Learning and Evaluating Clinical Decision Rules for Cervical Spine Injuries

Jaewon Saw, Jeffrey Cheng, Ahmed Eldeeb, and Kaitlin Smith

December, 2022

1 Introduction

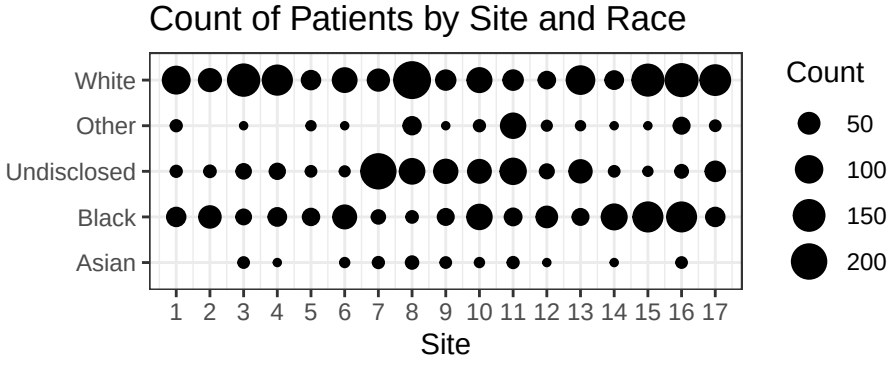
The goal of this project is to create a clinical decision rule to identify children who are most likely to have a cervical spine injury (CSI). The adverse effects of immobilizing children and subjecting children to ionizing radiation motivates such a rule, as there is a desire to minimize the number of children unnecessarily subject to radiographic assessment while continuing to maintain high sensitivity.

2 Data

2.1 Data Collection

The data was taken from the Pediatric Emergency Care Applied Research Network (PECARN) public use dataset titled "Predicting Cervical Spine Injury (CSI) in Children: A Multi-Centered Case-Control Analysis". A total of 17 PECARN sites participated in the study, with a total of 3,314 subjects included in this dataset. This dataset was collected between January 2000 and December 2004 and was initially procured for the purpose of creating a decision rule for identifying factors associated with CSI. The results for this study are presented in "Factors Associated With Cervical Spine Injury in Children After Blunt Trauma" by Leonard et al.

Of the 3,314 records, 540 are deemed positive cervical spine injuries from radiology reports or spine consultation. These positive injury records were verified by the principal investigator of Leonard et al. and by a pediatric neurosurgeon[Lenoard et al]. The remaining 2,774 controls fall into three control groups: 1,060 unmatched random controls, 1,012 mechanism-of-injury and age matched controls, and 702 age-matched EMS controls.



In the figure above, it is clear that the patients are not equally distributed across PECARN sites, as the racial distribution of patients varies visually across sites.

We also used the same features developed by Leanoard et al. These features provide data on how the patient was injured, any pre-existing condition the patient may have, and the symptoms of the patient, which are defined in Table 2 of Leonard et al. Leonard et al. determined 6 major variables that are associated with CSI: altered mental status, focal neurological deficit, complaint of neck pain, substantial injury to the torso, high-risk motor vehicle crash, and diving. Our analyses focused on these factors, as these were the only clinical variables that were provided for all 3,314 records. Each variable is not directly comparable with each other, but the variables have already been one-hot encoded in the PECARN dataset.

2.2 Cleaning

From the original PECARN public use dataset, the amount of cleaning depended on the classifier used. The baseline rule from Leonard et al. and decision trees were both tolerant of missing data, so no additional cleaning was done for these processes. The same applies to the tree and stub based approaches we explored. For the lasso selection of variables and logistic regression, records with missing values were removed from the analysis.

2.3 Training and Evaluation Split

We assume that in practice, our decision rule will be deployed at hospitals not included in the dataset. Then, to simulate the performance of the models in practice, the data was split into test and training sets based on sites. Sites 5, 16, and 17 were randomly chosen as the evaluation sites. Leave-one-out-cross-validation (LOOCV) was conducted over the remaining sites during training.

3 Modeling

3.1 Replicate Research Findings

First, we implemented the clinical decision rule found in Leonard et al. This study determined a clinical decision rule by selecting features using forward selection with logistic regression. For each iteration of the forward selection process, the algorithm would add the feature whose p-value from the Chi-squared statistic was smallest when that feature is added to the model, versus the p-value from when any other feature was added. The algorithm stopped adding features when no additional feature resulted in a p-value less than 0.05. 1000 samples were then bootstrapped and a new logistic regression model was computed each time. A covariate was included in the final decision rule if it appeared in over 50% of the bootstrapped models. This process was repeated for each control group: random, EMS, and mechanism of injury controls. The forward selection logistic regression models identified 6 common covariates between these 3 models: altered mental status, focal neurological deficit, complaint of neck pain, substantial injury to the torso, high-risk motor vehicle crash, and diving. The decision rule then classified a patient as likely to have cervical spine injury if any one of these factors was true, otherwise the patient was ruled to not have a cervical spine injury.

Although the report did not specify how missing values are handled, we removed samples where any of the covariates were missing, since our protocol in R was to remove NA values in logistic regression. We calculated this decision rule to have the following metrics, where patients from all 3 control groups were included:

Metric	Estimate	Lower CI Bound	Upper CI Bound
Sensitivity	0.906	0.880	0.933
Specificity	0.405	0.384	0.426

Although the value for sensitivity is 2% lower and the value for specificity is 5% higher than presented in the Leonard et al. paper, we were not able to exactly replicate results due to assumptions we had to make about handling missing values, and which control groups to include in the final calculations. However, this replication gave us a baseline to compare our own model results to.

