Predicting the need for CT imaging in children with minor head injury using an ensemble of Naive Bayes classifiers

William Klement[a,b], Szymon Wilk[b,c], Wojtek Michalowski[b],
Ken J. Farion[b,d,e], Martin H. Osmond[d, e], Vedat Verter[f]

[a] Ottawa-Carleton School of Computer Science, University of Ottawa
800 King Edward Ave., Ottawa, Ontario, K1N 6N5 Canada.

[b] MET Research Group, Telfer School of Management, University of Ottawa
55 Laurier Ave. E., Ottawa, Ontario, K1N 6N5 Canada.

[c] Institute of Computing Science, Poznan University of Technology
ul. Piotrowo 2, 60-965 Poznan, Poland.

[d] Division of Emergency Medicine, Children's Hospital of Eastern Ontario,
[e] Department of Pediatrics, University of Ottawa
401 Smyth Rd., Ottawa, Ontario, K1H 8L1 Canada.

[f] Desautels Faculty of Management, McGill University,
001 Sherbrooke St. W. Montreal, Quebec, H3A 1G5 Canada.

Present address for corresponding author:
Dr. William Klement
MET Research Group
Telfer School of Management
University of Ottawa
55 Laurier Ave. E.
Ottawa, Ontario K1N 6N5
Canada
Email: william.klement@gmail.com

Abstract

Objective

Using an automatic data-driven approach, this paper develops a prediction model that achieves more balanced performance (in terms of sensitivity and specificity) than the Canadian Assessment of Tomography for Childhood Head Injury (CATCH) rule, when predicting the need for computed tomography (CT) imaging of children after a minor head injury.

Methods and Material

CT is widely considered an effective tool for evaluating patients with minor head trauma who have potentially suffered serious intracranial injury. However, its use poses possible harmful effects, particularly for children, due to exposure to radiation. Safety concerns, along with issues of cost and practice variability, have led to calls for the development of effective methods to decide when CT imaging is needed. Clinical decision

rules represent such methods and are normally derived from the analysis of large prospectively collected patient data sets. The CATCH rule was created by a group of Canadian pediatric emergency physicians to support the decision of referring children with minor head injury to CT imaging. The goal of the CATCH rule was to maximize the sensitivity of predictions of potential intracranial lesion while keeping specificity at a reasonable level. After extensive analysis of the CATCH data set, characterized by severe class imbalance, and after a thorough evaluation of several data mining methods, we derived an ensemble of multiple Naive Bayes classifiers as the prediction model for CT imaging decisions.

Results

In the first phase of the experiment we compared the proposed ensemble model to other ensemble models employing rule-, tree- and instance-based member classifiers. Our prediction model demonstrated the best performance in terms of AUC, G-mean and sensitivity measures. In the second phase, using a bootstrapping experiment similar to that reported by the CATCH investigators, we showed that the proposed ensemble model achieved a more balanced predictive performance than the CATCH rule with an average sensitivity of 82.8% and an average specificity of 74.4% (vs. 98.1% and 50.0% for the CATCH rule respectively).

Conclusion

Automatically derived prediction models cannot replace a physician's acumen. However, they help establish reference performance indicators for the purpose of developing clinical decision rules so the trade-off between prediction sensitivity and specificity is better understood.

Keywords: ensemble model; Naive Bayes; class imbalance; clinical decision rule; pediatric head injury; computed tomography.

1 Introduction

Computed tomography (CT) is widely accepted as an effective diagnostic modality to detect rare but clinically significant intracranial injuries in patients suffering minor head injury. As such, it has been increasingly utilized as a routine test for these patients [1]. However, a seminal study by Brenner and Hall [2] warns against its harmful effects (particularly for children) due to the radiation exposure associated with CT. Independent CT imaging studies [1, 3, 4] advocate the adoption of a comprehensive approach that targets physicians' education to reduce the over-reliance on CT imaging for head injury patients.

1.1 The CATCH study

The diagnosis of a potentially serious brain injury following a minor head trauma is a well-documented challenge [5]. It is believed that clinical decision rules could help with this challenge and reduce unnecessary CT imaging. Broder [4] recommends that such decision rules rely on readily available patient data including physical examination and a patient's

history. In line with these recommendations and in response to a growing need to improve the management of pediatric patients with minor head trauma in the emergency department (ED), Osmond et al. [6] developed the Canadian Assessment of Tomography for Childhood Head Injury (CATCH) clinical decision rule. The prospective cohort study was conducted in ten Canadian pediatric teaching hospitals and enrolled children brought to the ED who had blunt head trauma characterized by loss of consciousness, amnesia, disorientation or repeated vomiting along with a score of at least 13 on the Glasgow Coma Scale. Such patients have often, but not always in a consistent manner [7], been referred to CT imaging to rule out a potential intracranial lesion that might necessitate a neurologic intervention.

The CATCH data set contains 3866 patient records described by 26 clinical attributes (standardized clinical findings from a patient's medical history, general examination and neurological status). These patient records are partitioned according to two classification schemes; the primary classification distinguishes between patients who had a brain injury and those who had no injury, where "brain injury" is defined as any acute intracranial finding revealed on a CT image and attributable to acute head trauma. Because this classification corresponds directly to the need for CT imaging (patients with the suspected injury require this test, and the remaining ones do not), we label these two classes as *CT = yes* and *CT = no* respectively. The secondary classification indicates whether or not a neurologic intervention was needed, and hence, we refer to these two secondary classes as *neurologic intervention = yes* and *neurologic intervention = no*. It is important to note that records in the *neurologic intervention = yes* class form a subset of the *CT = yes* class. Retrospectively, the need for neurological intervention was defined in the CATCH data set by the death of the patient within a week after the head injury or by the need of any of the following procedures within the same time period: craniotomy, an elevation of skull fracture, intracranial pressure monitoring, or intubation for head injury (demonstrated on the CT image).

