Lab 1 - PECARN TBI Data, STAT 215A, Fall 2026

September 20, 2026

1 Introduction

Identifying children at risk of clinically important traumatic brain injury (ciTBI) matters because clinicians must balance missed injuries against unnecessary computed tomography (CT) scans and radiation exposure. Better understanding the recorded symptoms could help identify useful information for this decision.

The analysis examines how symptom detail, missing answers, and cohort selection relate to ciTBI in the PECARN public-use dataset. The study population includes children younger than 18, and this analysis retains patients whose completed Glasgow Coma Scale (GCS) components sum to 14 or 15. The report covers data collection, cleaning, and exploration, followed by two findings, reality and stability checks, and two simple classifiers.

2 Data

The dataset is the Pediatric Emergency Care Applied Research Network (PECARN) public-use dataset associated with Kuppermann et al. [1]. The raw file contains 43,399 patient records and 125 columns covering demographics, injury mechanisms, symptoms, examination findings, imaging, and outcomes. The supplied documentation explains the variables and response codes [2].

These assessments and outcomes allow the analysis to examine which recorded characteristics accompany ciTBI. The study defines ciTBI as death from traumatic brain injury, neurosurgical intervention for traumatic brain injury, intubation lasting more than 24 hours because of traumatic brain injury, or hospitalization for at least two nights for traumatic brain injury with associated injury on CT. The analysis contains 42,388 patients, including 373 with ciTBI, after applying the GCS eligibility rule and requiring a recorded outcome.

2.1 Data Collection

The data were collected prospectively across 25 North American emergency departments from June 2004 through September 2006. The study enrolled children younger than 18 who presented within 24 hours of head trauma. Clinicians recorded standardized assessments before knowing imaging results. CT use was left to the treating clinicians, so not every patient received imaging. Outcomes were established through medical records and follow-up [1].

The measurements differ in how they are obtained. Age is recorded numerically, GCS components score eye, verbal, and motor responses, and many symptoms rely on clinician judgment or parent reports. Young children have limited ability to describe symptoms, and some assessments depend on age. These differences affect both measurement consistency and completeness. Since imaging was selective, a missing CT result cannot be treated as evidence that no injury occurred.

2.2 Data Cleaning

The raw data contains 20 missing ciTBI outcomes and 1,342 records with at least one missing GCS component. Other issues include disagreement between recorded GCS totals and GCS components, overlapping summary and detailed symptom variables, and special response codes that are fundamentally different from blank entries. A parent summary is a broad symptom indicator, while its child variables describe specific symptoms.

The cleaning and preprocessing steps were applied in the following order.

1. Use the documentation to translate numbers into readable labels, assign categorical types, and rename columns. Keep GCS components as their original numeric scores.
2. Keep explicit responses such as nonverbal, not applicable, and unclear examination separate from blanks. These responses describe the assessment process and should not automatically mean that a symptom is missing.
3. For each record with missing GCS components, find the component combinations that agree with the GCS total. Fill the components only when there is exactly one possible combination. Among retained patients, 1,264 records have components completed this way.
4. Retain records with complete components, a component sum of 14 or 15, and a nonmissing ciTBI outcome. These conditions leave 42,388 patients.
5. Remove patient ID, outcome and related outcome fields, CT findings and decision fields, emergency department disposition, and duplicate age and GCS representations from the predictors. This avoids using later information to predict an earlier clinical decision.
6. Use detailed symptoms instead of the parent symptoms and GCS components instead of the total.
7. Encode ordered categories by their documented order and other categories with one indicator column per category. Encode ciTBI as 0 or 1.

Completing GCS components assumes that the recorded total and observed components are reliable. Ordered encoding imposes numerical spacing between categories. Special responses are explicit categories, but blanks could still have several causes, and parent-child symptom inconsistencies are not automatically fixed.

2.3 Data Exploration

The retained population includes 10,710 children younger than 2 and 31,678 aged 2 or older. ciTBI is uncommon in both groups, with rates of 0.92% and 0.87%, respectively. Figure 1 shows the age distribution and the large imbalance between positive and negative ciTBI outcomes. An analysis based only on accuracy could hide missed injuries because 99.12% of patients do not have ciTBI.

GCS separates groups with very different outcome rates. The GCS 14 group contains 1,321 patients and 104 ciTBI cases, a rate of 7.87%, while the GCS 15 group contains 41,067 patients and 269 cases, a rate of 0.66%. This suggests checking whether findings from the combined population also remain across each GCS group. The presence of both broad symptom summaries and detailed symptoms motivated the first finding, while missing values motivated the second.