3.2 Modeling Approach

In our own modeling efforts we tried several classification methods coupled with two main approaches for features selection. Here we present two of the modeling approaches that we found amenable to interpretability and stability analysis, namely a single decision tree and linear logistic regression. For all of our modeling experiments we employed k-fold Cross Validation, with the folds determined by site (leaving one site out for each fold). We first outline our feature selection approaches, then present out classification results.

3.3 Feature Selection

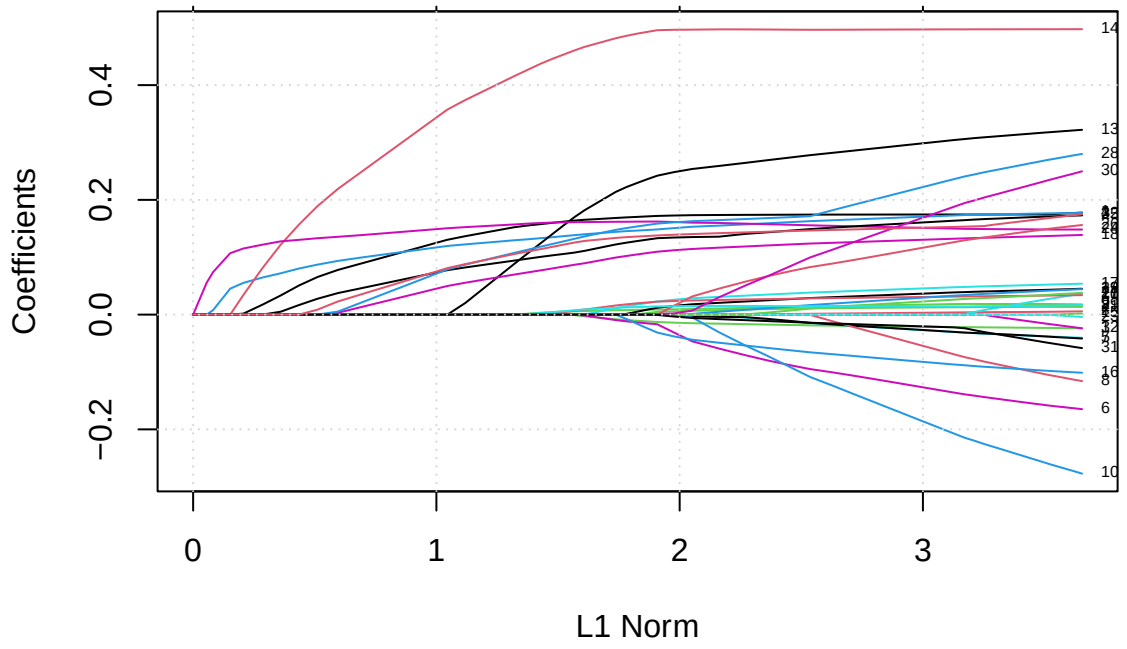
3.3.1 Bootstrapped Forward Selection

First, we selected features using forward selection with logistic regression, similar to the method used by Leonard et al. We started with an empty model, and added features sequentially, including the feature with the smallest p-value each iteration. However, we stopped when there was no feature with a p-value less than 0.15, instead of 0.05 used in the feature selection method by Leonard et al. We then proceeded with the same bootstrapping procedure, selecting features that appeared in over 50% of the bootstrapped models.

3.3.2 Lasso Logistic Regression (L1 regularization)

Next, we selected features from a Lasso logistic regression model. First, we completed 10 fold cross validation to find the value of *lambda* that minimizes the $L1$ loss for the training data. We then selected the features from this model that had non-zero regression coefficients.

In the following graph, you can see the order coefficients are added to the Lasso model, which provide insights into which features are most important to the final probability.

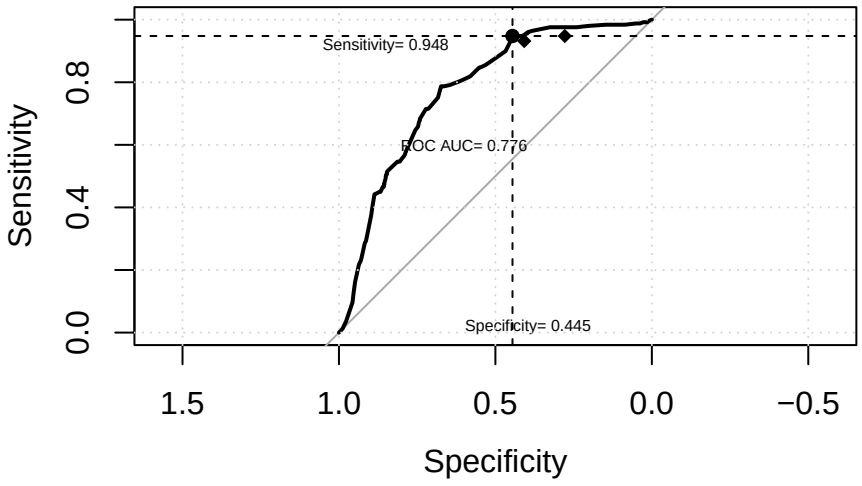


Order	Variable
1	FocalNeuroFindings2
2	FocalNeuroFindings
3	AlteredMentalStatus
4	HighriskDiving
5	PainNeck2
6	subinj_TorsoTrunk2
7	HighriskMVC
8	Torticollis2
9	Predisposed
10	SubInj_Head
11	subinj_Head2
12	AxialLoadAnyDoc
13	ambulatory
14	PosMidNeckTenderness
15	HighriskHitByCar
16	HighriskHanging

3.4 Classification Experiments

The first classification approach we will present is the single decision tree approach. For that we used the rpart R package [Therneau], and compared the cross validation results with the results from the Leonard et al paper.

