Final Project - Evaluation of Clinical Decision Rules for Pediatric Cervical Spine Injuries, Stat 215A, Fall 2021

Yaxuan Huang, Ishaan Srivastava, and William Torous

December 12, 2021

Abstract: This project aims to create interpretable decision rules for emergency room diagnoses of cervical spine injuries in children. These neck injuries can lead to paralysis or death if not treated correctly, making sensitivity a priority for our work. Data to build these rules comes from medical records abstracted for a retrospective case-control analysis of over 3000 ED visits for potential cervical spine injuries. The PCS framework is used to test the predictive power and stability of our proposed decision rules, as well as ones from other studies. We propose a Special Tree model which is a constrained tree-based model and a forward selection procedure to fit the model. The model is super interpretable and easy to construct. It performs as well as Random Forest and much better than CART. Based on this method, we construct decision making rules based on age and specific clinic indicators. Our final model has similar performance with baseline, sensitivity 96.8% (specificity 36.9%) on training set, and 91.3% (34.22%) on testing set. The corresponding numbers for baseline model are 92.1% (40.22%), and 96.1% (42.2%). Unlike the baseline, our model reflects the heterogeneity patients' age from toddler to teen and derives rules with age stratification. And we also detected that an important factor `GCSbelowThreshold` which is not include in Leonard et al. [2011].

1 Introduction

The advent of computed tomography (CT) scans has given emergency department (ED) doctors a powerful tool for non-invasive medical imaging. CT scans use repeated X-rays to construct cross-sectional images of a patient's body and provide more information than traditional suites of X-ray images. Because of this information gain, CT scans have become a popular and often standard tool for identifying injuries in adults. The role of CT scans in pediatric care is an active area of research and debate because the procedure delivers 30 times more ionizing radiation to a patient than traditional X-rays (Leonard et al. [2011]). Children are more sensitive to the effects of ionizing radiation because their cells divide more frequently than adults'; a review study has estimated 1 in every 4000 head or neck CT scans for children under 10 will result in a malignant brain tumor (Chen et al. [2014]).

A cervical spine (c-spine) injury (CSI) refers to an injury in the first seven vertebrae (C1-C7) of the neck. These injuries, which include fractures and ligament damage, are extremely important to diagnose and treat correctly because they can damage the nervous system and lead to paralysis or death. Common mechanisms of c-spine injuries in children include motor vehicle accidents, falls from heights, and sports collisions. CT scans are a standard diagnostic tools for adults suspected of having CSI. Due to differences in physiology, c-spine injuries are less common in children than adults (Devlin [2021c]), and less than 1% of pediatric trauma cases in the ED are caused by them (Leonard et al. [2011]). Because of this lower likelihood of injury and children's increased risk from radiation, the decision to use a CT scan must be carefully weighed against alternatives by ED doctors.

Clinical decision rules are designed to help doctors make more informed judgement calls by summarizing information the medical community has learned from prior cases. The research of Leonard et al. [2011] suggests 8 predictive covariates for when pediatric patients are at greater risk for a CSI. In this report, we use the PCS framework to stress test those covariates using the original dataset. This reanalysis of Leonard et al. [2011] is especially important because their model was built without withholding unseen test data, which can lead to overfitting. We also propose additional clinical covariates and build alternative decision rules for ED doctors. We verify the stability of our proposals by generating alternative versions of the data under different researcher

judgement calls and data perturbations. Dr. Gabriel Devlin of UCSF, a pediatrician who has worked in the ED, provided perspective on pediatric ED operations, answered our medical questions, and gave feedback about judgement calls.

A missed CSI diagnosis can have life-altering effects, so having very high sensitivity is the main priority for our work. Having high specificity helps achieve our project's goal of reducing unnecessary CT scans. Although we focus on predicting c-spine injuries, not whether a CT scan is required, answering this question can help clinicians make more informed imaging decisions. Considering the time-sensitive application of our rule in the ED, interpretability to doctors and ease of use are also important considerations.

We begin this report by describing the dataset we work with: abstracted medical records describing over 3000 visits to the ED for suspected c-spine injuries. We describe steps undertaken to clean and explore this data before modeling c-spine injuries with tree and logistic regression-based methods. We use a simple 8 covariate rule proposed by Leonard et al. [2011] as a baseline to hopefully improve upon. After proposing a suite of models, we select the most promising and test its stability under various researcher judgement calls and data perturbations.

2 Data

The data for this project was originally collected for the research of Leonard et al. [2011]. The researchers conducted a retrospective case-control study to identified clinically-observable factors related to c-spine injuries in children under 16. The dataset's observational units are patients who visited the emergency department for a potential cervical spine injury at one of 17 study site hospitals in the Pediatric Emergency Care Applied Research Network (PECARN). All within the United States, the study sites are large urban hospitals which serve as regional trauma centers and have specialists in pediatric care. Patients in this dataset include transfers from other hospitals. Each patient's medical records have been extracted by a doctor into 786 covariates across 11 datasets. These covariates thoroughly describe a patient's condition upon initial evaluation at the ED, their prior medical history, information from any X-rays, CTs, or MRIs performed, and their physical state on discharge. Data is available for 3314 patients who visited a study site ED between 2000 and 2004. We downloaded the data from PECARN's website. The dataset contains 540 case patients who had a c-spine injury. The remaining 2774 patients fall in three control groups. The largest control group with 1060 patients is comprised of randomly selected patients who did not have a c-spine injury after evaluation for one. A mechanism of injury (MOI) matched control group contains 1012 patients with a similar MOI and within a year of age to specific case patients. The third control group contains 702 matches for case patients who arrived to the hospital via EMS. The original researchers focus on the randomized group to build their rule and use the matched pairs to analyze whether age, MOI, or EMS confounds results. We use a similar approach.

As we aim for our decision rules to be used during initial evaluation at the ED, we attempt to identify which information would be available in that setting. The information we consider most relevant is recorded by the evaluating doctors. Medical records collected in the ED indicate the detailed physical condition of a patient, their responses to any questions doctors asked during examination, and information about how the injury may have occurred. If the patient arrived via EMS or was transfered from a referring hospital, similar information summarizing their previous condition may be available. According to Dr. Devlin, a patient's electronic health record (EHR) of prior hospital visits is often available during ED evaluation. This can provide useful information about pre-disposing conditions a patient may have. Demographic data, including physical characteristics and insurance information about a patient, could also be available in the ED. However it is unclear why patients with different demographics should receive different care, so we do not predict based off this information and only use it to evaluate subgroup performance.

The dataset contains sources of information not available to doctors making a diagnosis. This includes the diagnosis itself, the results of medical imaging that was ordered, what interventions including surgery, immobilization, medication, and rehab were undertaken, and how the injury impacted the patient's quality of life. We use this information to evaluate how well our models perform on different injury severities.

As part of Leonard et al.'s research effort, doctors reviewed the medical records associated with each case and filled out an extensive survey about the data. The covariates extracted from this survey were split into 10 distinct datasets described below. Each patient is associated with a unique ID, a case ID to denote matched groups, a case-control type, and a study site. We use these four variables as an index in our data processing.

Dataset	Description	Number Covariates
Index	Common between datasets, used to match rows	4
AnalysisVariables	Covariates derived by Leonard et. al (2011)	32
ClinicalPresentationField	Observations recorded by EMS in the field	97
ClinicalPresentationOutside	Observations recorded by referring outside hospital	110
ClinicalPresentationSite	Observations recorded at the study site hospital	132
Demographics	Demographic information	7
InjuryClassification	Details of injuries and their classification	204
InjuryMechanism	Details of how injuries occured	31
Kappa	Data re-abstracted into the survey by another doctor	120
MedicalHistory	Prior medical history from EHRs	27
RadiologyOutside	Results of radiology imaging from referring hospital	11
RadiologySite	Results of radiology imaging at study site	11

Table 1: Description of the 11 datasets created by Leonard et al. summarizing medical records for c-spine ED visits. There are 786 unique covariates across all 11.

Leonard et al. built their models with 32 derived binary covariates in the `AnalysisVariables` dataset. These aggregate groups of related indicator covariates from the full dataset; for example, `FocalNeurlogicalFindings` is indicated if the patient exhibits one of three sensory changes and `HighriskMVC` is a decision rule based off speed and other factors for high risk car accident. Because these features have been derived by domain experts and group related indicators, we work significantly with them. Of these, 22 are unique and 10 are robust versions which have a `2` suffix in their name. The baseline covariates only include information recorded at the ED, such as complaints of neck pain. The robust version indicates if the baseline condition was indicated at the study site OR by EMS OR at an outside hosptial. As discussed later, the relevance of information collected outside the hospital is a judgement call.

A dataset entitled `RadiologyReview` contains detailed information from a reanalysis of the 540 case patients' medical imaging in Nigrovic et al. [2012]. As we only consider the types of imaging ordered during our posthoc analysis and because the study does not consider control patients, we do not include this dataset. Another notable dataset is `Kappa`. The medical records for 365 patients were re-abstracted in Leonard et al. [2011] to evaluate agreement between reviewers. We use this data as a natural stability check and do not consider it during exploratory data analysis.

2.1 Preprocessing

For ease of analysis, we begin our project by merging all 10 datasets into a single data frame. Some datasets share covariate names; we add suffixes to differentiate between them and for straightforward filtering. Covariates collected by EMS or by a referring hospital have the suffixes `_ems` and `_outside`, respectively. Information about injury classification, radiology ordered at the study site, and long-term outcomes has the `_posthoc` suffix to denote they are not available at decision time.

The dataset contains 35 covariates with free-form text information. These include descriptions if an `other` category is marked, such as for surgical intervention; insurance billing codes describing injury mechanism and classification; and chart notes. Thankfully doctors have reviewed this text and abstracted most information into binary or categorical covariates. For this reason and because we have little experience working with text data, we do not consider these covariates. We also remove all information about date and time from the datasets because this information does not make sense in a decision rule. We confirmed with Dr. Devlin that this judgement call makes clinical sense Devlin [2021d].

Leonard et al. format their analysis variables such that 1 indicates an abnormal condition, and we standardize all the binary covariates to this format. We also create a binary treatment indicator `csi_injury` which is our outcome of interest. We also note that many of the dataset's categorical covariates can be encoded as binary without loss of information. When possible we automated this process by searching for covariates which only take on pairs such as `Y/N` and `A/N` (abnormal and normal). Categorical covariates are often blank if that information was not found during the medical record extraction, and we mark these blank entries with `NA`. Imputation will be handled later.

Almost all of non-binary categorical covariates have the _posthoc suffix, and there is no modeling incentive to binarize these in some fashion. An important categorical covariate is the AVPU test. Doctors categorize patients as alert, verbally responsive, painfully responsive, or unresponsive in this common ED procedure. As these categories are mutually exclusive, we convert AVPU test results to one-hot encoding. We include an indicator if AVPU was missing or recorded in an invalid way (e.g. "S").