3 Findings

3.1 First finding

Broad symptom summaries combine specific signs with different observed ciTBI rates. Figure 2 compares the parent summaries with their child signs across six groups. For example, the rate among patients with a scalp hematoma is 1.37%, but it is 0.59% for frontal hematomas and 3.07% for parietal or temporal hematomas. It is 0.75% for small hematomas and 3.21% for large hematomas. These differences suggest that specific child symptoms can contain information that the summary does not include.

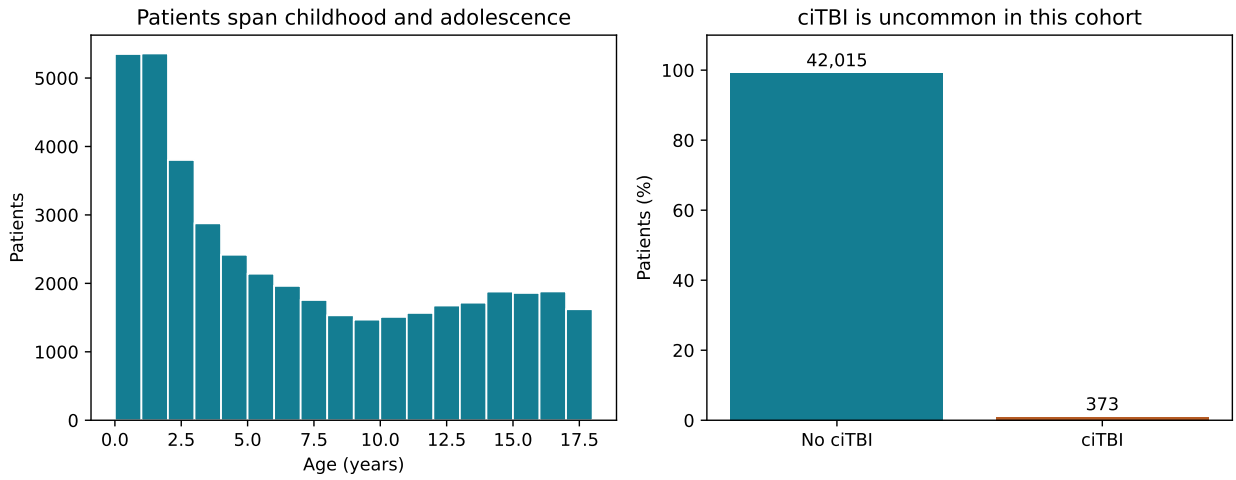


Figure 1: Age and ciTBI outcome distributions for the retained patients.

3.2 Second finding

Patients with blank answers often have higher observed ciTBI rates than patients with recorded answers. This occurs for eight of the nine selected variables in Figure 3. For headache, the rate is 4.95% among 444 patients with blank answers and 0.84% among 41,944 with recorded answers. For seizure history, the corresponding rates are 3.63% and 0.83%. Palpable skull fracture is the exception, with no ciTBI cases among the 101 patients whose answers are blank.

The same eight positive differences appear when the analysis is repeated separately for children younger than 2 and those aged 2 or older. This suggests that the combined pattern is not explained entirely by that age division. However, missingness could still reflect severity, the assessment process, or other patient differences. The practical implication is that replacing blanks without retaining any missingness information could discard useful information about how patients were assessed.

3.3 Reality Check

The published study provides a reference for checking the cohort size, outcome count, age distribution, and CT use [1]. Table 1 compares its reported counts with two selections from the public-use data.

This check assumes that the public-use fields represent the same outcomes and age groups as the published study. It is useful for detecting major errors in variable interpretation or cohort selection.

Cohort	Patients	ciTBI	Under 2	CT scans
Published study	42,412	376	10,718	14,969
Recorded GCS total 14–15	42,412	376	10,718	14,969
Completed component sum 14–15	42,388	373	10,710	14,947

Table 1: Comparison with the published study. Both selections from the public-use file require a recorded ciTBI outcome.

The selection based on recorded GCS totals matches all four published counts. The component-based analysis excludes 24 of those patients, including three with ciTBI, and adds no others. Its ciTBI rate is 0.88%, compared with 0.89% in the paper. The broad agreement supports the interpretation of the data, but the excluded injuries show that the eligibility choice is not inconsequential. The analysis should therefore not be described as using exactly the published cohort.