3.4.1 Single Decision Tree Results

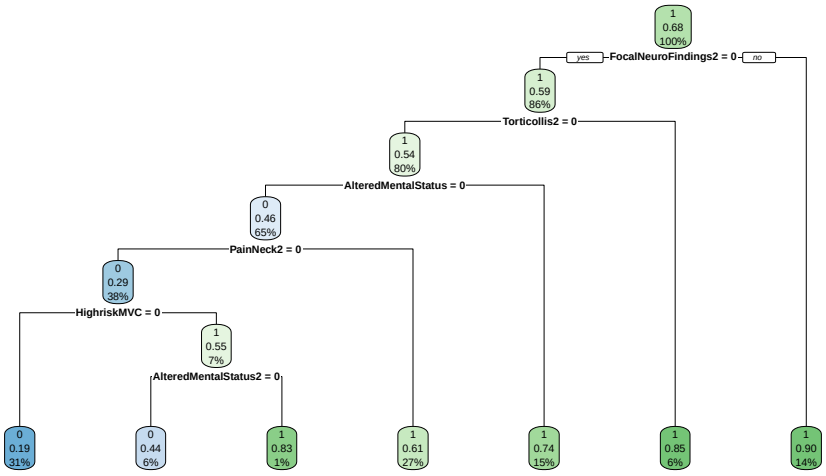


As we can see, classification tree prodcues an ROC curve with AUC = 0.78. The curve is strictly above the decision rule replicated from the published decision rule. Further selection of an appropriate decision threshold produces the following sensitivity/specificity metrics:

	Estimate	Lower CI Bound	Upper CI Bound
Sensitivity	0.948	0.920	0.975
Specificity	0.445	0.418	0.473

The list of predictors used and the decition tree itself are the following:

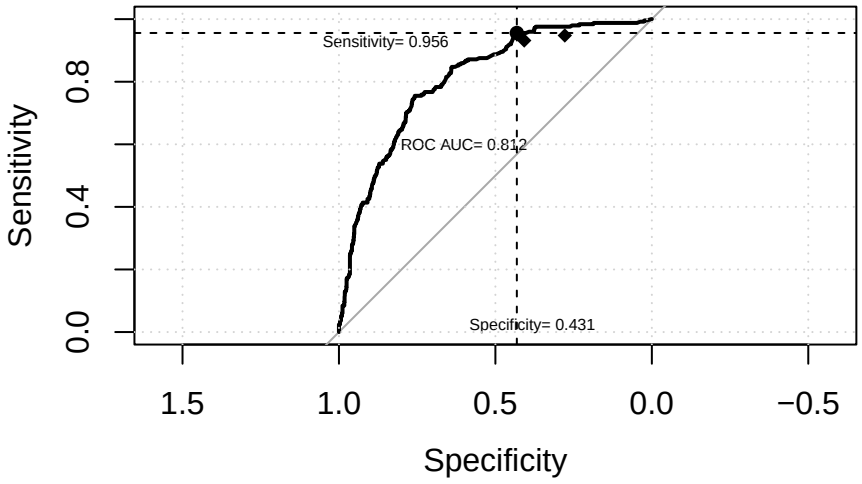
Used Covariates
FocalNeuroFindings2
Torticollis2
AlteredMentalStatus
PainNeck2
HighriskMVC
AlteredMentalStatus2



One notable advantage to using decisin tree induction, is that most tree induction algorithms can handle missing data, so for the purpose of this classification model we used the whole dataset without having to remove rows with missing data, which should lend the algorithm more statistical power.

3.4.2 Logistic Regression Results

Next we present the results from simple logistic regression, performed on a set of covariates chosen by the feature selection approaches we outlined above. We combined the sets of covariates selected by forward selection with the set of covariates selected by LASSO and used the intersection of the two sets for higher stability.



