2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

A New CPXR Based Logistic Regression Method and Clinical Prognostic Modeling Results Using the Method on Traumatic Brain Injury

Vahid Taslimitehrani, Guozhu Dong
Department of Computer Science and Engineering and Kno.e.sis Center
Wright State University, Dayton, Ohio 45435, USA
taslimitehrani.2@wright.edu, guozhu.dong@wright.edu

***Abstract*—Prognostic modeling is central to medicine, as it is often used to predict patients' outcome and response to treatments and to identify important medical risk factors. Logistic regression is one of the most used approaches for clinical prediction modeling. Traumatic brain injury (TBI) is an important public health issue and a leading cause of death and disability worldwide. In this study, we adapt CPXR (Contrast Pattern Aided Regression, a recently introduced regression method), to develop a *new logistic regression method* called CPXR(Log), for general binary outcome prediction (including *prognostic modeling*), and we use the method to *carry out prognostic modeling for TBI* using admission time data. The models produced by CPXR(Log) achieved AUC as high as 0.93 and specificity as high as 0.97, much better than those reported by previous studies. Our method produced interpretable prediction models for diverse patient groups for TBI, which show that different kinds of patients should be evaluated differently for TBI outcome prediction and the odds ratios of some predictor variables differ significantly from those given by previous studies; such results can be valuable to physicians.**

I. Introduction

Aims of this study include: (A) Provide accurate and informative prognostic models for traumatic brain injury. (B) Provide a powerful new generic logistic regression method. It advances the frontier of research on the following three important issues:

(1) Prognostic models are central to medicine; they are often used to predict patients' outcome and response to medical treatments. Physicians routinely make their decisions on patient treatment plan, screening, and ordering of tests and procedures, based on the prognosis or likelihood of a disease [17]. Prognosis models also help on the understanding of diseases, including identifying discriminating variables highly correlated with the outcome. Medicine is moving from a traditional subjective one to an evidence-based one, which uses prediction models built from population samples to inform clinical decision-making [7]. Indeed, prediction modeling in medicine has received increasing attention recently; e.g. the number of articles listed in PubMed [23] with "prediction model" in the title in 2012 is 7 times of that in 2000.

(2) Logistic regression is one of the most used approaches for building clinical prediction models and its usage has been increasing over the years. Indeed, most of those articles on medical prediction modeling mentioned above used logistic regression, as shown in Figure 1. Logistic regression models are desirable since they are flexible (they can incorporate both categorical and continuous predictors, as well as non-linear transformations) and they provide probabilities for the predicted outcomes.

(3) TBI is an important public health problem and a leading cause of death and disability worldwide: Every year, more than 1.5 million people die and hundred of millions need emergency treatment [14]. In the US, CDC estimated that 2.4 million emergency room visits, hospitalizations and deaths are related to TBI and $76.5 billion dollars including direct and indirect cost (excluding combat related treatments) in 2010 [21]. While confident predictions could usually be made 24 hours after the injury, they are hard to make at admission time [9]. Physicians need to make vital decisions ranging from whether to perform/withdraw certain treatment based on their prognosis evaluation [15], and they need accurate prognostic models that only use admission time data to make time-critical clinical decisions.

Challenges in clinical modeling include the following five (some are for general predictive clinical modeling and some are for prognostic modeling for traumatic brain injury). (a) Accuracy of prediction models is the most important aspect for clinical prediction modeling, as making a wrong decision in medicine may put a human's life in danger. (b) Prediction models for medicine should be easy to interpret, so that physicians can (i) explain critical medical decisions to patients and their families and (ii) can identify the important risk factors for the disease under consideration. (c) Prediction models for medicine should avoid overfitting as much as possible, so that they can be used to make accurate predictions on new cases. (d) Prediction models should allow physicians to make early decisions. This is often critical, as an early decision will allow hospitals to make early effort on patients who will likely benefit from the treatment. For traumatic brain injury, correct treatment decisions made at time of admission, with admission time data, will help the patient to recover better, while delayed decision will diminish their chance of recovery. (e) As will be discussed in the Related Work section, traumatic brain injury patients in different population groups require different prediction models. In fact, as will be shown in this study, this heterogeneity is not limited to known population groups; TBI is an illness having diverse predictor-response variable relationships[1] [5].

[1] When we say "an illness has diverse predictor-response relationships", we mean the data associated with the illness contains multiple logical data groups whose fitted regression models are highly different. Illness can be other things.

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.16

Our approach and advantages: In this study, (1) we adapt CPXR (Contrast Pattern Aided Regression, a recently introduced regression method), to develop a new logistic regression method called CPXR(Log), for general prognostic modeling and binary outcome prediction, and (2) we apply it to traumatic brain injury outcome prediction. The CPXR(Log) algorithm constructs a pattern aided logistic regression model defined by several patterns and several associated local logistic regression models. Like CPXR, CPXR(Log) has several significant advantages including high prediction accuracy and ability to handle data with diverse predictor-response relationships, often outperforming standard logistic regression and state-of-the-art classifiers on various accuracy measures. The prediction models produced by CPXR(Log) are easy to interpret. The ability for effectively handling data with diverse predictor-response relationships is especially useful in clinical applications, as modern medicine is becoming more and more personalized.

Goodness of our results on prediction modeling for TBI: For prognostic modeling on TBI using admission time data, the models produced by CPXR(Log) achieved AUC as high as 0.93 and specificity as high as 0.97, much better than those reported by previous studies. Each prediction model produced by CPXR(Log) contains several interpretable local prediction models for different patient groups for TBI, indicating that there are several different kinds of patients that should be evaluated differently for TBI outcome prediction. We present a complete CPXR(Log) prediction model, containing the patterns and the local logistic regression models, for the Unfavorable dichotomized version of GOS using 15 predictor variables. We also study the odds ratio differences of the predictor variables based on different logistic regression models; we provide predictor variables whose odds ratios in some local model (of the CPXR(Log) model) differ from that in the global model significantly (including variables whose odds ratios change by more than 6 folds). The example CPXR(Log) model demonstrates that CPXR(Log) can also extract informative multi-variable outcome-related interactions among subsets of variables, which are hard to identify by standard logistic regression when the number of variables is large.

Organization: Section II is on related work. Section III gives preliminaries. Section IV presents the CPXR(Log) algorithm. Section V reports results on TBI. Section VI concludes.

II. Related Work

The related works belong to two main groups.

(a) Studies on general clinical prediction models: Clinical prediction modeling is a very broad and active area of research. Most recent articles on clinical prediction modeling used Logistic Regression (e.g. [1]), while others used methods such as Decision Trees (e.g. [2]), Random Forest (RF) (e.g. [10]) and Support Vector Machine (SVM) (e.g. [20]); See Figure 1.

