STAT 215A, FINAL PROJECT
CLASSIFYING CITBI IN YOUTH

FLORICA CONSTANTINE, HYUNSUK KIM, MARK OUSSOREN, SAHIL SAXENA

1. Introduction

Clinically-important traumatic brain injuries, hereafter referred to as ciTBIs, are both common in children and require immediate medical attention, as lack of care can lead to death or permanent disability. However, diagnoses often require a CT scan to confirm the presence of a TBI [3]. In children, this need is problematic, as the radiation from a CT scan can lead to long-term adverse affects[1]; hence, given a child presenting with a potential TBI, it is desirable to find a way to decide whether they actually need a CT scan balanced with the knowledge that forgoing a CT scan in the presence of an actual TBI is to be avoided. In this report, we revisit the data from [7] and derive a new decision rule for identifying which child patients need a CT scan. That is, given details (data) of a patient's injury and condition, we translate these into numerical features that we may feed into a statistical model to predict the need for a CT scan.

In this report, we begin by discussing the data in section 2. We discuss its collection in section 2.1 and some of its details in section 2.2. We perform a lengthy exploratory data analysis (EDA) in section 2.2. In section 3, we fit models to the data and interpret the results. We discuss the baseline model from [7] in section 3.1 and results in section 3.2. We perform an analysis with an alternative data split (explained later) in section 3.4. Finally, we present conclusions in section 4.

2. Data

2.1. **Data collection.** The authors in [7] collected data in a prospective cohort study from 43,499 patients younger than 18 years of age that visited a hospital within 24 hours of experiencing head trauma. The study was run across 25 pediatric emergency departments over a span of approximately 2 years, where the last few months were used to collect samples for validating the decision rules derived in the original study. Only patients with GCS (Glasgow Coma Scale) scores of 14 or 15 were considered; those with scores 13 or less were enrolled but were not grouped with the others. For each patient, a trained investigator or other medical personnel recorded various prespecified details, e.g., mechanism of injury, medical history, and responses to standardized questions about the presence of several symptoms or signs of head trauma on a standardized data form.

For a small subset of patients (approximately 4%), a second assessment was performed for quality control purposes–note that we do not use this information, but that its presence is reassuring. Note that there will likely be uncaught entry errors or errors arising from incorrect patient reporting in the data, as well as subjective biases in reporting (e.g., what constitutes a severe injury might differ among physicians and between physicians, parents, and child patients). All of these are sources of randomness in the data. Moreover, there will be natural differences arising from the different hospitals and the different populations that they serve—we are not

Date: 2021 December 10 Friday.

[1]Per our clinical collaborators, we note that medical practitioners say that children are at greater risk for long-term adverse affects from radiation because they have a longer life expectancy (years left to live) than adults. I.e., their primary rationale is that of patients having longer to live, as opposed to something inherent different about childrens' reaction to radiation.

privy to this information, but it is a source of potential batch effects. Nonetheless, we believe that apart from age groupings (discussed later), all of the data may be analyzed together—we cannot correct for this unavoidable randomness, and this is the best that we can do[2]. Moreover, each sample in the data comes from a unique person-event combination, that is, there are no repeated samples or temporal dependencies that we know of. It is certainly possible that a patient is in the data twice, but we believe that this is likely a very rare event, if it occurs at all.

2.2. **Data Feature Meaning.** In this section, we discuss some of the important variables and features in the dataset.

The study defined the following as a positive, ciTBI outcome: death from a TBI, a hospital admission of over 2 days following a diagnosis of TBI from a CT scan, intubation for over 24 hours due to head trauma, or the need for neurosurgery following a CT scan. Other patients were assigned to the negative outcome; to find missed positives, the study coordinators performed telephone surveys to follow up with parents and tracked followup visits. If a positive outcome was missed, the patient's label was updated to positive.

An important variable in this data set is the GCS score. The Glasgow Coma Scale (GCS) is a common scoring system used in emergency departments to determine a patient's level of consciousness by rating their ability to pass certain tests for eye and motor movement along with verbal ability [14]. The scores from each of these three categories are summed to form a total GCS score, valued in the range 3-15. The lower the score a patient has in each category leads to a lower GCS total score (meaning the worse a state a patient is in).

Note that several variables or descriptors in the study require the ability to converse with the child for assignment, e.g., the presence of a headache or whether or not the child is suffering from amnesia. Similarly, a GCS score for a pre-verbal child is also calculated by slightly different metrics than those for an adult. Judging verbal ability, especially, is different with a condition like 'inappropriate words' being instead assessed as 'cries of pain' for those children who are pre-verbal. Even motor ability has some conditions assessed differently such as looking for 'spontaneous movement' rather than 'follow commands' in preverbal children. Hence, as in [7], we chose to separate patients under the age of two (presumed pre-verbal) from those aged two or older (presumed verbal) in our analysis. Moreover, as children under the age of two are considered more at risk for long-term adverse effects from radiation, it is reasonable to consider this group separately [3]. Note that this is not a perfect grouping as some children will be verbal by age two and some children are not verbal after age 2 (we look into this in our stability analysis in section 3.4), but, nonetheless, it is good developmental benchmark [1].

The variables in our data are all categorical or ordinal, except for age (however the categorial version of the age variable where it was discretized by < 2 years old and ≥ 2 years old was used in all our analyses for the reasons stated above). While it would be ideal if instead the continuous version of the variables were reported and they were not pre-sorted into sometimes arbitrarily chosen categories (i.e. the length of a seizure is binned as < 1 min, $1 - 5$ min, $5 - 15$ min, and > 15 min), we are restricted to the categorical data.

Several binary indicator variables exist in the dataset, looking at, respectively, whether a child suffered a loss of consciousness, seizure, headache, vomiting, altered mental state, palpable skull fracture, basilar skull fracture, hematoma, trauma above the clavicles, neurological deficits, or other (non-head) substantial injuries. Each of these variables also has more specific follow up questions, e.g. the type of basilar skull fracture if it is indicated a patient has one. Other important variables included in our data set are the injury mechanism, injury severity, and whether the child is acting normally, is intubated, is paralyzed, and/or is sedated.

[2]Indeed, we might prefer having some noise in the training data, as it may in fact improve generalization performance of our models [5].

Lastly, we also have several meta variables such as patient number, race, ethnicity, gender, position of medical professional, and certification of medical professional. These variables do not affect whether a patient will be positive for ciTBI. However, they may be useful to look at after our analyses are complete in case they are acting as a proxy for something deeper that is taking place but should not be used as feature inputs to our models.

2.3. **Exploratory Data Analysis.**

2.3.1. *Outcome.* First, we looked at our outcome variable. Recall that the study defined the following as a positive, ciTBI outcome: death from a TBI, a hospital admission of over 2 days following a diagnosis of TBI from a CT scan, intubation for over 24 hours due to head trauma, or the need for neurosurgery following a CT scan. That is, the presence of any of the four sub-outcomes constituted a positive outcome. The lack of all four sub-outcomes constituted a negative outcome. In the data, we noted that there were 20 patients that had a missing value for the final outcome: this is a discrepancy with [7], wherein 18 rather than 20 patients have a missing value. This difference could not be resolved. Of these 20 patients, 17 of them are negative for all four of the sub-outcome variables making up our outcome. We thus assign these patients as being negative for a ciTBI. For the three remaining patients, they had missing values for one or more of the four outcomes and based on clinical guidance, were dropped. The proportion of each of the four sub-outcomes is shown below in Figure 1. We can see that the vast majority of people were positive for a prolonged hospital stay.

