

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email editorial.bmjopen@bmj.com

BMJ Open

Mortality prediction of motorcycle riders using machine learning models

Journal:	<i>BMJ Open</i>
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)

SCHOLARONE™
Manuscripts

Mortality prediction of motorcycle riders using machine learning models

Pao-Jen Kuo^{1†}, M.D.; Cheng-Shyuan Rau^{2†}, M.D.; Peng-Chen Chien¹, M.Sc.;
Yi-Chun Chen¹, M.Sc.; Hsiao-Yun Hsieh¹, 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 mile¹, 30 times more likely to die in a motor vehicle crash², and 58 times more likely to be killed on a per-trip basis³. In Taiwan, motorcyclist fatalities account for nearly 60% of all driving fatalities⁴. The fatalities are often associated with men, advanced age, not wearing a helmet, unlicensed status, and riding under the influence of alcohol⁵⁻⁹. In addition, head injuries were the major factor leading to mortality, followed by thoracic and abdominal injuries⁶⁻⁹.

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 management¹². 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 patients¹⁴. Even after the incorporation of other or revised predictors, like blood

1
2
3 pressure¹⁵, co-morbidities, and separate categories for different age-groups¹⁶ into this
4
5 model, the addition of more predictors to the basic TRISS model did not always result
6
7 in higher performance^{13 17 18}. Although the revised TRISS, resulting from the
8
9 USA National Trauma Database is inaccurate for trauma systems, particularly in
10
11 the management of predominantly blunt injuries¹⁹, the further development of the
12
13 model based on advanced methodological quality, the performance of the model in
14
15 subsets of patient groups, and practical application is mandatory in the prediction of
16
17 mortality¹³.
18
19
20
21

22
23 Currently, machine learning (ML) had been successfully applied in the real world
24
25 in many fields including automatic medical diagnostics and personalized health care
26
27 ²⁰⁻²². There is an increasing interest in the application of supervised ML methods to
28
29 aid diagnosis and prognosis in trauma patients. ML is based on the way the human
30
31 brain approaches pattern recognition tasks, providing an artificial intelligence-based
32
33 approach to solve classification problems and improving their efficiency and
34
35 effectiveness over time²³. The usefulness of ML is bolstered by the versatility of its
36
37 techniques and its utility for artificial intelligence such as prediction, classification,
38
39 planning, recognition, and clustering^{23 24}. Comparisons of different learning strategies
40
41 have been conducted previously by others using field-specific datasets, many of which
42
43 have shown significantly better predictive power than the more conventional
44
45 alternatives²⁵. Examples of multivariate techniques for pattern recognition include,
46
47 but are not limited to, LR, support vector machine (SVM), decision trees (DT), and
48
49 artificial neural networks. LR is a widely used and accepted statistical analysis tool to
50
51 predict the probability of the occurrence of an event²⁶. It attempts to build a
52
53 functional relationship between two or more independent predictors and the one
54
55
56
57
58
59
60

1
2
3 dependent outcome variable, under the assumption that the response variables are
4 linearly related to the coefficients of the predictor variables ²⁶.
5
6
7

8
9 SVM uses a training set of data composed of one or more features to determine
10 an optimal boundary separating a set of cases. The binary SVM classifier constructs a
11 set of the optimal hyperplanes in high-dimensional space with the maximal margin of
12 the two classes ²⁷. In the case that all training points cannot be separated by the
13 hyperplane, a soft margin method is used to construct a hyperplane that separates the
14 training data points ^{28,29}. It has been found that the SVM model has a great capability
15 of dealing with classification problems ³⁰⁻³⁴.
16
17
18
19
20
21
22
23
24
25

26 A DT is a hierarchical model composed of decision rules based upon optimal
27 feature cutoff values that recursively split independent variables into different groups
28 ³⁵⁻³⁷. The purpose of DT building is to search for a set of decision rules to predict an
29 outcome from a set of input variables ^{33, 35, 36}. Some models are used to construct
30 decision-tree models, including classification and regression trees (CART), ID3s,
31 chi-square automatic interaction detector DTs (CHAIDs), and C4.5 and C5.0 DTs [26,
32 28]. Among these methods, the CART analysis is a combined approach based on
33 nonparametric and nonlinear variables for recursive partitioning analysis. CART
34 analysis is an innovative DT model in which several predictive variables are crucial to
35 identify patients at different levels of risk in various medical fields through
36 progressive binary splits to develop prediction models in order to enable better
37 prediction and clinical decision-making ³⁸⁻⁴⁰.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54

55 This study is aimed to construct a model for the mortality prediction of
56
57
58
59
60

1
2
3 motorcycle riders using ML algorithms and obtaining data from a population-based
4 trauma registry in a level I trauma center.
5
6
7
8

9 **METHODS**

10 **Ethics statement**

11
12 This study was preapproved by the institutional review board (IRB) of Chang Gung
13 Memorial Hospital with approval number 201600653B0. Informed consent was
14 waived according to the IRB regulations.
15
16
17
18
19
20
21

22 **Data preparation**

23
24 Detailed patient information between January 2009 and December 2015 was
25 retrieved from the Trauma Registry System of our institution, a 2,400-bed facility and
26 Level I regional trauma center. Only the trauma patients who sustained a traffic
27 accident as a motorcycle rider and were hospitalized for treatment were included in
28 the study. The patient information included the following variables: age, sex,
29 helmet-wearing status, co-morbidities such as coronary artery disease (CAD),
30 congestive heart failure (CHF), cerebral vascular accident (CVA), diabetes mellitus
31 (DM), end-stage renal disease (ESRD), and hypertension (HTN) as well as vital signs,
32 including temperature, systolic blood pressure (SBP), heart rate (HR), respiratory
33 rate (RR), ISS, Glasgow coma scale (GCS) score, abbreviated injury scale (AIS) in
34 different regions of the body, number of injured body regions according to AIS
35 (number of AIS locations), the in-hospital mortality, the blood level of white blood
36 cell count (WBC), red blood cell count (RBC), hemoglobin (Hb), hematocrit (Hct),
37 platelets, blood urine nitrogen (BUN), creatinine (Cr), alanine aminotransferase
38 (ALT), aspartate aminotransferase (AST), sodium (Na), potassium (K), blood alcohol
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 concentration (BAC), and glucose at emergency department.
4
5

6
7 These enrolled patients were divided into a training sample, which was used for
8 predictor discovery and supervised classification to generate a plausible model, and a
9 test sample, which was used to test the performance of the model generated in the
10 training sample. Those patients with missing data were not included for further
11 analysis. The patients who registered in a six-year span between January 2009 and
12 December 2014 were allocated in the training set, which comprised of a total of 6,306
13 patients. It included 6,161 survival and 145 mortality patients. In the test set, there
14 were 946 patients, including 923 survival and 23 mortality patients, from the one-year
15 span between January 2015 and December 2015. The sample similarity was assessed
16 by Euclidean distance for quantitative data to reduce the size of a sample designed for
17 use in data analysis⁴¹. The sample reduction used Euclidean distance of the dist
18 function in the stats package in R (R Foundation for Statistical Computing, Vienna,
19 Austria). During sample reduction, the data size can be reduced to speed up
20 calculations in the analysis⁴². However, considering the exploratory character of this
21 study, all samples (n=6,306) and reduced samples (n=1,510) in the training set of this
22 study would require to be analyzed in ML classification.
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43

44 **ML classifiers**

45 This work provides a performance comparison of three different ML classifiers
46 (LR, SVM, and DT).
47
48
49
50
51

52 *Logistic regression*

53 The LR classifier used the glm function in the stats package in R3.3.3 (R
54
55
56
57
58
59
60

1
2
3 Foundation for Statistical Computing, Vienna, Austria). Univariate LR analyses were
4 initially performed to identify the significant predictor variables of the mortality risk.
5
6 Stepwise LR analysis was used to control the effects of confounding variables to
7 identify independent risk factors for mortality. The selected independent risk factors
8
9 obtained from LR were also used as selected features to be implemented by the SVM
10
11 and the DT to explain their weights in determining the risk of mortality.
12
13
14
15

16 17 18 *Support vector machine*

19
20 The SVM classifier used the `tune.svm` & `svm` function in the `e1071` package in
21 R. In the training set, the SVM classifier was performed for the prediction of mortality
22 with regard to either all 32 variables or 12 selected features as well as all the samples
23 and reduced samples in the training set. The mapping procedure was accomplished by
24 the kernel function, which is a matrix of pair-wise similarities between data points,
25 such as a linear, polynomial, or radial basis function (RBF)⁴³. For this study, the RBF
26 kernel was chosen because it can handle non-linear interactions between class labels
27 and features⁴⁴. The two main parameters presented in SVM with RBF kernel were the
28 penalty parameter C and the kernel hyper-parameter γ . The penalty parameter C
29 determined the tradeoff between the fitting error minimization and model complexity,
30 while the hyper-parameter γ defined the nonlinear feature transformation onto a
31 higher dimensional space and controlled the tradeoff between error due to bias and
32 variance in the model.⁴⁵ The optimal operating point was estimated by varying the
33 parameters - C and γ using a grid search for each combination of feature selection and
34 dimension reduction with a 10-fold cross-validation⁴⁴.
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53

54 55 *Decision tree*

1
2
3 The DT by CART based on the Gini impurity index used the rpart function in the
4 rpart package in R. The CART analysis searched for the split on the variable that
5 would partition the data into two different groups—a group of mostly ‘0s’ (people
6 who survived) and a group of mostly ‘1s’ (people who died)^{46 47}. Using the best
7 overall split, the CART model partitioned the data and assigned a predicted class to
8 each subgroup. CART repeated this same process on each predictor in the model,
9 identifying the best split by iteratively testing all possible splits, and producing the
10 greatest reduction in impurity³⁸⁻⁴⁰. CART proceeded recursively in this way until the
11 specified stopping criteria were reached, a specified number of nodes were created, or
12 a further reduction in node impurity became impossible³⁸⁻⁴⁰.

26 **Performance evaluation**

27
28 We used receiver operator characteristic (ROC) curve analysis to assess and
29 compare the performances of the individual ML models. Model predictive ability was
30 evaluated using confusion matrix and the area under curve (AUC) analysis between
31 two approaches of ML models.
32
33
34
35
36
37
38
39
40

41 *Confusion matrix*

42 The confusion matrix calculates the accuracy, sensitivity, and specificity of a
43 given model with true negative, true positive, false positive, and false negative values
44 and presents as a result an accuracy, which represents the overall proportion of correct
45 classifications; a sensitivity, which refers to the proportion of true positives correctly
46 identified (e.g. percentage of people with fatality identified to be dead); and a
47 specificity, which refers to the proportion of true negatives correctly identified (e.g.
48 percentage of people who survived identified as not dead).
49
50
51
52
53
54
55
56
57
58
59
60

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 individuals⁴⁸.

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 \pm 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

1
2
3 difference in age, ISS, GCS, glucose, temperature, Hb, Hct, platelets, K, Cr, AST,
4 ALT, and incidences of CAD was found between patients with fatality and the ones
5 who survived respectively (Figure 2). Because the distribution pattern between Hb
6 and Hct as well as between AST and ALT is very similar, only one of these two
7 variables (i.e. Hct and AST) was selected for further ML classification to prevent the
8 inclusion of duplicate parameters. Therefore, a total of 32 variables were used for
9 imputation into ML classifiers as all variables, in contrast to considering selected
10 features obtained by using the independent risk factors identified by the LR given
11 below.

22 **Performance of ML classifiers in training set**

23 *Logistic regression*

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

27 The predictive models were listed as:

28 All samples (n=6,306)

$$29 \quad Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 4.71648 - 0.00846 * \text{Platelets} + 0.01189 * \text{Glucose} +$$

$$30 \quad 0.03459 * \text{BUN} + 0.10667 * \text{Cr} + 0.00195 * \text{AST} + 0.09513 * \text{Na} +$$

$$31 \quad 0.02533 * \text{Age} - 0.39968 * \text{GCS} - 0.56396 * \text{Temperature} - 0.93232 * \\ 32 \quad \text{Number of AIS locations} + 0.14098 * \text{ISS} - 0.95726 * \text{HTN}$$

33 Reduced samples (n=1,510)

$$34 \quad Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 5.76780 - 0.00763 * \text{Platelets} + 0.00953 * \text{Glucose} +$$

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

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 $\gamma = 0.00390625$. 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)

1
2
3 for all samples and reduced samples, respectively were (Table 1):
4
5
6

7 *Decision tree*

8
9 As shown in Figure 4, in the DT model, GCS was identified as the variable of
10 initial split with an optimal cut-off value of > 3 . Among patients with GCS higher
11 than 3, glucose was selected as the variable of second split at a discrimination level of
12 180 and 177 mg/dL for all samples and reduced samples, respectively. After the
13 glucose level < 180 or 177 mg/dL for all samples and reduced samples, respectively,
14 the next best predictor of mortality was platelets with an optimal cut-off of 201×10^3
15 / μ L. For the node, with patients having a GCS not greater than 3, ISS < 24 and glucose
16 < 218 mg/dL, these predictors were selected as significant variables for all samples
17 and reduced samples, with GCS > 8 , glucose < 198 mg/dL, and the number of AIS
18 locations ≥ 3 being an additional predictors for splitting for the reduced samples. With
19 all variables in the model, the DT achieved an accuracy of 98.92% (sensitivity of
20 62.76% and specificity of 99.77%) and 95.83% (sensitivity of 68.97% and specificity
21 of 98.68%) for the all samples and reduced samples, respectively. The AUCs for all
22 samples and reduced samples were 0.8872 and 0.9289, respectively. With selected
23 features in the model, the DT achieved an accuracy of 98.92% (sensitivity of 64.14%
24 and specificity of 99.74%) and 95.83% (sensitivity of 70.34% and specificity of
25 98.53%) for the all samples and reduced samples, respectively. The AUCs for all
26 samples and reduced samples were 0.8872 and 0.9289, respectively (Figure 3). In the
27 condition of using reduced samples but not all samples in the DT model, the number
28 of AIS locations would be added in the split of the node slightly increasing the
29 sensitivity from 62.76% to 68.97% and from 64.14% to 70.34% with input
30 comprising of all variables and selected variables, respectively. In addition, in the
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56

1
2
3 condition of using selected features but not all variables in the DT model, the level of
4
5 K was not used in the splitting of the node and was substituted by the cut-off value of
6
7 $AST \geq 104$ IU/L, slightly increasing the sensitivity from 62.76% to 64.14% and from
8
9 68.97% to 70.34% with input as all samples and reduced samples, respectively. In
10
11 addition, the AUCs for all samples and reduced sample were 0.8875 and 0.9292,
12
13 respectively (Figure 3)
14
15

16 17 18 *Comparison in AUC analysis*

19
20 In the comparisons of AUCs for LR, SVM, and DT for the training set (Table 2),
21
22 both LR and SVM had a significantly higher AUC than DT, regardless of using all
23
24 samples or reduced samples as well as for all variables or for selected features.
25
26 However, there was no significant difference of AUC between LR and SVM,
27
28 regardless of using all samples or reduced samples as well as for all variables or for
29
30 selected features. In addition to this, in DT sample reduction had a significantly
31
32 higher AUC than the one obtained using all samples, but there was no significant
33
34 difference of AUC between DT with all variables or with selected features.
35
36
37
38
39