(b) Studies on TBI related prediction models: Many studies have been reported on prediction modeling for predicting the outcome after traumatic brain injury. A pre-eminent study is the project of IMPACT (International Mission for Prognosis and Analysis of Clinical Trials in TBI) [24], which collected data for nearly 10 years and developed and validated many prognostic models for classification and prognostic risk calculation. In [12], Murray et al. examined important risk factors on TBI patients' outcome based on a cohort of 8686 patients from multiple clinical trials. They fitted a proportional odds model and found age, GSC score, pupil response and CT characteristics are the most powerful prognostic risk factors. Hukkelhoven et al. [8] performed a study to detect critical age threshold on TBI patients' outcome on a set of 5600 patients. CRASH (Corticosteroid Randomization After Significant Head Injury) [22] is another major trial on TBI which ended up with 10008 patients; it also developed prognostic models and risk calculators. Reference [19] built prognostic models to predict Mortality and Unfavorable, where the outcome classes are determined based on the GOS score at 6 months after the surgery, using Logistic Regression. Reference [2] developed similar models using decision tree analysis. Reference [13] found that heterogeneity of head injuries is a challenge in TBI prognostic models; CRASH [11] found that prognostic models for TBI patients' outcome for low, middle and high income countries differ significantly, which is in agreement with the heterogeneity findings of [13].

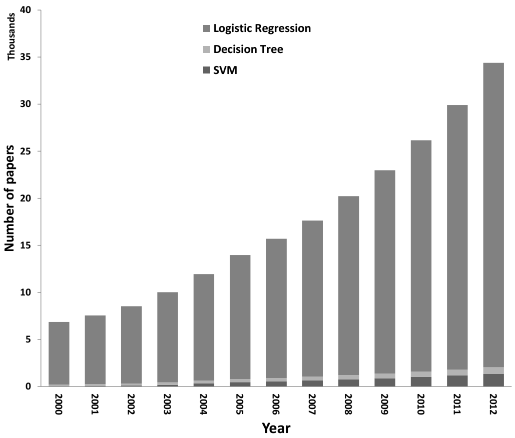


Fig. 1: Number of prediction-modeling articles published in PubMed with "Logistic Regression", "Decision Tree" and "Support Vector Machine" in the titles between 2000 and 2012.

III. Preliminaries

A. Preliminaries on logistic regression

Regression analysis aims to design regression models to predict response variable values, based on predictor variables' values. For clinical prediction modeling one often builds models to predict categorical response variable values (classes); for example, for diagnostic modeling one may want to predict "having disease" or "not" and for prognostic modeling one may want to predict "the outcome is mortal" or "not".

For logistic regression, the response variable has two values, 0 and 1, representing the two classes with 1 representing the class of interest. Let $x_1, ..., x_n$ be the predictor variables. A vector of values of the predictor variables is an *instance*, to be denoted by X below. The response variable will be denoted by Y. Logistic regression uses a set $\{(X_i, Y_i) \mid 1 \leq i \leq m\}$ of instance and response variable value pairs as *training data*.

The goal of logistic regression is to find a linear function $lp(X) = \beta_0 + \sum_{i=1}^{n} \beta_i \times x_i$ satisfying $log([P(\text{Y=1})/(P(\text{Y=1}) + 1)]) = lp(X)$. (The lefthand side is often written as $log(odds(P(\text{Y=1}))$ and $logit(P(\text{Y=1}))$.) Solving for $P(\text{Y=1})$ one gets $P(\text{Y=1}) = e^{lp(X)}/(1+e^{lp(X)}) = 1/(1 + e^{-lp(X)})$. The coefficients β_i are usually, though not always, estimated by iterative likelihood maximization. The resulting logistic regression model is $f(X) = 1/(1+e^{-lp(X)})$.

There are many ways to measure goodness of fit of any given logistic regression model g. One commonly used one is the *Chi-square* (χ^2) statistic. Standardized residual (also called Pearson's residual) of g on a particular instance X_i is

$$r_i = \frac{Y_i - g(X_i)}{\sqrt{g(X_i)(1 - g(X_i))}} \tag{1}$$

and then the Chi-square statistic (χ^2) of g is

$$\chi^2(g) = \sum_{i=1}^{m} r_i^2 \tag{2}$$

B. Preliminaries on patterns and discretization

Let D be a training data set for regression. For the pattern part, we ignore the outcome variable. To avoid using too many symbols, we will still denote the projection of D onto the predictor variables as D.

We usually use the entropy based method [6], to partition ranges of numerical variables into disjoint intervals (bins) that are as pure as possible. Bin boundaries will be used in patterns.

An *item* is a *single-variable condition* of the form "$A = a$" if A is a categorical variable, or "$v_1 \leq A < v_2$", where v_1 and v_2 are constants (usually the bin boundaries), if A is numerical. A *pattern* or *itemset* is a finite set of items. An instance X is said to *satisfy*, or *match*, a pattern P, denoted by $X \models P$, if X satisfies every item/condition in P. The *matching data* of P in D is $\mathsf{mds}(P, D) = \{X \in D \mid X \models P\}$. The *support* of P in D is $supp(P, D) = \frac{|mt(P,D)|}{|D|}$. We may generalize the above by using a subset D' of D to replace D, as in $supp(P, D')$.

Intuitively, a pattern is a contrast pattern if its supports in different classes are very different. Formally we have:

Definition 1: [4] Given two data classes C_1 and C_2, the *support ratio* (also called *growth rate*) of a pattern P from C_1 to C_2 is[2] $suppRatio_{C_1}^{C_2}(P) = \frac{supp(P,C_2)}{supp(P,C_1)}$. Given a support ratio threshold γ, a *contrast pattern* (also called *emerging pattern*) of class C_2 is a pattern P satisfying $suppRatio_{C_1}^{C_2}(P) \geq \gamma$.

In addition to γ, in this paper we will also use a threshold, $minSup$, on contrast patterns in the LE class (see below).

IV. The CPXR(Log) algorithm

This section presents the Pattern Aided Regression model (PXR) and the Logistic Contrast Pattern Aided Regression Algorithm (CPXR(Log)) that constructs highly accurate logistic PXR models. PXR and CPXR for linear regression were both introduced in [5] for the linear regression case. This section will discuss how to adapt CPXR, to build logistic PXR models instead of linear ones.

Let $D = \{(X_i, Y_i) \mid 1 \leq i \leq n\}$ be a given training data set for regression. Let f be a regression model built on D, which we will call the baseline model on D.