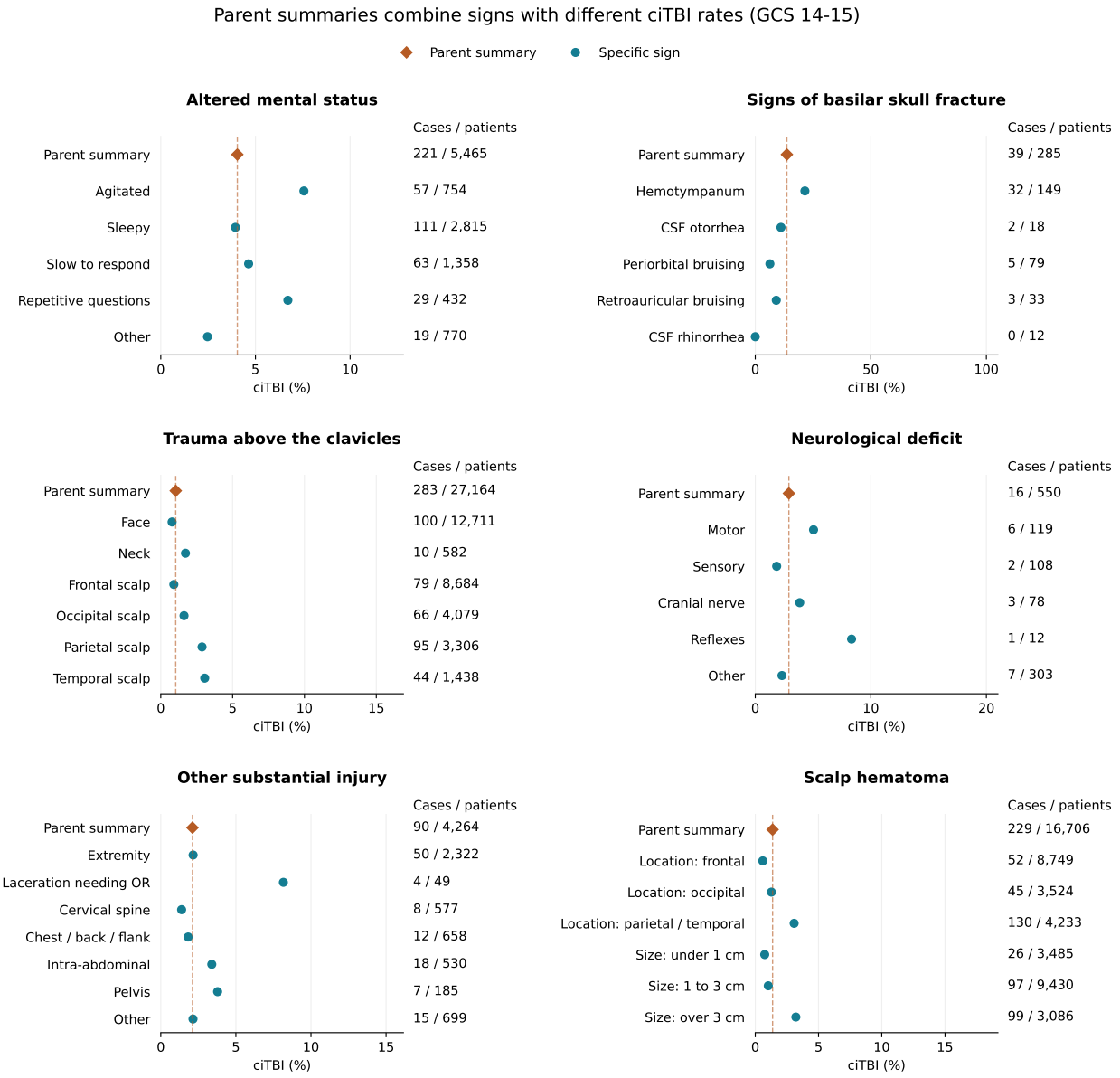


Figure 2: ciTBI rates among patients with each recorded summary and specific symptom. Scales differ between symptom groups.

