Improving and vetting clinical decision rules in cervical spine injury in children

Group3

December 13, 2021

1 Introduction

Clinical decision rules (CDR) are evidence-based tools to assist practitioners to make decisions about clinical diagnosis and interventions, etc. They provide standardized principles for clinical practice to reduce risk by systematically following the direction of evidence-based medicine. CDRs can be understood as a predictive model where the input is several entries of the patient's clinical presentations, examination results, medical history, etc. and output is a binary indicator for making certain clinical decisions.

CDRs are particularly useful in clinical practice of the emergency department given the fact that clinicians are expected to make many important decisions quickly and accurately. Therefore, CDRs can greatly help increase efficiency in decision-making. In this report, we try to create and vet different CDR models for assessing the risk of cervical spine injury (CSI) in children. When a patient with suspected CSI comes to the emergency department, after initial evaluation based on history-taking and physically examinations, clinicians need to quickly decide whether to give further examinations or therapies such as radiographic imaging and cervical spine stabilization to this patient or not. Blindly giving further examination or treatment to all suspected patients will lead to adverse effects such as pain, pressure wounds and ionizing radiation as well as unnecessary economic and environmental burdens. Therefore a CDR based on results from the initial evaluation will be very useful in this clinical decision-making scenario. Previously a few CDRs for assessing CSI risks in adults were published and thoroughly validated in clinically practice, including the Canadian C-Spine rule [3] and National Emergency X-Radiography Utilization Study (NEXUS) criteria [1]. As part of the Pediatric Emergency Care Applied Research Network (PECARN) project, Leonard et al. [2] conducted a multi-center retrospective case-control study to derive CDRs for CSI risks in children below 16 years old. This resulted in an eight-variable model where a patient is regarded as having high CSI risk if any of the variables is positive. The variables include altered mental status, focal neurologic findings, neck pain, torticollis, substantial torso injury, conditions predisposing to cervical spine injury, diving, and high-risk motor vehicle crash. This CDR is claimed to have 98% (95% confidence interval 96% to 99%) sensitivity and 26% (95% confidence interval 23% to 29%) specificity in validation. A significant problem in the validation of this CDR is that an independent test dataset was not used to generate the above metrics. All data were used for deriving as well as validating the CDR, which raises questions around generalizability to patient populations outside this study.

In our study, we try to create new models for CDRs to assess CSI risk in children with the dataset from Leonard et al. and validate our rules using independent test data. Further, we perturb the data and human judgement calls to investigate model stability in scenarios closer to real-world applications. We translate this question into a binary classification problems where the negative class is considered as not having CSI and the positive class is patients with CSI. If a patient is predicted to be positive by the CDR model, further clinical examination and intervention such as cervical spine stabilization and radiographic imaging is recommended.

2 Data

2.1 Data collection

Leonard et al. collected data from 17 medical centers enrolled in PECARN program for children before 16 with cervical spine radiography after blunt trauma from 2000 to 2004. Cases (taken as positive samples) are identified by searching the hospital database for International Classification of Diseases, 9th Revision (ICD-9) codes for CSI, essentially those patients with final CSI diagnosis. All records are reviewed and verified by principal investigator and a pediatric neurosurgeon. The controls (taken as negative samples) are found in three different strategies. The first group is the random control group, where children with cervical spine radiography record but no CSI diagnosis are randomly selected. The second group is the mechanism of injury (MOI) control group, in which each cases are matched to one or a few control patients with the same age group and mechanism of injury. The third group is the emergency medical service (EMS) control group, where the control patients are selected to have the same age group as the cases and have received EMS out-of-hospital care. Essentially the first and second group differ in whether the samples are matched or not and the third group focuses on the scenario where the patient is sent by EMS. For each patient or sample, certain results from the medical records were extracted to form the features of sample. They include clinical presentations, mechanisms of injury, demographics, past medical history, final injury classifications etc. This features will be described in detail in the next section. Some samples are further excluded from the dataset for having insufficient medical records, unmatched visit times, being transferred away for final diagnosis, etc. This results in 540 cases, 1,060 random controls, 1,012 MOI controls and 702 EMS contorls.

2.2 Feature description

There are in total 609 features describing various aspects of each patient after removing some irrelevant data and features unavailable when CDR should be applied in the clinical settings. Here we introduce them in five categories according to which aspects the features are supposed to characterize. In the original study of Leonard et al., there is a set of expertly engineered features based on these raw features called `AnalysisVariables` in the published dataset, which were used in their analysis for deriving the original CDR in the paper. Hence we also summarized how each category is represented by the features in `AnalysisVariables`.

I. Clinical presentation

This group of features are different symptoms and physical signs that can be obtained by history-taking and physical examination by clinical practitioners. It has the largest size and is likely to be most influential to the outcome. There are totally three sources provided in this dataset: the record from the study hospital (on-site), the record from the transferring hospital ED (outside) and the EMS out-of-hospital run sheet (field). The on-site results are usually considered the most reliable because those are first-hand results for on-site clinicians and the examinations took place under the most stable condition. After excluding irrelevant features (e.g. date, time, transfer methods, etc) and post-hoc features (Neurological outcome, Long-term rehabilitation, etc), we can generally divide the features into 10 different groups as shown below. We also mark each group with a (on-site), b (outside) and c (field) to show which sources the features are available from. Note that this group of features are manually collected by clinicians. While being convenient and easily obtainable, human errors may be present in the process.

1. Conciousness (a, b and c):

- `HxLOC`: history of lost of consciousness
- `GCS and its sections`: Glasgow Coma Scale, a commonly used scoring system to assess level of consciousness based on several physical examinations
- `AVPU and its sections`: Alert, Verbal, Pain, Unresponsive scale, another scoring system to assess consciousness, simpler than GCS.

This group of features are the past record of lost of consciousness in medical history and two scoring systems based on on-site physical examinations. Lost of consciousness can indicate spinal cord injury

which is a high-risk situation that will need further examination and intervention. This group is summarize as `AlteredMentalStatus` and `LOC` in `AnalysisVariables`.