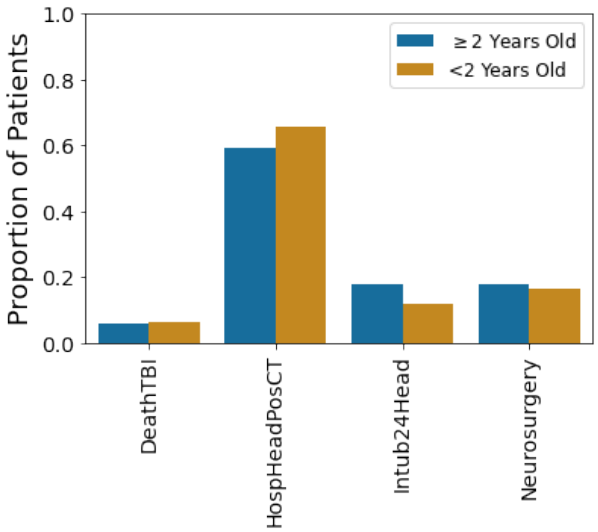
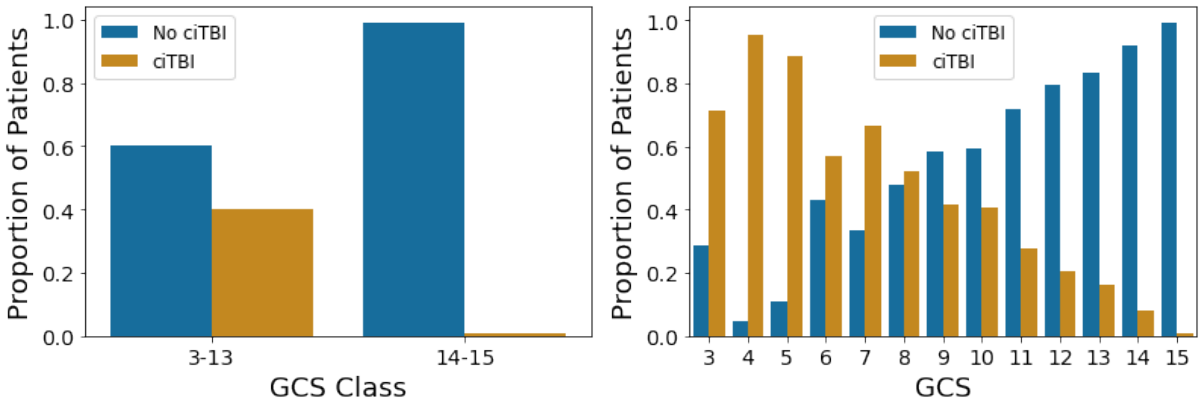


FIGURE 1. Outcome type for ciTBI patients by age group

2.3.2. *GCS Scores.* From [7], we learned that it is not controversial to perform a CT scan for patients with a GCS score ranging from 3 to 13 as in this group the risk of finding a TBI on a CT is more than 20%. For our data set, we looked at the proportion of patients positive for ciTBI with a GCS scores in the range of 3-13 and also for those in the range for 14-15 in Figure 2a. Looking at this, we can see that 40% of patients with a GCS score in the range of 3 to 13 were positive for ciTBI versus only 0.8% of those with a GCS score of 14 or 15—this is quite a dramatic difference. However, we wanted to know if separating the GCS score into classes with a cutoff GCS score of 14, in particular, was the best possible split. We broke up the previous plot further into individual GCS scores (Figure 2b). We can see that, in general, the lower the GCS score the higher the proportion is for a patient to be positive for ciTBI—as expected. Even at a GCS of 13, 20% of patients were positive for ciTBI. Thus, keeping the current cutoff of 3-13 and 14-15 as the two separate GCS classes seems reasonable. Hence, we remove any patients that have a GCS in the range of 3-13 (969 total patients), as the risk of having a positive ciTBI is too high and any reasonable or acceptable (to a practitioner) decision rule would suggest always performing a CT scan for this group.



(A) Proportion of patients with ciTBI by age and GCS Class

(B) Proportion of patients with ciTBI by age and GCS Score

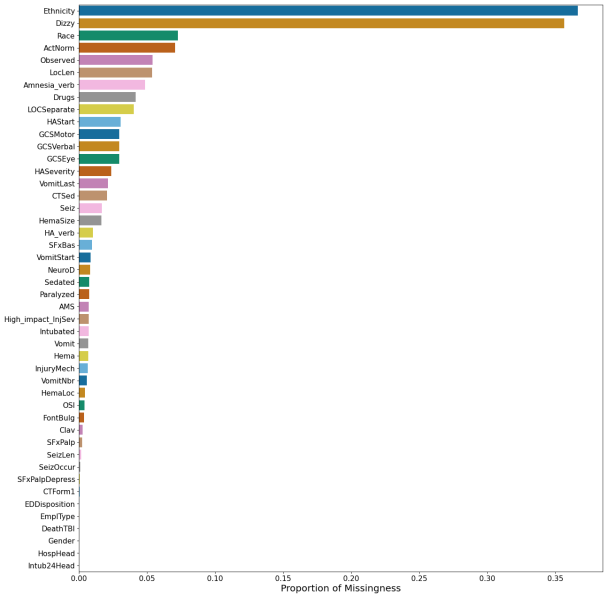
FIGURE 2. GCS Score and ciTBI

2.3.3. *Data Missingness.* Next, we look at the rate of missingness for each feature in Figure 3a. We note that the features 'Dizzy' and 'Ethnicity' are missing in more than 35% of patients. On the data form, ethnicity asks whether the patient is hispanic or not and may potentially be skipped over by a patient if they fill in the race field instead (or if they are too young to fill out a form and the medical personnel does not want to guess). However, we are already considering ethnicity to be a meta variable and did not use it in our analyses. After speaking with the clinicians, we learned that notating whether a patient is dizzy or not is not very relevant in diagnosing TBI and it is also a very subjective variable: it is highly susceptible to change from patient to patient based on their own personal definition of feeling dizzy. Thus, as there is no objective way to compare or impute this variable, we decided to drop it.

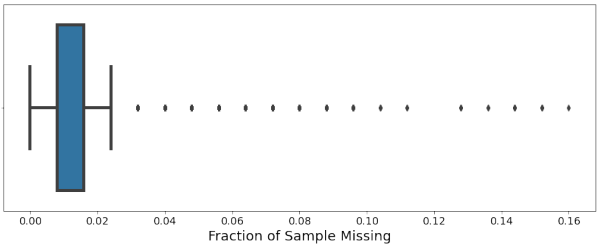
For the other variables with missingness, they were either imputed with what a 'healthy' response would be, e.g. a 'No' value would be imputed for missing paralyzed or intubation values—this choice was based on clinical guidance, as in the view of practitioners, serious events like paralysis or intubation are unlikely to be left unnotated. Otherwise, the response that was the mode was used for variables where there was no clear 'healthy' response. e.g. hematoma size. We note that, per Figure 3a, all features used in our analysis have a missingness under 10%, and hence we believe that imputation likely has a minor effect on our results—even if some values were actually positive (rather than negative or healthy).

Many variables have a parent question such as 'Seiz' for seizure that have follow up question such as the length of the seizure. If a patient has a response of 'No' for seizure then in the form 'Not applicable' is often marked for each follow up question. Without loss of generality, we convert these 'Not applicable' answers to be 'No' to make analyses easier to perform.

We further note that the majority of patients have only around 1% of data features missing, and are at maximum still under 20% (Figure 3b) and thus we do not drop any patients from our analyses.



(A) Fraction of samples missing a given feature



(B) Fraction of entries missing within a sample

FIGURE 3. Missingingness in the data

2.3.4. *Age Class Cutoff.* The age was a major factor in [7] for creating a decision rule. Two rules were created based on age categories of < 2 and ≥ 2 years of age. We can see a large portion of the patient population in our data set is younger and around 2 years of age in Figure 4.

Besides radiation exposure risks, one reason to demarcate age at 2 years is because of verbal ability: below this age, children typically do not talk coherently if at all. We wanted to check the number of pre-nonverbal subjects at each age to see if two years old is actually a good cutoff age for being verbal. In Figure 5, we can see that actually there are still a large proportion of subjects that are pre-verbal at ages 2, 3, and even 4 when calculated based on responses for whether a patient had a headache or amnesia in the data. Both of these features are the closest proxy we have to knowing how many pre-verbal patients are in our data set, as a binary variable for being pre-verbal does not exist. It is reassuring that the proportions between the two for each age are extremely similar.

We will revisit this division in section 3.4, where we explore the results of fitting models for different age splits.

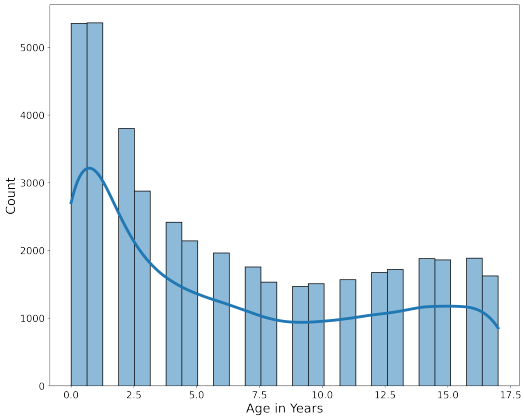
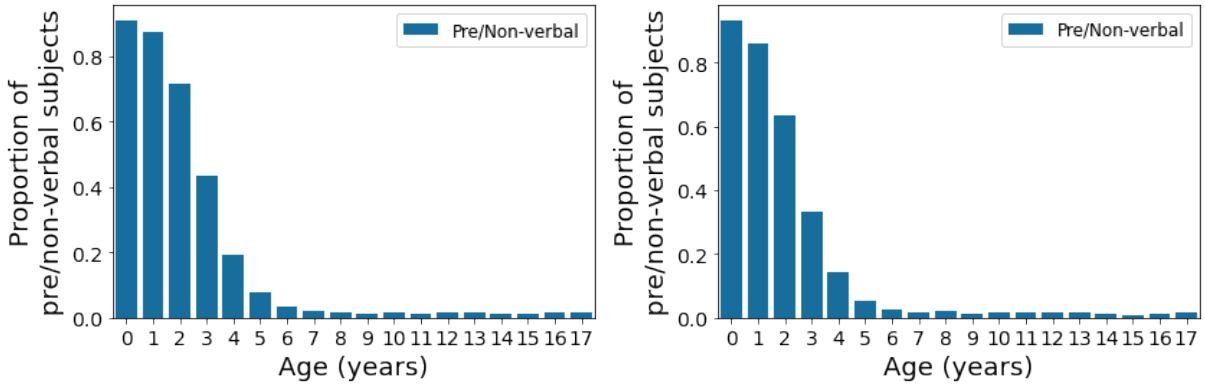