40 **Performance of ML classifiers in test set**

41
42 In test set, the LR model for all samples and reduced samples - both achieved an
43
44 accuracy of 98.41%, with a sensitivity of 73.91% and specificity of 99.02% in
45
46 predicting the mortality (Table 1). All of these four SVM models create an accuracy
47
48 more than 98% and a specificity near 99% but a sensitivity of 69.57%, 86.96%,
49
50 69.57%, and 73.91% for all samples and all variables, all samples and selected
51
52 features, reduced samples and all variables, and for reduced samples and selected
53
54 features, respectively, whereas all of these four DT models create an accuracy of
55
56
57
58
59

1
2
3 approximately 98% and a specificity of approximately 99% but a sensitivity of less
4 than 70%. Considering that the majority of patients survived except for a few with
5 fatality, would result into a very high accuracy and specificity index in predicting the
6 mortality, therefore the comparison should further focus on the sensitivity of different
7 ML models. We found that, in the test, all LR and SVM models, but not the DT
8 models, had an increased sensitivity than that in the test set. Furthermore, the SVM
9 model for all samples with selected features had a significantly highest sensitivity
10 (86.96%) in predicting the mortality.
11
12
13
14
15
16
17
18
19
20
21

22 **DISCUSSION**

23
24 LR is widely used in epidemiological studies for causal inference and, with the
25 selection of built-in features; it does not necessarily utilize all the predictors. With a
26 relatively limited number of variables i.e. variables less than 20, LR provides
27 estimates of odd ratios of the risk factors⁴⁹. However, its limitations become apparent
28 when analyzing a complex dataset with a high number of relevant exposures and
29 multiple interactions⁵⁰. With too many predictors, the availability of sufficient
30 information to specify all interactions would become nearly impossible⁵⁰. In addition,
31 the DT with CART analysis is exploratory and not based on the probabilistic method,
32 which may lead to overestimating the importance of included risk factors or cause
33 missing of other potential confounders that could influence each patient's actual risk
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 method ⁵³. These three ML models (LR, SVM, DT) all create an accuracy and a
4
5 specificity around 98% and 99%, respectively, but a sensitivity less than or around
6
7 70% in the training dataset. In this study, both LR and SVM resulted in a
8
9 significantly higher AUC than DT in the training set, regardless of using all samples
10
11 or reduced samples as well as for all variables or for selected features.
12
13
14
15

16 This study included different variants of SVM considering the sample size and
17
18 feature selection to show all possible improvements to the more conventional
19
20 strategies like LR or DT. Although sample reduction for SVM had been proposed to
21
22 greatly improve the training speed of the SVM and save a lot of storage space ^{54 55},
23
24 using the kernel is a more efficient technique in case of similarity of representation
25
26 between samples. Thus, the computational complexity of SVM is not wholly
27
28 governed by the number of samples, but by the number of features, which is
29
30 advantageous for analysis in the high-dimensional settings ⁵³. In addition, for SVM,
31
32 feature selection based on recursive feature elimination was performed with addition
33
34 and/or removal of predictors to determine the optimal combination that would
35
36 maximize the AUC ²⁵. Aided by feature selection, the proposed SVM method
37
38 identifies the most discriminating indexes for mortality prediction. We found that,
39
40 although both LR and SVM did not have a different AUC in the training procedure,
41
42 the SVM model for all samples with selected features had a significantly higher
43
44 sensitivity (86.96%) in predicting the mortality of motorcycle rides in the test set
45
46 compared to the rest of the models. The increased sensitivity of SVM in test set than
47
48 that in training set may be attributed to an improved quality of registered content and
49
50 less missing data in our registered data after continuous quality assessment and years
51
52 of experience of working with the registers. Such increased sensitivity was also found
53
54
55
56
57
58
59
60

1
2
3 in the LR model in the test set. With addition of more data in the model, the SVM
4
5 model has the potential to get an increased predictive power. This study demonstrates
6
7 the feasibility of using SVM classification with feature selection to predict the
8
9 mortality risk for motorcycle riders in the trauma care. However, the SVM model
10
11 generally works like a black-box and cannot identify the relationships between
12
13 mortality and the various explanatory variables and therefore, cannot be directly used
14
15 to validate our hypothesis of increased sensitivity in the test set.
16
17
18
19

20
21 There are several limitations to this study. Firstly, the patients who had
22
23 incomplete records were excluded from the analysis. This could have caused the
24
25 results to be biased and the results could have been different from the data acquired
26
27 by including the patients with incomplete records and replacing the missing data on a
28
29 variable by a value that is drawn from an estimate of the distribution of this variable
30
31 ⁵⁶⁻⁵⁸. The benefit of imputation is that we would be able to include patients who might
32
33 have relevant features for analysis, but were excluded owing to errors in data
34
35 collection or recording ⁵⁶⁻⁵⁸. Secondly, a source of potential bias may come from the
36
37 exclusion of patients declared dead (either on arriving at the hospital or at the accident
38
39 spot itself) and injured patients who were discharged against the advice of the
40
41 emergency department. Thirdly, there was lack of important data regarding injury
42
43 mechanism and circumstance, including motorcycle speed and type, helmet material,
44
45 and impact force during collision. In addition, the imputation of physiological and
46
47 laboratory data collected from the time of arriving at the emergency department
48
49 cannot reflect the dynamic changes in hemodynamic and metabolic variables of the
50
51 trauma patients under a possible resuscitation procedure. Finally, the study population
52
53 was limited to a single urban trauma center in southern Taiwan, which may not be
54
55
56
57
58
59
60

1
2
3 representative of other populations.
4
5

6 7 **CONCLUSION**

8
9 We demonstrate that ML is able to provide a feasible level of accuracy for
10 predicting mortality of the motorcycle riders. Whilst there are significant theoretical
11 and practical challenges to the translational implementation of this approach, the
12 results of the studies published so far are encouraging and may provide the first steps
13 towards the development of a prediction model integrated into the trauma care system
14 in order to identify an individual motorcycle rider's risk of mortality.
15
16
17
18
19
20
21
22
23

24 **COMPETING INTERESTS**

25 The authors declare that they have no competing interests.
26
27
28
29
30

31 **AUTHOR CONTRIBUTIONS**

32 PJK wrote the manuscript; CSR analyzed the tables; PCC performed the statistical
33 analyses and ML programming; YCC and HYH collected the data and are responsible
34 for the integrity of the registered data; and CHH designed the study and contributed to
35 the analysis and interpretation of data. All authors have read and approved the final
36 manuscript.
37
38
39
40
41
42
43
44
45

46 **DATA SHARING**

47 No additional data are available.
48
49
50
51

52 **ACKNOWLEDGEMENTS**

53 This research was supported by a grant from Chang Gung Memorial Hospital
54
55
56
57
58
59
60

1
2
3 CMRPG8F0891. We appreciate the Biostatistics Center, Kaohsiung Chang Gung
4 Memorial Hospital for helping us with the statistical work.
5
6
7

8 9 REFERENCES

- 10
11
12 1. Weiss H, Agimi Y, Steiner C. Youth motorcycle-related brain injury by state helmet
13 law type: United States, 2005-2007. *Pediatrics* 2010;**126**(6):1149-55.
14
15
- 16 2. National Highway Traffic Safety Administration (NHTSA) TSFDDH.
17
18
- 19 3. Beck LF, Dellinger AM, O'Neil ME. Motor vehicle crash injury rates by mode of
20 travel, United States: using exposure-based methods to quantify differences.
21
22
23
24
25
26
27
28
29 American journal of epidemiology 2007;**166**(2):212-8.
30
- 31 4. Chang HL, Lai CY. Using travel socialization and underlying motivations to better
32 understand motorcycle usage in Taiwan. *Accident; analysis and prevention*
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52 5. Jou RC, Yeh TH, Chen RS. Risk factors in motorcyclist fatalities in Taiwan. *Traffic*
53 injury prevention 2012;**13**(2):155-62.
54
- 55 6. Liang CC, Liu HT, Rau CS, et al. Motorcycle-related hospitalization of adolescents
56 in a Level I trauma center in southern Taiwan: a cross-sectional study. *BMC*
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1. 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.

- 1
2
3
4 8. Hsieh CH, Hsu SY, Hsieh HY, et al. Differences between the sexes in
5
6 motorcycle-related injuries and fatalities at a Taiwanese level I trauma center.
7
8
9 Biomed J 2017;**40**(2):113-20.
10
- 11
12 9. Hsieh CH, Liu HT, Hsu SY, et al. Motorcycle-related hospitalizations of the elderly.
13
14
15 Biomed J 2017;**40**(2):121-28.
16
- 17
18 10. Densmore JC, Lim HJ, Oldham KT, et al. Outcomes and delivery of care in
19
20 pediatric injury. Journal of pediatric surgery 2006;**41**(1):92-8; discussion 92-8.
21
- 22
23 11. Rogers SC, Campbell BT, Saleheen H, et al. Using trauma registry data to guide
24
25 injury prevention program activities. The Journal of trauma 2010;**69**(4
26
27 Suppl):S209-13.
28
- 29
30 12. Norrie J. Mortality prediction in ICU: a methodological advance. The Lancet
31
32 Respiratory medicine 2015;**3**(1):5-6.
33
- 34
35 13. de Munter L, Polinder S, Lansink KW, et al. Mortality prediction models in the
36
37 general trauma population: A systematic review. Injury 2017;**48**(2):221-29.
38
- 39
40 14. Demetriades D, Chan L, Velmanos GV, et al. TRISS methodology: an
41
42 inappropriate tool for comparing outcomes between trauma centers. J Am Coll
43
44 Surg 2001;**193**(3):250-4.
45
- 46
47 15. Jones JM, Skaga NO, Sovik S, et al. Norwegian survival prediction model in
48
49 trauma: modelling effects of anatomic injury, acute physiology, age, and
50
51
52
53
54
55
56

- 1
2
3
4 co-morbidity. *Acta Anaesthesiol Scand* 2014;**58**(3):303-15.
5
6
7 16. Bergeron E, Rossignol M, Osler T, et al. Improving the TRISS methodology by
8
9 restructuring age categories and adding comorbidities. *J Trauma*
10
11 2004;**56**(4):760-7.
12
13
14
15 17. Fueglistaler P, Amsler F, Schuepp M, et al. Prognostic value of Sequential Organ
16
17 Failure Assessment and Simplified Acute Physiology II Score compared with
18
19 trauma scores in the outcome of multiple-trauma patients. *Am J Surg*
20
21 2010;**200**(2):204-14.
22
23
24
25
26 18. Kroezen F, Bijlsma TS, Liem MS, et al. Base deficit-based predictive modeling of
27
28 outcome in trauma patients admitted to intensive care units in Dutch trauma
29
30 centers. *J Trauma* 2007;**63**(4):908-13.
31
32
33
34
35 19. Stoica B, Paun S, Tanase I, et al. Probability of Survival Scores in Different
36
37 Trauma Registries: A Systematic Review. *Chirurgia (Bucur)* 2016;**111**(2):115-9.
38
39
40
41 20. Cohen AM, Ambert K, McDonagh M. A Prospective Evaluation of an Automated
42
43 Classification System to Support Evidence-based Medicine and Systematic
44
45 Review. *AMIA Annual Symposium proceedings AMIA Symposium*
46
47 2010;**2010**:121-5.
48
49
50
51 21. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in
52
53 cardiovascular risk prediction: applying machine learning to address analytic
54
55
56
57
58
59
60

- 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, Douligieris 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.

- 1
2
3
4 50. Gu W, Vieira AR, Hoekstra RM, et al. Use of random forest to estimate population
5
6 attributable fractions from a case-control study of Salmonella enterica serotype
7
8 Enteritidis infections. *Epidemiol Infect* 2015;**143**(13):2786-94.
9
10
11
12 51. Lemon SC, Roy J, Clark MA, et al. Classification and regression tree analysis in
13
14 public health: methodological review and comparison with logistic regression. *nm*
15
16 *Behav Med* 2003;**26**(3):172-81.
17
18
19
20 52. Chen S, Zhou S, Yin FF, et al. Investigation of the support vector machine
21
22 algorithm to predict lung radiation-induced pneumonitis. *Med Phys*
23
24 2007;**34**(10):3808-14.
25
26
27
28
29 53. Orru G, Pettersson-Yeo W, Marquand AF, et al. Using Support Vector Machine to
30
31 identify imaging biomarkers of neurological and psychiatric disease: a critical
32
33 review. *Neurosci Biobehav Rev* 2012;**36**(4):1140-52.
34
35
36
37 54. Du Hongle LQ, and Cao Jing. Reduce the Samples for SVM Based on Euclidean
38
39 Distance. 3rd International Conference on System Science, Engineering Design
40
41 and Manufacturing Informatization 2013.
42
43
44
45
46 55. R. H. Laskar FAT, Biman Paul and Debmalya Chakrabarty. Sample reduction
47
48 using recursive and segmented data structure analysis. *Journal of Engineering and*
49
50 *Computer Innovations* Vol 2(4), pp 59-67, 2011.
51
52
53
54 56. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to
55
56

1
2
3
4 imputation of missing values. *J Clin Epidemiol* 2006;**59**(10):1087-91.

5
6
7 57. Shrive FM, Stuart H, Quan H, et al. Dealing with missing data in a multi-question
8
9 depression scale: a comparison of imputation methods. *BMC medical research*
10
11 methodology 2006;**6**:57.

12
13
14
15 58. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing
16
17 data. *J Clin Epidemiol* 2002;**55**(4):329-37.

21 22 23 24 **Figure Legend**

25
26 Figure 1. Demographics and injury characteristics of the patients regarding gender,
27
28 co-morbidities, injury region, number of injury regions, and helmet-wearing status.

29
30
31
32 Figure 2. Injury characteristics of the patients regarding laboratory data collected from
33
34 the time point when arrival at the emergency department.

35
36
37
38 Figure 3. ROC curves for LR, SVM, and DT models in predicting mortality of
39
40 motorcycle riders.

41
42
43
44
45 Figure 4. Illustration of DT model for mortality of motorcycle riders. The boxes
46
47 denote the percentage of patients with discriminating variables from CART analysis.
48
49 Those who were survival and fatal were indicated with green and red colors,
50
51 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 $p < 0.05$

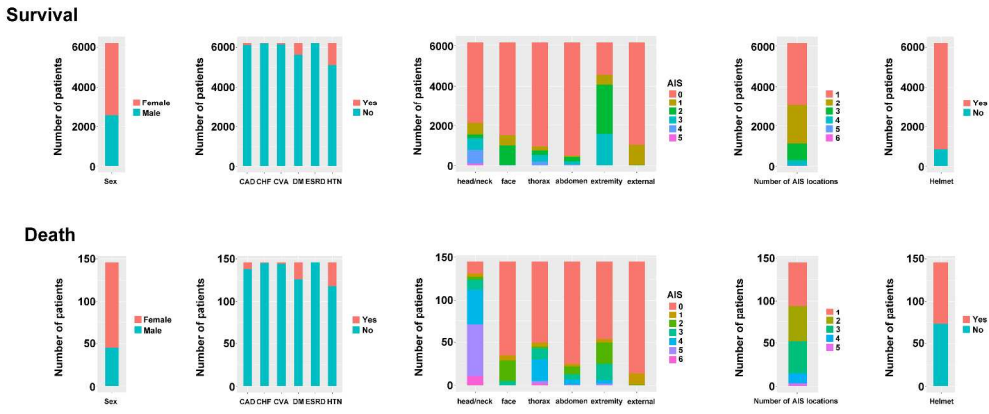


Figure 1. Demographics and injury characteristics of the patients regarding gender, co-morbidities, injury region, number of injury regions, and helmet-wearing status.