A. Pattern aided regression model (PXR) concepts

The main ideas in the pattern aided regression model type introduced in [5] are (a) to use a pattern P as a logical characterization of a specific group of data, and a local regression model f_P as a behavioral characterization of the intrinsic predictor-response relationship fitting that group of data, and (b) to use a small set of patterns and associated local regression models for several groups of data to define a regression model for all data.

We will say a PXR model is a linear PXR model if its local regression models are linear, and similarly for logistic PXR models.

PXR's strength lies with its flexible use of a pattern and local-regression-model pair to represent one particular predictor-response relationship for one group of data, and with its use of multiple pattern and local-model pairs to represent multiple predictor-response relationships, for different logical groups of data. The ability to represent diverse predictor-response variable relationships was considered as the main reason why linear PXR models obtained by CPXR significantly outperform state-of-the-art linear regression models [5]. As will be seen later, logistic PXR models can achieve much more accurate prediction than state-of-the-art logistic regression models.

Definition 2: [5] A *pattern aided regression* (PXR) model is a tuple $PM = ((P_1, f_1, w_1), ..., (P_k, f_k, w_k), f_d)$, where $k > 0$ is an integer, $P_1, ..., P_k$ are patterns, f_1, ..., f_k, f_d are regression models, and $w_1, ..., w_k > 0$ are weights. The set $\{P_1, ..., P_k\}$ is the *pattern set* of PM, f_i is the *local regression model* of P_i, and f_d is the *default regression model*. We define *the regression function of* PM, for each instance X, as

$$f_{PM}(X) = \begin{cases} \frac{\Sigma_{P_i \in \pi_X} w_i f_i(X)}{\Sigma_{P_i \in \pi_X} w_i} & \text{if } \pi_X \neq \emptyset \\ f_d(X) & \text{otherwise} \end{cases} \tag{3}$$

where $\pi_X = \{P_i \mid 1 \leq i \leq k,\ X \text{ satisfies } P_i\}$.

Remark: f_i is only applied to instances x satisfying P_i, and f_d is only used when x does not satisfy any P_i.

We typically define w_i based on f_i's residual reduction. Moreover, for each pattern P, in this paper we use a standard logistic regression method to build a local logistic regression model f_P. So f_i is f_{P_i} in the above definition.

B. Quality measures on patterns and pattern sets

CPXR(Log) needs to efficiently find a desirable pattern set, from huge search spaces of potential pattern sets, in order to build an accurate prediction model. We need quality measures on individual patterns and on pattern sets, to remove patterns that have little hope of being useful and to perform efficient incremental search of pattern sets respectively.

Let $r_X(g)$ denote a function g's residual on an instance X.

[2] If $supp(P, C_1) = 0$ we define the ratio as a large number such as $2|C_2|$ and we also call P a jumping emerging pattern.

Definition 3: The **average residual reduction** (arr) of a pattern P w.r.t. a prediction model f and a data set D is

$$\mathsf{arr}(P) = \frac{\Sigma_{X \in \mathsf{mds}(P)} |r_X(f)| - \Sigma_{X \in \mathsf{mds}(P)} |r_X(f_P)|}{|\mathsf{mds}(P)|} \tag{4}$$

The **total residual reduction** (trr) of a pattern set $PS = \{P_1, ..., P_k\}$ w.r.t. a prediction model f and a data set D is

$$\mathsf{trr}(PS) = \frac{\Sigma_{X \in \mathsf{mds}(PS)} |r_X(f)| - \Sigma_{X \in \mathsf{mds}(PS)} |r_X(f_{PM})|}{\Sigma_{X \in D} |r_X(f)|} \tag{5}$$

where $PM = ((P_1, f_{P_1}, w_1), ..., (P_k, f_{P_k}, w_k), f)$, $w_i = \mathsf{arr}(P_i)$, and $\mathsf{mds}(PS) = \cup_{P \in PS}\ \mathsf{mds}(P)$.

The arr measure will help filter out patterns of little value and will define weights for patterns in PXR models. The trr measure will be used for pattern set selection.

C. Description of the CPXR(Log) algorithm

To save space, we give a verbal description of CPXR(Log), omitting the pseudo-codes; it is very similar to CPXR [5] except that it uses logistic regression instead of linear regression, it builds the baseline regression model in the algorithm by default, and it uses the baseline model as the default model.

The CPXR(Log) algorithm takes three inputs: a training data set D for logistic regression, a ratio ρ for dividing D into large error (LE) and small error (SE) parts, and a $minSup$ threshold on contrast patterns.

First, CPXR(Log) builds the baseline regression model f using a standard logistic regression method (e.g. the implementation in the R package). Then it uses ρ to determine a split value κ, to partition D into LE (large error) and SE (small error) parts, so that $\rho \approx \frac{\sum_{X \in LE} |r_X(f)|}{\sum_{X \in D} |r_X(f)|}$. Then CPXR(Log) mines the contrast patterns of LE. Then it selects just one pattern having the shortest length from each equivalence class of patterns as candidate for subsequent consideration. (An equivalence class of patterns is a set of patterns with the same mds, defined by the equivalence relation given by: patterns P and Q are equivalent iff $\mathsf{mds}(P) = \mathsf{mds}(Q)$.) Then it uses some filters (see [5]) to remove patterns that are highly similar to others. Then it builds f_P for each remaining contrast pattern of LE, and removes patterns that yield little residual reduction. Then it uses a double loop to search for a desirable pattern set with large trr: The inner loop performs repeated pattern replacements, and the outer loop adds a new pattern to the pattern set and then calls the inner loop (both aimed to improve trr). The inner loop terminates when the improvement of the best replacement is too small. The outer loop terminates when the improvement of the previous iteration is too small. Let $PS = \{P_1, ..., P_k\}$ denote the resulting set of patterns. Then CPXR(Log) returns $PM = ((P_1, f_{P_1}, \mathsf{arr}(P_1)), ..., (P_k, f_{P_k}, \mathsf{arr}(P_k)), f)$ as the resulting PXR model. For this paper, all local regression models are built by the standard logistic regression algorithm. In general, the algorithm also builds a logistic regression model f_d for the set of instances that do not match any P_i ($1 \leq i \leq k$) and uses it in the place of f in PM.