In order to assess the physician's perception on the use of a clinical decision rule for minor head trauma patients, Osmond conducted a survey among Canadian pediatric ED physicians to determine a clinically acceptable level of prediction performance, so that, ED physicians will be confident with the rule. Results of this survey (personal communication, 80% response rate) revealed that the detection of a serious intracranial lesion is important for clinicians. Consequently, the CATCH study targeted to achieve a sensitivity of 95% when predicting the need for CT imaging. For those patients who subsequently required neurologic intervention, the CATCH rule aimed for 98% sensitivity.

With these findings in mind, Osmond and colleagues set to create a decision rule that maximized sensitivity of prediction at an inherent cost to specificity. Using recursive partitioning they developed a rule that is clear and intuitive for ED physicians to apply for the identification of two levels of risk among children with minor head trauma. According to the derived rule, the CT decision is made through a stratified evaluation of a patient's risk factors, where the presence of any of these factors indicates the need for CT imaging to detect a serious injury. The structure of the CATCH rule is presented in Figure 1. This rule can be interpreted as a disjunction of the risk factors as the rule's premise, and the decision

to perform CT imaging as a conclusion. In case of the high risk factors, the conclusion also indicates that neurosurgical intervention is necessary. Osmond et al. evaluated the performance of the CATCH rule on 1000 bootstrapped tests and reported the sensitivity and the specificity of the high risk (top four factors in the CATCH rule) for neurologic intervention as 97.9% and 70.2% respectively. They also reported the sensitivity and the specificity of all risk factors for the need of CT imaging to detect any brain injury as 98.1% and 50.0% respectively.

1.2 Research question

We were granted a unique opportunity to work with CATCH data to develop a prediction model that indicates the need for CT imaging. Following the CATCH study, Osmond and colleagues have initiated a prospective evaluation of the CATCH rule in selected Canadian hospitals. For this evaluation, patient information was limited to 17 out of the original 26 attributes. In order to maintain compatibility and continuity of the CATCH study, we decided to use the same 17 attributes for the construction of our prediction models from the CATCH data. In this way, the model discussed in the paper can be tested again when prospectively collected data becomes available.

The decision of whether a minor head injury patient requires CT imaging is a binary classification problem. The objective is to distinguish between patients who require a CT scan (*CT = yes*) and those who do not (*CT = no*). Thus, our research question is: *can a balanced (in terms of sensitivity and specificity) and well performing prediction model be automatically derived from the CATCH data?* As a corollary to this question, we do not constrain the prediction model with respect to its interpretability and comprehensibility by non-computer science experts.

An argument for having a balanced prediction model relies on the need to mitigate long-term effects of ionizing radiation associated with the potential overuse of CT imaging that might occur when maximization of sensitivity drives model's development. We are aware that the CATCH rule developed in conjunction with physicians' expertise according to a conservative approach is likely to outperform (in terms of sensitivity) an automatically constructed prediction model. However, we believe that such model may help in establishing reference performance indicators for the CATCH rule and estimating a trade-off between the sensitivity and specificity of prediction.

Additionally, we want to show how to automatically develop a prediction model from severely imbalanced data. This class imbalance situation is commonly encountered when analyzing clinical data where the population of patients with an acute health condition is usually significantly smaller than the population of relatively healthy ones. Our research demonstrates that well-performing model can be developed by utilizing data under-sampling when constructing an ensemble prediction classifier composed of multiple Naive Bayes (NB) classifiers.

While the CATCH study explicitly identifies a high-risk subgroup of those patients who need neurologic intervention (*neurologic intervention = yes* class), we do not make this distinction, and therefore, we do not consider maximal sensitivity of prediction for this group to be a driving objective for the development of our model. However, for the purpose

of consistency with Osmond's study, we report separately, the model's performance for patients in the *neurologic intervention = yes* class (i.e., the high risk patients according to the CATCH rule).

The paper is organized as follows. In the next section, we present related research on applying data mining techniques to clinical problems. Section 3 describes the data set used in this research, briefly characterizes data mining methods selected for the study, and reviews the experimental design. Section 4 presents experimental results, and the last section concludes with a discussion.

2 Related research

Data mining techniques allow for the development of sophisticated prediction models capable of analyzing high-dimensional data [8] without relying on domain expertise during the model development process. Techniques that are suitable for medical domains are discussed and summarized in [9] – they include rule and decision tree induction, instance-based learning, Bayesian classification, and inductive logic programming.

Clinical data that describes a specific patient condition or disease poses significant challenges for data mining. Typically, such data have few instances representing very sick patients that are overwhelmed by thousands of instances representing mildly sick or healthy patients [10] – this overwhelming effect is known as "class imbalance". It is important to stress that the presence of class imbalance significantly limits the ability of the model to make accurate predictions regardless of the data mining method used [11]. Thus, directly addressing class imbalance is the first step to be taken before searching for the best performing prediction model [11]. Methods that address the class imbalance problem follow two main approaches [12, 13]: modifying the class distribution in the data by sampling, or adjusting the mining method. These two approaches are also combined into hybrid techniques in [12, 13].