800x337mm (300 x 300 DPI)

er review only

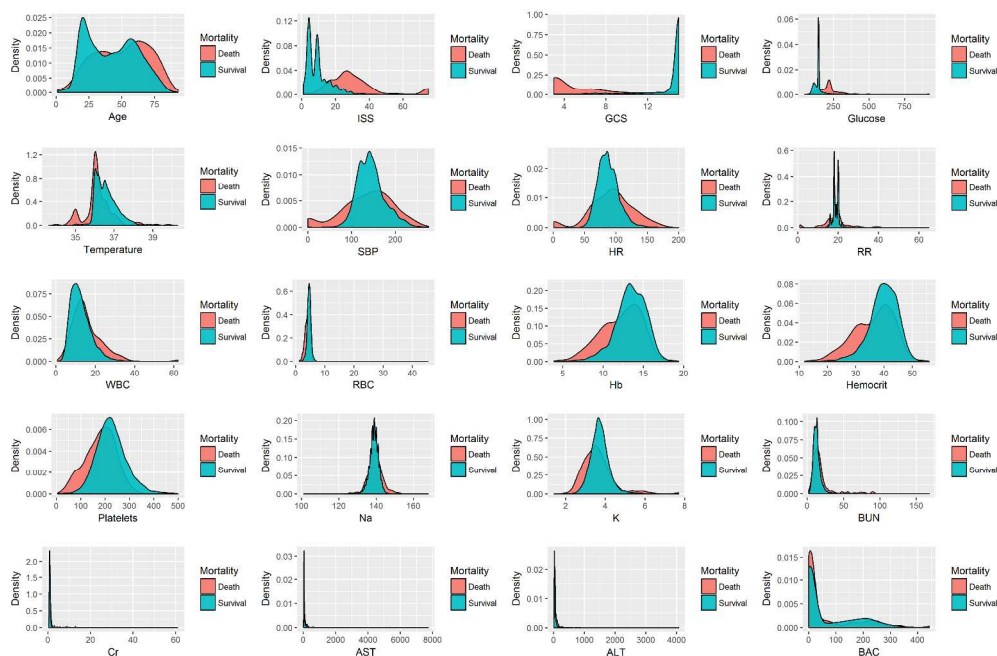


Figure 2. Injury characteristics of the patients regarding laboratory data collected from the time point when arrival at the emergency department

381x254mm (300 x 300 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

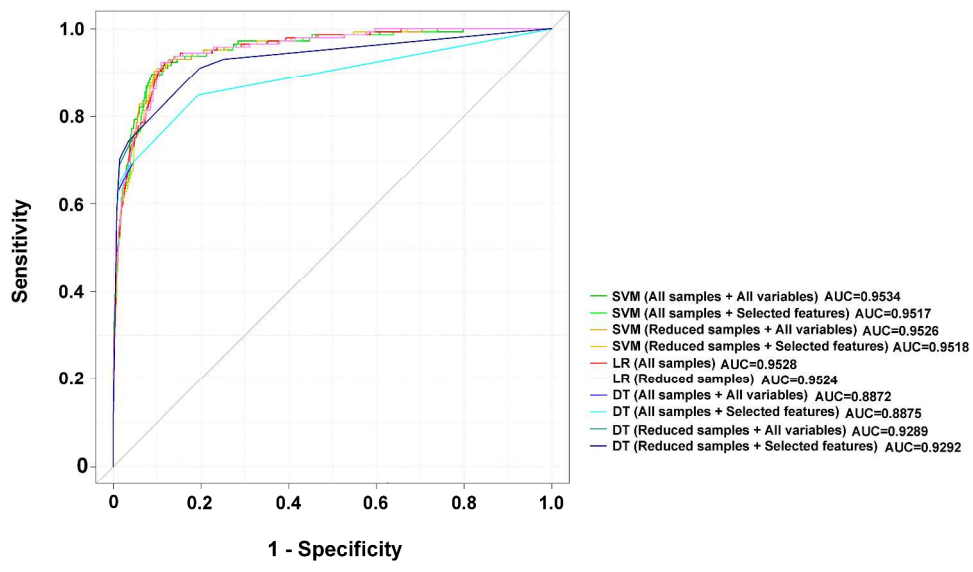


Figure 3. ROC curves for LR, SVM, and DT models in predicting mortality of motorcycle riders.

470x284mm (300 x 300 DPI)

Review only

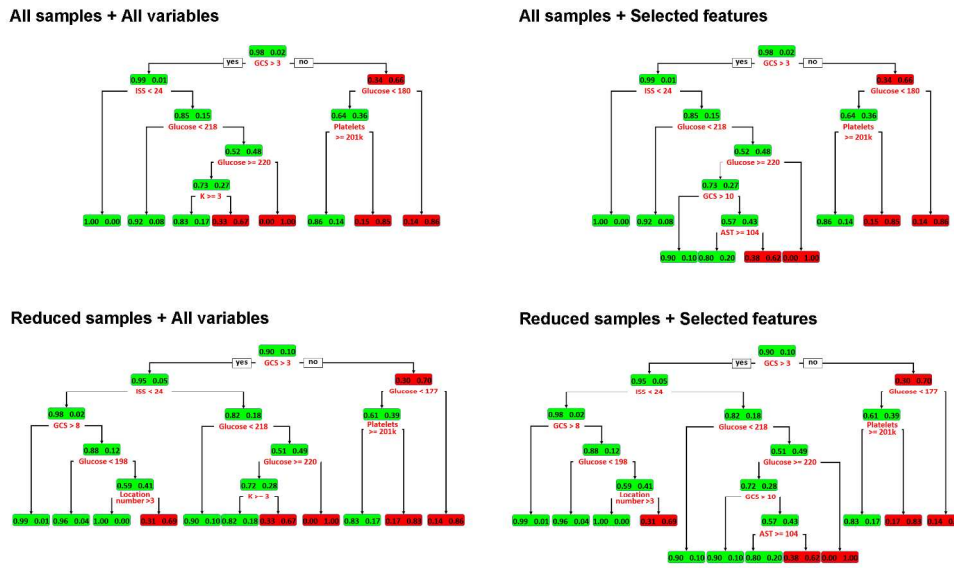


Figure 4. Illustration of DT model for mortality of motorcycle riders. The boxes denote the percentage of patients with discriminating variables from CART analysis. Those who were survival and fatal were indicated with green and red colors, respectively, in the boxes.

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:	<i>BMJ Open</i>
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

1
2
3 **Derivation and validation of different machine learning models in mortality**
4 **prediction of trauma motorcycle riders - a cross-sectional retrospective study in**
5 **southern Taiwan**
6
7

8
9 Pao-Jen Kuo^{1†}, M.D.; Shao-Chun Wu^{2†} M.D.; Peng-Chen Chien¹, M.Sc.;
10 Cheng-Shyuan Rau³, M.D.; Yi-Chun Chen¹, M.Sc.; Hsiao-Yun Hsieh¹, M.Sc.;
11 Ching-Hua Hsieh^{1*}, M.D., Ph.D.
12
13
14
15

16
17
18 † Indicates equal contribution in authorship as the first author.
19

20 ¹ Department of Plastic and Reconstructive Surgery, Kaohsiung Chang Gung
21 Memorial Hospital and Chang Gung University College of Medicine, Taiwan, 833
22

23 ² Department of Anesthesiology, Kaohsiung Chang Gung Memorial Hospital and
24 Chang Gung University College of Medicine, Taiwan, 833
25

26 ³ Department of Neurosurgery, Kaohsiung Chang Gung Memorial Hospital and Chang
27 Gung University College of Medicine, Taiwan, 833
28
29
30

31 Pao-Jen Kuo; email: bow110470@gmail.com
32

33 Shao-Chun Wu; email: shaochunwu@gmail.com
34

35 Peng-Chen Chien; e-mail: VENU_CHIEN@hotmail.com
36

37 Cheng-Shyuan Rau; e-mail: ersh2127@cloud.cgmh.org.tw
38

39 Yi-Chun Chen; e-mail: libe320@yahoo.com.tw
40

41 Hsiao-Yun Hsieh; e-mail: sylvia19870714@hotmail.com
42

43 Ching-Hua Hsieh; e-mail: m93chinghua@gmail.com
44
45

46 Corresponding author: Ching-Hua Hsieh, M.D., PhD.
47

48 Department of Trauma Surgery & Plastic and Reconstructive Surgery, Kaohsiung
49
50
51

1
2
3 Chang Gung Memorial Hospital and Chang Gung University College of Medicine,
4
5 Taiwan

6
7 No.123, Ta-Pei Road, Niao-Song District, Kaohsiung City 833, Taiwan

8
9 Tel: 886-7-7317123 ext: 8002; E-mail: m93chinghua@gmail.com

10
11
12
13
14
15
16 **ABSTRACT**

17
18 **Objectives:** We aimed to build and test models of machine learning (ML) to predict
19
20 the mortality of hospitalized motorcycle riders.

21
22 **Setting:** A Level I trauma center in southern Taiwan.

23
24 **Participants:** The hospitalized motorcycle riders between January 2009 and
25
26 December 2015 were allocated to be a training set (n=6,306) and a test set (n= 946).

27
28 Using the demographics and injury characteristics as well as laboratory data of
29
30 patients, logistic regression (LR), support vector machine (SVM), and decision tree
31
32 (DT) were performed to determine the mortality of the individual motorcycle riders,
33
34 under the different conditions of using all samples or reduced samples as well as using
35
36 all variables or selected features into the algorithm.

37
38 **Primary and secondary outcome measures:** Model predictive performance was
39
40 evaluated by accuracy, sensitivity, specificity, geometric mean, and by the analysis of
41
42 the area under curve of the receiver operator characteristic curves of the two different
43
44 models.
45
46

47
48 **Results:** In the training set, both LR and SVM had a significantly higher AUC than
49
50 that of DT, while there was no significant difference in the AUC of LR and SVM,
51
52 regardless of using all samples or reduced samples as well as all variables or selected
53
54 features. In the test set, SVM model for all samples with selected features presented a
55
56

1
2
3 better model than all the other models, with an accuracy of 98.73%, sensitivity
4 (86.96%), specificity (99.02%), geometric mean (92.79%) and AUC of 0.9517 for
5 mortality prediction.
6
7

8
9 **Conclusion:** We demonstrate that ML is able to provide a feasible level of accuracy
10 for predicting the mortality of the motorcycle riders. The integration of ML model,
11 particularly the SVM algorithm in trauma system may help identify high-risk patients
12 and therefore drive the appropriate response by the clinical staff.
13
14
15
16

17
18
19
20 **KEY WORDS:** Motorcycle accident; mortality; machine learning (ML); logistic
21 regression (LR), support vector machine (SVM), and decision tree (DT)
22
23
24
25

26 **ARTICLE SUMMARY**

27 **STRENGTHS AND LIMITATIONS OF THIS STUDY**

- 28
29
30
31 ■ This study demonstrates the feasibility of using support vector machine (SVM)
32 classification, one of machine learning models, to predict the mortality risk for
33 motorcycle riders.
34
35
36
37 ■ With addition of more data in the model, the SVM model has the potential to get
38 an increased predictive power and facilitate its clinical implement.
39
40
41
42 ■ The SVM model generally works like a black-box and cannot identify the
43 relationships between mortality and the various explanatory variables.
44
45
46
47 ■ The incomplete records of patients and the exclusion of patients declared dead
48 from the Trauma Registry System could bias the results.
49
50
51
52
53
54
55
56
57
58
59
60

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 mile¹, 30 times more likely to die in a motor vehicle crash², and 58 times more likely to be killed on a per-trip basis³. In Taiwan, motorcyclist fatalities account for nearly 60% of all driving fatalities⁴. The fatalities are often associated with men, advanced age, not wearing a helmet, unlicensed status, and riding under the influence of alcohol⁵⁻⁹. In addition, head injuries were the major factor leading to mortality, followed by thoracic and abdominal injuries⁶⁻⁹.

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 management¹². 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 patients¹³. Even after the incorporation

1
2
3 of other or revised predictors, like blood pressure ¹⁴, co-morbidities, and separate
4 categories for different age-groups ¹⁵ into this model, the addition of more predictors
5 to the basic TRISS model did not always result in higher performance ¹⁶⁻¹⁸. Although
6 the revised TRISS, resulting from the USA National Trauma Database is
7 inaccurate for trauma systems, particularly in the management of predominantly blunt
8 injuries ¹⁹, the further development of the model based on advanced methodological
9 quality, the performance of the model in subsets of patient groups, and practical
10 application is mandatory in the prediction of mortality ¹⁶.

21
22
23 Currently, machine learning (ML) had been successfully applied in the real world
24 in many fields including automatic medical diagnostics and personalized health care
25 ²⁰⁻²². There is an increasing interest in the application of supervised ML methods to
26 aid diagnosis and prognosis in trauma patients. ML is based on the way the human
27 brain approaches pattern recognition tasks, providing an artificial intelligence-based
28 approach to solve classification problems and improving their efficiency and
29 effectiveness over time ²³. The usefulness of ML is bolstered by the versatility of its
30 techniques and its utility for artificial intelligence such as prediction, classification,
31 planning, recognition, and clustering ^{23 24}. Comparisons of different learning strategies
32 have been conducted previously by others using field-specific datasets, many of which
33 have shown significantly better predictive power than the more conventional
34 alternatives ²⁵. Examples of multivariate techniques for pattern recognition include,
35 but are not limited to, LR, support vector machine (SVM), decision trees (DT), and
36 artificial neural networks. LR is a widely used and accepted statistical analysis tool to
37 predict the probability of the occurrence of an event ²⁶. It attempts to build a
38 functional relationship between two or more independent predictors and the one

1
2
3 dependent outcome variable, under the assumption that the response variables are
4 linearly related to the coefficients of the predictor variables ²⁶.
5
6
7

8
9 SVM uses a training set of data composed of one or more features to determine
10 an optimal boundary separating a set of cases. The binary SVM classifier constructs a
11 set of the optimal hyperplanes in high-dimensional space with the maximal margin of
12 the two classes ²⁷. In the case that all training points cannot be separated by the
13 hyperplane, a soft margin method is used to construct a hyperplane that separates the
14 training data points ^{28,29}. It has been found that the SVM model has a great capability
15 of dealing with classification problems ³⁰⁻³⁴.
16
17
18
19
20
21
22
23
24
25

26 A DT is a hierarchical model composed of decision rules based upon optimal
27 feature cutoff values that recursively split independent variables into different groups
28 ³⁵⁻³⁷. The purpose of DT building is to search for a set of decision rules to predict an
29 outcome from a set of input variables ^{33, 35, 36}. Some models are used to construct
30 decision-tree models, including classification and regression trees (CART), ID3s,
31 chi-square automatic interaction detector DTs (CHAIDs), and C4.5 and C5.0 DTs [26,
32 28]. Among these methods, the CART analysis is a combined approach based on
33 nonparametric and nonlinear variables for recursive partitioning analysis. CART
34 analysis is an innovative DT model in which several predictive variables are crucial to
35 identify patients at different levels of risk in various medical fields through
36 progressive binary splits to develop prediction models in order to enable better
37 prediction and clinical decision-making ³⁸⁻⁴⁰.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54