We noted above that Leonard et al. create robust versions of their covariates which indicate if a baseline study site indicator is satisfied by a wider net of study site, EMS or outside hospital data. Dr. Devlin suggested that ED doctors try to confirm abnormalities noted by outside sources (Devlin [2021b]), making the robust information not very useful in a decision rule. Using both the study site-only and robust covariates, we can determine if a patient's condition improved between outside evaluation and the ED. We add these derived covariates, which are also computed for some non-Leonard covariates, with the suffix _improved.

Many covariates that describe patients' specific outcomes are aggregated, which greatly reduces the dimensionality. We only use outcome information for our posthoc analysis, and Dr. Devlin suggested high-level aggregation would still be informative. From the InjuryClassification dataset, we keep only 3 of the 204 covariates. This binary dataset indicates very precise injury classifications, while we consider only the class of injury. We have indicators for CervicalSpineFractures, LigamentInjury, and CervicalSpineSignalChange, which indicates a general injury (Devlin [2021a]). In similar fashion, the three radiology aggregations we keep from the study site and referring hospital are CTPerformed, XRays, and MRIPerformed.

The full dataset contains very detailed mechanism of injury information through ICD9 codes and thorough details of motor vehicle crashes. We do not include these covariates and instead mainly rely on Leonard et al.'s aggregation into Highrisk MOI groups. We do include several additional MOI risk factors such as suspected child abuse and a seat belt indicator.

After these pre-processing steps, 317 covariates remain. The 17 non-binary numeric covariates contain information about a patient's age, Glasgow Coma Scores (GCS) (which range in integers from 3 to 15), and categories for number of steps fallen down. The 18 remaining categorical variables almost all map to posthoc patient outcomes, which we can leave in this format. We remove the two remaining categorical variables: one about where in the hospital patients arrive because this inidcates the ED for over 98% of units, and one about the position EMS found the patient in, which is missing in over 60% of units.

2.2 Conversion of Categorical to Binary Features

As the vast majority of our prediction covariates are binary, we make a modeling choice to add binary cutoffs to the remaining non-binary numeric covariates. From a patient's age, reported to two decimal places, we extract three binary covariates. We use as a baseline the cutoffs $[2, 5, 12]$ which are standard in drug discovery trials (Devlin [2021b]) and make each cutoff an automated judgement call. The variable VeryYoung denotes a patient under 2. Dr. Devlin shared that kids under 5 often have trouble verbalizing where their injury hurts and NonVerbal indicates age between 2 and 5. Finally, YoungAdult indicates children older than 12; this age group is more exposed to more violent collisions in sports and may take more unnecessary risks than younger children. Mirroring Leonard et al.'s creation of other Highrisk covariates, we denote a HighriskFallDownStairs as a fall down 6 or more stairs, and make a raised threshold of 15 an automated judgement call.

The dataset contains Glasgow Coma Score information collected at the study site, by EMS, and at a referring hospital. The total score is the sum of three subcategory scores: mental, eye, and verbal; the largest possible score is 15. Reviewing current c-spine decision rules in ED departments, total GCS is often used as a boundary. We make a judgement call to not consider GCS information collected outside the study site. We find that 140 patients have EMS or outside GCS, but not study site GCS. However, using AlteredMentalStatus $=$ 0 as imputation for total GCS $= 15$, as justified later, only 21 units remain. We leave these study site GCS values as NA and will impute them later. As our models are designed for binary data, we unfortunately cannot test this section's major judgement call of binary conversion.

3 Exploratory Data Analysis

The major goals of our exploratory data analysis are further shrinking the possible set of covariates before modeling, understanding patterns of missing data, and informing our translation of the c-spine injury prediction

task into a statistical question. The analysis variables summarised by Leonard et al. receive particular focus because they account for groups of outcomes that vary together and can potentially be imputed from each other. An initial question we investigate is how well this data represents different groups of patients. We will withhold prediction data for most of this section, but explore demographic trends on the entire dataset first for insights into how to split the data.

3.1 Demographics

A natural clustering to explore demographic trends is at the study site level. The 17 study sites are blinded and only referred to by their index. The demographic information is summarized by Figure 1. We observe that patients are consistently more likely to be male across all study sites. The other demographic covariates are less consistent. Patients over 12 years old are the largest demographic at most hospitals, and c-spine injuries occur in far fewer children under 5 in this dataset. Ethnicity information is rarely reported, and at least one hospital has over a quarter Hispanic patients. We can us insurance type as a weak proxy for socioeconomic status because some government insurances, such as Medicaid, are only available to those below a certain income level. At many hospitals in the dataset, private insurance is the most common class.

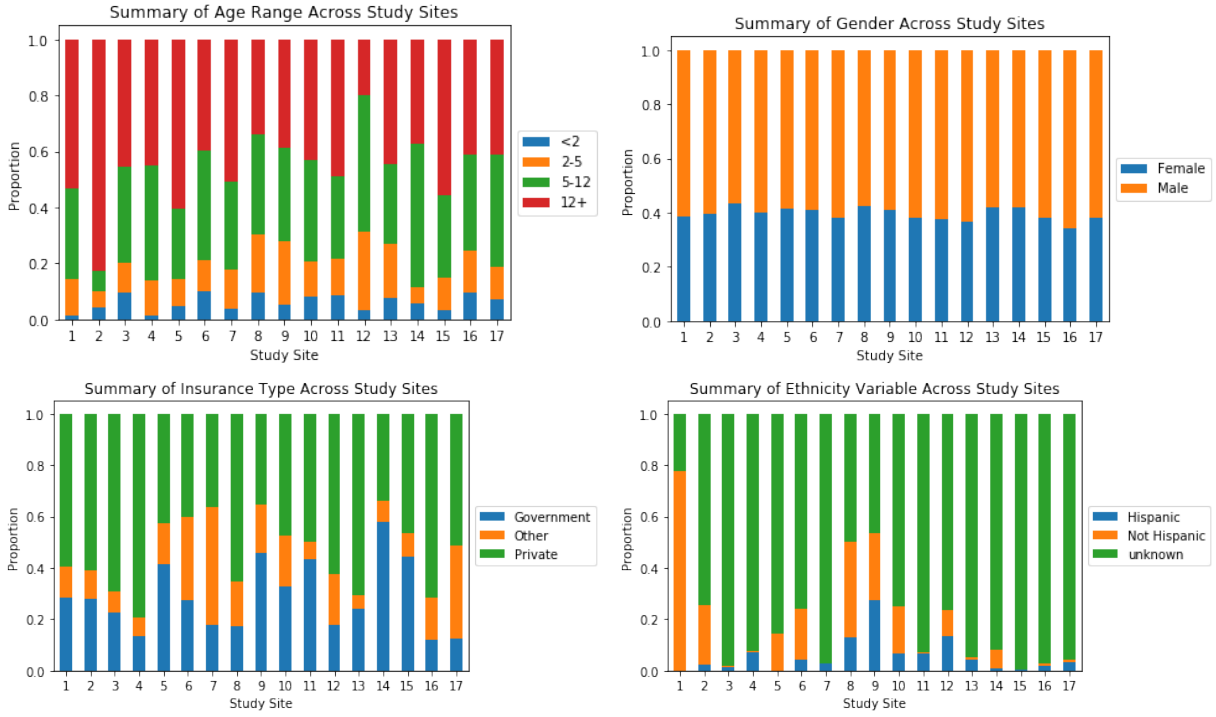


Figure 1: These bar charts summarize how age, gender, insurance type, and ethnicity vary across the 17 study sites. Other than gender, these demographic covariates have heterogeneity between hospitals.

Patients treated at study site hospitals may be wealthier than the average American because of research hospitals' urban locations. The location of a hospital within a city may also affect the demographics of its patients. Dr. Devlin shared the example of Seattle Children's Hospital, a member of PECARN, which is located in an area of the city that makes it more accessible to wealthier and whiter patients (Devlin [2021d]). These hypotheses are not testable, and we must make a judgement call that Leonard et al. [2011] took this factors into account during their study design. As regional trauma centers, PECARN hospitals injuries likely to treat severe injuries than rural hospitals, which may add bias to our dataset. This potential severity bias is unlikely to decrease the sensitivity of our models on unseen data. Based on the variation in many demographic characteristics, we believe it is advisable to generate training and testing splits within each study site.

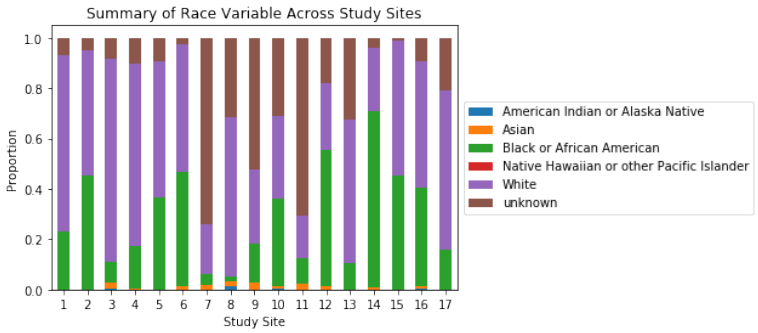
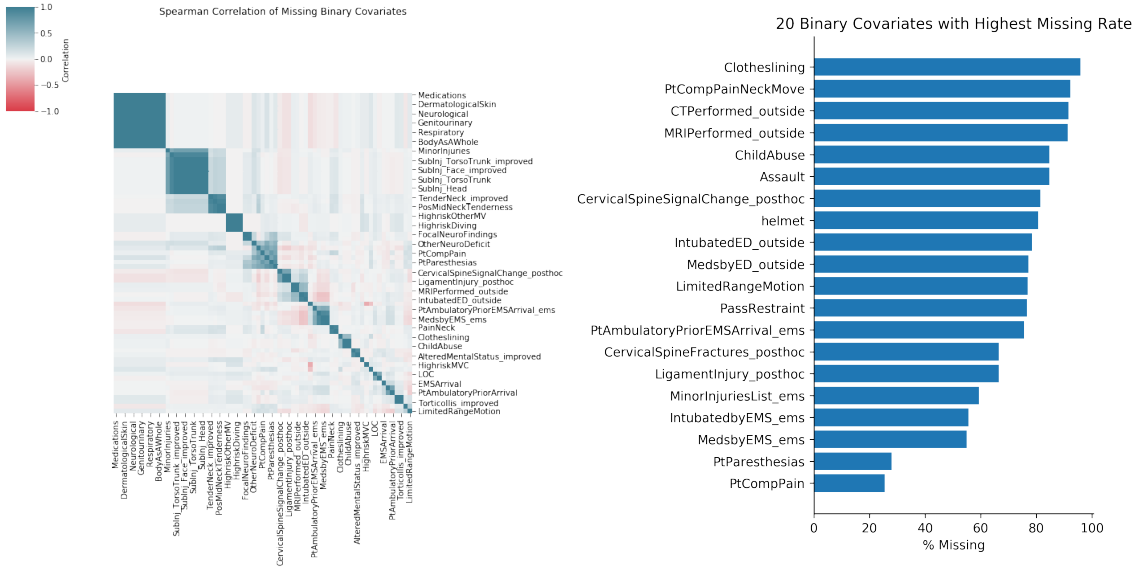


Figure 2: The empirical distribution of race for patients in the study.