FIGURE 4. Age Distribution

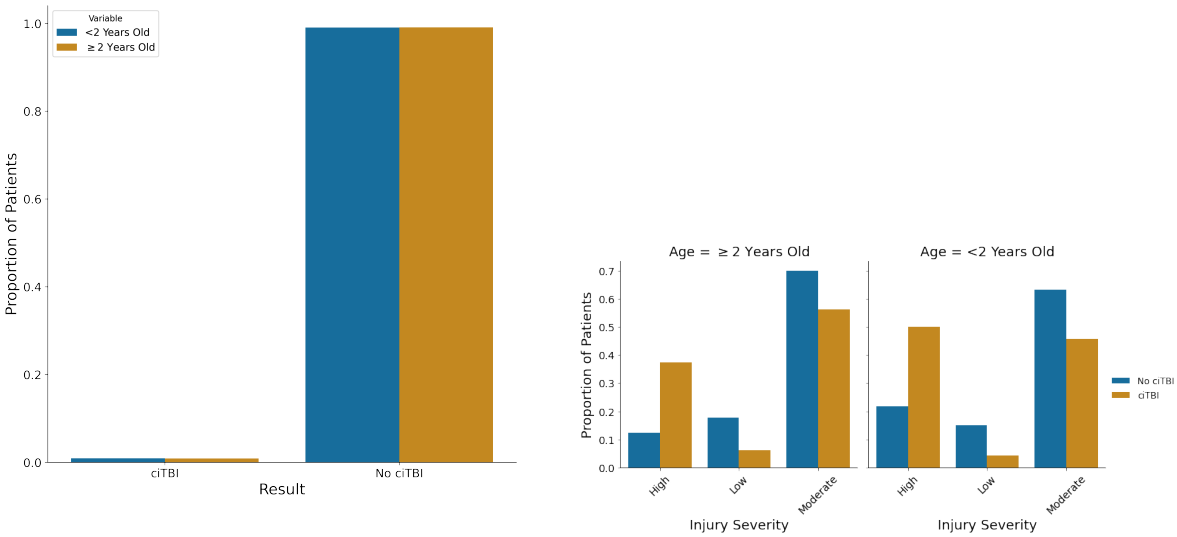


(A) Preverbal response to Amnesia question (B) Preverbal response to Headache question

FIGURE 5. Preverbal responses to Amnesia and Headache questions

2.3.5. *Distribution of Features by Age.* Next, we look at the occurrence of ciTBI in each of our two age categories in Figure 6a. We can see that the proportion of ciTBI in each age category is very close to being the same. The proportions looking at injury severity in Figure 6b are similar across age category.

However, there may still be other variables with different proportions of positive ciTBI across age categories in Figure 7. That is, for each age category and for each outcome, we look at the proportion of patients with the indicated symptom. This exercise might be indicative as to whether such a variable would potentially lead to a different decision rule between the two groups. We can see that the proportion of patients with ciTBI are noticeably different between age < 2 and age ≥ 2 for 'Vomit' and 'OSI' (other non-head injury). Also, we note that the variables measuring amnesia and headache cannot be answered by those that are pre-verbal and thus may be useful in a decision rule for those over age 2 but not under.



(A) Proportion of outcomes grouped by age (B) Proportion of injury severity grouped by age

FIGURE 6. Outcome and Injury Severity

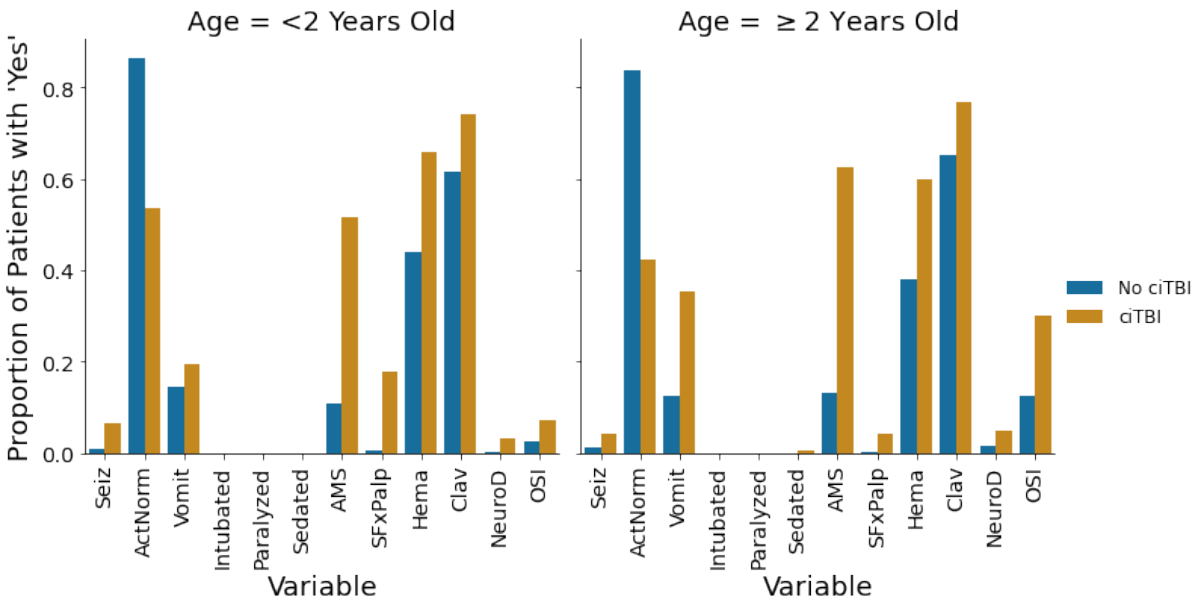


FIGURE 7. Proportion of positive feature identifiers by outcome (ciTBI) and age

2.3.6. *Correlation of Features to Outcome.* Next, we examine whether any of the features are particularly correlated to the outcome by calculating the Spearman's ρ coefficient on the ordinal variables against the binary outcome. We can see in Figure 8 that none of the features are well correlated with the outcome. A maximum correlation coefficient of 0.12 is attained by the altered mental state feature.

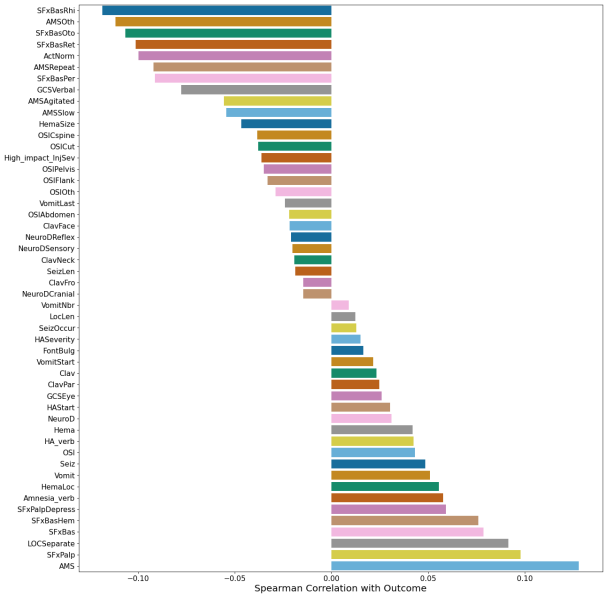


FIGURE 8. Spearman Correlation of Features to Outcome

2.3.7. *Principal Component Analysis.* We perform Principal Component Analysis (PCA) on the one-hot encoded data [8]. In Figure 9, we see that the nearly all of the variance is explained by the first 100 components, and that the first few components capture most of the variance (the first two components explain 13%, the first five explain 30%, and the first twenty explain 50% of variation in the data). That is, noting that the PCA eigenvalues (variances) decay rapidly, we might believe that this dataset behaves like a low-rank signal plus noise.

