score 1) show dramatically higher rates of both positive CT findings (68.4%) and neurosurgical intervention (36.5%).

Figure 4 explores the relationship between vomiting episodes and seizure occurrence. The heatmap reveals that among patients who experience seizures, the pattern of vomiting differs notably from those without seizures. This interaction between two key clinical indicators could help refine risk assessment protocols, particularly for cases where other neurological signs are ambiguous.

These visualizations together support a nuanced approach to pediatric TBI assessment, where age, neurological status (particularly GCS scores), and associated symptoms all contribute to risk stratification. The findings suggest potential opportunities for reducing unnecessary CT scans in specific patient subgroups while maintaining safety for high-risk cases.
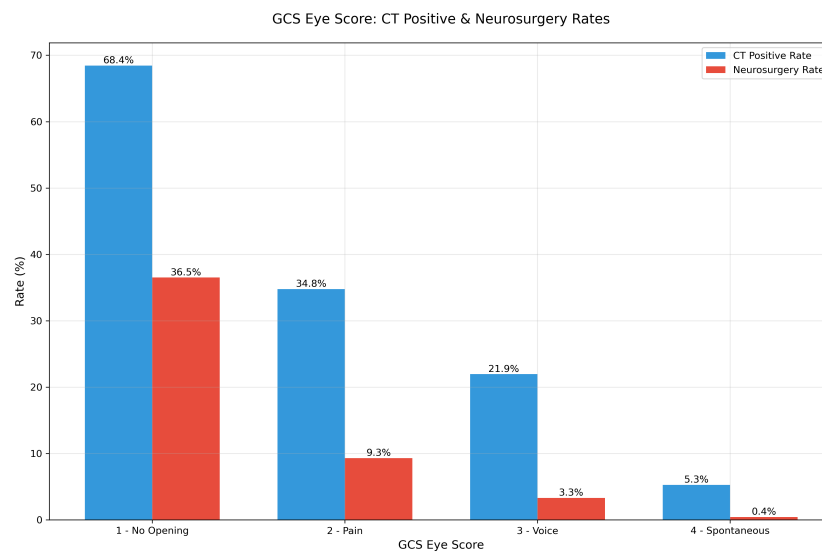
Figure 3: GCS Eye Score's Relationship with CT Findings and Neurosurgical Intervention
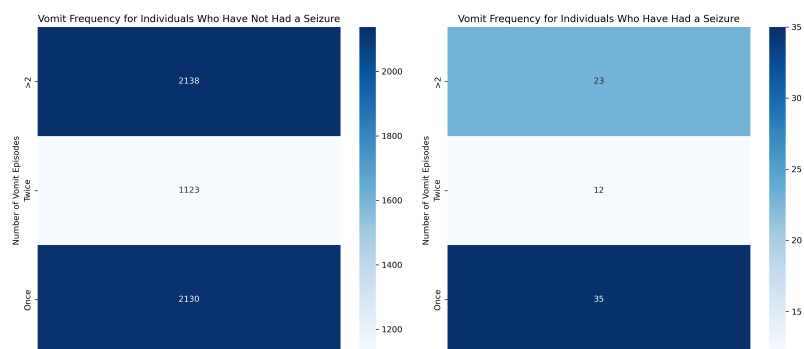

Figure 4: Relationship Between Vomiting Frequency and Seizure Occurrence

Reality Check

After conducting our data cleaning and visualization processes, it's crucial to validate our decisions against external reality and domain knowledge. Our approach to feature construction and data cleaning was strongly guided by established clinical practice, particularly regarding the key symptoms of TBI as documented by emergency care professionals. The decision to focus on seizures, vomiting, and GCS eye scores aligned with standard medical protocols, while our judgement calls for handling missing and contradictory data reflected the realities of emergency department and conservatism.

One particularly interesting finding that initially seemed counterintuitive – that patients with seizures tend to have fewer vomiting episodes – could align with clinical reality when considered more deeply. This pattern could be explained by the different temporal patterns of symptom onset (vomiting typically being immediate while seizures may develop later) and the likelihood of earlier medical intervention in seizure cases potentially preventing the progression of other symptoms. Discussing more with medical experts about this relationship could help further validate our cleaning.

In general, through both our cleaning procedures and subsequent analysis, the dataset maintained strong alignment with clinical reality. Our data cleaning decisions preserved important clinical relationships while addressing data quality issues in ways that reflect actual medical practice, providing a solid foundation for further analysis and model development.

Stability Check

In our analysis of GCS Eye scores and their relationship to CT positivity and neurosurgery rates, we initially presented our findings using a bar graph (Figure 3) that visually conveyed the percentage differences in CT positivity and neurosurgery rates across different GCS Eye score categories. While this graphical approach is effective for quickly communicating trends, we explore whether representing the same information as a frequency table affects our interpretation or the perceived significance of our results.