(a) A correlation matrix for covariates' indicator of a missing entry.

(b) The 20 binary covariates with the highest rate of missingness.

Figure 3: These figures summarize patterns of missing data for the dataset's binary covariates.

3.2 Missing data

The preprocessed dataset is characterized by covariates missing in groups and some binary indicators with very high rates of non-entry. To visualize patterns of missingness in our dataset, we generate a correlation matrix for indicators of missing entries. Along the diagonal of Figure 3a, we observe blocks of covariates which are always jointly missing. The largest groups include information about prior medical history, indicators of substantial injury, and patient responses during study site evaluation. Off the diagonal, there do not appear to be any strongly correlated groups. We use the Spearman correlation because it is designed for data with a monotonic relationship and can handle categorical data. We note the Frobenius-norm after differencing the Kendall rank correlation is on the magnitude of round-off errors.

We give special attention to the patterns of missingness in the analysis variables dataset because these covariates aggregate covariates which convey similar information, such as both AVPU and GCS scores into `AlteredMentalStatus`. Figure 3a demonstrates that almost all units have 0,1, or 2 analysis variables missing, a small fraction. This figure also illustrates that there are systematic differences in the rate of missing data across study sites. Figure 3b shows that analysis variables and their robust version tend to be missing at similar rates. Focal neurological findings, denoting abnormal sensations, and torticollis, denoting a twisted neck, are the most likely to be missing.

We drop patients with a missingness threshold based on analysis variables, not the full dataset. Using the

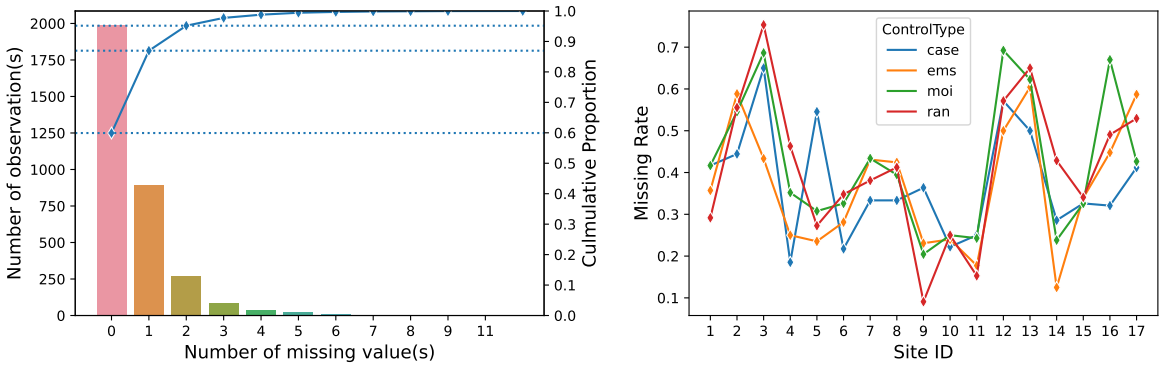


Figure 4: Left panel: Histogram of the number of missing values in a single observation. The line chart denotes the cumulative proportion. Right panel: line chart of number of observations with missing value(s) in different sites and different Control Type.

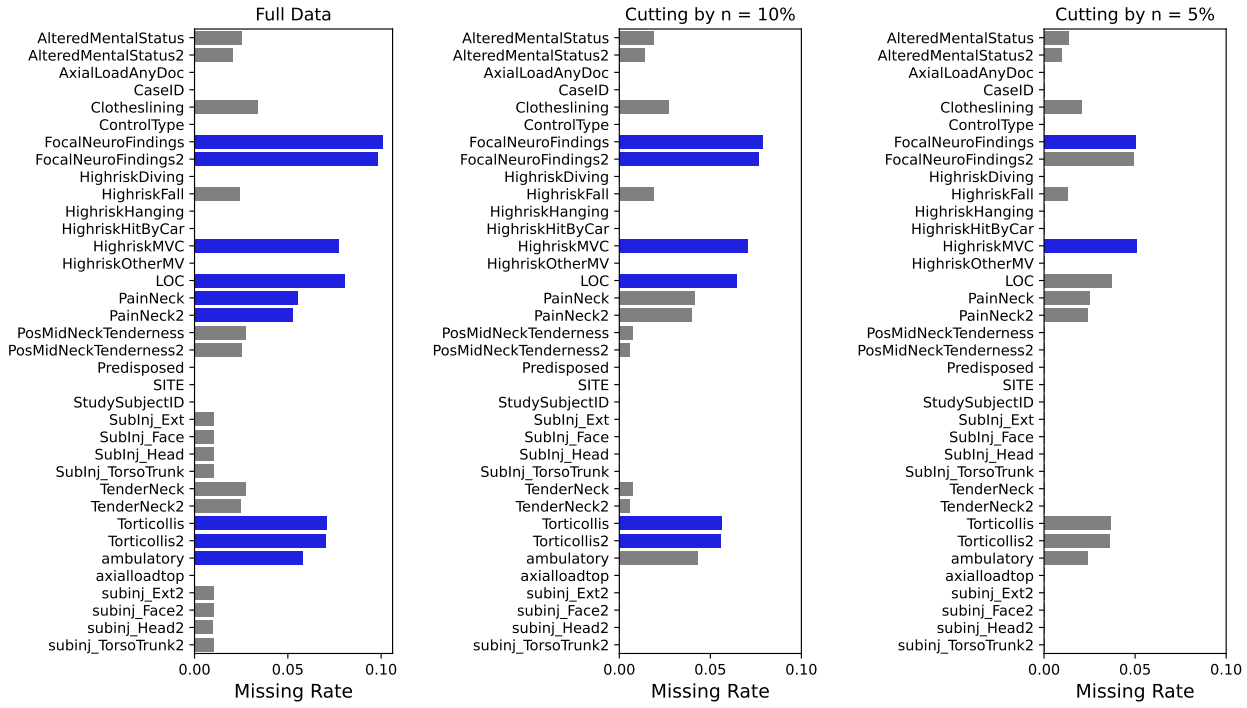


Figure 5: These two figures show some trends in how the analysis variables are missing. Blue bars denote variables with missing rate greater than 5%.

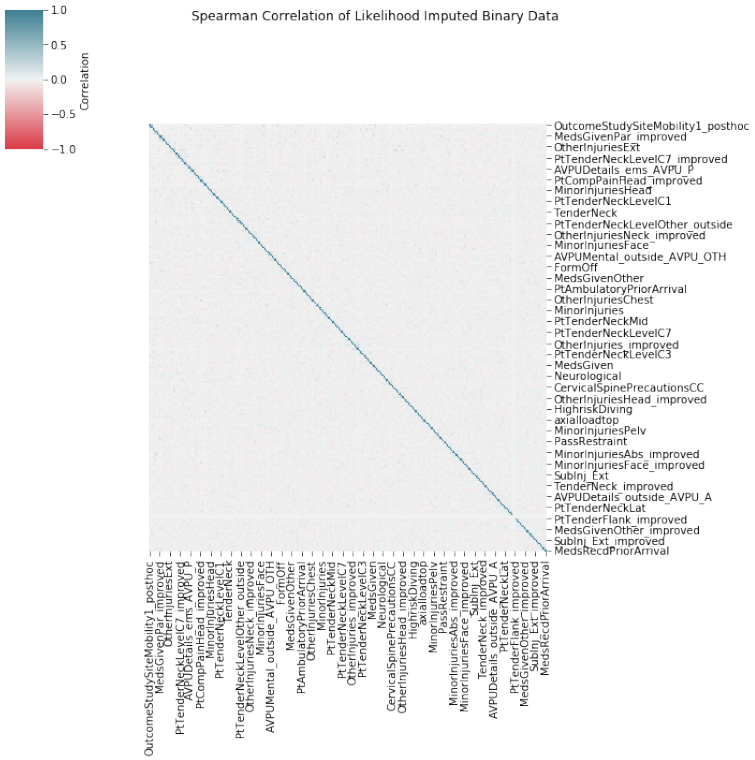


Figure 6: A likelihood-based random imputation removes almost all correlation between covariate outcomes.

analysis variables incorporates domain-expert information about important covariate groups with overlapping information, and this removal strategy . As a judgement call, we remove units with 0, 1, or 2 of their analysis variables missing, not including robust ones ending with 2.

3.3 Imputation

The vast majority of our binary covariates describe truly binary information, such as whether a patient complained of neck pain. These covariates are encoded such that an abnormal condition is always the 1 condition, which leads us to try zero imputation. We believe this approach is justified for a number of reasons. Leonard et al. [2011]'s study design leaves covariates blank that abstracting doctors do not find in medical records. As doctors are more likely to record abnormal information (Devlin [2021b]), this supports imputation with the normal state 0. Figure 3b demonstrates most of the variables with high rates of missingness indicate conditions that only apply to a specific subset of patients, such as having a CT performed at a referring hospital or wearing a helmet at the time of injury. It seems plausible, for instance, that a patient who is not involved in a sports accident will not have information recorded about their helmet status. This further signals a blocking pattern of randomness and supports zero imputation.

We considered a likelihood based imputation method as well. Treating each covariate as independent, we impute missing entries with random draws from a Bernoulli random variable parameterized by the observed likelihood. We note in Figure 6 that this imputation method removes almost all correlations between covariates. This suggests the information in this dataset is sensitive to false positives, further supporting zero imputation. If we had more time, we would explore likelihood-based imputation which takes advantage of the group-wise missingness. A notable group we impute are the patients without any medical history. This may have occured because their EHRs were not shared with the study site. Dr. Devlin suggested imputing this information with 0 is justified because the vast majority of children do not have abnormal health histories in his experience (Devlin [2021b]), and most indicators of abnormal medical history in this dataset occur in less than 5% of patients.

Some covariates proposed by Leonard et al. are binary cutoffs of continuous or categorical variables, such

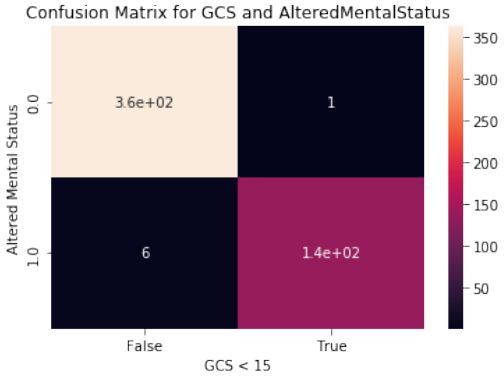


Figure 7: Confusion matrix between AlteredMentalStatus and perfect GCS score in patients with both covariates recorded.

as `HighriskFall` which bins a fall's elevation. Our zero imputation is less justified for these variables because the missing non-binary variables may lie very close to the decision boundary. To avoid these imputation issues on our own derived binary variables, we first impute the source continuous and categorical variables.