Again, the logistic regression produced an ROC curve that was strictly dominant to the published decision rule (according to our CV) with AUC = 0.81. A choice of a suitable decision threshold produced the following sensitivity/specificity metrics:

	Estimate	Lower CI Bound	Upper CI Bound
Sensitivity	0.956	0.930	0.981
Specificity	0.431	0.404	0.459

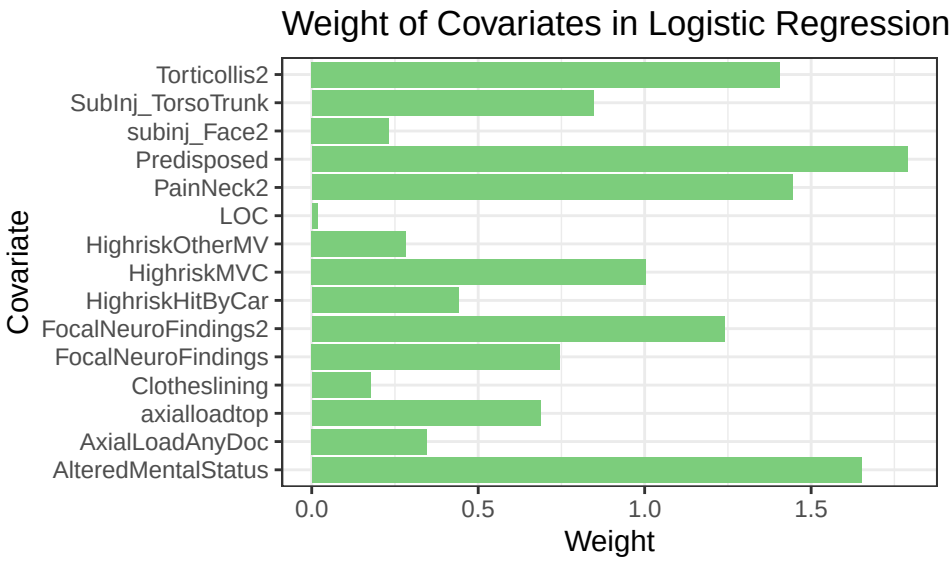
The list of predictors used for the logistic regression model is:

	Parameter Estimate
(Intercept)	-3.855
AlteredMentalStatus	1.680
FocalNeuroFindings	0.870
HighriskHitByCar	0.578
HighriskMVC	1.188
PainNeck2	1.502
Predisposed	2.125
SubInj_TorsoTrunk	1.024
FocalNeuroFindings2	1.156
axialloadtop	0.414
Torticollis2	1.298
LOC	0.001
HighriskOtherMV	0.406
Clotheslining	0.589
subinj_Face2	0.288
AxialLoadAnyDoc	0.471

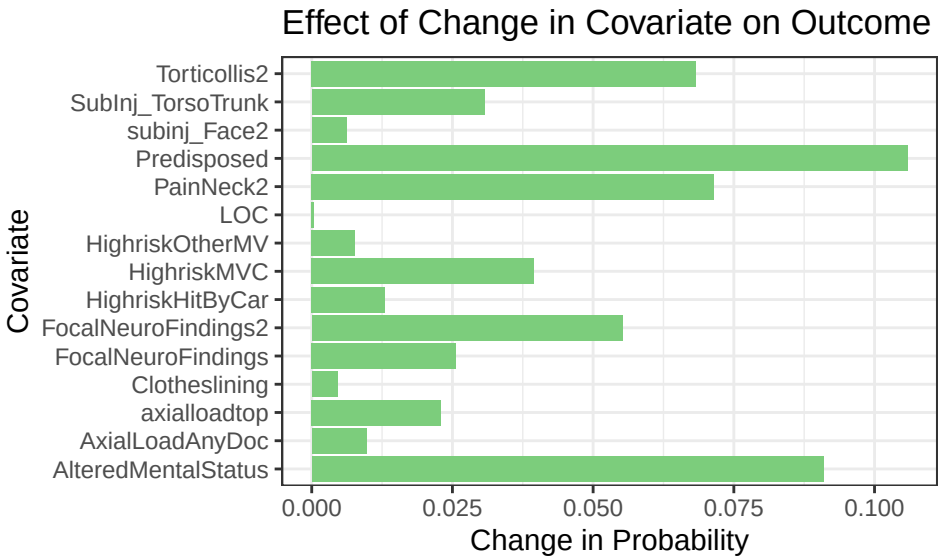
In addition to these two approaches we have experimented with neural networks and gradient boosting (on stubs and trees). Both approaches yielded results that were not much better than the approaches we presented here, and produced models that were more difficult to interpret, so we chose to omit them from this report.

3.5 Interpretation

Our model benefits from the interpretability of logistic regression. When a patient is inferenced using the model, the model gives a probability of that patient having a cervical spine injury. That said, due to the necessity for high sensitivity in the model, we declare any patient with a probability of an injury greater than 7.9% as a patient who needs further imaging. It is also possible to view the weights of the features, which demonstrate which features contribute the most to the final probability of injury:



It is important to note that the values of the coefficients in logistic regression cannot be interpreted the same way the coefficients in linear regression can be interpreted. An increase in any covariate by 1, does not results in linear change in the output by the value of B_j, where B_j is the value of the coefficient for that variable in the model. Hence, an increase in any of the covariates by one would then yield an increase in the odds ratio by $exp(B_j)$, where B_j is the weight for that particular covariate. However, coefficients with greater weight still can have a greater effect on the output probability from logistic regression, which we will examine below:



