BMJ Open

Mortality prediction of motorcycle riders using machine learning models

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018252
Article Type:	Research
Date Submitted by the Author:	16-Jun-2017
Complete List of Authors:	Kuo, Pao-Jen; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Rau, Cheng-Shyuan; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Neurosurgery Chien, Peng-Chen; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Chen, Yi-Chun; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Hsieh, Hsiao-Yun; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Hsieh, Ching-Hua; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery
Primary Subject Heading:	Public health
Secondary Subject Heading:	Public health, Research methods
Keywords:	Motorcycle accident, mortality, machine learning (ML), logistic regression (LR), support vector machine (SVM), decision tree (DT)

Mortality prediction of motorcycle riders using machine learning models

Pao-Jen Kuo[1†], M.D.; Cheng-Shyuan Rau[2†], M.D.; Peng-Chen Chien[1], M.Sc.; Yi-Chun Chen[1], M.Sc.; Hsiao-Yun Hsieh[1], M.Sc.; Ching-Hua Hsieh[1*], M.D., Ph.D.

[†] Indicates equal contribution in authorship as the first author.

[1] Department of Plastic and Reconstructive Surgery, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan, 833

[2] Department of Neurosurgery, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan, 833

Pao-Jen Kuo; email: bow110470@gmail.com

Cheng-Shyuan Rau; e-mail: ersh2127@cloud.cgmh.org.tw

Peng-Chen Chien; e-mail: VENU_CHIEN@hotmail.com

Yi-Chun Chen; e-mail: libe320@yahoo.com.tw

Hsiao-Yun Hsieh; e-mail: sylvia19870714@hotmail.com

Ching-Hua Hsieh; e-mail: m93chinghua@gmail.com

Corresponding author: Ching-Hua Hsieh, M.D., PhD.

Department of Trauma Surgery & Plastic and Reconstructive Surgery, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan

No.123, Ta-Pei Road, Niao-Song District, Kaohsiung City 833, Taiwan

Tel: 886-7-7317123 ext: 8002; E-mail: m93chinghua@gmail.com

ABSTRACT

Objectives: We aimed to build and test models of machine learning (ML) to predict the mortality of hospitalized motorcycle riders.

Setting: A Level I trauma center in southern Taiwan.

Participants: The hospitalized motorcycle riders between January 2009 and December 2015 were allocated to be a training set (n=6,306) and a test set (n= 946). Using the demographics and injury characteristics as well as laboratory data of patients, logistic regression (LR), support vector machine (SVM), and decision tree (DT) were performed to determine the mortality of the individual motorcycle riders, under the different conditions of using all samples or reduced samples as well as using all variables or selected features into the algorithm.

Primary and secondary outcome measures: Model predictive performance was evaluated by accuracy, sensitivity, and specificity, and by the analysis of the area under curve of the receiver operator characteristic curves of the two different models.

Results: In the training set, both LR and SVM had a significantly higher AUC than that of DT, while there was no significant difference in the AUC of LR and SVM, regardless of using all samples or reduced samples as well as all variables or selected features. In the test set, SVM model for all samples with selected features presented a better model than all the other models, with an accuracy of 98.73%, sensitivity (86.96%), specificity (99.02%) and AUC of 0.9517 for mortality prediction.

Conclusion: We demonstrate that ML is able to provide a feasible level of accuracy for predicting the mortality of the motorcycle riders. The integration of ML model, particularly the SVM algorithm in trauma system may help identify high-risk patients and therefore drive the appropriate response by the clinical staff.

Trial registration: Approval number 201600653B0 by the institutional review board

(IRB) of the hospital.

KEY WORDS: Motorcycle accident; mortality; machine learning (ML); logistic regression (LR), support vector machine (SVM), and decision tree (DT)

ARTICLE SUMMARY

STRENGTHS AND LIMITATIONS OF THIS STUDY

- This study demonstrates the feasibility of using support vector machine (SVM) classification, one of machine learning models, to predict the mortality risk for motorcycle riders.
- With addition of more data in the model, the SVM model has the potential to get an increased predictive power and facilitate its clinical implement.
- The SVM model generally works like a black-box and cannot identify the relationships between mortality and the various explanatory variables.
- The incomplete records of patients and the exclusion of patients declared dead from the Trauma Registry System could bias the results.

Background

As a less expensive and convenient means of transportation, motorcycle use is popular in many cities. However, despite being a small fraction of the travel, motorcycle riders involved in road traffic accidents often sustain severe morbidity and mortality. Compared to the occupants in a motor vehicle, motorcycle riders are 8 times more likely to be injured per vehicle mile [1], 30 times more likely to die in a motor vehicle crash [2], and 58 times more likely to be killed on a per-trip basis [3]. In Taiwan, motorcyclist fatalities account for nearly 60% of all driving fatalities [4]. The fatalities are often associated with men, advanced age, not wearing a helmet, unlicensed status, and riding under the influence of alcohol [5-9]. In addition, head injuries were the major factor leading to mortality, followed by thoracic and abdominal injuries [6-9].

Identifying patients with high risk of mortality is vital for the integration of trauma management to maximize resources and quality of care delivered [10 11]. More robust and accurate individual predictions of mortality from better models might give clinicians better information about the likelihood of good or poor outcomes and improve individual trauma and mortality management [12]. To identify the possibility of mortality, a frequently used model is the Trauma and Injury Severity Score (TRISS) [13], which gives a probability of death based on logistic regression (LR) with variables including age, anatomical variable (Injury Severity Score [ISS]), physiological variable (Revised Trauma Score [RTS]), and different coefficients for blunt and penetrating injuries. However, TRISS is imperfect and fails to determine a correct classification in 15-30% of the trauma patients [14]. Even after the incorporation of other or revised predictors, like blood pressure [15], co-morbidities, and separate categories for different age-groups [16] into this model, the addition of more predictors to the basic TRISS model did not always result in higher performance [13 17 18]. Although the revised TRISS, resulting from the USA National Trauma Database is inaccurate for trauma systems, particularly in the management of predominantly blunt injuries [19], the further development of the model based on advanced methodological quality, the performance of the model in subsets of patient groups, and practical application is mandatory in the prediction of mortality [13].

Currently, machine learning (ML) had been successfully applied in the real world in many fields including automatic medical diagnostics and personalized health care [20-22]. There is an increasing interest in the application of supervised ML methods to aid diagnosis and prognosis in trauma patients. ML is based on the way the human brain approaches pattern recognition tasks, providing an artificial intelligence-based approach to solve classification problems and improving their efficiency and effectiveness over time [23]. The usefulness of ML is bolstered by the versatility of its techniques and its utility for artificial intelligence such as prediction, classification, planning, recognition, and clustering [23 24]. Comparisons of different learning strategies have been conducted previously by others using field-specific datasets, many of which have shown significantly better predictive power than the more conventional alternatives [25]. Examples of multivariate techniques for pattern recognition include, but are not limited to, LR, support vector machine (SVM), decision trees (DT), and artificial neural networks. LR is a widely used and accepted statistical analysis tool to predict the probability of the occurrence of an event [26]. It attempts to build a functional relationship between two or more independent predictors and the one dependent outcome variable, under the assumption that the response variables are linearly related to the coefficients of the predictor variables [26].

SVM uses a training set of data composed of one or more features to determine an optimal boundary separating a set of cases. The binary SVM classifier constructs a set of the optimal hyperplanes in high-dimensional space with the maximal margin of the two classes [27]. In the case that all training points cannot be separated by the hyperplane, a soft margin method is used to construct a hyperplane that separates the training data points [28 29]. It has been found that the SVM model has a great capability of dealing with classification problems [30-34].

A DT is a hierarchical model composed of decision rules based upon optimal feature cutoff values that recursively split independent variables into different groups [35-37]. The purpose of DT building is to search for a set of decision rules to predict an outcome from a set of input variables [33 35 36]. Some models are used to construct decision-tree models, including classification and regression trees (CART), ID3s, chi-square automatic interaction detector DTs (CHAIDs), and C4.5 and C5.0 DTs [26, 28]. Among these methods, the CART analysis is a combined approach based on nonparametric and nonlinear variables for recursive partitioning analysis. CART analysis is an innovative DT model in which several predictive variables are crucial to identify patients at different levels of risk in various medical fields through progressive binary splits to develop prediction models in order to enable better prediction and clinical decision-making [38-40].

This study is aimed to construct a model for the mortality prediction of motorcycle riders using ML algorithms and obtaining data from a population-based trauma registry in a level I trauma center.

METHODS

Ethics statement

This study was preapproved by the institutional review board (IRB) of Chang Gung Memorial Hospital with approval number 201600653B0. Informed consent was waived according to the IRB regulations.

Data preparation

Detailed patient information between January 2009 and December 2015 was retrieved from the Trauma Registry System of our institution, a 2,400-bed facility and Level I regional trauma center. Only the trauma patients who sustained a traffic accident as a motorcycle rider and were hospitalized for treatment were included in the study. The patient information included the following variables: age, sex, helmet-wearing status, co-morbidities such as coronary artery disease (CAD), congestive heart failure (CHF), cerebral vascular accident (CVA), diabetes mellitus (DM), end-stage renal disease (ESRD), and hypertension (HTN) as well as vital signs, including temperature, systolic blood pressure (SBP), heart rate (HR), respiratory rate (RR), ISS, Glasgow coma scale (GCS) score, abbreviated injury scale (AIS) in different regions of the body, number of injured body regions according to AIS (number of AIS locations), the in-hospital mortality, the blood level of white blood cell count (WBC), red blood cell count (RBC), hemoglobin (Hb), hematocrit (Hct), platelets, blood urine nitrogen (BUN), creatinine (Cr), alanine aminotransferase (ALT), aspartate aminotransferase (AST), sodium (Na), potassium (K), blood alcohol concentration (BAC), and glucose at emergency department.

These enrolled patients were divided into a training sample, which was used for predictor discovery and supervised classification to generate a plausible model, and a test sample, which was used to test the performance of the model generated in the training sample. Those patients with missing data were not included for further analysis. The patients who registered in a six-year span between January 2009 and December 2014 were allocated in the training set, which comprised of a total of 6,306 patients. It included 6,161 survival and 145 mortality patients. In the test set, there were 946 patients, including 923 survival and 23 mortality patients, from the one-year span between January 2015 and December 2015. The sample similarity was assessed by Euclidean distance for quantitative data to reduce the size of a sample designed for use in data analysis [41]. The sample reduction used Euclidean distance of the dist function in the stats package in R (R Foundation for Statistical Computing, Vienna, Austria). During sample reduction, the data size can be reduced to speed up calculations in the analysis [42]. However, considering the exploratory character of this study, all samples (n=6,306) and reduced samples (n=1,510) in the training set of this study would require to be analyzed in ML classification.

ML classifiers

This work provides a performance comparison of three different ML classifiers (LR, SVM, and DT).

Logistic regression

The LR classifier used the glm function in the stats package in R3.3.3 (R

Foundation for Statistical Computing, Vienna, Austria). Univariate LR analyses were initially performed to identify the significant predictor variables of the mortality risk. Stepwise LR analysis was used to control the effects of confounding variables to identify independent risk factors for mortality. The selected independent risk factors obtained from LR were also used as selected features to be implemented by the SVM and the DT to explain their weights in determining the risk of mortality.

Support vector machine

The SVM classifier used the tune.svm & svm function in the e1071 package in R. In the training set, the SVM classifier was performed for the prediction of mortality with regard to either all 32 variables or 12 selected features as well as all the samples and reduced samples in the training set. The mapping procedure was accomplished by the kernel function, which is a matrix of pair-wise similarities between data points, such as a linear, polynomial, or radial basis function (RBF) [43]. For this study, the RBF kernel was chosen because it can handle non-linear interactions between class labels and features [44]. The two main parameters presented in SVM with RBF kernel were the penalty parameter C and the kernel hyper-parameter γ. The penalty parameter C determined the tradeoff between the fitting error minimization and model complexity, while the hyper-parameter γ defined the nonlinear feature transformation onto a higher dimensional space and controlled the tradeoff between error due to bias and variance in the model. [45]. The optimal operating point was estimated by varying the parameters - C and γ using a grid search for each combination of feature selection and dimension reduction with a 10-fold cross-validation [44].

Decision tree

The DT by CART based on the Gini impurity index used the rpart function in the rpart package in R. The CART analysis searched for the split on the variable that would partition the data into two different groups—a group of mostly ‘0s’ (people who survived) and a group of mostly ‘1s’ (people who died) [46 47]. Using the best overall split, the CART model partitioned the data and assigned a predicted class to each subgroup. CART repeated this same process on each predictor in the model, identifying the best split by iteratively testing all possible splits, and producing the greatest reduction in impurity [38-40]. CART proceeded recursively in this way until the specified stopping criteria were reached, a specified number of nodes were created, or a further reduction in node impurity became impossible [38-40].

Performance evaluation

We used receiver operator characteristic (ROC) curve analysis to assess and compare the performances of the individual ML models. Model predictive ability was evaluated using confusion matrix and the area under curve (AUC) analysis between two approaches of ML models.

Confusion matrix

The confusion matrix calculates the accuracy, sensitivity, and specificity of a given model with true negative, true positive, false positive, and false negative values and presents as a result an accuracy, which represents the overall proportion of correct classifications; a sensitivity, which refers to the proportion of true positives correctly identified (e.g. percentage of people with fatality identified to be dead); and a specificity, which refers to the proportion of true negatives correctly identified (e.g. percentage of people who survived identified as not dead).

AUC analysis

In order to compare the performance of multiple ML classifiers in multiple training data sets, a nonparametric approach to the analysis of areas under correlated ROC curves using the roc & roc.test function in the pROC package in R is pursued. This nonparametric approach takes into account the correlated nature of the data that two or more empirical curves are constructed based on tests performed on the same individuals [48].

All statistical analyses were performed using SPSS 20.0 (IBM Inc., Chicago, IL, USA) and R 3.3.3. For categorical variables, Chi-square tests were used to determine the significance of the association between the predictor and outcome variables. For continuous variables, student t-tests were applied to analyze normally distributed data, while Kolmogorov-Smirnov tests or Mann-Whitney U tests were used to compare non-normally distributed data. All of the results were presented in the form of the mean ± standard deviation. A p-value < 0.05 was considered statistically significant.

RESULTS

Demographics and injury characteristics of the patients

The patients with fatality had a higher AIS score at the head and neck region but lower AIS score at the extremities compared to the patients who survived (Figure 1). The patients with fatality had sustained more number of injured body regions (number of AIS locations) than the ones who survived. In addition, the patients with fatality comprised more of females and fewer of them were observed to be wearing a helmet compared to the patients who survived (Figure 1). A statistically significant difference in age, ISS, GCS, glucose, temperature, Hb, Hct, platelets, K, Cr, AST, ALT, and incidences of CAD was found between patients with fatality and the ones who survived respectively (Figure 2). Because the distribution pattern between Hb and Hct as well as between AST and ALT is very similar, only one of these two variables (i.e. Hct and AST) was selected for further ML classification to prevent the inclusion of duplicate parameters. Therefore, a total of 32 variables were used for imputation into ML classifiers as all variables, in contrast to considering selected features obtained by using the independent risk factors identified by the LR given below.

Performance of ML classifiers in training set

Logistic regression

LR identified 12 predictors (platelets, glucose, BUN, Cr, AST, Na, Age, GCS, temperature, number of AIS locations, ISS, and HTN) as independent risk factors for mortality in motorcycle riders from either all samples or the reduced samples.

The predictive models were listed as:

All samples (n=6,306)

$$Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 4.71648 - 0.00846 * \text{Platelets} + 0.01189 * \text{Glucose} + 0.03459 * \text{BUN} + 0.10667 * \text{Cr} + 0.00195 * \text{AST} + 0.09513 * \text{Na} + 0.02533 * \text{Age} - 0.39968 * \text{GCS} - 0.56396 * \text{Temperature} - 0.93232 * \text{Number of AIS locations} + 0.14098 * \text{ISS} - 0.95726 * \text{HTN}$$

Reduced samples (n=1,510)

$$Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 5.76780 - 0.00763 * \text{Platelets} + 0.00953 * \text{Glucose} +$$

$$0.03773 * \mathrm{BUN} + 0.00152 * \mathrm{AST} + 0.08630 * \mathrm{Na} + 0.02014 * \mathrm{Age} - 0.34116 * \mathrm{GCS} - 0.53370 * \mathrm{Temperature} - 0.91439 * \text{Number of AIS locations} + 0.12191 * \mathrm{ISS} - 1.00522 * \mathrm{HTN}$$

The LR achieved an accuracy of 98.64% (sensitivity of 59.31% and specificity of 99.56%) and 94.44% (sensitivity of 60.00% and specificity of 98.10%) for all samples and reduced samples, respectively. The AUCs for all samples and reduced sample were 0.9528 and 0.9524, respectively (Table 1).

Support vector machine

In the training set, the SVM classifier was performed for the prediction of mortality taking input as either all 32 variables or the 12 selected features in all samples and reduced samples, respectively. With the RBF as the kernel function, the SVM model has two parameters (C, γ) that need to be determined. The accuracy was highly robust to small changes in the hyper-parameters, so reasonable choices were obtained by a grid search of 2^x where x is an integer between -8 and 4 for C and between -10 and -2 for γ. The values which gave the highest 10-fold cross-validation accuracy are reported to be C = 0.25 and γ= 0.00390625. Under the input of all variables into the model, the SVM achieved an accuracy of 98.62% (sensitivity of 62.07% and specificity of 99.48%) and 94.37% (sensitivity of 59.31% and specificity of 98.10%) for all samples and reduced samples, respectively (Table 1). The AUCs for all samples and reduced sample were 0.9534 and 0.9526, respectively (Figure 3). With selected features in the model, the SVM achieved an accuracy of 98.62% (sensitivity of 64.14% and specificity of 99.43%) and 93.84% (sensitivity of 62.76% and specificity of 97.14%) (Table 1) as well as 0.9517 and 0.9518 AUCs (Figure 3) for all samples and reduced samples, respectively were (Table 1):

Decision tree

As shown in Figure 4, in the DT model, GCS was identified as the variable of initial split with an optimal cut-o value of > 3. Among patients with GCS higher than 3, glucose was selected as the variable of second split at a discrimination level of 180 and 177 mg/dL for all samples and reduced samples, respectively. After the glucose level < 180 or 177 mg/dL for all samples and reduced samples, respectively, the next best predictor of mortality was platelets with an optimal cut-off of $201X10^3$ /µL. For the node, with patents having a GCS not greater than 3, ISS < 24 and glucose < 218 mg/dL, these predictors were selected as significant variables for all samples and reduced samples, with GCS > 8, glucose < 198 mg/dL, and the number of AIS locations ≥ 3 being an additional predictors for splitting for the reduced samples. With all variables in the model, the DT achieved an accuracy of 98.92% (sensitivity of 62.76% and specificity of 99.77%) and 95.83% (sensitivity of 68.97% and specificity of 98.68%) for the all samples and reduced samples, respectively. The AUCs for all samples and reduced samples were 0.8872 and 0.9289, respectively. With selected features in the model, the DT achieved an accuracy of 98.92% (sensitivity of 64.14% and specificity of 99.74%) and 95.83% (sensitivity of 70.34% and specificity of 98.53%) for the all samples and reduced samples, respectively. The AUCs for all samples and reduced samples were 0.8872 and 0.9289, respectively (Figure 3). In the condition of using reduced samples but not all samples in the DT model, the number of AIS locations would be added in the split of the node slightly increasing the sensitivity from 62.76% to 68.97% and from 64.14% to 70.34% with input comprising of all variables and selected variables, respectively. In addition, in the condition of using selected features but not all variables in the DT model, the level of K was not used in the splitting of the node and was substituted by the cut-off value of AST ≥ 104 IU/L, slightly increasing the sensitivity from 62.76% to 64.14% and from 68.97% to 70.34% with input as all samples and reduced samples, respectively. In addition, the AUCs for all samples and reduced sample were 0.8875 and 0.9292, respectively (Figure 3)

Comparison in AUC analysis

In the comparisons of AUCs for LR, SVM, and DT for the training set (Table 2), both LR and SVM had a significantly higher AUC than DT, regardless of using all samples or reduced samples as well as for all variables or for selected features. However, there was no significant difference of AUC between LR and SVM, regardless of using all samples or reduced samples as well as for all variables or for selected features. In addition to this, in DT sample reduction had a significantly higher AUC than the one obtained using all samples, but there was no significant difference of AUC between DT with all variables or with selected features.