In Figure 10, we project the one-hot encoded data to two dimensions to study if the classes (age and outcome) are visually separable. First, we see that the classes do not separate, but that there are two distinct clusters in the data—since the data were taken from 25 hospitals and that there was no laboratory or experimental processing of the data, we did not suspect a batch effect (there would only be two potential batches). Instead, we see that the presence of an OSI (other, non-head-related injury) leads to the two clusters. Note that the prevalence of OSI in the data is low (10%), but that it is enough to strongly affect the results of PCA. We will keep this in mind when doing our analyses and look to see if the majority of the misclassified points come from patients who had an OSI injury. This means we may want to consider forming a separate decision rule for this subgroup of the patient population.

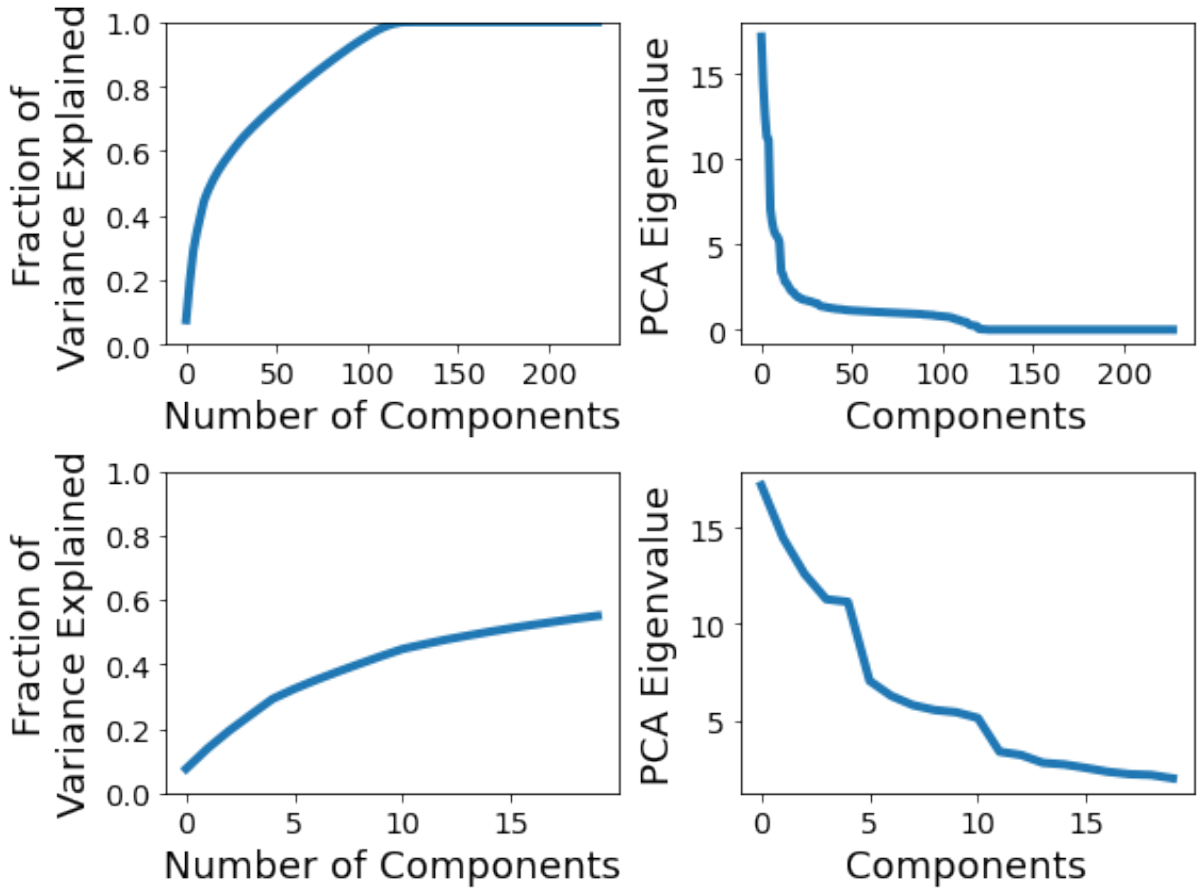
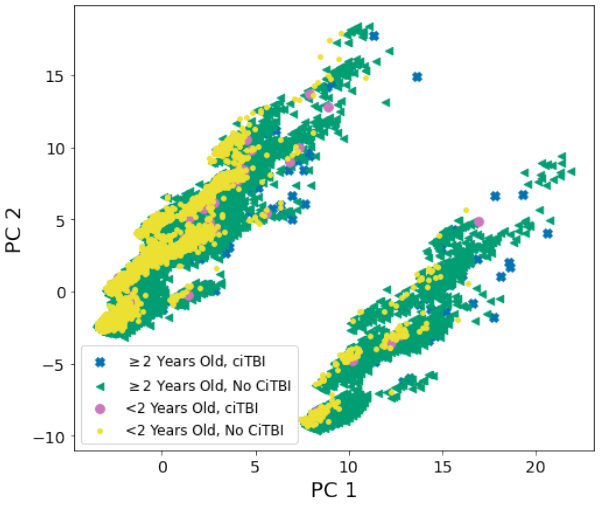
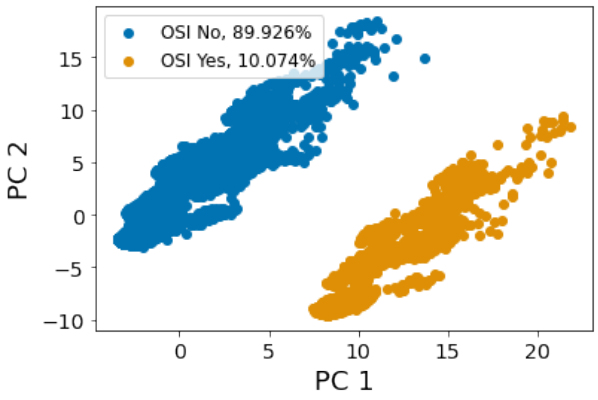


FIGURE 9. PCA: Cumulative Variance Explained and PCA Eigenvalues; the second row is a zoomed-in version of the top row.



(A) 2-dimensional PCA projection colored by Age and Outcome



(B) 2-dimensional PCA projection colored by OSI

FIGURE 10. 2-dimensional PCA projection and two natural clusters

3. Modeling: Results and Discussion

In this section, we fit models to predict whether patients have a ciTBI outcome from their covariates. We consider several models: ℓ_1-penalized logistic regression [12], group ℓ_1-penalized logistic regression [12], a single decision tree [10], a random forest [6], AdaBoost [11], LogitBoost [4], and a linear SVM [2]. All of these models were chosen for their relative ease of implementation and interpretability. We focus on the metric of sensitivity as, per our clinical collaborators, clinicians will only accept or trust decision rules with $\sim 0.1 - 1\%$ missed positives (ciTBI cases) in children, and on NPV as we want to minimize unneeded CT scans. Note that minimizing unneeded CT scans minimizes unneeded exposure to long-term adverse effects from radiation, and also increases patient throughput and hospital efficiency. As all of these algorithms have one or more parameters that can be tuned, based on practitioner feedback, we chose to operate at a point wherein the sensitivity was at least 0.99 (or as close to it as possible) and the negative predictive value (NPV) was as close to 1 as possible. However, we found that operating at a point corresponding to 0.99 validation sensitivity lead to poor generalization on the test set (to be discussed later), and present results corresponding to a validation sensitivity of 0.95. We also present results corresponding to a validation sensitivity of 0.99 for comparison.

We continue with the data split of age < 2 and ≥ 2 and fit models for each group; we also fit models for the entire, unsplit dataset. The unsplit models can be thought of as a stability and reality check for whether the age-based demarcation is significant and necessary, as our exploratory data analysis in sections 2.3.4 and 2.3.5 indicate that it may not be. In section 3.4, we study the effect of a different data split (pre-verbal v. verbal) and compare the results with the age-2 split.

3.1. **Baseline Model.**

Before proceeding, we briefly describe the model from [7], hereafter referred to as the baseline model. The model consists of two decision trees: one for patients with an age under 2 years, and one for patients aged two years or older. The tree for patients younger than 2 consists of questions about the presence of: an altered mental status (AMS), scalp hematoma, loss of consciousness for greater than 5 seconds, a severe cause of injury, a palpable or possible skull fracture, and abnormal behavior per a parent. The questions are asked in the given order, and if all answers are 'no', no scan is recommended. For patients older than 2, the tree consists of questions about the presence of: an altered mental status (AMS), loss of consciousness, a history of vomiting, a severe cause of injury, a palpable or possible skull fracture, signs of a basilar skull fracture, and the presence of a severe headache. Once again, all 'no' answers leads to no scan.