Sampling the data can balance the class distribution by either increasing (over-sampling) the frequency of the minority class, or by decreasing (under-sampling) the frequency of the majority class. Simple sampling techniques involve the duplication or the removal of randomly selected instances. It has been shown that random under-sampling outperforms random over-sampling [11, 14], and that random over-sampling can lead to over-fitting [15]. More advanced sampling approaches rely on an informed selection of instances to remove (for example by targeting noisy or redundant instances from the majority class [16]) or introduce synthetic instances in specific regions of the minority class [17].

Common techniques adjusting the mining method are ensemble learning and cost-based learning [18]. Ensemble learning involves training multiple single classifiers (so-called member classifiers) on various subsets of the data for broader coverage, and subsequently, combining them to form one prediction model. An ensemble can be developed using bagging where individual member classifiers are trained on different, randomly selected sub-samples of the training data, and when combined to form the ensemble, a voting or averaging process combines their predictions to determine the overall prediction. Boosting [19] is another ensemble method that combines several single classifiers by weighting

(boosting) their classification results to improve the accuracy by repeatedly constructing consecutive member classifiers from misclassified training instances. However, it is documented that prediction models developed with boosting have a tendency to over fit the data [19].

In cost-based learning, the learning algorithm is modified to accommodate for varying misclassification costs. Instances in the minority class usually have higher misclassification costs than those in the majority class. This approach involves specialized methods developed to explicitly use costs when constructing and applying a prediction model [20-23], as well as generic methods that act as wrappers around any data mining method to make it cost sensitive [24, 25]. In the medical domain however, it is often difficult to determine the relative cost of misclassification. Most physicians would argue that failing to identify acute patients (members of small positive class) should be avoided at any cost, as demonstrated by the results of the survey communicated to us by Osmond.

The majority of hybrid approaches integrate sampling with ensemble learning, where various sampling techniques are applied to the training data sub-samples that are used to construct individual prediction models [26, 27]. Hybrid approaches have been shown to produce models that outperform ensemble models alone and single classifiers constructed from sampled data [27].

In medical domains, measuring the performance of a prediction model is a complex problem and involves multiple performance criteria to evaluate the model's accuracy, robustness and confidence [9, 28]. Computing the accuracy metric fails to deliver proper performance assessment due to the insensitivity to the imbalance in the class distribution [29]. Therefore, sensitivity and specificity metrics are used instead. These two measures collectively assess the ability of the model to detect positive instances (sensitivity) while being able to reject negative ones (specificity). Kubat et al. [30] proposed calculating the geometric mean of sensitivity and specificity (G-mean in short) to assess how balanced the performance of the prediction model is. Moreover, the inherent trade-off between sensitivity and specificity is captured by the receiver operating characteristics (ROC) curve [31]. The ROC curve can be summarized with a scalar metric by computing the area under the curve (AUC) [32]. The latter measures how well the prediction model separates the two classes.

3 Material, methods, and techniques

3.1 Data set

Attributes describing the CATCH data, which were used in our analysis, are listed in Table 1. We applied an automatic approach to discretizing values for *Age* and with the aid of a clinical expert we discretized values for *VomitNum*. Both discretizations were verified and approved by physicians involved in the CATCH study. To replicate the CATCH study design, we imputed missing attribute values with clinically reasonable values (they usually corresponded to a negative answer, e.g., *no* or *none* in lieu of a missing value).

After imputing missing values, the data was evaluated for possible inconsistencies among records from the *CT = no* class. We considered a *CT = no* record to be inconsistent if there

existed an identical (i.e., described with exactly the same values of 17 clinical attributes) record in the *CT = yes* class. There were 134 such records and we eliminated them from the data as potential bias. Subsequent analysis showed that this data cleaning did not impact the final results and comparison to the CATCH rule.

The basic characteristics of data used in the study are given in Table 2. The data set was heavily imbalanced - only 4.3% of all children included in the analysis required CT imaging (*CT = yes* class), and just 0.6% were in need of neurologic intervention. Apart from class imbalance, the CATCH data demonstrated other challenging characteristics that further complicated the task of class separation by a prediction model. They included "rare" cases in the *CT = yes* class (i.e., very small clusters of instances), as well as overlapping boundaries between classes. These two issues are often more problematic than the class imbalance itself [33, 34]. Evaluation of the attributes used to describe CATCH data revealed that they were characterized by low information gain. The attribute *HemSize* with the largest value of this measure (0.033, the second best was 0.022) decreased entropy by only 12%.

Figure 2 presents the plot of a self-organizing map (SOM [35]) that visualizes the CATCH data and illustrates issues with data. Clearly, the *CT = yes* and *CT = no* classes are difficult to separate (as indicated by overlapping points on the plot), and *CT = yes* instances are few and scattered throughout the space. Moreover, the critical instances (*neurosurgical intervention = yes*) are rare and dispersed.

All the characteristics described above are inherent for medical data [10], and they amplify the fact that CATCH data truly reflects a very difficult diagnostic problem of identifying an acute medical condition in a patient population that is clinically unlikely to suffer from this condition.