Performance of ML classifiers in test set

In test set, the LR model for all samples and reduced samples - both achieved an accuracy of 98.41%, with a sensitivity of 73.91% and specificity of 99.02% in predicting the mortality (Table 1). All of these four SVM models create an accuracy more than 98% and a specificity near 99% but a sensitivity of 69.57%, 86.96%, 69.57%, and 73.91% for all samples and all variables, all samples and selected features, reduced samples and all variables, and for reduced samples and selected features, respectively, whereas all of these four DT models create an accuracy of approximately 98% and a specificity of approximately 99% but a sensitivity of less than 70%. Considering that the majority of patients survived except for a few with fatality, would result into a very high accuracy and specificity index in predicting the mortality, therefore the comparison should further focus on the sensitivity of different ML models. We found that, in the test, all LR and SVM models, but not the DT models, had an increased sensitivity than that in the test set. Furthermore, the SVM model for all samples with selected features had a significantly highest sensitivity (86.96%) in predicting the mortality.

Discussion

LR is widely used in epidemiological studies for causal inference and, with the selection of built-in features; it does not necessarily utilize all the predictors. With a relatively limited number of variables i.e. variables less than 20, LR provides estimates of odd ratios of the risk factors [49]. However, its limitations become apparent when analyzing a complex dataset with a high number of relevant exposures and multiple interactions [50]. With too many predictors, the availability of sufficient information to specify all interactions would become nearly impossible [50]. In addition, the DT with CART analysis is exploratory and not based on the probabilistic method, which may lead to overestimating the importance of included risk factors or cause missing of other potential confounders that could influence each patient's actual risk [51]. In contrast to LR, which is heavily influenced by outliers in its linear discriminant analysis method, the SVM boundary is only minimally influenced by outliers that are difficult to separate, despite the complexity of data [52]. In addition, employing kernels in the SVM would be advantageous to learn non-linear decision boundaries allowing the classifier to solve more difficult classification problems than the linear analysis method [53]. These three ML models (LR, SVM, DT) all create an accuracy and a specificity around 98% and 99%, respectively, but a sensitivity less than or around 70% in the training dataset. In this study, both LR and SVM resulted in a significantly higher AUC than DT in the training set, regardless of using all samples or reduced samples as well as for all variables or for selected features.

This study included different variants of SVM considering the sample size and feature selection to show all possible improvements to the more conventional strategies like LR or DT. Although sample reduction for SVM had been proposed to greatly improve the training speed of the SVM and save a lot of storage space [54 55], using the kernel is a more efficient technique in case of similarity of representation between samples. Thus, the computational complexity of SVM is not wholly governed by the number of samples, but by the number of features, which is advantageous for analysis in the high-dimensional settings [53]. In addition, for SVM, feature selection based on recursive feature elimination was performed with addition and/or removal of predictors to determine the optimal combination that would maximize the AUC [25]. Aided by feature selection, the proposed SVM method identifies the most discriminating indexes for mortality prediction. We found that, although both LR and SVM did not have a different AUC in the training procedure, the SVM model for all samples with selected features had a significantly higher sensitivity (86.96%) in predicting the mortality of motorcycle rides in the test set compared to the rest of the models. The increased sensitivity of SVM in test set than that in training set may be attributed to an improved quality of registered content and less missing data in our registered data after continuous quality assessment and years of experience of working with the registers. Such increased sensitivity was also found in the LR model in the test set. With addition of more data in the model, the SVM model has the potential to get an increased predictive power. This study demonstrates the feasibility of using SVM classification with feature selection to predict the mortality risk for motorcycle riders in the trauma care. However, the SVM model generally works like a black-box and cannot identify the relationships between mortality and the various explanatory variables and therefore, cannot be directly used to validate our hypothesis of increased sensitivity in the test set.

There are several limitations to this study. Firstly, the patients who had incomplete records were excluded from the analysis. This could have caused the results to be biased and the results could have been different from the data acquired by including the patients with incomplete records and replacing the missing data on a variable by a value that is drawn from an estimate of the distribution of this variable [56-58]. The benefit of imputation is that we would be able to include patients who might have relevant features for analysis, but were excluded owing to errors in data collection or recording [56-58]. Secondly, a source of potential bias may come from the exclusion of patients declared dead (either on arriving at the hospital or at the accident spot itself) and injured patients who were discharged against the advice of the emergency department. Thirdly, there was lack of important data regarding injury mechanism and circumstance, including motorcycle speed and type, helmet material, and impact force during collision. In addition, the imputation of physiological and laboratory data collected from the time of arriving at the emergency department cannot reflect the dynamic changes in hemodynamic and metabolic variables of the trauma patients under a possible resuscitation procedure. Finally, the study population was limited to a single urban trauma center in southern Taiwan, which may not be representative of other populations.

Conclusion

We demonstrate that ML is able to provide a feasible level of accuracy for predicting mortality of the motorcycle riders. Whilst there are significant theoretical and practical challenges to the translational implementation of this approach, the results of the studies published so far are encouraging and may provide the first steps towards the development of a prediction model integrated into the trauma care system in order to identify an individual motorcycle rider's risk of mortality.

Competing interests

The authors declare that they have no competing interests.

Author contributions

PJK wrote the manuscript; CSR analyzed the tables; PCC performed the statistical analyses and ML programming; YCC and HYH collected the data and are responsible for the integrity of the registered data; and CHH designed the study and contributed to the analysis and interpretation of data. All authors have read and approved the final manuscript.

Data Sharing

No additional data are available.

Acknowledgements

This research was supported by a grant from Chang Gung Memorial Hospital

CMRPG8F0891. We appreciate the Biostatistics Center, Kaohsiung Chang Gung Memorial Hospital for helping us with the statistical work.

References

1. Weiss H, Agimi Y, Steiner C. Youth motorcycle-related brain injury by state helmet law type: United States, 2005-2007. Pediatrics 2010;**126**(6):1149-55.

2. National Highway Traffic Safety Administration (NHTSA) TSFDDH.

3. Beck LF, Dellinger AM, O'Neil ME. Motor vehicle crash injury rates by mode of travel, United States: using exposure-based methods to quantify differences. American journal of epidemiology 2007;**166**(2):212-8.

4. Chang HL, Lai CY. Using travel socialization and underlying motivations to better understand motorcycle usage in Taiwan. Accident; analysis and prevention 2015;**79**:212-20.

5. Jou RC, Yeh TH, Chen RS. Risk factors in motorcyclist fatalities in Taiwan. Traffic injury prevention 2012;**13**(2):155-62.

6. Liang CC, Liu HT, Rau CS, et al. Motorcycle-related hospitalization of adolescents in a Level I trauma center in southern Taiwan: a cross-sectional study. BMC pediatrics 2015;**15**:105.

7. Liu HT, Liang CC, Rau CS, et al. Alcohol-related hospitalizations of adult motorcycle riders. World journal of emergency surgery : WJES 2015;**10**(1):2.

8. Hsieh CH, Hsu SY, Hsieh HY, et al. Differences between the sexes in motorcycle-related injuries and fatalities at a Taiwanese level I trauma center. Biomed J 2017;**40**(2):113-20.

9. Hsieh CH, Liu HT, Hsu SY, et al. Motorcycle-related hospitalizations of the elderly. Biomed J 2017;**40**(2):121-28.

10. Densmore JC, Lim HJ, Oldham KT, et al. Outcomes and delivery of care in pediatric injury. Journal of pediatric surgery 2006;**41**(1):92-8; discussion 92-8.

11. Rogers SC, Campbell BT, Saleheen H, et al. Using trauma registry data to guide injury prevention program activities. The Journal of trauma 2010;**69**(4 Suppl):S209-13.

12. Norrie J. Mortality prediction in ICU: a methodological advance. The Lancet Respiratory medicine 2015;**3**(1):5-6.

13. de Munter L, Polinder S, Lansink KW, et al. Mortality prediction models in the general trauma population: A systematic review. Injury 2017;**48**(2):221-29.

14. Demetriades D, Chan L, Velmanos GV, et al. TRISS methodology: an inappropriate tool for comparing outcomes between trauma centers. J Am Coll Surg 2001;**193**(3):250-4.

15. Jones JM, Skaga NO, Sovik S, et al. Norwegian survival prediction model in trauma: modelling effects of anatomic injury, acute physiology, age, and

co-morbidity. Acta Anaesthesiol Scand 2014;**58**(3):303-15.

16. Bergeron E, Rossignol M, Osler T, et al. Improving the TRISS methodology by restructuring age categories and adding comorbidities. J Trauma 2004;**56**(4):760-7.
17. Fueglistaler P, Amsler F, Schuepp M, et al. Prognostic value of Sequential Organ Failure Assessment and Simplified Acute Physiology II Score compared with trauma scores in the outcome of multiple-trauma patients. Am J Surg 2010;**200**(2):204-14.
18. Kroezen F, Bijlsma TS, Liem MS, et al. Base deficit-based predictive modeling of outcome in trauma patients admitted to intensive care units in Dutch trauma centers. J Trauma 2007;**63**(4):908-13.
19. Stoica B, Paun S, Tanase I, et al. Probability of Survival Scores in Different Trauma Registries: A Systematic Review. Chirurgia (Bucur) 2016;**111**(2):115-9.
20. Cohen AM, Ambert K, McDonagh M. A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. AMIA Annual Symposium proceedings AMIA Symposium 2010;**2010**:121-5.
21. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic

challenges. Eur Heart J 2016.

22. Szlosek DA, Ferrett J. Using Machine Learning and Natural Language Processing Algorithms to Automate the Evaluation of Clinical Decision Support in Electronic Medical Record Systems. EGEMS (Washington, DC) 2016;**4**(3):1222.

23. Oquendo MA, Baca-Garcia E, Artes-Rodriguez A, et al. Machine learning and data mining: strategies for hypothesis generation. Mol Psychiatry 2012;**17**(10):956-9.

24. Kotoku J. An Introduction to Machine Learning. Igaku butsuri : Nihon Igaku Butsuri Gakkai kikanshi = Japanese journal of medical physics : an official journal of Japan Society of Medical Physics 2016;**36**(1):18-22.

25. Yahya N, Ebert MA, Bulsara M, et al. Statistical-learning strategies generate only modestly performing predictive models for urinary symptoms following external beam radiotherapy of the prostate: A comparison of conventional and machine-learning methods. Med Phys 2016;**43**(5):2040.

26. Siuly, Yin X, Hadjiloucas S, et al. Classification of THz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers. Comput Methods Programs Biomed 2016;**127**:64-82.

27. V. V. Statistical learning theory. New York, NY: John Wiley & Sons 1998.

28. de Boves Harrington P. Support Vector Machine Classification Trees. Anal Chem

2015;**87**(21):11065-71.

29. Lee Y. Support vector machines for classification: a statistical portrait. Methods Mol Biol 2010;**620**:347-68.

30. Chen C, Zhang G, Qian Z, et al. Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accid Anal Prev 2016;**90**:128-39.

31. Galatzer-Levy IR, Karstoft KI, Statnikov A, et al. Quantitative forecasting of PTSD from early trauma responses: a Machine Learning application. J Psychiatr Res 2014;**59**:68-76.

32. Li Z, Liu P, Wang W, et al. Using support vector machine models for crash injury severity analysis. Accid Anal Prev 2012;**45**:478-86.

33. Marucci-Wellman HR, Corns HL, Lehto MR. Classifying injury narratives of large administrative databases for surveillance-A practical approach combining machine learning ensembles and human review. Accid Anal Prev 2017;**98**:359-71.

34. Patil BM, Joshi RC, Toshniwal D, et al. A new approach: role of data mining in prediction of survival of burn patients. J Med Syst 2011;**35**(6):1531-42.

35. Farion K, Michalowski W, Wilk S, et al. A tree-based decision model to support prediction of the severity of asthma exacerbations in children. J Med Syst

2010;**34**(4):551-62.

36. Zintzaras E, Bai M, Douligeris C, et al. A tree-based decision rule for identifying profile groups of cases without predefined classes: application in diffuse large B-cell lymphomas. Comput Biol Med 2007;**37**(5):637-41.

37. Kasbekar PU, Goel P, Jadhav SP. A Decision Tree Analysis of Diabetic Foot Amputation Risk in Indian Patients. Frontiers in endocrinology 2017;**8**:25.

38. Guilbault RWR, Ohlsson MA, Afonso AM, et al. External Validation of Two Classification and Regression Tree Models to Predict the Outcome of Inpatient Cardiopulmonary Resuscitation. J Viral Hepat 2017;**32**(5):333-38.

39. Shi KQ, Zhou YY, Yan HD, et al. Classification and regression tree analysis of acute-on-chronic hepatitis B liver failure: Seeing the forest for the trees. 2017;**24**(2):132-40.

40. Zimmerman RK, Balasubramani GK, Nowalk MP, et al. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. BMC Infect Dis 2016;**16**(1):503.

41. Amaratunga D, Cabrera J, Lee YS. Resampling-based similarity measures for high-dimensional data. J Comput Biol 2015;**22**(1):54-62.

42. Bhattacharya S, Mariani TJ. Transformation of expression intensities across generations of Affymetrix microarrays using sequence matching and regression

modeling. Nucleic Acids Res 2005;**33**(18):e157.

43. Vapnik VN. The Nature of Statistical Learning Theory. New York, 2nd ed. 2000.

44. Gultepe E, Green JP, Nguyen H, et al. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. J Am Med Inform Assoc 2014;**21**(2):315-25.

45. Chen H, Hu L, Li H, et al. An Effective Machine Learning Approach for Prognosis of Paraquat Poisoning Patients Using Blood Routine Indexes. Basic Clin Pharmacol Toxicol 2017;**120**(1):86-96.

46. Chang LY, Wang HW. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid Anal Prev 2006;**38**(5):1019-27.

47. Ripley B. tree: Classification and regression trees. R package version 1.0-34. URL： http://CRAN.R-project.org/package=tree. 2013.

48. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;**44**(3):837-45.

49. Knol MJ, Vandenbroucke JP, Scott P, et al. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. Am J Epidemiol 2008;**168**(9):1073-81.

50. Gu W, Vieira AR, Hoekstra RM, et al. Use of random forest to estimate population attributable fractions from a case-control study of Salmonella enterica serotype Enteritidis infections. Epidemiol Infect 2015;**143**(13):2786-94.

51. Lemon SC, Roy J, Clark MA, et al. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. nn Behav Med 2003;**26**(3):172-81.

52. Chen S, Zhou S, Yin FF, et al. Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. Med Phys 2007;**34**(10):3808-14.

53. Orru G, Pettersson-Yeo W, Marquand AF, et al. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci Biobehav Rev 2012;**36**(4):1140-52.

54. Du Hongle LQ, and Cao Jing. Reduce the Samples for SVM Based on Euclidean Distance. 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization 2013.

55. R. H. Laskar FAT, Biman Paul and Debmalya Chakrabarty. Sample reduction using recursive and segmented data structure analysis. Journal of Engineering and Computer Innovations Vol 2(4), pp 59-67, 2011.

56. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to

imputation of missing values. J Clin Epidemiol 2006;**59**(10):1087-91.

57. Shrive FM, Stuart H, Quan H, et al. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC medical research methodology 2006;**6**:57.

58. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. J Clin Epidemiol 2002;**55**(4):329-37.

Figure Legend

Figure 1. Demographics and injury characteristics of the patients regarding gender, co-morbidities, injury region, number of injury regions, and helmet-wearing status.

Figure 2. Injury characteristics of the patients regarding laboratory data collected from the time point when arrival at the emergency department.

Figure 3. ROC curves for LR, SVM, and DT models in predicting mortality of motorcycle riders.

Figure 4. Illustration of DT model for mortality of motorcycle riders. The boxes denote the percentage of patients with discriminating variables from CART analysis. Those who were survival and fatal were indicated with green and red colors, respectively, in the boxes.

Tables

Table 1. Summarizes mortality prediction performances regarding accuracy, sensitivity, and specificity with LR, SVM, and DT models in the training and test sets.

			All samples n=6306		Reduced samples n=1510	
			All variables		All variables	
LR	Train	Accuracy	98.64		94.44	
		Sensitivity	59.31		60.00	
		Specificity	99.56		98.10	
	Test	Accuracy	98.41		98.41	
		Sensitivity	73.91		73.91	
		Specificity	99.02		99.02	
			All variables	Selected features	All variables	Selected features
SVM	Train	Accuracy	98.62	98.62	94.37	93.84
		Sensitivity	62.07	64.14	59.31	62.76
		Specificity	99.48	99.43	98.10	97.14
	Test	Accuracy	98.41	98.73	98.41	98.31
		Sensitivity	69.57	86.96	69.57	73.91
		Specificity	99.13	99.02	99.13	98.92
DT	Train	Accuracy	98.92	98.92	95.83	95.83
		Sensitivity	62.76	64.14	68.97	70.34
		Specificity	99.77	99.74	98.68	98.53
	Test	Accuracy	98.31	98.52	97.67	97.89
		Sensitivity	65.22	69.57	65.22	69.57
		Specificity	99.13	99.24	98.48	98.59

Table 2. Comparison of AUC between LR, SVM, and DT models in the training set. A * indicated p < 0.05. AS, all samples; RS, reduced samples; AV, all variables; SF, selected features.

		LR		SVM				DT			
		AS	RS	(AS + AV)	(AS + SF)	(RS + AV)	(RS + SF)	(AS + AV)	(AS + SF)	(RS+ AV)	(RS + SF)
LR	AS										
	RS	0.6575									
SVM	(AS + AV)	0.7481	0.6785								
	(AS + SF)	0.4121	0.7075	0.2473							
	(RS + AV)	0.9151	0.9161	0.6619	0.6652						
	(RS + SF)	0.3502	0.5965	0.4135	0.9939	0.5346					
DT	(AS + AV)	0.0001*	0.0001*	0.0001*	0.0002*	0.0002*	0.0002*				
	(AS + SF)	0.0001*	0.0002*	0.0001*	0.0002*	0.0002*	0.0002*	0.3578			
	(RS + AV)	0.0542	0.0618	0.0543	0.0713	0.0658	0.0703	0.0009*	0.0010*		
	(RS + SF)	0.0566	0.0643	0.0567	0.0743	0.0684	0.0731	0.0008*	0.0009*	0.3570	

LR: Logistic regression; SVM: support vector machine; DT: decision tree; AS: all samples; RS: reduced samples; AV: all variables; SF: selected features. * indicated p < 0.05

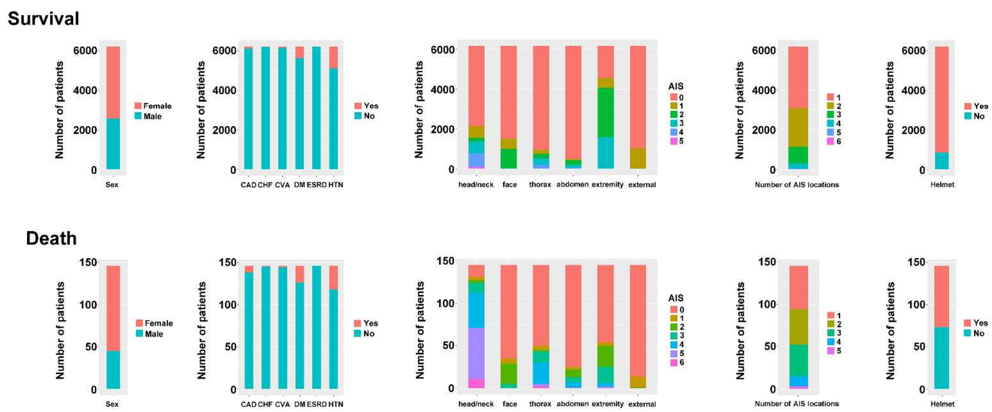


Figure 1. Demographics and injury characteristics of the patients regarding gender, co-morbidities, injury region, number of injury regions, and helmet-wearing status.