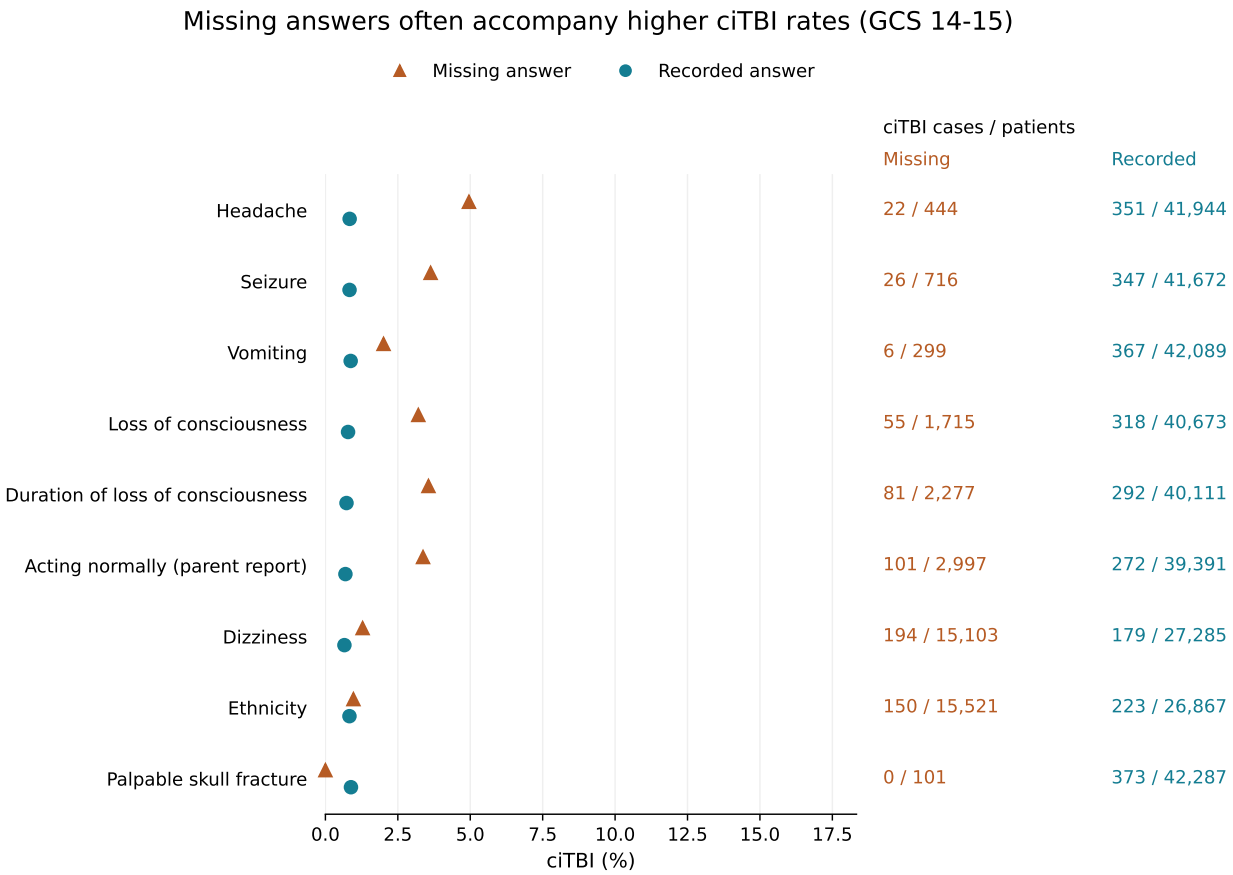


Figure 3: ciTBI rates for blank and recorded responses in the pooled GCS 14–15 population. Recorded responses include explicit special categories, not just yes-or-no answers. Counts help distinguish patterns supported by many observations from those based on small groups.

3.4 Stability Check

The original analysis pools GCS 14 and GCS 15 patients. To test this choice, the missing vs recorded comparison is repeated within both GCS 14 and GCS 15 seperately. This is a clinically relevant perturbation because the groups have different injury rates. It changes both patient composition and sample size, especially for GCS 14, so it is a substantial restriction rather than a small random change. Figure 3 gives the original comparison, and Figures 4 and 5 show the individual versions on the same horizontal scale.