2. Complaints of pain (a, b and c):

- `PtCompPainNeck`: Complaint of pain in the neck
- `PtCompPainNeckMove`: Increase of pain when moving neck
- `PtCompPain*`: Complaint of pain in other regions (face, chest, back, flank, abdomen, pelvis, extremities)

Complaints of pain is one of the most common symptoms of traumatic injuries. This group includes complaint of pain in either neck and other regions of the body. Note that pain from other regions could also be informative for clinical decisions because they could be distracting and lead to underestimation of the severity of neck injuries. This group is summarized as `PainNeck` in `AnalysisVariables`. Complaint of pain in other regions are not included.

3. Tenderness (a, b and c):

- `PtTenderNeck`: tenderness in neck
- `PtTenderNeckLevel (C1-7)`: at which level of cervical spine is the tenderness is observed
- `PtTenderNeckLocation`: location of neck tenderness, either anterior, posterior, lateral or midline
- `PtTender*`: tenderness in other regions (face, chest, back, flank, abdomen, pelvis, extremities)

Tenderness is a common sign of traumatic injuries which means pain or discomfort upon touching. Tenderness can be characterized at specific locations, which is indicative of the injury. For example, tenderness in the posterior or midline region are considered to be more indicative of spinal fracture. This group is summarized as `TenderNeck` and `PosMidNeckTenderness` in `AnalysisVariables`. Tenderness in other regions are not included.

4. Torticollis (a, b and c):

- `LimitedRangeMotion`: torticollis or limited range of of motion of the neck

Torticollis means abnormal twisting of the neck causing head to be persistently tilted to one side and usually comes with limited range of head motion, which could be caused by CSI. It is summarized as `Torticollis` in `AnalysisVariables`.

5. Other injuries (a, b and c):

- `OtherInjuries*`: has observed substantial injuries in other regions (face, chest, back, flank, abdomen, pelvis, extremities)
- `MinorInjuries*`: has observed minor injuries in other regions (face, chest, back, flank, abdomen, pelvis, extremities)

Injuries in other regions of the body could also affect our evaluation of the risk of CSI. Similar to complaint of pain, injury in other region can be distracting and masking the symptom of CSI. Also, injuries in adjacent regions are sometimes associated with neck injuries because the trauma happens in similar areas. This group is summarized as `SubInj_*` in `AnalysisVariables`, where only substantial injuries are considered.

6. Focal neurological deficit (a, b and c):

- `PtParesthesias`: has parethesias
- `PtSensoryLoss`: has loss of sensation in any region of the body
- `PtExtremityWeakness`: has extremity weakness
- `OtherNeuroDeficitDesc`: has other focal neurologic signs

Focal neurological deficit is a group of signs that indicate impairment of specific part of the neurological system. Here they dataset includes parethesias (abnormal sensation), loss of sensation, limb weakness, etc. In the case of CSI, it could suggest that injury has affected the spinal cord, which is a condition cannot be ignored. In `AnalysisVariables`, these are summarized as `FocalNeuroFindings` to indicate if there is any positive findings.

7. Ambulatory (a, b and c):

- `PtAmbulatoryPriorEMSArrival`: was patient ambulatory before EMS arrival

Nonambulatory may be a result of loss of consciousness or neurological dysfunction, which are both indicators of severe CSI. This feature is included in `AnalysisVariables` as `Ambulatory`.

8. Position (c):

- `PatientsPosition`: the patient's position on EMS arrival (sitting, walking/standing, lying down, immobilized, pre-ambulatory)

This is a feature not included in `AnalysisVariables`. Position can reflect the patient's neurological function after the injury. In the Canadian C-spine Rule, whether a patient has sitting position is taken as one of the low-risk factors.

9. Intervention (a and b):

- `CervicalSpineImmobilization`: has C-spine immobolization before arrival
- `CervicalSpineInterv*`: specific types of immobilization methods
- `MedsRecdPriorArrival`: has received medication before arrival
- `MedsRecd*`: specifics types of medication
- `ArrPtIntub`: was intubated before arrival

This group of features is about the medical interventions the patient received before arriving at the study sites or outside hospitals, which is not included in `AnalysisVariables`. It indirectly reflects the type and severity of the patient's injury based on the judgement calls of previous clinicians. It depends heavily on the people who made the decisions about intervention thus could be unstable across different cases. So we take special care of this group and run perturbations that either including it or not.

10. Post-hoc features (a):

- `EDDisposition`: the patient's disposition (home, icu, general inpatient, $<$ 24h observation unit, transfer, death)
- `IntervForCervicalStab`: has received intervention at the study site
- `IntervForCervicalStab*`: types of intervention received (soft collar, rigid collar, brace, traction, surgical, other)
- `LongTermRehab`: has received long-term inpatient rehabilitation
- `OutcomeStudySite`: patient's neurological outcome at discharge
- `OutcomeStudySiteNeuro`: cognitive function at discharge
- `OutcomeStudySiteMobility`: mobility at discharge
- `OutcomeStudySiteBowel`: bowel function at discharge
- `OutcomeStudySiteUrine`: bladder function at discharge

This group of features are clinical treatments the patients eventually received at the study site and neurological outcomes of the CSIm which are essentially actual clinical decisions made by practitioners. They are generally unavailable when our CDR is supposed to be applied and therefore should not be included in our predictive model. Here we do post-hoc analysis of our classification results using this features to investigate the characteristics of the misclassified patients.

II. Injury mechanism

The risk of CSI from different injury mechanisms can potentially be different. Therefore it could be considered for building CDR. This group of features are generally more accurate because the process of how injury happened is not usually mistaken.

- `InjuryPrimaryMechanism`: how the injury happened. Categorized into 15 types of mechanism (e.g motor vehicle collision, pedestrian struck by moving vehicle, blunt injury to head/neck, falling, diving, hanging, etc.).
- `clotheslining`: was the injury a result of an object striking the neck
- `HeadFirst`: was the impact first sticking head
- `HeadFirstRegion`: region of head that was struck

III. Medical history

This group of features are collected from the patient's medical history. This group of feature is summarized as the `Predisposed` in `AnalysisVariables`. Note that it is common to miss certain aspects in medical history in clinical practice given the wide range of conditions each patient can have.

- `BodyAsAWhole*`: conditions known to predipose to CSI: Down Syndrome, Klippel-Feil Syndrome, Achondrodysplasia, Mucopolysaccaridosis, Ehlers-Danlos Syndrome, Marfan Syndrome, Osteogensis Imperfecta
- `Genitourinary1`: has renal osteodystrophy
- `Endocrinological1`: has Ricketts
- `HematologicLymphatic1`: has Juvenile Rheumatoid Arthritis
- `HematologicLymphatic2`: has Juvenile Ankylosing Spondylitis
- `Neurological`: has preexisting neurological abnormalities
- `Musculoskeletal`: has preexisting musculoskeletal abnormalities

IV. Demography

The distributions of many diseases and conditions vary significantly between genders and across different ages. Therefore gender and age can be important predictors in clinical decision rules. These features are not included in `AnalysisVariables`.

- `Gender`
- `Age`: we discretize the ages in to four groups: infant/toddler ($<$ 2 yrs), preschool (2-5 yrs), school-aged (6-13 yrs) and adolescent ($>$ 14 yrs).

V. Injury classification

This group of features is only available for cases. Detailed classification of all CSI by experts combining all information including radiograph. There are levels, locations and types of cervical spine fractures, ligamental injuries and signal changes of spinal cord on MRI. This is also a group of post-hoc features and we use them to examine the misclassified patients.

2.3 Feature derivation

As mentioned in the previous section, the original study has an expertly engineered set of features (i.e. `AnalaysisVariables`) that summarized features in most categories above. For example, the `AlteredMentalStatus` feature combines the information from both the GCS and AVPU scores from physical examininations. Either GCS is not getting full scores or AVPU is not getting A will result in a positive or 1 value in this feature. This totally results in 22 features.

We further augmented `AnalaysisVariables` with features from categories not covered. We included `Position` to characterize patient position after injury. We included `CervicalSpineImmobilization`, `MedsRecdPriorArrival`, `ArrPtIntub` to characterize the clinical intervention the patients received prior to arrival. We included `Position` to characterize the patient positions after injury. We also introduced `PtCompPainHead`, `PtCompPainFace`, `PtCompPainExt`, `PtCompPainTorsoTrunk`, `PtTenderHead`, `PtTenderFace`, `PtTenderExt2`, `PtTenderTorsoTrunk` to characterize the pain and tenderness detected in other region of the body. We also included age and gender. Lastly we created a feature `is\_ems` to represent whether the patient is sent by emergency medical service. This results in 45 features in total. We call this set of features `augmented AnalysisVariables`.

Note that for the clinical presentation features, some are available from either the study site, outside hospital or the field record. Data from the study site is considered to be the most reliable because the record is more comprehensive in large medical centers and the examinations were performed under the most favorable conditions. So the authors created two versions for this set of features. The $[feature\_name]$ uses only the on-site data. And $[feature\_name2]$ is build upon $[feature\_name]$ where negative and missing values are altered to positive if it is shown as positive in either the outside data or field data. We follow the same procedures for our augmented features. Then there are in total 21 features processed in this manner, including 11 of our augmented features.

We also derived a de novo set feature basically just binarizing all the aforementioned features in different categories without any further engineering. This strategy potentially provides us more information about the patients but could also harm generalizability and interpretability. Because the engineering process in which several basic features are summarized in to one composite feature can potentially reduce variation and increase robustness. This set of features turned out not to be giving better performance to the predictive model and the results are not shown for brevity.

2.4 Feature visualization

Next we visualize several groups of features to investigate correlations between features in each group and their correlation with the outcome (defined as the sample is case or control). Here we removed all samples with missing values and Pearson correlation coefficient is calculated. (Figure 1)

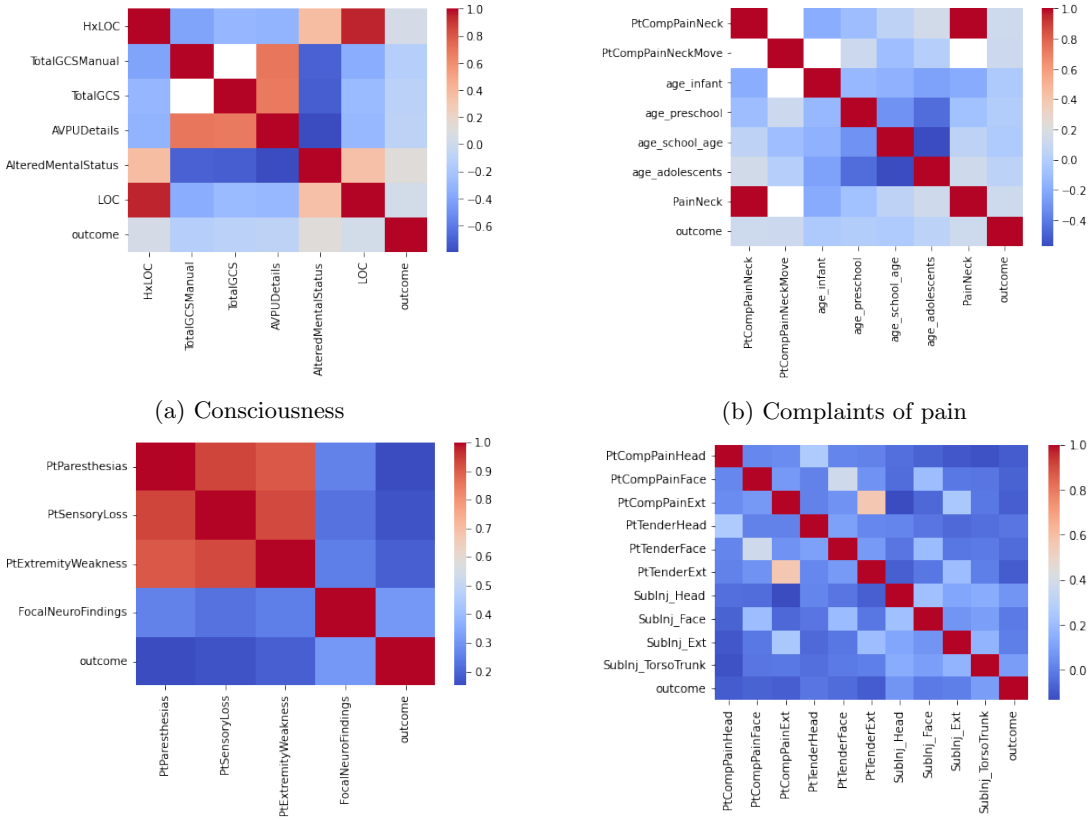
The first group are features about the level of consciousness. Here We can see the history of lost of consciousness (`HxLOC`) is almost the same as the `LOC` feature in `AnalysisVariables`. The GCS and AVPU scores showed high correlation, and both showed strong negative correlation with the `AlteredMentalStatus` feature in `AnalysisVariables`. This show that the two features in `AnalysisVariables` summarized the level of consciousness in patients quite well.

The second group is about complaint of neck pain and association with age. We can see that younger age groups are negatively correlated with complaint of neck pain, since it is more difficult to obtain complaints from younger kids. And no children in the <2 years group has complaint of increased pain when moving neck, which is expected because they are not capable of this level of self-expression.

The third group is about neck tenderness. Here we only investigate their correlation with outcome. We can see that tenderness in neck generally has positive correlation with the outcome. Tenderness in C6, C7 levels and posterior, midline regions tend to have the highest correlation with the outcome. Therefore the features `PosMidNeckTenderness` and `TenderNeck` in `AnalysisVariables` is an appropriate summary of this group.

The fourth group is about focal neurological findings. We can see that the first three features have strong correlation with each other. Therefore summarizing the features as one `FocalNeuroFindings` features is reasonable.

In the fifth group we aim to investigate the correlation between complaint of pain, tenderness and observable injury in other region of the body. In `AnalysisVariables`, only observable injuries are included (`SubInj_*` features). But here we see that complaint of pain and tenderness in certain



(a) Consciousness

(b) Complaints of pain

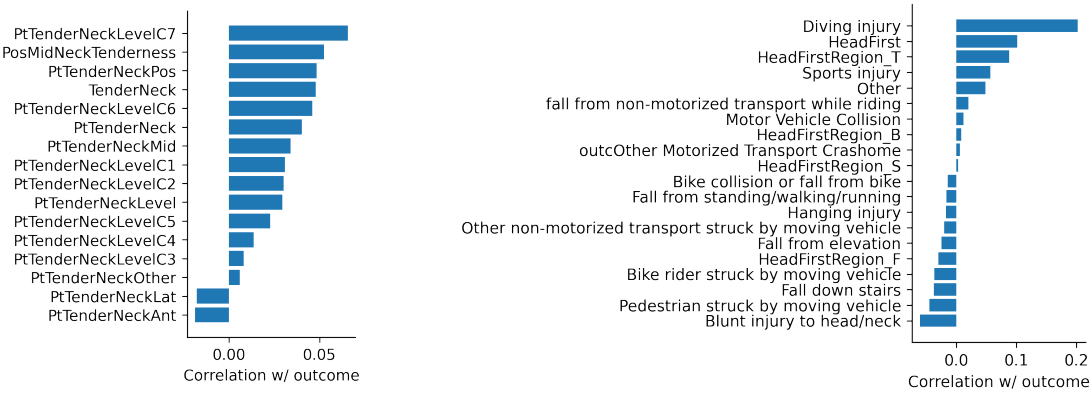
(c) Torticollis

(d) Other injuries

Figure 1: Correlation among various feature groups

region are not strongly correlated with the observable injury features. Therefore it is reasonable to also include complaint of pain and tenderness in our analysis.

In the last group, we want to investigate which mechanisms of injuries are most strongly correlated with the outcome. It turns out that diving injuries and injuries from strikes first hitting the head and particularly in the top region are most risky mechamisms, which are included in `AnalysisVariables` as `HighriskDiving`, `AxialLoadAnyDoc` and `AxialLoadTop`.



(a) Tenderness

(b) Injury mechanism

Figure 2: Correlations of various tenderness conditions and injury mechanisms with CSI outcome

2.5 Imputation

There is a large fraction of missing values in this dataset (Fig. 3). Excluding all patients with missing values will result in a overly reduced sample size. Therefore a plausible imputation strategy in necessary here. We therefore consulted clinical experts to make human judgement calls on imputation. Basically we choose to impute each feature either liberally (taking missing values as 0 or negative) or conservatively (taking missing values as 1 or positive) based on domain knowledge, which is based on the most likely situation in actual clinical settings. For example, focal neurological deficits are critical indicators of injuries of the brain or the nervous system. So every clinician will pay utmost attention to it. If any signs are detected during examination, it is very unlikely that the record will be missing. So we consider it to be a liberal feature and missing values are imputed to be negative. We ended up with most feature considered to be imputed liberally and only `LOC` and `clotheslining` to be imputed conservatively. Still, there are a group of features whose imputation strategy is unclear based solely on this domain knowledge. These features include `AlteredMentalStatus`, `ambulatory`, `PainNeck`, `PosMidNeckTenderness`, and `TenderNeck`. As a solution we choose to perturb the direction of imputation for them in later analysis.

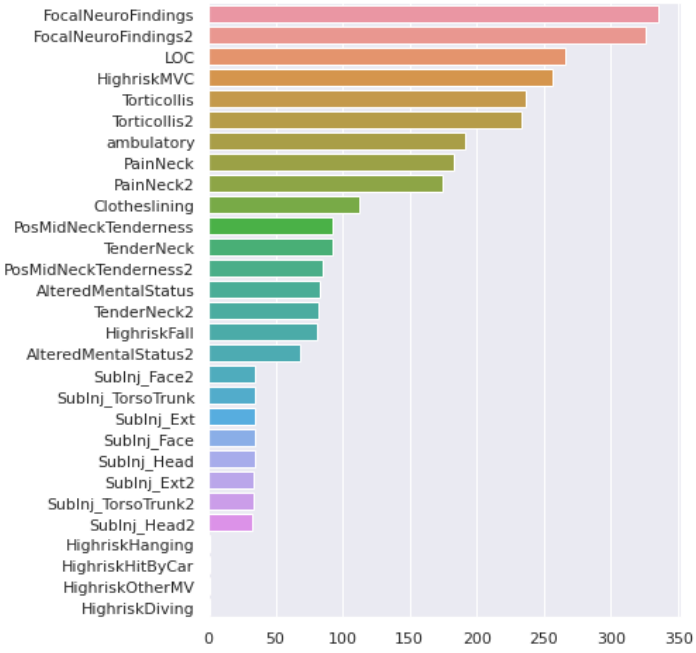


Figure 3: Visualization of amounts of missing values in *Analysis Variables*.

3 Modeling

We compared the performance of the PECARN clinical decision rule to 8 different off-the-shelf ML models. Our model selection and evaluation process was motivated by two high-level guidelines stemming from the fact that we are working within the medical domain. First, our analysis was limited to interpretable machine learning methods, since it is crucial for a medical practitioner to understand how predictions are make when using a clinical decision tool. Second, we explicitly considered each model's complexity, limiting our analysis to models with low enough complexity that they could reasonably be applied by a practitioner within the short time frame allotted to decide whether a CT scan is necessary.

To construct a test set that would capture the true generalizability of the models, we held out the data for 4 randomly selected sites that together constituted about 20% of the data. As opposed to a random split, a site split creates a higher bar for models, and additionally replicates the conditions of an external validation study more closely.

For each model, we tried a wide variety of hyperparameters and evaluated the effectiveness of each hyperparameter setting using 10-fold cross validation on the training set. We then selected the best

hyperparameter setting among those which produced a model with low complexity, before evaluating on the test set.

The primary metric used to select the best models in validation is the best specificity achieved at a given level of sensitivity. Since each model (with the exception of the PECARN rule) provides some kind of class probability or confidence score for each predicted sample, one can vary the threshold at which a sample will be classified as positive until a given sensitivity is reached. This is motivated by the fact that near-perfect sensitivity is required for a model which predicts whether a CT scan is necessary in the case of cervical spinal injury, since a false negative can mean a life-threatening injury goes untreated. Essentially, the task for each of our candidate models is to be as selective as possible while having sensitivity close to 1.0.

3.1 Model descriptions

- **CART**: Simple decision trees. For CART, complexity is defined as the total number of splits in the tree. Varied hyperparameters included the maximum number of leaf nodes, the relative weight given to each class, and the criterion for split selection (gini impurity or entropy).
- **Random Forest**: Ensemble of decision trees with subsampling. Complexity is defined as the total number of splits across all trees. Varied hyperparameters included the number of trees in the forest, the amount of subsampling, and the maximum depth of component trees.
- **Gradient Boosted Trees**: Ensemble of decision trees with boosting. Complexity is defined as the total number of splits across all trees. Varied hyperparameters included the number of trees in the ensemble, the type of loss used (deviance or exponential, the latter recovering AdaBoost), and the maximum depth of component trees.
- **RuleFit**: Friedman and Popescu's algorithm from 2008 which first fits an ensemble of trees, extracts the resulting rules, then fits a sparse (L1 regularized) linear model on rules. Complexity is defined as the total number of rule terms, such that a rule with two conditions (i.e. "feature1 and not feature2") has twice the complexity of a rule with one condition. Varied hyperparameters included the number of trees in the initial ensemble and the regularization parameter.
- **Greedy Rule List**: Like a decision tree, a greedy rule list greedily selects the rule that will decrease entropy or gini impurity the most. However, instead of generating a tree by creating further splits on both outcomes of existing splits, it picks one outcome to further split on. Complexity is defined as the number of rules, since rules can only have one term. Varied hyperparameters include the maximum number of rules, relative class weight, and the criterion for split selection (gini impurity or entropy).
- **Boosted Rule Set**: Boosted rule sets generate rules by first fitting boosted trees, then extracting the resulting rules and scoring the rules based on their corresponding confidence scores in the trees. Complexity is defined as the total number of rule terms. Varied hyperparameters include the number of trees and the relative class weight.
- **Skope Rules**: Skope Rules fits an ensemble of trees, extracts the resulting rules, then filters out rules which do not meet precision and recall minimums. Complexity is defined as the total number of rule terms. Varied hyperparameters include the number of trees in the ensemble, minimum rule precision, minimum rule recall, and maximum tree depth.
- **Repeated Rules**: This model is similar to RuleFit, except the input rules for the linear model do not come from a tree ensemble but are generated by fitting an ensemble of rule-based models and extracting the repeated rules. Complexity is defined as the total number of rule terms. Varied hyperparameters include all those of RuleFit, boosted rule sets, and skope rules, which were used as the weak learners for our experiments.
- **PECARN Rule**: The PECARN rule is a simple logical-or of 8 features in the dataset, created by the original authors of the CSI study.

Model	0.9 sensitivity	0.95 sensitivity	0.96 sensitivity	0.98 sensitivity
CART	0.4712	0.0623	0.0495	0.0175
Random Forest	0.4904	0.3035	0.2827	0.1070
Gradient Boosting	0.0000	0.0000	0.0000	0.0000
RuleFit	**0.4920**	**0.3993**	0.3674	**0.2204**
Greedy Rule List	0.4536	0.3434	0.316294	0.0000
Boosted Rule Set	0.453674	0.0000	0.000000	0.0000
Skope Rules	0.4648	0.0335	0.0335	0.0335
Repeated Rules L1	0.4888	0.1070	0.1070	0.1070
PECARN Rule	0.3658	0.3658	**0.3658**	0.0000

Table 1: Specificity achieved at given sensitivity level, complexity ≤ 25.

3.2 Results

After the best models with low complexity were selected, each was fitted to the entire training set and evaluated on the test set. In particular, we limited complexity to 25, which we determined to be a generous upper bound on the number of rule terms or decision tree splits that would be practical for a clinical decision rule of this kind. Specificity-sensitivity metrics on the test set for each model are provided in the table below.

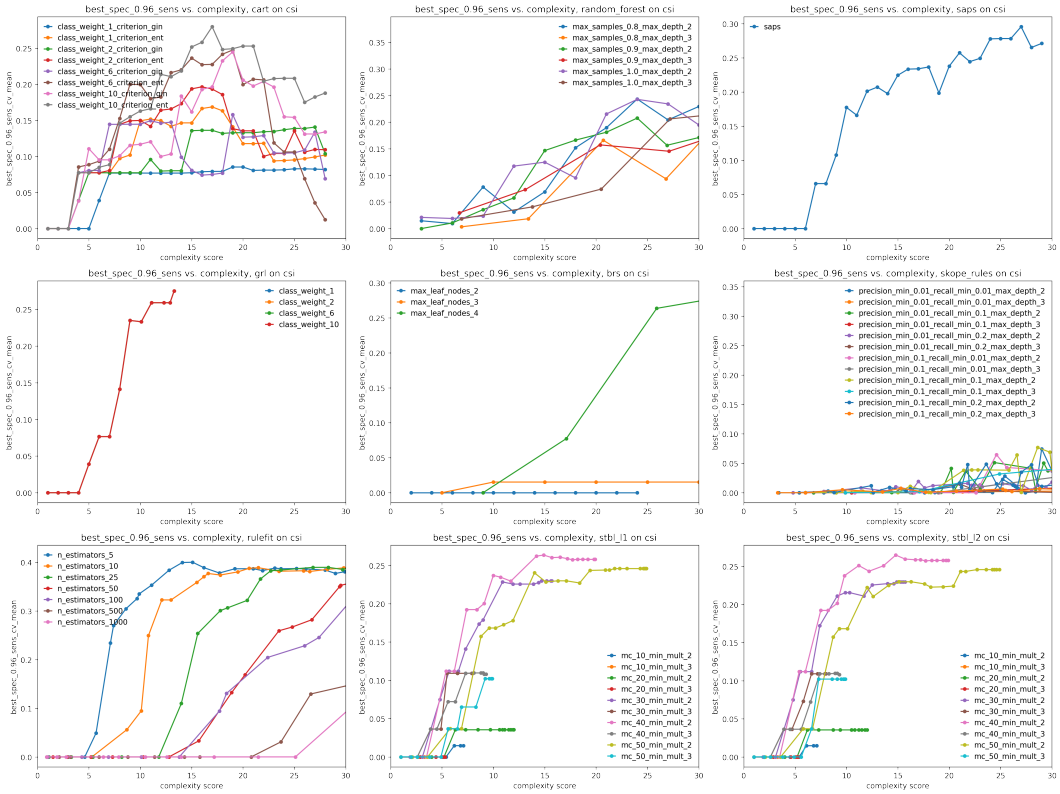


Figure 4: We assessed effects of different hyperparameters to the performance of a list of candidate models. The black dot denotes performance of the baseline CDR model described in previous study.

We find that the RuleFit model is the best performing, outperforming all other models and the PECARN rule at the 0.9, 0.95, and 0.98 levels of sensitivity. The PECARN rule, which achieves 0.966 sensitivity and 0.365 specificity on the test set, outperforms RuleFit at the 0.96 minimum sensitivity.

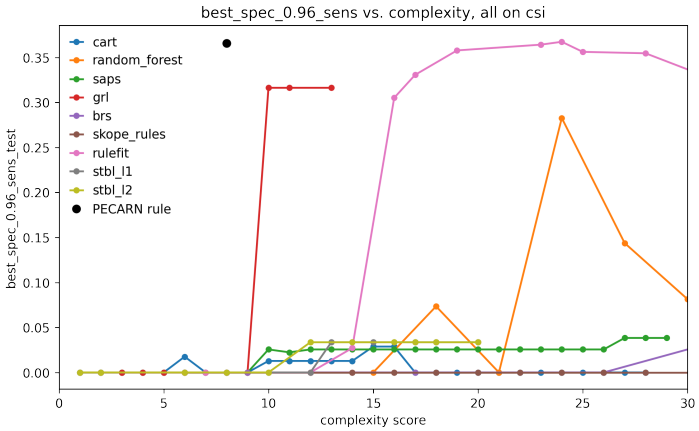


Figure 5: Performances of various models after hyperparameter tuning as a function of model complexity defined in Section 3

4 Stability analysis

From raw data to final data models, large amounts of decisions are often made; a quantitative evaluations of the effects of these decisions, or **judgement calls**, are critical for providing practitioners with optimality guarantees. After the major decision steps are finalized, to be qualified as good statistical applications that can refine scientific hypothesis and generate reliable predictions about future data, the relevant data models need to be robust against data perturbations, and generate stable predictions despite changes in input data.

Specifically, throughout the process of designing our clinical decision rule model for the CSI Pecarn dataset, we made the following decisions:

- `include_intervention` in `clean_data`: Whether to include the presence of medical intervention like immobilization of patients or air pathway intubation in the dataset, default argument for the best model described here is `True`. We chose this as default along with `True` for `augmented_features` described below with the assumptions that the model would likely benefit from more features.

- `unclear_feat_default` in `preprocess_data`: For imputation of missing data values, there are a large amount of variables for which from clinicians' domain knowledge it is not immediately clear whether we should do a conservative (1) or liberal (0) imputation, default argument for the best model described here is 0. We chose this as our initial default because it is equivalent to median imputation, a common strategy with data analysis.

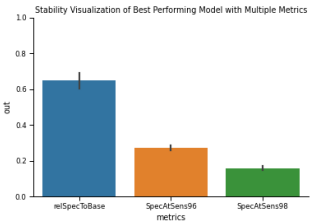
- `only_site_data` in `preprocess_data`: Whether to use only data from the study site (2) or also include field and outside hospital data (1), or use both of them combined: default argument for the best model described here is 2. We picked this as a result of preliminary exploratory analysis.

- `augmented_features` in `preprocess_data`: Whether to use augmented features from other files or just original `AnalysisVariables`, default argument for the best model described here is `True`

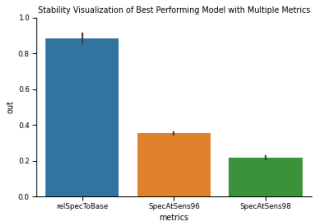
4.1 Stability evaluation of performance of selected best model

With the default judgement calls determined with exploratory analysis and domain knowledge, we selected the best model according to Section 3.2. To evaluate the stability of the modeling performance, we performed bootstrap sampling on training set and tuning set combined as an input to fit data models. Across different random instances, we obtained the bootstrapped distribution of specificities evaluated with sensitivity held at 96% and 98% respectively. In addition, we evaluated the baseline CDR obtained in previous studies for each random bootstrapped dataset, and calculated a relative specificity of our model when it is held at the same specificity as the baseline CDR. Since CDR was trained on the whole dataset and with boostrapping we are feeding our data model with repetitive

instances, the performances naturally deflated compared to cross validation results reported above. Through this procedures, we were able to obtain confidence intervals of the model performances, shown in Figure 6a.



(a) model stability result of the best performing model in Section 3.2 with default judgement call



(b) model stability result with best judgement call decisions following analysis 4.2

Figure 6: We tested the stability of the model performances through bootstrapping, evaluated via three different metrics. With the curated judgement call selections determined with analysis in Section 4.2, higher performance is achieved compared to the default judgement call selections. In general across the two selections, the model performance is quite stable.

4.2 Effects of judgement calls to data model

To evaluate the effects of our selections in judgement calls, we perturbed the dataset by considering other combinations of judgement call options. Then we calculated the specificity for the best performing model obtained from Section 3.2 with its sensitivity held at 96% and 98% level respectively. To obtain confidence intervals of the metrics, we performed bootstrap sampling on training set and tuning set combined as an input to fit data models. Similarly to the stability analysis above, we obtained the bootstrapped distribution of specificities at the two levels through evaluating each random model's performance on the held out test dataset. Intuitively, we can think of the held-out test set as new patients that the clinical decision rules will be used on; and the bootstrapping process can be thought of alternative data collection process, where more or less patients with the exact same conditions as the original data would be incorporated in the dataset.

Through the perturbation experiment, we generated performance distributions of the best selected model with different judgement call choices (Figure 7). Much to our surprise, the augmented features and intervention variables are in fact not very informative compared to the original `AnalysisVariables`, and actually negatively affected performance. With this novel information obtained through this perturbation analysis, we re-evaluated the stability of the model with the new set of judgement call variables. As expected, performances across the three metrics increased as shown in Figure 6b.

4.3 Methodology

To perform the stability and perturbation analysis, we adopted the `veridical-flow` framework developed in Yu Group in UC Berkeley. We modified the original `Dataset` template to utilize the most updated developer version of `vflow`. For the stability analysis of the selected best model described in Section 4.1, the vflow pipeline visualization is shown in Figure 8a. For the judgement call perturbation analysis, the vflow pipeline is visualized in figure 8b.

5 Post-hoc analysis of results

Finally, we examined the false-negative samples from our final model to verify if the actual CSI patients we missed are under severe conditions. When applying our model in actual clinical practice, we would likely set a strict threshold to control the sensitivity to be above 96% or 98%. But here we set a relaxed threshold to achieve 90% sensitivity in order to examine a larger group of false-negatives. As shown in Table 2, we can first observe that none of the patients received long-term rehabilitation and one out of ten patients has abnormal neurological outcome, which will also be correctly predicted if a more

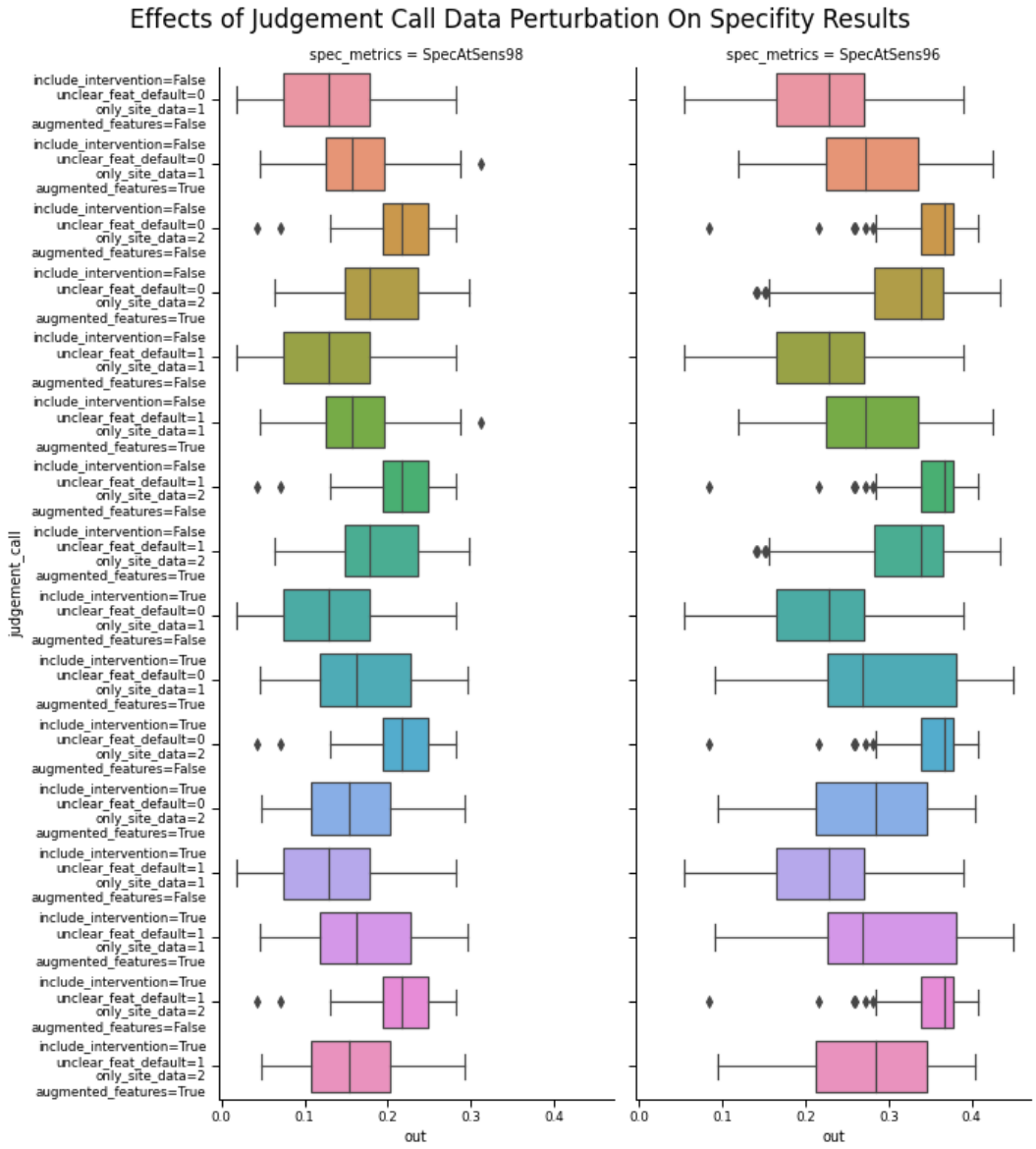
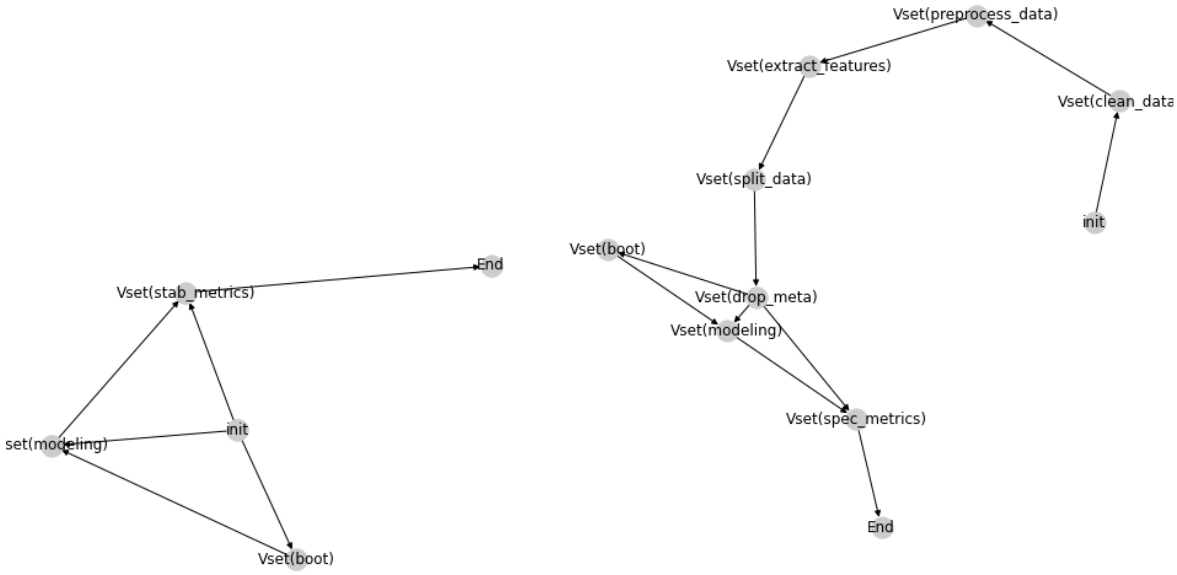


Figure 7: Effects of using different judgement call combinations to data model performance. Augmented features turned out to be mostly extraneous and not effective at improving model performance.

strict threshold is applied. Again only one patient has shown change of MRI signal of spinal cord and received surgical stabilization. Overall, these patients have acceptable clinical outcomes at discharge.

6 Conclusion and discussion

In this study, we carefully investigated a dataset from a multi-center retrospective case-control study for deriving CDRs for assessing CSI risks in children. With the original expertly engineered features from Leonard et al. plus the group of augmented feature we derived, we evaluated eight different types of prediction models against the PECARN CDR from the paper. We found that only the optimized Rulefit model has achieved comparable performance to the original CDR. We further perturbed the input data by altering human judgement calls and bootstrapping to investigate the stability of the model. We surprisingly found that using original smaller set of features actually produced better performance in the stability analysis. Post-hoc analysis of the undetected CSI patients has also shown the patients had mostly acceptable clinical outcomes.



(a) pipeline for stability analysis in Section 4.1 (b) pipeline for judgement call perturbation in Section 4.2

Figure 8: Stability analysis pipelines with veridical flow

Study subject ID	Injury	Disposition	Treatment	Neural outcome	LongTerm Rehab
110016	C7 teardrop fracture,	ICU	R-Collar	No	No
110020	C1 lateral mass fracture	GEN		No	No
210153	C7 fracture spinous Process	ICU	R-Collar	No	No
210165	Ligamentous injuryC4-5,5-6	ICU	R-Collar	WD	No
610710	C7 fracture transverse process	GEN	Brace,	No	No
1612771	LigamentousInjuryC1-2,5-6,6-7,C7-T1, C2 spinal cord MRI signal change	ICU	Halo, Surgery	No	No
1612780	C2 odontoid fracture	ICU	Halo	No	No
1612783	C7 wedge/compression frature	ICU	R-Collar	No	No
1612810	C7 spinous process fracture, Ligamentous injury C7-T1	ICU	S-Collar	No	No
1612811	C6 wedge/compression fractureBY	ICU	S-Collar	No	No

Table 2: Characteristics of children with CSI in the test set who was not predicted by our model (the sensitivity was chosen as 91.5%).

Overall, we did not find a new CDR model with obvious superior performance compared to the PECARN CDR. However, as mentioned in Introduction, PECARN CDR were derived using all the available data from while we only used a fraction of the data to train our models, leaving the rest as independent test set. We argue that the test performance of our models are more faithful to real-world situation when applied to future data.

The CSI PECARN data set is sampled from 17 different hospital sites. From exploratory analysis and modeling, for a same set of decision rules, we observed a great variations in performance across site locations. To yield a conservative performance assessment of our model, we performed site split for held-out test set. Accordingly, the performance measure we provided above is less sensitive to site differences. More rigorous analysis can be done under the veridical flow framework to assess the influence of cross-site differences to the model accuracy. Such analysis is crucial for future work, as it will provide a detailed generalizability assessment of the clinical decision rules derived in the curated

data model for clinical applications in arbitrary hospital sites.

Author contributions

All members conceived the study and designed the analysis workflow. C.Y and A.Q. studied the dataset and documentation, conducted feature derivation and implemented imputation strategies. Y.D. performed feature visualization and tested preprocessing scripts. K.N. implemented the modeling pipeline and performed model selection and validation. A.Q designed and implemented the stability analysis. C.Y. and Y.D. performed post-hoc analysis. All members wrote and reviewed the report.

References

[1] Jerome R Hoffman, Allan B Wolfson, Knox Todd, William R Mower, NEXUS Group, et al. Selective cervical spine radiography in blunt trauma: methodology of the national emergency x-radiography utilization study (nexus). *Annals of emergency medicine*, 32(4):461–469, 1998.

[2] 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.

[3] Ian G Stiell, George A Wells, Katherine L Vandemheen, Catherine M Clement, Howard Lesiuk, Valerie J De Maio, Andreas Laupacis, Michael Schull, R Douglas McKnight, Richard Verbeek, et al. The canadian c-spine rule for radiography in alert and stable trauma patients. *Jama*, 286(15):1841–1848, 2001.