V. Experimental Results

This section presents the results of CPXR(Log) on prediction of 6-month outcome after moderate or severe TBI. It has two focuses, one on accuracy of CPXR(Log) models and the other on new insights on TBI offered by CPXR(Log) models (which could be useful to medical scientists and physicians). (a) For the former, we mostly compare CPXR(Log) against standard logistic regression (denoted by SLogR), while briefly comparing against state-of-the-art classification algorithms such as SVM and RF. The results indicate that CPXR(Log) is more accurate, outperforming SLogR and others significantly. (b) For the latter, the CPXR(Log) models present new patterns that capture outcome-related interactions among variables and that define groups of patients whose outcome should be predicted using their own local prediction models instead of the SLogR model. Moreover, based on CPXR(Log) we present variables having high odds ratio for certain patient groups (defined by patterns used by the CPXR(Log) models) and having low odds ratio based on the SLogR model. We believe that important variables for TBI include those whose CPXR(Log) based odds ratios differ from their SLogR based odds ratio by large margins, and those that occur in patterns used by the CPXR(Log) models.

Regarding parameters, we used fixed values $minSup = 0.02$ and $\rho = 0.4$ for[3] CPXR(Log), and we used default settings of the R [16] packages for SLogR, SVM and RF. Running time and memory usage are given in this footnote[4].

A. Background information

The TBI dataset considered in this study[5] is from [17], which will be called TBI in this paper, on patients from an International and US Tirilazad trials. It contains 2159 instances and 15 predictor variables; its missing predictor variable values were treated using multiple imputation as suggested by [17].

The outcome variable of TBI is assessed with the Glasgow outcome scale (GOS), which has been widely used in brain injury studies. The scale ranges from dead (GOS 1), vegetative state (GOS 2), severe disability (GOS 3), moderate disability (GOS 4), to good recovery (GOS 5). The predictor variables belong to three groups: **Basic variables (4):** cause of injury, age of patient, GCS motor score, and pupil reactivity. **Computed-tomography variables (7):** hypoxia, hypotension, CT characteristics (Marshall CT classification), traumatic subarachnoid hemorrhage (tSAH), epidural hematoma (EDH),

[3] While CPXR(Log) is very similar to CPXR given in [5], there are some important technical differences, including the choice of ρ, used for splitting data into LE and SE. For linear regression, we found that [5] having ρ in the range of 0.45 and 0.65 usually yields the best PXR models. For logistic regression, it turns out that the best choice is around $\rho = 0.4$ for the TBI data for the three sets of predictor variables. (Different choices of ρ yielded AUCs as follows, for Mortality using (Basic+CT) variables (see below): $\rho = 0.25$:AUC= 0.84, 0.30:0.84, 0.35:0.88, 0.38:0.882, 0.4:0.88, 0.45:0.87, 0.55:0.85.) This might be caused by the difference between the distribution of logistic residuals (Pearson's residuals) and that of linear residuals.

[4] When building CPXR(Log) models for Mortality, on a Windows 7 machine with 2.90 GHz CPU and 8GB RAM, with CPXR(Log) coded in Java: 5 minutes and 4 MB of memory for the Basic variables, 20 minutes and 9 MB for Basic+CT, and 45 minutes and 17 MB for Basic+CT+Lab.

[5] Dr Steyerberg kindly shared this dataset with us. In general, datasets on traumatic brain injury are not publicly available and hence we are limited to this dataset in this paper. We plan to use our method to carry out prognostic modeling on other traumatic brain injury datasets if they become available.

compressed cistern at CT, and midline shift more than 5mm. **Lab variables (4):** glucose, ph, sodium and hb (hemoglobin) Details on these predictor variables can be found in [11], [17].

Prognostic models studied consist of all six prognostic models for the six combinations of two dichotomized versions of GOS and three subsets of variables. These combinations were examined in previous studies [11], [17] on TBI, allowing us to compare the performance of our method against [11], [17].

The two dichotomized versions of GOS are: mortality (versus survival) and unfavorable outcome (versus favorable outcome). For the first, called "Mortality", all cases with the dead outcome (GOS 1) belong to the "mortal" class and all others (GOS 2–5) are in the "survival" class. For the second, called "Unfavorable", all cases with dead, vegetative and severe disability outcomes (GOS 1–3) are in the "unfavorable" class, whereas all cases with moderate disability and good recovery outcomes (GOS 4–5) are in the "favorable" class.

The three variable sets considered are: Basic, Basic+CT, and Basic+CT+Lab, consisting 4, 11, 15 variables respectively.

Notations: We add qualifiers to the names of models to avoid confusion. Specifically, we will use Method-DichotomizedName-VariableSet as model names. For example, the SLogR-Mortality-(Basic+CT) model refers to the model built by SLogR for Mortality using the Basic+CT variables.

B. *Evaluation on prognostic model accuracy measures*

We now compare CPXR(Log) against SLogR on several measures concerning model accuracy. The results show that CPXR(Log) outperforms SLogR consistently and by big margins. The strong outperformance implies that TBI has diverse predictor-response relationships [5].

Tables I and II present[6] [7] [8] the performance of prognostic models built by SLogR and CPXR(Log). Figure 2 shows the ROC curves of all six models developed by SLogR and CPXR(Log); solid lines represent curves of SLogR and dashed lines represent curves for CPXR(Log). (Figure 3 compares ROC curves of CPXR(Log) against that of SVM and RF.)

Apparently, all six CPXR(Log) models outperformed corresponding SLogR models on all performance measures. In particular, AUC of all CPXR(Log) models improved that of SLogR models by 11.7% on average (all six models), and the improvement was 12.9% for Mortality models and 10.4% for Unfavorable models. Moreover, on average over all six models, specificity, sensitivity, accuracy, and χ^2 of the Basic+CT+Lab models built by CPXR(Log) improved over those of SLogR by 8%, 18%, 16% and 28% respectively. Interestingly, CPXR(Log) achieved more improvement over SLogR on χ^2 concerning Mortality models than Unfavorable models: The average improvement for Mortality models is 29.4% and it is 26.3% for Unfavorable models.

TABLE III: AUC improvement when more variables are used by CPXR(Log) and SLogR

	Mortality		Unfavorable	
Variable set change	CPXR(Log)	SLogR	CPXR(Log)	SLogR
Basic → Basic+CT	10%	7.7%	6%	5.2%
Basic+CT → Basic+CT+Lab	4.5%	2.5%	6.8%	1.25%
Basic → Basic+CT+Lab	15.0%	11.1%	13.4%	6.6%

TABLE IV: AUC improvement by CPXR(Log) over SLogR for given variable sets

Mortality			Unfavorable		
Basic	Basic+CT	Basic+CT+Lab	Basic	Basic+CT	Basic+CT+Lab
11.1%	12.8%	15%	7.9%	8.8%	14.8%

One main strength of CPXR(Log) is its ability to effectively utilize more variables to derive more accurate models, which is similar to CPXR for linear regression [5]. Indeed, while both CPXR(Log) and SLogR obtained improvement on AUC when more variables are used, CPXR(Log) obtained larger improvement in all cases (see Table III). Moreover, when more variables are used, CPXR(Log) achieved larger improvement on AUC over SLogR, as shown in Table IV. CPXR can also effectively extract useful information capturing interactions among multiple predictor variables that are often missed by standard logistic regression methods and other classification algorithms (see §V-C).