55 This study aimed to construct a model for the mortality prediction of motorcycle
56
57
58
59
60

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

1
2
3 concentration (BAC), and glucose at emergency department.
4
5
6

7
8 These enrolled patients were divided into a training sample, which was used for
9 predictor discovery and supervised classification to generate a plausible model, and a
10 test sample, which was used to test the performance of the model generated in the
11 training sample. Those patients with missing data were not included for further
12 analysis. The patients who registered in a six-year span between January 2009 and
13 December 2014 were allocated in the training set, which comprised of a total of 6,306
14 patients. It included 6,161 survival and 145 mortality patients. In the test set, there
15 were 946 patients, including 923 survival and 23 mortality patients, from the one-year
16 span between January 2015 and December 2015. The sample similarity was assessed
17 by Euclidean distance for quantitative data to reduce the size of a sample designed for
18 use in data analysis⁴¹. The sample reduction used Euclidean distance of the `dist`
19 function in the `stats` package in R (R Foundation for Statistical Computing, Vienna,
20 Austria). During sample reduction, the data size can be reduced to speed up
21 calculations in the analysis⁴². However, considering the exploratory character of this
22 study, all samples (n=6,306) and reduced samples (n=1,510) in the training set of this
23 study would require to be analyzed in ML classification.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43

44 **ML classifiers**

45
46 This work provides a performance comparison of three different ML classifiers
47 (LR, SVM, and DT).
48
49
50

51 *Logistic regression*

52
53
54 The LR classifier used the `glm` function in the `stats` package in R3.3.3 (R
55
56
57
58
59
60

1
2
3 Foundation for Statistical Computing, Vienna, Austria). Univariate LR analyses were
4 initially performed to identify the significant predictor variables of the mortality risk.
5
6 Stepwise LR analysis was used to control the effects of confounding variables to
7 identify independent risk factors for mortality. The selected independent risk factors
8
9 obtained from LR were also used as selected features to be implemented by the SVM
10
11 and the DT to explain their weights in determining the risk of mortality.
12
13
14
15
16
17

18 *Support vector machine*

19
20 The SVM classifier used the `tune.svm` & `svm` function in the `e1071` package
21 in R. In the training set, the SVM classifier was performed for the prediction of
22 mortality with regard to either all 32 variables or 12 selected features as well as all the
23 samples and reduced samples in the training set. The mapping procedure was
24 accomplished by the kernel function, which is a matrix of pair-wise similarities
25 between data points, such as a linear, polynomial, or radial basis function (RBF)⁴³.
26 For this study, the RBF kernel was chosen because it can handle non-linear
27 interactions between class labels and features⁴⁴. The two main parameters presented
28 in SVM with RBF kernel were the penalty parameter C and the kernel
29 hyper-parameter γ . The penalty parameter C determined the tradeoff between the
30 fitting error minimization and model complexity, while the hyper-parameter γ defined
31 the nonlinear feature transformation onto a higher dimensional space and controlled
32 the tradeoff between error due to bias and variance in the model.⁴⁵ The optimal
33 operating point was estimated by varying the parameters - C and γ using a grid search
34 for each combination of feature selection and dimension reduction with a 10-fold
35 cross-validation⁴⁴.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 impurity³⁸⁻⁴⁰. 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 impossible³⁸⁻⁴⁰.

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 individuals⁴⁹.

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 \pm 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

1
2
3 lower AIS score at the extremities compared to the patients who survived (Table 1
4 and supplemental Figure 1). The patients with fatality had sustained more number of
5 injured body regions (number of AIS locations) than the ones who survived. In
6 addition, the patients with fatality comprised more of females and fewer of them were
7 observed to be wearing a helmet compared to the patients who survived (Table 1 and
8 supplemental Figure 1). A statistically significant difference in age, ISS, GCS,
9 glucose, temperature, Hb, Hct, platelets, K, Cr, AST, ALT, and incidences of CAD
10 was found between patients with fatality and the ones who survived respectively
11 (Table 2 and supplemental Figure 2). Because the distribution pattern between Hb and
12 Hct as well as between AST and ALT is very similar, only one of these two variables
13 (i.e. Hct and AST) was selected for further ML classification to prevent the inclusion
14 of duplicate parameters. Therefore, a total of 32 variables were used for imputation
15 into ML classifiers as all variables, in contrast to considering selected features
16 obtained by using the independent risk factors identified by the LR given below.
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

35 **Performance of ML classifiers in training set**

36 *Logistic regression*

37
38 LR identified 12 predictors (platelets, glucose, BUN, Cr, AST, Na, Age, GCS,
39 temperature, number of AIS locations, ISS, and HTN) as independent risk factors for
40 mortality in motorcycle riders from either all samples or the reduced samples.
41
42
43
44
45
46
47

48 The predictive models were listed as:

49 All samples (n=6,306)

$$50 Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 4.71648 - 0.00846 * \text{Platelets} + 0.01189 * \text{Glucose} +$$

$$51 0.03459 * \text{BUN} + 0.10667 * \text{Cr} + 0.00195 * \text{AST} + 0.09513 * \text{Na} +$$

$$0.02533 * \text{Age} - 0.39968 * \text{GCS} - 0.56396 * \text{Temperature} - 0.93232 * \\ \text{Number of AIS locations} + 0.14098 * \text{ISS} - 0.95726 * \text{HTN}$$

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 $\gamma = 0.00390625$. 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

1
2
3 of 98.10%) for all samples and reduced samples, respectively (Table 3). The AUCs for
4
5 all samples and reduced sample were 0.9534 and 0.9526, respectively (Figure 1). With
6
7 selected features in the model, the SVM achieved an accuracy of 98.62% (sensitivity
8
9 of 64.14% and specificity of 99.43%) and 93.84% (sensitivity of 62.76% and
10
11 specificity of 97.14%) (Table 3) as well as 0.9517 and 0.9518 AUCs for all samples
12
13 and reduced samples, respectively (Figure 1).
14
15

16 17 18 *Decision tree*

19
20 As shown in Figure 2, in the DT model, GCS was identified as the variable of
21
22 initial split with an optimal cut-off value of > 3 . Among patients with GCS higher
23
24 than 3, glucose was selected as the variable of second split at a discrimination level of
25
26 180 and 177 mg/dL for all samples and reduced samples, respectively. After the
27
28 glucose level < 180 or 177 mg/dL for all samples and reduced samples, respectively,
29
30 the next best predictor of mortality was platelets with an optimal cut-off of 201×10^3
31
32 / μ L. For the node, with patients having a GCS not greater than 3, ISS < 24 and glucose
33
34 < 218 mg/dL, these predictors were selected as significant variables for all samples
35
36 and reduced samples, with GCS > 8 , glucose < 198 mg/dL, and the number of AIS
37
38 locations ≥ 3 being an additional predictors for splitting for the reduced samples. With
39
40 all variables in the model, the DT achieved an accuracy of 98.92% (sensitivity of
41
42 62.76% and specificity of 99.77%) and 95.83% (sensitivity of 68.97% and specificity
43
44 of 98.68%) for the all samples and reduced samples, respectively. The AUCs for all
45
46 samples and reduced samples were 0.8872 and 0.9289, respectively. With selected
47
48 features in the model, the DT achieved an accuracy of 98.92% (sensitivity of 64.14%
49
50 and specificity of 99.74%) and 95.83% (sensitivity of 70.34% and specificity of
51
52 98.53%) for the all samples and reduced samples, respectively. The AUCs for all
53
54
55
56
57
58
59
60

1
2
3 samples and reduced samples were 0.8872 and 0.9289, respectively (Figure 1). In the
4
5 condition of using reduced samples but not all samples in the DT model, the number
6
7 of AIS locations would be added in the split of the node slightly increasing the
8
9 sensitivity from 62.76% to 68.97% and from 64.14% to 70.34% with input
10
11 comprising of all variables and selected variables, respectively. In addition, in the
12
13 condition of using selected features but not all variables in the DT model, the level of
14
15 K was not used in the splitting of the node and was substituted by the cut-off value of
16
17 $AST \geq 104$ IU/L, slightly increasing the sensitivity from 62.76% to 64.14% and from
18
19 68.97% to 70.34% with input as all samples and reduced samples, respectively. In
20
21 addition, the AUCs for all samples and reduced sample were 0.8875 and 0.9292,
22
23 respectively (Figure 1)
24
25

26 27 28 29 *Comparison in AUC analysis*

30
31 In the comparisons of AUCs for LR, SVM, and DT for the training set (Table 4
32
33 and Figure 1), both LR and SVM had a significantly higher AUC than DT, regardless
34
35 of using all samples or reduced samples as well as for all variables or for selected
36
37 features. However, there was no significant difference of AUC between LR and SVM,
38
39 regardless of using all samples or reduced samples as well as for all variables or for
40
41 selected features. In addition to this, in DT sample reduction had a significantly
42
43 higher AUC than the one obtained using all samples, but there was no significant
44
45 difference of AUC between DT with all variables or with selected features.
46
47
48
49

50 51 **Performance of ML classifiers in test set**

52
53 In test set, the LR model for all samples and reduced samples - both achieved an
54
55 accuracy of 98.41%, with a sensitivity of 73.91% and specificity of 99.02% in
56
57

1
2
3 predicting the mortality (Table 3). All of these four SVM models create an accuracy
4 more than 98% and a specificity near 99% in predicting the mortality. Whereas the
5 SVM model for all samples with selected features had a significantly highest
6 sensitivity (86.96%) and geometric mean (92.79%). All of these four DT models
7 create an accuracy of approximately 98% and a specificity of approximately 99% but
8 a sensitivity of less than 70%. Considering that the majority of patients survived
9 except for a few with fatality, would result into a very high accuracy and specificity
10 index in predicting the mortality, therefore the comparison should further focus on the
11 sensitivity and geometric mean of different ML models. We found all LR and SVM
12 models, but not the DT models, had an increased sensitivity in the test set. In addition,
13 the SVM model for all samples with selected features had a significantly highest
14 sensitivity and geometric mean.
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

31 **DISCUSSION**

32 LR is widely used in epidemiological studies for causal inference and, with the
33 selection of built-in features; it does not necessarily utilize all the predictors. With a
34 relatively limited number of variables i.e. variables less than 20, LR provides
35 estimates of odd ratios of the risk factors⁵⁰. However, its limitations become apparent
36 when analyzing a complex dataset with a high number of relevant exposures and
37 multiple interactions⁵¹. With too many predictors, the availability of sufficient
38 information to specify all interactions would become nearly impossible⁵¹. In addition,
39 the DT with CART analysis is exploratory and not based on the probabilistic method,
40 which may lead to overestimating the importance of included risk factors or cause
41 missing of other potential confounders that could influence each patient's actual risk
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 analysis method, the SVM boundary is only minimally influenced by outliers that are
4 difficult to separate, despite the complexity of data⁵³. In addition, employing kernels
5 in the SVM would be advantageous to learn non-linear decision boundaries allowing
6 the classifier to solve more difficult classification problems than the linear analysis
7 method⁵⁴. These three ML models (LR, SVM, DT) all create an accuracy and a
8 specificity around 98% and 99%, respectively, but a sensitivity less than or around
9 70% in the training dataset. In this study, both LR and SVM resulted in a
10 significantly higher AUC than DT in the training set, regardless of using all samples
11 or reduced samples as well as for all variables or for selected features.
12
13
14
15
16
17
18
19
20
21
22
23

24 This study included different variants of SVM considering the sample size and
25 feature selection to show all possible improvements to the more conventional
26 strategies like LR or DT. Although sample reduction for SVM had been proposed to
27 greatly improve the training speed of the SVM and save a lot of storage space^{55 56},
28 using the kernel is a more efficient technique in case of similarity of representation
29 between samples. Thus, the computational complexity of SVM is not wholly
30 governed by the number of samples, but by the number of features, which is
31 advantageous for analysis in the high-dimensional settings⁵⁴. In addition, feature
32 selection in SVM may maximize the AUC²⁵. Aided by feature selection, the proposed
33 SVM method identifies the most discriminating indexes for mortality prediction. We
34 found that, although both LR and SVM did not have a different AUC in the training
35 procedure, the SVM model for all samples with selected features had a significantly
36 higher sensitivity (86.96%) in predicting the mortality of motorcycle rides in the test
37 set compared to the rest of the models. The increased sensitivity of SVM in test set
38 than that in training set may be attributed to an improved quality of registered content
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 and less missing data in our registered data after continuous quality assessment and
4 years of experience of working with the registers. Such increased sensitivity was also
5 found in the LR model in the test set. With addition of more data in the model, the
6 SVM model has the potential to get an increased predictive power. This study
7 demonstrates the feasibility of using SVM classification with feature selection to
8 predict the mortality risk for motorcycle riders in the trauma care. However, the SVM
9 model generally works like a black-box and cannot identify the relationships between
10 mortality and the various explanatory variables and therefore, cannot be directly used
11 to validate our hypothesis of increased sensitivity in the test set.
12
13
14
15
16
17
18
19
20
21
22
23

24 There are several limitations to this study. Firstly, the patients who had
25 incomplete records were excluded from the analysis. This could have caused the
26 results to be biased and the results could have been different from the data acquired
27 by including the patients with incomplete records and replacing the missing data on a
28 variable by a value that is drawn from an estimate of the distribution of this variable
29 ⁵⁷⁻⁵⁹. The benefit of imputation is that we would be able to include patients who might
30 have relevant features for analysis, but were excluded owing to errors in data
31 collection or recording ⁵⁷⁻⁵⁹. Secondly, a source of potential bias may come from the
32 exclusion of patients declared dead (either on arriving at the hospital or at the accident
33 spot itself) and injured patients who were discharged against the advice of the
34 emergency department. Thirdly, there was lack of important data regarding injury
35 mechanism and circumstance, including motorcycle speed and type, helmet material,
36 and impact force during collision. In addition, the imputation of physiological and
37 laboratory data collected from the time of arriving at the emergency department
38 cannot reflect the dynamic changes in hemodynamic and metabolic variables of the
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 trauma patients under a possible resuscitation procedure. Further, some other
4 DT-related methods like DT by C4.5 ⁶⁰, combined classifiers of LR and DT by C4.5
5 ⁴⁸, and random forest ⁶¹ had been reported to provide a very good performance in
6
7 dealing with the classification problem; however, these techniques were not explored
8
9 in this study. Finally, the study population was limited to a single urban trauma center
10
11 in southern Taiwan, which may not be representative of other populations.
12
13
14
15

16 17 18 **CONCLUSION**

19
20 We demonstrate that ML is able to provide a feasible level of accuracy for
21 predicting mortality of the motorcycle riders. Whilst there are significant theoretical
22 and practical challenges to the translational implementation of this approach, the
23 results of the studies published so far are encouraging and may provide the first steps
24 towards the development of a prediction model integrated into the trauma care system
25 in order to identify an individual motorcycle rider's risk of mortality.
26
27
28
29
30
31
32
33
34

35 **COMPETING INTERESTS**

36 The authors declare that they have no competing interests.
37
38
39
40

41 **AUTHOR CONTRIBUTIONS**

42 PJK wrote the manuscript; SCW revised the manuscript; PCC performed the
43 statistical analyses and machine learning programming; CSR analyzed the tables;
44 YCC and HYH collected the data and are responsible for the integrity of the
45 registered data; and CHH designed the study and contributed to the analysis and
46 interpretation of data. All authors have read and approved the final manuscript.
47
48
49
50
51
52
53
54
55
56
57

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.

