STAT 215A, FINAL PROJECT: CLASSIFYING CITBI IN YOUTH

MARK OUSSOREN, SAHIL SAXENA, HYUNSUK KIM, FLORICA CONSTANTINE

1. Introduction

Clinically-important traumatic brain injuries, hereafter referred to as ciTBIs, are both commonplace and require immediate medical attention. However, diagnoses often require a CT scan to confirm the presence of a TBI [CITE 14-16 from paper]. In children, this need is problematic, as the radiation from a CT scan can lead to long-term adverse affects; 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. Note that forgoing a CT scan in the presence of an actual TBI is also undesirable. In this report, we revisit the data from [CITE] and derive an updated decision rule for identifying which child patients need a CT scan. That is, given several descriptors of a patient's status, we translate these into numerical features that we may feed into a statistical model to predict the need for a CT scan.

TODO: OUTLINE OF SECTIONS

2. Data

2.1. **Data collection.** The authors in [CITE] 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 between physicians and parents). 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 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. 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.

Date: 12/10/2021.

2.2. **Meaning.** The study defined death, a hospital admission of over 2 days, intubation for over 24 hours, or the need for surgery following a CT scan as a positive, ciTBI outcome. 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 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. The scores from each of these three categories are summed to form a total GCS score. 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). A GCS score ranges from 3-15.

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 to 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, both the authors of [CITE] and us chose to separate patients under the age of two (pre-verbal) from those aged two or older (verbal) in their analysis. Moreover, as children under the age of two are more sensitive to radiation, it is reasonable to consider this group separately [CITE SOMETHING]. 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) but, nonetheless, it is good developmental benchmark [CITE SOMETHING].

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). 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).

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.

Several binary indicator variables exist 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.

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 and noted that there were 20 patients that had a missing value. Note, there is a discrepancy where with [CITE], where they mention 18 not 20 patients have a missing value. However, the reasoning for this difference could not be resolved. Of these 20 patients, 17 of them are negative for all four of the variables making up our outcome (death, hospital admission of longer than 2 days post CT, intubation for more than 24 hours, or neurological surgery). We thus assign these patients as being negative for a

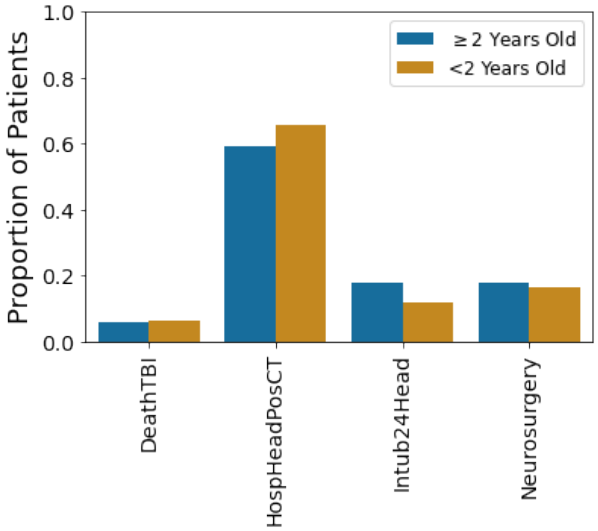
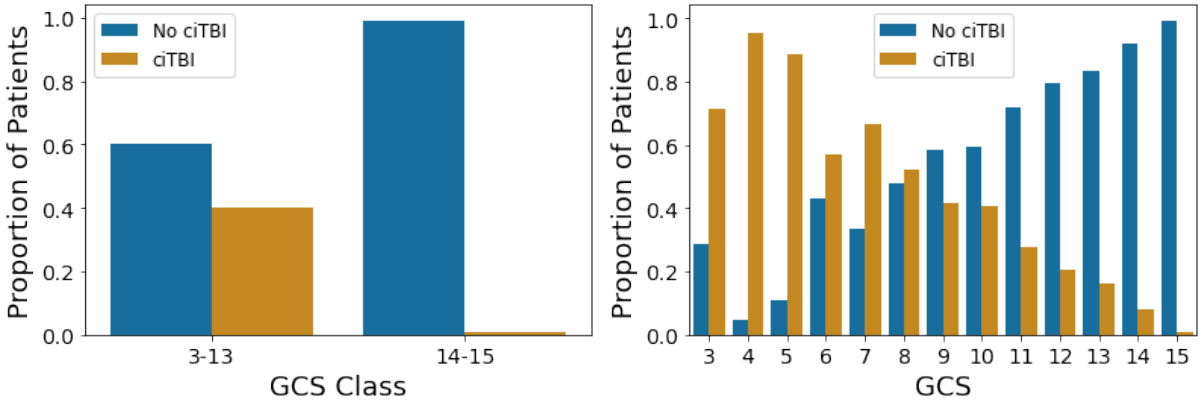


FIGURE 1. Outcome type for ciTBI patients by age group

ciTBI. For the three remaining patients, they had missing values for one or more of the four outcomes and were dropped. The last remaining patient has missing values for intubation and death by TBI. We impute all three of these patients to also be negative for ciTBI after talking to our clinician collaborators who suggest it is unlikely death by TBI or intubation would not be marked on a medical form if they did in fact take place. The proportion of each of the four outcomes is shown below in Figure 1. We can see that the vast majority of people were positive for a prolonged hospital stay.