3.2 Methods for constructing prediction models

Our objective was to construct a well-performing and balanced (in comparison to the CATCH rule) prediction model from class-imbalanced medical data characterized by scattered instances in the minority class, "noisy" boundaries between the classes, and low information gain of the attributes. In order to address these issues we used a hybrid approach that builds an ensemble of NB classifiers whose classification thresholds are adjusted and which are constructed from under-sampled training sub-samples as we described in [36]. In doing so, we also followed the recommendation by Tanwani et al. [28] who advocate the use of ensemble methods when dealing with imbalanced class distributions in biomedical data. Our reliance on NB was in line with what Sajda [8] writes about Bayesian methods for biomedical applications: "*Analysis and classification of biomedical data is challenging because it must be done in the face of uncertainty; datasets are often noisy, incomplete, and prior knowledge may be inconsistent with the measurements. Bayesian decision theory is a principal approach for inferring underlying properties of data in the face of such uncertainty. More recently Bayesian methods have become a cornerstone in machine learning, and in learning theory in general, and have been able to account for a range of inference problems relevant to biological learning*." This was further confirmed in a

series of auxiliary experiments, which we conducted, where NB came out as the best performing classifier among all classifiers tested.

Specific details of proposed hybrid approach as well as discussion of the performance of an ensemble prediction model are described in detail in [36]. Below we briefly summarize how each member of the ensemble (denoted further as E-NB) was constructed:

1. The training data was balanced by under-sampling the majority class (*CT = no*). This sampling technique preserved the entire minority class (*CT = yes*) and randomly selected the same number of instances from the majority class without replacement. After experimenting with different class distribution ratios, we decided to proceed with the class distribution ratio of 1:1 (same number of instances coming from each class). Under-sampling can potentially cause a loss of information due to the reduction of the *CT = no* class. However, our experiment showed that this loss was relatively small due to the fact that many instances in a majority class were very similar to each other, and in a sense, they are deemed redundant from the perspective of the prediction model. Thus, the majority class was well described.

2. The class membership probabilities for instances in the training set were calculated according to Naïve Bayes classification method using Bayes rule of conditional probability (see [9] for details).

3. Class membership probabilities (computed in 2) were adjusted to maximize the individual NB model's performance on both classes [37]. This adjustment was carried out by the selection of a mid-point threshold on the probability output obtained from the NB classification method. The mid-point threshold was set to maximize the F-measure (a weighted harmonic mean of the precision and recall of predictions) [38]. This approach slightly preferred sensitivity to specificity in a situation when there was a tie between these two measures

According to our research reported in [36], a number of members in an ensemble can be approximated as n/n^+, where n is the total number of instances and n^+ is the number of instances in the minority class. Thus for the CATCH data this would produce an ensemble composed of 23 members. However, conducted experiments with various sizes of ensembles allowed us to reduce the number of members to 10, decreasing the complexity of the E-NB model. When predicting the class membership for a new instance, members of the E-NB (i.e. 10 NB classifiers) were applied, and their outcomes – class membership probabilities – were averaged. The resulting average probability became the outcome of the E-NB model for the instance in question.

Our expectation was that the E-NB model would achieve similar sensitivity but better specificity than its individual members. This is justified because each member of the ensemble was trained on data composed of the entire *CT = yes* class and on a random sample (of equal size) of the *CT = no* class. Thus, each member was developed using the same instances from *CT = yes* class and varying instances from the *CT = no* class leading to improved "coverage" of the majority class.

3.3 Experimental design

The experiment consisted of two phases. Phase 1 involved evaluating and comparing performance of E-NB to other prediction models built on the same ensemble scheme and employing different types of member classifiers. Phase 2 dealt with assessing the performance of E-NB in relation to the CATCH rule following the evaluation schema reported in [6].

3.3.1 Phase 1

In order to assess performance of the E-NB model we compared it to other ensemble models constructed according to the same scheme (10 member classifiers constructed from under-sampled training sets), but employing different types of member classifiers. Specifically, we used the following member classifiers that are typically considered in medical domain [9]: rule-based, tree-based and instance-based. We refer to the resulting ensemble models as to E-RB – the ensemble of rule-based classifiers constructed using the RIPPER algorithm [39], E-TB – the ensemble of tree-based classifiers constructed using C4.5 [40], and E-IB – the ensemble of instance-based classifiers constructed with the *k*-nearest neighbor (kNN) technique. All models were implemented using WEKA [41] and default values of learning algorithms' parameters. Only in the case of kNN we set *k* to 3 following recommendations from other studies on imbalanced data ([27, 42]).

The performance of all ensembles was measured by averaging the results over ten rounds of 10-fold cross-validation. In each round, all models were tested on unseen patient records (holdout data) and the quality of their predictions was measured by calculating the values of AUC, sensitivity and specificity of prediction (for the *CT = yes* class), and G-mean. Differences between values of these measures were assessed for statistical significance (using a paired t-test) at the 5% significance level. Primary evaluation measures were AUC and G-mean followed by sensitivity and specificity.

3.3.2 Phase 2

In order to compare the performance of the E-NB model with the CATCH rule, we used the bootstrap method with 1000 iterations because this evaluation strategy was used and reported by Osmond and colleagues [6]. The bootstrap method followed the .632 schema [43] that includes the following steps, assuming the data set *D* with *n* instances:

1. Select *n* instances of *D* using random selection with replacement into the training set.
2. Select these instances of *D* that haven't been selected for the training set into the testing set.
3. Train the prediction model using data obtained from step 1.
4. Test the prediction model obtained in step 2 on the training set from step 1 and the testing set from set 2.
5. Compute performance *p* according to $p = 0.368 \cdot p_{train} + 0.632 \cdot p_{test}$, where p_{train} and p_{test} is the performance observed on the training and testing sets respectively.