- 1
2
3 8. Hsieh CH, Hsu SY, Hsieh HY, et al. Differences between the sexes in
4 motorcycle-related injuries and fatalities at a Taiwanese level I trauma center.
5 Biomed J 2017;**40**(2):113-20.
6
7
- 8
9 9. Hsieh CH, Liu HT, Hsu SY, et al. Motorcycle-related hospitalizations of the elderly.
10 Biomed J 2017;**40**(2):121-28.
11
12
- 13 10. Densmore JC, Lim HJ, Oldham KT, et al. Outcomes and delivery of care in
14 pediatric injury. Journal of pediatric surgery 2006;**41**(1):92-8; discussion 92-8.
15
16
- 17 11. Rogers SC, Campbell BT, Saleheen H, et al. Using trauma registry data to guide
18 injury prevention program activities. The Journal of trauma 2010;**69**(4
19 Suppl):S209-13.
20
21
- 22 12. Norrie J. Mortality prediction in ICU: a methodological advance. The Lancet
23 Respiratory medicine 2015;**3**(1):5-6.
24
25
- 26 13. Demetriades D, Chan L, Velmanos GV, et al. TRISS methodology: an
27 inappropriate tool for comparing outcomes between trauma centers. J Am Coll
28 Surg 2001;**193**(3):250-4.
29
30
- 31 14. Jones JM, Skaga NO, Sovik S, et al. Norwegian survival prediction model in
32 trauma: modelling effects of anatomic injury, acute physiology, age, and
33 co-morbidity. Acta Anaesthesiol Scand 2014;**58**(3):303-15.
34
35
- 36 15. Bergeron E, Rossignol M, Osler T, et al. Improving the TRISS methodology by
37 restructuring age categories and adding comorbidities. J Trauma 2004;**56**(4):760-7.
38
39
- 40 16. de Munter L, Polinder S, Lansink KW, et al. Mortality prediction models in the
41 general trauma population: A systematic review. Injury 2017;**48**(2):221-29.
42
43
- 44 17. Fueglistaler P, Amsler F, Schuepp M, et al. Prognostic value of Sequential Organ
45 Failure Assessment and Simplified Acute Physiology II Score compared with
46 trauma scores in the outcome of multiple-trauma patients. Am J Surg
47
48
49
50
51
52
53
54
55
56

- 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

- 1
2
3 machine-learning methods. *Med Phys* 2016;**43**(5):2040.
- 4
5 26. Siuly, Yin X, Hadjiloucas S, et al. Classification of THz pulse signals using
6 two-dimensional cross-correlation feature extraction and non-linear classifiers.
7
8 *Comput Methods Programs Biomed* 2016;**127**:64-82.
- 9
10
11 27. V. V. Statistical learning theory. New York, NY: John Wiley & Sons 1998.
- 12
13 28. de Boves Harrington P. Support Vector Machine Classification Trees. *Anal Chem*
14
15 2015;**87**(21):11065-71.
- 16
17 29. Lee Y. Support vector machines for classification: a statistical portrait. *Methods*
18
19 *Mol Biol* 2010;**620**:347-68.
- 20
21 30. Chen C, Zhang G, Qian Z, et al. Investigating driver injury severity patterns in
22
23 rollover crashes using support vector machine models. *Accid Anal Prev*
24
25 2016;**90**:128-39.
- 26
27 31. Galatzer-Levy IR, Karstoft KI, Statnikov A, et al. Quantitative forecasting of
28
29 PTSD from early trauma responses: a Machine Learning application. *J Psychiatr*
30
31 *Res* 2014;**59**:68-76.
- 32
33 32. Li Z, Liu P, Wang W, et al. Using support vector machine models for crash injury
34
35 severity analysis. *Accid Anal Prev* 2012;**45**:478-86.
- 36
37 33. Marucci-Wellman HR, Corns HL, Lehto MR. Classifying injury narratives of
38
39 large administrative databases for surveillance-A practical approach combining
40
41 machine learning ensembles and human review. *Accid Anal Prev* 2017;**98**:359-71.
- 42
43 34. Patil BM, Joshi RC, Toshniwal D, et al. A new approach: role of data mining in
44
45 prediction of survival of burn patients. *J Med Syst* 2011;**35**(6):1531-42.
- 46
47 35. Farion K, Michalowski W, Wilk S, et al. A tree-based decision model to support
48
49 prediction of the severity of asthma exacerbations in children. *J Med Syst*
50
51 2010;**34**(4):551-62.
- 52
53
54
55
56
57
58
59
60

- 1
2
3 36. Zintzaras E, Bai M, Douligieris C, et al. A tree-based decision rule for identifying
4 profile groups of cases without predefined classes: application in diffuse large
5 B-cell lymphomas. *Comput Biol Med* 2007;**37**(5):637-41.
6
7
8
9 37. Kasbekar PU, Goel P, Jadhav SP. A Decision Tree Analysis of Diabetic Foot
10 Amputation Risk in Indian Patients. *Frontiers in endocrinology* 2017;**8**:25.
11
12 38. Guilbault RWR, Ohlsson MA, Afonso AM, et al. External Validation of Two
13 Classification and Regression Tree Models to Predict the Outcome of Inpatient
14 Cardiopulmonary Resuscitation. *J Viral Hepat* 2017;**32**(5):333-38.
15
16
17 39. Shi KQ, Zhou YY, Yan HD, et al. Classification and regression tree analysis of
18 acute-on-chronic hepatitis B liver failure: Seeing the forest for the trees.
19 2017;**24**(2):132-40.
20
21
22 40. Zimmerman RK, Balasubramani GK, Nowalk MP, et al. Classification and
23 Regression Tree (CART) analysis to predict influenza in primary care patients.
24 *BMC Infect Dis* 2016;**16**(1):503.
25
26
27 41. Amaratunga D, Cabrera J, Lee YS. Resampling-based similarity measures for
28 high-dimensional data. *J Comput Biol* 2015;**22**(1):54-62.
29
30
31 42. Bhattacharya S, Mariani TJ. Transformation of expression intensities across
32 generations of Affymetrix microarrays using sequence matching and regression
33 modeling. *Nucleic Acids Res* 2005;**33**(18):e157.
34
35
36 43. Vapnik VN. *The Nature of Statistical Learning Theory*. New York, 2nd ed. 2000.
37
38
39 44. Gultepe E, Green JP, Nguyen H, et al. From vital signs to clinical outcomes for
40 patients with sepsis: a machine learning basis for a clinical decision support
41 system. *J Am Med Inform Assoc* 2014;**21**(2):315-25.
42
43
44 45. Chen H, Hu L, Li H, et al. An Effective Machine Learning Approach for
45 Prognosis of Paraquat Poisoning Patients Using Blood Routine Indexes. *Basic Clin*
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Pharmacol Toxicol 2017;**120**(1):86-96.
- 4
5 46. Chang LY, Wang HW. Analysis of traffic injury severity: an application of
6 non-parametric classification tree techniques. *Accid Anal Prev*
7 2006;**38**(5):1019-27.
- 8
9
10
11 47. Ripley B. tree: Classification and regression trees. R package version 1.0-34. URL
12 : <http://CRAN.R-project.org/package=tree>. 2013.
- 13
14
15
16 48. Sanz J, Paternain D, Galar M, et al. A new survival status prediction system for
17 severe trauma patients based on a multiple classifier system. *Comput Methods*
18 *Programs Biomed* 2017;**142**:1-8.
- 19
20
21
22 49. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or
23 more correlated receiver operating characteristic curves: a nonparametric approach.
24 *Biometrics* 1988;**44**(3):837-45.
- 25
26
27
28 50. Knol MJ, Vandenbroucke JP, Scott P, et al. What do case-control studies estimate?
29 Survey of methods and assumptions in published case-control research. *Am J*
30 *Epidemiol* 2008;**168**(9):1073-81.
- 31
32
33
34 51. Gu W, Vieira AR, Hoekstra RM, et al. Use of random forest to estimate population
35 attributable fractions from a case-control study of Salmonella enterica serotype
36 Enteritidis infections. *Epidemiol Infect* 2015;**143**(13):2786-94.
- 37
38
39
40 52. Lemon SC, Roy J, Clark MA, et al. Classification and regression tree analysis in
41 public health: methodological review and comparison with logistic regression. *nn*
42 *Behav Med* 2003;**26**(3):172-81.
- 43
44
45
46 53. Chen S, Zhou S, Yin FF, et al. Investigation of the support vector machine
47 algorithm to predict lung radiation-induced pneumonitis. *Med Phys*
48 2007;**34**(10):3808-14.
- 49
50
51
52 54. Orru G, Pettersson-Yeo W, Marquand AF, et al. Using Support Vector Machine to
53
54
55
56

- 1
2
3 identify imaging biomarkers of neurological and psychiatric disease: a critical
4 review. *Neurosci Biobehav Rev* 2012;**36**(4):1140-52.
5
6
7 55. Du Hongle LQ, and Cao Jing. Reduce the Samples for SVM Based on Euclidean
8 Distance. 3rd International Conference on System Science, Engineering Design
9 and Manufacturing Informatization 2013.
10
11
12
13 56. R. H. Laskar FAT, Biman Paul and Debmalya Chakrabarty. Sample reduction
14 using recursive and segmented data structure analysis. *Journal of Engineering and*
15 *Computer Innovations Vol 2*(4), pp 59-67, 2011.
16
17
18
19 57. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to
20 imputation of missing values. *J Clin Epidemiol* 2006;**59**(10):1087-91.
21
22
23
24 58. Shrive FM, Stuart H, Quan H, et al. Dealing with missing data in a multi-question
25 depression scale: a comparison of imputation methods. *BMC medical research*
26 *methodology* 2006;**6**:57.
27
28
29
30 59. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing
31 data. *J Clin Epidemiol* 2002;**55**(4):329-37.
32
33
34
35 60. Wiharto W, Kusnanto H, Herianto H. Interpretation of Clinical Data Based on
36 C4.5 Algorithm for the Diagnosis of Coronary Heart Disease. *Healthcare*
37 *informatics research* 2016;**22**(3):186-95.
38
39
40
41 61. Rigatti SJ. Random Forest. *J Insur Med* 2017;**47**(1):31-39.
42
43
44
45

46 **Figure Legend**

47
48 Figure 1. ROC curves for LR, SVM, and DT models in predicting mortality of
49 motorcycle riders.
50
51

52
53
54 Figure 2. Illustration of DT model for mortality of motorcycle riders. The boxes
55
56

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 (°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

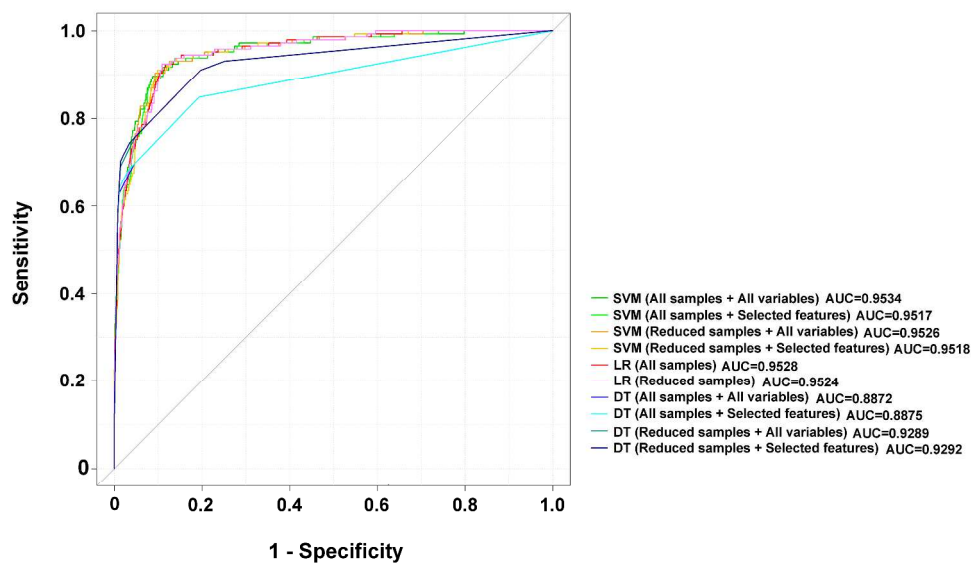


Figure 1. ROC curves for LR, SVM, and DT models in predicting mortality of motorcycle riders.

470x284mm (300 x 300 DPI)

Review only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

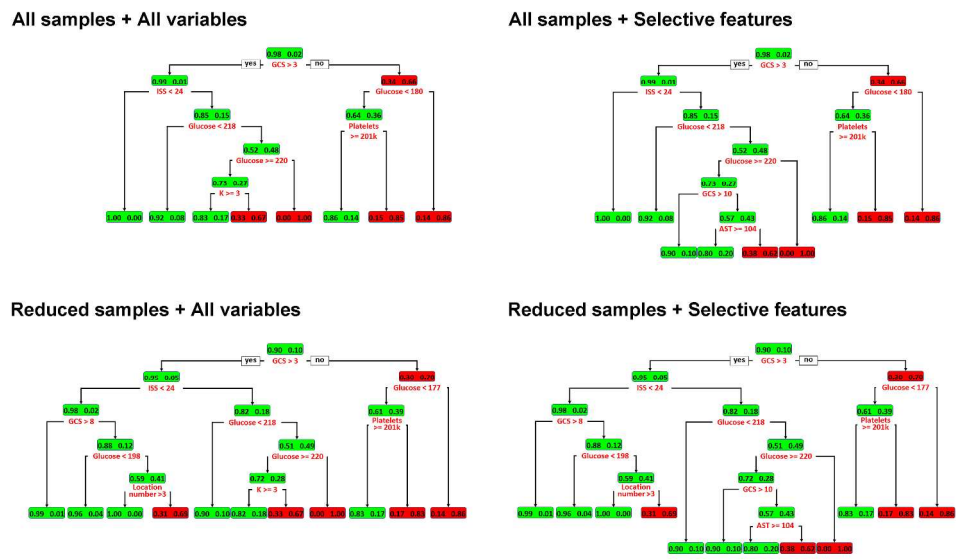
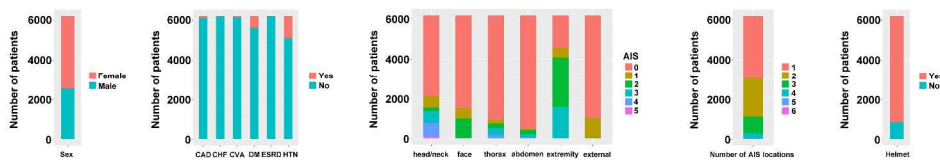


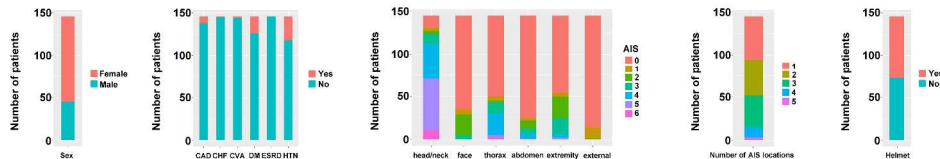
Figure 2. Illustration of DT model for mortality of motorcycle riders. The boxes denote the percentage of patients with discriminating variables from CART analysis. Those who were survival and fatal were indicated with green and red colors, respectively, in the boxes.