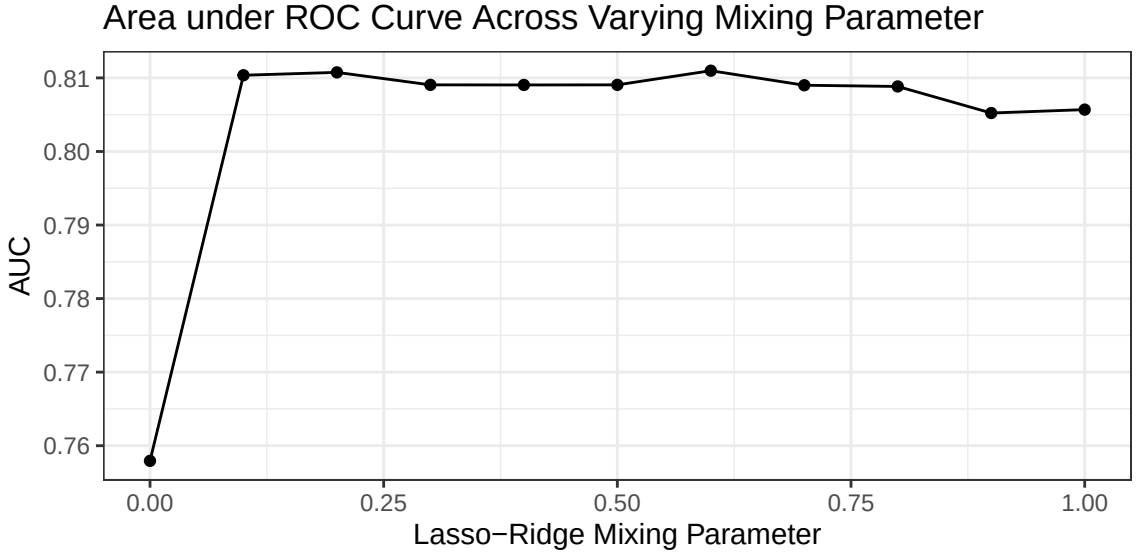
In the above plot, we find the difference in the outcome probability of logistic regression when the value of each covariate is changed from 0 to 1, while all other covariates are held at 0. We chose to hold all other covariates at 0 as the mode for each variable is 0. As you can see, the largest coefficients from logistic regression have the largest changes in probability when that particular covariate is present in the patient.

However, these particular changes in probability are dependent on all of the other covariates being held at 0, so a medical professional could not easily state that a patient being predisposed for a cervical spine injury would have a 10% increase in probability of having a cervical spine injury versus if they were not predisposed. However, these values do serve to give a sense on how certain variables affect the outcome more than others.

4 Stability

4.1 Model Perturbation

The effect of varying the elasticnet mixing parameter, or the mixture of the L1 lasso penalty and L2 ridge penalty, was analyzed. The top 15 variables selected by the specific mixing parameter were then used to perform logistic regression. To compare the models, the area under the ROC curves (AUC) were calculated.



The nonzero α parameters yield similar AUC values, but the pure ridge regression α yields significantly worse performance.

As for introducing a perturbation to logistic regression classification, one can adjust the classification threshold. The effect of this is captured in the ROC curve of the model presented in the previous section.

4.2 Data Perturbation

4.2.1 Change in Covariate Distribution During Testing

To examine the effect of changing the covariate distribution of the test set, the model was evaluated on the held-out site 7. As seen in Figure 1, site 7 has a higher proportion of of subjects of race group what was not disclosed (ND).

	Estimate	Lower CI Bound	Upper CI Bound
Orig. Training Data	0.944	0.915	0.972
Site 7 Data	0.857	0.728	0.987

As expected, worse performance is observed, as the records from site 7 were excluded from training.

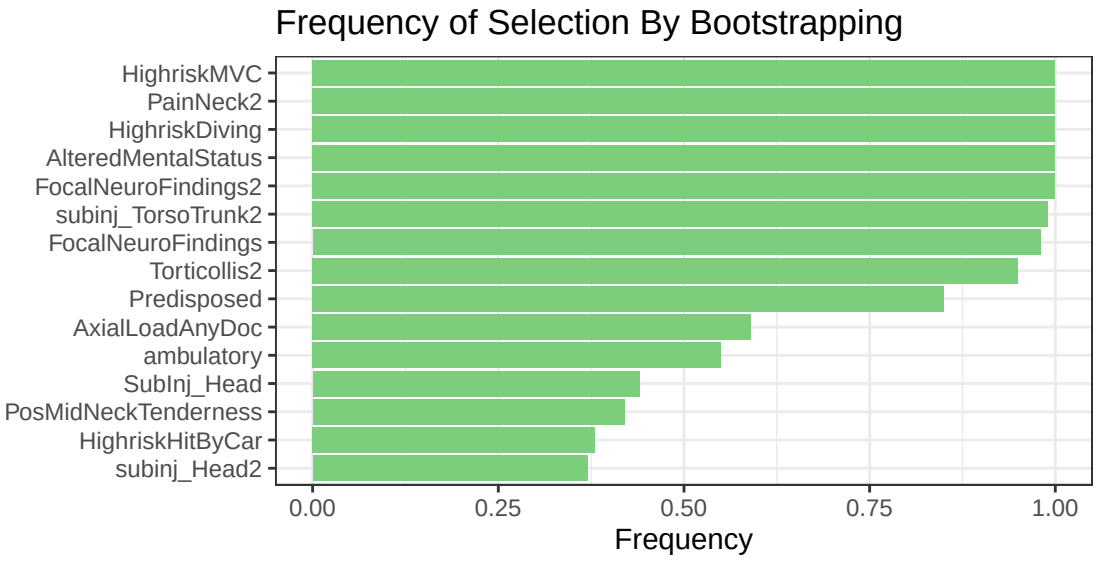
4.2.2 Stability under Subsampling

The model was evaluated on just the positive injury records and the EMS control group.