Eye Score	Total Cases	CT Positive Rate (%)	Neurosurgery Rate (%)
No Opening	263	68.4	36.5
Pain	161	34.8	9.3
Voice	483	21.9	3.3
Spontaneous	14538	5.3	0.4

Table 1: GCS Eye Score: CT Positive and Neurosurgery Rates

The frequency table presents the same findings in tabular form, explicitly listing the total number of cases, CT positive rates, and neurosurgery rates for each eye score category. The structured nature of the table allows for a direct numerical comparison, reinforcing the stark

contrast in risk between patients with no eye opening (68.4% CT positive, 36.5% neurosurgery rate) versus those with spontaneous eye opening (5.3% CT positive, 0.4% neurosurgery rate). Unlike the bar graph, which emphasizes proportional differences visually, the table highlights the absolute case counts, providing a clearer sense of sample size distribution across categories.

This change in presentation does not alter the underlying findings but offers a complementary perspective. The frequency table may be preferable in contexts where precise numerical values are needed for decision-making, such as medical guideline development, whereas the graph may be more effective for conveying overall trends at a glance. The robustness of our conclusions remains intact across both representations, affirming the reliability of our findings regarding the predictive value of GCS Eye scores in TBI assessment.

Modeling

Model Implementation

For our classification task of determining CT scan necessity, we implemented two primary decision tree models and a complementary risk score model. The first decision tree model prioritizes extremely high sensitivity (TPR: 99.87%, FPR: 37.97%), adhering to the critical requirement of not missing any serious injuries. This model uses a focused set of predictors: GCS Eye score, seizure occurrence, and vomiting patterns. The second, more conservative decision tree model (TPR: 90.60%, FPR: 16.54%) incorporates additional clinical indicators including loss of consciousness, scalp findings, and alertness measures, achieving better specificity while maintaining clinically acceptable sensitivity.
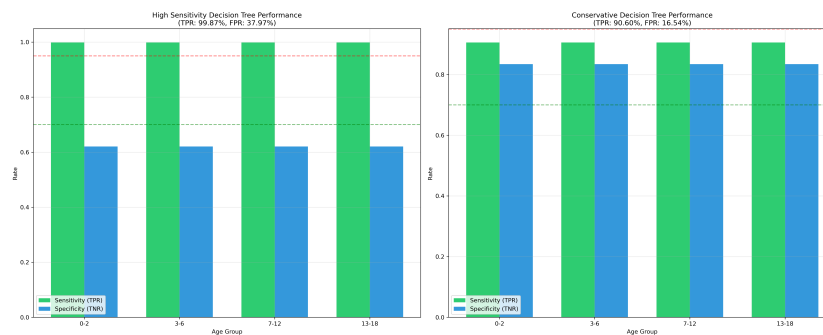

Figure 5: Decision Tree Model Performances

Figure 5 shows how our decision tree models perform across different age groups. The high sensitivity model (left) maintains near-perfect sensitivity across all age groups at the cost of higher false positives, while the conservative model (right) achieves better specificity while maintaining clinically acceptable sensitivity levels.

We also developed a novel risk score model was also developed to provide a different perspective on risk assessment. This model employs age-stratified thresholds and continuous risk scoring, achieving varying performance across age groups. However, this model has high specificity for those between the ages of 3-18, meaning it could be useful for cases where the decision trees show disagreement.
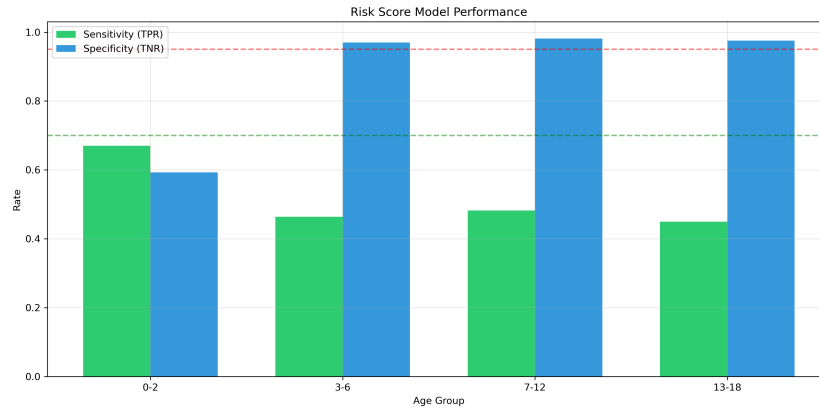

Figure 6: Risk Score Model Performance

Interpretability

The decision tree models offer strong interpretability which is especially helpful for clinical implementation. The high-sensitivity model makes primary splits on GCS Eye scores and seizure presence, reflecting established clinical indicators. As shown in our earlier analysis of GCS Eye scores (Figure 3), this approach is well-grounded in clinical evidence, with clear relationships between eye response and both CT findings and neurosurgical intervention rates.

The conservative model's additional features allow for more nuanced risk stratification while maintaining clear decision paths that clinicians can easily follow. These transparent decision processes align well with current medical practice, where clinicians need to understand and explain their assessment rationale.