340x199mm (300 x 300 DPI)

Survival



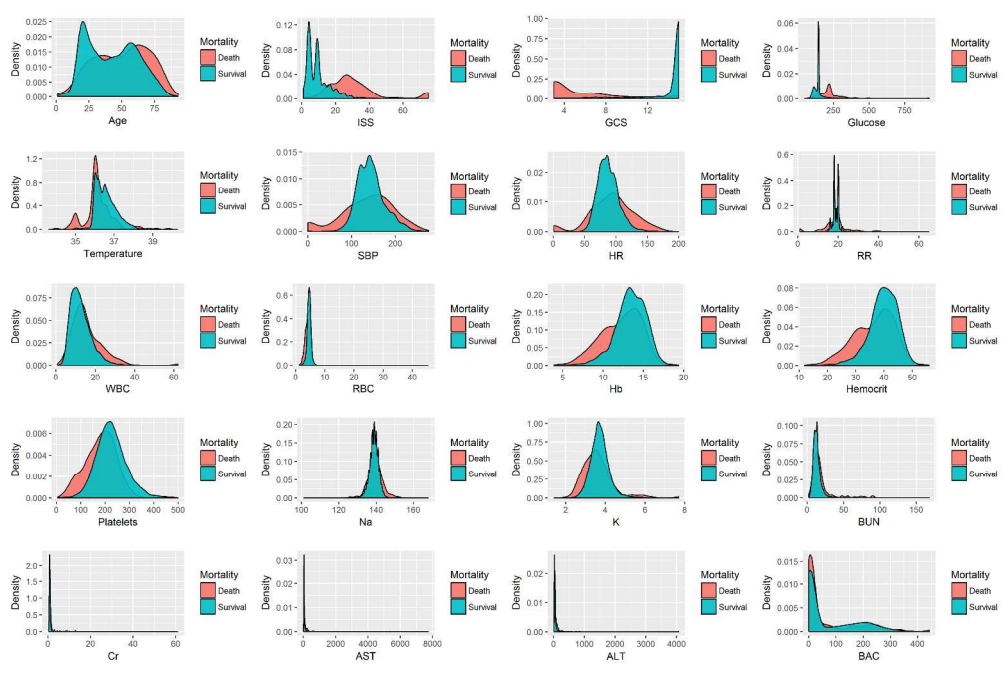
Death



800x337mm (300 x 300 DPI)

Peer review only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



381x254mm (300 x 300 DPI)

view only

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 (b) Give reasons for non-participation at each stage (c) Consider use of a flow diagram	7 - -
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders (b) Indicate number of participants with missing data for each variable of interest	9-11 -
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 (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	11 - -
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:	<i>BMJ Open</i>
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)

SCHOLARONE™
Manuscripts

1
2
3 **Derivation and validation of different machine learning models in mortality**
4 **prediction of trauma motorcycle riders - a cross-sectional retrospective study in**
5 **southern Taiwan**
6
7

8
9 Pao-Jen Kuo^{1†}, M.D.; Shao-Chun Wu^{2†} M.D.; Peng-Chen Chien¹, M.Sc.;
10 Cheng-Shyuan Rau³, M.D.; Yi-Chun Chen¹, M.Sc.; Hsiao-Yun Hsieh¹, M.Sc.;
11 Ching-Hua Hsieh^{1*}, M.D., Ph.D.
12
13
14
15

16
17
18 † Indicates equal contribution in authorship as the first author.
19

20 ¹ Department of Plastic and Reconstructive Surgery, Kaohsiung Chang Gung
21 Memorial Hospital and Chang Gung University College of Medicine, Taiwan, 833
22

23 ² Department of Anesthesiology, Kaohsiung Chang Gung Memorial Hospital and
24 Chang Gung University College of Medicine, Taiwan, 833
25

26 ³ Department of Neurosurgery, Kaohsiung Chang Gung Memorial Hospital and Chang
27 Gung University College of Medicine, Taiwan, 833
28
29
30

31 Pao-Jen Kuo; email: bow110470@gmail.com
32

33 Shao-Chun Wu; email: shaochunwu@gmail.com
34

35 Peng-Chen Chien; e-mail: VENU_CHIEN@hotmail.com
36

37 Cheng-Shyuan Rau; e-mail: ersh2127@cloud.cgmh.org.tw
38

39 Yi-Chun Chen; e-mail: libe320@yahoo.com.tw
40

41 Hsiao-Yun Hsieh; e-mail: sylvia19870714@hotmail.com
42

43 Ching-Hua Hsieh; e-mail: m93chinghua@gmail.com
44
45
46
47

48 Corresponding author: Ching-Hua Hsieh, M.D., PhD.
49

50 Department of Trauma Surgery & Plastic and Reconstructive Surgery, Kaohsiung
51
52
53

1
2
3 Chang Gung Memorial Hospital and Chang Gung University College of Medicine,
4
5 Taiwan

6
7 No.123, Ta-Pei Road, Niao-Song District, Kaohsiung City 833, Taiwan

8
9 Tel: 886-7-7317123 ext: 8002; E-mail: m93chinghua@gmail.com

10
11
12
13
14
15
16 **ABSTRACT**

17
18 **Objectives:** This study aimed to build and test the models of machine learning (ML)
19
20 to predict the mortality of hospitalized motorcycle riders.

21
22 **Setting:** The study was conducted in a level 1 trauma center in southern Taiwan.

23
24 **Participants:** Motorcycle riders who were hospitalized between January 2009 and
25
26 December 2015 were classified into a training set (n=6,306) and test set (n= 946).

27
28 Using the demographic information, injury characteristics, and laboratory data of
29
30 patients, logistic regression (LR), support vector machine (SVM), and decision tree
31
32 (DT) analyses were performed to determine the mortality of individual motorcycle
33
34 riders, under different conditions, using all samples or reduced samples as well as all
35
36 variables or selected features in the algorithm.

37
38 **Primary and secondary outcome measures:** The predictive performance of the
39
40 model was evaluated based on accuracy, sensitivity, specificity, and geometric mean,
41
42 and an analysis of the area under the receiver operating characteristic curves of the
43
44 two different models was carried out.

45
46
47
48 **Results:** In the training set, both LR and SVM had a significantly higher AUC than
49
50 DT; no significant difference was observed in the AUC of LR and SVM, regardless of
51
52 whether all samples or reduced samples and whether all variables or selected features
53
54 were used. In the test set, the performance of the SVM model for all samples with
55
56

1
2
3 selected features was better than that of all other models, with an accuracy of 98.73%,
4 sensitivity of 86.96%, specificity of 99.02%, geometric mean of 92.79%, and AUC of
5 0.9517, in mortality prediction.
6
7

8
9 **Conclusion:** ML can provide a feasible level of accuracy in predicting the mortality
10 of motorcycle riders. Integration of the ML model, particularly the SVM algorithm in
11 the trauma system, may help identify high-risk patients, and therefore, guide
12 appropriate interventions by the clinical staff.
13
14
15
16

17
18
19
20 **KEY WORDS:** Motorcycle accident; mortality; machine learning (ML); logistic
21 regression (LR); support vector machine (SVM); and decision tree (DT)
22
23
24
25

26 **ARTICLE SUMMARY**

27 **STRENGTHS AND LIMITATIONS OF THIS STUDY**

- 28
29
30
31 ■ This study first used machine learning to predict the mortality risk of motorcycle
32 riders.
33
34
35 ■ The SVM model generally works like a black box and cannot identify the
36 relationship between mortality and various explanatory variables.
37
38
39 ■ The incomplete records of patients and exclusion of those who were declared dead
40 in the trauma registry system could cause result bias.
41
42
43
44 ■ The single-center setting may limit the generalizability of the results.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 crash² 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

1
2
3 other or revised predictors, such as blood pressure,¹⁴ co-morbidities, and separate
4 categories for different age groups,¹⁵ into this model, the addition of more predictors
5 to the basic TRISS model did not always result in higher performance.¹⁶⁻¹⁸ Although
6 the revised TRISS derived from the USA National Trauma Database for trauma
7 systems is inaccurate, particularly in the management of predominantly blunt
8 injuries,¹⁹ further development of the model based on advanced methodological
9 quality, performance in the subsets of patient groups, and practical application is
10 required for the prediction of mortality.¹⁶

21
22 Currently, machine learning (ML) had been successfully applied in real-life
23 settings in several fields of study, including automatic medical diagnostics and
24 personalized health care.²⁰⁻²² The application of supervised ML methods to aid
25 diagnosis and prognosis in trauma patients has been a topic of interest. ML is based
26 on how the human brain approaches pattern recognition tasks, thus providing an
27 artificial intelligence-based approach to solve classification problems and improving
28 their efficiency over time.²³ The usefulness of ML is bolstered by the versatility of its
29 techniques and utility for artificial intelligence, such as prediction, classification,
30 planning, recognition, and clustering.^{23,24} Different learning strategies were previously
31 compared using field-specific datasets, of which several had a significantly better
32 predictive power than the more conventional alternatives.²⁵ Examples of multivariate
33 techniques for pattern recognition include but are not limited to LR, support vector
34 machine (SVM), decision tree (DT), and artificial neural networks. LR is a widely
35 used and accepted statistical analysis tool that predicts the probability of the
36 occurrence of an event.²⁶ It aims to build a functional relationship between two or
37 more independent predictors and one dependent outcome variable, with the

1
2
3 assumption that the response variables are linearly related to the coefficients of the
4 predictor variables.²⁶
5
6
7

8
9 SVM uses a training set of data with one or more features to determine an
10 optimal boundary that separates a set of cases. The binary SVM classifier establishes
11 a set of optimal hyperplanes in a high-dimensional space with the maximal margin of
12 the two classes.²⁷ When all training points cannot be separated by the hyperplane, a
13 soft margin method is used to establish a hyperplane that can separate the training
14 data points.^{28 29} Moreover, the SVM model can be used for the classification of
15 problems.³⁰⁻³⁴
16
17
18
19
20
21
22
23
24
25

26 DT is a hierarchical model that is composed of decision rules based on the
27 optimal feature cutoff values that recursively classify independent variables into
28 different groups.³⁵⁻³⁷ It has been built to search for a set of decision rules that can
29 predict an outcome from a set of input variables.^{33 35 36} Some models are used to
30 construct DT models, including classification and regression trees (CART), ID3s,
31 chi-square automatic interaction detector DTs (CHAIDs), and C4.5 and C5.0 DTs [26,
32 28]. CART analysis is a combined approach based on nonparametric and nonlinear
33 variables for recursive partitioning analysis. In addition, it is an innovative DT model
34 in which several predictive variables are used in identifying high-risk patients in
35 various medical fields through progressive binary splits to develop prediction models
36 and to enable better prediction and clinical decision-making.³⁸⁻⁴⁰
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

52 Thus, this study aimed to establish a model for the mortality prediction of
53 motorcycle riders using ML algorithms based on data from a population-based trauma
54
55
56

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.

1
2
3
4
5 Patient samples were divided into a training sample, which was used for
6 predictor discovery and supervised classification to generate a plausible model, and a
7 test sample, which was used to test the performance of the model that was generated
8 in the training sample. Patients with missing data were not included for further
9 analysis. Those who registered within the 6-year period between January 2009 and
10 December 2014 were included in the training set, with a total of 6,306 patients. The
11 group was composed of 6,161 survivors and 145 patients who died. In the test set, 946
12 patients were included, of which 923 survived and 23 died, within the 1-year period
13 between January 2015 and December 2015. The sample similarity was assessed based
14 on Euclidean distance for the quantitative data to reduce the sample that was designed
15 for data analysis.⁴¹ The sample reduction used the Euclidean distance of the `dist`
16 function in the `stats` package in R (R Foundation for Statistical Computing, Vienna,
17 Austria). During sample reduction, the data size can be reduced to speed up
18 calculations in the analysis.⁴² However, considering the exploratory nature of this
19 study, all samples (n=6,306) and reduced samples (n=1,510) in the training set of this
20 study must be analyzed during ML classification.
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

41 **ML classifiers**

42 The present study provides a performance comparison of the three different ML
43 classifiers (LR, SVM, and DT).
44
45
46
47
48
49

50 *Logistic regression*

51 The LR classifier used the `glm` function in the `stats` package in R3.3.3 (R
52 Foundation for Statistical Computing, Vienna, Austria). Univariate LR analyses were
53
54
55
56
57
58
59
60

1
2
3 initially performed to identify the significant predictor variables of the mortality risk.
4
5 A stepwise LR analysis was carried out to control the effects of the confounding
6
7 variables that help identify the independent risk factors of mortality. The selected
8
9 independent risk factors obtained from LR were also used as selected features for the
10
11 implementation of the SVM and DT to explain their importance in determining
12
13 mortality risk.
14
15

16 17 18 *Support vector machine*

19
20 The SVM classifier used the `tune.svm` and `svm` function in the `e1071`
21
22 package in R. In the training set, the SVM classifier was used for the prediction of
23
24 mortality with regard to either all 32 variables or 12 selected features as well as all
25
26 samples and reduced samples in the training set. The mapping procedure was
27
28 performed using the kernel function, which is a matrix of pair-wise similarities
29
30 between data points, such as a linear, polynomial, or radial basis function (RBF).⁴³ In
31
32 the present study, the RBF kernel was used because it can control non-linear
33
34 interactions between class labels and features.⁴⁴ The two main parameters presented in
35
36 the SVM with RBF kernel were the penalty parameter C and kernel hyper-parameter
37
38 γ . The penalty parameter C determined the tradeoff between the fitting error
39
40 minimization and model complexity, whereas the hyper-parameter γ defined the
41
42 nonlinear feature transformation onto a higher dimensional space and controlled the
43
44 tradeoff between errors due to bias and variance in the model.⁴⁵ The optimal operating
45
46 point was estimated by differentiating the parameter C and γ using a grid search for
47
48 each combination of feature selection and dimension reduction with a 10-fold
49
50 cross-validation.⁴⁴
51
52
53
54
55
56
57
58
59
60

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

1
2
3 accurately identified (e.g., percentage of people who survived and were declared
4 dead). In addition, because the geometric mean can provide a good trade-off between
5 sensitivity and specificity in a manner that a better accuracy in both classes leads to a
6 larger value, it was calculated in this study according to the methods used by Sanz J et
7 al.⁴⁸
8
9
10
11
12

13 14 15 *AUC analysis*

16
17 To compare the performance of multiple ML classifiers in multiple training data
18 sets, a nonparametric approach was used to analyze the areas under the correlated
19 ROC curves using the `roc` and `roc.test` function in the `pROC` package in R. This
20 nonparametric approach considers the correlated nature of the data that two or more
21 empirical curves are established based on tests performed on the same individual.⁴⁹
22
23
24
25
26
27
28
29
30