2.3.2. *GCS Scores.* From [CITE], 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. We remove any patients from here on out 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 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

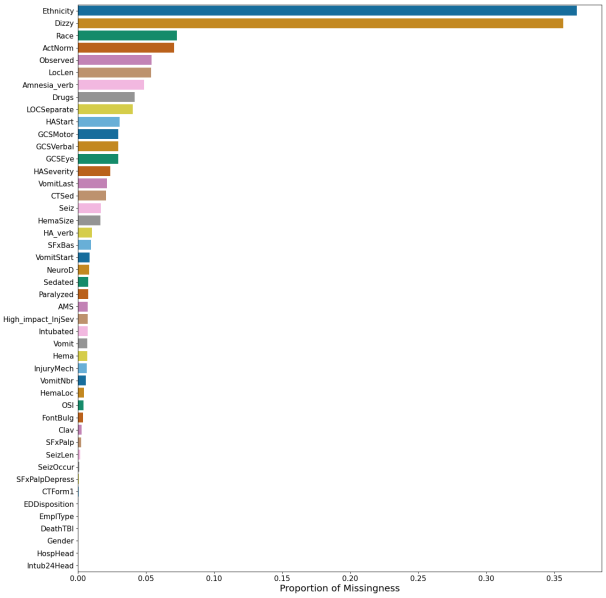
FIGURE 2. GCS Score and ciTBI

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 medicial 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 anyway. 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 as it is highly susceptible to change from patient to patient based on their own personal definition of feeling dizzy. Thus, we decide to drop this variable.

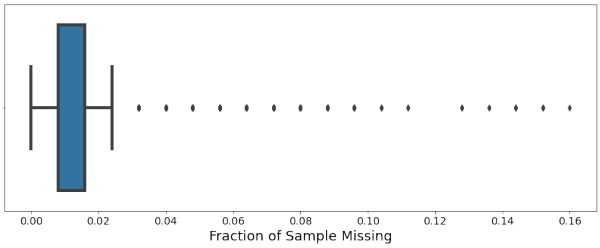
For the other variables with missingness, they were either imputed with what a 'healthy' response would be, e.g. a 'No' value would be imputed for missing paralyzed or sedated values. Otherwise, the response that was the mode was used for variables where there is no clear 'healthy' response. e.g. hematoma size.

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. 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 at most 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 [CITE] in terms of creating a decision rule. Two rules were created based on age categories of < 2 and ≥ 2. 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.

We want to check the number of pre-nonverbal subjects at each age to see if two years old is a good cutoff age for being verbal, however. Below, in Figure 5, we can see that actually there are still a large proportion of subjects that are pre-verbal at ages 2 and 3 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 age age are extremely similar.

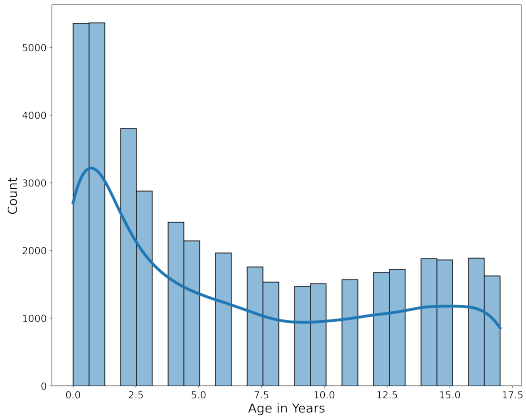
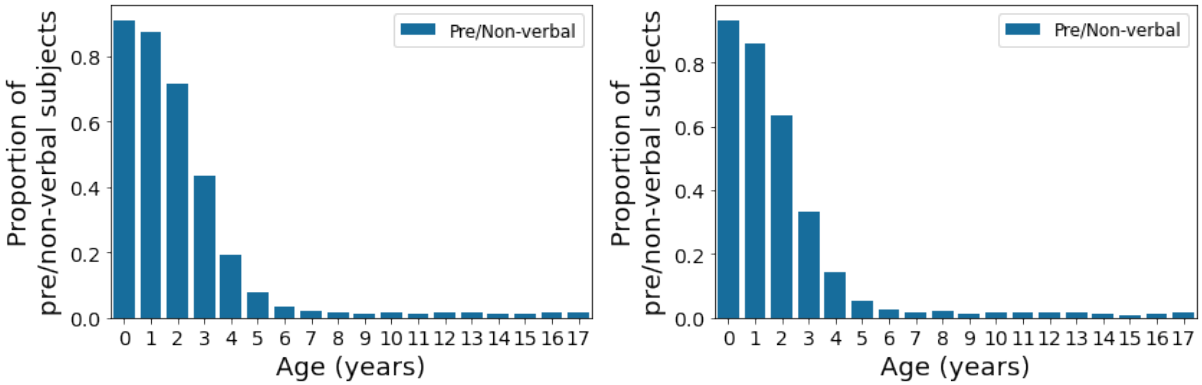