	Estimate	Lower CI Bound	Upper CI Bound
Orig. Training Data	0.944	0.915	0.972
Case and EMS Data	0.731	0.694	0.769

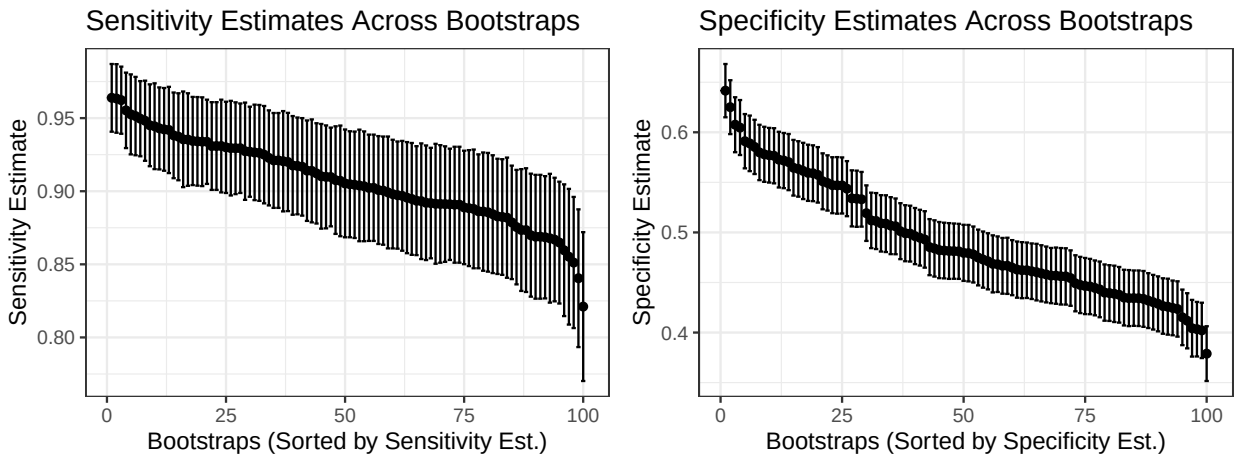
4.2.3 Stability under Bootstrapping

In addition, we observed the stability our model under bootstraps. For the lasso selection of variables, we analyze the frequency of the original 15 factors in our model in the top 15 factors selected by lasso generated by the bootstrapped data.

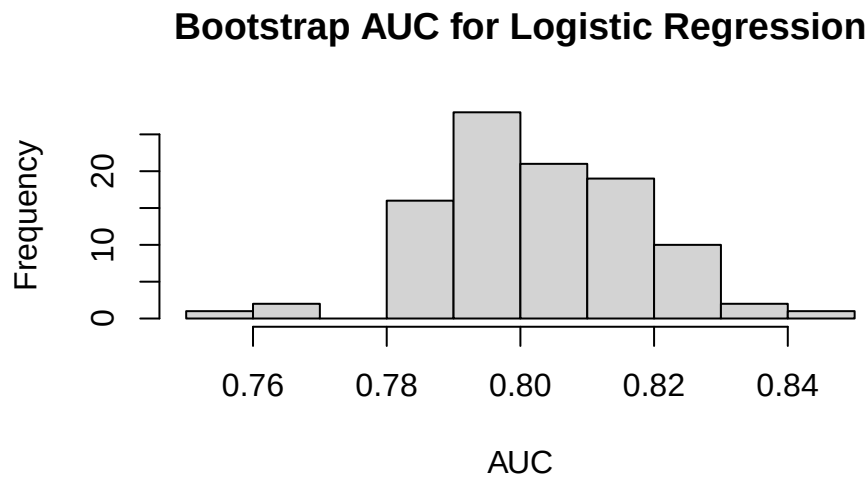


We can see that the top of the original factors appear in the bootstrapped data's top factors very frequently, whereas the lower of the original factors appear less frequently.

The stability of the logistic regression classifier under bootstrapping was also analyzed. The sensitivity and specificity confidence intervals of each bootstrap were recorded and are plotted below.



The sensitivity estimates appear to be more stable than the specificity estimates, as the sensitivity confidence intervals share more overlap.



To better visualize the stability across bootstraps, we plot the area under the ROC curve for each logistic regression classifier generated for each bootstrap. We see the AUC is roughly normally distributed and centered around an area of 0.8, which seems stable enough.

5 Evaluation

5.1 Three Test Sites

As stated in the subsection *Training and Evaluation Split*, three hospitals were randomly chosen as the sites for evaluation: 5, 16, and 17. These three sites were left out during the exploratory data analysis, model training, and model stability assessment.

5.2 Baseline Performance

We first applied the decision rule derived by Leonard et al. (2021) to the three test sites to evaluate the baseline performance, to which the relative performance of our models would later be evaluated. Using the 6-variable decision rule (altered mental status, focal neurologic deficit, complaint of neck pain, substantial injury to the torso, high-risk motor vehicle crash, and diving), the sensitivity and specificity of identifying cervical spine injury by the presence of at least one of these six factors were 85% (95% CI 78% to 92%) and 44% (95% CI 39% and 49%), respectively.