Interestingly, Figures 2 and 3 show that the ROC curves of CPXR(Log) always have larger true positive rate for every false positive rate, than that of SLogR, SVM, and RF.

Remark: For both SLogR and CPXR(Log), prognostic models for Mortality are more accurate than those for Unfavorable, suggesting that the unfavorable class is harder to model.

Overfitting: Overfitting is a major issue in clinical prediction modeling; it happens when a prediction model (or classifier) is much more accurate on training data than on test data that are unknown to the model. Overfitting models are not desirable, as end users including physicians cannot be very confident in using them to make predictions on new cases. When comparing prediction models, more accurate models are preferred; among equally accurate models, the less overfitting ones are preferred.

One method to evaluate prediction models with respect to overfitting is to examine how big is the drop on given goodness measures, from training data to test data. To that end, we split the TBI dataset into training (2/3) and test (1/3), for the Unfavorable dichotomized version using Basic+CT variables. We train prognostic models using CPXR(Log), SLogR, SVM and RF on the training data. Figure 3 shows the ROC curves and the AUCs of the models built by the mentioned algorithms on the test data. Interestingly, the AUCs of SVM, RF, CPXR(Log) and SLogR on the training data are 0.94, 0.92, 0.94 and 0.84 respectively. Hence, the AUCs of SVM, RF, CPXR(Log) and SLogR dropped, from training data to test data, by 25%, 22%, 8.5% and 5% respectively. CPXR(Log) and SLogR have much smaller drops than SVM and RF. Combined with the fact CPXR(Log) has much higher AUC and other measures concerning accuracy than the other three, CPXR(Log) is the winner on overfitting.

[6] The results reported here agree with those reported in [18]

[7] Specificity, sensitivity and accuracy are for the classifier using the logistic regression models with 0.5 as cutoff: If the predicted value is larger than 0.5 then the predicted class is 1, otherwise the predicted class is 0.

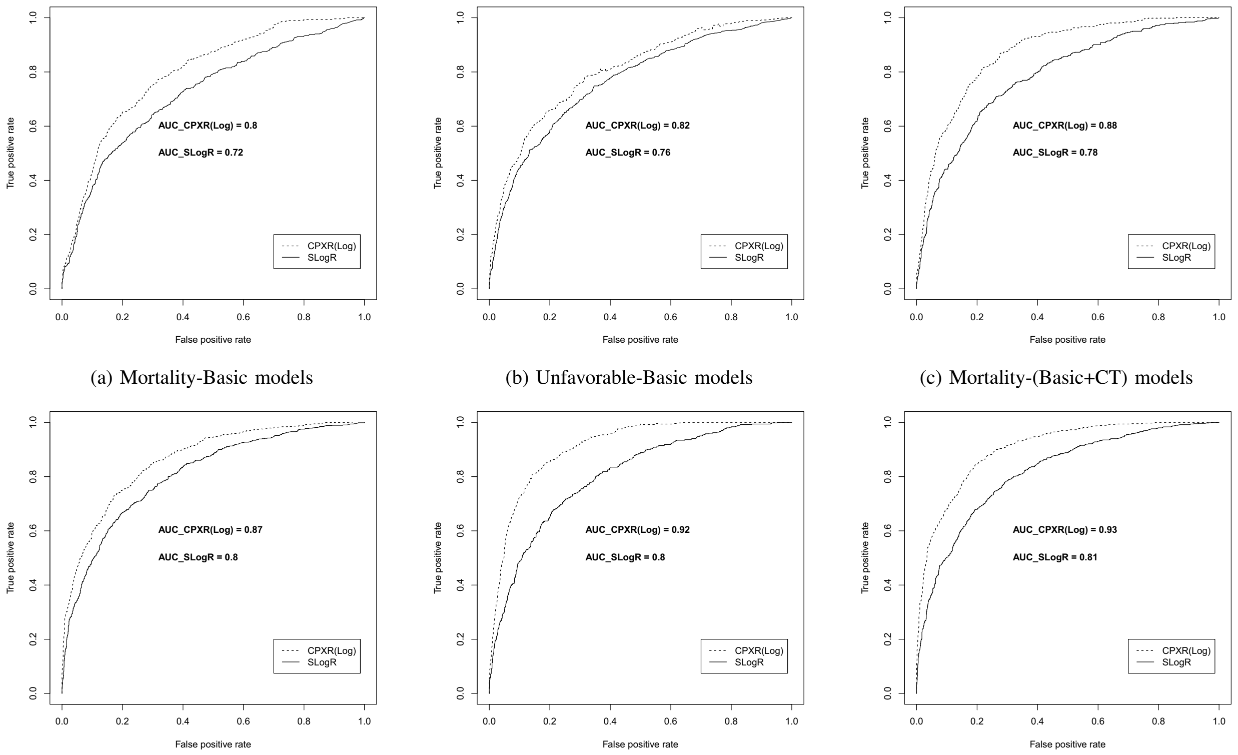
[8] Sensitivity is the proportion of actual positives which are correctly identified as such. Specificity is the proportion of negatives which are correctly identified as such. We omit definitions of other measures to save space.

TABLE I: SLogR performance on accuracy

	Mortality						Unfavorable					
Model	Specificity	Sensitivity	F_1	Accuracy	AUC	χ^2	Specificity	Sensitivity	F_1	Accuracy	AUC	χ^2
Basic	0.95	0.18	0.27	0.77	0.72	2192	0.85	0.52	0.59	0.72	0.76	2174
Basic+CT	0.95	0.32	0.42	0.8	0.78	2183	0.85	0.6	0.66	0.75	0.8	2172
Basic+CT+Lab	0.94	0.36	0.46	0.8	0.8	2094	0.84	0.61	0.66	0.75	0.81	2137

TABLE II: CPXR(Log) performance on accuracy

	Mortality						Unfavorable					
Model	Specificity	Sensitivity	F_1	Accuracy	AUC	χ^2	Specificity	Sensitivity	F_1	Accuracy	AUC	χ^2
Basic	0.96	0.18	0.28	0.78	0.8	1801	0.89	0.54	0.63	0.75	0.82	1848
Basic+CT	0.96	0.42	0.53	0.85	0.88	1483	0.87	0.65	0.7	0.79	0.87	1601
Basic+CT+Lab	0.97	0.46	0.58	0.89	0.92	1290	0.91	0.72	0.76	0.87	0.93	1327