The steps described above were repeated 1000 times and results from the iterations were averaged in order to obtain the final performance in terms of sensitivity and specificity (similarly to what Osmond and colleagues reported). We need to point out that it was impossible for us to use the exact training/testing partitions of the data as those used in evaluating the CATCH rule. However, we believe that repeating the bootstrap 1000 times produces a reliable estimate of the model's performance.

Phase 2 terminated with presenting the results of sensitivity, specificity and G-mean (this is because it was not possible to compute AUC for the CATCH rule). These results were expanded with the number of correctly classified critical patients (the *neurologic intervention = yes* class). Collectively, they allow for drawing conclusions with regards to how the automatically developed E-NB model measured up when compared with the CATCH rule.

4 Results

4.1 Evaluation of the E-NB model

Table 3 contains evaluations of E-NB and the three other ensemble models. It shows the mean and standard deviations of the evaluation measures as well as their confidence intervals (CI) with 95% confidence.

The E-NB model outperformed other models in terms of AUC values, thus it demonstrated the best capability to separate decision classes, and all differences between AUC values were statistically significant. E-NB was also best in terms of G-mean, only the E-TB model achieved comparable value – 77.7% for E-TB vs. 78.0% for E-NB with statistically insignificant difference. Differences of G-mean between E-NB and the other two models were statistically significant.

In terms of the sensitivity, the E-NB model outperformed all three other models and the differences were statistically significant. At the same time E-NB was the least performing model in terms of specificity (differences between E-NB and all other models were statistically significant). Such performance of the E-NB model was associated with adjusted probabilities (auxiliary analysis revealed that the E-NB model with no probability adjustment demonstrated an opposite behavior, i.e., lower sensitivity and higher specificity). While this result might be perceived as a drawback, we decided that the superior performance of the E-NB in terms of the AUC, G-mean, and sensitivity measures supported its selection as the best performing ensemble model. The E-NB model was also more stable, as indicated by the lowest standard deviations of AUC, sensitivity and G-mean values among all compared models.

4.2 Comparison of the E-NB model to the CATCH rule

In order to be consistent with the experimental design reported by Osmond and colleagues, we also assessed performance of the E-NB model using the bootstrap method described earlier. We only calculated values for sensitivity, specificity and G-mean – in case of the CATCH rule the latter value was calculated from overall results given in [6]. The results are reported in Table 4.

In the bootstrap experiment, the E-NB model demonstrated much higher value of G-mean than the CATCH rule (78.4% vs. 70.0%) confirming that it is a better-balanced model. The sound performance of the E-NB model was evident when detecting patients who did not require CT imaging (*CT = no* class) – it achieved an average specificity of 74.4% compared to 50.0% reported for the CATCH rule. If we consider the sensitivity of predictions, the E-NB model remained reasonably effective – it produced an average prediction sensitivity of 82.8% compared to 98.1% for the CATCH rule.

As for the group of critical patients (*neurologic intervention = yes*), in a bootstrap experiment E-NB correctly captured 20 out of 24 such patients, and the CATCH rule captured almost all of them (reported sensitivity of the CATCH rule for these patients is 97.9%). To better understand this less than desired performance of the E-NB model on critical patients, we set on identifying the source of the misclassification errors. After repeated experiments, it became clear that the same two critical patients were misclassified in almost all experiments and two others were frequently misclassified. This suggests the presence of a systematic difficulty when E-NB attempts to correctly classifying these four patients.

To learn more about the reasons behind this difficulty we carried a kNN analysis (for the neighborhoods with varying size as defined by k = 3,....,10) on all of the 24 critical patients. Without the use of under-sampling, most instances of critical patients were very close (i.e. similar) to patients from the *CT = no* class, and an instance of only one critical patient appeared to be surrounded by several instances of patients from the *CT = yes* class (correctly, it had more neighbors from the *CT = yes* class than from *CT = no*). The number of such critical patients increased to twelve when the under-sampling was used. In addition, the same two critical patients (which were almost always misclassified by E-NB) had the majority of *CT= no* patients in their neighborhoods for all considered values of k. A plausible explanation for the apparent weak separation between critical and other patients is attributed to a clinical similarity of these patients to those who do not require CT imaging. This indicates a need for a substantial tacit knowledge involved in separating these two groups of patients. Thus, we posit that prediction errors for critical patients will persist for the CATCH data regardless of the type of a prediction model that is being used.

5 Discussion

The decision to order a diagnostic test and the timing of this test are two important facets of medical decision-making. The CATCH rule was developed to help identify children with minor head injury who require CT imaging. It was created from prospectively collected data and designed to eliminate the false negatives for critical patients who require neurologic intervention. Therefore, the rule's performance is characterized by an almost perfect sensitivity at a cost of low specificity.

Having been granted access to the CATCH data, we conducted research on the feasibility of developing a balanced prediction model from this data. We followed a hybrid approach that combines ensemble learning with data sampling. Specifically, we developed the E-NB model being an ensemble of 10 NB classifiers [36], where each member was constructed from under-sampled training sub-sample and its probabilities (as well as decision

threshold) were adjusted for balanced performance. Under-sampling combined with ensemble learning successfully dealt with class imbalance (member classifiers were constructed from balanced subsamples), minimized possible information loss [26] in the majority *CT = no* class (10 different subsets of the majority class were used to construct specific member models) and improved the stability of the resulting E-NB model. Under-sampling also helped with rare cases by cleaning their neighborhood, as confirmed by the kNN analysis. Moreover, member NB classifiers were able to deal with noisy concepts [44] and overlapping regions between classes [45] and to provide relatively accurate predictions for the minority *CT = yes* class. Their performance was further improved by adjusting probabilities and the decision threshold [37].

