

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

Pao-Jen Kuo,<sup>1</sup> Shao-Chun Wu,<sup>2</sup> Peng-Chen Chien,<sup>1</sup> Cheng-Shyuan Rau,<sup>3</sup> Yi-Chun Chen,<sup>1</sup> Hsiao-Yun Hsieh,<sup>1</sup> Ching-Hua Hsieh<sup>1</sup>

**To cite:** Kuo P-J, Wu S-C, Chien P-C, *et al.* Derivation and validation of different machine-learning models in mortality prediction of trauma in motorcycle riders: a cross-sectional retrospective study in southern Taiwan. *BMJ Open* 2018;**8**:e018252. doi:10.1136/bmjopen-2017-018252

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2017-018252>).

P-JK and S-CW contributed equally.

Received 16 June 2017

Revised 6 November 2017

Accepted 24 November 2017



CrossMark

For numbered affiliations see end of article.

## Correspondence to

Dr Ching-Hua Hsieh;  
m93chinghua@gmail.com

## ABSTRACT

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

**Setting** The study was conducted in a level-1 trauma centre in southern Taiwan.

**Participants** Motorcycle riders who were hospitalised between January 2009 and December 2015 were classified into a training set (n=6306) and test set (n=946). Using the demographic information, injury characteristics and laboratory data of patients, logistic regression (LR), support vector machine (SVM) and decision tree (DT) analyses were performed to determine the mortality of individual motorcycle riders, under different conditions, using all samples or reduced samples, as well as all variables or selected features in the algorithm.

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

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

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

## BACKGROUND

Motorcycle use is popular in numerous cities because it is a less expensive and convenient

## Strengths and limitations of this study

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

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 with other riders of motor vehicles, motorcycle riders are eight times more likely to be injured per vehicle mile,<sup>1</sup> and they are also 30 times more likely to die in a motor vehicle crash<sup>2</sup> and 58 times more likely to be killed on a per-trip basis.<sup>3</sup> In Taiwan, motorcyclist fatalities account for nearly 60% of all driving fatalities,<sup>4</sup> which are often associated with gender (men), advanced age, lack of helmet use, unlicensed status and driving under the influence of alcohol.<sup>5-9</sup> In addition, head injury is the leading cause of mortality, followed by thoracic and abdominal injuries.<sup>6-9</sup>

Identifying patients who are at high risk is important for the integration of trauma management to maximise resources and improve quality of care.<sup>10 11</sup> 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.<sup>12</sup> 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 patient with trauma based on logistic regression (LR) analysis of variables, including age, anatomical variable (Injury Severity Score; ISS), physiological variable (Revised Trauma Score) and different coefficients for blunt and penetrating injuries. However, TRISS has limitations and fails to determine an accurate classification in 15%–30% of patients with trauma.<sup>13</sup> Even after the incorporation of other or revised predictors, such as blood pressure,<sup>14</sup> comorbidities and separate categories for different age groups,<sup>15</sup> into this model, the addition of more predictors to the basic TRISS model did not always result in higher performance.<sup>16–18</sup> Although the revised TRISS derived from the USA National Trauma Database for trauma systems is inaccurate, particularly in the management of predominantly blunt injuries,<sup>19</sup> further development of the model based on advanced methodological quality, performance in the subsets of patient groups and practical application is required for the prediction of mortality.<sup>16</sup>

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

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

DT is a hierarchical model that is composed of decision rules based on the optimal feature cut-off values that recursively classify independent variables into different groups.<sup>35–37</sup> It has been built to search for a set of decision rules that can predict an outcome from a set of input variables.<sup>33 35 36</sup> Some models are used to construct DT models, including classification and regression trees (CART), iterative dichotomiser 3 (ID3),  $\chi^2$  automatic interaction detector DTs and C4.5 and C5.0 DTs.<sup>26 28</sup> CART analysis is a combined approach based on non-parametric and non-linear variables for recursive partitioning analysis. In addition, it is an innovative DT model in which several predictive variables are used in identifying high-risk patients in various medical fields through progressive binary splits to develop prediction models and to enable better prediction and clinical decision-making.<sup>38–40</sup>

Thus, this study aimed to establish a model for the mortality prediction of motorcycle riders using ML algorithms based on data from a population-based trauma registry in a level 1 trauma centre.