FIGURE 4. Age Distribution

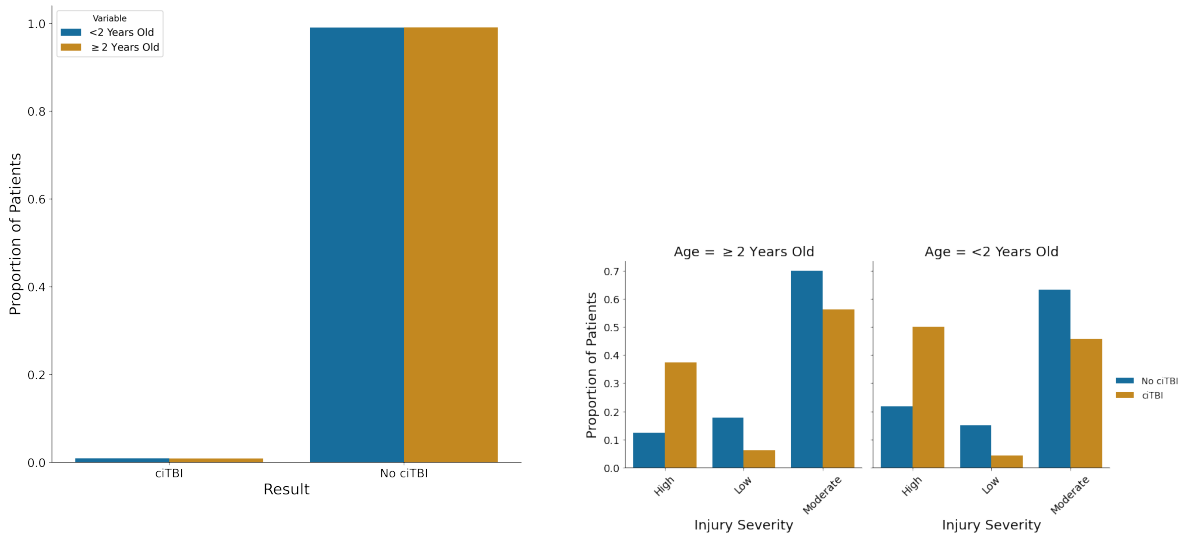


(A) Preverbal response to Amnesia question (B) Preverbal response to Headache question

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, remember 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

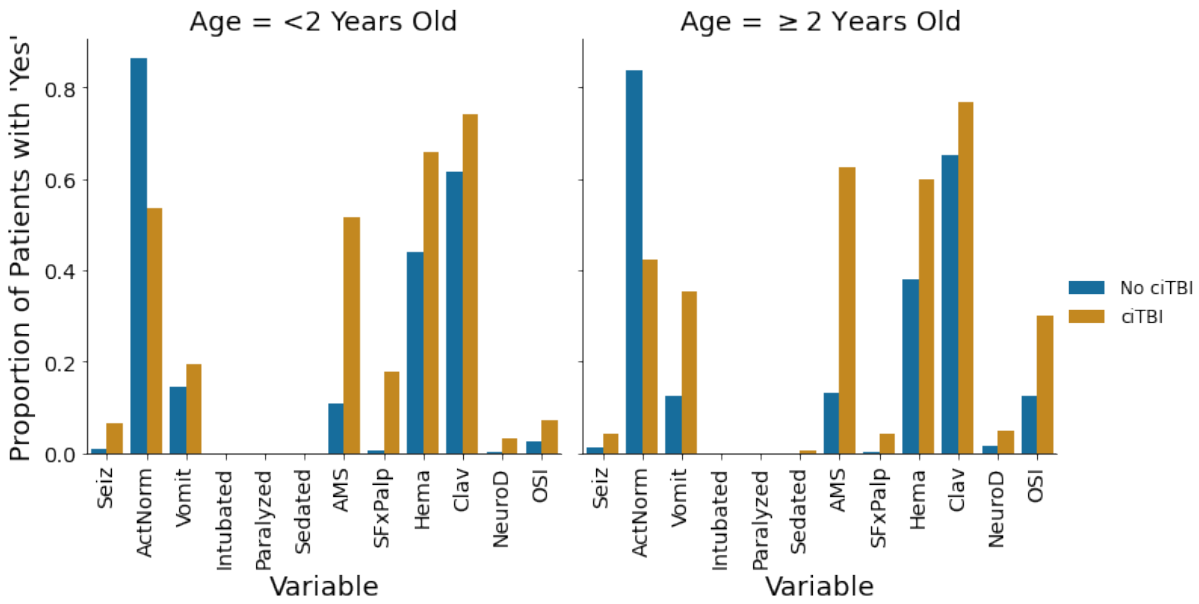


FIGURE 7. Proportion of positive feature identifiers by outcome (ciTBI) and age

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 particularly highly correlated with the outcome. A maximum correlation coefficient of 0.12 is attained by the altered mental state feature.