Both decision trees can be interpreted as clinical flowcharts, making them particularly suitable for emergency department use. The risk score model, while slightly more complex in its continuous scoring approach, still maintains interpretability through its clear age-stratified thresholds and explicit risk contribution from each clinical indicator.

The risk score model combines clinical indicators in a transparent, weighted manner that aligns with medical reasoning. The model starts with GCS Eye scores as the foundation, using actual observed CT positive rates we found in our previous section (68.4% for score 1, 34.8% for score 2, 21.9% for score 3, and 5.3% for score 4) as base risk values. This choice reflects the strong predictive power of eye response in neurological assessment. Additional risk is then added

for concerning symptoms: seizures contribute a significant risk increase (0.3) due to their strong association with serious injury, while vomiting adds graduated risk (0.25 for multiple episodes, 0.15 for any vomiting) based on severity. When multiple symptoms are present, the model applies a 20% increase to the final risk score, reflecting the clinical understanding that symptom combinations often indicate higher severity. These scores are then compared against age-specific thresholds (0.30 for young children, 0.35 for middle childhood, and 0.40 for adolescents), with lower thresholds for younger patients reflecting their higher risk and more challenging assessment. This structure allows clinicians to understand exactly how each symptom contributes to the final recommendation, making it particularly useful for cases where decision trees might disagree or when a more nuanced risk assessment is needed.

The effectiveness of vomiting, GCS eye testing, and seizure features in both a weighted average and a decision tree approach illustrates these features' predictive robustness.

Stability

The perturbation introduced in our stability check – switching from a bar graph to a frequency table – did not impact our model's predictions or overall interpretations. This is expected, as the underlying numerical relationships between GCS Eye scores, CT positivity rates, and neurosurgery rates remained unchanged. The representation choice affects how the data is presented to human interpreters but does not alter the fundamental statistical properties of the dataset that our models rely on.

Our decision trees and risk score model were trained on raw numerical values, not visualizations, meaning that their decision boundaries remained the same before and after the perturbation. This confirms that our modeling approach is robust to changes in data representation, as long as the input distributions and key features remain consistent. Additionally, it underscores the fact that our visual choices primarily affect communication rather than the substantive conclusions drawn from the analysis.

The robustness of our models under this perturbation suggests that the relationships between GCS Eye scores and CT outcomes are strong and stable. This strengthens our confidence that the insights derived from our models are not artifacts of a specific visualization method but rather reflect genuine patterns in the data. It also reinforces that our choices in data preprocessing and feature engineering were sound, ensuring that the predictive power of our models remains consistent regardless of how the findings are presented.

Discussion

The vast size of the PECARN dataset (43,399 cases) provided challenges for our analysis. While the large sample size allowed for robust analysis and age-stratified modeling, picking features from the smorgasbord provided proved impossible without relying heavily on domain

expertise. While domain expertise is necessary for any data science project, over-reliance could make us less inclined to look for predictive feature relationships unknown to the general medical community: us being satisfied with the local minima we have nestled into.

In considering the three realms of data science - data/reality, algorithms/models, and future data/reality - our analysis demonstrates both strengths and limitations. In the data/reality realm, our cleaning process preserved key clinical relationships while addressing quality issues. The algorithms/models realm benefited from our cleaned dataset, enabling the development of both high-sensitivity and conservative prediction approaches. For future data/reality, our cleaned dataset structure and documented preprocessing decisions provide a framework for handling new cases, though local validation would be needed for implementation in new clinical settings. Transferability of our model should not be implied nor taken as granted.

The relationship between our data and reality shows strong concordance in many areas, particularly in age-related risk patterns and neurological assessment metrics. However, we should acknowledge that emergency department data collection can never perfectly capture the complete clinical picture. Moreover, using our models outside of the narrow scope of the circumstances the data was collected would be negligent without again empirically proving transferability.

Next, our visualization approaches – particularly in showing model performance across age groups – help bridge the gap between ground truth and our model by presenting complex relationships in ways that align with clinical decision-making processes.

The high interpretability of our models – from the simple decision paths of our trees to the clear risk contribution of each factor in our scoring model – makes them particularly valuable for real-world medical practice where physicians must be able to explain and justify their clinical decisions to patients and colleagues.

References

Kuppermann, (2009). Identification of children at very low risk of clinically-important brain injuries after head trauma: A prospective cohort study. The Lancet, 374(9696), 1160–1170. https://doi.org/10.1016/S0140-6736(09)61558-0

Registered Nurse RN. EKG Interpretation (Part 1) - Introduction to EKGs, Rate, Rhythm! [Video]. YouTube. https://www.youtube.com/watch?v=e7Yl6urL7n0

Registered Nurse RN. NCLEX Review: Nursing Cardiac Medications Pharmacology – Antiarrhythmic, Beta Blockers, Calcium Channel Blockers [Video]. YouTube. https://www.youtube.com/watch?v=a33Hq89sELc