# Predicting Cervical Spine Injuries in Children

Final Project: Statistics 215A, Fall 2021

Andy Shen, Licong Lin, Seunghoon Paik

# Contents

# 1    Introduction and Domain Problem

Cervical spine injuries (CSI) include injuries sustained to the spinal cord connecting the head and neck to the rest of the body. These injuries, albeit rare, are extremely severe for children whose bodies are still developing. CSI detection strategies include the use of a **computed tomography (CT) scan**, among others. While these interventions are generally effective at pinpointing injuries, they have the potential to inadvertently expose children to ionizing radiation and cause unnecessary pain (Leonard et al. (2011)).

The rarity of a CSI, in conjunction with the potentially harmful effects of CSI detection methods, force clinicians and specialists to assess the risk of performing a CT scan against the potential of missing critical injury diagnoses. The decision of whether to perform a CT scan or not is supported through a **clinical decision rule (CDR)** which uses prior data of potential CSI injuries to predict whether a patient has sustained a true CSI or not. Doctors use the results of the CDR and their own assessment of the situation to decide if the risk of a CT scan outweighs its potential benefits.

Current predictive capabilities in this regime have high sensitivity but low specificity values. This means that, while more children are undergoing CT scans, these scans do not detect injury, resulting in the aforementioned consequences. **Therefore, our objective is to improve on current CDR methods by increasing specificity without sacrificing sensitivity, all while maintaining interpretability throughout the process.**

In this project, we develop a CDR for predicting CSI in children, based on data provided by the Pediatric Emergency Care Applied Research Network (PECARN). The data are described in Section 2. **Our objective is to predict the true outcome of a patient's trauma (whether they truly have a CSI).** We ensure all components of our CDR are interpretable and realistically accessible to physicians when speculation of a CSI is called into question. Throughout the modeling process, we ensure that our predictions are accurate, stable, and capture the breadth of possible injuries and the varying medical and demographic history of each patient.

The structure of this report is as follows: In Section 2, we explain our data and the features used to predict CSI. In...

# 2    Data Collection

Our data is provided by the Pediatric Emergency Care Applied Research Network (PECARN), which contains patient CSI information from a previous CDR study. The previous study contained children both with and without CSI. The children without CSI were grouped into various controls based on their injury mechanism and whether they had received out-of-patient emergency medical services (EMS) treatment. In our study, we are only interested in distinguishing true injury from simple trauma. We group all of the sub-control groups into a single control group for more refined predictions.

The data consist of 12 datasets, each containing various categories of information about the patients, such as demographics, medical history, mechanism of injury, their appearance at the time of response and arrival at the hospital, among others.

Some of the datasets have the same name, but are denoted with an `field`, `out` or `site` at the end of the file name (e.g `clinicalpresentationfield` vs `clinicalpresentationsite`). The suffix "field" denotes measurements of the patient at the location where the injury was sustained, such as the site of the car crash or the patient's home. The suffix "out" refers to information reported by an outside hospital referring the patient to the PECARN system. Note that this information is not available for patients who were not referred by an outside hospital. The suffix "site" refers to information collected at the PECARN hospital site. **The site information is always collected last**, with the outside hospital information collected between field and site if pertinent.

It is also important to recognize that this data is extremely prone to human and/or machine error. We describe potential lapses in the data collection process in Section 2.1. The data cleaning process is discussed

in Section 3.2.

## 2.1   Meaning

In order to better group the variables in our data, we separate them into **three distinct categories: demographic information, injury mechanism, and trauma presentation.** Demographic information simply refers to the patient's basic information, such as age, gender, and ethnicity. While these features may have very little predictive power, they can be useful in grouping injured or non-injured patients post-hoc.

Injury mechanism refers to how the patient sustained their trauma. Examples of injury mechanism include a vehicle accident, child abuse, assault, or falling. Each category of injury is then further elucidated into more specific modes of injury (such as a rear-end collision vs a side impact collision). Most mechanisms of injury in the given data are injuries sustained from various youth sports or from a motor vehicle accident.

Finally, trauma presentation refers to the condition of the patient when they arrive to the hospital or at the site of trauma. These features are spanned across multiple datasets. Features in this category include whether the patient is conscious or not, whether they report neck or facial pain, whether their mental status appears to be normal, among others.

Note here that some features across these three categories preclude accurate reporting of others. For example, if a patient is unconscious or cannot communicate properly, they will be unable to properly describe whether they are in pain.

For our study, we prioritize using the "site" data if it is available, imputing necessary missing values from outside hospital data and site data in that order. We select the site data since it is most recent in terms of a patient's trauma presentation. Features in the other two categories do not differ across location and are named in an unsuffixed dataset.

The non-demographic variables all measure some aspect of the patient's trauma and allows us to assess how serious it is. Each piece of data was recorded from multiple different perspectives. The measurements were collected differently across 2-3 different locations and at 2-3 different time periods - the actual values could differ across time and location. Moreover, the instruments used to take each measurement could be calibrated differently or the individual taking the measurement could have made a mistake. It is very important to keep these sources of error in mind in the modeling process and when questioning unusual observations.

By assuming our data is correct, we are assuming that no errors were made in the data recording process. We also make other assumptions stated in more detail in Section 2.2.

## 2.2   Relevance and Randomness

The data provided are relevant to our study. The features described in our model are all accessible to a physician when they make their clinical decision. If any additional data is necessary, it could include other potentially useful features that are typically made available to a physician at the time of making the clinical decision. We do not collect any more data since we are provided with all of the data recorded by PECARN and it is highly unlikely that additional data that is pertinent to this study even exists.

What is the randomness in this PECARN data set? Is it a random sample from a population? Which one? Why can the data be viewed as a random sample? What assumptions are being made? Can one check these conditions using the info on the data collection process?

# 3   Exploratory Data Analysis

## 3.1   Reducing the feature space

**ZZQ I want to make this a bit better.**

3

Our initial feature space included over 300 features and an interpretable model only requires a small fraction of these features. We trim down our initial subset of features through the consultation of domain experts in cervical spine injury. Utilizing the assistance of Dr. Michael Boyle, a physician in the UC San Francisco School of Medicine, we reduced our feature space to roughly 51 variables to perform variable selection on.

The omitted features were mostly redundant information or information that would not be provided to a physician when making a clinical decision. These omissions were largely determined through domain-expert judgment calls. When in doubt, we retained a predictor to avoid potentially losing crucial predictive information. As discussed in Section 4.2, we use a more rigorous variable selection process to determine our best features for modeling.

## 3.2   Data Cleaning

After reducing the feature space, we implement several layers of judgment calls in our data cleaning procedure. Any judgment call that directly involved medical subject-matter expertise was verified by Dr. Boyle.

***Missing values***: Certain columns in the data contain blank or missing values. Features with a proportion of missing values above 15% are removed from this analysis. The remaining missing values are imputed using *defensive imputation* which assumes the worst-case scenario for the patient. For instance, if we do not have data on whether a patient has tenderness in their neck, we assume they do. Note that this does not default to imputing all missing values with "yes." We must consider what each feature is asking for and determine the worst-case scenario accordingly.

***Categorical features***: In order to create an interpretable decision rule, we map our categorical features into binary 0-1 features, **with 1 corresponding to "yes" and 0 corresponding to "no."** To do this, we maintain the defensive imputation strategy by determining the worst-case outcome for each feature. This consisted of mapping unknown values using defensive imputation and mapping other variables using common sense. For example, a value of `S` which stands for "suspected, but unknown" would map to "yes." Fortunately, the data contains documentation which provides meaning to the possible values for each feature.

***Ordinal features***: Our feature space contains two ordinal features: `AVPU` and `TotalGCS`. AVPU stands for "Alert, Verbal, Pain, Unresponsive," and is a widely accepted health care protocol to determine a patient's overall status. Each letter in "AVPU" corresponds to a more urgent patient status. Alertness and unresponsiveness are simply whether the patient is alert or unresponsive. "Verbal" and "Pain" correspond to whether the patient is receptive to verbal or painful stimulation. In our case, we treat anything that is not alert as the worst-case scenario and impute accordingly. Failing to respond to any stimuli whatsoever should be heavily scrutinized and our imputation takes this into account.

The Glasglow Coma Scale is a similar score used to determine one's consciousness level. It is a sum of three sub-tests and is scored on a scale from 3 (completely unconscious) to 15 (fully conscious). Here, we classify any score above 3 as the worst-case scenario.

Our final cleaned data contains 5037 control patients and 927 patients who were truly injured with CSI.

## 3.3   Correlation between features

Figure 1 shows a correlation heatmap of the reduced feature space of size 52. While most features are not strongly correlated with each other, there are some features that exhibit high correlations. These features are typically those of the same type, such as whether the patient reported pain in their neck and whether tenderness was observed in the patient's neck. It is expected for features like these to be influenced by one another.

Moreover, most features have absolute correlation $|\rho| < 0.5$, indicating weak association throughout the feature space in general. Because of this, we feel confident that we will not encounter any issues surrounding variance inflation and multicollinearity in the modeling stage.

Figure 1: Correlation heatmap of reduced feature space.

## 3.4 Correlation with outcome

After seeing which features are associated with each other, we then examine the correlation of each feature with respect to the outcome variable (whether the patient truly has a CSI). We find that most features are not highly correlated with the outcome in either direction, with some exceptions: the `IntervForCervicalStab` feature, which tells whether the patient underwent any cervical stabilization measures at the site, has the highest correlation (above 0.60) with the outcome. This should not be surprising since a cervical stabilization measure is only imposed on a patient if a medical official on-site thinks it is necessary, thus we expect most true injuries to be highly associated with this intervention.

Similarly, `DxCspineInjury` measures whether the patient is suspected of having a CSI. This has the second-highest correlation of roughly 0.45, which also should not be surprising. We discuss this variable in more detail in Section 4.4.

## 3.5 Frequency of feature values

We further explore our data by examining the frequency of "yes" and "no" values for the outcome and controls group. We plot the five features with the greatest and smallest *absolute differences* in relative frequency for the outcome group in Figure 2 and for the controls group in Figure 3.
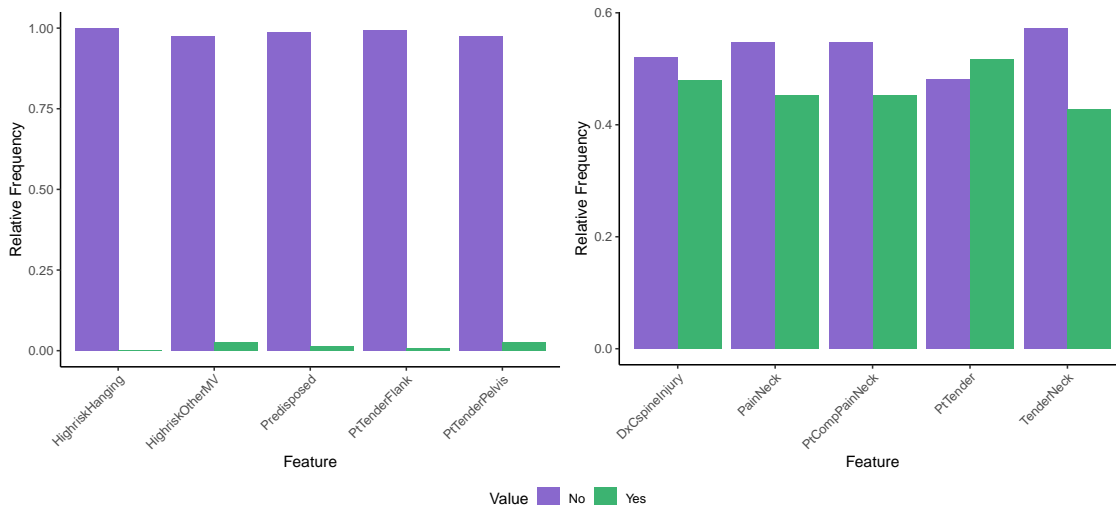


Figure 2: Top five features with the greatest and smallest absolute differences in relative frequency for the OUTCOME group.
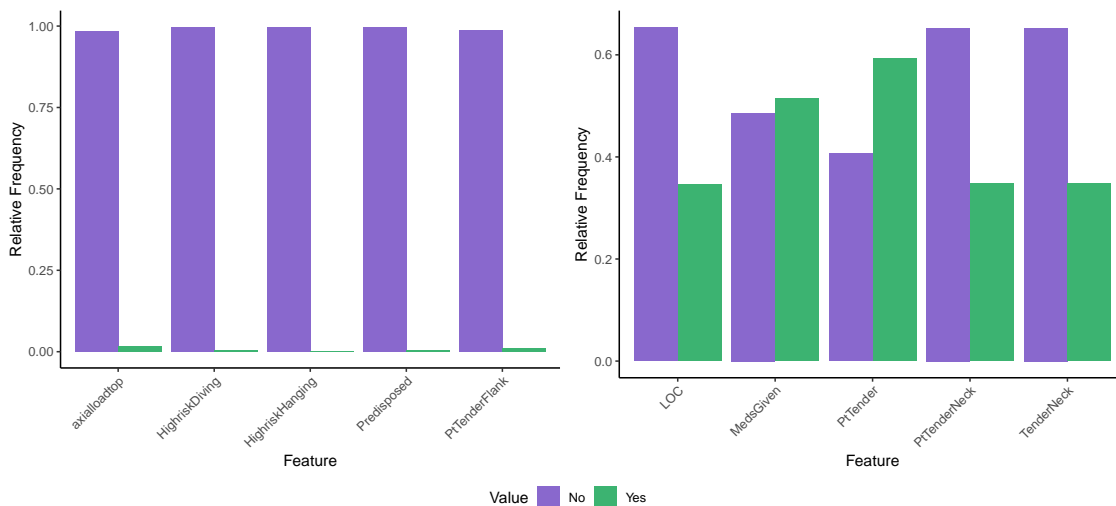


Figure 3: Top five features with the greatest and smallest absolute differences in relative frequency for the CONTROLS group.

These figures show that, while there is some overlap across the two sub-groups, we cannot say that there is complete uniformity in which features stand out between the outcome and control groups. For instance, the `DxCspineInury` feature, which tells us whether the patient is arriving with a diagnosis or suspicion of a CSI, has the second-lowest difference in yes/no rate for the outcome group (48% vs 52%), whereas it is somewhere in the middle for the control group (where roughly 92% were not suspected). This feature proves to be extremely crucial in the modeling stage and it is discussed further in Section 4.5.

# 4 Modeling

## 4.1 Translation and Comparability

<span style="color:red">How should one translate the question in (1) into a statistical question regarding the data to best answer the original question? Are there multiple translations? For example, can we translate the question into a prediction problem or an inference problem regarding a statistical model? List the pros and cons of each translation relative to answering the substantive question before choosing a model.</span>

Our statistical goal is to accurately diagnose a cervical spine injury while keeping the Type II error rate as low as possible. In the context of CSI, a Type II error results when our CDR determines no CSI is present, when this is truly not the case. This statistical flaw bears serious and life-threatening consequences on children and we must pay special attention to this rate.

However, if the false negative rate is our only area of concern, there is no need for any decision rule since a clinician would diagnose a CSI and perform a CT scan for all situations. As mentioned in the Introduction (Section 1), CT scans could result in unnecessary exposure to ionizing radiation, which is also harmful. Therefore, we must strike a balance between taking the conservative route of always diagnosing someone with CSI and the worst-case scenario of not diagnosing a serious injury.

Regarding comparability, the training and test data come from the PECARN sites and are independent. We have no reason to assume that the data comes from another distribution. Moreover, the injury status of one patient is independent of another patient's injury status. Even if two children were injured from one another (colliding heads in soccerm for example), whether one child sustains a CSI has no bearing on whether the other child does. This is an extreme example and most data points are indeed completely independent.

Prior to modeling, we split our data into *training*, *tuning*, and *testing* sets. The training set will be used to construct our candidate models. The tuning set will be used to select an optimal model, and the testing set will be used to corroborate the performance of our best model.

These splits were determined randomly as there is no underlying temporal or spatial phenomena that makes the outcome of one patient dependent on another.

## 4.2 Feature and hyperparameter selection

We use **logistic regression** followed by **backwards selection** to select our model features. We first fit the outcome against all features with logistic regression and then omit the feature with the smallest absolute fitted coefficient value. There is no need to scale our features since they are all binary. This process repeats until $m$ features remain, where $m$ is a pre-specified number. This variable selection process is called recursive feature elimination (RFE) in Python and can be done automatically.

To select the number of features $m$, we run logistic regression model with $m = 1, 2, ..., 10$ features on the training data and select the one with the lowest misclassification rate on the tuning data, resulting in an optimal value of $m = 9$. These 9 features correspond to:

- Whether the patient was intubated or not (`ArrPtIntub`),

- Whether the patient is suspected of having a CSI (`DxCspineInjury`),

- Whether the patient has focal neurological deficits such as a spinal cord issue (`FocalNeuroFindings`),

- If the sustained trauma was the result of diving (`HighriskDiving`),

- Whether the patient required a cervical stability intervention such as a collar or brace (`IntervForCervicalStab`),

- Whether the patient has extremity weakness (`PtExtremityWeakness`),

- Whether the patient has sensory loss (`PtSensoryLoss`),

- Whether the patient has tenderness in the extremities (`PtTenderExt`), and

- If the trauma was sustained to the patient's torso or trunk (`SubInj_TorsoTrunk`).

These features were discussed with Dr. Boyle, who verified their legitimacy in a real-life clinical context. Moreover, common sense knowledge tells us that these features are meaningful and serious with regards to CSI's.

## 4.3 Models

### 4.3.1 AdaBoost

**ZZQ Licong or Hoon, please provide a description for each, as well as diagrams if applicable.**

### 4.3.2 Decision Tree

**ZZQ Licong or Hoon, please provide a description for each, as well as diagrams if applicable.**

I think a diagram for the decision tree would be helpful.

### 4.3.3 Logistic Regression

**ZZQ Licong or Hoon, please provide a description for each, as well as diagrams if applicable.**

## 4.4 Best model

The sensitivity and specificity rates for all three models are shown in 4. The values in the figure are rounded to two significant figures but the sensitivity rates for all three models are identical. Moreover, the specificity rates for AdaBoost and Logistic Regression are identical.

The Receiver Operating Characteristic (ROC) curves for all three models are shown in Figure 5. We show here that a simple decision tree can achieve effectively the same error rate as something more powerful. We achieve over three times the specificity of Leonard et al. (2011) while maintaining the same sensitivity. Statistically speaking, we increase the power over threefold without sacrificing the potential for a Type I error.

**Ultimately, we believe the decision tree is the best model for a clinical decision rule.** The model is interpretable, easy-to-follow, while sacrificing a truly negligible amount of specificity. A clinician at a patient's bedside can quickly reach a decision simply by answering a series of yes or no questions, which is effectively what a classification tree is.
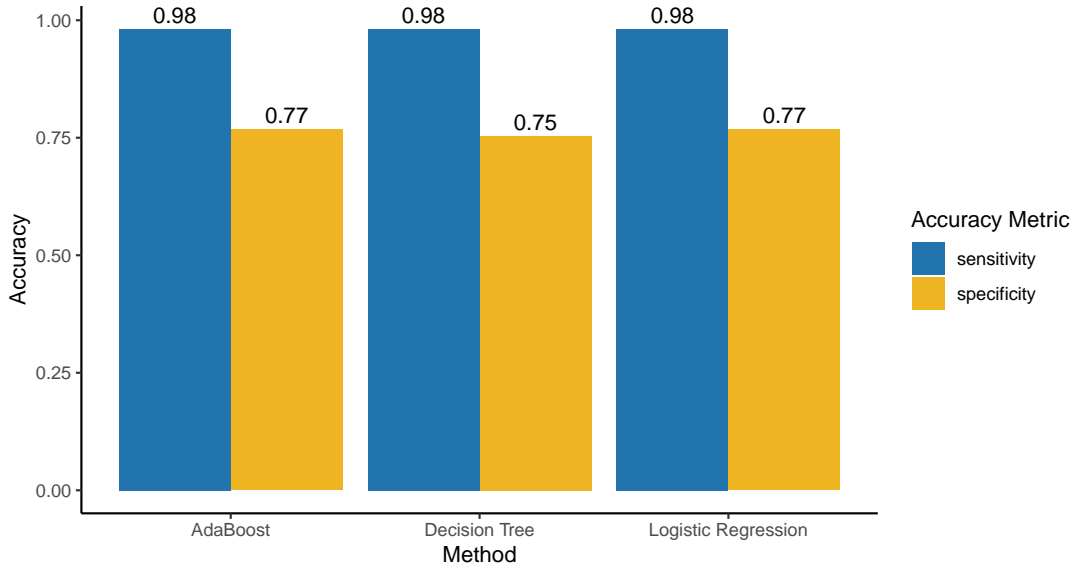
Figure 4: Sensitivity and specificity rates for three candidate models: logistic regression, decision tree, and AdaBoost. Accuracy rates are assessed at each model's optimal hyperparameter value.
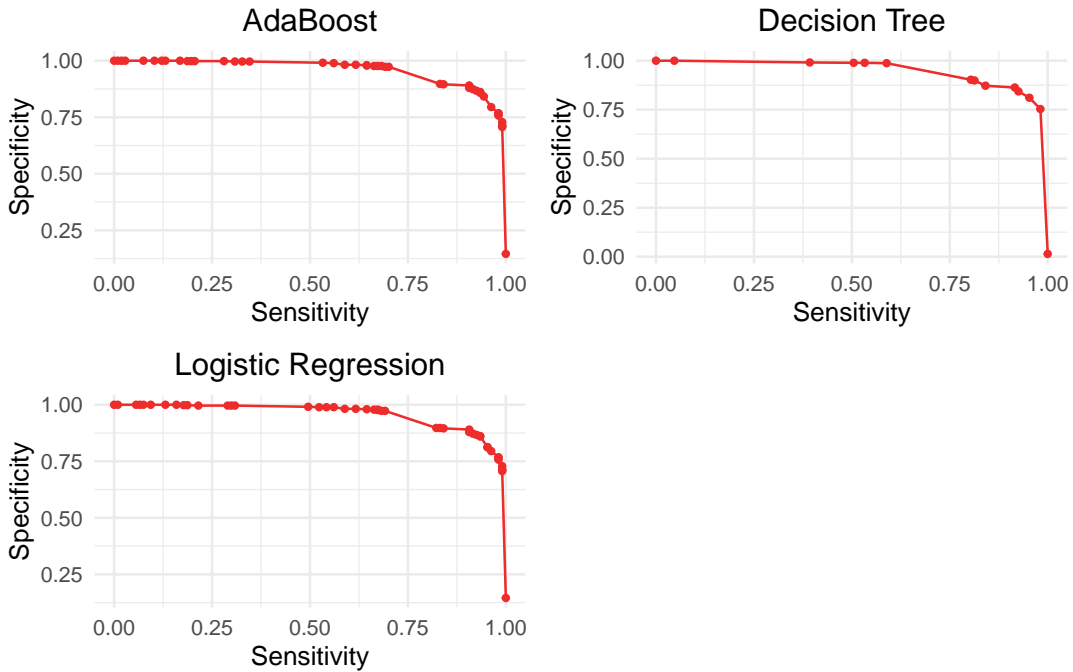


Figure 5: ROC curves for the three candidate models.

## 4.5 Caveats

Despite its strong performance, our model has domain-specific caveats that we must adjust for and address.

First, the `DxCspineInjury` feature (explained in Section 3.4) has a unique context. This feature is literally a guess as to whether this person has a CSI or if they arrived with a diagnosis. Such a guess information could be revealing of the true outcome even though it is a feature made available to a physician prior to

making the CDR. In order to use a feature like this, we must provide a thorough investigation of the feature and ensure that it is indeed reasonable.

A rigorous inspection of the raw data reveals that this feature has **zero** missing or indeterminate values. Thus, we can rely on this information to be provided in general. We then examined how this feature was recorded. Fortunately, our data documentation contains a copy of the form that a nurse or EMT fills out when examining the patient. The question of CSI suspicion appears in the survey along with the other questions that are part of our feature space. If we picture ourselves as the nurse or physician, we can see them record their personal judgment of the patient just like they will record other measurements, such as AVPU or neck tenderness, for example. A copy of the questionnaire with the `DxCspineInjury` feature is shown in Figure 6.



Figure 6: Copy of data recording mechanism for CSI suspicion.

Moreover, as depicted in Figures 2 and 3, the true diagnosis is only confirmed about 50% of the time in the outcomes group compared to 92% of the time in the controls group. We can infer from these results that most of these guesses are not in favor of a CSI. This is indeed true, and roughly 85% of all patients are not predicted to have a CSI.

All of the arguments above were discussed in detail with Dr. Boyle. Therefore, we are confident that this feature can be utilized in a real-world context. Not only is its distribution statistically reasonable, but we have no reason to believe that this feature is fundamentally different from other features in terms of how it is documented. The only difference is that it relies on a human judgment call, but such a judgment call is being made by a trained professional such as a nurse or EMT.

# 5 Stability

What off-the-shelf method will you use? Do different methods give the same qualitative conclusion? Perturb one's data, for example, by adding noise or subsampling if data units are exchangeable (in general, make sure the subsamples respect the underlying structures, e.g. dependence, clustering, heterogeneity, so the subsamples are representative of the original data). Do the conclusions still hold? Only trust those that pass

the stability test, which is an easy-to-implement, first defense against over-fitting or too many false positive discoveries.

**ZZQ Seunghoon**

# 6 Conclusion

# References

Leonard, Julie C, Nathan Kuppermann, Cody Olsen, Lynn Babcock-Cimpello, Kathleen Brown, Prashant Mahajan, Kathleen M Adelgais, et al. 2011. "Factors Associated with Cervical Spine Injury in Children After Blunt Trauma." *Annals of Emergency Medicine* 58 (2): 145–55.