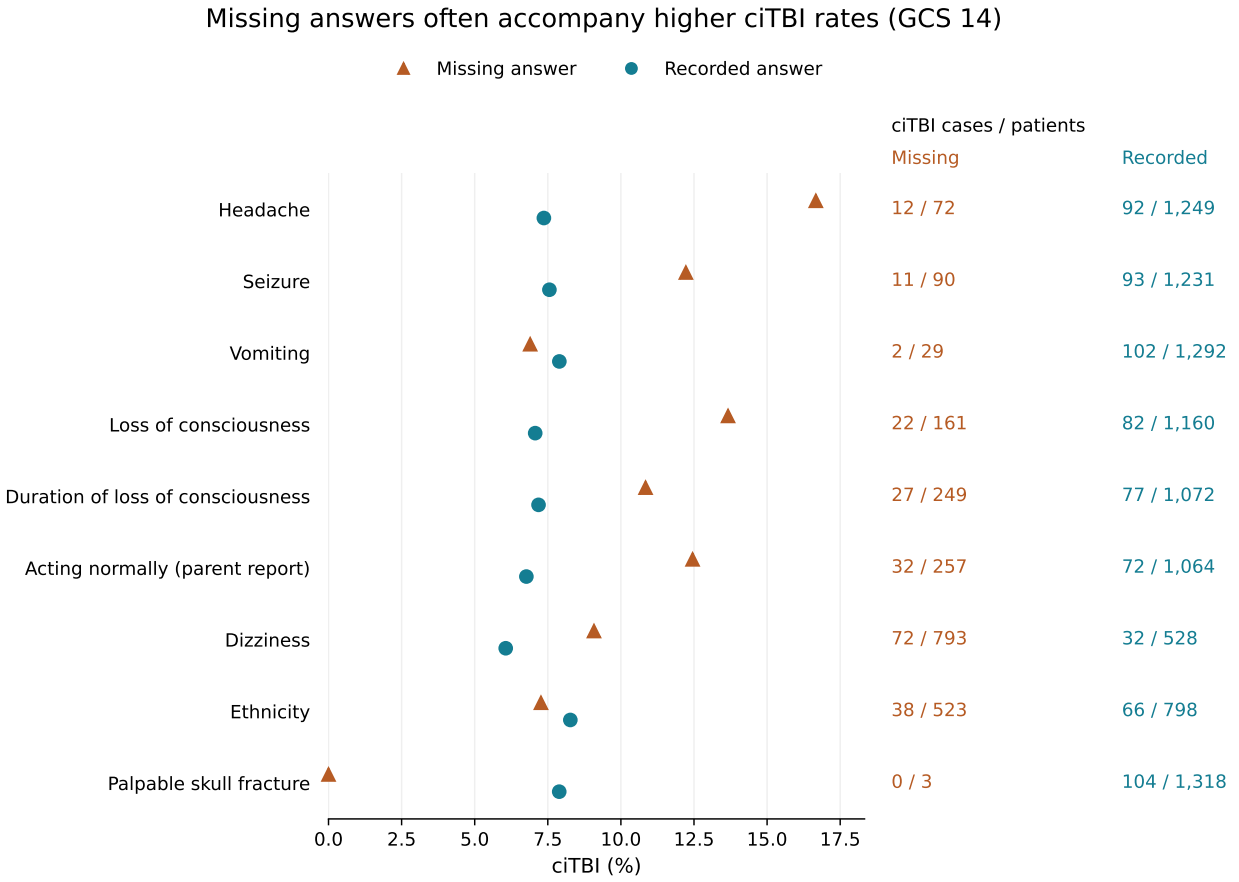


Figure 4: The second finding after restricting to GCS 14, with 1,321 patients and 104 ciTBI cases. Six of nine variables have a higher ciTBI rate among blank responses.

The general missingness pattern persists, but it is not stable for every variable. The headache difference between blank and recorded responses is 4.12 percentage points in the pooled data, 9.30 in GCS 14, and 2.05 in GCS 15. Vomiting and ethnicity reverse direction in GCS 14. In conclusion several missing-response associations persist across GCS groups, while their sizes and some directions depend on the cohort.

4 Modeling

4.1 Implementation

The two classifiers are logistic regression and a small decision tree. A positive flag is treated as a recommendation for CT. Logistic regression penalizes large coefficients using L2 regularization with $C = 1$ and a maximum of 3,000 iterations. The tree has maximum depth 2 and at least 20 training patients per leaf. Both use balanced class weights to give the uncommon ciTBI cases more influence. These hyperparameters were fixed to keep the models simple, rather than selected through a search.