Information about a patient's age is never missing. Information about GCS scores is highly missing for many patients. At the suggestion of Dr. Devlin, we investigate `AlteredMentalStatus=0` as an imputation rule for `GCS = 15`. Figure 7 demonstrates that in patients with both of those covariates recorded, this imputation rule leads to a single false negative. We conclude this strategy is well justified. A perfect GCS score also allows us to impute the score's subcategories. After this imputation rule, only 18 patients have GCS still missing. Five of these patients have c-spine injuries, suggesting an imputation of 15 is not justified. Instead we impute the GCS subcategories with the mean of patients who have `AlteredMentalStatus=1`. This leads to an imputed GCS around 13. We consider median imputation as well and the maximum possible GCS scores are the median for patients with `AlteredMentalStatus=1`. We make this choice an automated judgement call for our stability checks. As with AVPU, we add an indicator if GCS was missing before mean or median imputation.

3.4 Relationships with Cervical Spine Injury Outcome

Applying the zero imputation scheme argued for above, we investigate the associations between our binary covariates and the outcome of interest: a c-spine injury. We naturally do not include covariates which may capture information about a patient's outcome. The top left chart in Figure 8 suggests that covariates recorded at a referring hospital have the highest correlation with c-spine injury. This can be explained by selection bias for patients who require transfer between hospitals. As these patients are a minority of units (roughly 20%) and even less likely at more rural hospitals, we are wary that including information from outside hospitals and EMS in our models may bias their performance towards units with this information collected. For this reason we decide to remove `DxCspineInjury`, which indicates if a patient has a diagnosis of a suspected c-spine injury from another hospital, and `ReceivedInTransfer`. Considering the lower left chart, the study site information most correlated with c-spine injuries are almost all analysis variables derived by Leonard et al. [2011]. It is notable that many of the least correlated outcomes are collected during physical examination, denoted by `Pt`. Given that many analysis variables are highly correlated with c-spine injuries, we prioritize them in our modeling.

4 Modeling

4.1 Final Features

After pre-processing and imputing our data, only two groups of covariates are not binary: `AgeInYears` and those related to GCS. We make decisions about encoding this information in binary form after discussing with

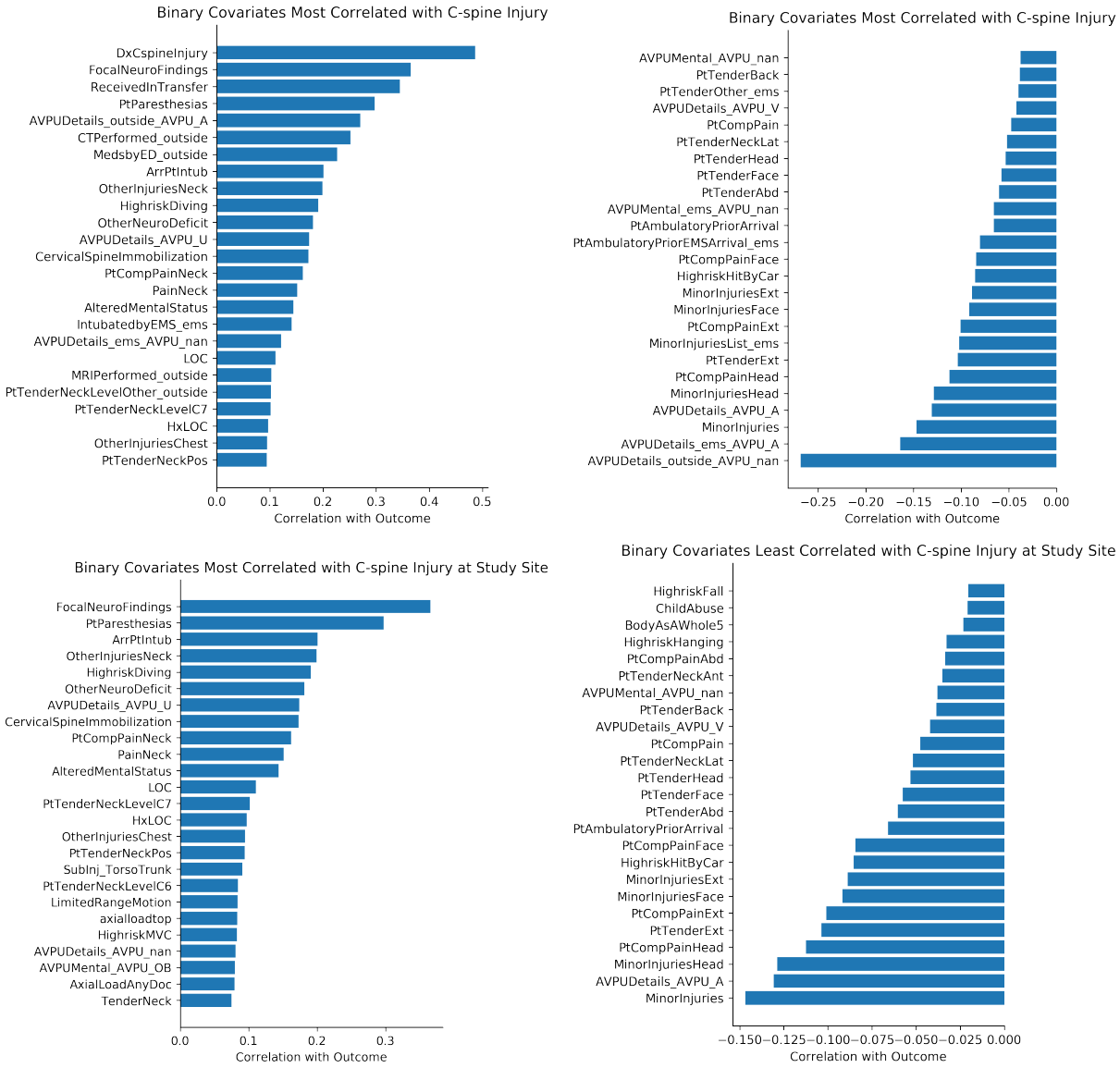


Figure 8: The binary covariates with the largest and smallest Spearman correlation with c-spine injuries. We present all covariates available to ED doctors and only those collected at the study site.

Dr. Devlin and reviewing c-spine decision rules implemented at other hospitals. These decision rules come from a research survey collected and kindly shared by Dr. Devlin.

It is common in c-spine decision rules to have age cutoffs, which we apply to `AgeInYears`. Very young children under the age of 2 are indicated by `VeryYoung`. As previously mentioned, children below 5 have difficulty communicating how their body feels to doctors, and we indicate the age range from 2 to 5 with `NonVerbal`. We use this information to make the indicator for patient complaints of neck pain more robust, as suggested by Dr. Devlin. For verbal `NonVerbal` patients, we expand the covariates which indicate this status to include face, head, and chest pain as well. The demographic EDA above shows that at some study site hospitals, patients over 12 make up more than half of the patients in our dataset. We denote this condition as `YoungAdult`. We make each of these age cutoffs an automated judgement call because there is not much consensus about them between hospitals in the reviewed decision rules.

Almsot all of the c-spine decision rules we reviewed used total GCS scores instead of information from the subcategories. For that reason, we only consider binarization of `TotalGCS`. Two binary indicators that frequently appear in clinical decision rules are any abnormal GCS (GCS below 15) and very low GCS (threshold varies). We encode these cases as `GCSnot15` and `GCSbelowThreshold`. Most decision rules we viewed had a cutoff of 8 or below, but some use 10. We make this an automated judgement call in our data processing.

As the features Leonard et al. [2011] selected have high correlations with c-spine injuries and efficiently summarize the entire dataset, we include all of the non-robust ones. Recall that baseline covariates only include information recorded at the ED, while robust versions indicate if that condition was indicated at the study site or by EMS or at an outside hospital. Dr. Devlin suggested that ED doctors feel more comfortable making decisions based off information they collected themselves, making the robust variables' not so clinically useful.

From the study site, we chose to include covariates that contain information not well summarised by Leonard's analysis variables. These include five covariates which summarize how and in what condition a patient arrived at the study site. We also include indicators for pain in specific parts of the upper body, such as chest or head, which Leonard aggregate; we make a judgment call endorsed by Dr. Devlin to remove extremity and lower body pain indicators. We only consider GCS and AVPU test results collected at the study site, as justified above. Combined with Leonard's analysis variables, our covariates capture all the pre-evaluation information available at the study site short of minor injuries. We only add information collected by EMS or an outside hospital through the condition improved covariates. We make this judgement call to prioritize study site information and also to reduce the number of features which only apply to a subset of patients. This removal choice also extends to covariates about interventions performed before study site arrival: these are rare but highly correlated with c-spine injuries. The `Predisposed` covariate from analysis variables does a very thorough job of summarizing several variables which indicate a predisposing condition; therefore, we only include information about other abnormal health histories. Our final dataset include 94 covariates, of which 64 are available at decision time. A full list and description of these can be found in our `data_dictionary.md` file.

4.2 Question Translation

Our goal is to build decision rules which predict when patients admitted to the ED are at high risk for CSI. Leonard et al. [2011] frame this problem as a binary classification problem with binary covariates. We choose to follow this framework as well because only a handful of predictors are categorical or continuous. Conversion of these factors to binary allows us to implement models which require binary data and compare naturally with the baseline model.

The downside of this approach is that we lose some information when binarizing through cutoffs. We could alternatively have a probabilistic output instead of a hard classification. Logistic regression methods accommodate this, but still requires thresholding of these probabilities to compare against the baseline model. Instead of a binary prediction problem, this task can be reframed as a categorical prediction for type of c-spine injury or intervention received. We do not attempt this because the binary prediction task aggregates these more granular divisions and thus we hypothesize it is easier to learn from data. These alternative translations appear reasonable and would be worth exploring in future work. Due to the binary nature of the recorded outcomes, this problem is not well-suited for continuous prediction.

4.3 Statistical Assumptions

As described above, we make a strong assumption that missing binary variables are not recorded because the condition they indicate is normal. This justifies our mean imputation strategy. Group-wise missingness exists, such as the patients with no medical history and the lack of recorded ethnicity at many study sites. In Figure 3a major missing groups of covariates include substantial injury, factors recorded by EMS, and high risk MOI. These groups of more uncommon events support the unrecorded normal condition hypothesis. Figure 4 indicates that the percent of units with at least one analysis variable missing varies across study sites but is consistent across control types. This is potentially explained by different standards for what information to record about a visit at different hospitals. We use data re-abstracted by another doctor as a stability check, but generally assume that data was correctly abstracted from medical records. These assumptions allow us to model the randomness in our dataset as completely driven by factors inherent to a patient or caused by their injury.

