Learning and Evaluating Clinical Decision Rules for Cervical Spine Injury

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

December, 2022

1 Introduction

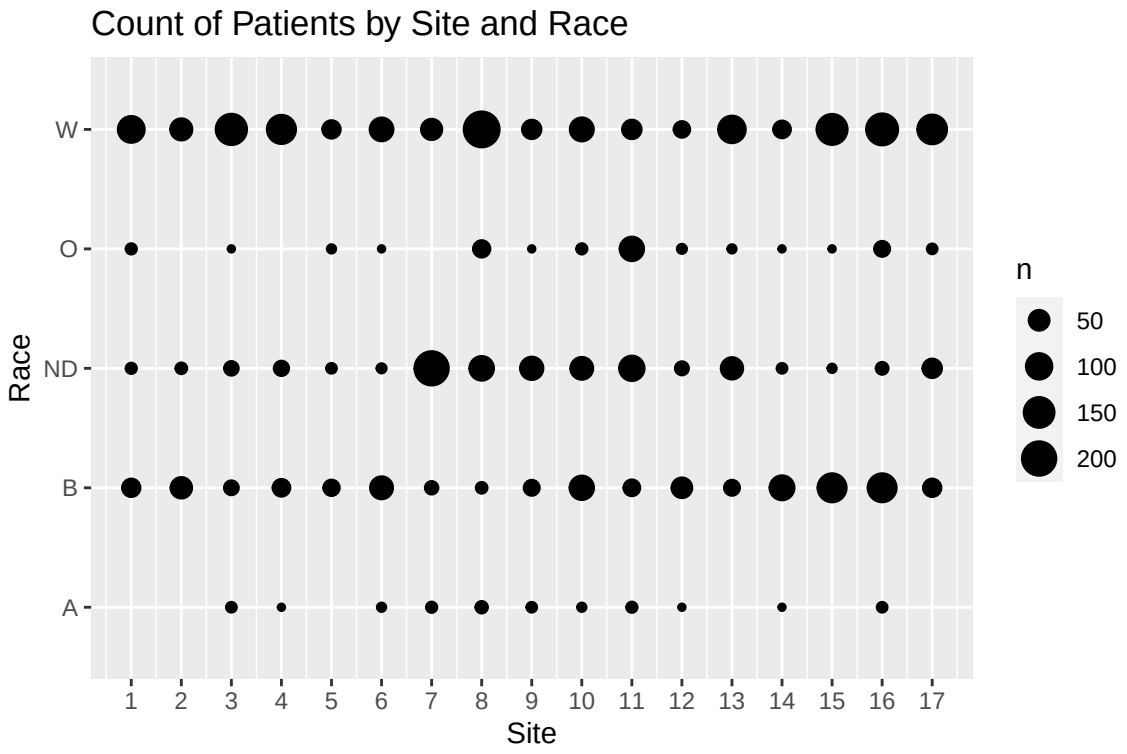
The goal of this project is to create a clinical decision rule to identify children who are most likely to have a cervical spine injury (CSI). The adverse effects of immobilizing children and subjecting children to ionizing radiation motivates such a rule, as there is a desire to minimize the number 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[CITATION]. 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.

Leonard et al. 8 major variables that are associated with CSI: altered mental status, focal neurological deficits, complaint of neck pain, torticollis, substantial injury to the torso, predisposing condition, 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. For the lasso selection of variables and logistic regression, records with missing values were removed from the analysis.

2.3 Training and Evaluation Split

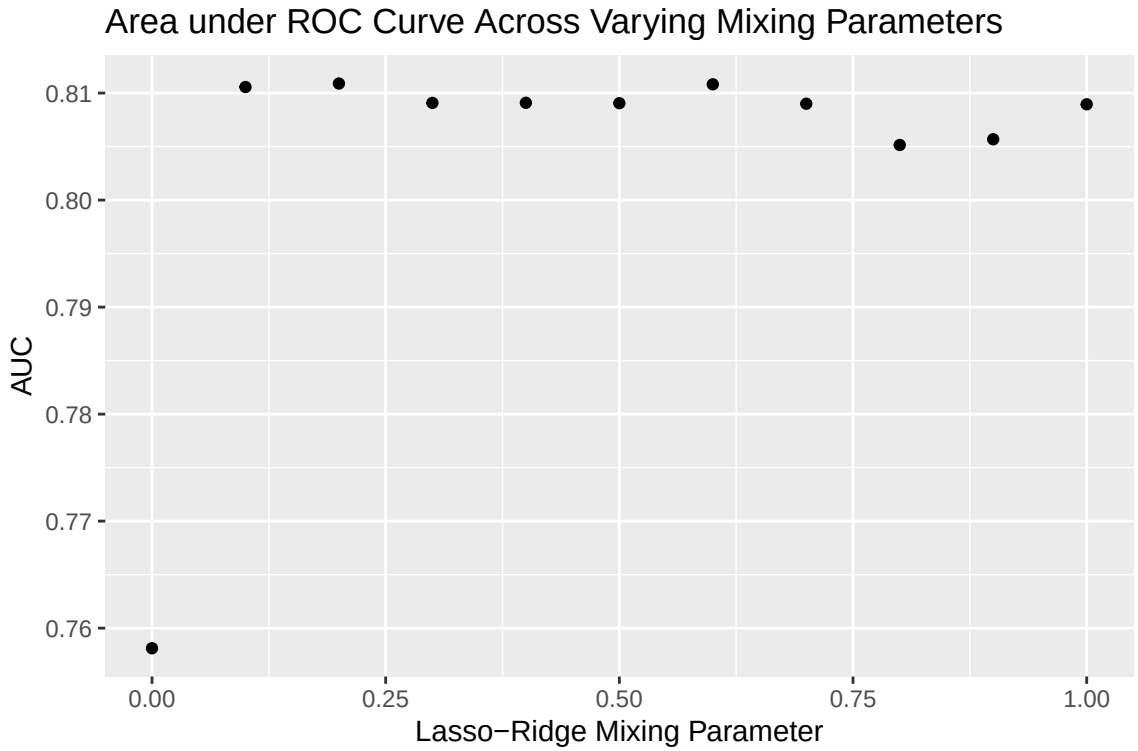
We assume that in practice, our decision rule will be deployed at hospitals not included in the dataset. Then, to simulate the performance of the models in practice, the data was split into test and training sets based on sites.