3.2. **Results.**

In Table 1, we present results for all of the algorithms on the validation set. That is, we trained all algorithms on the training set for a wide variety of parameters (if there were any), selected an appropriate operating point/threshold as described above on the validation set, and have summarized the selected operating points for each algorithm. We see that ℓ_1-penalized Logistic Regression has the highest AUC while having a sensitivity close to 0.95 and an NPV close to 1. Moreover, relative to other algorithms with similar characteristics (e.g., the Group ℓ_1-penalized Logistic Regression and AdaBoost), we see that the specificity is much higher. Hence, we selected ℓ_1-penalized Logistic Regression as our 'best' method. Moreover, we see that this method performs better than the baseline algorithm: we have better validation AUCs as well as better specificities and comparable NPVs at similar sensitivities. In general, we see that the models trained on the unsplit (by age) data perform slightly worse than both of the models trained on the individual halves.

We note that the group ℓ_1-penalized Logistic Regression performed almost as well as the ℓ_1-penalized Logistic Regression. However, this method is extremely sensitive to the regularization parameter, and we suspect that slight changes in the data used for training would lead to vastly

Algorithm	Age	AUC	Accuracy	Sensitivity	Specificity	NPV	Balanced Accuracy
ℓ_1-penalized Logistic Regression	young	0.938	0.764	1.0	0.762	1.0	0.881
ℓ_1-penalized Logistic Regression	old	0.931	0.751	0.957	0.75	1.0	0.854
ℓ_1-penalized Logistic Regression	all	0.917	0.75	0.952	0.748	0.999	0.85
Group ℓ_1-penalized Logistic Regression	young	0.908	0.728	1.0	0.726	1.0	0.863
Group ℓ_1-penalized Logistic Regression	old	0.917	0.68	0.957	0.678	1.0	0.818
Group ℓ_1-penalized Logistic Regression	all	0.917	0.745	0.952	0.743	0.999	0.848
AdaBoost	young	0.781	0.064	1.0	0.058	1.0	0.529
AdaBoost	old	0.872	0.25	0.957	0.245	0.999	0.601
AdaBoost	all	0.899	0.59	0.952	0.587	0.999	0.77
LogitBoost	young	0.825	0.889	0.714	0.891	0.998	0.802
LogitBoost	old	0.814	0.213	0.957	0.208	0.998	0.583
LogitBoost	all	0.746	0.198	0.952	0.191	0.998	0.572
Decision Tree	young	0.898	0.118	0.929	0.113	0.996	0.521
Decision Tree	old	0.875	0.724	0.957	0.722	1.0	0.84
Decision Tree	all	0.809	0.01	0.988	0.0	0.75	0.494
Random Forest	young	0.815	0.844	0.714	0.845	0.998	0.779
Random Forest	old	0.889	0.796	0.894	0.795	0.999	0.844
Random Forest	all	0.845	0.816	0.798	0.816	0.998	0.807
Linear SVM	young	0.275	0.014	1.0	0.008	1.0	0.504
Linear SVM	old	0.645	0.057	0.957	0.051	0.994	0.504
Linear SVM	all	0.644	0.063	0.952	0.054	0.991	0.503
Baseline	young	0.903	0.545	1.0	0.542	1.0	0.771
Baseline	old	0.869	0.615	0.957	0.613	0.999	0.785

TABLE 1. Algorithm performance on validation data for each data split

different results. This result is unfortunate, as in principle, the grouping would allow us to enforce sparsity across a group of covariates (e.g., everything vomit related) and hence improve interpretability. We note that the decision tree also performed well (recall that the baseline model is also a decision tree), but that the logistic regression was better. Also, decision trees can heavily depend on the training data in ways that regression models do not: they are more complex models and create multiple decision boundaries whereas logistic regression creates only one. The random forest and boosted models (AdaBoost and LogitBoost) do not perform as well; we noticed that the performance was not linear in the number of trees, and conjecture that there may be some degree of overfitting on the training data. Either way, an ensemble model is naturally harder to interpret than a linear model. The SVM performed quite poorly, in contrast—it is somewhat surprising that a logistic regression method performs well where a linear SVM does not, but we chose not to investigate further given time and space constraints.

Hence, we summarize results for ℓ_1-regularized Logistic Regression and the baseline models on the test set in Table 2. We found that the regularization parameter for both the young and old models was ≈ 0.336 and that the parameter for the model trained on the unsplit data was ≈ 1.129. We see that the test sensitivity is close to 0.95 and that the NPV is still close to 1, and that the AUC is close to 0.85; these numbers are a slight drop from the validation results, but are still good—in particular, they are comparable to the baseline model, with a better specificity. Once again, we see that the model trained on the unsplit (by age) data performs slightly worse than both of the models trained on the individual halves. Recalling Figure 10, wherein we saw a significant clustering effect based on OSI, we note that OSI and classification accuracy are not well correlated.

Algorithm	Age	AUC	Accuracy	Sensitivity	Specificity	NPV	Balanced Accuracy
ℓ_1-penalized Logistic Regression	young	0.846	0.772	0.923	0.77	0.999	0.846
ℓ_1-penalized Logistic Regression	old	0.848	0.762	0.937	0.76	0.999	0.848
ℓ_1-penalized Logistic Regression	all	0.822	0.794	0.85	0.793	0.999	0.822
Baseline	young	0.875	0.554	1.0	0.549	1.0	0.774
Baseline	old	0.873	0.639	0.937	0.636	0.999	0.786

TABLE 2. Algorithm performance on test data for each data split

3.3. **Discussion of the ℓ_1-regularized Logistic Regression model.** In this section, we provide some insights from studying the logistic regression models that we have fit. We note that this form of model is a good choice, as it is naturally interpretable: the ℓ_1-penalization leads to naturally sparse coefficient vectors, so that only a subset of features are used in prediction. The sparsity combined with the linear nature of the classifier means that the coefficients' magnitudes have meaning (our data is one-hot encoded, so each coefficient is easily interpretable as a contribution of that feature), and the form of the model means that the odds ratio is a linear function of the data—this model is hence easy to use.

In Figure 11, 12, and 13, we provide computed feature importances from the model. That is, we report the magnitude of coefficients times the standard deviation (on the validation set) of the features. We see that across all of the data splits, AMS (altered mental state) is an important variable, as are various features related to vomit, hematoma location, and loss of consciousness.

In Figure 14, we present ROC curves for the ℓ_1-regularized Logistic Regression for the validation and the test data. We see that all of the logistic regression models are better than the baseline model, and that the ROC curves are far from the 45-degree line.

Finally, in Figure 15, we study the validation AUC as a function of the regularization strength. We see that in a neighborhood of the chosen value ($[10^{-1}, 10^{1/2}]$), the AUC is relatively stable and high. That is, perturbations to the regularization parameter or not searching on a fine enough grid are not concerns in our analysis.

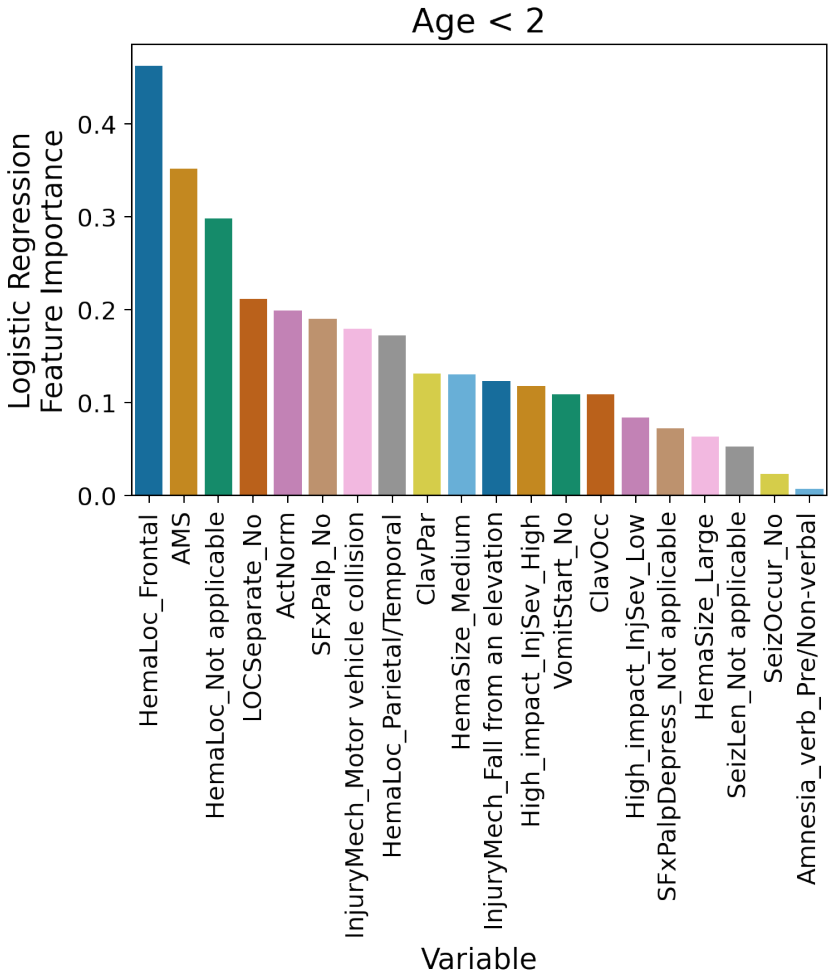


FIGURE 11. Feature importances from the ℓ_1-regularized Logistic Regression for young patients