4.4 Tree-based Models

As mentioned earlier, given the importance of detecting CSIs, our model must be highly sensitive. However we must also bear in mind that physicians make decisions based on a variety of factors such as patient condition, domain knowledge, as well as decision rules. Therefore our models must also be highly interpretable to be used to construct or evaluate clinical decision rules. Tree-based models are generally interpretable to people even outside statistics and machine learning, as in the case of the standard Classification and Regression Tree (CART) algorithm. In comparison, more complicated models such as random forests are less interpretable yet are still not in the same category as black box models such as neural networks. While random forests are not as interpretable as a single decision tree, they still offer valuable insight regarding feature importance. Hence we include both CART and random forest in our set of candidate models.

Additionally, given the setting of interest and bearing in mind the results from Leonard et al. [2011], we propose a novel decision tree specifically designed for this problem in this setting. For convenience, we call this special tree and propose a greedy algorithm to fit this model. Our special tree and training algorithm prioritise simplicity and interpretability without compromising model performance, based on results compared to those of our aforementioned models.

4.4.1 Special Tree

We begin by noting the natural relationship between tree-based models and decision rules: in particular, each can be written as the other. In Leonard et al. [2011], the authors performed variable selection using step-wise forward selection based on logistic regression, ultimately proposing 8 high-risk indicators. Their final decision rule is as follows: any patient that has at least one of these high-risk indicators is classified as having CSI. This rule can equivalently be written in the form of an 8-layer decision tree. Such a rule is simple, interpretable, and useful in clinical settings. While the aforementioned decision rule technically falls under the umbrella of tree-based models, most tree-based algorithms such as CART cannot find globally optimal solutions. The greedy algorithms used to train these models favor balanced models, hence they do not easily find asymmetric and deep trees, which the above decision rule maps to.

As such, we examine only a subsection of tree-based models that follow a particular structure that we determine through our EDA and domain knowledge, and fit a model using a greedy algorithm akin to forward selection. In terms of justifying our structure, we assert the importance of age stratification in this context since children in different age categories have vastly different capabilities in terms of motion, speech, and description from each other. Furthermore, the body's physiology changes rapidly throughout childhood. Therefore, we divide the children into four groups based on age: `VeryYoung` (0-2 years old), `NonVerbal` (0-5 years old), and `YoungAdult` (12+ years old), as we briefly touch upon earlier. Within each group, we build a decision tree with the same decision structure as Leonard et al. [2011] i.e. if at least one of the indicators is positive, we classify the patient as having CSI. Figure 9 illustrates the specific structure of our novel special tree algorithm. The first layer splits on the basis of age, and the following layers split on the basis of selected variables. When an indicator is positive, the patient will be labelled as 1 ie. has CSI group, otherwise, the algorithm will proceed to the next variable and split on it. If all the indicators are negative, then the patient will be classified as 0 ie. does not have CSI.

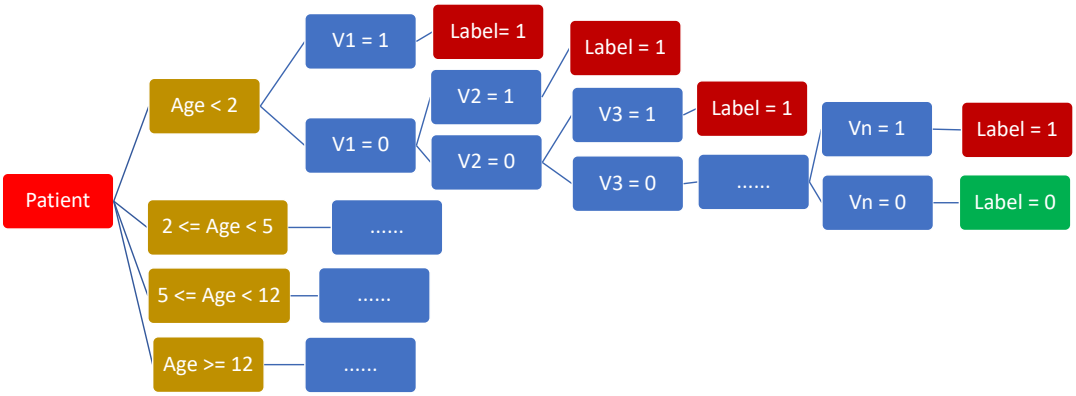


Figure 9: The structure of our novel special tree algorithm: the first layer splits on the basis of age, and the following layers split on the basis of selected variables. When an indicator is positive, the patient will be labelled as 1 ie. has CSI, otherwise, the algorithm will proceed to the next variable and split on it. If all the indicators are negative, then the patient will be classified as 0 ie. does not have CSI.

Even though we consider only a subset of tree-based models with a simplified structure, finding an optimal model is still computationally nontrivial due to the cardinality of our candidate variable set. Therefore to fit our novel tree, we propose a simple greedy algorithm akin to the step-wise forward selection procedure. Specifically at each step, we evaluate all the variables in the current candidate set and select the one with best performance on current unlabelled data, delete it from the candidate set, and classify the observation as positive if the indicator is positive. We then repeat this procedure until either there are no candidate variables or unlabelled observations remaining. With the addition of each variable, we boost sensitivity at the cost of specificity, and to some extent interpretability. Once the procedure has been completed, we generate a Receiver Operating Characteristic (ROC) curve in order to find an appropriate cutoff for classification and accordingly determine which variables are used in our final decision model.

During feature selection, there are different criteria that can be used to evaluate model performance in order to finally choose the variables. Here we primarily consider two scores based on possible classification results. For one score, we directly use the misspecification rate in the 'label = 1' case, and for another we consider the Gini index, which is the same metric that CART and random forests use during feature selection. In this particular context, the first score is a viable metric to use even though it weighs one class more heavily than the other. In particular it is likelier to select features that pose a high risk for CSI but have low prevalence in the data. In comparison, the Gini index is a weighted average of the impurity in both leaves after splitting on a particular feature, so it is less likely to select a low frequency variable in the first place. Bearing in mind sample size considerations, the relationship that we uncover between outcomes and low frequency features might be unreliable. Therefore, we use the Gini index at each step.

4.4.2 Model Performance and Analysis

To evaluate model performance and construct our best model, we randomly split our data to obtain a training and validation set, with the model trained on the former and evaluated on the latter. Furthermore, we consider the decision rules proposed by Leonard et al. [2011] as a baseline model for comparison. Given the heterogeneity of patient capabilities across age groups, we construct and compare models in each age group separately. Four different types of trees we considered are as follows:

- CART and Pruned CART (max-depth = 5);
- Random Forest (number of trees = 200);
- Special Tree.

Figures 10 and 11 visualise our results for patients older and younger than 5 years old, respectively. In each figure, we construct ROC curves for each model performance on both the training and the validation

Group	Training	Validation
12+	388	134
5∼12	261	84
2∼5	113	38
0 2	70	23

Table 2: Sample sizes of training sets and validation sets in each age group

set. We also mark the baseline performance in dashed blue lines to help with comparison. It should be noted that unlike our training-validation split, Leonard et al. [2011] trained their model on the whole dataset. Therefore, all values presented for the baseline model are actually training performance and we cannot ignore the likely over-fitting issue in the baseline. Theoretically, training set performance is not directly comparable with validation set performance, but we still present it for reference. We mainly evaluate models via ROC curves bearing in mind that sensitivity is prioritized over specificity here. Since it is difficult to quantify the overall performance of a model with just one metric in this context, we aim for the overall goal of creating a model that has near perfect sensitivity and relatively high specificity.

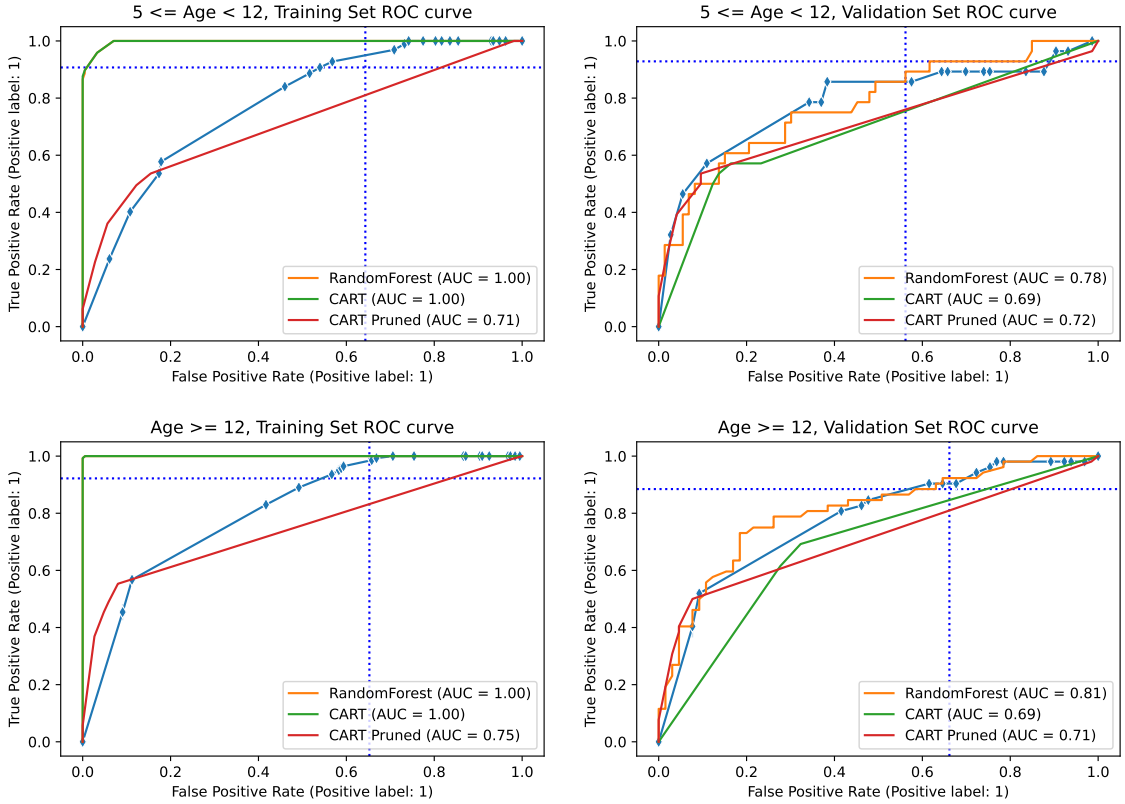


Figure 10: ROC curves for four tree-based methods on training set (left) and validation set (right) for patients older than 5 years old. The blue solid lines with diamond points are performance of our special tree. The blue dashed lines denote the performance of the baseline model.