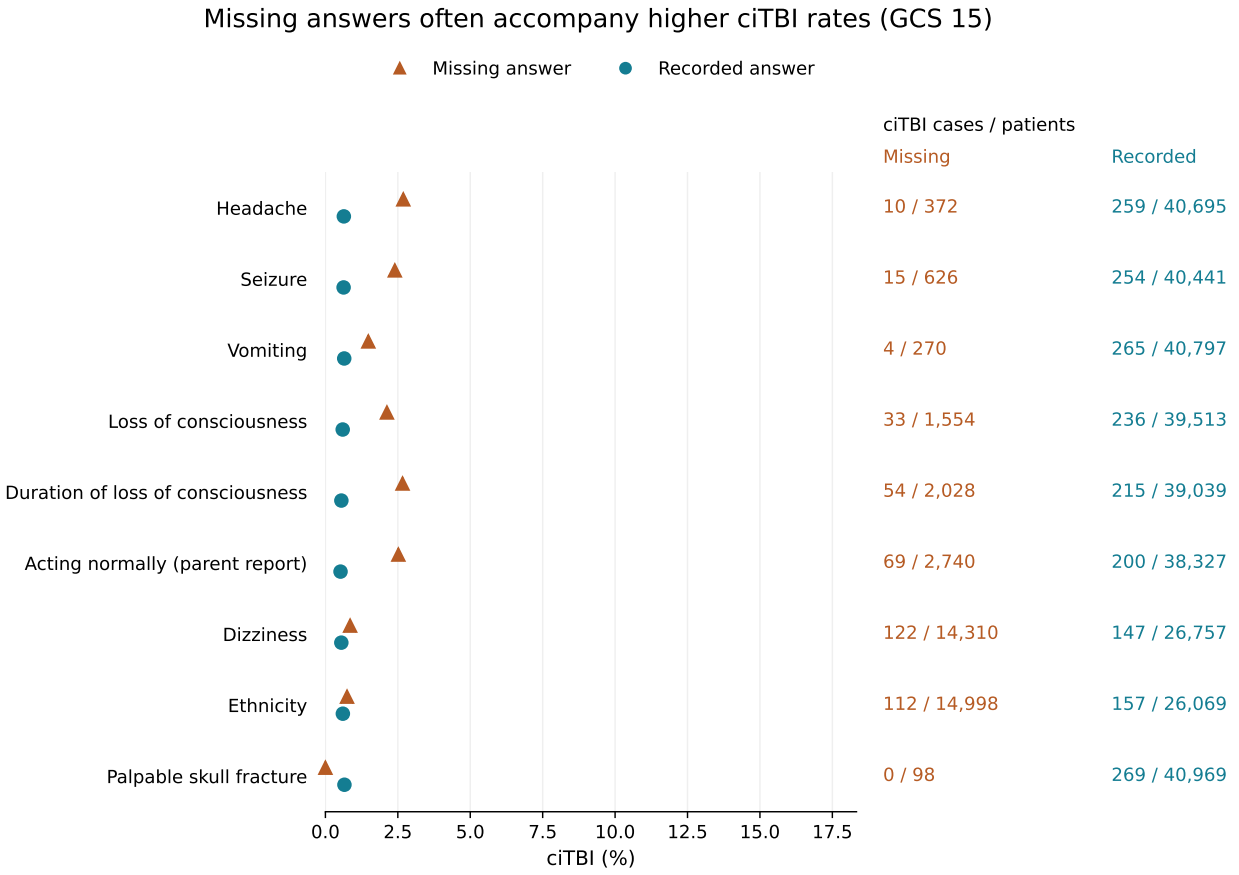


Figure 5: The same comparison after restricting to GCS 15, with 41,067 patients and 269 ciTBI cases. Eight of nine variables retain the combined direction.

The patients are divided into training, validation, and test sets using a 60/20/20 split. Proportional splitting by the combination of GCS group and outcome is used to preserve proportions. The training set contains 25,432 patients and 223 ciTBI cases. The validation and test sets each contain 8,478 patients and 75 cases. Missing values are replaced with the median value from the training data. Extra columns record which values were originally missing. Entirely blank training columns are retained and filled with zero. For logistic regression, each value is adjusted by subtracting the training average and dividing by the training standard deviation. The same averages and standard deviations are used to adjust the validation and test data.

For each model, the cutoff is the highest score that still catches at least 95% of ciTBI cases in the validation data. Sensitivity is the percentage of injury cases caught, while specificity is the percentage of patients without injury who are not flagged. The cutoffs are 0.0548 for logistic regression and 0.2537 for the tree, which catch 96% and 100% of validation cases, respectively. The 95% target was chosen to limit missed injuries.

Model	Missed cases	Sensitivity	Specificity	Flagged	ROC AUC
Logistic regression	8 of 75	89.33%	49.33%	51.01%	0.844
Decision tree	0 of 75	100.00%	0.00%	100.00%	0.770

Table 2: Results on the 8,478 test patients using the validation cutoffs. ROC AUC measures ranking across cutoffs, the other measures use the selected cutoff.

Logistic regression misses eight test injuries despite meeting the validation target. Its precision, the proportion of flagged patients with ciTBI, is only 1.55%. The tree misses none because it flags everyone, so it offers no reduction in the number flagged. Test patients were excluded from model fitting and cutoff selection.

4.2 Stability

The models are refitted using only GCS 14 or only GCS 15 training patients, matching the cohort restriction in Section 3.4. Each restricted model is evaluated on the same test patients as the corresponding subset of the combined model. The original validation cutoff is held fixed. This comparison checks how changing the training cohort affects scores and flags without also selecting a new cutoff.

Logistic regression changes its flag for 117 of 264 GCS 14 test patients, or 44.32%. Sensitivity falls from 95.24% with pooled training to 61.90% with GCS 14 training, and missed cases increase from one to eight out of 21. In GCS 15, 396 of 8,214 flags change, or 4.82%, and missed cases increase from seven to nine out of 54. The GCS 14 training group contains only 793 patients and 62 cases, compared with 24,639 patients and 161 cases in GCS 15.

The tree changes scores but continues to flag every test patient at the fixed cutoff. Unchanged flags therefore do not show useful stability. Overall, the cohort restriction reveals high sensitivity in the logistic model for GCS 14 and shows why score changes and decision changes should both be checked.

4.3 Discussion

Logistic regression and a shallow tree were chosen because they offer two simple ways to relate clinical measurements to ciTBI. Logistic regression combines many predictors into one score, while the tree uses a short sequence of splits and can represent interactions. Together, they show that a model can rank patients reasonably well while still giving an unhelpful decision rule at a demanding sensitivity target.

The tree has a directly interpretable sequence of at most two splits, shown in Figure 7. Logistic regression is less simple to review.