31 All statistical analyses were performed using SPSS 20.0 (IBM Inc., Chicago, IL,
32 USA) and R 3.3.3. For the categorical variables, the chi-square test was carried out to
33 determine the significance of the association between the predictor and outcome
34 variables. For the continuous variables, the student t-test was conducted to analyze
35 normally distributed data, whereas the Kolmogorov–Smirnov test or Mann–Whitney
36 U test was performed to compare non-normally distributed data. Results were
37 presented as mean \pm standard deviation. A p-value < 0.05 was considered statistically
38 significant.
39
40
41
42
43
44
45
46
47
48
49

50 **RESULTS**

51 **Demographic information and injury characteristics of the patients**

52
53
54
55 Patients with head and neck injury had a higher AIS score. However, patients
56
57
58
59
60

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)

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

$$0.03459 * \text{BUN} + 0.10667 * \text{Cr} + 0.00195 * \text{AST} + 0.09513 * \text{Na} + \\ 0.02533 * \text{age} - 0.39968 * \text{GCS} - 0.56396 * \text{temperature} - 0.93232 * \\ \text{number of AIS locations} + 0.14098 * \text{ISS} - 0.95726 * \text{HTN}$$

Reduced samples (n=1,510)

$$Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 5.76780 - 0.00763 * \text{platelet} + 0.00953 * \text{glucose} + \\ 0.03773 * \text{BUN} + 0.00152 * \text{AST} + 0.08630 * \text{Na} + 0.02014 * \text{age} - \\ 0.34116 * \text{GCS} - 0.53370 * \text{temperature} - 0.91439 * \\ \text{number of AIS locations} + 0.12191 * \text{ISS} - 1.00522 * \text{HTN}$$

The LR 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 $\gamma = 0.00390625$. Under the input of all variables into the model, the SVM achieved an accuracy of 98.62% (sensitivity of 62.07% and

1
2
3 specificity of 99.48%) and 94.37% (sensitivity of 59.31% and specificity of 98.10%)
4
5 for all samples and reduced samples, respectively, (Table 3). The AUCs for all
6
7 samples and reduced samples were 0.9534 and 0.9526, respectively, (Figure 1). With
8
9 the use of the selected features in the model, the SVM had an accuracy of 98.62%
10
11 (sensitivity of 64.14% and specificity of 99.43%) and 93.84% (sensitivity of 62.76%
12
13 and specificity of 97.14%) (Table 3) as well as and AUC values of 0.9517 and 0.9518
14
15 for all samples and reduced samples, respectively, (Figure 1).
16
17
18
19

20 *Decision tree*

21
22 As shown in Figure 2, in the DT model, GCS was identified as the variable of the
23
24 initial split with an optimal cut-off value of > 3 . Among the patients with a GCS
25
26 higher than 3, glucose level was selected as the variable of the second split at a
27
28 discrimination level of 180 mg/dL and 177 mg/dL for all samples and reduced
29
30 samples, respectively. Glucose level below 180 mg/dL or 177 mg/dL for all samples
31
32 and reduced samples, respectively, was the best predictor of mortality; the next best
33
34 predictor was platelet count, with an optimal cut-off value of $201 \times 10^3/\mu\text{L}$. For the
35
36 node, in patients with a GCS not greater than 3, ISS below 24, and glucose level
37
38 below 218 mg/dL, these predictors were considered as significant variables for all
39
40 samples and reduced samples along with a GCS > 8 and glucose level below 198
41
42 mg/dL, and the number of AIS locations ≥ 3 was considered as an additional predictor
43
44 for the splitting of the reduced samples. With all the variables used in the model, the
45
46 DT had an accuracy of 98.92% (sensitivity of 62.76% and specificity of 99.77%) and
47
48 95.83% (sensitivity of 68.97% and specificity of 98.68%) for all samples and reduced
49
50 samples, respectively. The AUC values for all samples and reduced samples were
51
52 0.8872 and 0.9289, respectively. With the selected features used in the model, the DT
53
54
55
56
57
58
59
60

1
2
3 had an accuracy of 98.92% (sensitivity of 64.14% and specificity of 99.74%) and
4
5 95.83% (sensitivity of 70.34% and specificity of 98.53%) for all samples and reduced
6
7 samples, respectively. The AUC values for all samples and reduced samples were
8
9 0.8872 and 0.9289, respectively, (Figure 1). In the condition wherein reduced samples
10
11 but not all samples were used in the DT model, the number of AIS locations would be
12
13 added in the split of the node, thus slightly increasing the sensitivity from 62.76% to
14
15 68.97% and from 64.14% to 70.34% with the input composed of all variables and
16
17 selected variables, respectively. In addition, in the condition wherein selected features
18
19 but not all variables were used in the DT model, the level of K was not used in the
20
21 splitting of the node and substituted by the cut-off value of AST (≥ 104 IU/L),
22
23 therefore slightly increasing the sensitivity from 62.76% to 64.14% and from 68.97%
24
25 to 70.34% with input composed of all samples and reduced samples, respectively. The
26
27 AUC values for all samples and reduced samples were 0.8875 and 0.9292,
28
29 respectively, (Figure 1).
30
31
32
33
34

35 *Comparison of the results of AUC analysis*

36
37 When the AUCs for LR, SVM, and DT were used for the training set (Table 4
38
39 and Figure 1), both LR and SVM had a significantly higher AUC than DT, regardless
40
41 of whether all samples or reduced samples and whether all variables or selected
42
43 features were used. However, no significant difference was observed in the AUC of
44
45 LR and SVM, regardless whether all samples or reduced samples as well as all
46
47 variables or selected features were used. In addition, the DT sample reduction had a
48
49 significantly higher AUC than that obtained using all samples. However, no
50
51 significant difference was observed in the AUC of DT, regardless whether all
52
53 variables or selected features were used.
54
55
56
57
58
59
60

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

1
2
3 overestimation of the importance of the risk factors or may cause other potential
4 confounders to be missed, thus affecting each patient's actual risk.⁵² In contrast to LR,
5 which is significantly affected by outliers using a linear discriminant analysis method,
6 the SVM boundary is only minimally affected by outliers that are difficult to separate,
7 despite the complexity of data.⁵³ In addition, the use of kernels in the SVM model is
8 beneficial for non-linear decision boundaries, thus allowing the classifier to solve
9 more difficult classification problems than the linear analysis method.⁵⁴ These three
10 ML models (LR, SVM, and DT) all had an accuracy and specificity of approximately
11 98% and 99%, respectively, but a sensitivity less than or approximately 70% in the
12 training dataset. In this study, both LR and SVM had a significantly higher AUC than
13 DT in the training set, regardless of whether all samples or reduced samples and
14 whether all variables or selected features were used.

15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31 This study included the different variants of SVM, considering the sample size
32 and feature selection, to show all possible improvements and conventional strategies,
33 such as LR or DT. Although the sample reduction for SVM had been proposed to
34 significantly improve the training speed of the SVM and save a lot of storage space,⁵⁵
35 ⁵⁶ kernel use is a more efficient technique for the representation between samples.
36 Thus, the computational complexity of SVM is not wholly governed by the number of
37 samples but by the number of features, which is advantageous for the analysis in
38 high-dimensional settings.⁵⁴ In addition, feature selection in SVM may maximize the
39 AUC.²⁵ When aided by feature selection, the proposed SVM method identifies the
40 most discriminating indexes for mortality prediction. Although both LR and SVM did
41 not have a different AUC in the training procedure, the SVM model for all samples
42 with selected features had a significantly higher sensitivity (86.96%) in predicting the
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 mortality of motorcycle riders in the test set compared to the rest of the models. The
4 higher sensitivity of SVM in the test set compared to that in the training set may be
5 attributed to an improved quality of registered content and less missing data in our
6 registered data after continuous quality assessment and years of working experience
7 with the registers. Such increased sensitivity was also found in the LR model in the
8 test set. With the addition of more data in the model, the SVM model may have an
9 increased predictive power. In the present study, the feasibility of using SVM
10 classification with feature selection can predict the mortality risk of motorcycle riders
11 admitted in trauma care centers. However, the SVM model generally works like a
12 black box, and it cannot identify the relationships between mortality and various
13 explanatory variables. Therefore, this model cannot be directly used to validate our
14 hypothesis on the increased sensitivity in the test set.
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

31 This study has several limitations. First, the patients who had incomplete records
32 were excluded from the analysis. This could have caused result bias, and the results
33 could have been different from the data acquired if the patients with incomplete
34 records were included and the missing data on a variable were replaced by a value
35 that is drawn from an estimate of the distribution of this variable.⁵⁷⁻⁵⁹ Imputation can
36 include patients who might have relevant features for analysis. However, these
37 patients were excluded due to errors in data collection or recording.⁵⁷⁻⁵⁹ Second, the
38 exclusion of patients who were declared dead (either upon arriving at the hospital or
39 at the accident area itself) and injured patients who were discharged against the advice
40 of physicians in the emergency department may cause a potential bias. Third,
41 important data regarding injury mechanism and circumstance, including motorcycle
42 speed and type, helmet material, and impact force during collision, were missing. In
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 addition, the imputation of physiological and laboratory data collected from the time
4 of arrival at the emergency department cannot reflect the dynamic changes in
5 hemodynamic and metabolic variables of the trauma patients when resuscitation is
6 possible. Furthermore, other DT-related methods, such as DT by C4.5,⁶⁰ combined
7 classifiers of LR and DT by C4.5,⁴⁸ and random forest,⁶¹ have extremely satisfying
8 performance in dealing with the classification problem. However, these techniques
9 were not investigated in this study. Lastly, the study population was limited to a single
10 urban trauma center in southern Taiwan, which may not be representative of other
11 populations.
12
13
14
15
16
17
18
19
20
21
22
23

24 **CONCLUSION**

25
26 ML can provide a feasible level of accuracy in predicting the mortality of
27 motorcycle riders. However, there are significant theoretical and practical challenges
28 to the translational implementation of this approach. The results of previous studies
29 are extremely helpful and may help in establishing the first step towards the
30 development of a prediction model that can be integrated into the trauma care system
31 to identify an individual motorcycle rider's risk of mortality.
32
33
34
35
36
37
38
39
40
41

42 **COMPETING INTERESTS**

43
44 The authors declare that they have no competing interests.
45
46
47

48 **AUTHOR CONTRIBUTIONS**

49
50 PJK wrote the manuscript. SCW revised the manuscript. PCC performed the
51 statistical analyses and machine learning programming. CSR analyzed the data in the
52 tables. YCC and HYH collected the data and ensured the integrity of the registered
53
54
55
56
57
58
59
60

1
2
3 data, and CHH designed the study and contributed to the analysis and interpretation of
4
5 data. All authors have read and approved the final manuscript.
6
7

8 9 **DATA SHARING**

10
11 No additional data are available.
12
13

14 15 **ACKNOWLEDGEMENTS**

16
17 This research was supported by a grant from Chang Gung Memorial Hospital
18
19 (CMRPG8F0891). We would like to thank the personnel of Biostatistics Center,
20
21 Kaohsiung Chang Gung Memorial Hospital, for helping us with the statistical work.
22
23
24
25

26 27 **REFERENCES**

- 28
29 1. Weiss H, Agimi Y, Steiner C. Youth motorcycle-related brain injury by state helmet
30
31 law type: United States, 2005-2007. *Pediatrics* 2010;**126**(6):1149-55.
32
33 2. National Highway Traffic Safety Administration (NHTSA) TSFDDH.
34
35 3. Beck LF, Dellinger AM, O'Neil ME. Motor vehicle crash injury rates by mode of
36
37 travel, United States: using exposure-based methods to quantify differences.
38
39 *American journal of epidemiology* 2007;**166**(2):212-8.
40
41 4. Chang HL, Lai CY. Using travel socialization and underlying motivations to better
42
43 understand motorcycle usage in Taiwan. *Accident; analysis and prevention*
44
45 2015;**79**:212-20.
46
47 5. Jou RC, Yeh TH, Chen RS. Risk factors in motorcyclist fatalities in Taiwan. *Traffic*
48
49 *injury prevention* 2012;**13**(2):155-62.
50
51 6. Liang CC, Liu HT, Rau CS, et al. Motorcycle-related hospitalization of adolescents
52
53 in a Level I trauma center in southern Taiwan: a cross-sectional study. *BMC*
54
55
56
57
58
59
60

- 1
2
3 pediatrics 2015;**15**:105.
4
5 7. Liu HT, Liang CC, Rau CS, et al. Alcohol-related hospitalizations of adult
6
7 motorcycle riders. World journal of emergency surgery : WJES 2015;**10**(1):2.
8
9 8. Hsieh CH, Hsu SY, Hsieh HY, et al. Differences between the sexes in
10
11 motorcycle-related injuries and fatalities at a Taiwanese level I trauma center.
12
13 Biomed J 2017;**40**(2):113-20.
14
15 9. Hsieh CH, Liu HT, Hsu SY, et al. Motorcycle-related hospitalizations of the elderly.
16
17 Biomed J 2017;**40**(2):121-28.
18
19 10. Densmore JC, Lim HJ, Oldham KT, et al. Outcomes and delivery of care in
20
21 pediatric injury. Journal of pediatric surgery 2006;**41**(1):92-8; discussion 92-8.
22
23 11. Rogers SC, Campbell BT, Saleheen H, et al. Using trauma registry data to guide
24
25 injury prevention program activities. The Journal of trauma 2010;**69**(4
26
27 Suppl):S209-13.
28
29 12. Norrie J. Mortality prediction in ICU: a methodological advance. The Lancet
30
31 Respiratory medicine 2015;**3**(1):5-6.
32
33 13. Demetriades D, Chan L, Velmanos GV, et al. TRISS methodology: an
34
35 inappropriate tool for comparing outcomes between trauma centers. J Am Coll
36
37 Surg 2001;**193**(3):250-4.
38
39 14. Jones JM, Skaga NO, Sovik S, et al. Norwegian survival prediction model in
40
41 trauma: modelling effects of anatomic injury, acute physiology, age, and
42
43 co-morbidity. Acta Anaesthesiol Scand 2014;**58**(3):303-15.
44
45 15. Bergeron E, Rossignol M, Osler T, et al. Improving the TRISS methodology by
46
47 restructuring age categories and adding comorbidities. J Trauma 2004;**56**(4):760-7.
48
49 16. de Munter L, Polinder S, Lansink KW, et al. Mortality prediction models in the
50
51 general trauma population: A systematic review. Injury 2017;**48**(2):221-29.
52
53
54
55
56
57
58
59
60