When our age stratification is based on patients being younger or older than 5 years old, model performance for each of the two groups is similar. From the training set plots in Figures 10 and 11, it is obvious that CART and Random Forest are overfitting with perfect ROC curves. Results for pruned CART and our special tree are more reasonable with the latter performing better than both the former and the baseline. The validation set plots are more faithful representations of true model predictive ability. The standard CART and pruned CART have ROC curves that exhibit substantial similarity, while the performance our special tree closely resembles that of the random forest. In the [5, 12) age group, we see the special tree outperforming the random forest in

some cases. Both these methods generate similar results compared to the baseline, performing slightly worse in the $[5, 12)$ age group and slightly better in the 12+ age group. Bearing in mind that the baseline model results are obtained by training and evaluating a model on the same data, our special tree and random forest models perform quite well.

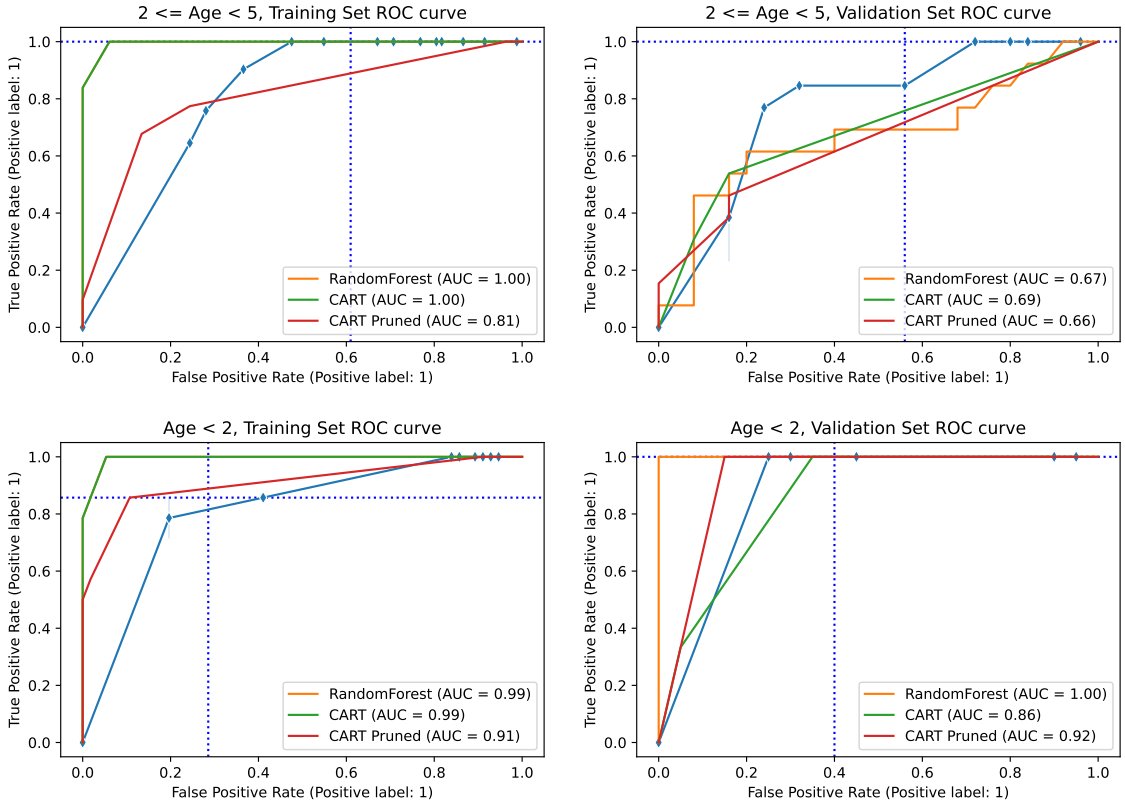


Figure 11: ROC curves for four tree-based methods on training set (left) and validation set (right) for patient younger than 5 years old. The blue solid lines with diamond points are performance of our special tree. The blue dashed lines denote the performance of the baseline model.

We obtain similar results in the case where the patients are younger than 5 years old, but in particular we face the problem of small sample sizes. Table 2 shows the sample size of the training set and the validation set in each age group. Given the small sample size for patients younger than 5 years old, the variables selected during training may be those that are present with low frequency. Since the validation sets are even smaller than the training sets, we cannot rely on the ROC curve as much as in the previous case. However even in this setting we note that our special tree and random forest outperform CART.

Comparing our different tree-based methods, we find that in terms of model performance, $ST \approx RF >> CART$, while from an interpretability standpoint, we see that $ST >= CART >> RF$. Considering the importance of both factors in this setting as discussed previously, we decide to use special tree as our final model. Having finalised our model, we train it on the entire dataset ie. no training-validation split, in order to alleviate the aforementioned issue of small sample sizes after age stratification.

It is noteworthy that our novel special tree algorithm gives results comparable to those of far more complex models such as random forest while preserving interpretability. This suggests that only certain sub-structures within the overall structure of the random forest have learned the underlying signal in the data while the rest of the model is overfitting. While random forests are trained using greedy algorithms, they explore more possible splits in feature space compared to CART, and are more stable due to the construction of a forest rather than relying on a single tree. On the other hand, we evaluate the performance CART, which is also trained using a greedy algorithm. We note that in our results, the CART model performs suboptimally even compared to the random forest which also trained using a greedy algorithm. In comparison, our novel special tree model proposed in this report uses both domain knowledge and human judgement calls in order to constrain the

tree structure, after which we fit it greedily. Overall, we achieve high model performance while preserving model interpretability. Thus we see that rather than becoming overly reliant on pre-existing algorithms and approaches, we are able to optimize for both performance and interpretability by focusing on features relevant to the the target problem and integrating domain knowledge throughout our modelling process.

4.5 Logistic Regression

In addition to tree-based methods, we also explore extensive logistic regression models, bearing in mind their use by Leonard et al. [2011]. While certain regularised logistic regression models perform well on the training and validation set with AUC exceeding 0.8, the models are not able to perform as well as either the baseline model or the tree-based models we have already examined. Furthermore there is the issue of interpreting feature importance in the context of logistic regression models. On one hand standardising the data leads to better performance in terms of AUC since features with different prevalences exhibit substantially different variances, yet we are not able to a priori justify standrardising binary data in this way. Thus bearing these considerations in mind and for concision's sake, we do not include the bulk of our work with logsitic regression models here, but provide our analysis notebooks for reference.

5 Analysis of the Best Model

5.1 Model description

We train our novel special tree model on the entire training data (without validation split), and select a proper cutoff by examining the training ROC curve. We describe our methods in greater depth later in the report. As mentioned before, in this setting sensitivity is much more important than specificity, and as such we choose a set of variables for each age group such that the overall model‘ sensitivity is greater than 95%. The selection results are listed below:

- $Age < 2$: `AlteredMentalStatus`, `PosMidNeckTenderness`, `Predisposed`, `AxialLoadAnyDoc`, `EMSArrival`;
- $2 \leq Age < 5$: `AlteredMentalStatus`, `FocalNeuroFindings`, `Torticollis`, `Predisposed`, `HighriskMVC`;
- $5 \leq Age < 12$: `FocalNeuroFindings`, `GCSbelowThreshold`, `Torticollis`, `PainNeck`, `Clotheslining`, `Predisposed`, `AxialLoadAnyDoc`, `HighriskFall`, `HEENT`;
- $Age \geq 12$: `FocalNeuroFindings`, `HighriskDiving`, `GCSbelowThreshold`, `PainNeck`, `SubInj_Head`, `axialloadtop`, `TenderNeck`, `HighriskMVC`.

Figure 12 shows ROC curves of our special tree model on the training set. The variables selected at each step are marked by black points. The red dashed lines denote the 95% threshold for sensitivity. This plot illustrates the order in which variables are selected and also shows how sensitivity and specificity change upon the addition of a new variable. The algorithm terminates when the sensitivity is greater than 95% and classifies the remaining patients as 0 or does not have CSI.

Further examining the plot, we note that the four age groups are substantially different from each other, especially when examining differences between patients older and younger than 5 years old, respectively. The most important variable for younger age groups (Age < 5) is `AlteredMentalStatus`, while `FocalNeuroFindings` is the most important one for children older than 5 years old. `GCSbelowThreshold` and `PainNeck` are selected as important factors for the older (Age ≥ 5) age groups but not the younger ones. This may be due to the inherent difficulty in accurately measuring GCS for non-verbal children, and also because it is harder for the, to describe and complain about the pain in their neck. Additionally, `Predisposed` is selected in groups younger than 12 years old and `FocalNeuroFindings` is selected in groups older than 2 years old. These results again justify our decision to stratify child patients into four age groups.

We further note that within the set of selected variables, the variables affect the overall model performance in different ways when they are used in making the final decision. In particular, these variables primarily act in one of two ways: when high-risk factors like `Predisposed`, `PosMidNeckTenderness` and `HighriskDiving` are added to the model, the ROC curve will have a steep but minor increase, which corresponds to an improvement

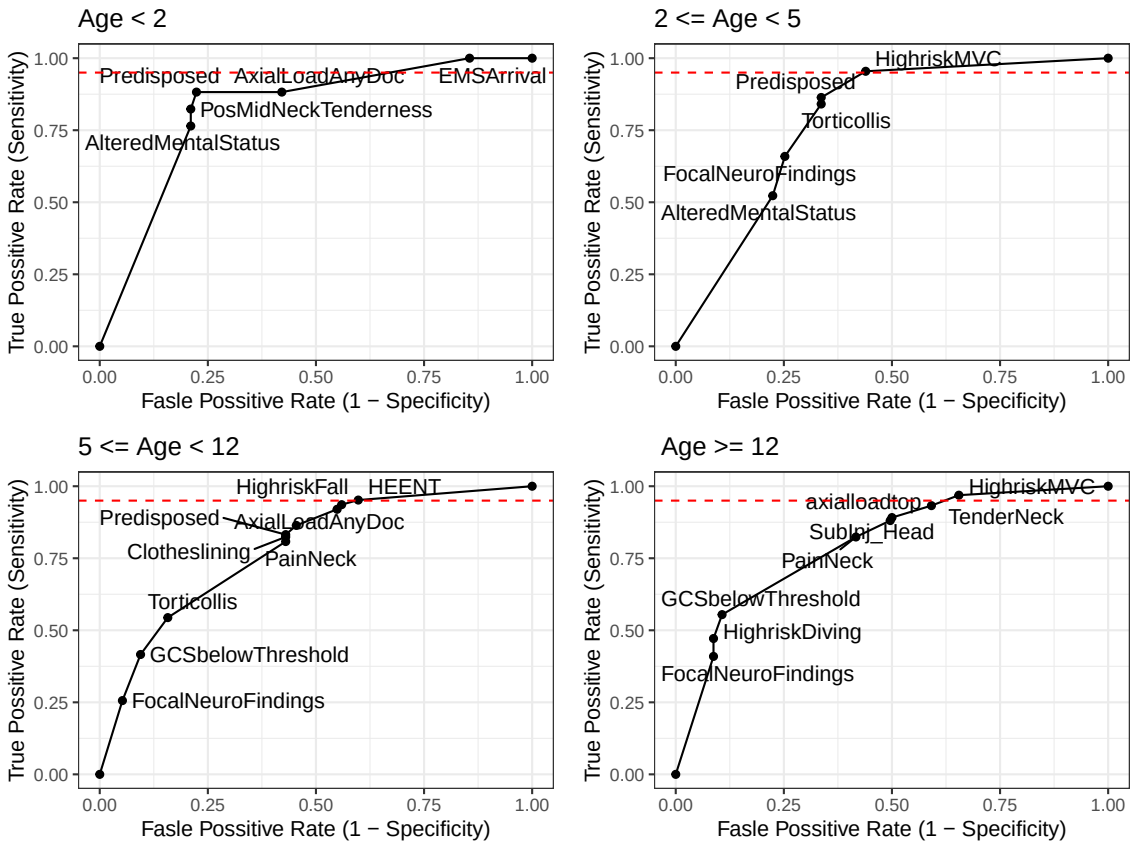


Figure 12: ROC curves for our special tree model evaluated on the training set. Variables selected at each step are marked by black points. The red dashed lines denote the 95% threshold for sensitivity.