Although the E-NB model is unable to achieve the same level of sensitivity as the CATCH rule (82.8% vs. 98.1%), it improves the specificity (74.4% vs. 50.0%) and improves the balance as measured by the G-mean (78.4% vs. 70.0%). While the model's predictive performance may not meet the expectations of ED physicians with regards to CT imaging decisions, it provides a good estimate of the loss in specificity when decision-making behavior is driven by the maximization of the sensitivity of prediction.

From a clinical perspective, the E-NB prediction model missed too many critical patients requiring neurologic intervention. Our analysis shows that despite removing clearly inconsistent instances there is an overlap in descriptions of critical patients and those who do not require CT imaging ("noisy" boundaries and rare instances, visible in Figure 2) that is not handled properly by the E-NB model despite all the mitigating actions we applied. Such an overlap can hardly be controlled when using automated mining methods without a significant impact on the model's balanced performance [42]. While under-sampling was able to partially address this issue, we believe further improvement would require a problem-specific informed re-sampling scheme that uses domain knowledge to evaluate similarity (or distance) between the instances.

In conclusion, our research demonstrates that making clinical decisions often is beyond capabilities of prediction models that are constructed by applying automatic discovery process in search for patterns in data. Frequently, these patterns are obstructed, extremely difficult to detect (as was the case with the neurologic intervention patients) and require interpretation by an experienced physician. Thus, clinical decision rules developed with help of domain experts will typically outperform a prediction model developed automatically, especially when a conservative approach is favored in the diagnostic process. However, an automated approach like ours may help in establishing reference performance indicators for these decision rules and assist their developers in better estimating a trade-off between the sensitivity and specificity of prediction.

Acknowledgements

The authors would like to thank Terry P. Klassen MD, George A. Wells PhD, Rhonda Correll RN, Anna Jarvis MD, Gary Joubert MD, Benoit Bailey MD, Laurel Chauvin-Kimoff MD CM, Martin Pusic MD, Don McConnell MD, Cheri Nijssen-Jordan MD, Norm Silver MD, Brett Taylor MD, Ian G. Stiell MD; of the Pediatric Emergency Research Canada (PERC) Head Injury Study Group for providing access to the CATCH data.

The current version of the paper benefited from the insightful comments of the reviewers.

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institutes of Health Research (CIHR) and NSERC-CREATE programs. The second author wishes to acknowledge support of the Polish Ministry of Science and Higher Education.

Dr. Klement and Dr. Wilk conducted this research while being postdoctoral fellows with the MET Research Group at the University of Ottawa.

Conflict of interest statement

No conflicts of interest exist.

References

[1]. M. Smits, D.W. Dippel, P.J. Nederkoorn, H.M. Dekker, P.E. Vos, D.R. Kool et al., Minor head injury: CT-based strategies for management - a cost-effectiveness analysis, *Radiology 254* (2010) 532-540.
[2]. D.J. Brenner and E.J. Hall, Computed tomography -- an increasing source of radiation exposure, *N Engl J Med 357* (2007) 2277-2284.
[3]. P.J. Bairstow, J. Persaud, R. Mendelson and L. Nguyen, Reducing inappropriate diagnostic practice through education and decision support, *Int J Qual Health Care 22* (2010) 194-200.
[4]. J.S. Broder, CT utilization: the emergency department perspective, *Pediatr Radiol 38 Suppl 4* (2008) 664-669.
[5]. F. Rivara, D. Tanaguchi, R.A. Parish, G.K. Stimac and B. Mueller, Poor prediction of positive computed tomographic scans by clinical criteria in symptomatic pediatric head trauma, *Pediatrics 80* (1987) 579-584.
[6]. M.H. Osmond, T.P. Klassen, G.A. Wells, R. Correll, A. Jarvis, G. Joubert et al., CATCH: a clinical decision rule for the use of computed tomography in children with minor head injury, *CMAJ 182* (2010) 341-348.
[7]. T.P. Klassen, M.H. Reed, I.G. Stiell, C. Nijssen-Jordan, M. Tenenbein, G. Joubert et al., Variation in utilization of computed tomography scanning for the investigation of minor head trauma in children: a Canadian experience, *Acad Emerg Med 7* (2000) 739-44.
[8]. P. Sajda, Machine learning for detection and diagnosis of disease, *Annu Rev Biomed Eng 8* (2006) 537-565.
[9]. N. Lavrac, Selected techniques for data mining in medicine, *Artif Intell Med 16* (1999) 3-23.
[10]. K.J. Cios and G.W. Moore, Uniqueness of medical data mining, *Artif Intell Med 26* (2002) 1-24.
[11]. C. Drummond and R.C. Holte, Severe class imbalance: Why better algorithms aren't the answer, in: J. Gama, R. Camacho, P. Brazdil, A. Jorge and L. Torgo, eds., *Proceedings of the 16th European Conference of Machine Learning, ECML 2005* (Springer, Berlin / Heidelberg / New York, 2005) 539-546.
[12]. H. He and E.A. Garcia, Learning from imbalanced data, *IEEE Trans Knowl Data Eng 21* (2009) 1263-1284.
[13]. N.V. Chawla, Data mining for imbalanced data sets: an overview, in: O. Maimon and L. Rokach, eds., *The Data Mining and Knowledge Discovery Handbook* (Springer, 2005) 853-867.
[14]. C. Drummond and R.C. Holte, C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, in: *Proceedings of the International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets II* 2003) 1-8.
[15]. N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, *J Artif Intell Res 16* (2002) 321-357.
[16]. M. Kubat and S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: *Proceedings of the 14th International Conference on Machine Learning, ICML'97* (Morgan Kaufmann, San Francisco, CA, 1997) 179-186.
[17]. H. Han, W.Y. Wang and B.H. Mao, Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, in: J.M. Benitez and M. Ali, eds., *Proceedings of the 23rd*

Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2010s (Springer, Berlin / Heidelberg / New York, 2005) 878-887.
[18]. V. Garcia, J.S. Sanchez, R.A. Mollineda, R. Alejo and J.M. Sotoca, The class imbalance problem in pattern classification and learning, in: *II Congreso Español de Informática, CEDI 2007* (Thomson, 2007) 283-291.
[19]. R.E. Schapire, The boosting approach to machine learning: An overview, in: D.D. Denison, M.H. Hansen, C. Holmes, B. Mallick and B. Yu, eds., *Nonlinear Estimation and Classification* (Springer, New York, 2003)
[20]. C. Drummond and R.C. Holte, Exploiting the cost (in)sensitivity of decision tree splitting criteria, in: *Proceedings of the 17th International Conference on Machine Learning, ICML'00* (Morgan Kaufmann, San Francisco, CA, 2000) 239-246.
[21]. W. Fan, S.J. Stolfo, J. Zhang and P.K. Chan, AdaCost: misclassification cost-sensitive boosting, in: *Proceedings of the 16th International Conference on Machine Learning, ICML'99* (Morgan Kaufmann, San Francisco, CA, 1999) 97-105.
[22]. D.D. Margineantu, Class probability estimation and cost-sensitive classification decisions, in: T. Elomaa, H. Mannila and H. Toivonen, eds., *Proceedings of the 13th European Conference on Machine Learning, ECML 2002* (Springer, Berlin / Heidelberg / New York, 2002) 270-281.
[23]. B. Zadrozny and C. Elkan, Learning and making decisions when costs and probabilities are both unknown, in: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-2001* (ACM Press, New York, NY, 2001) 204-213.
[24]. P. Domingos, MetaCost: a general method for making classifiers cost-sensitive, in: *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-99* (ACM Press, New York, NY, 1999) 155-164.
[25]. B. Zadrozny, J. Langford and N. Abe, Cost-sensitive learning by cost-proportionate example weighting, in: *Proceedings of the 3rd IEEE International Conference on Data Mining, ICDM'03* (IEEE Computer Society, Washington, DC, 2003) 435.
[26]. X.-Y. Liu, J. Wu and Z.-H. Zhou, Exploratory under-sampling for class-imbalance learning, *IEEE Trans Syst Man Cybern Part B Cybern 39* (2009) 539-550.
[27]. J. Błaszczyński, M. Deckert, J. Stefanowski and S. Wilk, Integrating selective pre-processing of imbalanced data with Ivotes ensemble, in: M. Szczuka, M. Kryszkiewicz, S. Ramanna, R. Jensen and Q. Hu, eds., *Proceesings of the 7th International Conference, RSCTC 2010* (Springer, Berlin / Heidelberg / New York, 2010) 148-157.
[28]. A.K. Tanwani, J. Afridi, M.Z. Shafiq and M. Farooq, Guidelines to select machine learning scheme for classification of biomedical datasets, in: *Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO'09* (Springer, Berlin / Heidelberg, 2009) 128-139.
[29]. F. Provost, T. Fawcett and R. Kohavi, The case against accuracy estimation for comparing induction algorithms, in: *Proceedings of the 15th International Conference on Machine Learning, ICML'97* (Morgan Kaufmann, San Francisco, CA, 1997) 445-453.
[30]. M. Kubat, R.C. Holte and S. Matwin, Learning when negative examples abound, in: M. van Someren and G. Widmer, eds., *Proceedings of the 9th European Conference on Machine Learning, ECML'97* (Springer, 1997) 146-153.
[31]. T. Fawcett, *ROC graphs: Notes and practical considerations for data mining researchers*. (HP-Labs, 2003).