Table 9: Decision Rule from Leonard et al. (2011)

	Estimate	Lower CI Bound	Upper CI Bound
Sensitivity	0.853	0.781	0.924
Specificity	0.438	0.390	0.485

5.3 Forward Variable Selection Performance

We then applied the decision rule derived by using forward variable selection on the decision rule by Leonard et al. (2011) to the three test sites. This decision rule consisted of four additional variables: conditions predisposing to cervical spine injury, high-risk hit by car, axial load to top of the head, and injury by clothes-lining. (Note that at each forward step, the forward selection process adds the one variable that gives the single best improvement to the model. Hence, it is possible for the final model to contain variables that are significant when added at each respective step but are no longer as significant in the presence of subsequently added variables (Leonard et al. 2021).) The sensitivity and specificity of identifying cervical spine injury by the presence of at least one of these ten factors was 88% (95% CI 82% to 95%) and 30% (95% CI 26% and 35%), respectively. Compared to the baseline, there was a slight increase in the sensitivity (3% increase) and decrease in the specificity (14% decrease).

Table 10: Forward Variable Selection

	Estimate	Lower CI Bound	Upper CI Bound
Sensitivity	0.884	0.820	0.949
Specificity	0.303	0.259	0.347

5.4 Single Decision Tree Performance

We tested the single decision tree model (using the aforementioned ten variables) to the three test sites. The sensitivity and specificity was 91% (95% CI 85% to 96%) and 47% (95% CI 42% and 52%), respectively. Compared to the baseline, there were increases in both the sensitivity (6% increase) and in the specificity (3% increase).

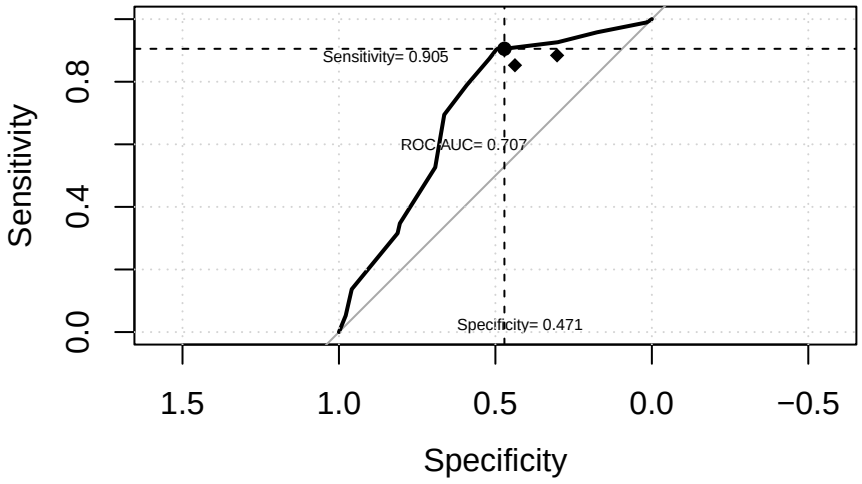


Table 11: Single Decision Tree with 10 Variables

	Estimate	Lower CI Bound	Upper CI Bound
Sensitivity	0.905	0.846	0.964
Specificity	0.471	0.423	0.519

5.5 Gradient-Boosted Tree Performance

We tested the gradient-boosted decision tree model to the three test sites. The sensitivity and specificity was 92% (95% CI 86% to 97%) and 41% (95% CI 36% and 46%), respectively. Compared to the single-tree model, there was a slight increase in the sensitivity (1% increase) and decrease in the specificity (6% decrease). Compared to the baseline, there was an increase in the sensitivity (7% increase) and a decrease in the specificity (3% decrease).

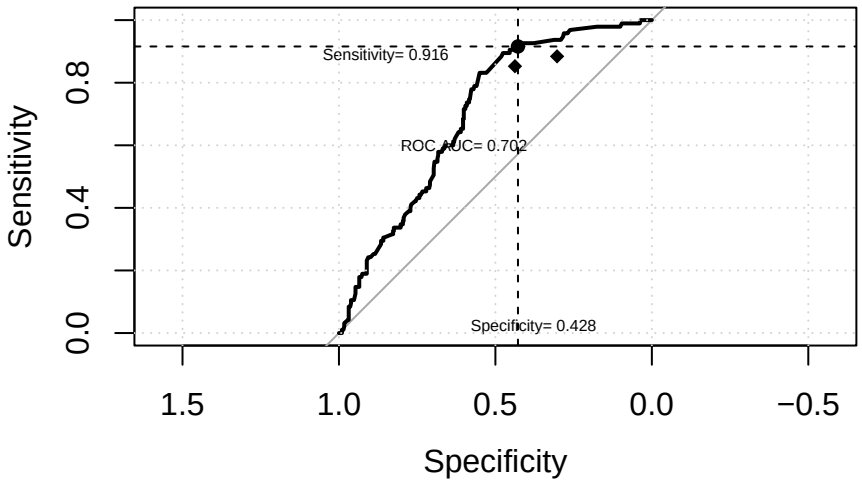


Table 12: Gradient-Boosted Tree

	Estimate	Lower CI Bound	Upper CI Bound
Sensitivity	0.916	0.86	0.972
Specificity	0.428	0.38	0.475

5.6 Logistic Regression

We tested the logistic regression model (based on all variables from feature selection) to the three test sites. The sensitivity and specificity was 88% (95% CI 82% to 95%) and 43% (95% CI 38% and 48%), respectively. Compared to the baseline, there were increases in both the sensitivity (3% increase) and in the specificity (1% increase).