in sensitivity without a substantial decrease in specificity. This indicates that while these factors are not highly prevalent, when they are present there is a high risk of CSI.

The other way variables act in the context of our model is in the case of common risk factors like `PainNeck`. When adding such variables, the ROC curve will have a relatively flat increase but make a large step along the x-axis, meaning such features substantially increase sensitivity but does so with a nontrivial decrease in specificity. These variables are more prevalent in the data than the aforementioned high risk variables. While these indicators in and of themselves are not suggestive of a high risk for CSI, patients without these indicators are far less likely to have CSI. By using the Gini index in our forward selection procedure, we are able to capture the effect of these two different kinds of variables in order to devise reasonable and well balanced decision rules.

5.2 Comparison with Leonard et al. [2011]

With the exception of age stratification, both our final decision rules and the baseline model proposed in Leonard et al. [2011] follow the same structure, wherein if any of the features is positive, the patient is classified as having CSI. Bearing this in mind, we carefully compare and contrast the two models in this section.

5.2.1 Model Performance

In terms of model performance, we evaluated our model both on the training set to mimic the evaluation done by Leonard et al. [2011], and on an untouched testing set to obtain results not affected by overfitting. The training performance of our special tree model is 96.8% sensitivity with 36.9% specificity, and the testing performance is 91.3% sensitivity with 34.22% specificity. For the baseline model, the numbers are 92.1% sensitivity and 40.22% specificity, and 96.1% sensitivity and 42.2% specificity on the training and testing set, respectively. When evaluating model performance on the training set, their model has higher specificity. If we lower our sensitivity threshold, we can obtain both higher sensitivity and higher specificity than their model, but given the severity of outcome for a false-negative case, we opt for a 95% cutoff.

Note that the data we use as the test set for our model is a subset of the data that the baseline model was trained on. Thus in both situations, the results we obtain for the baseline are actually its training set performance, and therefore we cannot evaluate its predictive performance on unseen real world data.

5.2.2 Variable Selection

Table 3 lists all the variable selection results. In particular, the first column lists all the variables considered during model construction. The second to fifth columns mark variables used in the four respective age groups. The sixth column marks variables used in the baseline model, and the last column marks all the analysis variables considered in Leonard et al. [2011].

We consider 35 variables in total, including 22 analysis variables proposed in the original paper, and 13 variables summarised from our data prepossessing procedure. We note a strong overlap between the selected variables and the analysis variables but find that three variables, namely `GCSbelowThreshold`, `EMSArrival`, and `HEENT` that are selected in our model are not analysis variables. `GCSbelowThreshold` in particular is quite a useful indicator in the older patient groups that we analyze. We note that all 8 variables considered in the baseline model proposed by Leonard et al. [2011] are used in the special tree model for at least one age group. The most frequently used variables are `FocalNeuroFindings` and `Predisposed`, with our model also including some other analysis variables that are not in the baseline model, such as `AxialLoadAnyDoc`, `axialloadtop`, `Clotheslining`, `HighriskFall`, `PostMidNeckTenderness`, `SubInj_Head`, and `TenderNeck`, among which `AxialLoadAnyDoc` is used in more than 2 age groups.

Thus, our special tree model and baseline model are at least somewhat consistent. Our model improves upon the baseline model in that we consider a wider range of variables, finding `GCSbelowThreshold` to be a useful feature even though it is not considered in the original paper. Furthermore, our model uses age stratification bearing in mind heterogeneity in data across different age groups and therefore proposes more specific decision rules.

Variable List	Age $\in (0, 2)$	Age $\in [2, 5)$	Age $\in [5, 12)$	Age $\in [12, 18)$	Baseline	AnalysisVariable
AlteredMentalStatus	+	+			+	+
Assault						
AxialLoadAnyDoc	+		+			+
axialloadtop				+		+
ChildAbuse						
Clotheslining			+			+
EMSArrival	+					
FocalNeuroFindings		+	+	+	+	+
GCSbelowThreshold			+	+		
GCSnot15						
HEENT			+			
helmet						
HighriskDiving				+	+	+
HighriskFall			+			+
HighriskFallDownStairs						
HighriskHanging						+
HighriskHitByCar						+
HighriskMVC		+		+	+	+
HighriskOtherMV						+
LOC						+
Medications						
Musculoskeletal						
Neurological						
NonAmbulatory						+
PainNeck			+	+	+	+
PassRestraint						
PosMidNeckTenderness	+					+
Predisposed	+	+	+		+	+
Respiratory						
SubInj_Ext						+
SubInj_Face						+
SubInj_ Head				+		+
SubInj_TorsoTrunk			+		+	+
TenderNeck				+		+
Torticollis		+	+		+	+

Table 3: Variable selection results. The first column lists all variables considered during model construction. The second to fifth columns mark variables used in the four age groups. The sixth column marks variables used in the baseline model, and the last column marks all the analysis variables considered in Leonard et al. [2011].

5.3 Post-EDA

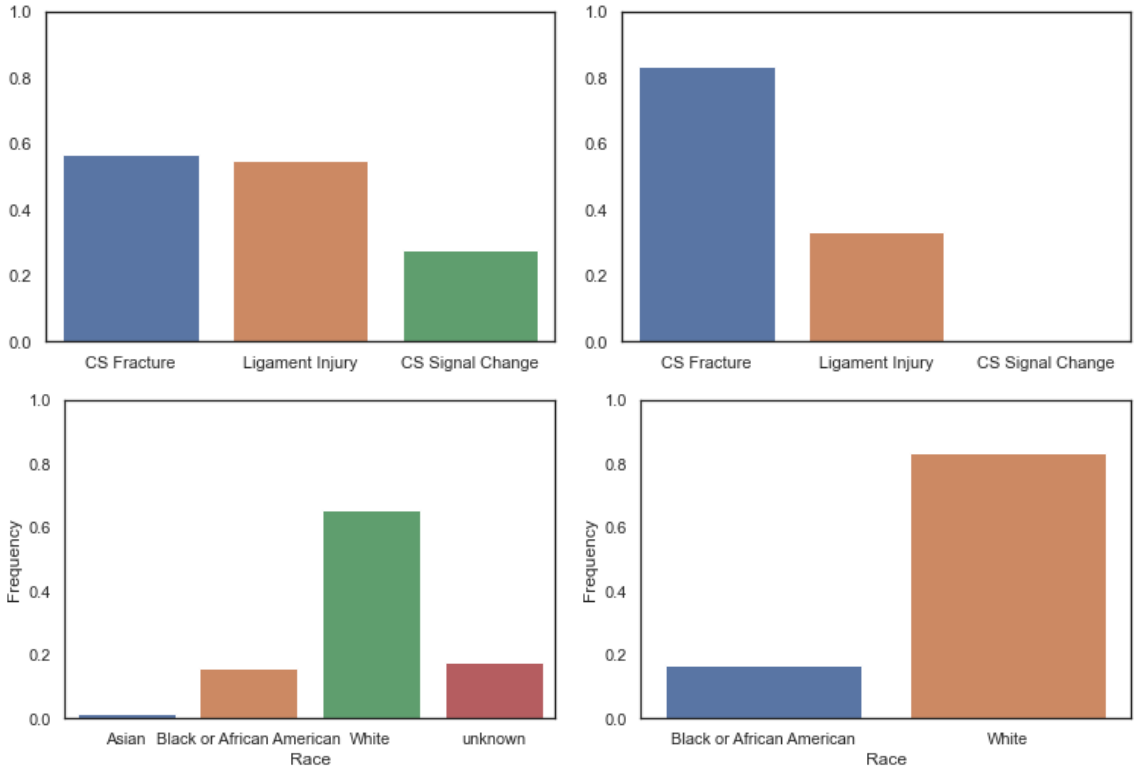


Figure 13: Post-hoc analysis of our final model performance on unseen test data

We now conduct post-hoc model evaluation on unseen test data in Figure 13. We note that a majority of false negative cases are associated with a cervical spine bone fracture, while our model correctly flags all patients who have spinal cord damage as having a CSI. Thus the model may be improved by considering features that in particular may be associated with cervical spine fractures. To find such features, we will consult with domain experts. In terms of race, since there are only 6 patients that are classified as false negative, we are not able to draw substantial conclusions based on how race may or may not be associated with medical care in the case of false negatives, and thus leave this for future work.

6 Stability Tests

Guided by the PCS framework, we analyze the stability of predictions made by our proposed special tree and Leonard et al.'s baseline models. To quantify stability, we use each model's sensitivity and specificity on unseen data. There are two main classes we check against: researcher judgement calls and data perturbation.

6.1 Stability with Respect to Researcher Judgement Calls

The researcher judgement calls we can analyze for stability are somewhat limited because both the baseline and proposed special tree use a fixed set of covariates. The binarized age and total GCS variables are used. For each of the age cutoffs, we test two values which show up frequently in decision rules. For GCS, we vary the cutoff low score as either 8 or 10 and impute with either mean or median. Because the baseline model uses analysis variables which we do not perturb, the only judgement call which affects these results is dropping missing units with 0, 1, or 2 analysis variables missing. In total, there are 96 possible combinations of judgement calls. Figure 14 demonstrates that the baseline model has higher mean sensitivity by several percentage points. The range of sensitivity is similar. Surprisingly, the baseline has higher average testing sensitivity than training sensitivity. The baseline model has consistently larger specificity with a very tight

range of possible values. The minimum testing baseline specificity across perturbations is larger than the maximum special tree specificity.