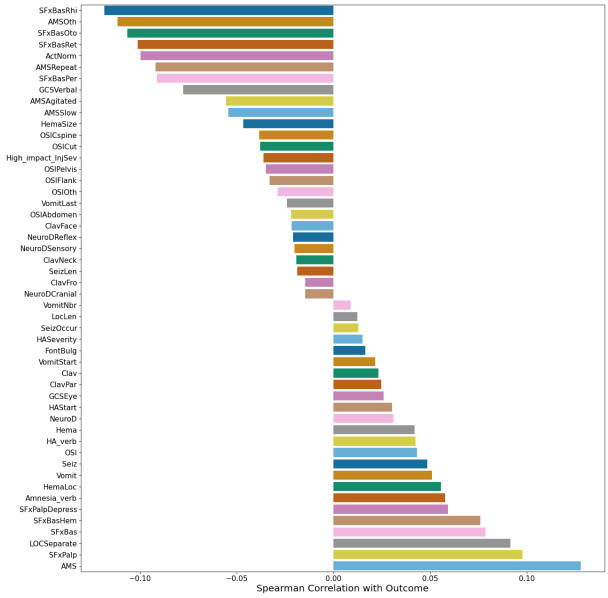


FIGURE 8. Spearman Correlation of Features to Outcome

2.3.7. *Principal Component Analysis.* We perform Principal Component Analysis (PCA) on the one-hot encoded data. 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, we did not suspect a batch effect. 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.

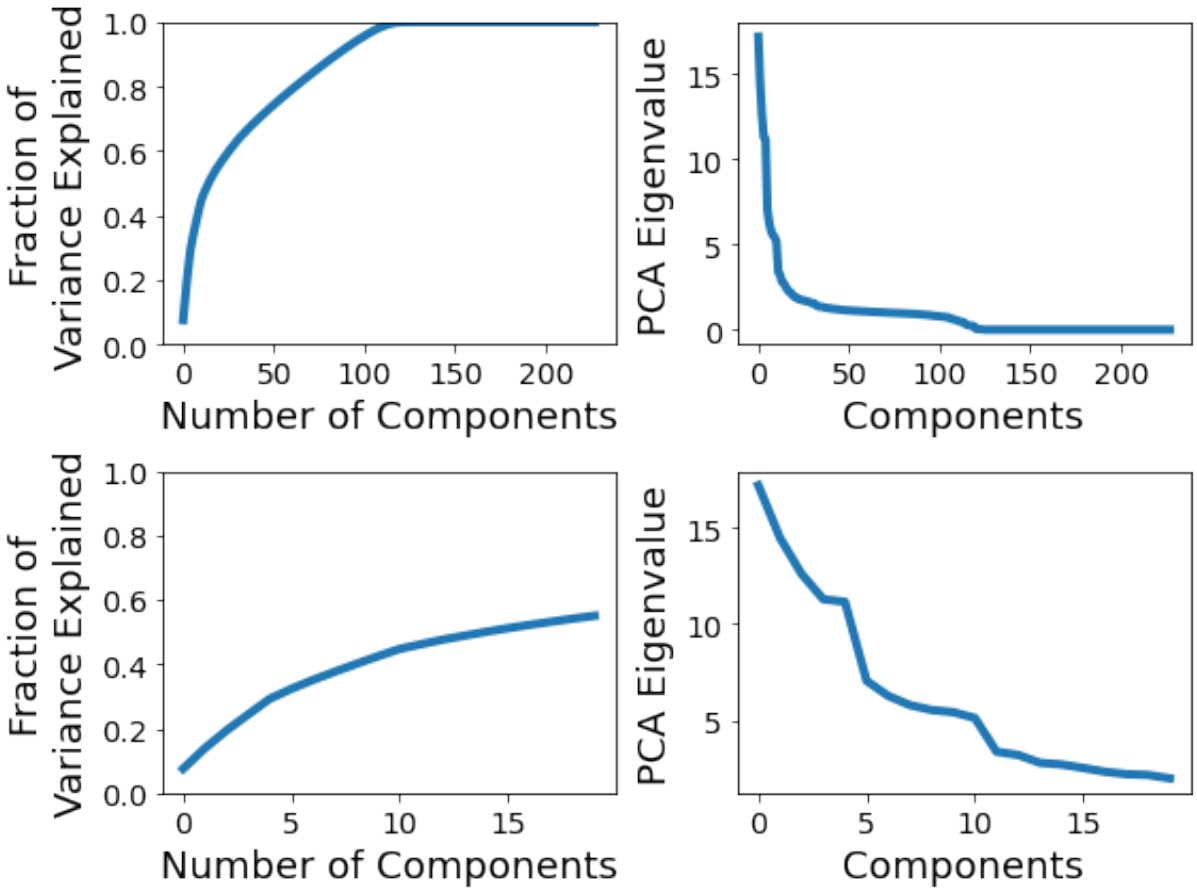
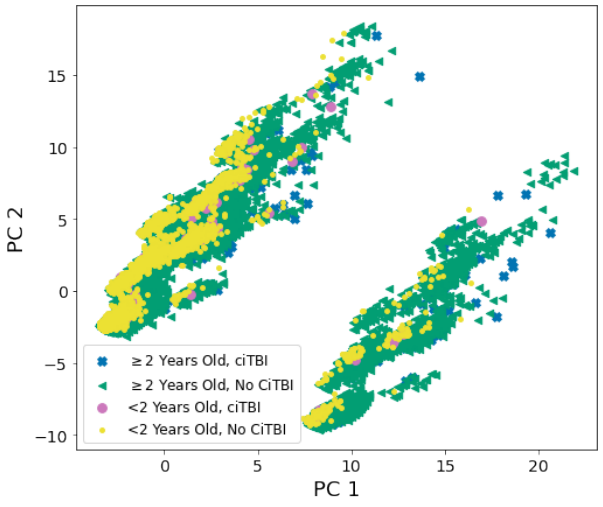
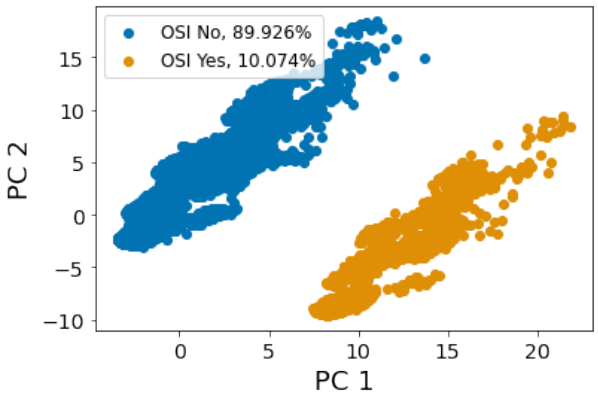