For logistic regression, coefficient signs and magnitudes can help explain how a score is constructed, but correlated signs, rare categories, and full sets of category indicators complicate individual interpretations. Reviewing predictions for concrete patient profiles would make the combined effects easier to understand.

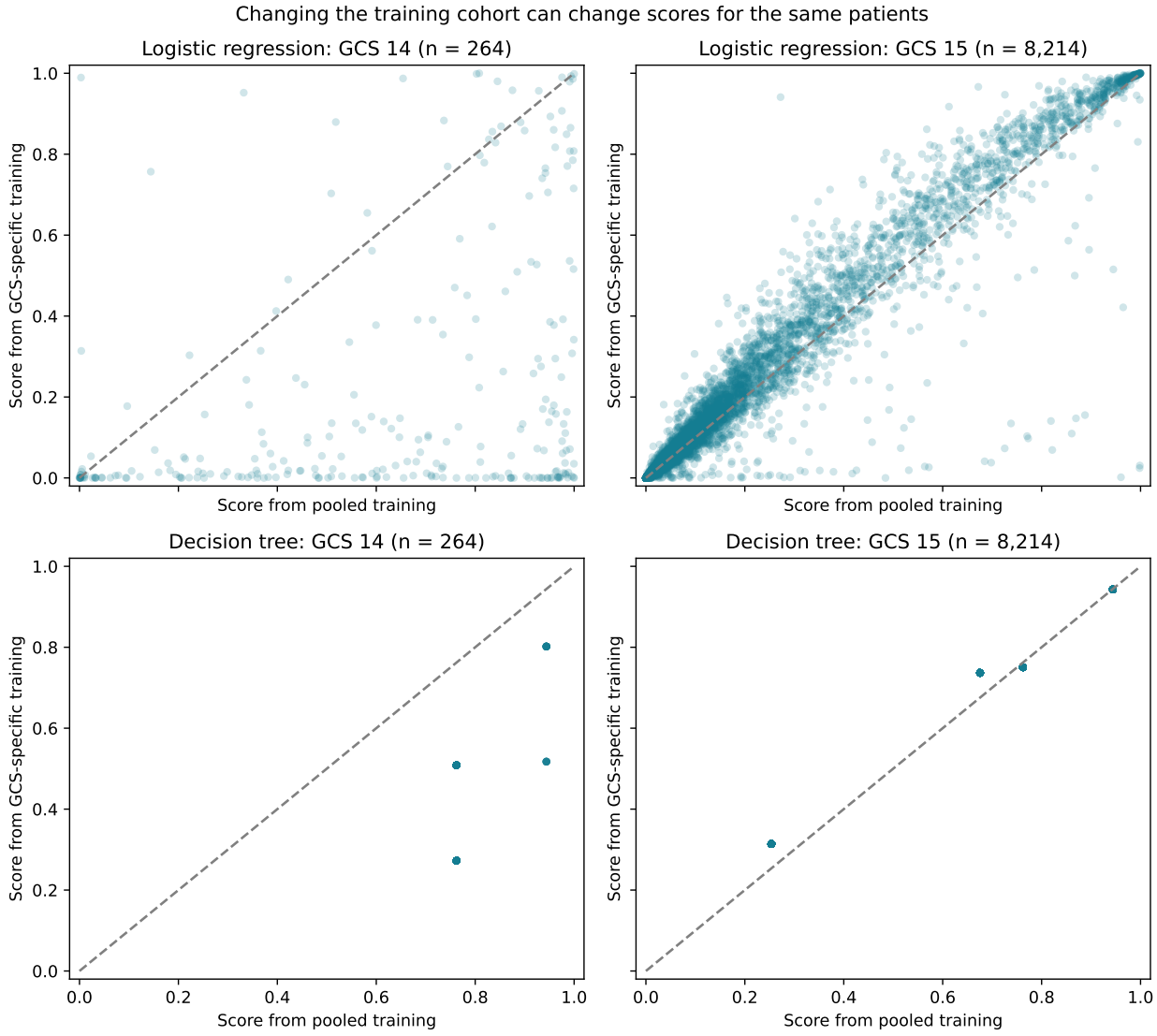


Figure 6: Scores for the same test patients before and after restricting the training cohort. Points on the dashed diagonal have unchanged scores. Identical scores overlap, especially for the trees, so visible point counts do not represent patient counts.

For the tree, following the splits explains the score, but the selected cutoff must still be applied to obtain the final recommendation for CT.