800x337mm (300 x 300 DPI)

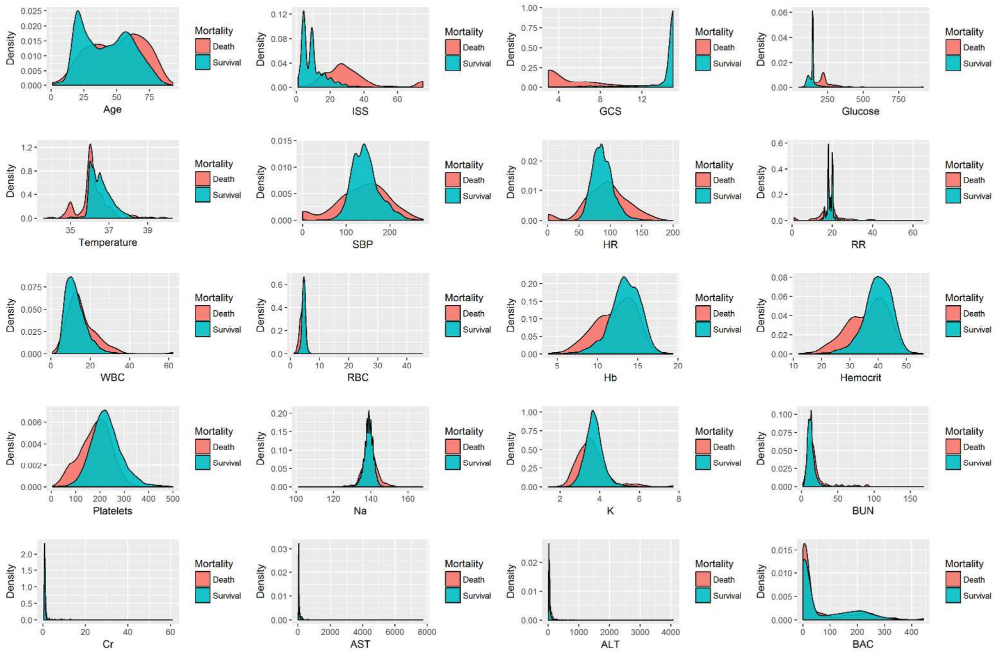


Figure 2. Injury characteristics of the patients regarding laboratory data collected from the time point when arrival at the emergency department

381x254mm (300 x 300 DPI)

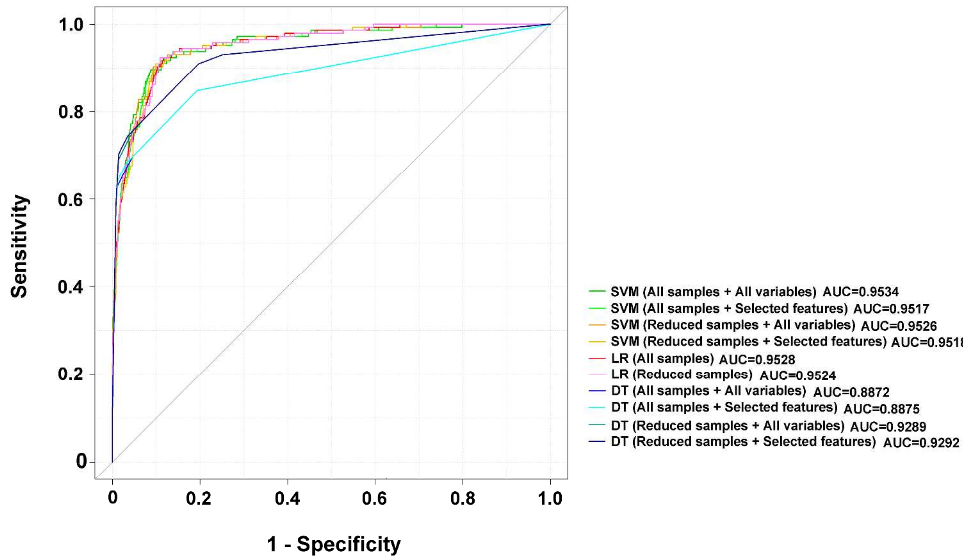


Figure 3. ROC curves for LR, SVM, and DT models in predicting mortality of motorcycle riders.

470x284mm (300 x 300 DPI)

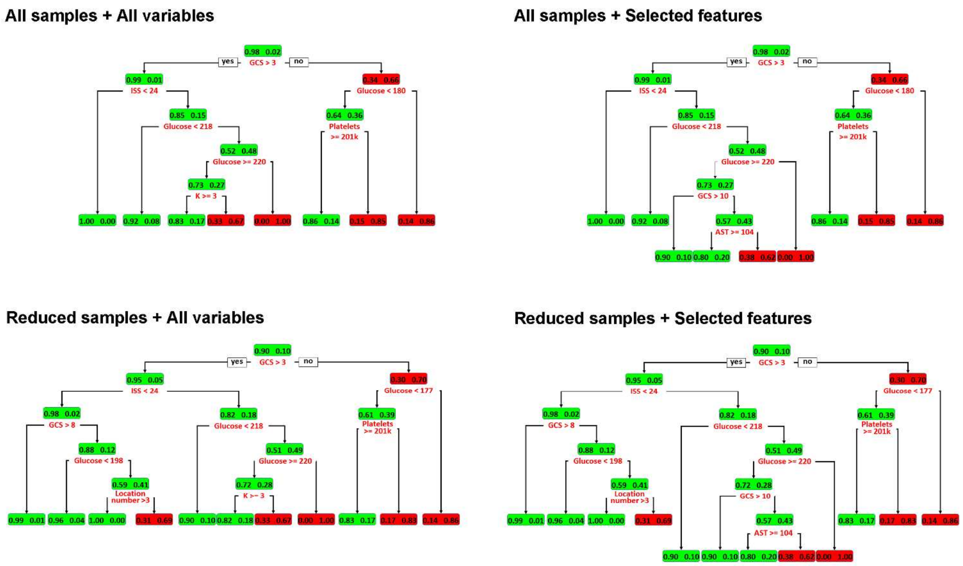


Figure 4. Illustration of DT model for mortality of motorcycle riders. The boxes denote the percentage of patients with discriminating variables from CART analysis. Those who were survival and fatal were indicated with green and red colors, respectively, in the boxes.

742x467mm (96 x 96 DPI)

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cross-sectional studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	4
Methods			
Study design	4	Present key elements of study design early in the paper	7
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	8
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	8
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	8-11
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	8
Bias	9	Describe any efforts to address potential sources of bias	-
Study size	10	Explain how the study size was arrived at	7
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	7-8
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	7-8
		(*b*) Describe any methods used to examine subgroups and interactions	7-8
		(*c*) Explain how missing data were addressed	-
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	7-8
		(*e*) Describe any sensitivity analyses	-
Results			

Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	7
		(b) Give reasons for non-participation at each stage	-
		(c) Consider use of a flow diagram	-
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	9-11
		(b) Indicate number of participants with missing data for each variable of interest	-
Outcome data	15*	Report numbers of outcome events or summary measures	-
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	11
		(*b*) Report category boundaries when continuous variables were categorized	-
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	-
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	-
Discussion			
Key results	18	Summarise key results with reference to study objectives	11-18
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	18
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	11
Generalisability	21	Discuss the generalisability (external validity) of the study results	-
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	19

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

BMJ Open

Derivation and validation of different machine learning models in mortality prediction of trauma motorcycle riders - a cross-sectional retrospective study in southern Taiwan

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018252.R1
Article Type:	Research
Date Submitted by the Author:	11-Oct-2017
Complete List of Authors:	Kuo, Pao-Jen; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Wu, Shao-Chun; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Anesthesiology Chien, Peng-Chen; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Rau, Cheng-Shyuan; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Neurosurgery Chen, Yi-Chun; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Hsieh, Hsiao-Yun; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Hsieh, Ching-Hua; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery
Primary Subject Heading:	Public health
Secondary Subject Heading:	Public health, Research methods
Keywords:	Motorcycle accident, mortality, machine learning (ML), logistic regression (LR), support vector machine (SVM), decision tree (DT)

SCHOLARONE™ Manuscripts

Derivation and validation of different machine learning models in mortality prediction of trauma motorcycle riders - a cross-sectional retrospective study in southern Taiwan

Pao-Jen Kuo[1†], M.D.; Shao-Chun Wu[2†] M.D.; Peng-Chen Chien[1], M.Sc.; Cheng-Shyuan Rau[3], M.D.; Yi-Chun Chen[1], M.Sc.; Hsiao-Yun Hsieh[1], M.Sc.; Ching-Hua Hsieh[1*], M.D., Ph.D.

[†] Indicates equal contribution in authorship as the first author.

[1] Department of Plastic and Reconstructive Surgery, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan, 833

[2] Department of Anesthesiology, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan, 833

[3] Department of Neurosurgery, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan, 833

Pao-Jen Kuo; email: bow110470@gmail.com

Shao-Chun Wu; email: shaochunwu@gmail.com

Peng-Chen Chien; e-mail: VENU_CHIEN@hotmail.com

Cheng-Shyuan Rau; e-mail: ersh2127@cloud.cgmh.org.tw

Yi-Chun Chen; e-mail: libe320@yahoo.com.tw

Hsiao-Yun Hsieh; e-mail: sylvia19870714@hotmail.com

Ching-Hua Hsieh; e-mail: m93chinghua@gmail.com

Corresponding author: Ching-Hua Hsieh, M.D., PhD.

Department of Trauma Surgery & Plastic and Reconstructive Surgery, Kaohsiung

Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan

No.123, Ta-Pei Road, Niao-Song District, Kaohsiung City 833, Taiwan

Tel: 886-7-7317123 ext: 8002; E-mail: m93chinghua@gmail.com

Abstract

Objectives: We aimed to build and test models of machine learning (ML) to predict the mortality of hospitalized motorcycle riders.

Setting: A Level I trauma center in southern Taiwan.

Participants: The hospitalized motorcycle riders between January 2009 and December 2015 were allocated to be a training set (n=6,306) and a test set (n= 946). Using the demographics and injury characteristics as well as laboratory data of patients, logistic regression (LR), support vector machine (SVM), and decision tree (DT) were performed to determine the mortality of the individual motorcycle riders, under the different conditions of using all samples or reduced samples as well as using all variables or selected features into the algorithm.

Primary and secondary outcome measures: Model predictive performance was evaluated by accuracy, sensitivity, specificity, geometric mean, and by the analysis of the area under curve of the receiver operator characteristic curves of the two different models.

Results: In the training set, both LR and SVM had a significantly higher AUC than that of DT, while there was no significant difference in the AUC of LR and SVM, regardless of using all samples or reduced samples as well as all variables or selected features. In the test set, SVM model for all samples with selected features presented a better model than all the other models, with an accuracy of 98.73%, sensitivity (86.96%), specificity (99.02%), geometric mean (92.79%) and AUC of 0.9517 for mortality prediction.

Conclusion: We demonstrate that ML is able to provide a feasible level of accuracy for predicting the mortality of the motorcycle riders. The integration of ML model, particularly the SVM algorithm in trauma system may help identify high-risk patients and therefore drive the appropriate response by the clinical staff.

KEY WORDS: Motorcycle accident; mortality; machine learning (ML); logistic regression (LR), support vector machine (SVM), and decision tree (DT)

ARTICLE SUMMARY

STRENGTHS AND LIMITATIONS OF THIS STUDY

- This study demonstrates the feasibility of using support vector machine (SVM) classification, one of machine learning models, to predict the mortality risk for motorcycle riders.
- With addition of more data in the model, the SVM model has the potential to get an increased predictive power and facilitate its clinical implement.
- The SVM model generally works like a black-box and cannot identify the relationships between mortality and the various explanatory variables.
- The incomplete records of patients and the exclusion of patients declared dead from the Trauma Registry System could bias the results.

BACKGROUND

As a less expensive and convenient means of transportation, motorcycle use is popular in many cities. However, despite being a small fraction of the travel, motorcycle riders involved in road traffic accidents often sustain severe morbidity and mortality. Compared to the occupants in a motor vehicle, motorcycle riders are 8 times more likely to be injured per vehicle mile [1], 30 times more likely to die in a motor vehicle crash [2], and 58 times more likely to be killed on a per-trip basis [3]. In Taiwan, motorcyclist fatalities account for nearly 60% of all driving fatalities [4]. The fatalities are often associated with men, advanced age, not wearing a helmet, unlicensed status, and riding under the influence of alcohol [5-9]. In addition, head injuries were the major factor leading to mortality, followed by thoracic and abdominal injuries [6-9].

Identifying patients with high risk of mortality is vital for the integration of trauma management to maximize resources and quality of care delivered [10 11]. More robust and accurate individual predictions of mortality from better models might give clinicians better information about the likelihood of good or poor outcomes and improve individual trauma and mortality management [12]. To identify the possibility of mortality, a frequently used model is the Trauma and Injury Severity Score (TRISS), which was established in 1987 to estimate an individual trauma patient's survival probability based on logistic regression (LR) with variables including age, anatomical variable (Injury Severity Score [ISS]), physiological variable (Revised Trauma Score [RTS]), and different coefficients for blunt and penetrating injuries. However, TRISS is imperfect and fails to determine a correct classification in 15-30% of the trauma patients [13]. Even after the incorporation of other or revised predictors, like blood pressure [14], co-morbidities, and separate categories for different age-groups [15] into this model, the addition of more predictors to the basic TRISS model did not always result in higher performance [16-18]. Although the revised TRISS, resulting from the USA National Trauma Database is inaccurate for trauma systems, particularly in the management of predominantly blunt injuries [19], the further development of the model based on advanced methodological quality, the performance of the model in subsets of patient groups, and practical application is mandatory in the prediction of mortality [16].

Currently, machine learning (ML) had been successfully applied in the real world in many fields including automatic medical diagnostics and personalized health care [20-22]. There is an increasing interest in the application of supervised ML methods to aid diagnosis and prognosis in trauma patients. ML is based on the way the human brain approaches pattern recognition tasks, providing an artificial intelligence-based approach to solve classification problems and improving their efficiency and effectiveness over time [23]. The usefulness of ML is bolstered by the versatility of its techniques and its utility for artificial intelligence such as prediction, classification, planning, recognition, and clustering [23 24]. Comparisons of different learning strategies have been conducted previously by others using field-specific datasets, many of which have shown significantly better predictive power than the more conventional alternatives [25]. Examples of multivariate techniques for pattern recognition include, but are not limited to, LR, support vector machine (SVM), decision trees (DT), and artificial neural networks. LR is a widely used and accepted statistical analysis tool to predict the probability of the occurrence of an event [26]. It attempts to build a functional relationship between two or more independent predictors and the one dependent outcome variable, under the assumption that the response variables are linearly related to the coefficients of the predictor variables [26].

SVM uses a training set of data composed of one or more features to determine an optimal boundary separating a set of cases. The binary SVM classifier constructs a set of the optimal hyperplanes in high-dimensional space with the maximal margin of the two classes [27]. In the case that all training points cannot be separated by the hyperplane, a soft margin method is used to construct a hyperplane that separates the training data points [28 29]. It has been found that the SVM model has a great capability of dealing with classification problems [30-34].

A DT is a hierarchical model composed of decision rules based upon optimal feature cutoff values that recursively split independent variables into different groups [35-37]. The purpose of DT building is to search for a set of decision rules to predict an outcome from a set of input variables [33 35 36]. Some models are used to construct decision-tree models, including classification and regression trees (CART), ID3s, chi-square automatic interaction detector DTs (CHAIDs), and C4.5 and C5.0 DTs [26, 28]. Among these methods, the CART analysis is a combined approach based on nonparametric and nonlinear variables for recursive partitioning analysis. CART analysis is an innovative DT model in which several predictive variables are crucial to identify patients at different levels of risk in various medical fields through progressive binary splits to develop prediction models in order to enable better prediction and clinical decision-making [38-40].

This study aimed to construct a model for the mortality prediction of motorcycle riders using ML algorithms and obtaining data from a population-based trauma registry in a level I trauma center.

METHODS

Ethics statement

This study was preapproved by the institutional review board (IRB) of Chang Gung Memorial Hospital with approval number 201600653B0. Informed consent was waived according to the IRB regulations.

Data preparation

Detailed patient information between January 2009 and December 2015 was retrieved from the Trauma Registry System of our institution, a 2,400-bed facility and Level I regional trauma center. Only the trauma patients who sustained a traffic accident as a motorcycle rider and were hospitalized for treatment were included in the study. The patient information included the following variables: age, sex, helmet-wearing status, co-morbidities such as coronary artery disease (CAD), congestive heart failure (CHF), cerebral vascular accident (CVA), diabetes mellitus (DM), end-stage renal disease (ESRD), and hypertension (HTN) as well as vital signs, including temperature, systolic blood pressure (SBP), heart rate (HR), respiratory rate (RR), ISS, Glasgow coma scale (GCS) score, abbreviated injury scale (AIS) in different regions of the body, number of injured body regions according to AIS (number of AIS locations), the in-hospital mortality, the blood level of white blood cell count (WBC), red blood cell count (RBC), hemoglobin (Hb), hematocrit (Hct), platelets, blood urine nitrogen (BUN), creatinine (Cr), alanine aminotransferase (ALT), aspartate aminotransferase (AST), sodium (Na), potassium (K), blood alcohol concentration (BAC), and glucose at emergency department.

These enrolled patients were divided into a training sample, which was used for predictor discovery and supervised classification to generate a plausible model, and a test sample, which was used to test the performance of the model generated in the training sample. Those patients with missing data were not included for further analysis. The patients who registered in a six-year span between January 2009 and December 2014 were allocated in the training set, which comprised of a total of 6,306 patients. It included 6,161 survival and 145 mortality patients. In the test set, there were 946 patients, including 923 survival and 23 mortality patients, from the one-year span between January 2015 and December 2015. The sample similarity was assessed by Euclidean distance for quantitative data to reduce the size of a sample designed for use in data analysis [41]. The sample reduction used Euclidean distance of the `dist` function in the `stats` package in R (R Foundation for Statistical Computing, Vienna, Austria). During sample reduction, the data size can be reduced to speed up calculations in the analysis [42]. However, considering the exploratory character of this study, all samples (n=6,306) and reduced samples (n=1,510) in the training set of this study would require to be analyzed in ML classification.

ML classifiers

This work provides a performance comparison of three different ML classifiers (LR, SVM, and DT).

Logistic regression

The LR classifier used the `glm` function in the `stats` package in R3.3.3 (R

Foundation for Statistical Computing, Vienna, Austria). Univariate LR analyses were initially performed to identify the significant predictor variables of the mortality risk. Stepwise LR analysis was used to control the effects of confounding variables to identify independent risk factors for mortality. The selected independent risk factors obtained from LR were also used as selected features to be implemented by the SVM and the DT to explain their weights in determining the risk of mortality.

Support vector machine

The SVM classifier used the `tune.svm & svm` function in the `e1071` package in R. In the training set, the SVM classifier was performed for the prediction of mortality with regard to either all 32 variables or 12 selected features as well as all the samples and reduced samples in the training set. The mapping procedure was accomplished by the kernel function, which is a matrix of pair-wise similarities between data points, such as a linear, polynomial, or radial basis function (RBF) [43]. For this study, the RBF kernel was chosen because it can handle non-linear interactions between class labels and features [44]. The two main parameters presented in SVM with RBF kernel were the penalty parameter C and the kernel hyper-parameter γ. The penalty parameter C determined the tradeoff between the fitting error minimization and model complexity, while the hyper-parameter γ defined the nonlinear feature transformation onto a higher dimensional space and controlled the tradeoff between error due to bias and variance in the model. [45]. The optimal operating point was estimated by varying the parameters - C and γ using a grid search for each combination of feature selection and dimension reduction with a 10-fold cross-validation [44].

Decision tree

The DT by CART based on the Gini impurity index used the `rpart` function in the `rpart` package in R. The CART analysis searched for the split on the variable that would partition the data into two different groups—a group of mostly '0s' (people who survived) and a group of mostly '1s' (people who died) [46 47]. Using the best overall split, the CART model partitioned the data and assigned a predicted class to each subgroup. CART repeated this same process on each predictor in the model, identifying the best split by iteratively testing all possible splits, and producing the greatest reduction in impurity [38-40]. CART proceeded recursively in this way until the specified stopping criteria were reached, a specified number of nodes were created, or a further reduction in node impurity became impossible [38-40].

Performance evaluation

We used receiver operator characteristic (ROC) curve analysis to assess and compare the performances of the individual ML models. Model predictive ability was evaluated using confusion matrix and the area under curve (AUC) analysis between two approaches of ML models.

Confusion matrix and geometric mean

The confusion matrix calculates the accuracy, sensitivity, and specificity of a given model with true negative, true positive, false positive, and false negative values and presents as a result an accuracy, which represents the overall proportion of correct classifications; a sensitivity, which refers to the proportion of true positives correctly identified (e.g. percentage of people with fatality identified to be dead); and a specificity, which refers to the proportion of true negatives correctly identified (e.g.

percentage of people who survived identified as not dead). In addition, because the geometric mean can provide a good trade-off between sensitivity and specificity in a way that a better accuracy in both classes leads to a larger value, the geometric mean between sensitivity and specificity was calculated in this study according to the methods used by Sanz J et al. [48].

AUC analysis

In order to compare the performance of multiple ML classifiers in multiple training data sets, a nonparametric approach to the analysis of areas under correlated ROC curves using the `roc & roc.test` function in the `pROC` package in R is pursued. This nonparametric approach takes into account the correlated nature of the data that two or more empirical curves are constructed based on tests performed on the same individuals [49].

All statistical analyses were performed using SPSS 20.0 (IBM Inc., Chicago, IL, USA) and R 3.3.3. For categorical variables, Chi-square tests were used to determine the significance of the association between the predictor and outcome variables. For continuous variables, student t-tests were applied to analyze normally distributed data, while Kolmogorov-Smirnov tests or Mann-Whitney U tests were used to compare non-normally distributed data. All of the results were presented in the form of the mean ± standard deviation. A p-value < 0.05 was considered statistically significant.

RESULTS

Demographics and injury characteristics of the patients

The patients with fatality had a higher AIS score at the head and neck region but lower AIS score at the extremities compared to the patients who survived (Table 1 and supplemental Figure 1). The patients with fatality had sustained more number of injured body regions (number of AIS locations) than the ones who survived. In addition, the patients with fatality comprised more of females and fewer of them were observed to be wearing a helmet compared to the patients who survived (Table 1 and supplemental Figure 1). A statistically significant difference in age, ISS, GCS, glucose, temperature, Hb, Hct, platelets, K, Cr, AST, ALT, and incidences of CAD was found between patients with fatality and the ones who survived respectively (Table 2 and supplemental Figure 2). Because the distribution pattern between Hb and Hct as well as between AST and ALT is very similar, only one of these two variables (i.e. Hct and AST) was selected for further ML classification to prevent the inclusion of duplicate parameters. Therefore, a total of 32 variables were used for imputation into ML classifiers as all variables, in contrast to considering selected features obtained by using the independent risk factors identified by the LR given below.

Performance of ML classifiers in training set

Logistic regression

LR identified 12 predictors (platelets, glucose, BUN, Cr, AST, Na, Age, GCS, temperature, number of AIS locations, ISS, and HTN) as independent risk factors for mortality in motorcycle riders from either all samples or the reduced samples.

The predictive models were listed as:

All samples (n=6,306)

$$Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 4.71648 - 0.00846 * \text{Platelets} + 0.01189 * \text{Glucose} +$$

$$0.03459 * \text{BUN} + 0.10667 * \text{Cr} + 0.00195 * \text{AST} + 0.09513 * \text{Na} +$$

$0.02533 * \text{Age} - 0.39968 * \text{GCS} - 0.56396 * \text{Temperature} - 0.93232 *$

$\text{Number of AIS locations} + 0.14098 * \text{ISS} - 0.95726 * \text{HTN}$

Reduced samples (n=1,510)

$$Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 5.76780 - 0.00763 * \text{Platelets} + 0.00953 * \text{Glucose} +$$

$$0.03773 * \text{BUN} + 0.00152 * \text{AST} + 0.08630 * \text{Na} + 0.02014 * \text{Age} -$$

$$0.34116 * \text{GCS} - 0.53370 * \text{Temperature} - 0.91439 *$$

$$\text{Number of AIS locations} + 0.12191 * \text{ISS} - 1.00522 * \text{HTN}$$

The LR achieved an accuracy of 98.64% (sensitivity of 59.31% and specificity of 99.56%) and 94.44% (sensitivity of 60.00% and specificity of 98.10%) for all samples and reduced samples, respectively. The AUCs for all samples and reduced sample were 0.9528 and 0.9524, respectively (Figure 1).

Support vector machine

In the training set, the SVM classifier was performed for the prediction of mortality taking input as either all 32 variables or the 12 selected features in all samples and reduced samples, respectively. With the RBF as the kernel function, the SVM model has two parameters (C, γ) that need to be determined. The accuracy was highly robust to small changes in the hyper-parameters, so reasonable choices were obtained by a grid search of 2^x where x is an integer between -8 and 4 for C and between -10 and -2 for γ. The values which gave the highest 10-fold cross-validation accuracy are reported to be C = 0.25 and γ= 0.00390625. Under the input of all variables into the model, the SVM achieved an accuracy of 98.62% (sensitivity of 62.07% and specificity of 99.48%) and 94.37% (sensitivity of 59.31% and specificity of 98.10%) for all samples and reduced samples, respectively (Table 3). The AUCs for all samples and reduced sample were 0.9534 and 0.9526, respectively (Figure 1). With selected features in the model, the SVM achieved an accuracy of 98.62% (sensitivity of 64.14% and specificity of 99.43%) and 93.84% (sensitivity of 62.76% and specificity of 97.14%) (Table 3) as well as 0.9517 and 0.9518 AUCs for all samples and reduced samples, respectively (Figure 1).

Decision tree

As shown in Figure 2, in the DT model, GCS was identified as the variable of initial split with an optimal cut-o value of > 3. Among patients with GCS higher than 3, glucose was selected as the variable of second split at a discrimination level of 180 and 177 mg/dL for all samples and reduced samples, respectively. After the glucose level < 180 or 177 mg/dL for all samples and reduced samples, respectively, the next best predictor of mortality was platelets with an optimal cut-off of $201X10^3$ /µL. For the node, with patents having a GCS not greater than 3, ISS < 24 and glucose < 218 mg/dL, these predictors were selected as significant variables for all samples and reduced samples, with GCS > 8, glucose < 198 mg/dL, and the number of AIS locations ≥ 3 being an additional predictors for splitting for the reduced samples. With all variables in the model, the DT achieved an accuracy of 98.92% (sensitivity of 62.76% and specificity of 99.77%) and 95.83% (sensitivity of 68.97% and specificity of 98.68%) for the all samples and reduced samples, respectively. The AUCs for all samples and reduced samples were 0.8872 and 0.9289, respectively. With selected features in the model, the DT achieved an accuracy of 98.92% (sensitivity of 64.14% and specificity of 99.74%) and 95.83% (sensitivity of 70.34% and specificity of 98.53%) for the all samples and reduced samples, respectively. The AUCs for all samples and reduced samples were 0.8872 and 0.9289, respectively (Figure 1). In the condition of using reduced samples but not all samples in the DT model, the number of AIS locations would be added in the split of the node slightly increasing the sensitivity from 62.76% to 68.97% and from 64.14% to 70.34% with input comprising of all variables and selected variables, respectively. In addition, in the condition of using selected features but not all variables in the DT model, the level of K was not used in the splitting of the node and was substituted by the cut-off value of AST ≥ 104 IU/L, slightly increasing the sensitivity from 62.76% to 64.14% and from 68.97% to 70.34% with input as all samples and reduced samples, respectively. In addition, the AUCs for all samples and reduced sample were 0.8875 and 0.9292, respectively (Figure 1)

Comparison in AUC analysis

In the comparisons of AUCs for LR, SVM, and DT for the training set (Table 4 and Figure 1), both LR and SVM had a significantly higher AUC than DT, regardless of using all samples or reduced samples as well as for all variables or for selected features. However, there was no significant difference of AUC between LR and SVM, regardless of using all samples or reduced samples as well as for all variables or for selected features. In addition to this, in DT sample reduction had a significantly higher AUC than the one obtained using all samples, but there was no significant difference of AUC between DT with all variables or with selected features.

Performance of ML classifiers in test set

In test set, the LR model for all samples and reduced samples - both achieved an accuracy of 98.41%, with a sensitivity of 73.91% and specificity of 99.02% in predicting the mortality (Table 3). All of these four SVM models create an accuracy more than 98% and a specificity near 99% in predicting the mortality. Whereas the SVM model for all samples with selected features had a significantly highest sensitivity (86.96%) and geometric mean (92.79%). All of these four DT models create an accuracy of approximately 98% and a specificity of approximately 99% but a sensitivity of less than 70%. Considering that the majority of patients survived except for a few with fatality, would result into a very high accuracy and specificity index in predicting the mortality, therefore the comparison should further focus on the sensitivity and geometric mean of different ML models. We found all LR and SVM models, but not the DT models, had an increased sensitivity in the test set. In addition, the SVM model for all samples with selected features had a significantly highest sensitivity and geometric mean.

DISCUSSION

LR is widely used in epidemiological studies for causal inference and, with the selection of built-in features; it does not necessarily utilize all the predictors. With a relatively limited number of variables i.e. variables less than 20, LR provides estimates of odd ratios of the risk factors [50]. However, its limitations become apparent when analyzing a complex dataset with a high number of relevant exposures and multiple interactions [51]. With too many predictors, the availability of sufficient information to specify all interactions would become nearly impossible [51]. In addition, the DT with CART analysis is exploratory and not based on the probabilistic method, which may lead to overestimating the importance of included risk factors or cause missing of other potential confounders that could influence each patient's actual risk [52]. In contrast to LR, which is heavily influenced by outliers in its linear discriminant analysis method, the SVM boundary is only minimally influenced by outliers that are difficult to separate, despite the complexity of data [53]. In addition, employing kernels in the SVM would be advantageous to learn non-linear decision boundaries allowing the classifier to solve more difficult classification problems than the linear analysis method [54]. These three ML models (LR, SVM, DT) all create an accuracy and a specificity around 98% and 99%, respectively, but a sensitivity less than or around 70% in the training dataset. In this study, both LR and SVM resulted in a significantly higher AUC than DT in the training set, regardless of using all samples or reduced samples as well as for all variables or for selected features.

This study included different variants of SVM considering the sample size and feature selection to show all possible improvements to the more conventional strategies like LR or DT. Although sample reduction for SVM had been proposed to greatly improve the training speed of the SVM and save a lot of storage space [55 56], using the kernel is a more efficient technique in case of similarity of representation between samples. Thus, the computational complexity of SVM is not wholly governed by the number of samples, but by the number of features, which is advantageous for analysis in the high-dimensional settings [54]. In addition, feature selection in SVM may maximize the AUC [25]. Aided by feature selection, the proposed SVM method identifies the most discriminating indexes for mortality prediction. We found that, although both LR and SVM did not have a different AUC in the training procedure, the SVM model for all samples with selected features had a significantly higher sensitivity (86.96%) in predicting the mortality of motorcycle rides in the test set compared to the rest of the models. The increased sensitivity of SVM in test set than that in training set may be attributed to an improved quality of registered content and less missing data in our registered data after continuous quality assessment and years of experience of working with the registers. Such increased sensitivity was also found in the LR model in the test set. With addition of more data in the model, the SVM model has the potential to get an increased predictive power. This study demonstrates the feasibility of using SVM classification with feature selection to predict the mortality risk for motorcycle riders in the trauma care. However, the SVM model generally works like a black-box and cannot identify the relationships between mortality and the various explanatory variables and therefore, cannot be directly used to validate our hypothesis of increased sensitivity in the test set.

There are several limitations to this study. Firstly, the patients who had incomplete records were excluded from the analysis. This could have caused the results to be biased and the results could have been different from the data acquired by including the patients with incomplete records and replacing the missing data on a variable by a value that is drawn from an estimate of the distribution of this variable [57-59]. The benefit of imputation is that we would be able to include patients who might have relevant features for analysis, but were excluded owing to errors in data collection or recording [57-59]. Secondly, a source of potential bias may come from the exclusion of patients declared dead (either on arriving at the hospital or at the accident spot itself) and injured patients who were discharged against the advice of the emergency department. Thirdly, there was lack of important data regarding injury mechanism and circumstance, including motorcycle speed and type, helmet material, and impact force during collision. In addition, the imputation of physiological and laboratory data collected from the time of arriving at the emergency department cannot reflect the dynamic changes in hemodynamic and metabolic variables of the trauma patients under a possible resuscitation procedure. Further, some other DT-related methods like DT by C4.5 [60], combined classifiers of LR and DT by C4.5 [48], and random forest [61] had been reported to provide a very good performance in dealing with the classification problem; however, these techniques were not explored in this study. Finally, the study population was limited to a single urban trauma center in southern Taiwan, which may not be representative of other populations.

CONCLUSION

We demonstrate that ML is able to provide a feasible level of accuracy for predicting mortality of the motorcycle riders. Whilst there are significant theoretical and practical challenges to the translational implementation of this approach, the results of the studies published so far are encouraging and may provide the first steps towards the development of a prediction model integrated into the trauma care system in order to identify an individual motorcycle rider's risk of mortality.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

PJK wrote the manuscript; SCW revised the manuscript; PCC performed the statistical analyses and machine learning programming; CSR analyzed the tables; YCC and HYH collected the data and are responsible for the integrity of the registered data; and CHH designed the study and contributed to the analysis and interpretation of data. All authors have read and approved the final manuscript.

Data Sharing

No additional data are available.

Acknowledgements

This research was supported by a grant from Chang Gung Memorial Hospital CMRPG8F0891. We appreciate the Biostatistics Center, Kaohsiung Chang Gung Memorial Hospital for helping us with the statistical work.

References

1. Weiss H, Agimi Y, Steiner C. Youth motorcycle-related brain injury by state helmet law type: United States, 2005-2007. Pediatrics 2010;**126**(6):1149-55.
2. National Highway Traffic Safety Administration (NHTSA) TSFDDH.
3. Beck LF, Dellinger AM, O'Neil ME. Motor vehicle crash injury rates by mode of travel, United States: using exposure-based methods to quantify differences. American journal of epidemiology 2007;**166**(2):212-8.
4. Chang HL, Lai CY. Using travel socialization and underlying motivations to better understand motorcycle usage in Taiwan. Accident; analysis and prevention 2015;**79**:212-20.
5. Jou RC, Yeh TH, Chen RS. Risk factors in motorcyclist fatalities in Taiwan. Traffic injury prevention 2012;**13**(2):155-62.
6. Liang CC, Liu HT, Rau CS, et al. Motorcycle-related hospitalization of adolescents in a Level I trauma center in southern Taiwan: a cross-sectional study. BMC pediatrics 2015;**15**:105.
7. Liu HT, Liang CC, Rau CS, et al. Alcohol-related hospitalizations of adult motorcycle riders. World journal of emergency surgery : WJES 2015;**10**(1):2.

8. Hsieh CH, Hsu SY, Hsieh HY, et al. Differences between the sexes in motorcycle-related injuries and fatalities at a Taiwanese level I trauma center. Biomed J 2017;**40**(2):113-20.

9. Hsieh CH, Liu HT, Hsu SY, et al. Motorcycle-related hospitalizations of the elderly. Biomed J 2017;**40**(2):121-28.

10. Densmore JC, Lim HJ, Oldham KT, et al. Outcomes and delivery of care in pediatric injury. Journal of pediatric surgery 2006;**41**(1):92-8; discussion 92-8.

11. Rogers SC, Campbell BT, Saleheen H, et al. Using trauma registry data to guide injury prevention program activities. The Journal of trauma 2010;**69**(4 Suppl):S209-13.

12. Norrie J. Mortality prediction in ICU: a methodological advance. The Lancet Respiratory medicine 2015;**3**(1):5-6.

13. Demetriades D, Chan L, Velmanos GV, et al. TRISS methodology: an inappropriate tool for comparing outcomes between trauma centers. J Am Coll Surg 2001;**193**(3):250-4.

14. Jones JM, Skaga NO, Sovik S, et al. Norwegian survival prediction model in trauma: modelling effects of anatomic injury, acute physiology, age, and co-morbidity. Acta Anaesthesiol Scand 2014;**58**(3):303-15.

15. Bergeron E, Rossignol M, Osler T, et al. Improving the TRISS methodology by restructuring age categories and adding comorbidities. J Trauma 2004;**56**(4):760-7.

16. de Munter L, Polinder S, Lansink KW, et al. Mortality prediction models in the general trauma population: A systematic review. Injury 2017;**48**(2):221-29.

17. Fueglistaler P, Amsler F, Schuepp M, et al. Prognostic value of Sequential Organ Failure Assessment and Simplified Acute Physiology II Score compared with trauma scores in the outcome of multiple-trauma patients. Am J Surg

2010;**200**(2):204-14.

18. Kroezen F, Bijlsma TS, Liem MS, et al. Base deficit-based predictive modeling of outcome in trauma patients admitted to intensive care units in Dutch trauma centers. J Trauma 2007;**63**(4):908-13.

19. Stoica B, Paun S, Tanase I, et al. Probability of Survival Scores in Different Trauma Registries: A Systematic Review. Chirurgia (Bucur) 2016;**111**(2):115-9.

20. Cohen AM, Ambert K, McDonagh M. A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. AMIA Annual Symposium proceedings AMIA Symposium 2010;**2010**:121-5.

21. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2016.

22. Szlosek DA, Ferrett J. Using Machine Learning and Natural Language Processing Algorithms to Automate the Evaluation of Clinical Decision Support in Electronic Medical Record Systems. EGEMS (Washington, DC) 2016;**4**(3):1222.

23. Oquendo MA, Baca-Garcia E, Artes-Rodriguez A, et al. Machine learning and data mining: strategies for hypothesis generation. Mol Psychiatry 2012;**17**(10):956-9.

24. Kotoku J. An Introduction to Machine Learning. Igaku butsuri : Nihon Igaku Butsuri Gakkai kikanshi = Japanese journal of medical physics : an official journal of Japan Society of Medical Physics 2016;**36**(1):18-22.

25. Yahya N, Ebert MA, Bulsara M, et al. Statistical-learning strategies generate only modestly performing predictive models for urinary symptoms following external beam radiotherapy of the prostate: A comparison of conventional and

machine-learning methods. Med Phys 2016;**43**(5):2040.

26. Siuly, Yin X, Hadjiloucas S, et al. Classification of THz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers. Comput Methods Programs Biomed 2016;**127**:64-82.

27. V. V. Statistical learning theory. New York, NY: John Wiley & Sons 1998.

28. de Boves Harrington P. Support Vector Machine Classification Trees. Anal Chem 2015;**87**(21):11065-71.

29. Lee Y. Support vector machines for classification: a statistical portrait. Methods Mol Biol 2010;**620**:347-68.

30. Chen C, Zhang G, Qian Z, et al. Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accid Anal Prev 2016;**90**:128-39.

31. Galatzer-Levy IR, Karstoft KI, Statnikov A, et al. Quantitative forecasting of PTSD from early trauma responses: a Machine Learning application. J Psychiatr Res 2014;**59**:68-76.

32. Li Z, Liu P, Wang W, et al. Using support vector machine models for crash injury severity analysis. Accid Anal Prev 2012;**45**:478-86.

33. Marucci-Wellman HR, Corns HL, Lehto MR. Classifying injury narratives of large administrative databases for surveillance-A practical approach combining machine learning ensembles and human review. Accid Anal Prev 2017;**98**:359-71.

34. Patil BM, Joshi RC, Toshniwal D, et al. A new approach: role of data mining in prediction of survival of burn patients. J Med Syst 2011;**35**(6):1531-42.

35. Farion K, Michalowski W, Wilk S, et al. A tree-based decision model to support prediction of the severity of asthma exacerbations in children. J Med Syst 2010;**34**(4):551-62.

36. Zintzaras E, Bai M, Douligeris C, et al. A tree-based decision rule for identifying profile groups of cases without predefined classes: application in diffuse large B-cell lymphomas. Comput Biol Med 2007;**37**(5):637-41.

37. Kasbekar PU, Goel P, Jadhav SP. A Decision Tree Analysis of Diabetic Foot Amputation Risk in Indian Patients. Frontiers in endocrinology 2017;**8**:25.

38. Guilbault RWR, Ohlsson MA, Afonso AM, et al. External Validation of Two Classification and Regression Tree Models to Predict the Outcome of Inpatient Cardiopulmonary Resuscitation. J Viral Hepat 2017;**32**(5):333-38.

39. Shi KQ, Zhou YY, Yan HD, et al. Classification and regression tree analysis of acute-on-chronic hepatitis B liver failure: Seeing the forest for the trees. 2017;**24**(2):132-40.

40. Zimmerman RK, Balasubramani GK, Nowalk MP, et al. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. BMC Infect Dis 2016;**16**(1):503.

41. Amaratunga D, Cabrera J, Lee YS. Resampling-based similarity measures for high-dimensional data. J Comput Biol 2015;**22**(1):54-62.

42. Bhattacharya S, Mariani TJ. Transformation of expression intensities across generations of Affymetrix microarrays using sequence matching and regression modeling. Nucleic Acids Res 2005;**33**(18):e157.

43. Vapnik VN. The Nature of Statistical Learning Theory. New York, 2nd ed. 2000.

44. Gultepe E, Green JP, Nguyen H, et al. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. J Am Med Inform Assoc 2014;**21**(2):315-25.

45. Chen H, Hu L, Li H, et al. An Effective Machine Learning Approach for Prognosis of Paraquat Poisoning Patients Using Blood Routine Indexes. Basic Clin

Pharmacol Toxicol 2017;**120**(1):86-96.

46. Chang LY, Wang HW. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid Anal Prev 2006;**38**(5):1019-27.

47. Ripley B. tree: Classification and regression trees. R package version 1.0-34. URL ：http://CRAN.R-project.org/package=tree. 2013.

48. Sanz J, Paternain D, Galar M, et al. A new survival status prediction system for severe trauma patients based on a multiple classifier system. Comput Methods Programs Biomed 2017;**142**:1-8.

49. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;**44**(3):837-45.

50. Knol MJ, Vandenbroucke JP, Scott P, et al. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. Am J Epidemiol 2008;**168**(9):1073-81.

51. Gu W, Vieira AR, Hoekstra RM, et al. Use of random forest to estimate population attributable fractions from a case-control study of Salmonella enterica serotype Enteritidis infections. Epidemiol Infect 2015;**143**(13):2786-94.

52. Lemon SC, Roy J, Clark MA, et al. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. nn Behav Med 2003;**26**(3):172-81.

53. Chen S, Zhou S, Yin FF, et al. Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. Med Phys 2007;**34**(10):3808-14.

54. Orru G, Pettersson-Yeo W, Marquand AF, et al. Using Support Vector Machine to

identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci Biobehav Rev 2012;**36**(4):1140-52.

55. Du Hongle LQ, and Cao Jing. Reduce the Samples for SVM Based on Euclidean Distance. 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization 2013.

56. R. H. Laskar FAT, Biman Paul and Debmalya Chakrabarty. Sample reduction using recursive and segmented data structure analysis. Journal of Engineering and Computer Innovations Vol 2(4), pp 59-67, 2011.

57. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;**59**(10):1087-91.

58. Shrive FM, Stuart H, Quan H, et al. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC medical research methodology 2006;**6**:57.

59. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. J Clin Epidemiol 2002;**55**(4):329-37.

60. Wiharto W, Kusnanto H, Herianto H. Interpretation of Clinical Data Based on C4.5 Algorithm for the Diagnosis of Coronary Heart Disease. Healthcare informatics research 2016;**22**(3):186-95.

61. Rigatti SJ. Random Forest. J Insur Med 2017;**47**(1):31-39.

Figure Legend

Figure 1. ROC curves for LR, SVM, and DT models in predicting mortality of motorcycle riders.

Figure 2. Illustration of DT model for mortality of motorcycle riders. The boxes

denote the percentage of patients with discriminating variables from CART analysis. Those who were survival and fatal were indicated with green and red colors, respectively, in the boxes.

Supplemental Figure 1. Demographics and injury characteristics of the patients regarding categorical variables.

Supplemental Figure 2. Injury characteristics of the patients regarding continuous variables.

TABLES

Table 1. Demographics and injury characteristics of the patients regarding gender, helmet-wearing status, co-morbidities, injury region, and number of injury regions.

Variables		**Total (n = 7252)**	**Survival (n = 7084)**	**Mortality (n = 168)**	***P*-value**
Sex	Female	4291 (59.2%)	4174 (58.9%)	117 (69.6%)	0.005
	Male	2961 (40.8%)	2910 (41.1%)	51 (30.4%)	
Helmet	NO	1011 (13.9%)	929 (13.1%)	82 (48.8%)	<0.001
	YES	6241 (86.1%)	6155 (86.9%)	86 (51.2%)	
DM	NO	6562 (90.5%)	6414 (90.5%)	148 (88.1%)	0.286
	YES	690 (9.5%)	670 (9.5%)	20 (11.9%)	
HTN	NO	5939 (81.9%)	5802 (81.9%)	137 (81.5%)	0.919
	YES	1313 (18.1%)	1282 (18.1%)	31 (18.5%)	
CAD	NO	7120 (98.2%)	6960 (98.2%)	160 (95.2%)	0.011
	YES	132 (1.8%)	124 (1.8%)	8 (4.8%)	
CHF	NO	7228 (99.7%)	7061 (99.7%)	167 (99.4%)	0.431
	YES	24 (0.3%)	23 (0.3%)	1 (0.6%)	
CVA	NO	7168 (98.8%)	7002 (98.8%)	166 (98.8%)	0.722
	YES	84 (1.2%)	82 (1.2%)	2 (1.2%)	

ESRD	NO	7250 (100%)	7082 (100%)	168 (100%)	1.000
	YES	2 (0.0%)	2 (0.0%)	0 (0.0%)	
AIS (Head/Neck)	0	4642 (64%)	4627 (65.3%)	15 (8.9%)	<0.001
	1	665 (9.2%)	661 (9.3%)	4 (2.4%)	
	2	192 (2.6%)	189 (2.7%)	3 (1.8%)	
	3	713 (9.8%)	699 (9.9%)	14 (8.3%)	
	4	840 (11.6%)	795 (11.2%)	45 (26.8%)	
	5	189 (2.6%)	113 (1.6%)	76 (45.3%)	
	6	11 (0.2%)	0 (0%)	11 (6.5%)	
AIS (Face)	0	5472 (75.4%)	5347 (75.5%)	125 (74.4%)	<0.001
	1	574 (7.9%)	568 (8%)	6 (3.6%)	
	2	1173 (16.2%)	1141 (16.1%)	32 (19%)	
	3	33 (0.5%)	28 (0.4%)	5 (3%)	
AIS (Thorax)	0	6081 (83.9%)	5973 (84.3%)	108 (64.3%)	<0.001
	1	234 (3.2%)	229 (3.3%)	5 (3%)	
	2	260 (3.6%)	258 (3.6%)	2 (1.2%)	
	3	423 (5.8%)	404 (5.7%)	19 (11.3%)	
	4	245 (3.4%)	217 (3.1%)	28 (16.7%)	
	5	7 (0.1%)	3 (<0.1%)	4 (2.4%)	
	6	2 (<0.1%)	0 (0%)	2 (1.1%)	
AIS (Abdomen)	0	6654 (91.8%)	6516 (92%)	138 (82.1%)	<0.001
	1	57 (0.8%)	54 (0.8%)	3 (1.8%)	
	2	288 (4%)	277 (3.9%)	11 (6.5%)	
	3	170 (2.2%)	163 (2.3%)	7 (4.2%)	
	4	66 (0.9%)	58 (0.8%)	8 (4.8%)	
	5	17 (0.2%)	16 (0.2%)	1 (0.6%)	
AIS (Extremity)	0	2000 (27.6%)	1897 (26.8%)	103 (61.3%)	<0.001
	1	528 (7.3%)	524 (7.4%)	4 (2.4%)	
	2	2886 (39.8%)	2853 (40.3%)	33 (19.6%)	
	3	1822 (25.1%)	1800 (25.4%)	22 (13.1%)	
	4	12 (0.2%)	8 (0.1%)	4 (2.4%)	
	5	4 (0.1%)	2 (0.0%)	2 (1.2%)	
AIS (External)	0	6155 (84.9%)	6001 (84.7%)	154 (91.7%)	0.003
	1	1072 (14.8%)	1059 (14.9%)	13 (7.7%)	
	2	25 (0.3%)	24 (0.3%)	1 (0.6%)	
Number of AIS locations	1	3687 (50.8%)	3631 (51.3%)	56 (33.3%)	<0.001
	2	2255 (31.1%)	2205 (31.1%)	50 (29.8%)	
	3	982 (13.5%)	939 (13.3%)	43 (25.6%)	

4	280 (3.9%)	265 (3.7%)	15 (8.9%)
5	43 (0.6%)	39 (0.6%)	4 (2.4%)
6	5 (0.1%)	5 (0.1%)	0 (0.0%)

Table 2. Injury characteristics of the patients regarding laboratory data collected from the time point when arrival at the emergency department.

Variables	**Total (n = 7252)**	**Survival (n = 7084)**	**Mortality (n = 168)**	***P*-value**
Age (years)	38 (29)	37 (29)	47 (32)	<0.001
HR (beats/min)	89 (23)	89 (23)	93 (43)	<0.001
SBP (mmHg)	137 (38)	137 (37)	143 (79)	0.374
RR (times/min)	19 (2)	19 (2)	19 (5)	0.660
Temperature (^{o}C)	36.4 (0.8)	36.4 (0.8)	36.0 (0.5)	<0.001
GCS	15 (5)	15 (3)	3 (3)	<0.001
ISS	13 (12)	13 (13)	29 (11)	<0.001
RBC (10^6/uL)	4.6 (0.8)	4.6 (0. 8)	4.3 (1.1)	<0.001
WBC (10^3/uL)	12.9 (7.7)	12.9 (7.7)	13.2 (8.7)	<0.001
Hb (g/dL)	13.9 (2.5)	13.9 (2.5)	12.9 (3.5)	<0.001
Hct (%)	40.9 (6.8)	41.1 (6.6)	38.6 (9.4)	<0.001
Platelets (10^3/uL)	228 (79)	230 (79)	190 (78)	<0.001
Glucose (mg/dL)	145 (27)	145 (23)	218 (60)	<0.001
Na (mEq/L)	139 (3)	139 (3)	139 (4)	0.094
K (mEq/L)	3.5 (0.6)	3.5 (0.6)	3.4 (0.9)	<0.001
BUN (mg/dL)	12 (6)	12 (5)	14 (8)	<0.001
Cr (mg/dL)	0.8 (0.3)	0.8 (0.3)	1.0 (0.5)	<0.001
AST (U/L)	47 (50)	45 (48)	65 (76)	<0.001
ALT (U/L)	34 (35)	34 (33)	39 (55)	<0.001
BAC (mg/dL)	4.9 (133.0)	4.9 (136.4)	4.9 (62.5)	0.698

Table 3. Summarizes mortality prediction performances regarding accuracy, sensitivity, specificity, and geometric mean with LR, SVM, and DT models in the training and test sets.

			All samples n=6306		Reduced samples n=1510	
			All variables		All variables	
LR	Train	Accuracy	98.64		94.44	
		Sensitivity	59.31		60	
		Specificity	99.56		98.1	
		Geometric mean	76.84		76.72	
	Test	Accuracy	98.41		98.41	
		Sensitivity	73.91		73.91	
		Specificity	99.02		99.02	
		Geometric mean	85.55		85.55	
			All variables	Selected features	All variables	Selected features
SVM	Train	Accuracy	98.62	98.62	94.37	93.84
		Sensitivity	62.07	64.14	59.31	62.76
		Specificity	99.48	99.43	98.1	97.14
		Geometric mean	78.58	79.86	76.28	78.08
	Test	Accuracy	98.41	98.73	98.41	98.31
		Sensitivity	69.57	86.96	69.57	73.91
		Specificity	99.13	99.02	99.13	98.92
		Geometric mean	83.05	92.79	83.05	85.51
DT	Train	Accuracy	98.92	98.92	95.83	95.83
		Sensitivity	62.76	64.14	68.97	70.34
		Specificity	99.77	99.74	98.68	98.53
		Geometric mean	79.13	79.98	82.50	83.25
	Test	Accuracy	98.31	98.52	97.67	97.89
		Sensitivity	65.22	69.57	65.22	69.57
		Specificity	99.13	99.24	98.48	98.59
		Geometric mean	80.41	83.09	80.14	82.82

Table 4. Comparison of AUC between LR, SVM, and DT models in the training set. A * indicated p < 0.05. AS, all samples; RS, reduced samples; AV, all variables; SF, selected features.

		LR		SVM				DT			
		AS	RS	(AS + AV)	(AS + SF)	(RS + AV)	(RS + SF)	(AS + AV)	(AS + SF)	(RS+ AV)	(RS + SF)
LR	AS										
	RS	0.6575									
SVM	(AS + AV)	0.7481	0.6785								
	(AS + SF)	0.4121	0.7075	0.2473							
	(RS + AV)	0.9151	0.9161	0.6619	0.6652						
	(RS + SF)	0.3502	0.5965	0.4135	0.9939	0.5346					
DT	(AS + AV)	0.0001*	0.0001*	0.0001*	0.0002*	0.0002*	0.0002*				
	(AS + SF)	0.0001*	0.0002*	0.0001*	0.0002*	0.0002*	0.0002*	0.3578			
	(RS + AV)	0.0542	0.0618	0.0543	0.0713	0.0658	0.0703	0.0009*	0.0010*		
	(RS + SF)	0.0566	0.0643	0.0567	0.0743	0.0684	0.0731	0.0008*	0.0009*	0.3570	

LR: Logistic regression; SVM: support vector machine; DT: decision tree; AS: all samples; RS: reduced samples; AV: all variables; SF: selected features. * indicated p < 0.05

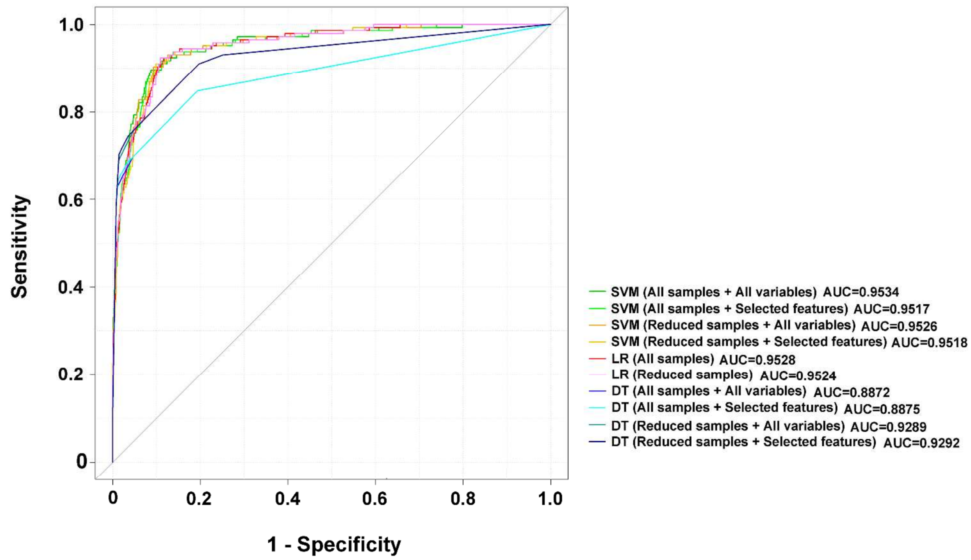


Figure 1. ROC curves for LR, SVM, and DT models in predicting mortality of motorcycle riders.

470x284mm (300 x 300 DPI)

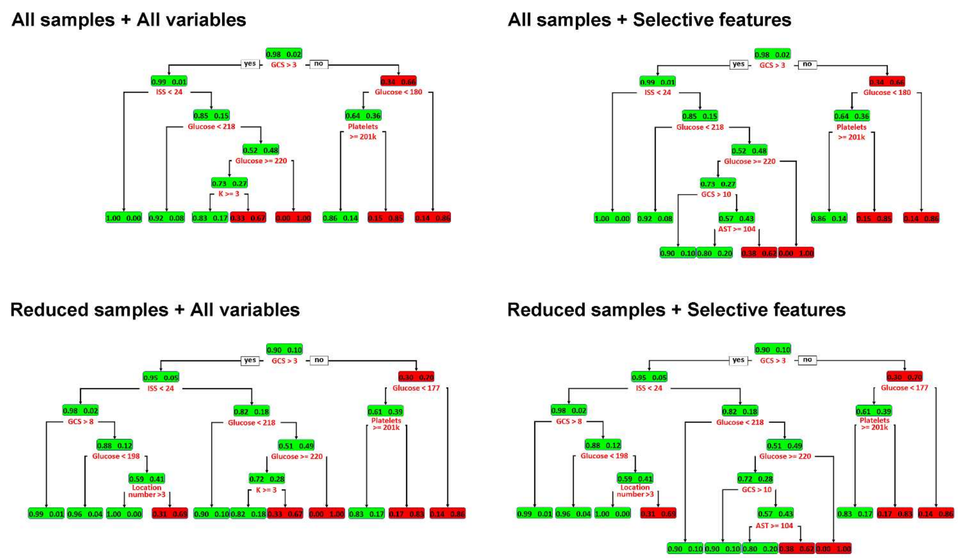
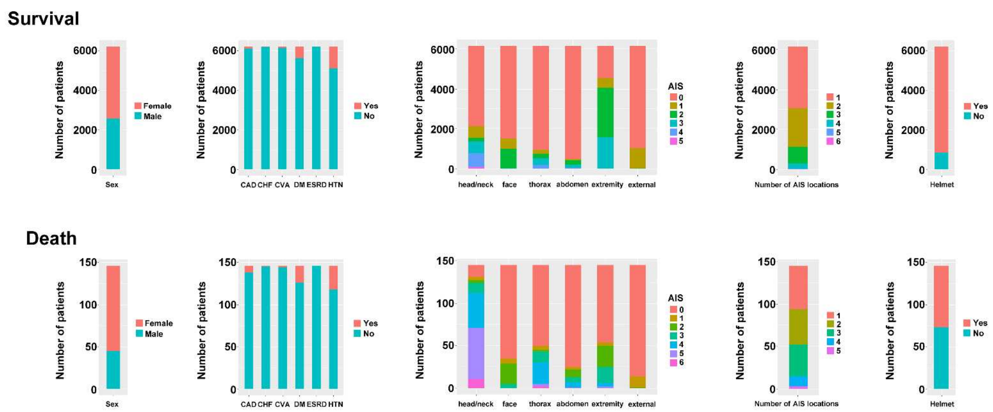
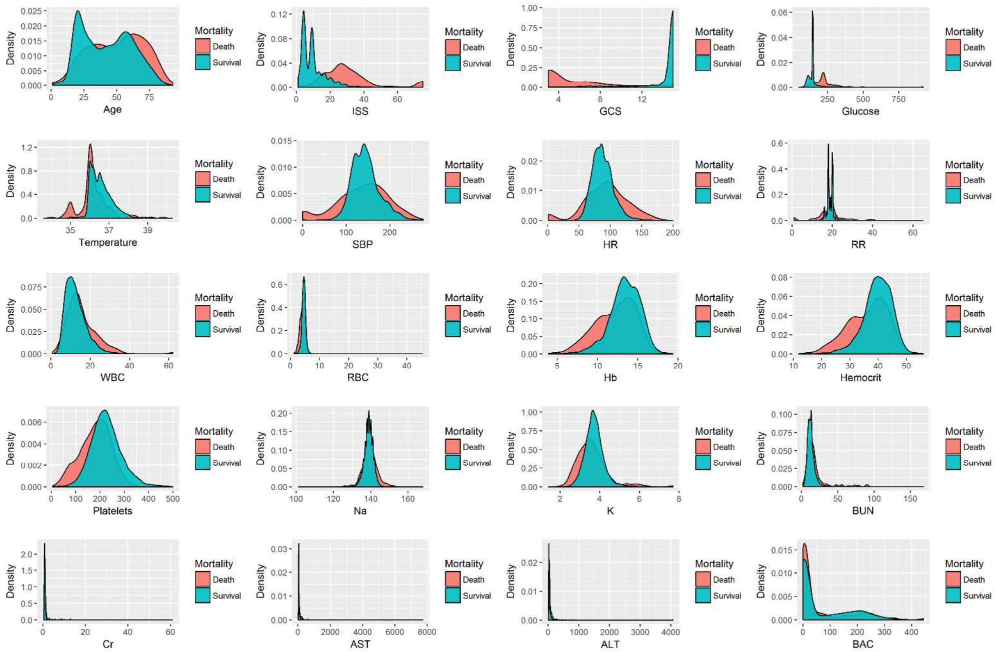


Figure 2. Illustration of DT model for mortality of motorcycle riders. The boxes denote the percentage of patients with discriminating variables from CART analysis. Those who were survival and fatal were indicated with green and red colors, respectively, in the boxes.

340x199mm (300 x 300 DPI)



800x337mm (300 x 300 DPI)



381x254mm (300 x 300 DPI)

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cross-sectional studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	4
Methods			
Study design	4	Present key elements of study design early in the paper	7
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	8
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	8
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	8-11
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	8
Bias	9	Describe any efforts to address potential sources of bias	-
Study size	10	Explain how the study size was arrived at	7
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	7-8
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	7-8
		(*b*) Describe any methods used to examine subgroups and interactions	7-8
		(*c*) Explain how missing data were addressed	-
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	7-8
		(*e*) Describe any sensitivity analyses	-
Results			

Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	7
		(b) Give reasons for non-participation at each stage	-
		(c) Consider use of a flow diagram	-
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	9-11
		(b) Indicate number of participants with missing data for each variable of interest	-
Outcome data	15*	Report numbers of outcome events or summary measures	-
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	11
		(*b*) Report category boundaries when continuous variables were categorized	-
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	-
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	-
Discussion			
Key results	18	Summarise key results with reference to study objectives	11-18
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	18
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	11
Generalisability	21	Discuss the generalisability (external validity) of the study results	-
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	19

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

BMJ Open

Derivation and validation of different machine learning models in mortality prediction of trauma motorcycle riders - a cross-sectional retrospective study in southern Taiwan

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018252.R2
Article Type:	Research
Date Submitted by the Author:	06-Nov-2017

Complete List of Authors:	Kuo, Pao-Jen; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Wu, Shao-Chun; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Anesthesiology Chien, Peng-Chen; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Rau, Cheng-Shyuan; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Neurosurgery Chen, Yi-Chun; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Hsieh, Hsiao-Yun; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery Hsieh, Ching-Hua; Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Department of Plastic and Reconstructive Surgery

Primary Subject Heading:	Public health
Secondary Subject Heading:	Public health, Research methods
Keywords:	Motorcycle accident, mortality, machine learning (ML), logistic regression (LR), support vector machine (SVM), decision tree (DT)

Derivation and validation of different machine learning models in mortality prediction of trauma motorcycle riders - a cross-sectional retrospective study in southern Taiwan

Pao-Jen Kuo[1†], M.D.; Shao-Chun Wu[2†] M.D.; Peng-Chen Chien[1], M.Sc.; Cheng-Shyuan Rau[3], M.D.; Yi-Chun Chen[1], M.Sc.; Hsiao-Yun Hsieh[1], M.Sc.; Ching-Hua Hsieh[1*], M.D., Ph.D.

[†] Indicates equal contribution in authorship as the first author.

[1] Department of Plastic and Reconstructive Surgery, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan, 833

[2] Department of Anesthesiology, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan, 833

[3] Department of Neurosurgery, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan, 833

Pao-Jen Kuo; email: bow110470@gmail.com

Shao-Chun Wu; email: shaochunwu@gmail.com

Peng-Chen Chien; e-mail: VENU_CHIEN@hotmail.com

Cheng-Shyuan Rau; e-mail: ersh2127@cloud.cgmh.org.tw

Yi-Chun Chen; e-mail: libe320@yahoo.com.tw

Hsiao-Yun Hsieh; e-mail: sylvia19870714@hotmail.com

Ching-Hua Hsieh; e-mail: m93chinghua@gmail.com

Corresponding author: Ching-Hua Hsieh, M.D., PhD.

Department of Trauma Surgery & Plastic and Reconstructive Surgery, Kaohsiung

Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Taiwan

No.123, Ta-Pei Road, Niao-Song District, Kaohsiung City 833, Taiwan

Tel: 886-7-7317123 ext: 8002; E-mail: m93chinghua@gmail.com

ABSTRACT

Objectives: This study aimed to build and test the models of machine learning (ML) to predict the mortality of hospitalized motorcycle riders.

Setting: The study was conducted in a level 1 trauma center in southern Taiwan.

Participants: Motorcycle riders who were hospitalized between January 2009 and December 2015 were classified into a training set (n=6,306) and test set (n= 946). Using the demographic information, injury characteristics, and laboratory data of patients, logistic regression (LR), support vector machine (SVM), and decision tree (DT) analyses were performed to determine the mortality of individual motorcycle riders, under different conditions, using all samples or reduced samples as well as all variables or selected features in the algorithm.

Primary and secondary outcome measures: The predictive performance of the model was evaluated based on accuracy, sensitivity, specificity, and geometric mean, and an analysis of the area under the receiver operating characteristic curves of the two different models was carried out.

Results: In the training set, both LR and SVM had a significantly higher AUC than DT; no significant difference was observed in the AUC of LR and SVM, regardless of whether all samples or reduced samples and whether all variables or selected features were used. In the test set, the performance of the SVM model for all samples with selected features was better than that of all other models, with an accuracy of 98.73%, sensitivity of 86.96%, specificity of 99.02%, geometric mean of 92.79%, and AUC of 0.9517, in mortality prediction.

Conclusion: ML can provide a feasible level of accuracy in predicting the mortality of motorcycle riders. Integration of the ML model, particularly the SVM algorithm in the trauma system, may help identify high-risk patients, and therefore, guide appropriate interventions by the clinical staff.

KEY WORDS: Motorcycle accident; mortality; machine learning (ML); logistic regression (LR); support vector machine (SVM); and decision tree (DT)

ARTICLE SUMMARY

STRENGTHS AND LIMITATIONS OF THIS STUDY

- This study first used machine learning to predict the mortality risk of motorcycle riders.
- The SVM model generally works like a black box and cannot identify the relationship between mortality and various explanatory variables.
- The incomplete records of patients and exclusion of those who were declared dead in the trauma registry system could cause result bias.
- The single-center setting may limit the generalizability of the results.

BACKGROUND

Motorcycle use is popular in numerous cities because it is a less expensive and convenient means of transportation. However, despite the less travel time, motorcycle riders who are involved in road traffic accidents tend to have a significantly high morbidity and mortality rate. Compared to other riders of motor vehicles, motorcycle riders are 8 times more likely to be injured per vehicle mile,[1] and they are also 30 times more likely to die in a motor vehicle crash[2] and 58 times more likely to be killed on a per-trip basis.[3] In Taiwan, motorcyclist fatalities account for nearly 60% of all driving fatalities,[4] which are often associated with gender (men), advanced age, lack of helmet use, unlicensed status, and driving under the influence of alcohol.[5-9] In addition, head injury is the leading cause of mortality, followed by thoracic and abdominal injuries.[6-9]

Identifying patients who are at high risk is important for the integration of trauma management to maximize resources and improve quality of care.[10, 11] More robust and accurate individual predictions of mortality using better models might provide clinicians with more precise information about the likelihood of good or poor outcomes and improve individual trauma and mortality management.[12] To identify the possibility of mortality, the Trauma and Injury Severity Score (TRISS) is frequently used, which was established in 1987, to estimate the survival probability of an individual trauma patient based on logistic regression (LR) analysis of variables, including age, anatomical variable (Injury Severity Score [ISS]), physiological variable (Revised Trauma Score [RTS]), and different coefficients for blunt and penetrating injuries. However, the TRISS has limitations and fails to determine an accurate classification in 15–30% of trauma patients.[13] Even after the incorporation of other or revised predictors, such as blood pressure,[14] co-morbidities, and separate categories for different age groups,[15] into this model, the addition of more predictors to the basic TRISS model did not always result in higher performance.[16-18] Although the revised TRISS derived from the USA National Trauma Database for trauma systems is inaccurate, particularly in the management of predominantly blunt injuries,[19] further development of the model based on advanced methodological quality, performance in the subsets of patient groups, and practical application is required for the prediction of mortality.[16]

Currently, machine learning (ML) had been successfully applied in real-life settings in several fields of study, including automatic medical diagnostics and personalized health care.[20-22] The application of supervised ML methods to aid diagnosis and prognosis in trauma patients has been a topic of interest. ML is based on how the human brain approaches pattern recognition tasks, thus providing an artificial intelligence-based approach to solve classification problems and improving their efficiency over time.[23] The usefulness of ML is bolstered by the versatility of its techniques and utility for artificial intelligence, such as prediction, classification, planning, recognition, and clustering.[23 24] Different learning strategies were previously compared using field-specific datasets, of which several had a significantly better predictive power than the more conventional alternatives.[25] Examples of multivariate techniques for pattern recognition include but are not limited to LR, support vector machine (SVM), decision tree (DT), and artificial neural networks. LR is a widely used and accepted statistical analysis tool that predicts the probability of the occurrence of an event.[26] It aims to build a functional relationship between two or more independent predictors and one dependent outcome variable, with the assumption that the response variables are linearly related to the coefficients of the predictor variables.[26]

SVM uses a training set of data with one or more features to determine an optimal boundary that separates a set of cases. The binary SVM classifier establishes a set of optimal hyperplanes in a high-dimensional space with the maximal margin of the two classes.[27] When all training points cannot be separated by the hyperplane, a soft margin method is used to establish a hyperplane that can separate the training data points.[28 29] Moreover, the SVM model can be used for the classification of problems.[30-34]

DT is a hierarchical model that is composed of decision rules based on the optimal feature cutoff values that recursively classify independent variables into different groups.[35-37] It has been built to search for a set of decision rules that can predict an outcome from a set of input variables.[33 35 36] Some models are used to construct DT models, including classification and regression trees (CART), ID3s, chi-square automatic interaction detector DTs (CHAIDs), and C4.5 and C5.0 DTs [26, 28]. CART analysis is a combined approach based on nonparametric and nonlinear variables for recursive partitioning analysis. In addition, it is an innovative DT model in which several predictive variables are used in identifying high-risk patients in various medical fields through progressive binary splits to develop prediction models and to enable better prediction and clinical decision-making.[38-40]

Thus, this study aimed to establish a model for the mortality prediction of motorcycle riders using ML algorithms based on data from a population-based trauma registry in a level I trauma center.

METHODS

Ethics statement

This study was approved by the institutional review board (IRB) of Chang Gung Memorial Hospital (referencing number: 201600653B0). Requirement for informed consent was waived according to the IRB regulations.

Data preparation

Detailed patient information was retrieved from the trauma registry system of our institution, a 2,400-bed facility and level 1 regional trauma center, between January 2009 and December 2015. Only trauma patients who sustained injuries from a motorcycle accident and were hospitalized for treatment were included in the study. Patient information included the following variables: age; sex; use of a helmet; co-morbidities, such as coronary artery disease (CAD), congestive heart failure (CHF), cerebral vascular accident (CVA), diabetes mellitus (DM), end-stage renal disease (ESRD), and hypertension (HTN); vital signs, including temperature, systolic blood pressure (SBP), heart rate (HR), and respiratory rate (RR); ISS; Glasgow coma scale (GCS) score; abbreviated injury scale (AIS) in the different regions of the body; number of injured body regions according to AIS (number of AIS locations); in-hospital mortality; and laboratory values (white blood cell [WBC], red blood cell [RBC], and platelet count; hemoglobin [Hb], hematocrit [Hct], blood urine nitrogen [BUN], creatinine [Cr], alanine aminotransferase [ALT], aspartate aminotransferase [AST], sodium [Na], potassium [K], and glucose level; and blood alcohol concentration [BAC]) upon emergency admission.

Patient samples were divided into a training sample, which was used for predictor discovery and supervised classification to generate a plausible model, and a test sample, which was used to test the performance of the model that was generated in the training sample. Patients with missing data were not included for further analysis. Those who registered within the 6-year period between January 2009 and December 2014 were included in the training set, with a total of 6,306 patients. The group was composed of 6,161 survivors and 145 patients who died. In the test set, 946 patients were included, of which 923 survived and 23 died, within the 1-year period between January 2015 and December 2015. The sample similarity was assessed based on Euclidean distance for the quantitative data to reduce the sample that was designed for data analysis.[41] The sample reduction used the Euclidean distance of the `dist` function in the `stats` package in R (R Foundation for Statistical Computing, Vienna, Austria). During sample reduction, the data size can be reduced to speed up calculations in the analysis.[42] However, considering the exploratory nature of this study, all samples (n=6,306) and reduced samples (n=1,510) in the training set of this study must be analyzed during ML classification.

ML classifiers

The present study provides a performance comparison of the three different ML classifiers (LR, SVM, and DT).

Logistic regression

The LR classifier used the `glm` function in the `stats` package in R3.3.3 (R Foundation for Statistical Computing, Vienna, Austria). Univariate LR analyses were initially performed to identify the significant predictor variables of the mortality risk. A stepwise LR analysis was carried out to control the effects of the confounding variables that help identify the independent risk factors of mortality. The selected independent risk factors obtained from LR were also used as selected features for the implementation of the SVM and DT to explain their importance in determining mortality risk.

Support vector machine

The SVM classifier used the `tune.svm and svm` function in the `e1071` package in R. In the training set, the SVM classifier was used for the prediction of mortality with regard to either all 32 variables or 12 selected features as well as all samples and reduced samples in the training set. The mapping procedure was performed using the kernel function, which is a matrix of pair-wise similarities between data points, such as a linear, polynomial, or radial basis function (RBF).[43] In the present study, the RBF kernel was used because it can control non-linear interactions between class labels and features.[44] The two main parameters presented in the SVM with RBF kernel were the penalty parameter C and kernel hyper-parameter γ. The penalty parameter C determined the tradeoff between the fitting error minimization and model complexity, whereas the hyper-parameter γ defined the nonlinear feature transformation onto a higher dimensional space and controlled the tradeoff between errors due to bias and variance in the model.[45] The optimal operating point was estimated by differentiating the parameter C and γ using a grid search for each combination of feature selection and dimension reduction with a 10-fold cross-validation.[44]

Decision tree

DT by CART that was based on the Gini impurity index used the `rpart` function in the `rpart` package in R. The CART analysis searched for the split on the variable that would partition the data into two different groups: a group of mostly "0s" (people who survived) and "1s" (people who died).[46 47] Using the best overall split, the CART model partitioned the data and assigned a predicted class to each subgroup. CART repeated this same process on each predictor in the model, thus identifying the best split by iteratively testing all possible splits and producing the most significant reduction in impurity.[38-40] CART proceeded recursively in this manner until the specified stopping criteria were met, a specified number of nodes were created, or a further reduction in node impurity was obtained.[38-40]

Performance evaluation

An analysis of the receiver operating characteristic (ROC) curve was carried out to assess and compare the performance of the individual ML models. The predictive ability of the model was evaluated using confusion matrix and via an analysis of the area under the curve (AUC) between the two approaches of ML models.

Confusion matrix and geometric mean

The confusion matrix was used to calculate the accuracy, sensitivity, and specificity of a given model with true negative, true positive, false positive, and false negative values, and thus, it presents accuracy, which represents the overall proportion of correct classifications; sensitivity, which refers to the proportion of true positives that were accurately identified (e.g., percentage of people who were declared dead); and specificity, which refers to the proportion of true negatives that were accurately identified (e.g., percentage of people who survived and were declared dead). In addition, because the geometric mean can provide a good trade-off between sensitivity and specificity in a manner that a better accuracy in both classes leads to a larger value, it was calculated in this study according to the methods used by Sanz J et al.[48]

AUC analysis

To compare the performance of multiple ML classifiers in multiple training data sets, a nonparametric approach was used to analyze the areas under the correlated ROC curves using the `roc and roc.test` function in the `pROC` package in R. This nonparametric approach considers the correlated nature of the data that two or more empirical curves are established based on tests performed on the same individual.[49]

All statistical analyses were performed using SPSS 20.0 (IBM Inc., Chicago, IL, USA) and R 3.3.3. For the categorical variables, the chi-square test was carried out to determine the significance of the association between the predictor and outcome variables. For the continuous variables, the student t-test was conducted to analyze normally distributed data, whereas the Kolmogorov–Smirnov test or Mann–Whitney U test was performed to compare non-normally distributed data. Results were presented as mean ± standard deviation. A p-value < 0.05 was considered statistically significant.

RESULTS

Demographic information and injury characteristics of the patients

Patients with head and neck injury had a higher AIS score. However, patients with injury in the extremities had a lower AIS score compared to those who survived (Table 1 and Supplemental Figure 1). Patients who sustained more body region injuries in the(number of AIS locations) tended to have a higher mortality risk than those who survived. In addition, women and those who did not wear helmets had a higher risk of mortality compared to those who survived (Table 1 and Supplemental Figure 1). A statistically significant difference was observed between patients who died and those who survived in terms of age, ISS, GCS, temperature, platelet count, glucose, Hb, Hct, K, Cr, AST, and ALT levels, as well as CAD incidence (Table 2 and Supplemental Figure 2). As the distribution patterns of Hb and Hct levels as well as AST and ALT levels are highly similar, only one of these two variables (i.e., Hct and AST) was selected for further ML classification to prevent the inclusion of duplicate parameters. Therefore, a total of 32 variables were used for imputation into ML classifiers rather than considering selected features that were obtained by using the independent risk factors identified by the LR given below.

Performance of ML classifiers in the training set

Logistic regression

LR considered 12 predictors (platelet count, glucose, BUN, Cr, AST, and Na levels, age, GCS, temperature, number of AIS locations, ISS, as well as HTN) as independent risk factors for mortality in motorcycle riders for either all samples or reduced samples.

The predictive models were listed as

All samples (n=6,306)

$$Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 4.71648 - 0.00846 * \text{platelet} + 0.01189 * \text{glucose} +$$

$$0.03459 * \mathrm{BUN} + 0.10667 * \mathrm{Cr} + 0.00195 * \mathrm{AST} + 0.09513 * \mathrm{Na} \ +$$

$$0.02533 * \mathrm{age} - 0.39968 * \mathrm{GCS} - 0.56396 * \mathrm{temperature} \ - 0.93232 *$$

$$\text{number of AIS locations} \ + 0.14098 * \mathrm{ISS} - 0.95726 * \ \mathrm{HTN}$$

Reduced samples (n=1,510)

$$Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 5.76780 - 0.00763 * \mathrm{platelet} + 0.00953 * \mathrm{glucose} +$$

$$0.03773 * \mathrm{BUN} + 0.00152 * \mathrm{AST} + 0.08630 * \mathrm{Na} \ + 0.02014 * \mathrm{age} -$$

$$0.34116 * \mathrm{GCS} - 0.53370 * \mathrm{temperature} \ - 0.91439 *$$

$$\text{number of AIS locations} + 0.12191 * \mathrm{ISS} - 1.00522 * \ \mathrm{HTN}$$

The LR had an accuracy of 98.64% (sensitivity of 59.31% and specificity of 99.56%) and 94.44% (sensitivity of 60.00% and specificity of 98.10%) for all samples and reduced samples, respectively. The AUCs for all samples and reduced samples were 0.9528 and 0.9524, respectively, (Figure 1).

Support vector machine

In the training set, the SVM classifier was performed for the prediction of mortality considering either all 32 variables or the 12 selected features in all samples and reduced samples, respectively. With the use of the RBF kernel, the two parameters (C and γ) of the SVM model must be determined. The accuracy was highly robust to small changes in the hyper-parameters. Thus, reasonable choices were obtained by a grid search of 2^x where x is an integer between -8 and 4 for C and between -10 and -2 for γ. The values with the highest 10-fold cross-validation accuracy were C = 0.25 and γ= 0.00390625. Under the input of all variables into the model, the SVM achieved an accuracy of 98.62% (sensitivity of 62.07% and specificity of 99.48%) and 94.37% (sensitivity of 59.31% and specificity of 98.10%) for all samples and reduced samples, respectively, (Table 3). The AUCs for all samples and reduced samples were 0.9534 and 0.9526, respectively, (Figure 1). With the use of the selected features in the model, the SVM had an accuracy of 98.62% (sensitivity of 64.14% and specificity of 99.43%) and 93.84% (sensitivity of 62.76% and specificity of 97.14%) (Table 3) as well as and AUC values of 0.9517 and 0.9518 for all samples and reduced samples, respectively, (Figure 1).

Decision tree

As shown in Figure 2, in the DT model, GCS was identified as the variable of the initial split with an optimal cut-o value of > 3. Among the patients with a GCS higher than 3, glucose level was selected as the variable of the second split at a discrimination level of 180 mg/dL and 177 mg/dL for all samples and reduced samples, respectively. Glucose level below 180 mg/dL or 177 mg/dL for all samples and reduced samples, respectively, was the best predictor of mortality; the next best predictor was platelet count, with an optimal cut-off value of $201 \times 10^3/\mu L$. For the node, in patients with a GCS not greater than 3, ISS below 24, and glucose level below 218 mg/dL, these predictors were considered as significant variables for all samples and reduced samples along with a GCS > 8 and glucose level below 198 mg/dL, and the number of AIS locations ≥ 3 was considered as an additional predictor for the splitting of the reduced samples. With all the variables used in the model, the DT had an accuracy of 98.92% (sensitivity of 62.76% and specificity of 99.77%) and 95.83% (sensitivity of 68.97% and specificity of 98.68%) for all samples and reduced samples, respectively. The AUC values for all samples and reduced samples were 0.8872 and 0.9289, respectively. With the selected features used in the model, the DT had an accuracy of 98.92% (sensitivity of 64.14% and specificity of 99.74%) and 95.83% (sensitivity of 70.34% and specificity of 98.53%) for all samples and reduced samples, respectively. The AUC values for all samples and reduced samples were 0.8872 and 0.9289, respectively, (Figure 1). In the condition wherein reduced samples but not all samples were used in the DT model, the number of AIS locations would be added in the split of the node, thus slightly increasing the sensitivity from 62.76% to 68.97% and from 64.14% to 70.34% with the input composed of all variables and selected variables, respectively. In addition, in the condition wherein selected features but not all variables were used in the DT model, the level of K was not used in the splitting of the node and substituted by the cut-off value of AST (≥ 104 IU/L), therefore slightly increasing the sensitivity from 62.76% to 64.14% and from 68.97% to 70.34% with input composed of all samples and reduced samples, respectively. The AUC values for all samples and reduced samples were 0.8875 and 0.9292, respectively, (Figure 1).

Comparison of the results of AUC analysis

When the AUCs for LR, SVM, and DT were used for the training set (Table 4 and Figure 1), both LR and SVM had a significantly higher AUC than DT, regardless of whether all samples or reduced samples and whether all variables or selected features were used. However, no significant difference was observed in the AUC of LR and SVM, regardless whether all samples or reduced samples as well as all variables or selected features were used. In addition, the DT sample reduction had a significantly higher AUC than that obtained using all samples. However, no significant difference was observed in the AUC of DT, regardless whether all variables or selected features were used.

Performance of ML classifiers in test set

In test set, the LR model for all samples and reduced samples had an accuracy of 98.41%, with a sensitivity of 73.91% and specificity of 99.02%, in predicting mortality (Table 3). These four SVM models had an accuracy of more than 98% and a specificity of approximately 99% in predicting mortality. In contrast, the SVM model for all samples with selected features had the highest sensitivity (86.96%) and geometric mean (92.79%). These four DT models had an accuracy of approximately 98% and a specificity of approximately 99% but a sensitivity of less than 70%. Considering that most patients survived and had a significantly high accuracy and specificity index in predicting mortality, the comparison should therefore focus on the sensitivity and geometric mean of the different ML models. All LR and SVM models, but not the DT models, had an increased sensitivity in the test set. In addition, the SVM model for all samples with selected features had the highest sensitivity and geometric mean.

DISCUSSION

LR is widely used in epidemiological studies for causal inference, and with the selection of built-in features, it does not necessarily utilize all the predictors. With a relatively limited number of variables, i.e., variables less than 20, LR provides the estimates of the odd ratios of the risk factors.[50] However, its limitations became apparent when a complex dataset with a high number of relevant exposures and multiple interactions was analyzed.[51] With the use of several predictors, data that can specify all interactions may not be obtained.[51] In addition, the DT with CART analysis was exploratory and not based on a probabilistic method, which may lead to an overestimation of the importance of the risk factors or may cause other potential confounders to be missed, thus affecting each patient's actual risk.[52] In contrast to LR, which is significantly affected by outliers using a linear discriminant analysis method, the SVM boundary is only minimally affected by outliers that are difficult to separate, despite the complexity of data.[53] In addition, the use of kernels in the SVM model is beneficial for non-linear decision boundaries, thus allowing the classifier to solve more difficult classification problems than the linear analysis method.[54] These three ML models (LR, SVM, and DT) all had an accuracy and specificity of approximately 98% and 99%, respectively, but a sensitivity less than or approximately 70% in the training dataset. In this study, both LR and SVM had a significantly higher AUC than DT in the training set, regardless of whether all samples or reduced samples and whether all variables or selected features were used.

This study included the different variants of SVM, considering the sample size and feature selection, to show all possible improvements and conventional strategies, such as LR or DT. Although the sample reduction for SVM had been proposed to significantly improve the training speed of the SVM and save a lot of storage space,[55, 56] kernel use is a more efficient technique for the representation between samples. Thus, the computational complexity of SVM is not wholly governed by the number of samples but by the number of features, which is advantageous for the analysis in high-dimensional settings.[54] In addition, feature selection in SVM may maximize the AUC.[25] When aided by feature selection, the proposed SVM method identifies the most discriminating indexes for mortality prediction. Although both LR and SVM did not have a different AUC in the training procedure, the SVM model for all samples with selected features had a significantly higher sensitivity (86.96%) in predicting the mortality of motorcycle riders in the test set compared to the rest of the models. The higher sensitivity of SVM in the test set compared to that in the training set may be attributed to an improved quality of registered content and less missing data in our registered data after continuous quality assessment and years of working experience with the registers. Such increased sensitivity was also found in the LR model in the test set. With the addition of more data in the model, the SVM model may have an increased predictive power. In the present study, the feasibility of using SVM classification with feature selection can predict the mortality risk of motorcycle riders admitted in trauma care centers. However, the SVM model generally works like a black box, and it cannot identify the relationships between mortality and various explanatory variables. Therefore, this model cannot be directly used to validate our hypothesis on the increased sensitivity in the test set.

This study has several limitations. First, the patients who had incomplete records were excluded from the analysis. This could have caused result bias, and the results could have been different from the data acquired if the patients with incomplete records were included and the missing data on a variable were replaced by a value that is drawn from an estimate of the distribution of this variable.[57-59] Imputation can include patients who might have relevant features for analysis. However, these patients were excluded due to errors in data collection or recording.[57-59] Second, the exclusion of patients who were declared dead (either upon arriving at the hospital or at the accident area itself) and injured patients who were discharged against the advice of physicians in the emergency department may cause a potential bias. Third, important data regarding injury mechanism and circumstance, including motorcycle speed and type, helmet material, and impact force during collision, were missing. In addition, the imputation of physiological and laboratory data collected from the time of arrival at the emergency department cannot reflect the dynamic changes in hemodynamic and metabolic variables of the trauma patients when resuscitation is possible. Furthermore, other DT-related methods, such as DT by C4.5,[60] combined classifiers of LR and DT by C4.5,[48] and random forest,[61] have extremely satisfying performance in dealing with the classification problem. However, these techniques were not investigated in this study. Lastly, the study population was limited to a single urban trauma center in southern Taiwan, which may not be representative of other populations.

CONCLUSION

ML can provide a feasible level of accuracy in predicting the mortality of motorcycle riders. However, there are significant theoretical and practical challenges to the translational implementation of this approach. The results of previous studies are extremely helpful and may help in establishing the first step towards the development of a prediction model that can be integrated into the trauma care system to identify an individual motorcycle rider's risk of mortality.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

PJK wrote the manuscript. SCW revised the manuscript. PCC performed the statistical analyses and machine learning programming. CSR analyzed the data in the tables. YCC and HYH collected the data and ensured the integrity of the registered data, and CHH designed the study and contributed to the analysis and interpretation of data. All authors have read and approved the final manuscript.

DATA SHARING

No additional data are available.

ACKNOWLEDGEMENTS

This research was supported by a grant from Chang Gung Memorial Hospital (CMRPG8F0891). We would like to thank the personnel of Biostatistics Center, Kaohsiung Chang Gung Memorial Hospital, for helping us with the statistical work.

REFERENCES

1. Weiss H, Agimi Y, Steiner C. Youth motorcycle-related brain injury by state helmet law type: United States, 2005-2007. Pediatrics 2010;**126**(6):1149-55.
2. National Highway Traffic Safety Administration (NHTSA) TSFDDH.
3. Beck LF, Dellinger AM, O'Neil ME. Motor vehicle crash injury rates by mode of travel, United States: using exposure-based methods to quantify differences. American journal of epidemiology 2007;**166**(2):212-8.
4. Chang HL, Lai CY. Using travel socialization and underlying motivations to better understand motorcycle usage in Taiwan. Accident; analysis and prevention 2015;**79**:212-20.
5. Jou RC, Yeh TH, Chen RS. Risk factors in motorcyclist fatalities in Taiwan. Traffic injury prevention 2012;**13**(2):155-62.
6. Liang CC, Liu HT, Rau CS, et al. Motorcycle-related hospitalization of adolescents in a Level I trauma center in southern Taiwan: a cross-sectional study. BMC

pediatrics 2015;**15**:105.

7. Liu HT, Liang CC, Rau CS, et al. Alcohol-related hospitalizations of adult motorcycle riders. World journal of emergency surgery : WJES 2015;**10**(1):2.

8. Hsieh CH, Hsu SY, Hsieh HY, et al. Differences between the sexes in motorcycle-related injuries and fatalities at a Taiwanese level I trauma center. Biomed J 2017;**40**(2):113-20.

9. Hsieh CH, Liu HT, Hsu SY, et al. Motorcycle-related hospitalizations of the elderly. Biomed J 2017;**40**(2):121-28.

10. Densmore JC, Lim HJ, Oldham KT, et al. Outcomes and delivery of care in pediatric injury. Journal of pediatric surgery 2006;**41**(1):92-8; discussion 92-8.

11. Rogers SC, Campbell BT, Saleheen H, et al. Using trauma registry data to guide injury prevention program activities. The Journal of trauma 2010;**69**(4 Suppl):S209-13.

12. Norrie J. Mortality prediction in ICU: a methodological advance. The Lancet Respiratory medicine 2015;**3**(1):5-6.

13. Demetriades D, Chan L, Velmanos GV, et al. TRISS methodology: an inappropriate tool for comparing outcomes between trauma centers. J Am Coll Surg 2001;**193**(3):250-4.

14. Jones JM, Skaga NO, Sovik S, et al. Norwegian survival prediction model in trauma: modelling effects of anatomic injury, acute physiology, age, and co-morbidity. Acta Anaesthesiol Scand 2014;**58**(3):303-15.

15. Bergeron E, Rossignol M, Osler T, et al. Improving the TRISS methodology by restructuring age categories and adding comorbidities. J Trauma 2004;**56**(4):760-7.

16. de Munter L, Polinder S, Lansink KW, et al. Mortality prediction models in the general trauma population: A systematic review. Injury 2017;**48**(2):221-29.

17. Fueglistaler P, Amsler F, Schuepp M, et al. Prognostic value of Sequential Organ Failure Assessment and Simplified Acute Physiology II Score compared with trauma scores in the outcome of multiple-trauma patients. Am J Surg 2010;**200**(2):204-14.

18. Kroezen F, Bijlsma TS, Liem MS, et al. Base deficit-based predictive modeling of outcome in trauma patients admitted to intensive care units in Dutch trauma centers. J Trauma 2007;**63**(4):908-13.

19. Stoica B, Paun S, Tanase I, et al. Probability of Survival Scores in Different Trauma Registries: A Systematic Review. Chirurgia (Bucur) 2016;**111**(2):115-9.

20. Cohen AM, Ambert K, McDonagh M. A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. AMIA Annual Symposium proceedings AMIA Symposium 2010;**2010**:121-5.

21. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2016.

22. Szlosek DA, Ferrett J. Using Machine Learning and Natural Language Processing Algorithms to Automate the Evaluation of Clinical Decision Support in Electronic Medical Record Systems. EGEMS (Washington, DC) 2016;**4**(3):1222.

23. Oquendo MA, Baca-Garcia E, Artes-Rodriguez A, et al. Machine learning and data mining: strategies for hypothesis generation. Mol Psychiatry 2012;**17**(10):956-9.

24. Kotoku J. An Introduction to Machine Learning. Igaku butsuri : Nihon Igaku Butsuri Gakkai kikanshi = Japanese journal of medical physics : an official journal of Japan Society of Medical Physics 2016;**36**(1):18-22.

25. Yahya N, Ebert MA, Bulsara M, et al. Statistical-learning strategies generate only modestly performing predictive models for urinary symptoms following external beam radiotherapy of the prostate: A comparison of conventional and machine-learning methods. Med Phys 2016;**43**(5):2040.

26. Siuly, Yin X, Hadjiloucas S, et al. Classification of THz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers. Comput Methods Programs Biomed 2016;**127**:64-82.

27. V. V. Statistical learning theory. New York, NY: John Wiley & Sons 1998.

28. de Boves Harrington P. Support Vector Machine Classification Trees. Anal Chem 2015;**87**(21):11065-71.

29. Lee Y. Support vector machines for classification: a statistical portrait. Methods Mol Biol 2010;**620**:347-68.

30. Chen C, Zhang G, Qian Z, et al. Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accid Anal Prev 2016;**90**:128-39.

31. Galatzer-Levy IR, Karstoft KI, Statnikov A, et al. Quantitative forecasting of PTSD from early trauma responses: a Machine Learning application. J Psychiatr Res 2014;**59**:68-76.

32. Li Z, Liu P, Wang W, et al. Using support vector machine models for crash injury severity analysis. Accid Anal Prev 2012;**45**:478-86.

33. Marucci-Wellman HR, Corns HL, Lehto MR. Classifying injury narratives of large administrative databases for surveillance-A practical approach combining machine learning ensembles and human review. Accid Anal Prev 2017;**98**:359-71.

34. Patil BM, Joshi RC, Toshniwal D, et al. A new approach: role of data mining in prediction of survival of burn patients. J Med Syst 2011;**35**(6):1531-42.

35. Farion K, Michalowski W, Wilk S, et al. A tree-based decision model to support prediction of the severity of asthma exacerbations in children. J Med Syst 2010;**34**(4):551-62.

36. Zintzaras E, Bai M, Douligeris C, et al. A tree-based decision rule for identifying profile groups of cases without predefined classes: application in diffuse large B-cell lymphomas. Comput Biol Med 2007;**37**(5):637-41.

37. Kasbekar PU, Goel P, Jadhav SP. A Decision Tree Analysis of Diabetic Foot Amputation Risk in Indian Patients. Frontiers in endocrinology 2017;**8**:25.

38. Guilbault RWR, Ohlsson MA, Afonso AM, et al. External Validation of Two Classification and Regression Tree Models to Predict the Outcome of Inpatient Cardiopulmonary Resuscitation. J Viral Hepat 2017;**32**(5):333-38.

39. Shi KQ, Zhou YY, Yan HD, et al. Classification and regression tree analysis of acute-on-chronic hepatitis B liver failure: Seeing the forest for the trees. 2017;**24**(2):132-40.

40. Zimmerman RK, Balasubramani GK, Nowalk MP, et al. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. BMC Infect Dis 2016;**16**(1):503.

41. Amaratunga D, Cabrera J, Lee YS. Resampling-based similarity measures for high-dimensional data. J Comput Biol 2015;**22**(1):54-62.

42. Bhattacharya S, Mariani TJ. Transformation of expression intensities across generations of Affymetrix microarrays using sequence matching and regression modeling. Nucleic Acids Res 2005;**33**(18):e157.

43. Vapnik VN. The Nature of Statistical Learning Theory. New York, 2nd ed. 2000.

44. Gultepe E, Green JP, Nguyen H, et al. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support

system. J Am Med Inform Assoc 2014;**21**(2):315-25.

45. Chen H, Hu L, Li H, et al. An Effective Machine Learning Approach for Prognosis of Paraquat Poisoning Patients Using Blood Routine Indexes. Basic Clin Pharmacol Toxicol 2017;**120**(1):86-96.

46. Chang LY, Wang HW. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid Anal Prev 2006;**38**(5):1019-27.

47. Ripley B. tree: Classification and regression trees. R package version 1.0-34. URL ：http://CRAN.R-project.org/package=tree. 2013.

48. Sanz J, Paternain D, Galar M, et al. A new survival status prediction system for severe trauma patients based on a multiple classifier system. Comput Methods Programs Biomed 2017;**142**:1-8.

49. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;**44**(3):837-45.

50. Knol MJ, Vandenbroucke JP, Scott P, et al. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. Am J Epidemiol 2008;**168**(9):1073-81.

51. Gu W, Vieira AR, Hoekstra RM, et al. Use of random forest to estimate population attributable fractions from a case-control study of Salmonella enterica serotype Enteritidis infections. Epidemiol Infect 2015;**143**(13):2786-94.

52. Lemon SC, Roy J, Clark MA, et al. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. nn Behav Med 2003;**26**(3):172-81.

53. Chen S, Zhou S, Yin FF, et al. Investigation of the support vector machine

algorithm to predict lung radiation-induced pneumonitis. Med Phys 2007;**34**(10):3808-14.

54. Orru G, Pettersson-Yeo W, Marquand AF, et al. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci Biobehav Rev 2012;**36**(4):1140-52.

55. Du Hongle LQ, and Cao Jing. Reduce the Samples for SVM Based on Euclidean Distance. 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization 2013.

56. R. H. Laskar FAT, Biman Paul and Debmalya Chakrabarty. Sample reduction using recursive and segmented data structure analysis. Journal of Engineering and Computer Innovations Vol 2(4), pp 59-67, 2011.

57. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;**59**(10):1087-91.

58. Shrive FM, Stuart H, Quan H, et al. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC medical research methodology 2006;**6**:57.

59. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. J Clin Epidemiol 2002;**55**(4):329-37.

60. Wiharto W, Kusnanto H, Herianto H. Interpretation of Clinical Data Based on C4.5 Algorithm for the Diagnosis of Coronary Heart Disease. Healthcare informatics research 2016;**22**(3):186-95.

61. Rigatti SJ. Random Forest. J Insur Med 2017;**47**(1):31-39.

Figure Legend

Figure 1. ROC curves for LR, SVM, and DT models in predicting mortality of

motorcycle riders.

Figure 2. Illustration of DT model for mortality of motorcycle riders. The boxes denote the percentage of patients with discriminating variables from CART analysis. Those who were survival and fatal were indicated with green and red colors, respectively, in the boxes.

Supplemental Figure 1. Demographics and injury characteristics of the patients regarding categorical variables.

Supplemental Figure 2. Injury characteristics of the patients regarding continuous variables.

Tables

Table 1. Demographics and injury characteristics of the patients regarding gender, helmet-wearing status, co-morbidities, injury region, and number of injury regions.

Variables		Total (n = 7252)	Survival (n = 7084)	Mortality (n = 168)	*P*-value
Sex	Female	4291 (59.2%)	4174 (58.9%)	117 (69.6%)	0.005
	Male	2961 (40.8%)	2910 (41.1%)	51 (30.4%)	
Helmet	NO	1011 (13.9%)	929 (13.1%)	82 (48.8%)	<0.001
	YES	6241 (86.1%)	6155 (86.9%)	86 (51.2%)	
DM	NO	6562 (90.5%)	6414 (90.5%)	148 (88.1%)	0.286
	YES	690 (9.5%)	670 (9.5%)	20 (11.9%)	
HTN	NO	5939 (81.9%)	5802 (81.9%)	137 (81.5%)	0.919
	YES	1313 (18.1%)	1282 (18.1%)	31 (18.5%)	
CAD	NO	7120 (98.2%)	6960 (98.2%)	160 (95.2%)	0.011
	YES	132 (1.8%)	124 (1.8%)	8 (4.8%)	

CHF	NO	7228 (99.7%)	7061 (99.7%)	167 (99.4%)	0.431
	YES	24 (0.3%)	23 (0.3%)	1 (0.6%)	
CVA	NO	7168 (98.8%)	7002 (98.8%)	166 (98.8%)	0.722
	YES	84 (1.2%)	82 (1.2%)	2 (1.2%)	
ESRD	NO	7250 (100%)	7082 (100%)	168 (100%)	1.000
	YES	2 (0.0%)	2 (0.0%)	0 (0.0%)	
AIS (Head/Neck)	0	4642 (64%)	4627 (65.3%)	15 (8.9%)	<0.001
	1	665 (9.2%)	661 (9.3%)	4 (2.4%)	
	2	192 (2.6%)	189 (2.7%)	3 (1.8%)	
	3	713 (9.8%)	699 (9.9%)	14 (8.3%)	
	4	840 (11.6%)	795 (11.2%)	45 (26.8%)	
	5	189 (2.6%)	113 (1.6%)	76 (45.3%)	
	6	11 (0.2%)	0 (0%)	11 (6.5%)	
AIS (Face)	0	5472 (75.4%)	5347 (75.5%)	125 (74.4%)	<0.001
	1	574 (7.9%)	568 (8%)	6 (3.6%)	
	2	1173 (16.2%)	1141 (16.1%)	32 (19%)	
	3	33 (0.5%)	28 (0.4%)	5 (3%)	
AIS (Thorax)	0	6081 (83.9%)	5973 (84.3%)	108 (64.3%)	<0.001
	1	234 (3.2%)	229 (3.3%)	5 (3%)	
	2	260 (3.6%)	258 (3.6%)	2 (1.2%)	
	3	423 (5.8%)	404 (5.7%)	19 (11.3%)	
	4	245 (3.4%)	217 (3.1%)	28 (16.7%)	
	5	7 (0.1%)	3 (<0.1%)	4 (2.4%)	
	6	2 (<0.1%)	0 (0%)	2 (1.1%)	
AIS (Abdomen)	0	6654 (91.8%)	6516 (92%)	138 (82.1%)	<0.001
	1	57 (0.8%)	54 (0.8%)	3 (1.8%)	
	2	288 (4%)	277 (3.9%)	11 (6.5%)	
	3	170 (2.2%)	163 (2.3%)	7 (4.2%)	
	4	66 (0.9%)	58 (0.8%)	8 (4.8%)	
	5	17 (0.2%)	16 (0.2%)	1 (0.6%)	
AIS (Extremity)	0	2000 (27.6%)	1897 (26.8%)	103 (61.3%)	<0.001
	1	528 (7.3%)	524 (7.4%)	4 (2.4%)	
	2	2886 (39.8%)	2853 (40.3%)	33 (19.6%)	
	3	1822 (25.1%)	1800 (25.4%)	22 (13.1%)	
	4	12 (0.2%)	8 (0.1%)	4 (2.4%)	
	5	4 (0.1%)	2 (0.0%)	2 (1.2%)	
AIS (External)	0	6155 (84.9%)	6001 (84.7%)	154 (91.7%)	0.003
	1	1072 (14.8%)	1059 (14.9%)	13 (7.7%)	

	2	25 (0.3%)	24 (0.3%)	1 (0.6%)	
Number of AIS locations	1	3687 (50.8%)	3631 (51.3%)	56 (33.3%)	<0.001
	2	2255 (31.1%)	2205 (31.1%)	50 (29.8%)	
	3	982 (13.5%)	939 (13.3%)	43 (25.6%)	
	4	280 (3.9%)	265 (3.7%)	15 (8.9%)	
	5	43 (0.6%)	39 (0.6%)	4 (2.4%)	
	6	5 (0.1%)	5 (0.1%)	0 (0.0%)	

Table 2. Injury characteristics of the patients regarding laboratory data collected from the time point when arrival at the emergency department.

Variables	**Total (n = 7252)**	**Survival (n = 7084)**	**Mortality (n = 168)**	***P*-value**
Age (years)	38 (29)	37 (29)	47 (32)	<0.001
HR (beats/min)	89 (23)	89 (23)	93 (43)	<0.001
SBP (mmHg)	137 (38)	137 (37)	143 (79)	0.374
RR (times/min)	19 (2)	19 (2)	19 (5)	0.660
Temperature (^{o}C)	36.4 (0.8)	36.4 (0.8)	36.0 (0.5)	<0.001
GCS	15 (5)	15 (3)	3 (3)	<0.001
ISS	13 (12)	13 (13)	29 (11)	<0.001
RBC (10^6/uL)	4.6 (0.8)	4.6 (0. 8)	4.3 (1.1)	<0.001
WBC (10^3/uL)	12.9 (7.7)	12.9 (7.7)	13.2 (8.7)	<0.001
Hb (g/dL)	13.9 (2.5)	13.9 (2.5)	12.9 (3.5)	<0.001
Hct (%)	40.9 (6.8)	41.1 (6.6)	38.6 (9.4)	<0.001
Platelets (10^3/uL)	228 (79)	230 (79)	190 (78)	<0.001
Glucose (mg/dL)	145 (27)	145 (23)	218 (60)	<0.001
Na (mEq/L)	139 (3)	139 (3)	139 (4)	0.094
K (mEq/L)	3.5 (0.6)	3.5 (0.6)	3.4 (0.9)	<0.001
BUN (mg/dL)	12 (6)	12 (5)	14 (8)	<0.001
Cr (mg/dL)	0.8 (0.3)	0.8 (0.3)	1.0 (0.5)	<0.001
AST (U/L)	47 (50)	45 (48)	65 (76)	<0.001
ALT (U/L)	34 (35)	34 (33)	39 (55)	<0.001
BAC (mg/dL)	4.9 (133.0)	4.9 (136.4)	4.9 (62.5)	0.698

Table 3. Summarizes mortality prediction performances regarding accuracy, sensitivity, specificity, and geometric mean with LR, SVM, and DT models in the training and test sets.

			All samples n=6306		Reduced samples n=1510	
			All variables		All variables	
LR	Train	Accuracy	98.64		94.44	
		Sensitivity	59.31		60	
		Specificity	99.56		98.1	
		Geometric mean	76.84		76.72	
	Test	Accuracy	98.41		98.41	
		Sensitivity	73.91		73.91	
		Specificity	99.02		99.02	
		Geometric mean	85.55		85.55	
			All variables	Selected features	All variables	Selected features
SVM	Train	Accuracy	98.62	98.62	94.37	93.84
		Sensitivity	62.07	64.14	59.31	62.76
		Specificity	99.48	99.43	98.1	97.14
		Geometric mean	78.58	79.86	76.28	78.08
	Test	Accuracy	98.41	98.73	98.41	98.31
		Sensitivity	69.57	86.96	69.57	73.91
		Specificity	99.13	99.02	99.13	98.92
		Geometric mean	83.05	92.79	83.05	85.51
DT	Train	Accuracy	98.92	98.92	95.83	95.83
		Sensitivity	62.76	64.14	68.97	70.34
		Specificity	99.77	99.74	98.68	98.53
		Geometric mean	79.13	79.98	82.50	83.25
	Test	Accuracy	98.31	98.52	97.67	97.89
		Sensitivity	65.22	69.57	65.22	69.57
		Specificity	99.13	99.24	98.48	98.59
		Geometric mean	80.41	83.09	80.14	82.82

Table 4. Comparison of AUC between LR, SVM, and DT models in the training set. A * indicated p < 0.05. AS, all samples; RS, reduced samples; AV, all variables; SF, selected features.

		LR		SVM				DT			
		AS	RS	(AS + AV)	(AS + SF)	(RS + AV)	(RS + SF)	(AS + AV)	(AS + SF)	(RS+ AV)	(RS + SF)
LR	AS										
	RS	0.6575									
SVM	(AS + AV)	0.7481	0.6785								
	(AS + SF)	0.4121	0.7075	0.2473							
	(RS + AV)	0.9151	0.9161	0.6619	0.6652						
	(RS + SF)	0.3502	0.5965	0.4135	0.9939	0.5346					
DT	(AS + AV)	0.0001*	0.0001*	0.0001*	0.0002*	0.0002*	0.0002*				
	(AS + SF)	0.0001*	0.0002*	0.0001*	0.0002*	0.0002*	0.0002*	0.3578			
	(RS + AV)	0.0542	0.0618	0.0543	0.0713	0.0658	0.0703	0.0009*	0.0010*		
	(RS + SF)	0.0566	0.0643	0.0567	0.0743	0.0684	0.0731	0.0008*	0.0009*	0.3570	

LR: Logistic regression; SVM: support vector machine; DT: decision tree; AS: all samples; RS: reduced samples; AV: all variables; SF: selected features. * indicated p < 0.05

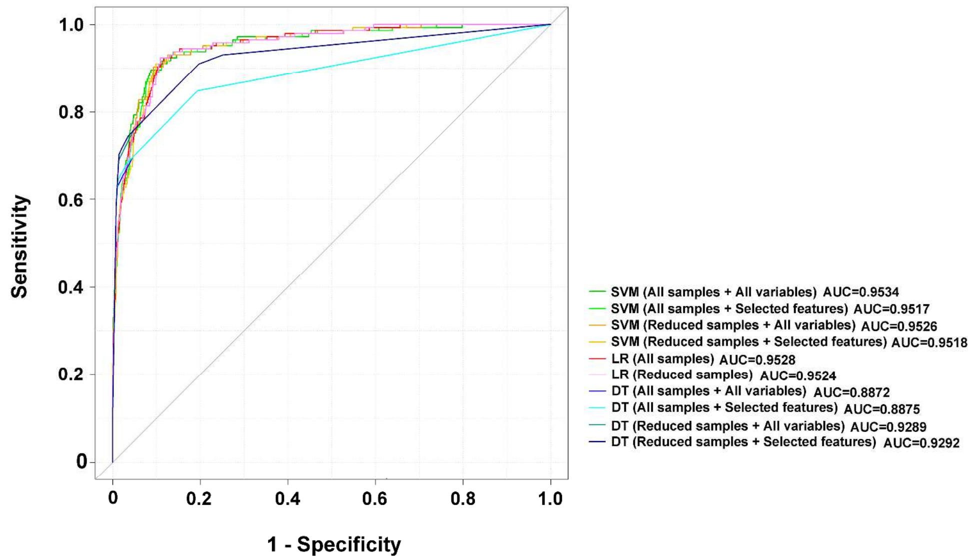


Figure 1. ROC curves for LR, SVM, and DT models in predicting mortality of motorcycle riders.

470x284mm (300 x 300 DPI)

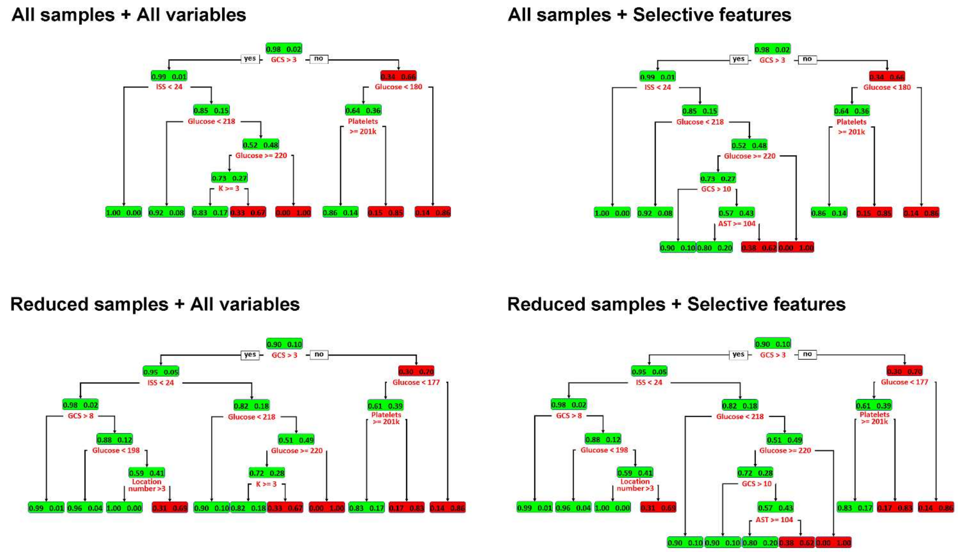
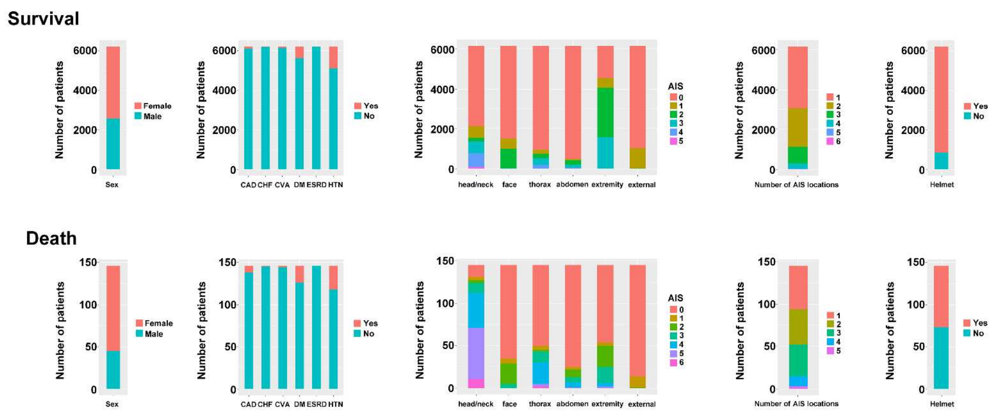
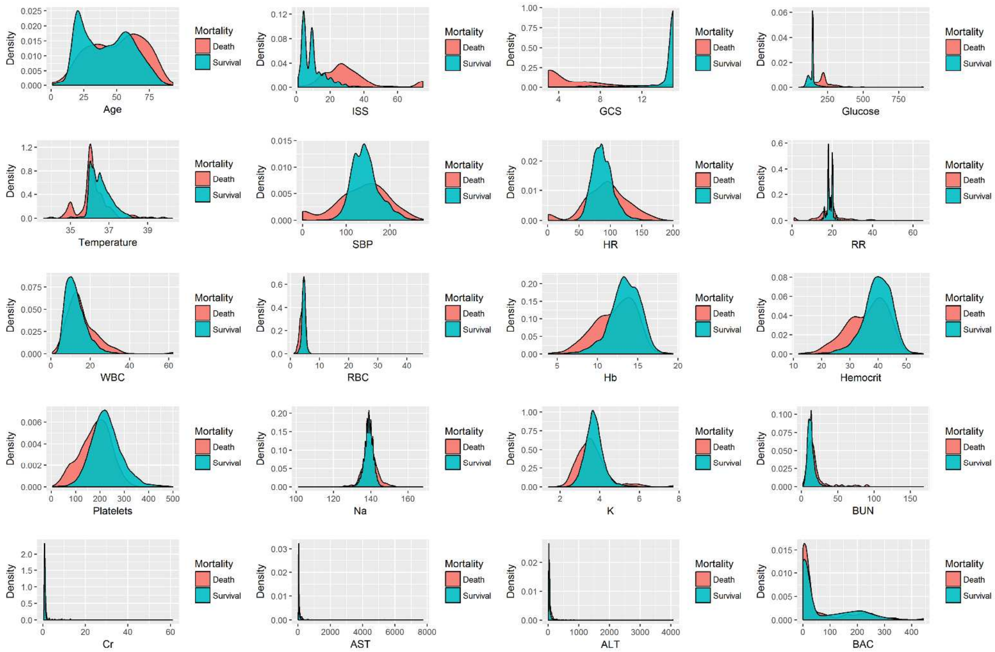


Figure 2. Illustration of DT model for mortality of motorcycle riders. The boxes denote the percentage of patients with discriminating variables from CART analysis. Those who were survival and fatal were indicated with green and red colors, respectively, in the boxes.

340x199mm (300 x 300 DPI)



800x337mm (300 x 300 DPI)



381x254mm (300 x 300 DPI)

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cross-sectional studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	4
Methods			
Study design	4	Present key elements of study design early in the paper	7
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	8
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	8
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	8-11
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	8
Bias	9	Describe any efforts to address potential sources of bias	-
Study size	10	Explain how the study size was arrived at	7
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	7-8
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	7-8
		(*b*) Describe any methods used to examine subgroups and interactions	7-8
		(*c*) Explain how missing data were addressed	-
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	7-8
		(*e*) Describe any sensitivity analyses	-
Results			

Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	7
		(b) Give reasons for non-participation at each stage	-
		(c) Consider use of a flow diagram	-
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	9-11
		(b) Indicate number of participants with missing data for each variable of interest	-
Outcome data	15*	Report numbers of outcome events or summary measures	-
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	11
		(*b*) Report category boundaries when continuous variables were categorized	-
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	-
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	-
Discussion			
Key results	18	Summarise key results with reference to study objectives	11-18
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	18
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	11
Generalisability	21	Discuss the generalisability (external validity) of the study results	-
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	19

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.