FIGURE 9. PCA: Cumulative Variance Explained and PCA Eigenvalues



(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. Baseline Model

4. Modeling

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, group ℓ_1-penalized logistic regression, a single decision tree, a random forest, AdaBoost, LogitBoost, and a linear SVM. All of these models were chosen for their relative ease of implementation and interpretability. 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 0.95 (or as close to it as possible) and the negative predictive value (NPV) was as close to 1 as possible. 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.

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 the desired value of 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 much better than the baseline algorithm. 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.

Algorithm	Age	AUC	Accuracy	Sensitivity	Specificity	NPV	Balanced Accuracy
ℓ_1-Logistic Regression	young	0.938	0.764	1.0	0.762	1.0	0.881
ℓ_1-Logistic Regression	old	0.931	0.751	0.957	0.75	1.0	0.854
ℓ_1-Logistic Regression	all	0.917	0.75	0.952	0.748	0.999	0.85
Group ℓ_1-Logistic Regression	young	0.908	0.728	1.0	0.726	1.0	0.863
Group ℓ_1-Logistic Regression	old	0.917	0.68	0.957	0.678	1.0	0.818
Group ℓ_1-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
SVM	young	0.275	0.014	1.0	0.008	1.0	0.504
SVM	old	0.645	0.057	0.957	0.051	0.994	0.504
SVM	all	0.644	0.063	0.952	0.054	0.991	0.503
Baseline	young	0.771	0.545	1.0	0.542	1.0	0.771
Baseline	old	0.785	0.615	0.957	0.613	0.999	0.785
Baseline	all	0.77	0.626	0.917	0.623	0.999	0.77

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

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 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. 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 generally 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 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 much better than the baseline model. 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.

Algorithm	Age	AUC	Accuracy	Sensitivity	Specificity	NPV	Balanced Accuracy
ℓ_1-Logistic Regression	young	0.846	0.772	0.923	0.77	0.999	0.846
ℓ_1-Logistic Regression	old	0.848	0.762	0.937	0.76	0.999	0.848
ℓ_1-Logistic Regression	all	0.822	0.794	0.85	0.793	0.999	0.822
Baseline	young	0.774	0.554	1.0	0.549	1.0	0.774
Baseline	old	0.786	0.639	0.937	0.636	0.999	0.786
Baseline	all	0.777	0.623	0.933	0.621	0.999	0.777

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

4.1. **Discussion of the ℓ_1-regularized Logistic Regression.** In this section, we provide some insights from studying the logistic regression models that we have fit. We note that this from 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, 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.