- 1
2
3 17. Fueglistaler P, Amsler F, Schuepp M, et al. Prognostic value of Sequential Organ
4 Failure Assessment and Simplified Acute Physiology II Score compared with
5 trauma scores in the outcome of multiple-trauma patients. *Am J Surg*
6 2010;**200**(2):204-14.
7
8
9
- 10
11 18. Kroezen F, Bijlsma TS, Liem MS, et al. Base deficit-based predictive modeling of
12 outcome in trauma patients admitted to intensive care units in Dutch trauma
13 centers. *J Trauma* 2007;**63**(4):908-13.
14
15
- 16
17 19. Stoica B, Paun S, Tanase I, et al. Probability of Survival Scores in Different
18 Trauma Registries: A Systematic Review. *Chirurgia (Bucur)* 2016;**111**(2):115-9.
19
20
- 21
22 20. Cohen AM, Ambert K, McDonagh M. A Prospective Evaluation of an Automated
23 Classification System to Support Evidence-based Medicine and Systematic
24 Review. *AMIA Annual Symposium proceedings AMIA Symposium*
25 2010;**2010**:121-5.
26
27
- 28
29 21. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in
30 cardiovascular risk prediction: applying machine learning to address analytic
31 challenges. *Eur Heart J* 2016.
32
33
- 34
35 22. Szlosek DA, Ferrett J. Using Machine Learning and Natural Language Processing
36 Algorithms to Automate the Evaluation of Clinical Decision Support in Electronic
37 Medical Record Systems. *EGEMS (Washington, DC)* 2016;**4**(3):1222.
38
39
- 40
41 23. Oquendo MA, Baca-Garcia E, Artes-Rodriguez A, et al. Machine learning and
42 data mining: strategies for hypothesis generation. *Mol Psychiatry*
43 2012;**17**(10):956-9.
44
45
- 46
47 24. Kotoku J. An Introduction to Machine Learning. *Igaku butsuri : Nihon Igaku*
48 *Butsuri Gakkai kikanshi = Japanese journal of medical physics : an official journal*
49 *of Japan Society of Medical Physics* 2016;**36**(1):18-22.
50
51
52
53
54
55
56

- 1
2
3 25. Yahya N, Ebert MA, Bulsara M, et al. Statistical-learning strategies generate only
4 modestly performing predictive models for urinary symptoms following external
5 beam radiotherapy of the prostate: A comparison of conventional and
6 machine-learning methods. *Med Phys* 2016;**43**(5):2040.
7
8
9
10
11 26. Siuly, Yin X, Hadjiloucas S, et al. Classification of THz pulse signals using
12 two-dimensional cross-correlation feature extraction and non-linear classifiers.
13 *Comput Methods Programs Biomed* 2016;**127**:64-82.
14
15
16
17
18 27. V. V. Statistical learning theory. New York, NY: John Wiley & Sons 1998.
19
20 28. de Boves Harrington P. Support Vector Machine Classification Trees. *Anal Chem*
21 2015;**87**(21):11065-71.
22
23
24 29. Lee Y. Support vector machines for classification: a statistical portrait. *Methods*
25 *Mol Biol* 2010;**620**:347-68.
26
27
28
29 30. Chen C, Zhang G, Qian Z, et al. Investigating driver injury severity patterns in
30 rollover crashes using support vector machine models. *Accid Anal Prev*
31 2016;**90**:128-39.
32
33
34
35 31. Galatzer-Levy IR, Karstoft KI, Statnikov A, et al. Quantitative forecasting of
36 PTSD from early trauma responses: a Machine Learning application. *J Psychiatr*
37 *Res* 2014;**59**:68-76.
38
39
40
41 32. Li Z, Liu P, Wang W, et al. Using support vector machine models for crash injury
42 severity analysis. *Accid Anal Prev* 2012;**45**:478-86.
43
44
45
46 33. Marucci-Wellman HR, Corns HL, Lehto MR. Classifying injury narratives of
47 large administrative databases for surveillance-A practical approach combining
48 machine learning ensembles and human review. *Accid Anal Prev* 2017;**98**:359-71.
49
50
51
52 34. Patil BM, Joshi RC, Toshniwal D, et al. A new approach: role of data mining in
53 prediction of survival of burn patients. *J Med Syst* 2011;**35**(6):1531-42.
54
55
56
57
58
59
60

- 1
2
3 35. Farion K, Michalowski W, Wilk S, et al. A tree-based decision model to support
4 prediction of the severity of asthma exacerbations in children. *J Med Syst*
5 2010;**34**(4):551-62.
6
7
8
9 36. Zintzaras E, Bai M, Douligeris C, et al. A tree-based decision rule for identifying
10 profile groups of cases without predefined classes: application in diffuse large
11 B-cell lymphomas. *Comput Biol Med* 2007;**37**(5):637-41.
12
13
14 37. Kasbekar PU, Goel P, Jadhav SP. A Decision Tree Analysis of Diabetic Foot
15 Amputation Risk in Indian Patients. *Frontiers in endocrinology* 2017;**8**:25.
16
17
18 38. Guilbault RWR, Ohlsson MA, Afonso AM, et al. External Validation of Two
19 Classification and Regression Tree Models to Predict the Outcome of Inpatient
20 Cardiopulmonary Resuscitation. *J Viral Hepat* 2017;**32**(5):333-38.
21
22
23 39. Shi KQ, Zhou YY, Yan HD, et al. Classification and regression tree analysis of
24 acute-on-chronic hepatitis B liver failure: Seeing the forest for the trees.
25 2017;**24**(2):132-40.
26
27
28 40. Zimmerman RK, Balasubramani GK, Nowalk MP, et al. Classification and
29 Regression Tree (CART) analysis to predict influenza in primary care patients.
30 *BMC Infect Dis* 2016;**16**(1):503.
31
32
33 41. Amaratunga D, Cabrera J, Lee YS. Resampling-based similarity measures for
34 high-dimensional data. *J Comput Biol* 2015;**22**(1):54-62.
35
36
37 42. Bhattacharya S, Mariani TJ. Transformation of expression intensities across
38 generations of Affymetrix microarrays using sequence matching and regression
39 modeling. *Nucleic Acids Res* 2005;**33**(18):e157.
40
41
42 43. Vapnik VN. *The Nature of Statistical Learning Theory*. New York, 2nd ed. 2000.
43
44
45 44. Gultepe E, Green JP, Nguyen H, et al. From vital signs to clinical outcomes for
46 patients with sepsis: a machine learning basis for a clinical decision support
47
48
49
50
51
52
53
54
55
56

- 1
2
3 system. *J Am Med Inform Assoc* 2014;**21**(2):315-25.
- 4
5 45. Chen H, Hu L, Li H, et al. An Effective Machine Learning Approach for
6
7 Prognosis of Paraquat Poisoning Patients Using Blood Routine Indexes. *Basic Clin*
8
9 *Pharmacol Toxicol* 2017;**120**(1):86-96.
- 10
11 46. Chang LY, Wang HW. Analysis of traffic injury severity: an application of
12
13 non-parametric classification tree techniques. *Accid Anal Prev*
14
15 2006;**38**(5):1019-27.
- 16
17 47. Ripley B. tree: Classification and regression trees. R package version 1.0-34. URL
18
19 : <http://CRAN.R-project.org/package=tree>. 2013.
- 20
21
22 48. Sanz J, Paternain D, Galar M, et al. A new survival status prediction system for
23
24 severe trauma patients based on a multiple classifier system. *Comput Methods*
25
26 *Programs Biomed* 2017;**142**:1-8.
- 27
28 49. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or
29
30 more correlated receiver operating characteristic curves: a nonparametric approach.
31
32 *Biometrics* 1988;**44**(3):837-45.
- 33
34 50. Knol MJ, Vandenbroucke JP, Scott P, et al. What do case-control studies estimate?
35
36 Survey of methods and assumptions in published case-control research. *Am J*
37
38 *Epidemiol* 2008;**168**(9):1073-81.
- 39
40
41 51. Gu W, Vieira AR, Hoekstra RM, et al. Use of random forest to estimate population
42
43 attributable fractions from a case-control study of Salmonella enterica serotype
44
45 Enteritidis infections. *Epidemiol Infect* 2015;**143**(13):2786-94.
- 46
47 52. Lemon SC, Roy J, Clark MA, et al. Classification and regression tree analysis in
48
49 public health: methodological review and comparison with logistic regression. *nn*
50
51 *Behav Med* 2003;**26**(3):172-81.
- 52
53
54 53. Chen S, Zhou S, Yin FF, et al. Investigation of the support vector machine
55
56
57
58
59
60

- 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%)	
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%)	
	1	3687 (50.8%)	3631 (51.3%)	56 (33.3%)	
	2	2255 (31.1%)	2205 (31.1%)	50 (29.8%)	
Number of AIS	3	982 (13.5%)	939 (13.3%)	43 (25.6%)	<0.001
locations	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 (°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

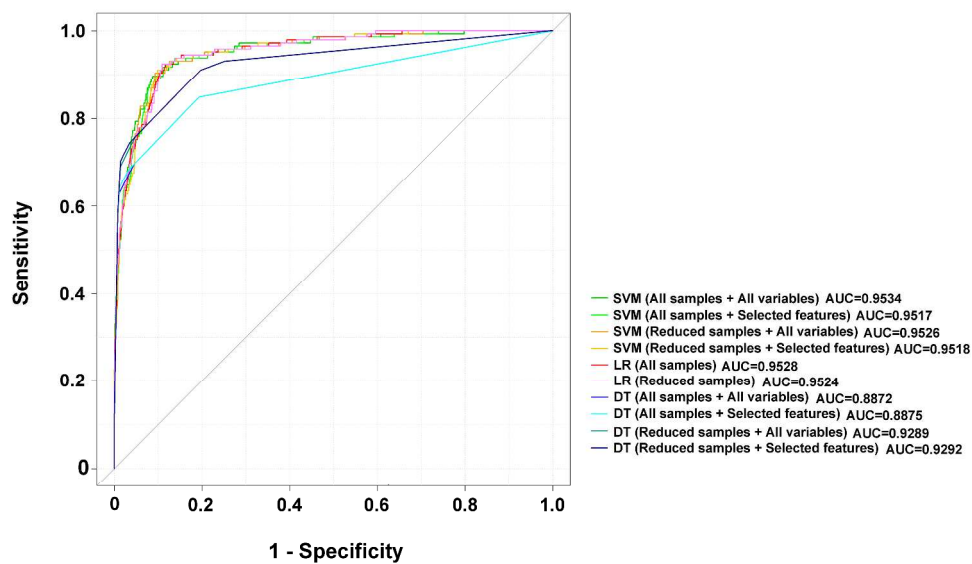


Figure 1. ROC curves for LR, SVM, and DT models in predicting mortality of motorcycle riders.

470x284mm (300 x 300 DPI)

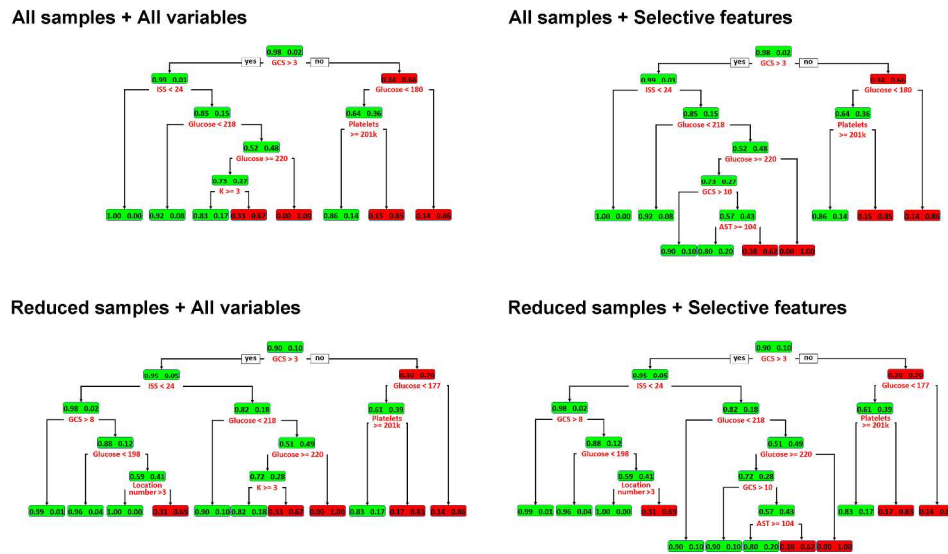
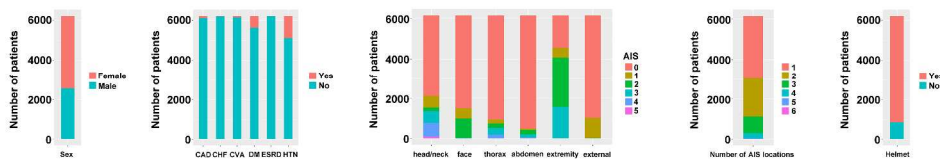


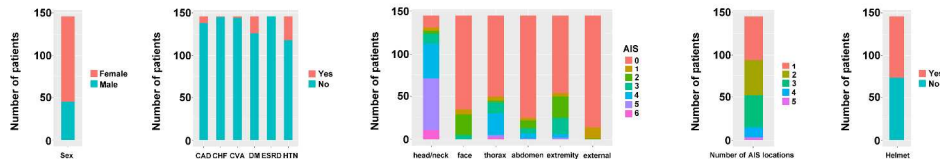
Figure 2. Illustration of DT model for mortality of motorcycle riders. The boxes denote the percentage of patients with discriminating variables from CART analysis. Those who were survival and fatal were indicated with green and red colors, respectively, in the boxes.

340x199mm (300 x 300 DPI)

Survival



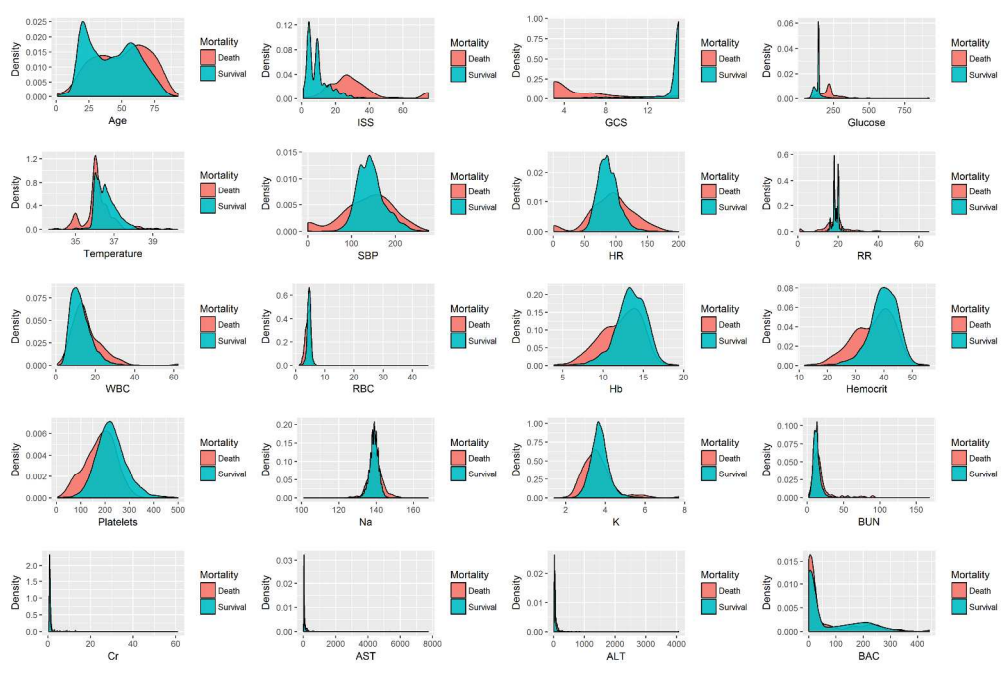
Death



800x337mm (300 x 300 DPI)

Peer review only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



381x254mm (300 x 300 DPI)

view only

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 (b) Give reasons for non-participation at each stage (c) Consider use of a flow diagram	7 - -
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders (b) Indicate number of participants with missing data for each variable of interest	9-11 -
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 (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	11 - -
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.