(a) Mortality-Basic models

(b) Unfavorable-Basic models

(c) Mortality-(Basic+CT) models

(d) Unfavorable-(Basic+CT) models

(e) Mortality-(Basic+CT+Lab) models

(f) Unfavorable-(Basic+CT+Lab) models

Fig. 2: Comparison of CPXR(Log) and SLogR: ROC curves and AUC

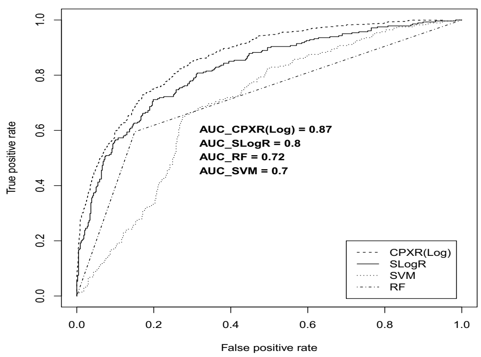


Fig. 3: ROC curves and AUC of CPXR(Log), SLogR, SVM, and RF on test data of models built from training data

C. The CPXR(Log)-Unfavorable-(Basic+CT+Lab) model and important predictor variables for TBI

Tables VI and VII give patterns and the local regression and baseline models of the CPXR(Log)-Unfavorable-(Basic+CT+Lab) model, respectively. Table V compares odds ratios[9] [10] of variables in the SLogR and CPXR(Log) models.

[9] Odds ratio (OR) is a popular measure to quantify how strongly the level of a predictor variable x_i is associated to the response variable [3]. The odds ratio of x_i is often estimated from a logistic regression model as $OR(x_i) = e^{\beta_i}$, where β_i is the coefficient of x_i in the logistic regression model.

[10] In logistic regression, we need to convert categorical variables into dummy variables. To avoid redundancy, one of the values of each categorical variable should be omitted ("treated as reference category"). We chose the most common value of each categorical variable as the reference category. A categorical variable for a local regression model becomes constant if the variable occurs in the pattern of the local model and is hence a constant. Reference categories and categorical variables involved in the patterns are both specified as "ref" in Tables V and VII.

Large differences in odds ratio can be of interest to physicians, as they indicate that for certain large population groups risk should be evaluated in a manner different from how risk is evaluated based on the SLogR model. Due to the popularity of SLogR, one can assume that physicians are familiar with SLogR models and they may have been using the information implied by such models in practice. Large difference can be for cases where odds ratio in CPXR(Log) models is significantly higher or lower than that in SLogR models.

There are quite a number of variable and value pairs where odds ratio differences are large. In Table V, we use the bold font to indicate such pairs where the odds ratio in the CPXR(Log) model is at least twice of that in the SLogR model, and we use the italic font to indicate cases where the odds ratio in the CPXR(Log) model is at most half of that in the SLogR model. We use underline to indicate some other cases where the odds ratio in the CPXR(Log) model is much larger than that in the SLogR model although not at least twice as much. To save space we omit rows having no large differences.

The largest odds ratio difference is 6.40 fold, for $6.79 < PH \leq 7.67$, whose odds ratio is 0.84 according to SLogR model and it is **5.38** for Model I of CPXR(Log). The largest odds ratio difference in absolute value is 7.07, for $reactivity = “No\ reactive''$, whose odds ratio is 2.66 according to SLogR model and it is **9.73** for Model II of CPXR(Log). The largest odds ratio decrease is 16.8 fold, for $6.79 \leq PH < 7.67$ whose odds ratio is 0.84 according to SLogR model and it is **0.05** for local Model VI of CPXR(Log).

TABLE VI: Pattern, arr, coverage of local models of CPXR(Log)-Unfavorable-(Basic+CT+Lab) model

Patterns	arr	Cov	Model
(CT classification = III)	15%	20%	I
(CT classification = V) AND (midline shift more than 5mm) AND ($0.56 <$ glucose ≤ 10.4)	12%	15%	II
(No compressed cistern) AND (No midline shift more than 5mm) AND ($7.22 <$ PH ≤ 7.45)	10%	40%	III
($10.77 <$ glucose ≤ 21.98)AND ($134 <$ sodium ≤ 144)	18%	18%	VI
(No Hypotension)AND ($134 <$ sodium ≤ 144) AND ($10.55 <$ HB ≤ 14.57) AND (With tSAH)	19%	20%	V
(No tSAH) AND ($134 <$ sodium ≤ 144) AND ($10.77 <$ glucose ≤ 21.98) AND (No Hypotension) AND (No midline shift) AND (One reactive pupil)	19%	20%	VI
(No tSAH) AND (One reactive pupil)	18%	40%	VII

Table VI presents[11] patterns used in CPXR(Log) model, together with their arr and support. (The support of the baseline model is 23%.) Multiple items in a pattern are joined by AND.

Important variables having large odds ratio changes or occurring in patterns: Our analysis with the CPXR(Log) model also shows that certain predictor variables are important. We found that all numerical variables with large ranges are important. Moreover, all variables occurring in patterns of the CPXR(Log) model are important. Finally, all categorical variables with large odds ratios in some local model or in the baseline model are important; the top six of such variables are (starting from the one with largest odds ratio):

> Pupillary reactivity, CT classification, PH, Cisterns compression, Cause, and Hypoxia.

An example where the CPXR(Log) model corrected a large prediction error: Consider this 15 year old patient with TBI due to a motorbike accident, who has the following characteristics at the admission time:

> GCS motor score = 5. No reactive pupil. No hypoxia. No hypotension. CT scan = V. No tSAH. Has epidural hematoma. Has fully compressed cisterns. Has midline shift. Glucose = 9.06 mmol/l. PH = 7.37. Sodium = 141 mmol/l. Hb = 14.4 g/dl.

Since the patient is young, having no evidence of hypoxia, hypotension and tSAH, the baseline regression model built by SLogR predicts that the patient's probability of survival is 0.78. However the observed outcome (after 6 month) is dead. In contrast, since this patient matches pattern II of the CPXR(Log) model, i.e., "(CT classification = V) AND (midline shift more than 5mm) AND ($0.56 <$ glucose $\leq$ 10.4)" is true, the CPXR(Log) model more accurately predicts probability of survival to be just 0.31, using local model II in Table VII associated with this pattern. ("CT scan = V" means "high lesion $>$ 25 mm and not surgically evacuated.") Incidently, in our analysis, we found that the SLogR model made many big prediction errors on young patients.

VI. Conclusion

We provided an effective new method, CPXR(Log) for logistic regression and for clinical predictive modeling. CPXR(Log) achieved much higher accuracy than standard logistic regression on traumatic brain injury (TBI). We also presented our CPXR(Log) model on TBI on admission time data, including patterns and local models, and presented new odds ratios of predictor variables based on CPXR(Log), including those whose odds ratio differ from the SLogR model based odds ratios. We hope that these findings will have significant value in accurate clinical prognostic decision making, including on TBI. In general, CPXR(Log) can effectively handle data with diverse predictor-response relationships.

Acknowledgement: We thank the anonymous reviewers for their very constructive comments.

References

[1] S.C. Bagley, W. Halbert, and B.A. Golomb: Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of clinical epidemiology*, 54(10): 979-985, 2010.

[2] A.W. Brown, J.F. Malec, R.L. McClelland, N.N. Diehl, J. Englander, and D.X. Cifu: Clinical elements that predict outcome after traumatic brain injury: a prospective multicenter recursive partitioning (decision-tree) analysis. *Journal of neurotrauma*, 22(10): 1040-1051, 2005.

[3] J. Cornfield: A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, 11(6): 1269-1275, 1951.

[4] G. Dong, J. Li: Efficient mining of emerging patterns: Discovering trends and differences. *In Proc. KDD*, 43-52, 1999.

[5] G. Dong, V. Taslimitehrani: Pattern-Aided Regression Modeling and Prediction Model Analysis. Under review at *IEEE Transactions on Knowledge and Data Engineering*.

[6] U. Fayyad, K. Irani: Multi-interval discretization of continuous-valued attributes for classification learning. *Proc. Int'l Joint Conf. on Uncertainty in AI*, 1993.

[7] T. Greenhalgh: How to read a paper: The basics of evidence-based medicine. *John Wiley & Sons*, 2010.

[8] C.W.P.M. Hukkelhoven, E.W. Steyerberg, A.J.J. Rampen, E. Farace, J.D.F. Habbema, L.F. Marshall, G.D. Murray, and A.I.R. Maas: Patient age and outcome following severe traumatic brain injury: an analysis of 5600 patients. *Journal of neurosurgery*, 99(4): 666-673, 2003.

[11] tSAH stands for Traumatic Subarachnoid Hemorrhage. Moreover, Cov denotes the support of patterns in the whole data set.

TABLE V: Odds ratios of predictor variables in the SLogR and CPXR(Log) models

Variables	Coding	Odds Ratios (95% CI)							
		SLogR	Model I	Model II	Model III	Model IV	Model V	Model VI	Model VII
Cause	Motorbike	0.87 (0.75-1.0)	0.85 (0.73-1.0)	**2.19** (1.8-2.6)	0.75 (0.65-0.85)	*0.36* (0.3-0.4)	0.5 (0.4-0.6)	1.08 (0.9-1.3)	1.2 (1.0-1.4)
	Assault	1.07 (0.9-1.2)	**2.86** (2.5-3.4)	0.94 (0.8-1.1)	0.69 (0.6-0.8)	0.62 (0.55-0.75)	*0.3* (0.25-0.35)	**2.94** (2.65-3.2)	**3.5** (3.0-4.0)
	Other	1.37 (1.1-1.6)	1.64 (1.5-1.8)	1.83 (1.6-2.1)	1.02 (0.8-1.2)	1.21 (1.0-1.4)	*0.47* (0.4-0.55)	1.72 (1.5-1.9)	1.35 (1.2-1.5)
Motor score	IV	0.37 (0.3-0.5)	*0.16* (0.1-0.2)	**0.77** (0.65-0.8)	0.47 (0.4-0.55)	0.32 (0.25-0.4)	0.31 (0.25-0.35)	0.48 (0.4-0.6)	0.58 (0.5-0.7)
	V/VI	0.23 (0.2-0.3)	*0.12* (0.1-0.15)	**0.55** (0.5-0.6)	0.22 (0.2-0.25)	0.18 (0.15-0.2)	*0.1* (0.05-0.15)	0.4 (0.35-0.45)	**0.46** (0.4-0.5)
Pupillary reactivity	No reactive	2.66 (2.3-3.0)	1.7 (1.5-1.9)	**9.73** (8.0-11.0)	2.37 (2.0-2.8)	1.69 (1.4-2.0)	2.87 (2.5-3.3)	1.0 (ref)	1.0 (ref)
Hypoxia	Yes	1.64 (1.45-1.8)	1.32 (1.1-1.5)	1.35 (1.2-1.5)	1.59 (1.4-1.8)	1.58 (1.4-1.8)	1.41 (1.2-1.6)	**3.18** (2.8-3.6)	1.58 (1.4-1.8)
Hypotens	Yes	1.19 (1.0-1.4)	2.25 (1.9-2.5)	1.0 (ref)	1.08 (0.9-1.3)	**2.44** (2.1-2.7)	1.0 (ref)	1.0 (ref)	1.18 (1.0-1.4)
CT classification	II	2.35 (2.0-2.7)	1.0 (ref)	1.0 (ref)	1.87 (1.6-2.2)	1.3 (1.1-1.5)	*0.33* (0.3-0.4)	1.79 (1.6-2.0)	2.23 (2.0-2.5)
	III	3.99 (3.5-4.5)	1.0 (ref)	1.0 (ref)	3.63 (3.1-4.1)	2.0 (1.8-2.2)	*0.6* (0.5-0.7)	7.11 (6.0-8.0)	4.32 (3.7-4.9)
	IV	3.74 (3.0-4.4)	1.0 (ref)	1.0 (ref)	5.05 (4.5-5.5)	*1.18* (1.0-1.4)	*0.71* (0.6-0.8)	3.7 (3.3-4.1)	*1.1* (0.9-1.3)
	V	4.72 (4.0-5.4)	1.0 (ref)	1.0 (ref)	2.79 (2.4-3.2)	*1.68* (1.5-1.9)	*0.68* (0.6-0.8)	6.6 (5.6-7.6)	4.02 (3.5-4.5)
	VI	5.04 (4.0-6.0)	1.0 (ref)	1.0 (ref)	3.94 (3.5-4.3)	*2.24* (2.0-2.4)	*0.8* (0.7-0.9)	4.63 (4.0-5.2)	4.76 (4.1-5.3)
Cisterns compression	Slightly	1.03 (0.9-1.1)	0.82 (0.75-0.9)	1.34 (1.2-1.5)	1.0 (ref)	0.86 (0.8-0.9)	*0.4* (0.35-0.45)	1.65 (1.45-1.85)	*0.57* (0.5-0.6)
	Fully	2.05 (1.7-2.3)	1.89 (1.7-2.1)	3.43 (3.0-4.0)	1.0 (ref)	**4.38** (3.8-5.0)	*0.75* (0.6-0.9)	2.53 (2.2-2.8)	2.09 (1.9-2.3)
Shift	Yes	1.03 (0.9-1.2)	1.18 (1.0-1.4)	1.04 (0.9-1.2)	1.0 (ref)	1.51 (1.3-1.7)	**2.49** (2.2-2.8)	1.0 (ref)	1.29 (1.1-1.5)
PH	6.79-7.67	0.84 (0.75-0.95)	**5.38** (5.0-5.8)	**4.45** (4.1-4.8)	*0.4* (0.35-0.45)	1.32 (1.2-1.45)	0.72 (0.6-0.8)	*0.05* (0.04-0.06)	*0.18* (0.15-0.21)