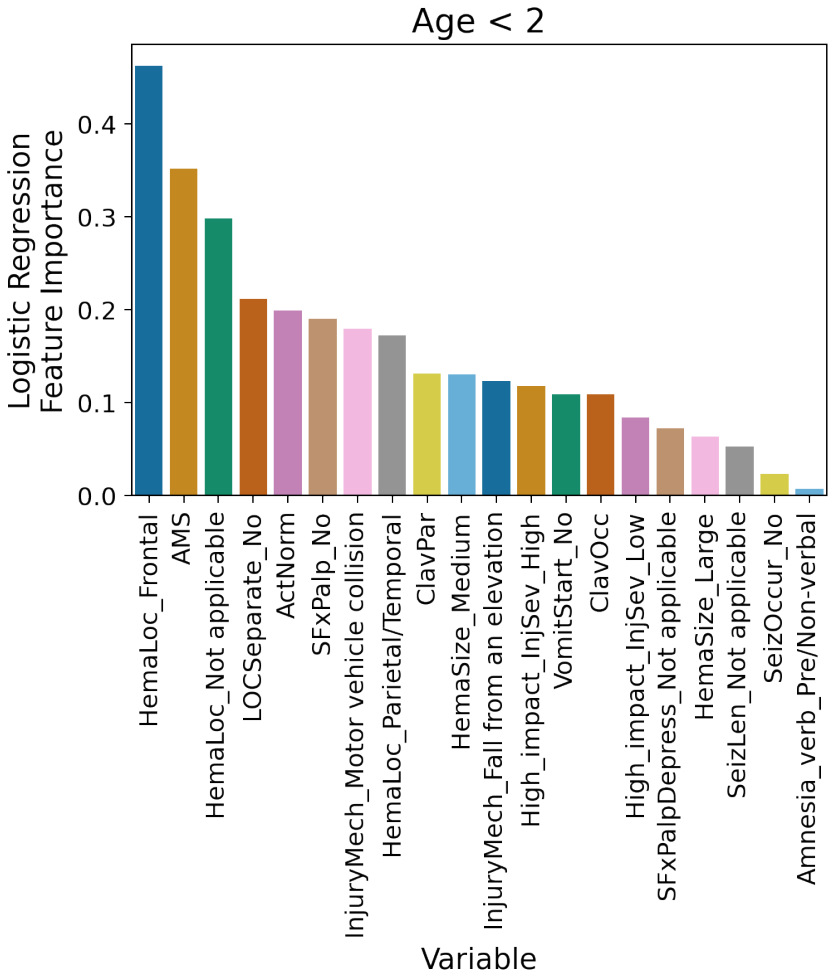


FIGURE 11. Feature importances from the ℓ_1-regularized Logistic Regression for young patients