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 Stability

3.1 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.

3.2 Data Perturbation

3.2.1 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 ND.

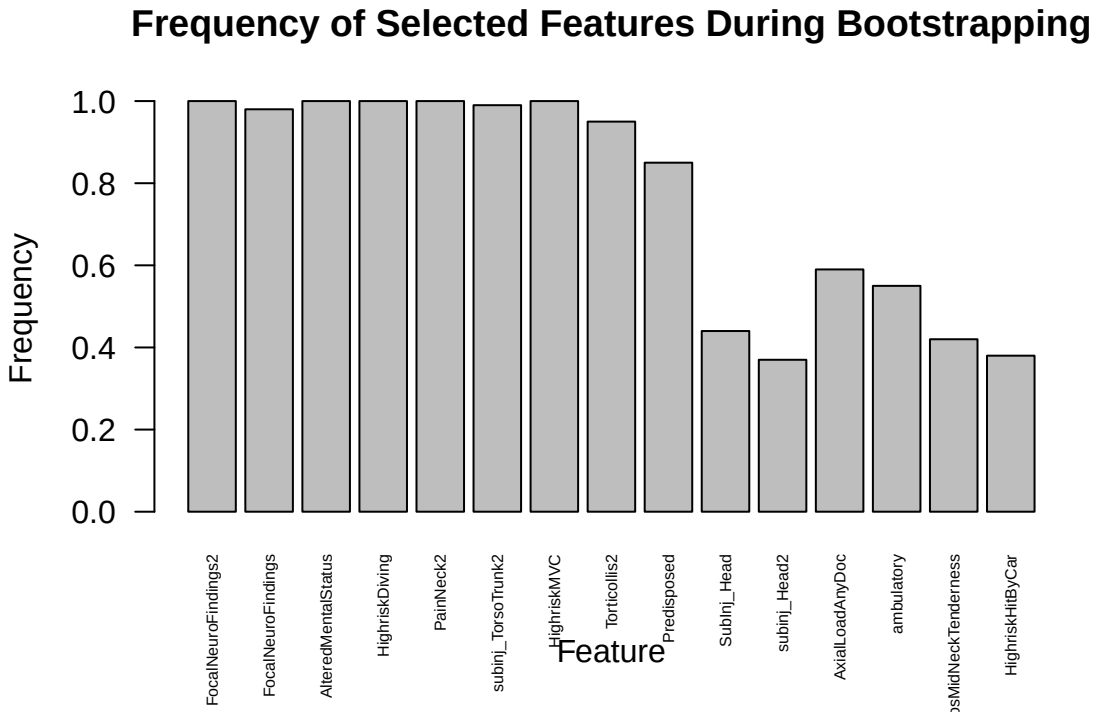
As expected, worse performance is observed, as the records from site 7 were excluded from training.