TABLE VII: The baseline model and local models of CPXR(Log) (dummy reference variables omitted)

Variables	Coding	Coefficients of variables in models)							
		Baseline	Model I	Model II	Model III	Model IV	Model V	Model VI	Model VII
Cause	Motorbike	-0.135	-0.161	0.784	-0.284	-1.013	-0.686	0.078	0.183
	Assault	0.07	1.051	-0.054	-0.358	-0.472	-1.199	1.08	1.253
	Domestic	-0.047	0.314	-0.318	0.193	-0.183	-0.143	-0.207	0.459
	Other	0.316	0.499	0.608	0.024	0.197	-0.735	0.543	0.306
Age, years	14-79	0.034	0.009	0.056	0.032	0.012	0.049	0.05	0.005
Motor score	III	-0.256	-0.654	0.244	0.214	0.281	-0.889	-0.902	0.297
	IV	-0.977	-1.832	-0.258	-0.752	-1.134	-1.142	-0.717	-0.531
	V/VI	-1.431	-2.094	-0.59	-1.512	-1.7	-2.238	-0.911	-0.758
Pupillary reactivity	One reactive	0.339	-0.194	0.24	0.096	0.022	-0.03	0 (ref)	0 (ref)
	No reactive	0.981	0.533	2.275	0.865	0.525	1.055	0 (ref)	0 (ref)
Hypoxia	Yes	0.494	0.279	0.306	0.465	0.459	0.344	1.158	0.46
Hypotens	Yes	0.18	0.814	0 (ref)	0.085	0.892	0 (ref)	0 (ref)	0.173
CT classification	II	0.855	0 (ref)	0 (ref)	0.63	0.264	-1.088	0.583	0.802
	III	1.384	0 (ref)	0 (ref)	1.289	0.697	-0.508	1.962	1.464
	IV	1.319	0 (ref)	0 (ref)	1.621	0.169	-0.335	14.923	0.101
	V	1.552	0 (ref)	0 (ref)	1.027	0.522	-0.376	1.89	1.392
	VI	1.619	0 (ref)	0 (ref)	1.372	0.81	-0.216	1.532	1.562
tSAH	Yes	0.658	0.792	0.02	0.82	0.892	0 (ref)	0 (ref)	0 (ref)
Epidural hematoma	Yes	-0.765	-0.164	-1.053	-0.62	-0.812	-0.377	-0.895	-0.776
Cisterns compression	Slightly	0.034	-0.195	0.294	0 (ref)	-0.145	-0.899	0.503	-0.552
	Fully	0.721	0.64	1.233	0 (ref)	1.477	-0.287	0.929	0.739
Shift	Yes	0.038	0.165	0.044	0 (ref)	0.415	0.915	0 (ref)	0.255
Glucose	0.56-41.4	0.075	0.142	0.197	0.026	-0.04	0.126	-0.012	0.102
PH	6.79-7.67	-0.169	1.68	1.494	-0.916	0.278	-0.318	-2.929	-1.661
Sodium	114-154	0.011	0.024	0.036	0.024	-0.05	-0.029	-0.04	0.063
HB	2.5-18.6	-0.112	-0.189	-0.111	-0.19	-0.134	-0.102	-0.161	-0.129

[9] B. Jennett, G. Teasdale, R. Braakman, J. Minderhoud, and R. Knill-Jones: Predicting outcome in individual patients after severe head injury. *The Lancet*, 307(7968): 1031-1034, 1976.

[10] R. Kennedy, L. Livingston, A. Riddick, J.H. Marwitz, J.S. Kreutzer, and N.D. Zasler: Evaluation of the Neurobehavioral Functioning Inventory as a depression screening tool after traumatic brain injury. *The Journal of head trauma rehabilitation*, 20(6): 512-526, 2005.

[11] MRC CRASH Trial Collaborators: Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *BMJ*, 336(7641): 425-9, 2008.

[12] G.D. Murray, I. Butcher, G.S. McHugh, J. Lu, N.A. Mushkudiani, A.I.R. Maas, A. Marmarou, and E.W. Steyerberg: Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study. *Journal of neurotrauma*, 24(2): 329-337, 2007.

[13] A. Maas, A. Marmarou, G.D. Murray, S.G.M. Teasdale, and E.W. Steyerberg: Prognosis and clinical trial design in traumatic brain injury: the IMPACT study. *Journal of neurotrauma*, 24(2): 232-238, 2007.

[14] P. Perel, P. Edwards, R. Wentz, and I. Roberts: Systematic review of prognostic models in traumatic brain injury. *BMC medical informatics and decision making*, 6(1): 1-10, 2006.

[15] P. Perel, J. Wasserberg, R.R. Ravi, H. Shakur, P. Edwards, and I. Roberts: Prognosis following head injury: a survey of doctors from developing and developed countries. *Journal of evaluation in clinical practice*, 13(3): 464-465, 2007.

[16] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Version 2.14.1 (2011-12-22).

[17] E. Steyerberg: Clinical prediction models. *Springer*, 2009.

[18] E. Steyerberg, H. Lingsma: Background information for TBI case study, Clinical prediction models *http://www.clinicalpredictionmodels.org*.

[19] E.W. Steyerberg, N. Mushkudiani, P. Perel, I. Butcher, J. Liu, G.S. McHugh, G.D. Murray, A. Marmarou, I. Robers, J.D.F. Habbema, and A.I.R. Maas: Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS medicine*, 5(8): e165, 2008.

[20] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M.J. Khoury: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making*, 10(1): 16, 2010.

[21] www.cdc.gov

[22] http://www.crash.lshtm.ac.uk/

[23] http://www.ncbi.nlm.nih.gov/pubmed/

[24] http://www.tbi-impact.org/