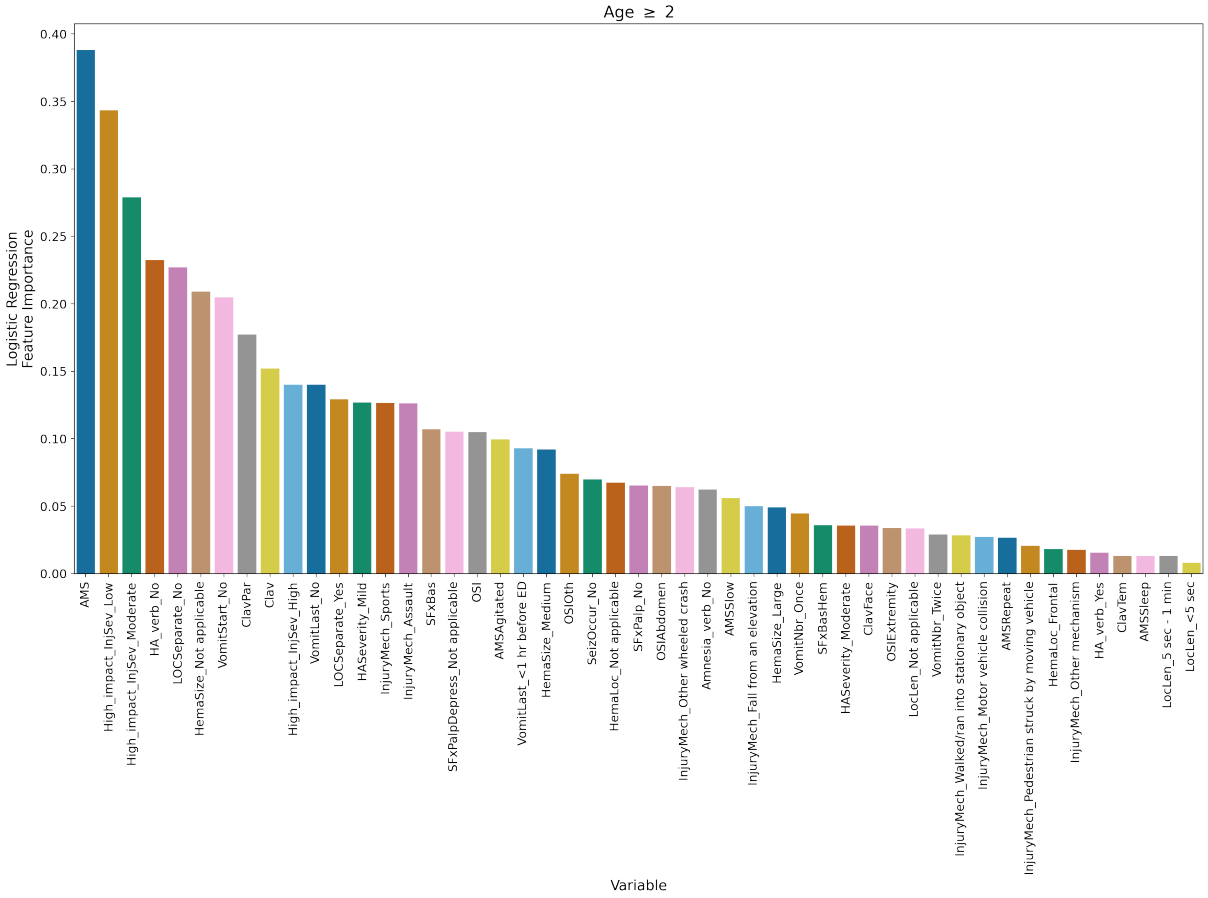


FIGURE 12. Feature importances from the ℓ_1-regularized Logistic Regression for older patients

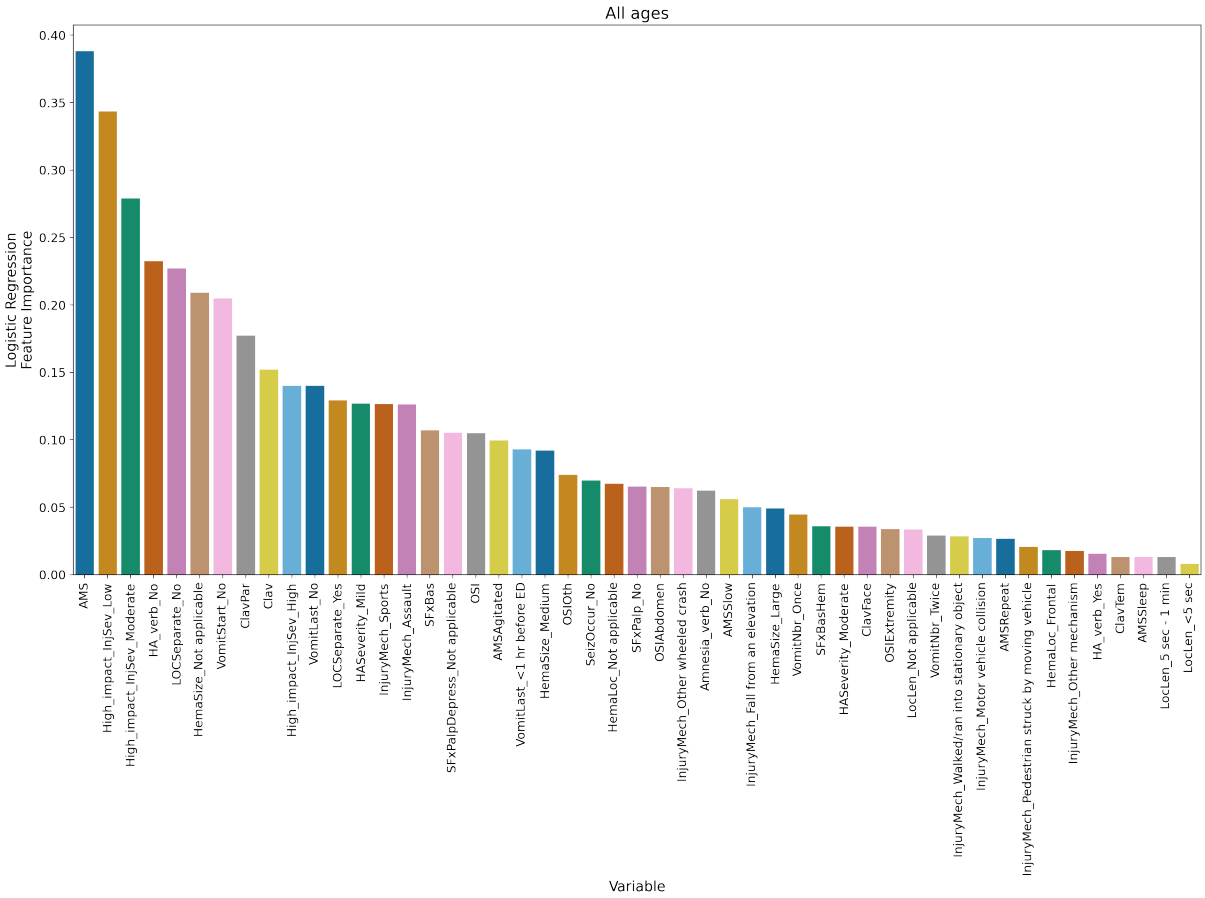
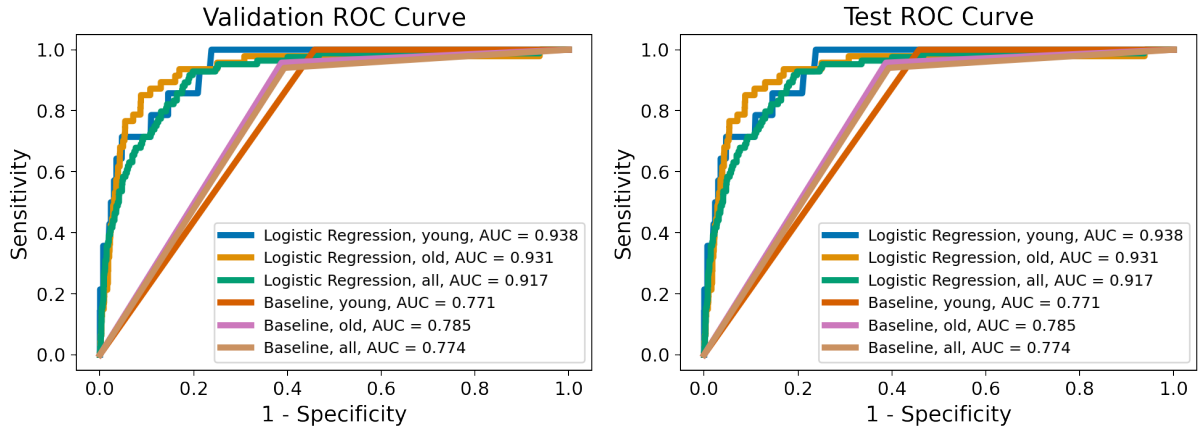


FIGURE 13. Feature importances from the ℓ_1-regularized Logistic Regression for all patients



(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

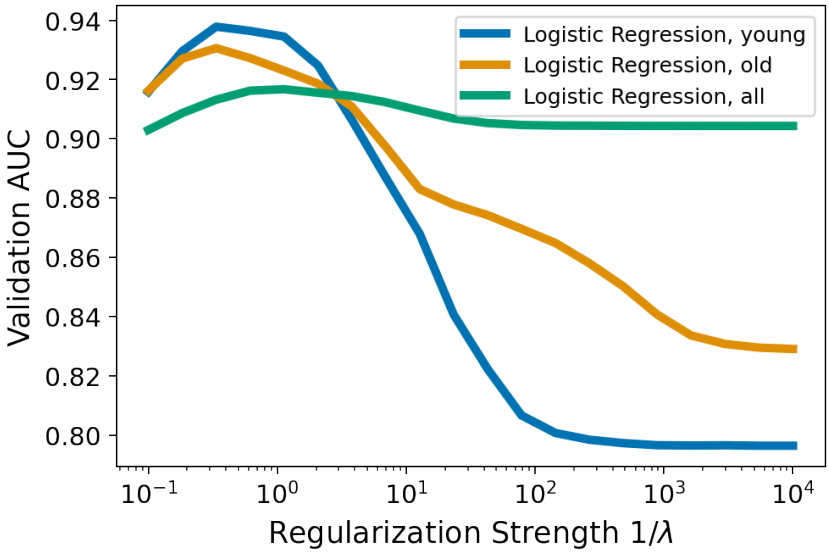
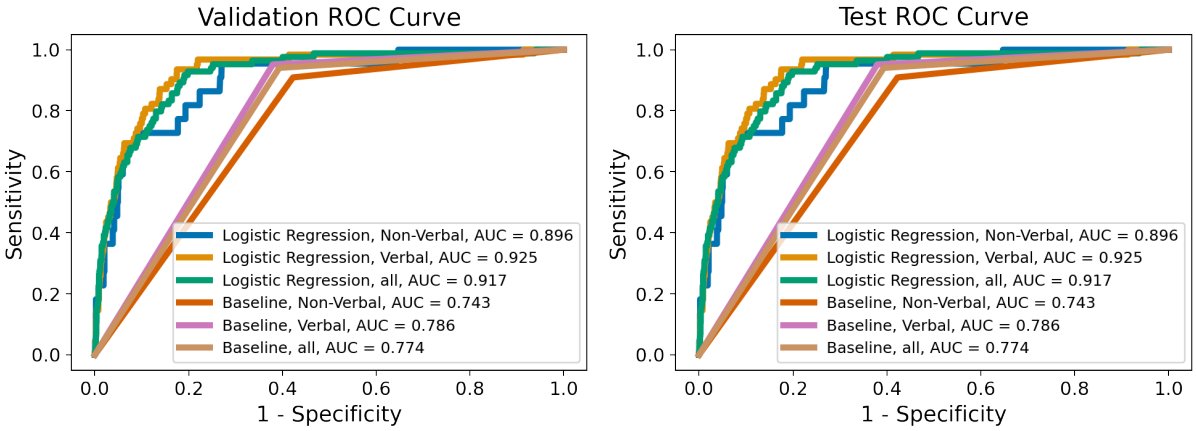


FIGURE 15. Validation AUC for the ℓ_1-regularized Logistic Regression as a function of regularization strength

4.2. **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 16, 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.



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

5. Conclusions

asdf.

5.1. **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, 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 much of the data processing and implementation of dataset.py, worked on coding the baseline model, summarized his findings for EDA and the baseline model, and documented the judgement calls made working with this data set, 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 and how SVM works on the first few PCA components, contributed to the data dictionary and README files, 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, wrote the final report, implemented `model_best.py` and `baseline.py`, and offered comments on the final report.

6. References