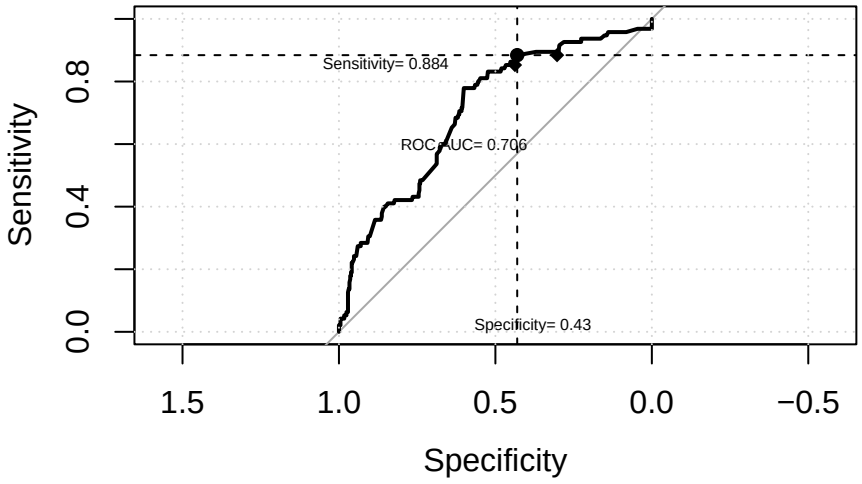
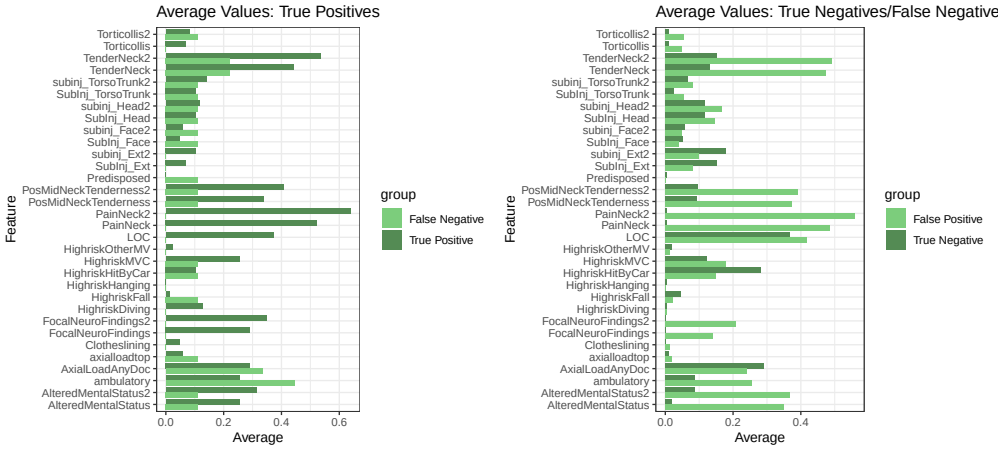


Table 13: Logistic Regression

	Estimate	Lower CI Bound	Upper CI Bound
Sensitivity	0.884	0.820	0.949
Specificity	0.430	0.383	0.478

To interpret potential patterns in the errors, we examine the means of each feature for the misclassified groups and correctly classified groups from the regression tree model, as this model resulted in the best performance compared to the baseline. In the bar graphs below, we analyze the average value of each feature in the set the model correclty labeled, versus the set on data points the model miss-classified for both true positives and true negatives. Although the decision tree split the data on 4 features, the average values for all of the features are presented below, to help identify any patterns that the model may have missed. For example, more cases where the "ambulatory" feature was present in the data were misclassified: both in the true positive and true negative case. However, the top node of the decision tree, FocalNeuroFindings2, is present in a high percentage of the correctly labelled positive cases, while it is present in a smaller percentage of the incorrectly labelled negative cases. Ultimately, these bar plots demonstrate the trade-off between sensitivity and specificity. If we were to decrease the incorrectly labelled false positives, we would increase the number of false negatives.



6 Summary

By investigating primary source data for cervical spine injury (CSI) in children, we derived and validated a clinical decision rule that aims to guide imaging decisions for children who experienced blunt trauma (e.g., altered mental status, focal neurologic deficits, complaint of neck pain, torticollis). Although we see on that the single decision tree performs the best on the test data, we chose to complete the bulk of our analysis on the logistic regression model since it performed the best on the training data. However, our stability analysis demonstrates that the model is generally stable: although we occasionally see some extreme values in sensitivity and specificity on the bootstrapped samples, the majority of the values fall within a tolerable range. These results emphasize the importance of the PCS framework in creating reliable modelling results. We chose to complete the bulk of our analysis without examining the model on the test data, so we made no decisions in the modelling process on the outcome of our test data. However, after examining the evaluation results, we would recommend further research on the single decision tree to examine if it would provide a stable, superior clinical decision rule.

7 Citations

[1] Leonard, J. C., Browne, L. R., Ahmad, F. A., Schwartz, H., Wallendorf, M., Leonard, J. R., Lerner, E. B., & Kuppermann, N. (2019). Cervical Spine Injury Risk Factors in Children With Blunt Trauma. Pediatrics, 144(1), e20183221. https://doi.org/10.1542/peds.2018-3221

[2] Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart