Predicting Low Risk of Clinically-Important TBI in Children: Vetting and Improving Existing Classifiers

Jimmy Butler, Andrej Leban, Ian Shen, Xin Zhou

Introduction and Problem Motivation

Among children ages 0-14, traumatic brain injuries (TBIs) are a leading cause of injury and death, accounting for over 2,100 deaths, 35,000 hospital admissions, and 470,000 emergency department visits annually (Faul (2010)). Accurately diagnosing TBIs is thus a crucial problem, especially since the earlier a TBI is identified, the sooner potentially life-saving treatment can be administered.

Typically, TBIs are diagnosed via computed tomography (CT) scans of a patient's brain. However, performing CT scans carries the risk of radiation-induced malignancy in the patient. This risk is especially pertinent for children, as younger individuals are at a higher risk of developing such malignancies. Therefore, while it is crucial to rapidly and accurately diagnose TBIs, it is also imperative that CT scans are performed only when there is reasonable suspicion the individual has a TBI. Otherwise, the risk of radiation-induced complications from the CT scan will outweigh any potential benefit of having CT scan results.

Nathan Kuppermann and his colleagues explored this problem in their 2009 Lancet paper (*Identification of children at very low risk of clinically-important traumatic brain injuries after head trauma: a prospective cohort study*), seeking to develop interpretable rules that could be used to identify children, who, despite presenting to the emergency room for head trauma, are at low risk of having TBIs and thus can avoid CT scans (Kuppermann (2008)). The purpose of this present study is to build on their work, developing alternative classification schemes to identify children for whom CT scans are unnecessary. With an alternative scheme with high predictive accuracy, one can compare with Kuppermann's to see if similar features were isolated, either reinforcing the importance of their identified covariates, or providing an alternative perspective on what is important in predicting TBI. Additionally, in this analysis, we employ our own data cleaning and preprocessing scheme that we find reasonable, allowing us to assess the stability of their findings subject to perturbations at this stage of the data science life cycle.

In addition to vetting their classification rules, another goal of this analysis is to offer some improvement on their classification strategy. Kuppermann and his colleagues developed rule sets for children aged younger than two and children aged two or older, the reason being that younger children are pre-verbal and thus require a different set of standards to identify TBIs than older children who can verbalize their symptoms. We seek to develop a single age-invariant classifier using covariates common to both primary age groups. We hope that such a classifier will not only identify common predictors of TBI across all ages, but also simplify the decision-making process by obviating the need for a second classifier. Nonetheless, in developing new classifiers, priority will be placed on interpretability, brevity and simplicity, and a low false negative rate without having too high of a false positive rate.

Data

Data Collection

Before going into extensive detail about the dataset used in this analysis, as well as our cleaning and preprocessing steps, we first offer our thoughts on what data would be most relevant for vetting and improving upon previous classifiers.

Besides needing a dataset similar in nature to the dataset on which the original classifiers were based (comprised of children presenting to ER with head trauma, etc.), to vet Kuppermann's classifiers, we need a dataset that will enable us to develop classifiers with high predictive accuracy, a low false negative rate,

and maximal generalizability. For instance, we would need a dataset containing covariates that are both relevant to the presence of TBIs and can be measured at the time a child is ushered into the ER. From common sense, as well as consulting with Dr. Robert Inglis of UCSF, it seems important information might be the injury mechanism (and a rating of severity), any symptoms such as headache or amnesia, and any results of a physical examination that may indicate neurological impact or skull injury. For both vetting the clinical decision rule and developing an age-invariant rule, it would be preferable to have covariates that are as objective as possible so as to avoid issues with interrater reliability and improve generalizability of the rules to novel, real-world settings. It would also be preferable to both have details on covariates invariant to gender, race, ethnicity, or any other such characteristic of the population of interest and a population with gender, age, and racial diversity, to reduce the potential of systematic misdiagnoses in these groups should the rule be applied in future settings. Finally, we wish to identify children who do not need CT scans among a population where there is already some doubt as to whether the CT scan will be worth the risk of radiation-induced malignancy. So, the ideal dataset will consist of children who are not critically injured and who could conceivably be TBI-free.

To develop an age-invariant classifier, the above points are certainly relevant. However, specific to an age-invariant classifier, we would also need covariates whose presence or values are not age-dependent.

The dataset used in this analysis is the same one used by Kuppermann and his colleagues to develop their classification rules, consisting of 43399 children who presented to the emergency room for head trauma at one of the PECARN network hospitals between 2004 and 2006. For each child, the data were collected by physicians who filled out standardized sheets detailing the child's condition, including both questions for the child ('Do you have a headache?') and results of a physician examination ('Palpable skull fracture?'). Additionally, the dataset contains information that would be available after the decision to pursue a CT scan has been made, such as the reasons why a CT scan is being pursued, whether a child will need to be sedated for a CT scan, and whether there were any traumatic findings on the scan. The outcome variable (TBI) was obtained by tracking the child's progress in the hospital after the initial examination (whether they were admitted for neurological surgery, or died in the ED, for example). If the child left the hospital, the researchers followed-up with the child's parents, the medical record, emergency department process improvement records, and county morgue records to ensure the child did not end up having a clinically-important TBI according to their definition. In total, there are 125 covariates, including the outcome.

It seems that in this dataset, most of the variables we conjecture are important for predicting TBIs are present, but that there are many covariates whose information would not be known at the time our classifier would be implemented. Additionally, many covariates seem more objective than others (*Palpable skull fracture?* vs. *Is the child acting normally?*). There are also many variables which are not invariant to age. For instance, the ability to say you have a headache is reserved for verbal children who must be above a certain age. These variables and decisions on what to do with them will be discussed at length in Cleaning and Preprocessing, as well as Post-Processing EDA. The dataset also provides demographic characteristics such as age, gender, race, and ethnicity. This information will be discussed at length in Post-Processing EDA.

Cleaning and Preprocessing

In order to achieve a suitable dataset for the predicting low risk of TBI in children according to the above parameters, we first implement a set of cleaning and preprocessing steps. The overarching goal of these steps is to keep only covariates relevant to predicting TBI at the time a child would present to the ER, keep only observations for whom this classification rule would be relevant (children for whom it's a toss-up as to whether they'll need a CT scan), and to deal with missing values. We present our procedure below, along with relevant judgement calls and their perturbations where applicable. Many of these decisions were made using common sense, as well as consultation with Dr. Inglis.

Steps and Explanations

- Remove columns irrelevant to predicting TBI at the time a child enters the emergency room
  - *EmplType*, *Certification*, *Ind...*, *CT...*, *Finding1*, ... *Finding23*, *EDDisposition*, *Observed*, *AgeTwoPlus*, *AgeInMonth*
  - Judgment Call [`injMech`]: drop *InjuryMech*
    - Perturbation: keep *InjuryMech*

The above-listed columns consist of the type of employee performing the exam as well as their certification, the reasons why a CT scan was ordered, whether sedation would be needed when performing a CT scan, the traumatic findings from the CT scan, the reason why the child left the ER, or whether the child was observed again to see if they needed a CT scan. We remove the former two variables because domain knowledge says that we should not expect the qualifications of the physician to systematically affect the responses collected. The rest of the variables would either not be available at the time a child first enters the ER, or would be known only after the decision to obtain a CT scan has been made. So, these are not of interest. We also remove an indicator for the individuals being above age 2, as well as the age provided in months, opting to keep the age provided in years. Similarly, according to judgment call [`injMech`], we drop the injury mechanism, a 13-level categorical variable detailing how the injury occurred, opting to keep a severity variable, a 3-tiered categorical variable rating the severity of the injury mechanism as 'low,' 'moderate,' or 'high' (*High_impact_InjSev*).

- Remove columns with high levels of missingness
  - *Dizzy*, *Ethnicity*

Dizzy, an indicator of whether the individual felt dizzy, and *Ethnicity*, whether the individual was Hispanic, are both missing in approximately 35% of the observations. *Ethnicity* would be useful for posthoc analysis, but with such a large number of missing values, it may not be particularly informative. *Dizzy* could be a good indicator of TBI, but was deemed too subjective and prone to inter-rater reliability issues according to both Kuppermann and Dr. Inglis.

- Remove observations whose *GCS* scores are less than 14

GCS, or Glasgow Coma Score, measures an individual's level of consciousness on a scale from 3 to 15, and is the sum total of three subscores: an eye score, a verbal score, and a motor score. If GCS is less than 14, the individual is quite injured and would likely need a CT scan anyway. We only keep individuals where the total GCS is 14 or 15.

- Impute and Drop based on GCS scores
  - Judgment Call [`missSubGCS`]: drop observations with total GCS of 14, but missing any of the GCS subcategory scores (*GCSVerbal*, *GCSMotor*, *GCSEye*)
    - Perturbation: keep the observations
  - Impute a full score into each GCS subcategory when total GCS is 15
  - Judgment Call [`fake15GCS`]: drop observations with total GCS of 15, but suboptimal score in a subcategory
    - Perturbation: keep the observations
  - Judgment Call [`fake14GCS`]: drop observations with total GCS of 14, but optimal score in every subcategory
    - Perturbation: keep the observations

GCS score is likely an important predictor of TBI, so it is imperative this variable is as completely filled-out as possible in this dataset. However, it should be noted that there are different standards for obtaining GCS scores for individuals two and younger and those older than two, since individuals two and younger cannot verbalize. We note this as a potential problem to developing an age-invariant classifier.

- Remove observations who were pharmacologically paralyzed, sedated, or intubated at the time of evaluation, or for whom any of this information is missing
  - *Paralyzed*, *Sedated*, *Intubated*

According to domain knowledge, individuals who must be pharmacologically paralyzed, sedated, or intubated at the time of evaluation are likely quite injured and do not represent the population of interest for these classification rules.

- Drop and impute based on variables of altered mental status
  - *AMS*, *AMSAgitated*, *AMSSleep*, *AMSSlow*, *AMSRepeat*, *AMSOth*
  - Judgment Call [`AMS`]: drop observations missing any of the above variables
    - Perturbation: impute missing *AMS* based on presence of any of the subvariables, and drop the subvariables and any remaining missing *AMS* observations

AMS is an indicator representing if the child has altered mental status, and the remaining variables are subvariables indicating the reason why the child has altered mental status (agitated, sleepy, slow to respond, asking repetitive questions, etc.) This is likely an important predictor, so it is important to

handle missing values. By judgment call `AMS`, we drop any observation with a missing value in any of these categories, since if the observation is missing it is impossible to truly know a value to impute for *AMS*. As a perturbation, we could impute a 1 in *AMS* if any if any of the subcategories is affirmative, and 0 otherwise. However, it would be impossible to impute for missing values in the subcategories, so we drop these observations.

Note that the above judgement call and its perturbation will be commonly used for other sets of variables where there is some parent quantity (like *AMS*) and children quantities that provide more detail on the parent quantity. We refer to the judgement call where we drop observations missing values in *any* subcategory as `Strong`, and the hudgement call where we impute missing values of the parent variable and drop the children as `Weak`.

- Drop and impute based on variables of other non-head substantial injuries
  - *OSI*, *OSIExtremity*, *OSICut*, *OSICspine*, *OSIFlank*, *OSIAbdomen*, *OSIPelvis*, *OSIOth*
  - Judgment Call [`OSI`]: `Strong`
    - Perturbation: `Weak`

OSI is an indicator on whether the physician determined the child sustained substantial non-head injuries. This could be an injury to the extremities like a fracture, a cut requiring surgery, a spinal cord injury, a flank injury, an intra-abdominal injury, or a pelvis injury. Although not directly related to head trauma, we find this variable might be an indicator of the injury severity. If another variable relevant to head trauma fails to indicate a TBI for a particular individual, including this variable offers another opportunity to diagnose the TBI, possibly suppressing the false negative rate. By judgement call `OSI`, we drop any observations missing any of these values or their subvariables for the same reason as judgment call `AMS`.

- Drop and impute based on hematoma information
  - *Hema*, *HemaSize*, *HemaLoc*
  - Judgment Call [`HEMA`]: `Strong`
    - Perturbation 1: `Weak`
    - Perturbation 2: Only drop missing *HemaLoc* and *Hema*, impute missing *Hema* from *HemaLoc* presence, drop remaining missing *Hema*

Hema indicates the presence of a scalp hematoma, or swelling due to some head trauma, *HemaSize* is a categorical variable representing the size ($<$ 1cm, 1-3cm, $>$ 3cm) and *HemaLoc* is a categorical variable representing the location (frontal, occipital, parietal/temporal). According to domain knowledge, scalp hematomas are fairly good predictors of TBI. By judgment call [`HEMA`], we drop all observations missing any of these categories. By Perturbation 1, we only keep the indicators for the presence of hematoma. By Perturbation 2, we keep the indicator as well as the location, since according to domain knowledge, the impact of head trauma to the brain is dependent on the location of the trauma.

- Drop and impute based on skull fracture variables
  - *SFxPalp*, *SFxPalpDepress*
  - Judgment Call [`SFx`]: recode an unclear *SFxPalp* observation as true, drop observations missing either *SFxPalp* or *SFxPalpDepress*
    - Perturbation 1: `Weak`
    - Perturbation 2: recode an unclear *SFxPalp* observation as false, drop observations missing either *SFxPalp* or *SFxPalpDepress*

SFxPalp is an indicator for a palpable skull fracture, which can also be marked as 'unclear' if there is too much swelling to make a determination. *SFxPalpDepress* is an indicator on whether the fracture feels depressed. Both of these are likely important predictors via domain knowledge. As a judgment call, we take an unclear observation as being a palpable skull fracture, since even if the examination was unclear, there is still substantial head trauma.

- Drop missing values for bulging Anterior Fontanelle
  - *FontBulg*

FontBulg is an indicator for whether the child's anterior fontanelle is bulging, indicating pressure within the brain. This feature would typically only be present in infants, since for older individuals their anterior fontanelle will have closed as their skull hardens with age. While possibly a good predictor for TBI in infants, this variable does not have much predictive value for older children.

- Drop and impute based on basilar skull fracture variables

  - *SFxBas*, *SFxBasHem*, *SFxBasOto*, *SFxBasPer*, *SFxBasRet*, *SFxBasRhi*
  - Judgment Call [`SFxBas`]: `Strong`
    * Perturbation: `Weak`

SFxBas is an indicator if the physician has determined there is basilar skull fracture present, or a fracture at the base of the skull. The rest of the variables indicate common indicators of basilar skull fracture (hemotympanum, spinal fluid otorrhea, raccoon eyes, Battle's sign, spinal fluid rhinorrhea). More specifically, hemotympanum is when blood is found in the middle ear, otorrhea is when spinal fluid drains from the ear, raccoon eyes is when internal bleeding occurs below the eye, Battle's sign is a bruise at the bottom of the skull, and rhinorrhea is when spinal fluid leaks out the nose. This is known to be quite an important predictor of TBI, so we wish to make these variables as complete as possible for an ideal dataset.

- Drop and impute based on above-the-clavicle injury variables
  - *Clav*, *ClavNeck*, *ClavFro*, *ClavOcc*, *ClavPar*, *ClavTem*
  - Judgment Call [`Clav`]: `Strong`
    * Perturbation: `Weak`

Clav is an indicator on whether there was a laceration, abrasion, or hematoma above the clavicles (neck). The subvariables indicate whether there was trauma on the neck or the frontal, occipital, parietal or temporal regions of the scalp. In the same way as *OSI*, this variable will be useful to keep around because it could indicate a TBI if one of the head trauma variables fails to.

- Drop and impute based on neurological deficit variables
  - *NeuroD*, *NeuroDMotor*, *NeuroDSensory*, *NeuroDCranial*, *NeuroDReflex*, *NeuroDOth*
  - Judgment Call [`Neuro`]: `Strong`
    * Perturbation: `Weak`

NeuroD is an indicator for whether or not there is some neurological deficit besides altered mental status, as determined by the physician. This includes motor, sensory, cranial nerve (pupil reactivity), or reflex issues. Generally, these categories are quite important for predicting TBI according to domain knowledge. As with previous groups of variables, we implement a judgment call to drop any observation missing any of these values.

- Drop and impute based on vomiting variables
  - *Vomit*, *VomitNbr*, *VomitStart*, *VomitLast*
  - Judgment Call [`Vomit`]: `Weak`
    * Perturbation: `Strong`

Vomit is an indicator for whether the individual vomited (as reported by child or parents), *VomitNbr* is the number of times the child vomited post-injury (1, 2, $>$ 2 times), *VomitStart* is how soon after the head injury the vomiting started (within 1 hour, 1 - 4 hours, $>$ 4 hours after), and *VomitLast* is how soon before the evaluation was the last episode ($<$ 1 hour, 1 - 4 hours, $>$ 4 hours before evaluation). As a judgement call, we only keep *Vomit* and impute missing values from the subvariables, since according to domain knowledge, more details about vomiting is generally unhelpful. As a perturbation, we keep these covariates.

- Drop and impute based on headache variables
  - *HA_verb*, *HAStart*, *HASeverity*
  - Judgment Call [`HA`]: Drop missing *HASeverity*, impute missing *HA_verb* from *HASeverity*, and drop remaining missing *HA_verb*, ignore *HAStart* column
    * Perturbation 1: `Weak`
    * Perturbation 2: `Strong`

HA_verb is an indicator on whether the child had a headache, and is marked differently if the child is too young to verbalize that they had a headache, or unable to speak. *HAStart* and *HASeverity* are both variables indicating when the headache began (before injury, within 1 hour after, 1-4 hours after, $>$ 4 hours after) and the severity (mild, moderate, and severe). As a judgment call, we keep *HA_verb* and *HASeverity* since these are both more likely to predict TBI than *HAStart*. However, it should be noted that this is an age-dependent variable as discussed in the prior section, since the ability to verbalize headache symptoms requires the child to be of a certain age.

- Drop and impute based on seizure variables

  - *Seiz*, *SeizLen*, *SeizOccur*
  - Judgement Call [`Seiz`]: Drop missing *SeizLen*, impute missing *Seiz* from *SeizLen*, drop missing *Seiz*, and ignore *SeizOccur* column
    * Perturbation: Drop missing *SeizLen* and *Seiz*, exclude *SeizOccur* column

Seiz is an indicator if the child had a seizure, as reported by the parent or child. *SeizLen* is how long the seizure lasted (< 1 min, 1-5 min, 5-15 min, >15 min) and *SeizOccur* is when did it occur relative to the injury time (immediately on contact, within 30 minutes, or > 30 minutes after injury). Whether a seizure occured is important to consider, and the length is also somewhat important. So, we opt to keep *SeizLen* and *Seiz*, but drop *SeizOccur*, since this is not as important according to domain knowledge.

- Drop and impute based on loss of consciousness variables
  - *LOCSeparate*, *LOCLen*
  - Judgement Call [`LOC`]: recode 'suspected' as 'yes' in *LOCSeparate*, and drop observations with missing *LOCSeparate* and *LOCLen*
    * Perturbation 1: `Strong`
    * Perturbation 2: recode 'suspected' as 'no' in *LOCSeparate*, drop missing *LOCSeparate* and *LOCLen*

LOCSeparate is an indicator if the child lost consciousness at all after the injury, and *LOCLen* is how long the child lost consciousness (< 5 sec, 5 sec - 1 min, 1 - 5 min, > 5 min). Loss of consciousness is generally a good predictor of head trauma, so we wish to include this to predict TBI. As a judgement call, we consider a suspected loss of consciousness as affirmative, since if there is any doubt as to whether the child lost consciousness, this should be of concern to err on the side of caution.

- Drop missing values for *Amnesia* and *High_impact_InjSev*

Amensia indicates whether the child has amnesia for the event, but can only be answered by verbal children. This is an example of a variable that is not age-invariant, although it could be a good predictor for verbal children. *High_impact_InjSev* is the rating of the injury mechanism severity, as mentioned above.

- Impute missing values for *ActNorm*
  - Judgment Call [`ActNorm`]: recode null values as affirmative
    * Perturbation: recode null values as not affirmative

ActNorm is a subjective assessment by the parent as to whether the parent thinks the child is acting normally. Although subjective, Dr. Inglis concluded that a parents' assessment of their child's behavior is a good indicator of how severely the child may be injured. Because this is important, we wish to impute as many values as possible. As a judgement call, we consider a missing value as being normal, since if the child was acting normally, it's possible this question could have simply been skipped over. As a perturbation, we reverse this judgement call.

- Make a pooled outcome variable

We define the outcome, a clinically-important TBI, in the same way as Kuppermann and his colleagues, except we impute missing values in the outcome by the presence of any of the following: death due to TBI, intubated for more than 24 hours, neurosurgery was performed, and the individual was hospitalized for more than 2 nights due to head injury. This should give a complete assessment as to whether an individual had a TBI.

Data Summary

Once we have cleaned the data according to the above steps, we wish to extract binary features from the categorical variables, where applicable. For every subvariable/ parent variable pairing mentioned above (for example, `NeuroDMotor` and `NeuroD`), if the parent variable is marked as 'False,' the subvariable has a value indicating it is missing. As a default, we maintain this missing category, and binarize the variables by one-hot encoding. Alternatively, we could impute 'False' into the missing values and drop the parent variables, so now each subvariable indicates the presence of the subvariable AND the parent variable as opposed to the subvariable GIVEN the parent variable.

After performing this one-hot encoding, we have 137 features in total. We split the data into a training set (19947 observations), a validation set (6649 observations), and a test set (6649 observations). For

exploratory data analysis, we only consider the training set, saving the test set to check model accuracy. We should note that there may be cluster structure in the dataset by virtue of the data being collected from different hospitals, presenting a possible comparability issue. However, we are not provided with hospital information, so we simply take the first 60% of observations as training, and divide the remaining 40% evenly into validation and test sets.

Exploratory Data Analysis

Data Collection detailed a list of characteristics of an ideal dataset to vet and improve upon the original decision rule, according to the goals outlined in Introduction and Problem Motivation. In this section, we explore the relevance of the cleaned and preprocessed data to answer the question at hand, assessing how the data meets these ideals via exploratory plots. The results of our exploratory analysis will then be used to motivate decisions regarding feature inclusion and sample splitting as we build our classifiers.

Of course, the dataset already meets some of these conditions by virtue of our cleaning process: the cleaned dataset only includes individuals with GCS scores of 14 or 15, and individuals who were not so injured as to be intubated, pharmacologically paralyzed, or sedated at the time of evaluation. Additionally, since we are using the same original dataset as Kuppermann, where children presented to the ER with head trauma, our results can be suitably compared to theirs. Furthermore, since our data was cleaned differently from theirs, we get an added stability assessment of their decision rule for free, a crucial element of vetting their work.

As mentioned in Data Collection, one of the primary criteria of a good dataset for this problem is a dataset which is diverse in demographic characteristics, including age, race, and gender. This is crucial for generalizability of our work, as a classifier built from a dataset that is demographically lopsided could exhibit unforeseen biases when exposed to more diverse data.

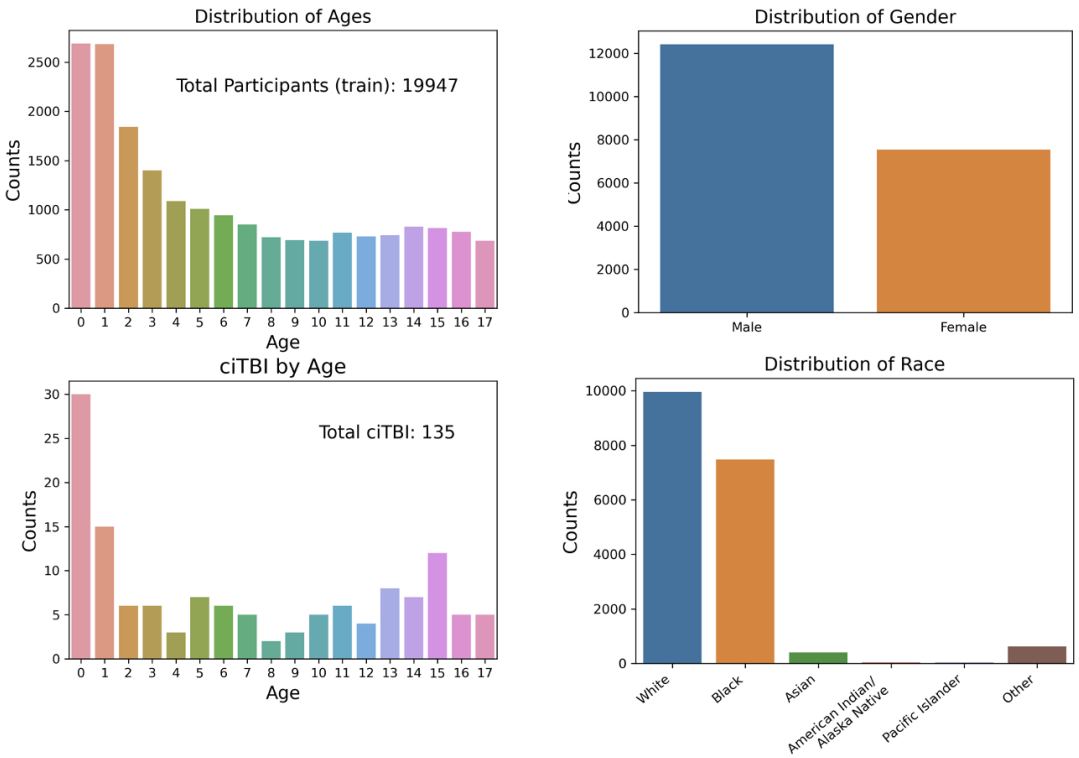

Figure 1: Distributions of various demographic variables in the training dataset, after cleaning and preprocessing.

Figure 1 shows distributions of various demographic variables in the cleaned training dataset. In the top-left plot, we see that younger ages are generally represented more than older ages. However, younger

children (2 or younger) share many characteristics, including that they are likely pre-verbal, whereas older children (older than 2) are likely verbal. With this split, it seems like there would be sufficient representation for each age group. Regardless, there are at least 1500 observations for each age out of 19947 total observations, which we find to be sufficient representation.

In the top-right plot, we see that there are about 8000 females and 12000 males in the dataset. While we wouldn't expect different indicators for traumatic brain injury by gender, the dataset is a bit lopsided in this regard; perhaps young boys tend to engage in more dangerous activities than young girls, and thus are at greater risk of head injury. Similarly, in the bottom-right plot we see that the vast majority of observations are White or Black, while other races are quite underrepresented by comparison. Once again, we would not expect a race-dependence of the indicators for TBIs, but like with the observed gender imbalance, we will keep this imbalance in mind as we perform a posthoc analysis of our best model's false negative classifications.

In the bottom left plot, we see that almost 20% of the TBIs observed are in children aged 0, and that the number of TBIs decreases as age increases, flatlining on average for children older than 2. Similarly with the age imbalance, if we were to split by young children versus older children, the number of TBIs in each age group would be more balanced and likely sufficient to generate a classification rule.

Since an imbalance in outcome across ages may skew our age-invariant classifier to detect rules relevant for one age group versus another, we should be careful to only include predictors whose relationships with the outcome are invariant to age. The next few plots provide evidence of age dependence and lack thereof for certain variables, motivating the use of certain features for an age-invariant classifier.

First, Figure 2 shows the distribution of 'verbal' vs. 'non-verbal' or 'pre-verbal' in response to the physician asking the child if they have amnesia or any headache symptoms. We clearly see that, among children aged two or younger, more than 80% of them were marked as non-verbal in response to both questions. So, a rule based on whether or not the individual has a headache or amnesia would be useless for these individuals. However, among children older than aged two, more than 90% of individuals were able to verbalize answers to these questions. These variables would likely be good predictors of a TBI, so for a classifier only on older individuals, these variables would be useful to include. However, they should not be included in any classifier for younger children, or for an age-invariant classifier.

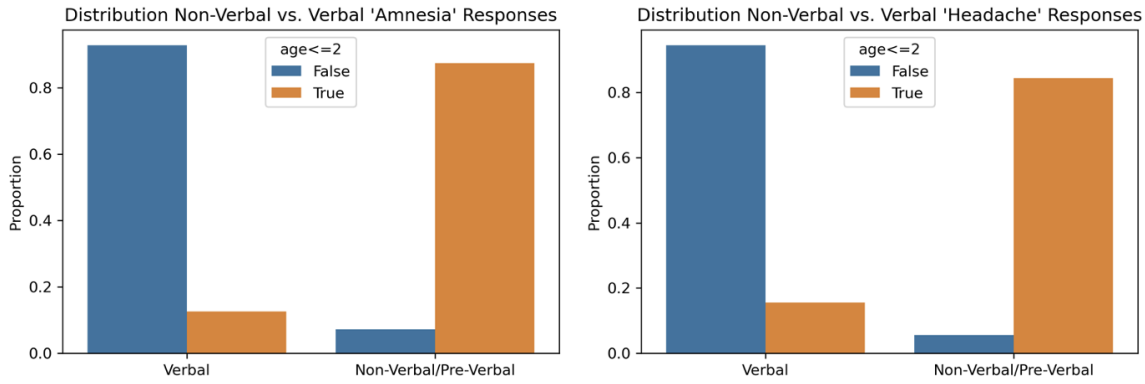

Figure 2: Distributions of 'Verbal' or 'Non-Verbal/Pre-Verbal' responses to questions on amnesia and headache symptoms, by age.

Next, we recall that, according to Kuppermann (2008), for children two or younger, a Pediatric GCS score was assigned, whereas for children older than two, a standard GCS score was assigned. In essence, these are two different standards for evaluating the consciousness of children, since children of different ages have different abilities in expressing their consciousness. However, because the standards are different across these two age groups, we fear that a suboptimal score with regard to one standard might be more indicative of a TBI than a suboptimal score with respect to the other standard. If GCS scores are included in an age-invariant classifier, this presents a potential comparability issue, where the rule that is learned may be applicable to one age-group but not the other. Figure 3 shows the distribution of the

outcome variable for suboptimal versus optimal scores of each GCS category (row) for the different age groups (column).

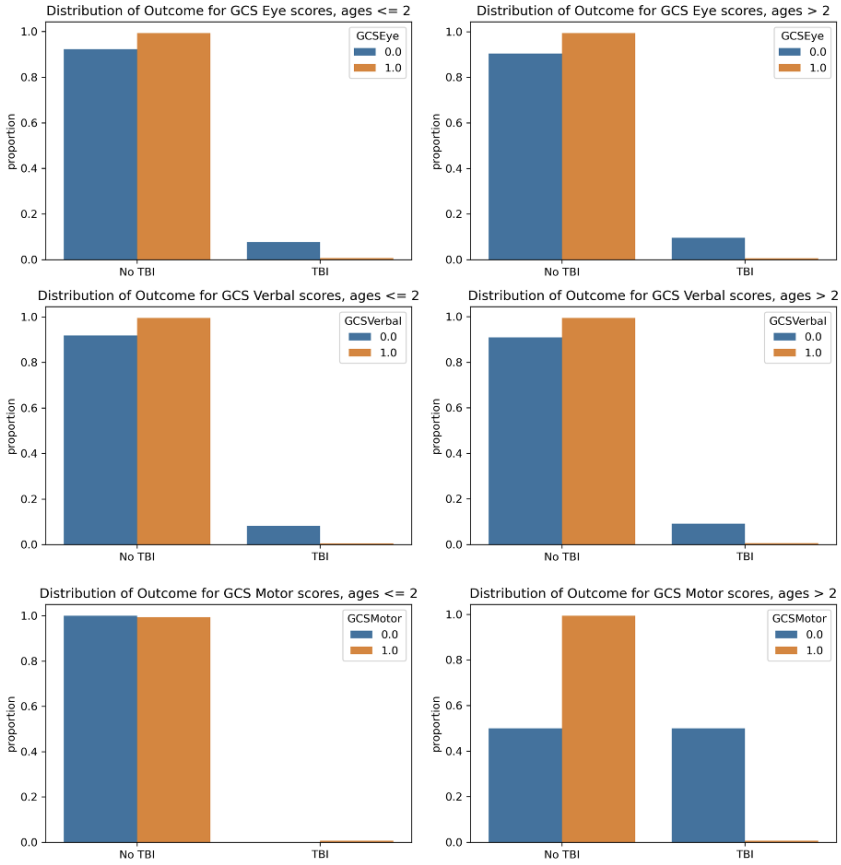

Figure 3: Distributions of outcome among different GCS scores (row) for different age groups (column). 1 indicates optimal score in a GCS category, 0 is one point suboptimal (to achieve total GCS of 14 instead of 15).

In this figure, we see that, for the GCS Eye scores and the GCS Verbal scores (first two rows), the outcome conditional on suboptimal versus optimal scores is distributed similarly across both age groups. This indicates that the relationship between GCS score in these categories and the outcome is age invariant. However, for the GCS Motor scores, we see that the distributions are not the same. In fact, it seems like for individuals older than age 2, half of individuals that had a suboptimal GCS Motor score ended up having a TBI, whereas among children two or younger, a much smaller proportion of those with a suboptimal GCS Motor score had a TBI. This tells us that there is likely some age dependence in how this variable is related to TBI. So, when building an age-invariant classifier we also wish to exclude the GCS scores. However, if we were to build classifiers for different age groups and include these variables, we must split by older than two versus two or younger. Note that Kuppermann instead builds classifiers for children two or older and children younger than two.

Before moving on, we should also establish the age-invariant relationships of other variables of interest with the outcome variable. Figure 4 shows that the distributions of the outcome conditional on having neurological deficits or not, and having basilar skull fracture or not, are not age-dependent. This is as we would expect, since these injuries are observed directly by doctors and are not dependent upon verbal ability or any other factor directly associated with age. Since the relationship of these indicators with the outcome are invariant to age, we should certainly include variables like these in an age-invariant classifier.

Finally, in making a classifier to vet and possibly improve upon that of Kuppermann et al., we explore the potential impact of including and excluding subvariable details. Recall from Cleaning and Preprocessing that there are many categories which can be thought of as parent categories (Altered mental status? Neurological Deficit? Basilar skull fracture?), with children subcategories that provide more detailed information about the parent category (if altered mental status, is it because you are sleepy? Agitated?

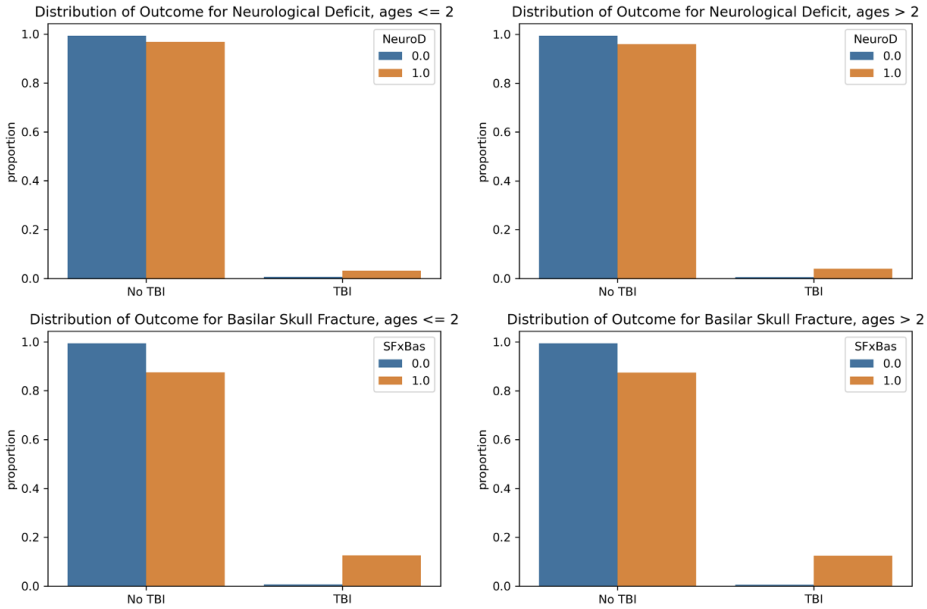

Figure 4: Distributions of outcome conditional on neurological deficit and basilar skull fracture, across both age groups (1 indicates the presence of neurological deficit/basilar skull fracture).

Asking repetitive questions?). In making an interpretable classifier with high predictive accuracy and a low false negative rate with which to compare to Kuppermann's, there are arguments for and against including this subvariable information.

On the one hand, it is possible that certain subvariables are more predictive of the outcome than other subvariables. Consider Figure 5, which shows the correlations of the different subvariables for two parent variables, altered mental status and basilar skull fracture, with the outcome. Each subvariable is binary encoded: if the individual has any of the qualities, it is marked as 1, otherwise, it is marked as 0. For AMS, it seems 'Sleepy,' 'Slow to Respond,' and 'Agitated' are much more correlated with the outcome than 'Repetitive.' Even more so, for basilar skull fracture, it seems 'Hemotympanum' is much more highly correlated with the outcome than the other variables. If we were to exclude this subvariable information in lieu of only including an indicator for basilar skull fracture, it is possible that we could over-indicate TBIs. Imagine if a child came into the emergency room with CSF Rhinorrhea, but our classifier only considers whether there is a sign of skull fracture. Skull fracture, as a parent variable, might be an important predictor, but only because hemotympanum is so correlated with the outcome. Thus, the child would probably be recommended to have a CT scan. However, the child may not have had a TBI in the first place because CSF Rhinorrhea is not very correlated with the outcome. Since we would be giving a child an unneeded CT scan, it could be important to include this subvariable information in an ideal dataset.

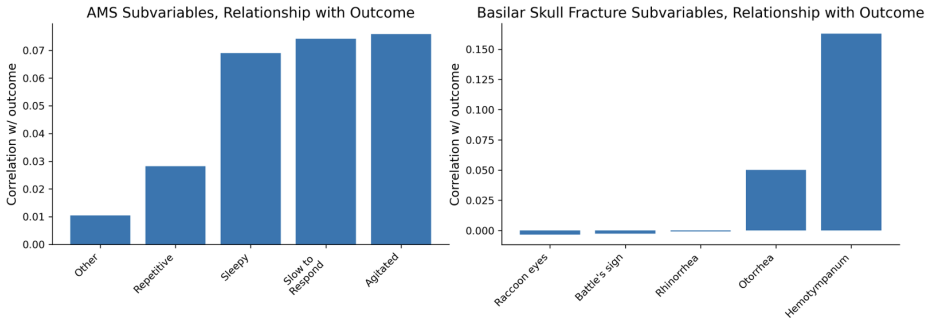

Figure 5: Subvariables of AMS (altered mental status) and basilar skull fracture are correlated differently with the outcome.

However, a clear argument against including this subvariable information is that including too many variables may lead to very uninterpretable and nonsensical rules that are the result of overfitting. Additionally, it seems like the parent variables alone are substantially correlated with the outcome, as is shown in Figures 6 and 7. So, in order to achieve a suitable dataset to make interpretable and accurate classifiers with which to vet and improve upon Kuppermann's, it seems like it may be suitable to just include the parent variables. Note also that the following two plots further show the dataset is suitable for making an age-invariant classifier: among the age-invariant variables, the correlations with the outcome are almost identical across age groups.

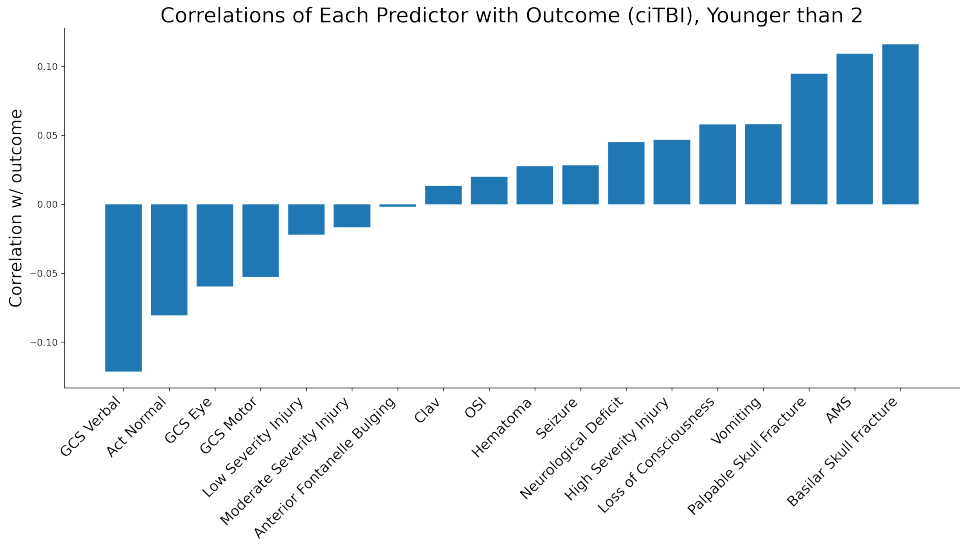

Figure 6: Among children 2 and younger, relevant parent variables seem substantially correlated with the outcome.

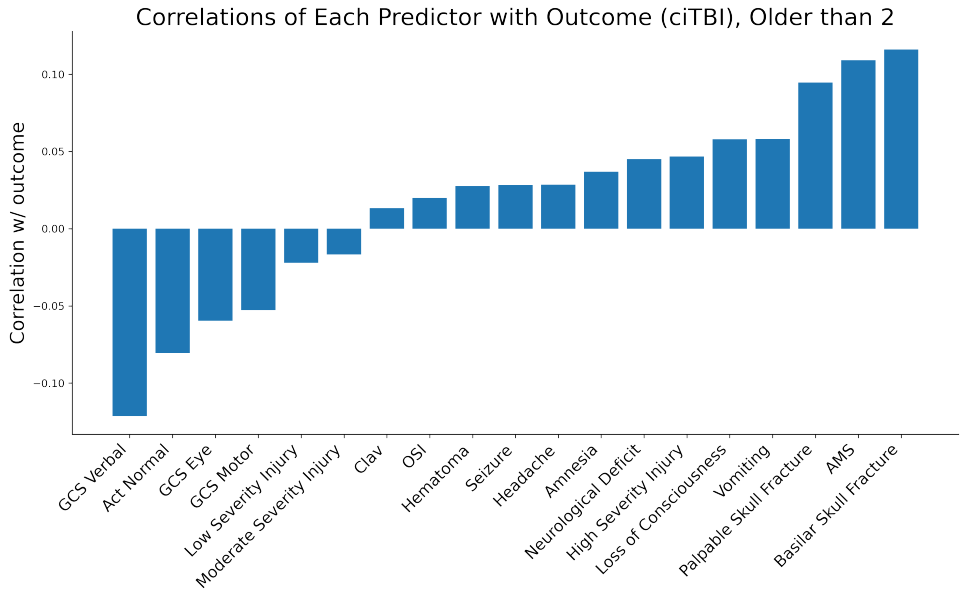

Figure 7: Among children older than 2, relevant parent variables seem substantially correlated with the outcome.

Models

Baseline Classifier

We use the model proposed by Kuppermann and his colleagues as the baseline classifier. To evaluate their model, we use the provided dataset cleaned according to their specifications. In total, the dataset consists of 42412 children presenting to the emergency room due to the head trauma with GCS scores 14 and above and with the final outcome non-missing. 376 of these observations experienced clinically-important TBIs. In addition, we use the same age splitting strategy as Kuppermann, where we use their baseline classifier for preverbal children (< 2 years of age) to predict TBI in this subgroup, and the classifier for verbal children (2 years and older) to predict TBI in this subgroup.

Description

The age-based classifiers in Kuppermann et al. are rule-list models, or models consisting of a list of rules ordered by importance in indicating a TBI. A child is labelled as having a TBI if and only if this child satisfies at least one of the rules, where the rules are checked in the aforementioned order of importance. Once we come across a rule that the child satisfies, we stop checking. Therefore, based on the order, each rule is applicable to different numbers of children. The list below gives the rules for Kuppermann's pre-verbal (younger than 2) classifier. Refer to Data Collection for a more in-depth description of the variables mentioned in each rule.

- Altered mental status (*AMS*)? If yes, TBI. If not, continue:
- Occipital, Parietal, or Temporal scalp hematoma (*HemaLoc*)? If yes, TBI. If not, continue:
- Loss of consciousness for more than 5 seconds (*LOCLen*)? If yes, TBI. If not, continue:
- Injury mechanism rated severe (*High_impact_InjSev*)? If yes, TBI. If not, continue:
- Palpable or unclear skull fracture (*SFxPalp*)? If yes, TBI. If not, continue:
- Acting normally according to parent (*ActNorm*)? If yes, TBI. If not, no TBI.

The list below now gives the rules for Kuppermann's verbal (2 or older) classifier.

- Altered mental status (*AMS*)? If yes, TBI. If not, continue:
- Loss of consciousness for more than 5 seconds (*LOCLen*)? If yes, TBI. If not, continue:
- Injury mechanism rated severe (*High_impact_InjSev*)? If yes, TBI. If not, continue:
- Basilar skull fracture (*SFxBas*)? If yes, TBI. If not, continue:
- Severe headache (*HA_Severity*)? If yes, TBI. If not, no TBI.

Procedure and Performance

The result of baseline model is shown in 8. The yellow blocks are identified as positive while the green blocks are identified as negative. In each block, the first number represents the number of children experiencing clinically-important TBI, the second number is the number of children in this block for whom the next rule will apply. We use the whole cleaned dataset to evaluate the model performance, while Kuppermann splits the data into a training set and validation set. The performance of the baseline classifiers is similar to the result in Kuppermann (2008). The baseline classifiers do identify almost all of the cases as the sensitivity is 98.9% and 96.8% for preverbal children and verbal children, respectively. However, the corresponding specificities are somewhat disappointing: 48.3% and 57.7%. In the following section, we propose a model that, run on our preprocessed data, achieves a similar sensitivity and higher specificity.

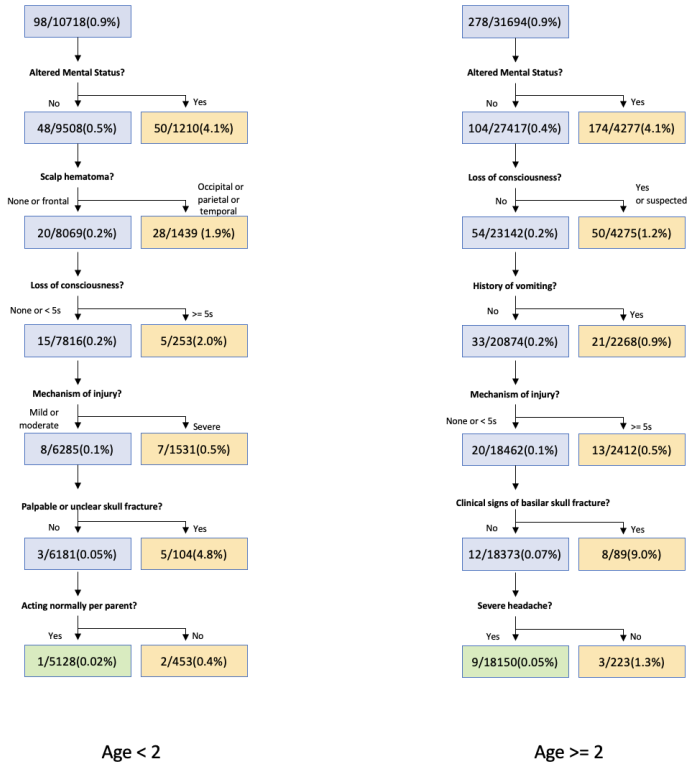

Figure 8: Decision rule lists for clinically-important traumatic brain injury in children younger than 2 years and in those aged 2 years and older.

Proposed Model

`Rulefit`: Introduction and Basic Illustrative Implementation

`Rulefit` was proposed by Friedman and Popescu (2008) and is an example of a wider category of *ensemble methods*. At a high level, we wish to describe our output function with a linear combination of simpler learners $\mathcal{T}_k$ (Hastie, Friedman, and Tisbshirani 2017):

$$f(x) = \sum_{k=1}^{K} \alpha_k \ \mathcal{T}_k(x) \tag{1}$$

The $\mathcal{T}$'s need not necessarily be analytic functions, so a common choice is an ensemble of *trees*. Suppose we generated a large number of such trees in the first step, e.g. via a boosting procedure. We then wish to make the linear combination as concise as possible. A usual procedure is to perform a *LASSO* regression of regularization strength λ on the space of $\mathcal{T}$s for the coefficients α_k:

$$\alpha(\lambda) = \underset{\vec{\alpha}}{argmin} \ \sum_{i=1}^{N} \mathcal{L}(y_i, \vec{\alpha_0} + \sum_{k=1}^{N} \alpha_k \ \mathcal{T}_k(x)) \ + \ \lambda \sum_{k=1}^{N} |\alpha_k| \tag{2}$$

We also added a linear term in all the original features $\vec{\alpha_0}$, which will come into play especially with the more complex models. The *LASSO* will for sufficiently large λ set a large portion of the coefficients to zero.

`Rulefit` follows the above procedure with an additional twist. After fitting the trees via a gradient-boosting algorithm, it extracts *rules* from the generated tree. Looking at a tree as a *trie*, each path down a tree encodes a specific combined rule (each node is split solely on a single variable). Naturally, the deeper the tree, the more complex a rule can get. It can be shown (Hastie, Friedman, and Tisbshirani 2017) that only the paths ending in leaf nodes describe all the rules for the whole tree.

Thus the input to the sum above are the sets of such rules from each tree in the original gradient-boosting sequence, which are most often overlapping.

The *LASSO* thus performs regression on the *space of rules* and outputs a parsimonious set with potentially added linear terms.

As a concrete illustration of `Rulefit` using our preprocessing and judgment call procedure, we first present the following two models. Note that these are solely illustrative: our best model presentation comes later.

- **"basic":** The judgement calls are selected as to be maximally restrictive, combining all sub-variables into their parent variable and throwing away as many of the rest as possible, as well as opting to remove observations whenever possible. The number of rules is restricted to 10.
- **"full":** The inverse of the above, including as many variables and throwing away as few observations as possible. The number of rules is allowed to go to 20.

These two classifiers should, therefore, illustrate the effects of our judgement calls on the structure of the model.

Below are presented the top five rules for each in terms of their importance. The latter is similar to the metric in the ensemble of trees, but calculated from the *LASSO* re-fitting (Molnar 2019):

Table 1: Top 5 basic model rules in order of importance

	rule	type	coef	support	importance
2	Seiz <= 0.5 and AMS <= 0.5	rule	-0.83	0.83	0.31
6	Vomit <= 0.5	rule	-1.03	0.93	0.26
9	GCSTotal > 0.5 and SFxPalp <= 0.5 and SFxBas <= 0.5	rule	-0.83	0.93	0.22
1	LOCSeparate <= 0.5	rule	-0.41	0.70	0.19
5	Seiz <= 0.5 and Hema <= 0.5	rule	-0.23	0.68	0.11

Table 2: Top 5 full model rules in order of importance

	rule	type	coef	sup	imp
176	SFxBasRet_0 <= 0.5 and NeuroDCranial_1 <= 0.5 and OSIPelvis_1 <= 0.5 and SeizLen_2 <= 0.5 and VomitNbr_2 <= 0.5 and AMSAgitated_0 <= 0.5 and AMSAgitated_1 <= 0.5 and AMSOth_1 <= 0.5 and SFxPalpDepress_1 <= 0.5	rule	-1.21	0.88	0.40
1	ActNorm	linear	-0.76	1.00	0.28
180	SeizLen_2 <= 0.5 and SFxPalp_0 > 0.5 and SFxBasHem_1 <= 0.5	rule	-1.22	0.97	0.20
175	ClavPar_1 <= 0.5 and SeizLen_3 <= 0.5 and AMSSlow_1 <= 0.5 and SFxPalpDepress_92 > 0.5	rule	-0.63	0.90	0.19
181	HemaSize_3 <= 0.5 and SFxBasHem_1 <= 0.5	rule	-0.71	0.93	0.18

We can immediately sense the increase in complexity for the *full* model. Also note that the *LASSO*-introduced linear terms become important in the complex model, perhaps substituting some of the effect of simpler rules.

To see how these relatively basic models perform, we examine their *sensitivity* - *specificity* curve in Figure 9 with labeling 'Intro Minimal' and 'Intro Full.' The minimal model is able to achieve 80%+ in both sensitivity and specificity for one threshold value; since we care much more about false negatives than positives, we consequently value sensitivity far above specificity. Therefore in this case, the fuller model with sensitivities above 90% would be preferred.

Comparing these basic rules with the performance of the final model in Kuppermann (2008), which

obtained approximately 96% sensitivity with 58% specificity, we can surmise that in the case of the full model we are approaching that sensitivity and exceeding it in specificity, without any specific tweaking to the models being performed at this stage!

As shown in the next section, with some careful construction of the models and selection of judgement calls, we can improve this performance considerably.

Procedure and Performance

As discussed in Exploratory Data Analysis, there are arguments for-and-against making a single age-invariant classifier on the whole dataset vs. classifiers for different age groups, and making classifiers including and excluding subvariable information. Taking the cartesian product of these possible decisions, there are 6 possible datasets on which to develop classifiers (children <= 2 with/without subvariables, children > 2 with/without subvariables, all children with/without subvariables). The datasets consist of the data preprocessed according to our default judgement calls (outlined in Cleaning and Preprocessing). For the younger age-split datasets, we exclude variables relevant to headaches or amnesia, per our EDA. For the older age-split datasets, we include these variables, but exclude the variable for anterior fontanelle bulging, per our Cleaning and Preprocessing justification. For the age-invariant dataset with all children, we exclude all of these variables and the GCS scores altogether.

To find a best classifier, we first determine on which of these datasets does our `Rulefit` model perform the best. It should be noted the dataset of children two and younger without subvariables contains too few TBIs, so we cannot get a classification result. Therefore, we will train and evaluate for the other five datasets.

For each possible dataset, we split 20% of the data into test data, which will only be used in the model evaluation, and use the remaining 80% to fit the model and tune hyperparameters via a four-fold cross-validation. Since this is an unbalanced dataset, raw classification accuracy will be a poor means of evaluating the model, as a model that simply predicts that no children have a TBI would be more than 99% accurate. A good classifier should first and foremost identify almost all of TBI cases, but also correctly identify most of the children without TBIs. According to these specifications, we define our evaluation measure as follows.

$$\text{score} = \begin{cases} 5 \times \text{sensitivity} + \text{specificity}, & \text{if specificity} > 0.5. \\ 0, & \text{if specificity} \leq 0.5. \end{cases}$$

For the `Rulefit` model, the hyperparameter is the maximum length for each rule. In addition, since our `Rulefit` model involves subsampling in generating the rules, the performance may vary even if we choose same hyperparameters. Therefore, for our `Rulefit` model, we also need to run the models with different random states and pick the best one as the result, given the maximum length for each rule. Based on the cross-validation results, we pick the best hyperparameters and evaluate the performance on our test data.

The test-data performance of our `Rulefit` model for the different datasets can be found in Figure 9. We find that the classifier on the age-invariant dataset with subvariables is quite good, as the sensitivity and specificity can simultaneously reach approximately 100% and 70%, respectively. When classifying on verbal children, including the subvariables increases the model's performance relative to excluding them. For instance, for the two verbal children datasets, when the sensitivity is larger than 95%, the specificity is about 70% when we include subvariables, while it drops to 55% if we exclude them. Finally, for preverbal children, the specificity is around 60% when the sensitivity is about 95%. These results suggest we develop an age-invariant classifier with subvariables as opposed to developing separate classifiers based on age group. For this classifier, based on our discussion of sensitivity and stability, the threshold is set to 1.3×10^{-5}, which performs well not only on training and validation data, but also on test data. For test data classification with this threshold, we achieve a sensitivity of 97.9% and a specificity of 70.4%.

In Figure 9, we list the important rules for the age-invariant classifier. Note that each categorical variable that was not already binary has been binarized in Figure 9 so each variable is an indicator, and the underscore after the variable name denotes which choice in the original categorical variable is new binary variable is an indicator for. For instance, `AMSOth_0` is an indicator for where `0` was marked in the `AMSOth` subvariable for `AMS` (altered mental status). Please refer to our data dictionary readme documentation for more specifics on what each numbered category represents.

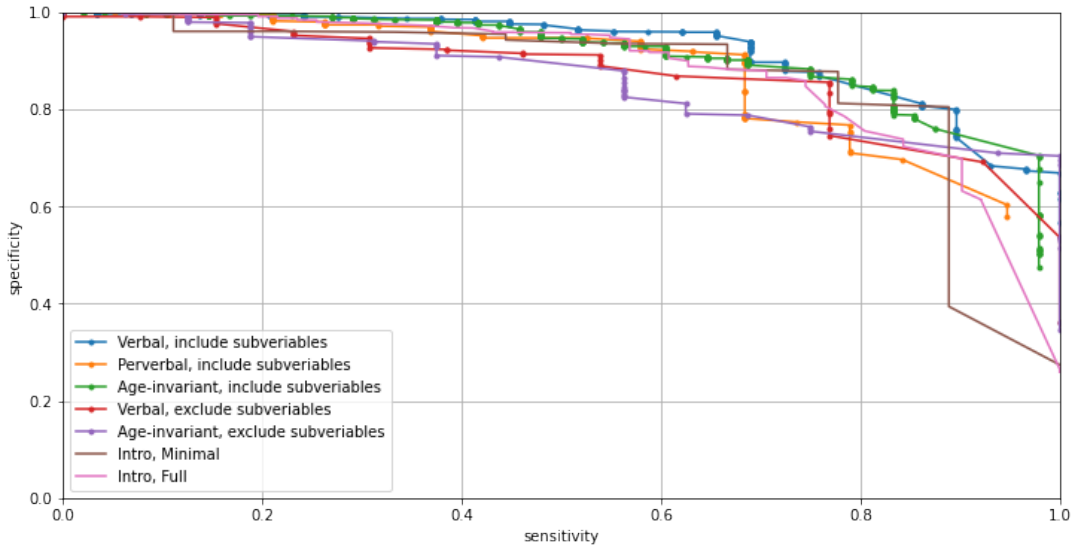

Figure 9: ROC curves of fine-tuned Rulefit models on the test data for all five datasets.

	rule	coef	support	importance
127	AMSOth_0 <= 0.5	-0.993656	0.914574	0.277742
124	Vomit <= 0.5 and SFxBasHem_1 <= 0.5	-0.601918	0.873966	0.199769
123	High_impact_InjSev_3 <= 0.5 and LocLen_4 <= 0.5 and SFxPalpDepress_0 <= 0.5 and NeuroDCranial_1 <= 0.5	-0.526590	0.855668	0.185058
121	OSIAbdomen_1 <= 0.5 and HemaLoc_3 <= 0.5	-0.558559	0.897779	0.169209
119	AMSOth_1 <= 0.5 and SFxBasHem_1 <= 0.5 and HemaSize_3 <= 0.5	-0.541056	0.915727	0.150304
18	LocLen_92	-0.391449	1.000000	0.117573
2	ActNorm	-0.296187	1.000000	0.098651
38	AMSOth_92	-0.219569	1.000000	0.065860
125	SFxPalpDepress_92 > 0.5 and SFxBasPer_0 <= 0.5	-0.641207	0.992781	0.054283
129	OSIOth_1 <= 0.5 and SeizLen_2 <= 0.5 and SFxPalpDepress_92 > 0.5 and SFxBasHem_1 <= 0.5 and SFxBasOto_1 <= 0.5	-0.230229	0.974683	0.036166
131	SFxPalpDepress_1 <= 0.5 and SFxBasHem_1 <= 0.5 and NeuroDCranial_1 <= 0.5	-0.314930	0.993483	0.025341
130	ClavPar_1 <= 0.5	-0.061043	0.926054	0.015974
126	OSICut_1 <= 0.5 and SFxPalp <= 0.5	-0.101669	0.978242	0.014833
128	OSICut_1 <= 0.5 and LocLen_4 <= 0.5 and SeizLen_4 <= 0.5 and SFxBasHem_1 <= 0.5 and NeuroDMotor_1 <= 0.5	-0.120065	0.989372	0.012312
122	SeizLen_1 <= 0.5 and SFxPalp <= 0.5	-0.075998	0.973129	0.012289
120	AMSSlow_1 <= 0.5 and SFxPalpDepress_92 > 0.5	-0.043991	0.976989	0.006596

Figure 10: The important rules for the age-invariant classifier

We find that other signs of altered mental status is the most important linear factor. Other important linear factors include how long there was loss of consciousness, as well as any history of vomiting. In addition to our linear predictors, we have numerous binary rules that appear in our classifier, the most important of which are history of vomiting and basilar skull fracture (specific sign: hemotympanum). Other important rules involve the severity of injury, loss of consciousness length, a depression in a palpable skull fracture, and neurological deficit in the cranial nerve. Note that the above coefficients are all negative, meaning that if a criterion is satisfied, then this is an indication that the individual may not have a TBI.

In addition, we also tried another candidate classifier, that is, the greedy rule list model, whose structure is similar to the baseline model and consists of a list of rules. We performed same cross-validation procedure to tune its hyperparameters and test on the withheld data. But it performed much worse than our `Rulefit` model.

Randomness Assessment

The randomness in the data can be divided into two categories: *intrinsic* and *measurement-related*.

Under intrinsic randomness we can include the natural variation in people's physiologies, where the same kind of injury might lead to a TBI in one child, but not in another, as well as the minute differences in the injury mechanics beyond the broader categories evaluated that might spare one child and not another from developing a TBI. Even if our rules somehow covered all the potential avenues to *ciTBI*, these would, together with potential additional confounding factors, still make the prediction a probabilistic exercise, not a deterministic one.

To the above, we can add the many sources of *measurement-related* randomness: many of the variables are subjective evaluations performed by the clinical personnel, the parents, or the children themselves (verbal indication of a headache, for example). These variables containing a degree of random "noise" should be obvious. As Dr. Inglis pointed out, however, an experienced physician might also *encode* a degree of domain knowledge - colloquially, "gut feel" - in some of these categories, making them a stronger instead of a weaker predictor.

A potential trap in understanding the model is to see it as a causal "if-then" model, such as the baseline one. As shown in `Rulefit`: Introduction and Basic Illustrative Implementation, the rules instead act in an additive fashion, each adding a degree of certainty in the outcome. Additionally, the model itself is random. While the *LASSO* procedure in step 2 is deterministic, it can only select rules from the initial bag given.

The latter is generated via a tree generation procedure, which is most commonly *not deterministic* (gradient boosting, for example, performs subsampling). This randomness in rule generation is somewhat ameliorated by the large number of trees we are generating, which should work towards obtaining the set of all feasible rules.

There is also a kind of arbitrariness in using the L_1 penalty, i.e. *LASSO*, as opposed to others in the feature selection step, which is going to privilege some rules above the others without a domain-based reason.

However, we can argue that physicians themselves follow an analogous heuristic when evaluating alternative hypotheses for a diagnosis: since the realm of possible causes is usually exceedingly large, they restrict themselves to a set of plausible causes using a "gut-feeling" that could be likened to the L_1 shrinkage employed by the *LASSO*.

Stability Analysis

In our stability analysis, we focus our efforts on checking both the stability of the isolated features subject to data and model perturbations, as well as the stability of our classification results subject to data perturbations of the test set.

Feature Stability

Kuppermann Comparison To address the former, we can first compare the isolated features with those of the classifier from Kuppermann, whose results are presented in Baseline Model. We cleaned our

data differently, so this is a stability assessment of the rules subject to data perturbations. Additionally, we use different classifier methods, giving us a sense of stability subject to model perturbations. Note that there is significant overlap between the rules of Kupperman's two age-dependent classifiers, and most of the isolated rules are for variables we consider to be age-invariant. So, comparisons between these two models will be meaningful.

Looking at the first few variables isolated in Figure 10, we see that the individual having altered mental status, a history of vomiting, hemotympanum (sign of basilar skull fracture), an injury rated 'severe,' loss of consciousness for more than 5 seconds, and neurological deficits related to the cranial nerve, and a parietal or temporal scalp hematoma are of utmost importance in determining if someone has a clinically-important TBI. Compared to the baseline classifier's variables in Figure 8, we see that all of these variables are included in their rules to some extent. In fact, sometimes our classifier isolates a more specific version of a parent variable. For instance, while the baseline classifier isolates general signs of basilar skull fracture, we isolate a particular sign (hemotympanum), and while the baseline simply checks for parietal, temporal, or occipital scalp hematoma, we find that only parietal and temporal hematomas are of most use.

Random Forest Comparison The next model to which we will compare our best classifier is an age-invariant random forest classifier. For our random forest, we exclude the parent variables entirely and impute missing values into the subvariables as 'False,' according to the alternative feature extraction scheme discussed in Data Summary. Comparisons to these results thus also represent a data perturbation. To create our cross-validation structure, we withheld a random sample of 4739 observations for validation, stratifying on outcome level. From there, the remaining 30,000 observations were randomly split into three equal folds, likewise stratified, and used to run a three-fold CV in order to evaluate a given model's sensitivity and specificity, averaged over the three possible train-test splits.

With the folds set up, we use this average sensitivity and specificity to tune our model's hyperparameters. One possible hyperparameter, number of trees, was found during exploratory analysis to have functionally no effect on anything except computation time beyond around 200 trees, so all models were run with 200 trees. The forest's maximum depth, on the other hand, was optimized over, as an increase in max depth would generally result in an increase in specificity at the expense of sensitivity. The unbalanced nature of our output (less than 1% of children actually had a TBI) leads us to another hyperparameter: weight. Essentially, if we do not weight our model in order to favor true positives over true negatives, any random forest will always predict that the child does not have a TBI; thus we add a weight factor to our "1" outcome in order to increase our model's sensitivity.

The values evaluated for our hyperparameters (chosen via cursory exploratory analysis) include a max depth ranging from 5 to 11 and a weight parameter evaluated at 1 and then every multiple of 100, up to 1000. We optimized over both of these parameters, testing a total of 77 models to find the ideal model that maximizing our aforementioned evaluation meausure for our best classifier. We also generated ROC curves for each max depth.

Our final random forest model ended up having depth 8 and weight 400, and when applied to the validation dataset, it demonstrated an impressive sensitivity of 0.971 and specificity of 0.617- slightly worse than our best classifier, but better than Kuppermann's classifier and overall good given our aforementioned preference for sensitivity over specificity. In order of importance, we isolate the following features:

- *SFxBasHem*: hemotympanum as an indicator for basilar skull fracture
- *AMSSlow*: whether the child was slow to respond
- *ActNorm*: whether the parent thinks the child is acting normally
- *AMSSleep*: whether the child is sleepy
- *ClavPar*: whether the child sustained an parietal scalp injury

Somewhat curiously, no predictors relevant to vomiting were found in the top six most important predictors, despite its high prevalence in our rule-fitting classifier. This may be a consequence of the flattening of the vomit variable, leaving us with many minor effects instead of the one overall binary indicator of whether or not the child was vomiting. However, we notice that variations of the above variables are found among the top features in our best `Rulefit` model.

Classification Result Stability To check the classification result stability, we run through a cartesian product of our possible perturbations for our judgement calls and, using our best `Rulefit` classifier,

predict on this perturbed test data for each combination of judgement calls. Because many of our judgement calls had to do with the inclusion/exclusion of certain sets of columns, for which it would be impossible to evaluate a model trained on a different feature space, we choose a set of 5 judgement calls, along with their perturbations, that involve imputing/dropping missing values and recoding categories. Please refer to Data Collection for a dictionary of the judgement calls and their perturbations.

- original judgement call: [`SFx`]; perturbation 2
- original judgement call: [`Seiz`]; perturbation 2
- original judgement call: [`LOC`]; perturbation 2
- original judgement call: [`ActNorm`]; perturbation (only perturbation listed)
- original judgement call: [`Vomit`]; perturbation (only perturbation listed)

This set of judgement calls effectively covers perturbations due to the inclusion or exclusion of missing values, imputation of missing values, as well as recoding observations. For instance, [`SFx`] and [`LOC`] are both judgement calls where a result marked 'unclear' was coded as 'True.' An interesting perturbation would be to see if coding these observations as 'False' changes the results substantially.

We loop through every possible combination of the above 5 judgement calls and their perturbations and obtain a test dataset (32 test datasets in total). For each of those sets, we evaluate the sensitivity and specificity using our best `Rulefit` model with the corresponding threshold. A histogram of specificities and sensitivities are provided in Figure 11

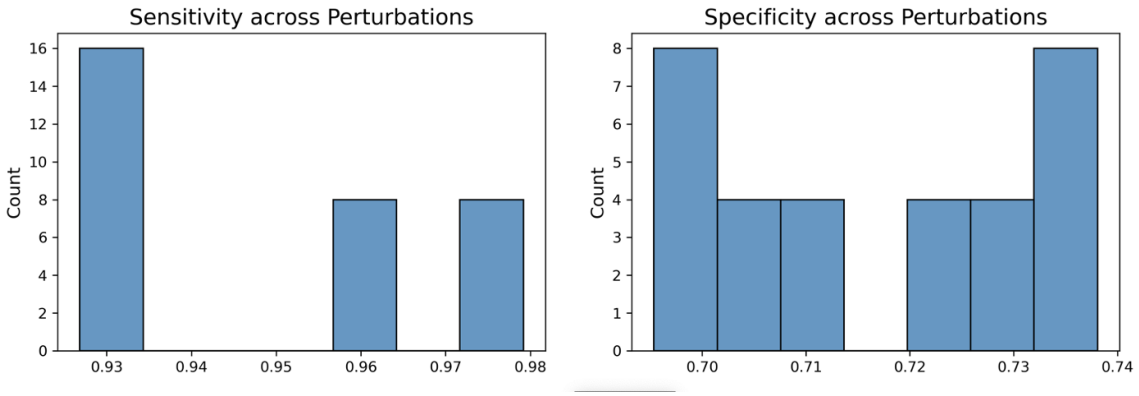

Figure 11: The sensitivity and specificity are fairly stable, with specificity being slightly more so than sensitivity.

In Figure 11, we see that, subject to these different combinations of judgement calls and their perturbations, the sensitivities and specificities are quite stable. We note that, in a large number of the judgement call combinations, the sensitivity drops to around 0.92 to 0.93. Upon further examination, it seems this sensitivity occurs when the perturbation for [`Vomit`] is applied. Instead of imputing missing values from the subvariable columns as in the [`Vomit`] judgement call, the perturbation drops any observation for which any of the subvariables or parent variables are missing, the reason being that if any of the criteria is missing, any provided results cannot be trusted. However, a counterargument would be that the physician could have been tired and possibly skipped through some of the subcategories, or felt that marking the main vomit category after having marked a subcategory was redundant. So, while our results are slightly impacted by this perturbation, we should take this with a grain of salt.

On the testing set, our best classifier only had one false negative: patient number 5295, a four year-old Black male. Upon examining his covariates, it seems the only abnormal responses were the presence of a large frontal hematoma, an above-the-clavicle injury on the frontal part of the scalp, an injury mechanism of moderate severity. Every other covariate was listed as normal. With such a patient, Kuppermann's classifier would also have listed him as not having a TBI. So, while it is unfortunate that we would have misclassified this individual and decided not to pursue a CT scan, the classifier we wish to improve upon would have also missed this diagnosis. This may indicate the need for new features to observe when collecting data to ensure no TBIs are missed in the future.

Discussion

Our results indicate that a `Rulefit` classifier on the age-pooled data with age-invariant columns and subvariable information performs the best in terms of sensitivity and specificity, and even outperforms both age-specific classifiers in Kuppermann (2008). Examining Figure 10 in more depth, we see altered mental status, whether or not the child vomited, hemotympanum as a sign of basilar skull fracture, severity of injury, loss of consciousness for more than 5 seconds, and neurological deficit of the cranial nerve are of vital importance when classifying the clinically-important TBI. As mentioned in stability section, the importance of these features largely coincides with that of a random forest, as well as the baseline classifiers from Kuppermann (2008), so we find the results stable subject to model and data perturbations.

Given the results and their stability, we find that our isolated features are possible age-invariant indicators of TBI in children. Moreover, in many instances, our classifier isolated particular subvariables of parent variables as being important in predicting TBI. For instance, hemotympanum specifically, as opposed to general signs of basilar skull fracture, is of great importantance. This also matches our exploratory results in Figure 5, where we found hemotympanum was much more highly correlated with the outcome than the other indicators of basilar skull fracture in the training dataset. This work supports the idea that including more detailed information about the parent variables could lead to more accurate diagnoses.

For future work, we could fit more interpretable models. Our best classifier is a `Rulefit` model, which generally provides less interpretability than the rule list models in Kuppermann (2008). Although we choose one kind of rule list model, that is, the greedy rule list model,as a candidate classifier, it does not perform well. It had lower a sensitivity and was not stable to data perturbations. In the future, we can try other types of rule list models to see whether we could obtain a stable model with high sensitivity and specificity, as well as more interpretability. Additionally, we may apply other combinations of judgement calls to fit the models, such as checking the importance of including and excluding individual subvariable groups for a particular parent variable.

Conclusion

In this paper, we utilize the same data as Kuppermann (2008) (subject to cleaning perturbations) to develop classifiers to identify TBIs in children presenting to the emergency room for head trauma. We train and evaluate `Rulefit` models both for the age-split data as well as data that contains and excludes subvariable information, finding that a `Rulefit` model on the age-pooled data (excluding age-dependent covariates) with subvariable information performs the best in terms of sensitivity and specificity, even outperforming each of the age-specific classifiers in Kuppermann (2008). In comparing our results to those in Kuppermann (2008) as well as the results of random forests, we find the corresponding important features selected by our model (altered mental status, the history of vomiting, hemotympanum as an indicator of basilar skull fracture, severity of injury, loss of consciousness for more than 5 seconds, and cranial nerve neurological deficits, etc.) to be stable subject to modelling and data perturbations. The sensitivity and the specificity is also quite stable to test dataset perturbations. Given the performance of our best classifier on a single dataset, we hope these results offer a meaningful improvement of those from Kuppermann (2008).

References

Special thanks to Dr. Rob Inglis of UCSF for graciously offering his time and expertise to discuss and interpret our findings!

Faul, et al., M. 2010. "Traumatic Brain Injury in the United States: Emergency Department Visits, Hospitalizations and Deaths 2002–2006." *Centers for Disease Control and Prevention.*

Friedman, Jerome H., and Bogdan E. Popescu. 2008. "Predictive Learning via Rule Ensembles." *The Annals of Applied Statistics* 2 (3). https://doi.org/10.1214/07-aoas148.

Hastie, Trevor, Jerome Friedman, and Robert Tisbshirani. 2017. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Kuppermann, et al., N. 2008. "Identification of Children at Very Low Risk of Clinically-Important Brain Injuries After Head Trauma: A Prospective Cohort Study." *The Lancet* 374: 1160–70. https://doi.org/10.1016/S0140-6736(09)61558-0.

Molnar, Christoph. 2019. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.*