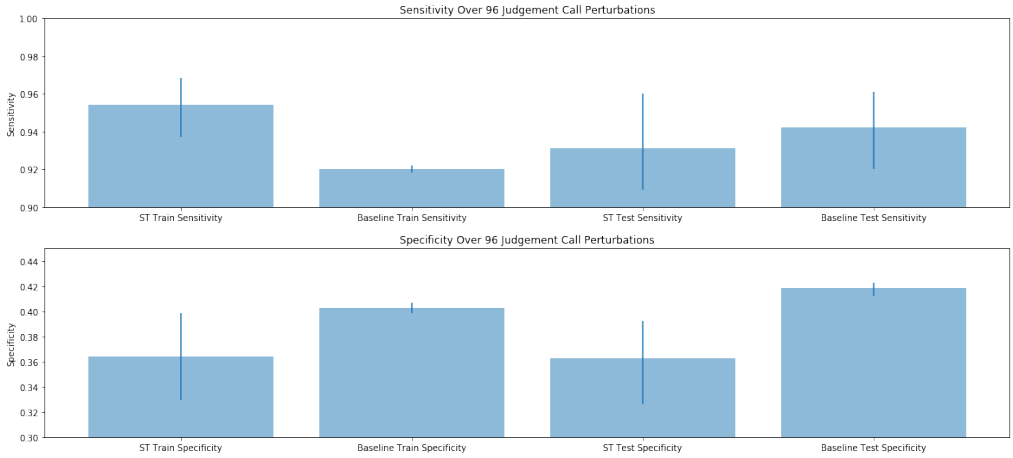


Figure 14: This bar chart displays the mean sensitivity and specificity of the baseline and special tree models over our judgement call perturbations. The error bars indicate the extreme values.

6.2 Stability with Respect to Data Perturbation

A natural way to evaluate stability is with controls matched by EMS arrival and mechanism of injury, up to now unused. The EMS control group matches case patients who arrived by ambulance with control patients who did as well and are within a year of age; the MOI group matches on that covariate as well as age. Other covariates are not matched. If possible, multiple controls are assigned to each case. We first ignore the matched structure and compare specificity across these control types. Note sensitivity will be constant across these experiments because it is a function of the cases, which are always included. Figure 15 demonstrates that the baseline model has larger training and testing specificity than our special tree model for all control groups.

To incorporate the case-control matched structure into our analysis, we summarize each group by two indicators: case patient predicted as injury and at least one control predicted as injury. These indicators allow us to derive which groups had all units correctly predicted and confusion matrices by treatment arm. In Figure 16 we present the two models matched confusion matrices for the MOI controls. We note that the baseline model has 2.7% higher sensitivity and almost never incorrectly classifies both the case and its controls. The baseline model labels nearly half of matched groups completely correctly, a significant improvement over the special tree. This difference is largely comprised of controls predicted with injuries, indicating the the baseline has higher specificity as well. These observations about specificity hold for the EMS controls and when the MOI and EMS control groups are merged. The corresponding confusion matrices can be generated by running the notebook `matched_control_analysis.ipynb`.

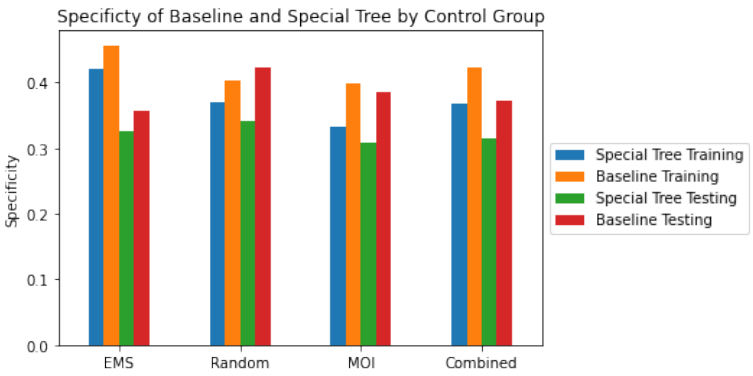
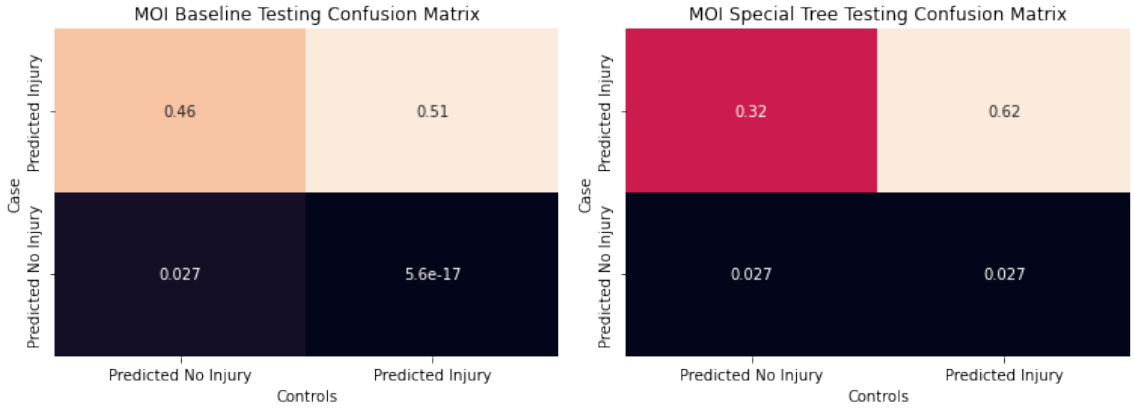


Figure 15: The specificity for both our special tree model and the baseline model for different control groups. Controls are pooled and the matched structure is disregarded.

Another natural data perturbation involves the 365 units re-abstracted by another doctor to calculate a kappa statistic for interrater reliability. Using the random controls we originally evaluate the models with,



(a) Baseline model

(b) Special Tree model

Figure 16: These confusion matrices summarize the two models performance within matched mechanism of injury case-control groups created by Leonard et al. The results are similar of EMS arrival matched groups. In the event of multiple controls matched to a case, the injury prediction indicator is 1 when any control is incorrectly predicted.

365 units have 18 covariates replaced. These covariates include GCS score, indicators of localized pain on physical examination, and injury mechanism details. This perturbation affects 33, just over 10%, of units in the testing set. We find this perturbed dataset increases the baseline models sensitivity and specificity by 4% and 2% respectively, while the special tree decreases by over 5% and 2% for these metrics. This result provides evidence for the baseline model's robustness. With more time to work on this project, we would explore bootstrap-based model perturbation.

7 Conclusion

In this project we explored interpretable decision rules for the prediction of cervical spine injuries in children from medical records. We utilized over 3000 medical records abstracted into covariates by Leonard et al. [2011] and compared our results against their proposed decision rule. We were able to find tree-based decision rules which achieve slightly better sensitivity, but were not able to improve upon their specificity. We also explored logistic regression-based approaches. We find both our proposed best model and the baseline a very stable with respect to researcher judgement calls and data perturbation. This is an import finding because Leonard et al. [2011] do not use withheld data to originally fit their model.

7.1 Contributions

WT focused on building the data pipeline and selecting the most relevant covariates from the full dataset. He explored missing data and imputation strategies, and also ran stability checks. YH mainly focused on tree-based modeling part, including realizing CART, RF, baseline model and proposed Special Tree. And she constructed the final model and did some corresponding analysis. IS primarily focused on generating profiling reports for datasets in EDA. He then explored a variety of logistic regression models for feature selection and to prevent overfitting. He then conducted post hoc analysis and focused on writing and editing the modeling section of the report.

8 Acknowledgements

We are deeply grateful for the domain expertise, feedback, suggestions, and time shared by Dr. Gabriel Devlin and Dr. Aaron Kornblith. We would also like to thank Chandan Singh for building a data pipeline template

and answering related questions, as well as Omer Ronen for helping with debugging and answering other questions.

References

J. X. Chen, B. Kachniarz, S. Gilani, and J. J. Shin. Risk of malignancy associated with head and neck CT in children: a systematic review. *Otolaryngol Head Neck Surg*, 151(4):554–566, Oct 2014.

Gabriel Devlin. Email 12/06, December 2021a.

Gabriel Devlin. Group Meeting 12/03, December 2021b.

Gabriel Devlin. Group Meeting 12/10, December 2021c.

Gabriel Devlin. Group Meeting 11/19, November 2021d.

Julie C Leonard, Nathan Kuppermann, Cody Olsen, Lynn Babcock-Cimpello, Kathleen Brown, Prashant Mahajan, Kathleen M Adelgais, Jennifer Anders, Dominic Borgialli, Aaron Donoghue, et al. Factors associated with cervical spine injury in children after blunt trauma. *Annals of emergency medicine*, 58(2):145–155, 2011.

L. E. Nigrovic, A. J. Rogers, K. M. Adelgais, C. S. Olsen, J. R. Leonard, D. M. Jaffe, J. C. Leonard, K. A. Lillis, P. Mahajan, C. Stankovic, A. Donoghue, K. Brown, S. D. Reeves, J. D. Hoyle, D. Borgialli, J. Anders, G. Rebella, E. C. Powell, E. Kim, N. Kuppermann, L. Babcock-Cimpello, C. Bhogte, G. Teshome, K. Call, J. M. Dean, R. Enriquez, R. Holubkov, B. Yu, S. J. Zuspan, N. Kuppermann, E. Alpern, D. Borgialli, K. Brown, J. Chamberlain, J. Dean, G. Foltin, M. Gerardi, M. Gorelick, J. Hoyle, C. Johns, K. Lillis, P. Mahajan, R. Maio, S. Miller, D. Monroe, R. Ruddy, R. Stanley, M. Tunik, A. Walker, D. Kavanaugh, M. Dean, R. Holubkov, S. Knight, A. Donaldson, S. Zuspan, T. Singh, A. Drongowski, L. Fukushima, M. Shults, J. Suhajda, M. Tunik, S. Zuspan, M. Gorelick, E. Alpern, G. Foltin, R. Holubov, J. Joseph, S. Miller, F. Moler, O. Soldes, S. Teach, A. Cooper, J. Dean, C. Johns, R. Kanter, R. Maio, N. Mann, D. Monroe, K. Shaw, D. Treloar, R. Stanley, D. Alexander, J. Burr, M. Gerardi, R. Holubkov, K. Lillis, R. Ruddy, M. Shults, A. Walker, W. Schalick, J. Brennan, J. Burr, J. Dean, J. Hoyle, R. Ruddy, T. Singh, D. Snowdon, and J. Wright. Utility of plain radiographs in detecting traumatic injuries of the cervical spine in children. *Pediatr Emerg Care*, 28(5):426–432, May 2012.