[32]. J.A. Hanley and B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology 143* (1982) 29-36.
[33]. N. Japkowicz, Class imbalance: Are we focusing on the right issue?, in: *Proceedings of the International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets II* 2003) 17-23.
[34]. T. Jo and N. Japkowicz, Class imbalances versus small disjuncts, *ACM SIGKDD Explor Newsl 6* (2004) 40-49.
[35]. A. Flexer, On the use of self-organizing maps for clustering and visualization, in: J. Zytkow and J. Rauch, eds., *Principles of Data Mining and Knowledge Discovery* (Springer, Berlin / Heidelberg, 1999) 80-88.
[36]. W. Klement, S. Wilk, W. Michalowski and S. Matwin, Classifying severely imbalanced data, in: C. Butz and P. Lingras, eds., *Proceedings of the 24th Canadian Conference on Advances in Artificial Intelligence, CAI'11* (Springer, Heidelberg / Dordrecht / London / New York, 2011) 258-262.
[37]. F. Provost, Learning with imbalanced data sets 101, in: *Papers from the AAAI Workshop. Technical Report WS-00-05* (AAAI Press, Menlo Park, CA, 2000) 1-4.
[38]. C.V. van Rijsbergen, *Information Retrieval*. (Butterworth, London, Boston, 1979).
[39]. W.W. Cohen, Fast effective rule induction, in: *Proceedings of the 12th International Conference on Machine Learning, ICML'95* (Morgan Kaufmann, San Francisco, CA, 1995) 115-123.
[40]. R. Quinlan, *C4.5: Programs for Machine Learning*. (Morgan Kaufmann, San Mateo, CA, 1993).
[41]. I.H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques*. (Morgan Kaufmann, 2005).
[42]. J. Stefanowski and S. Wilk, Selective pre-processing of imbalanced data for improving classification performance, in: I.-Y. Song, J. Eder and T.M. Nguyen, eds., *Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2008* (Springer, Berlin / Heidelberg, 2008)
[43]. R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: *Proceedings of the 14th International Joint Conference on Artificial intelligence, IJCAI'95* (Morgan Kaufmann, San Francisco, CA, 1995) 1137-1143.
[44]. I. Rish, An empirical study of the naive Bayes classifier, in: *Proceedings of the IJCAI'01 Workshop on Empirical Methods in Artificial Intelligence* 2001) 41-46.
[45]. V. Garcia, J. Sanchez and R.A. Mollineda, An empirical study of the behaviour of classiers on imbalanced and overlapped data sets, in: L. Rueda, D. Mery and J. Kittler, eds., *Proceedings of 12th Iberoamerican Congress on Pattern Recognition, CIARP 2007* (Springer, Berlin / Heidelberg, 2007) 397-406.

Table 1. Attributes considered in the study.

	Code	Description	Domain
1	Age	Age	0 – 6 months, > 6 months
2	Sex	Gender	male, female
3	MechInj	Mechanism of injury	struck head, motor, significant fall, sport, **other**
4	HxWitLOC	Loss of consciousness (witnessed)	**no**, < 1 min, 1-5 min, > 5 min
5	HxDisCon	Disorientation or confusion	**no**, < 1 min, 1-5 min, 6-10 min, > 10 min
6	AmnHow	Any amnesia	**no**, yes
7	AmnBefr30Min	Amnesia for events ≥ 30 min before impact	**no**, yes
8	AmnAfter30Min	Amnesia for events ≥ 30 min after impact	**no**, yes
9	HxWHache	Worsening headache	**no**, yes
10	VomitNum	Vomiting, number of episodes	**0-2**, > 2
11	HxSeiz	Seizure	no, yes
12	Letharg	Lethargy	**no**, yes
13	HxIrrit	Irritability	**no**, yes
14	IGCS	Initial Glasgow Comma Scale score	13, 14, 15
15	HemSize	Hematoma of the scalp	**none**, small & localized, large & boggy
16	SkullPen	Suspected open or depressed fracture	**no**, yes
17	BSkFrac	Signs of basilar skull fracture	**no**, yes

Values marked in bold font in a Domain column represent imputed missing values. Attributes Age, Sex, HxSeiz and IGCS had no missing values.

Table 2. Characteristics of the data set used in the study.

Class	n	%	Class	n	%
CT = no	3573	95.7	*Neurologic intervention = no*	3708	99.4
CT = yes	159	4.3	*Neurologic intervention = yes*	24	0.6
Total	3732	100.0		3732	100.0

Table 3. Results averaged over 10 runs of 10-fold cross validation.

	E-NB model	E-RB model	E-TB model	E-IB model
AUC, % (95% CI)	86.3 ± 0.2 (86.2, 86.4)	84.1 ± 0.7 (83.7, 84.5)	83.7 ± 0.4 (83.5, 84.0)	82.6 ± 0.7 (82.2, 83.1)
Sensitivity, % (95% CI)	83.5 ± 0.6 (83.2, 83.9)	75.2 ± 1.9 (74.0, 76.4)	77.7 ± 1.3 (76.9, 78.5)	58.4± 0.6 (58.0, 58.8)
Specificity, % (95% CI)	72.7 ±0.9 (72.2, 73.3)	76.9 ± 0.9 (76.3, 77.4)	77.8 ± 0.7 (77.3, 78.2)	86.5±0.4 (86.3, 86.7)
G-mean, % (95% CI)	78.0 ± 0.3 (77.7, 78.2)	76.0 ± 0.8 (75.5, 76.5)	77.7±0.6 (77.4, 78.1)	71.1±0.4 (70.8, 71.3)

Table 4. Results averaged over 1000 bootstrap iterations.

	E-NB model	CATCH rule*
Sensitivity, % (95% CI)	82.8 ± 4.0 (82.5, 83.0)	98.1 (98.0, 98.2)
Specificity, % (95% CI)	74.4 ± 3.2 (74.2, 74.6)	50.0 (50.0, 50.1)
G-mean, % (95% CI)	78.4± 1.2 (78.3, 78.5)	70.0

* As reported in Osmond et. al [6].

Figure 1. The CATCH clinical decision rule [6]

High risk

1. Glasgow Coma Scale score < 15 at two hours after injury
2. Suspected open or depressed skull fracture
3. History of worsening headache
4. Irritability on examination

Medium risk

5. Any sign of basal skull fracture (e.g. hemotympanum, "raccoon" eyes, otorrhea or rhinorrhea of the cerebrospinal fluid, Battle's sign)
6. Large, boggy hematoma of the scalp
7. Dangerous mechanism of injury (e.g. motor vehicle crash, fall from elevation >= 3ft or 5 stairs, fall from bicycle with no helmet

Figure 2. Visualization of the CATCH data using SOM