3.2.2 Stability under Subsampling

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

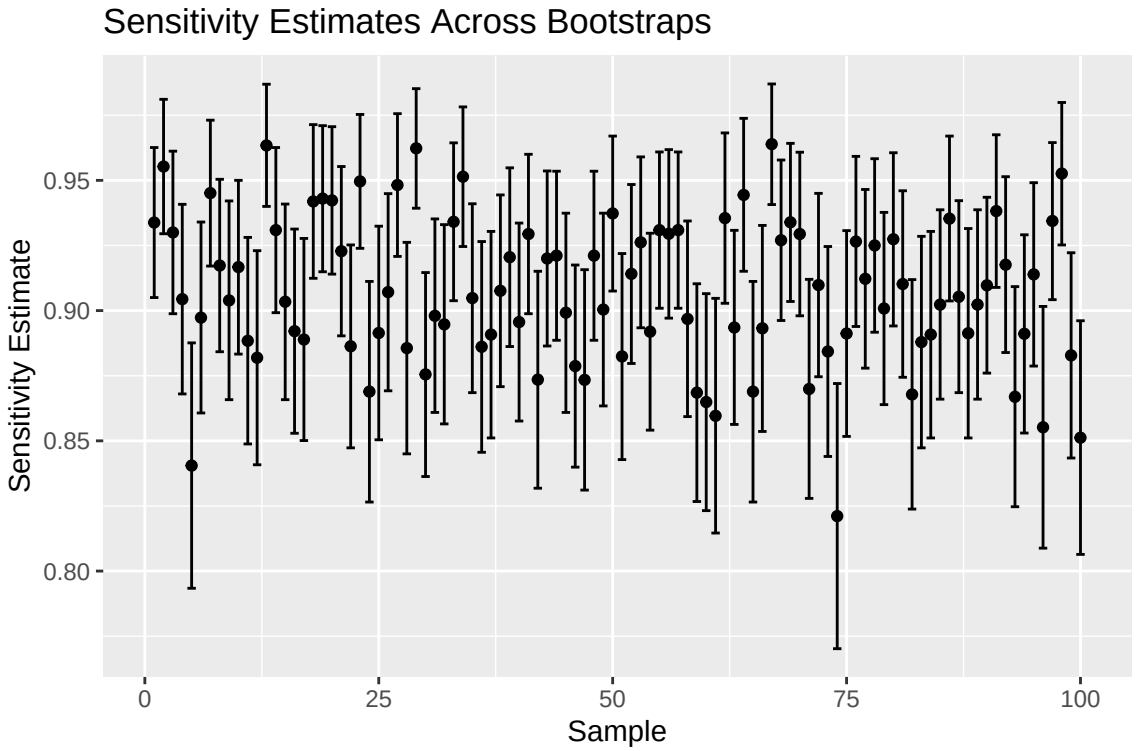
3.2.3 Stability under Bootstrapping

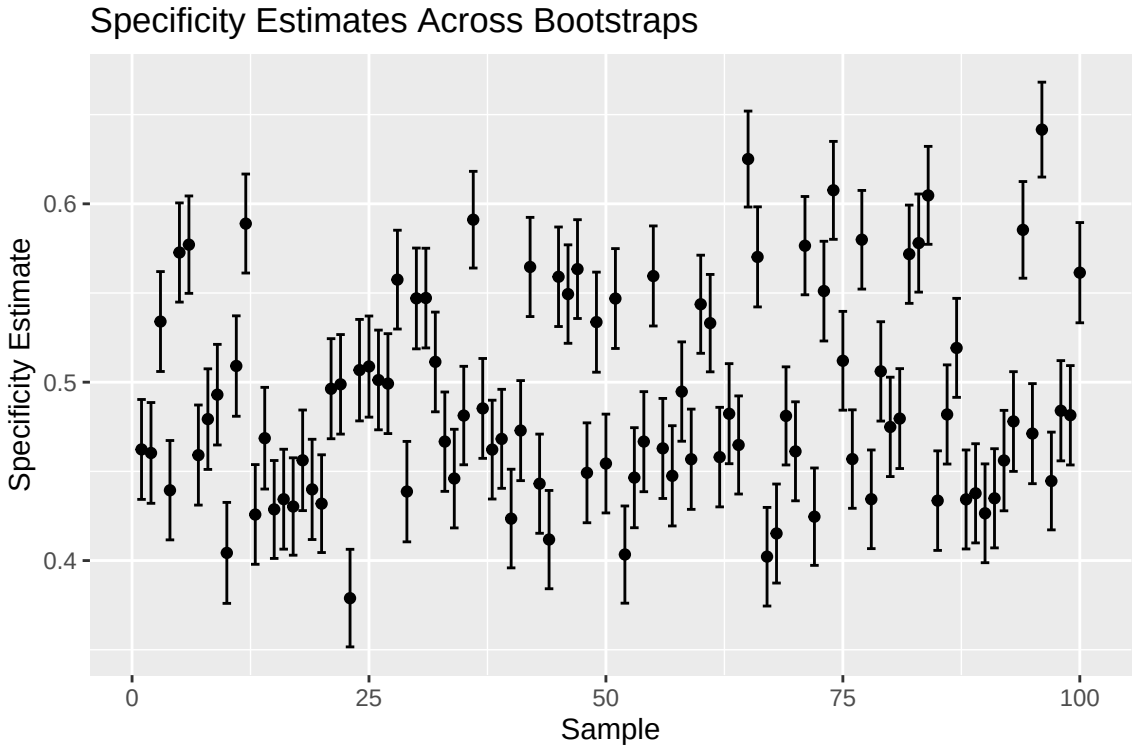
In addition, we observed the stability our model under bootstraps. For the lasso selection of variables, we analyze the frequency of the original 15 factors in our model in the top 15 factors selected by lasso generated by the bootstrapped data.



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.