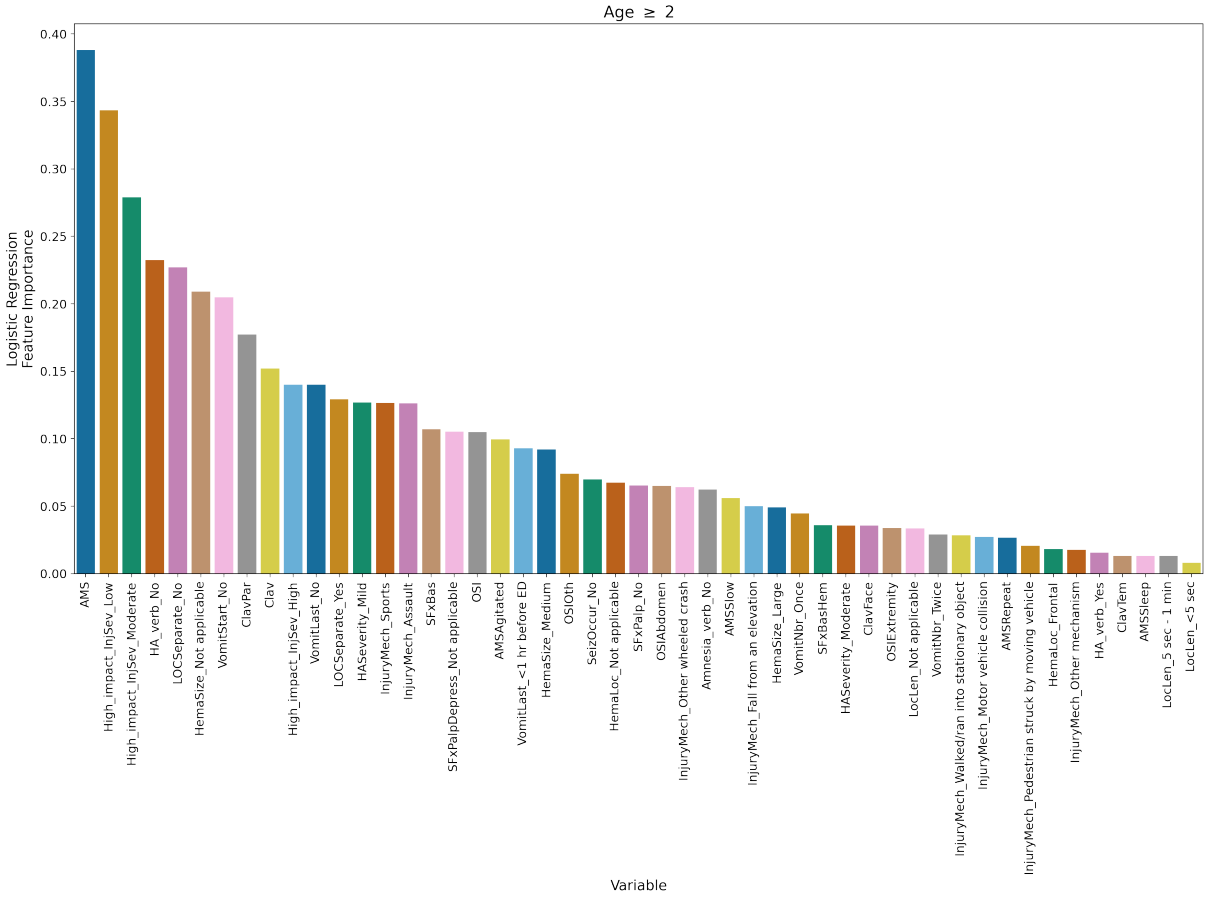


FIGURE 12. Feature importances from the ℓ_1-regularized Logistic Regression for older patients

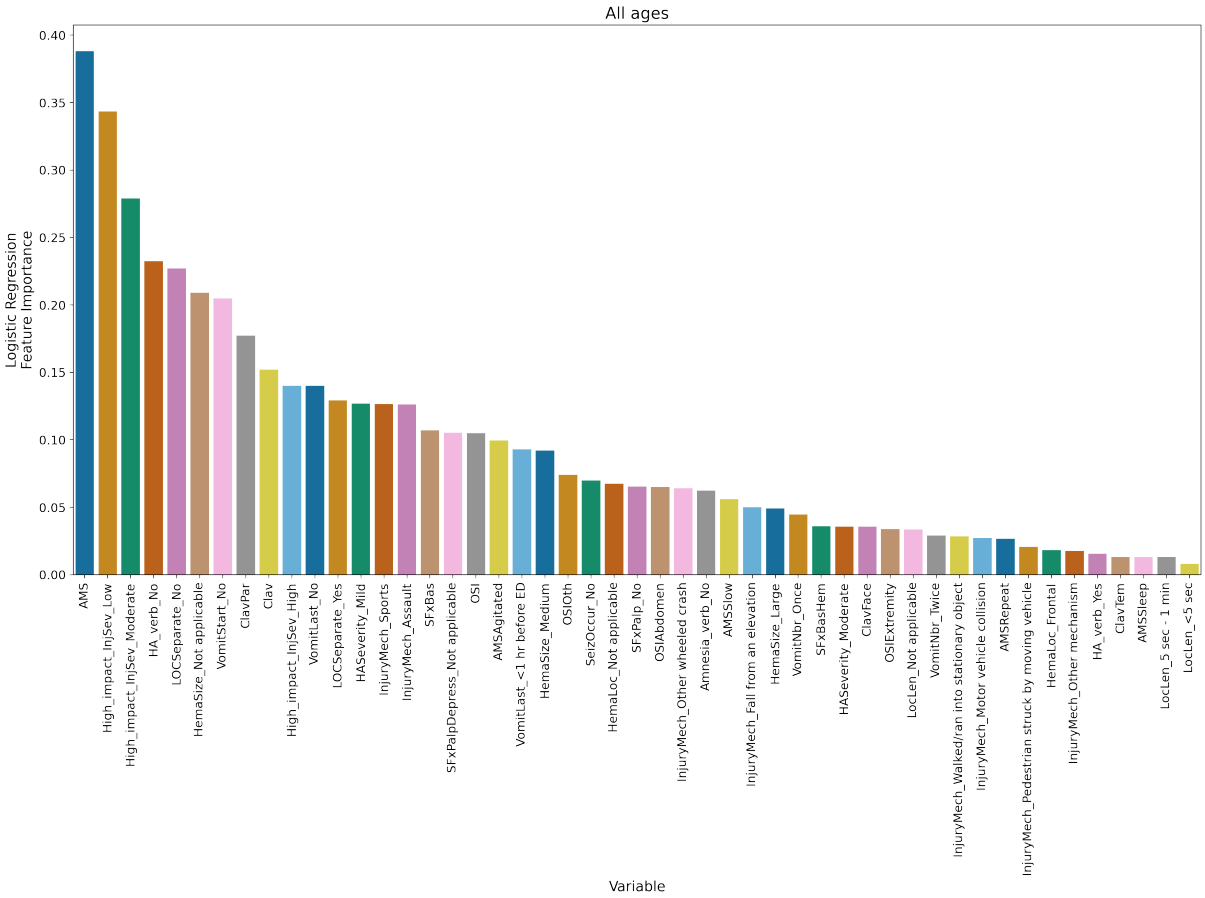
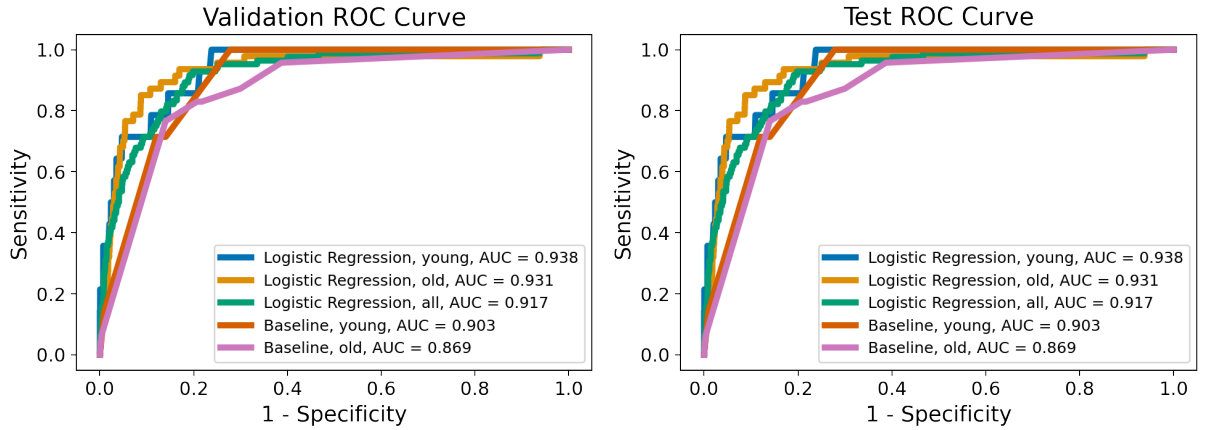


FIGURE 13. Feature importances from the ℓ_1-regularized Logistic Regression for all patients



(A) ROC curve on the validation data for the ℓ_1-regularized Logistic Regression model

(B) ROC curve on the test data for the ℓ_1-regularized Logistic Regression model

FIGURE 14. ROC curves for the ℓ_1-regularized Logistic Regression model

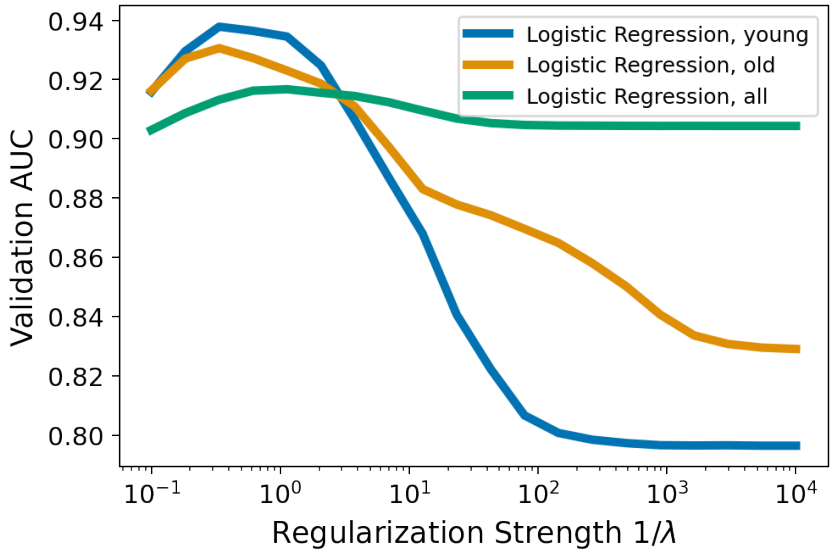


FIGURE 15. Validation AUC for the ℓ_1-regularized Logistic Regression as a function of regularization strength

3.3.1. *Model stability.* In this section, we report the results from bootstrapping the training dataset and evaluating the bootstrapped logistic regression models on the validation set. We measure the validation AUC as well as track which regularization parameter led to the best model. Ideally, we would see a tight concentration close to and around the previously observed values. As this step involves a parameter sweep, we used 20 bootstrap samples of size equal to a quarter of the original dataset and searched over 20 values of the regularization parameter. Our results appear in Figure 16, where we see that all of the AUCs have relatively tight concentration around the previously observed values. Moreover, the chosen regularization parameters are still close to the values picked for the chosen model. We present the mean and standard deviation of the bootstrapped validation AUCs in Table 3: the values are consistent with what we saw earlier on the full dataset.

Age	Mean AUC	Standard Deviation of AUC	Model AUC (full data)
Young	0.909	0.0172	0.938
Old	0.905	0.0124	0.931
All	0.892	0.009933	0.917

TABLE 3. Bootstrapped mean and standard deviation of validation AUCs for the ℓ_1-regularized Logistic Regression model; note that the 20 bootstrapped datasets were of smaller size (one quarter) than the original data.

For comparison, we present bootstrapped AUCs for the baseline model on the validation dataset. We report results from 50 bootstrapped datasets with sample sizes equal to that of the original dataset. We present our results in Table 4 and Figure 17: the values are consistent with what we saw earlier on the full dataset.

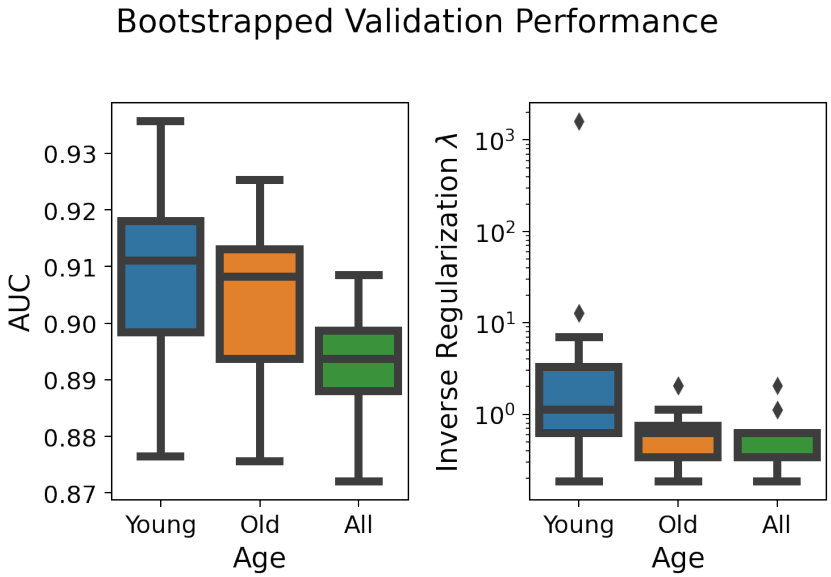


FIGURE 16. Bootstrapped validation AUC for the ℓ_1-regularized Logistic Regression and the corresponding regularization parameters. Recall that the young and old models previously had parameters of 0.336 and the all model had a parameter value of 1.129.

Age	Mean AUC	Standard Deviation of AUC	Model AUC (full data)
Young	0.870	0.0211	0.903
Old	0.902	0.0191	0.869

TABLE 4. Bootstrapped mean and standard deviation of validation AUCs for the baseline model

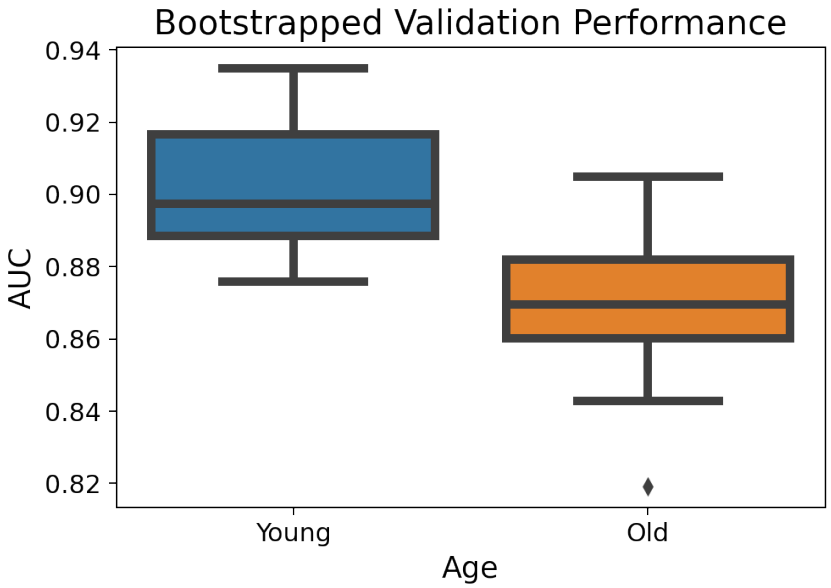


FIGURE 17. Bootstrapped validation AUC for the baseline model.

3.3.2. *A different threshold.* In this section, we look at the effect of setting the operating point based on a different threshold. Elsewhere in this report, we choose an operating point such that the validation sensitivity is as close to 0.95 as possible. Here, we look at choosing an operating point such that the sensitivity is as close to 0.99 as possible. This choice is motivated by discussions we had with our clinical practitioner contacts.

In Table 5, we present results on the validation data. These results are comparable to those with the previous threshold, though the accuracy on the all model is slightly worse. In Table 6, we present results on the test data. These results are noticeably worse, especially for the old and all models: we conjecture that the high sensitivity constraints set on the validation data lead to poor generalization on the test data, especially given the low prevalence of ciTBI cases in the data (0.8%).

Algorithm	Age	AUC	Accuracy	Sensitivity	Specificity	NPV	Balanced Accuracy
ℓ_1-penalized Logistic Regression	young	0.938	0.763	1.000	0.761	1.000	0.881
ℓ_1-penalized Logistic Regression	old	0.931	0.0683	1.000	0.0614	1.000	0.531
ℓ_1-penalized Logistic Regression	all	0.917	0.0685	1.000	0.0593	1.000	0.530

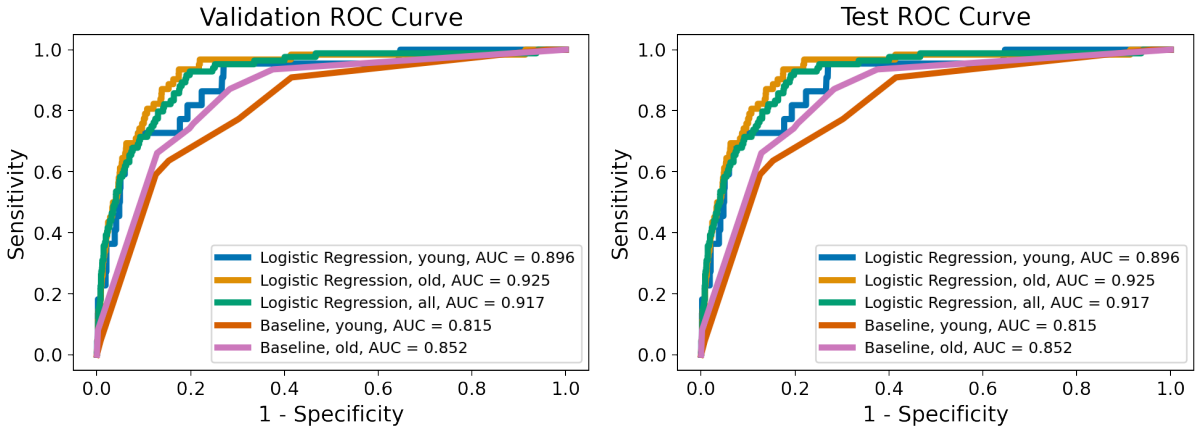
TABLE 5. Algorithm performance on validation data for each data split with a differently set threshold (0.99 sensitivity)

Algorithm	Age	AUC	Accuracy	Sensitivity	Specificity	NPV	Balanced Accuracy
ℓ_1-penalized Logistic Regression	young	0.846	0.771	0.923	0.769	0.998	0.866
ℓ_1-penalized Logistic Regression	old	0.530	0.0703	1.000	0.0610	1.000	0.530
ℓ_1-penalized Logistic Regression	all	0.587	0.197	0.983	0.191	0.999	0.587

TABLE 6. Algorithm performance on test data for each data split with a differently set threshold (0.99 sensitivity)

3.4. **A different data split: Pre-verbal v. verbal.** In this section, we return to Figure 5, where we saw that many patients under the age of 5 were pre-verbal, but the standard practice is to separate subjects before and after the age of 2. Moreover, there were relatively few patients over the age of 5 that were not verbal (104 subjects in validation), so we conjecture that a rule split based on pre-verbal v. verbal status might be a better choice. Using the same train/validation/test splits, we re-trained our ℓ_1-regularized Logistic Regression models.

In Figure 18, we present ROC curves for the ℓ_1-regularized Logistic Regression for the validation and the test data with the new data split. We see that all of the logistic regression models are better than the baseline model, and that the ROC curves are far from the 45-degree line, but that this split leads to slightly worse performance than the original division at the age of 2. Nonetheless, we believe that this or similar data splits merit further investigation: there are children under the age of two that are verbal and can hence communicate their mental status, but there are also those over the age of two that are not and hence cannot communicate. Per our clinical collaborators, childrens' verbal abilities fall along a spectrum, especially before the age of 5, and it is hence difficult to separate patients based on age.



(A) ROC curve on the validation data for the ℓ_1-regularized Logistic Regression model

(B) ROC curve on the test data for the ℓ_1-regularized Logistic Regression model

FIGURE 18. ROC curves for the ℓ_1-regularized Logistic Regression model with the pre-verbal/verbal split

4. Conclusions

In this report, we have looked at patient data from a prospective cohort study wherein patients between the ages of 0-17 who visited one of a series of hospitals presenting with a potential TBI were enrolled. We were given data from patient questionnaires that was converted to numerical features and fit models to predict the need for a CT scan. We found that an ℓ_1-penalized logistic regression model performed the best and was an improvement (better AUC, and comparable NPV and a higher specificity at similar sensitivities) on the model derived in [7]. Interestingly, the model trained on all of the data (unsplit by age) is only marginally worse (in terms of the AUC and other performance statistics) than the individual models trained on each age group: we are not surprised by this, based off our exploratory data analysis. Importantly, our logistic regression model (like the baseline model) can be computed quickly, as it is a simple, linear model. Hence, our model, like the baseline model, can be implemented easily on medical devices to help in practitioners' decision making.

There are many threads left unfinished in this work, mostly because of time and space constraints. First, it would be interesting to systematically investigate whether a better age cutoff (in terms of model performance) could be obtained, or whether an age cutoff combined with some other factor (like pre-verbal v. verbal) would be better. Additionally, it would be good to compare additional models and to investigate the performance and utility of some model explainability techniques, like LIME [9] or SHAP [13]. Finally, it would be important and interesting to understand why performance at an operating point of 0.99 validation sensitivity generalizes poorly, and if there is a better classifier or method that would lead to better results at this operating point. We believe that a larger dataset with more positive samples (more ciTBI cases) would be helpful in this task.

4.1. **Division of Labor.** Hyunsuk Kim contributed to the exploratory data analysis, implemented the ℓ_1-penalized logistic regression, grouped ℓ_1-penalized logistic regression, consolidated everyone's model code into one larger wrapper function, worked on the baseline model and on implementing `baseline.py`, edited and coded functions to find the statistical metrics saved by each of the models, and offered comments on the final report.

Mark Oussoren contributed to much of the exploratory data analysis, did most of the data processing and implementation of dataset.py, worked on coding the baseline model, summarized his findings for EDA and the baseline model, documented the judgement calls made working with this data set, contributed to the `data_dictionary.md` file, and offered comments on the final report.

Sahil Saxena created a slide deck to share with our clinician contact, implemented an SVM model and looked at the effect of different kernels on SVM performance, explored 3D visualizations of SVM and Logistic Regression on the first few principal components of our data, compiled the `data_dictionary.md` file, wrote much of the README file, and offered comments on the final report.

Florica Constantine acted as the project lead overseeing the project, editing others results across all aspects of the project, and also individually worked on much of the exploratory data analysis, implemented the boosting models, implemented the stability analysis, implemented `model_best.py` and some of `baseline.py`, documented judgement calls, contributed to the README file, and wrote the final report.

4.2. **Acknowledgements.** We would like to thank Dr. Aaron Kornblith and Nathan Velarde for their time and guidance, especially in answering our questions about clinical practices and the data.

References

[1] Charles D Blackwell et al. "Pediatric head trauma: Changes in use of computed tomography in emergency departments in the United States over time". In: *Annals of emergency medicine* 49.3 (2007), pp. 320–324.

[2] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152.

[3] David J Brenner. "Estimating cancer risks from pediatric CT: Going from the qualitative to the quantitative". In: *Pediatric radiology* 32.4 (2002), pp. 228–231.

[4] Yu-Dong Cai et al. "Using LogitBoost classifier to predict protein structural classes". In: *Journal of theoretical biology* 238.1 (2006), pp. 172–176.

[5] Nicholas Carlini et al. "On evaluating adversarial robustness". In: *arXiv preprint arXiv:1902.06705* (2019).

[6] Tin Kam Ho. "Random decision forests". In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[7] Nathan Kuppermann et al. "Identification of children at very low risk of clinically-important brain injuries after head trauma: A prospective cohort study". In: *The Lancet* 374.9696 (2009), pp. 1160–1170.

[8] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), pp. 559–572.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[10] S Rasoul Safavian and David Landgrebe. "A survey of decision tree classifier methodology". In: *IEEE transactions on systems, man, and cybernetics* 21.3 (1991), pp. 660–674.

[11] Robert E Schapire. "Explaining AdaBoost". In: *Empirical inference*. Springer, 2013, pp. 37–52.

[12] Noah Simon et al. "A sparse-group LASSO". In: *Journal of computational and graphical statistics* 22.2 (2013), pp. 231–245.

[13] Erik Štrumbelj and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions". In: *Knowledge and information systems* 41.3 (2014), pp. 647–665.

[14] Graham Teasdale et al. "The Glasgow Coma Scale at 40 years: Standing the test of time". In: *The Lancet Neurology* 13.8 (2014), pp. 844–854.

Email address: {florica, hyskim7, mark_oussoren, sahilsaxena18}@berkeley.edu