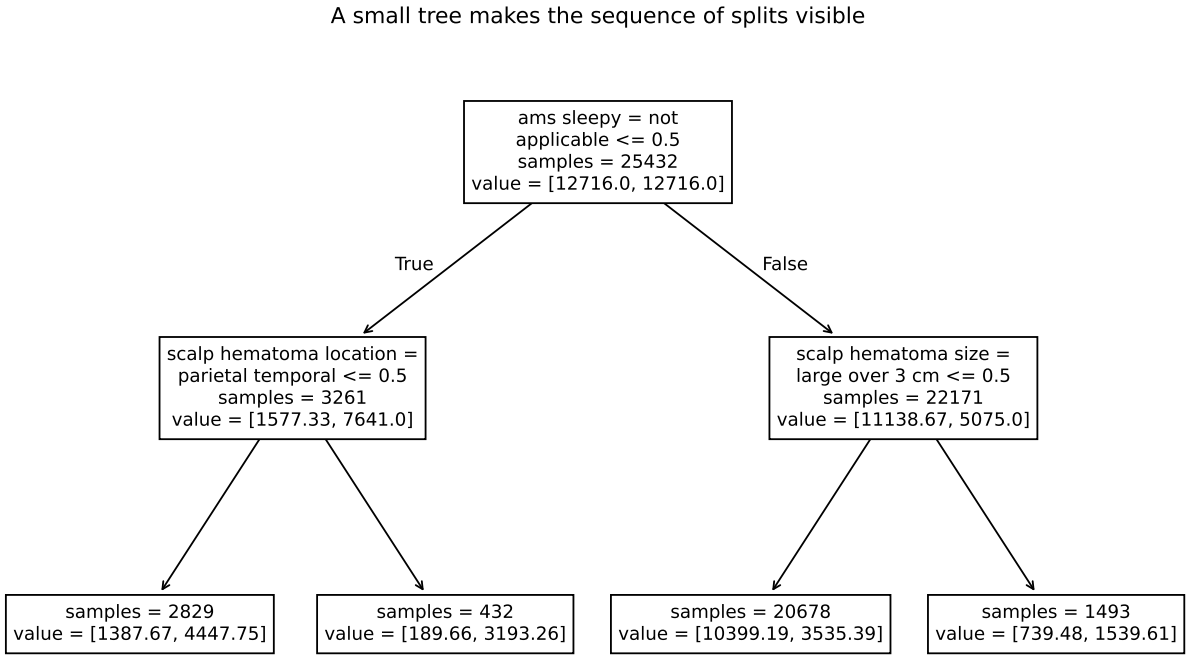


Figure 7: The fitted tree from the combined training data. Samples are patient counts, while values are class-weighted totals for no ciTBI and ciTBI. The weighted proportion in the final node determines the score. The validation cutoff flags patients in every leaf, despite the different leaf scores.

5 Conclusion

This analysis connects the data science life cycle to predictability, computability, and stability (PCS) [3]. In the data and reality part, the study design, codebook, cleaning choices, and published-count comparison connect recorded measurements to the clinical setting. In the algorithms and models part, the descriptive comparisons and two classifiers turn the measurements into findings and scores. The test results assess prediction, the reproducible code supports computability, and the GCS restrictions assess stability. Future data and reality would require checking these conclusions on new patients and in other hospitals before considering clinical use.

The total dataset was manageable for these simple analyses. The main size constraint was the small number of ciTBI cases, especially within rare signs and the GCS 14 training group. A few cases can change a subgroup rate or sensitivity by a large amount, so the large overall patient count does not remove uncertainty.

There is no one-to-one correspondence between these data and clinical reality. Symptoms can be recorded differently, parents and clinicians can disagree, some responses are unavailable, and not all children receive imaging. The cohort also depends on a choice between recorded GCS totals and component sums.

The two findings suggest that symptom detail and the pattern of missing answers deserve attention when studying ciTBI. However, neither finding establishes a clinical rule, and the modeling results show a difficult tradeoff between missed injuries and the number of patients flagged. Better prediction would require further assessment of these choices and validation beyond this dataset.

6 Academic honesty statement

Dear Bin,

Academic honesty matters to me because dishonest or misrepresented work makes academic communication and scientific progress less efficient. I want to contribute to making the world better, so I feel that it is my responsibility to be academically honest.

Claude was used during this work for explanations, coding help, revision, and report assistance. Claude Code does not have a public share link, but I have exported one of the conversations under the "other" folder in the repository.

7 Collaborators

I had no human collaborators. AI assistance is described above.

8 Bibliography

[1] Kuppermann, N., Holmes, J. F., Dayan, P. S., et al. (2009). Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. *The Lancet*, 374, 1160–1170. doi:10.1016/S0140-6736(09)61558-0.

[2] Pediatric Emergency Care Applied Research Network. *TBI PUD 10-08-2013.csv* and *TBI PUD Documentation 10-08-2013.xlsx*. Public-use data and documentation supplied for Lab 1.

[3] Yu, B., and Barter, R. L. *Veridical Data Science: The Practice of Responsible Data Analysis and Decision Making*. Online edition. `https://vdsbook.com/`.