## METHODS

### Ethics statement

Requirement for informed consent was waived according to the institutional review board regulations.

### Data preparation

Detailed patient information was retrieved from the trauma registry system of our institution, a 2400-bed facility and level 1 regional trauma centre, between January 2009 and December 2015. Only patients with trauma who sustained injuries from a motorcycle accident and were hospitalised for treatment were included in the study. Patient information included the following variables: age; sex; use of a helmet; comorbidities, such as coronary artery disease (CAD), congestive heart failure, cerebral vascular accident, diabetes mellitus, end-stage renal disease and hypertension (HTN); vital signs, including temperature, systolic blood pressure, heart rate and respiratory rate; 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); inhospital mortality and laboratory values (white cell count, red blood cell and platelet count; haemoglobin (Hb), haematocrit (Hct), blood urine nitrogen (BUN), creatinine (Cr), alanine aminotransferase (ALT), aspartate aminotransferase (AST), sodium (Na), potassium (K) and glucose level; and blood alcohol concentration) on emergency admission.

Patient samples were divided into a training sample, which was used for predictor discovery and supervised classification to generate a plausible model, and a test sample, which was used to test the performance of the model that was generated in the training sample. Patients with missing data were not included for further analysis. Those who registered within the 6-year period between January 2009 and December 2014 were included in the

training set, with a total of 6306 patients. The group was composed of 6161 survivors and 145 patients who died. In the test set, 946 patients were included, of which 923 survived and 23 died, within the 1-year period between January 2015 and December 2015. The sample similarity was assessed based on Euclidean distance for the quantitative data to reduce the sample that was designed for data analysis.<sup>41</sup> The sample reduction used the Euclidean distance of the `dist` function in the `stats` package in R (R Foundation for Statistical Computing, Vienna, Austria). During sample reduction, the data size can be reduced to speed up calculations in the analysis.<sup>42</sup> However, considering the exploratory nature of this study, all samples ( $n=6306$ ) and reduced samples ( $n=1510$ ) in the training set of this study must be analysed during ML classification.

### ML classifiers

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

#### Logistic regression

The LR classifier used the `glm` function in the `stats` package in R V.3.3.3. Univariate LR analyses were initially performed to identify the significant predictor variables of the mortality risk. A stepwise LR analysis was carried out to control the effects of the confounding variables that help identify the independent risk factors of mortality. The selected independent risk factors obtained from LR were also used as selected features for the implementation of the SVM and DT to explain their importance in determining mortality risk.

#### Support vector machine

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

#### 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).<sup>46,47</sup> 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.<sup>38–40</sup> 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.<sup>38–40</sup>

#### 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 (eg, percentage of people who were declared dead) and specificity, which refers to the proportion of true negatives that were accurately identified (eg, percentage of people who survived and were declared dead). In addition, because the geometric mean can provide a good trade-off between sensitivity and specificity in a manner that a better accuracy in both classes leads to a larger value, it was calculated in this study according to the methods used by Sanz *et al.*<sup>48</sup>

#### AUC analysis

To compare the performance of multiple ML classifiers in multiple training datasets, a non-parametric approach was used to analyse the areas under the correlated ROC curves using the `roc` and `roc.test` functions in the `pROC` package in R. This non-parametric approach considers the correlated nature of the data that two or more empirical curves are established based on tests performed on the same individual.<sup>49</sup>

All statistical analyses were performed using SPSS V.20.0 (IBM) and R V.3.3.3. For the categorical variables, the  $\chi^2$  test was carried out to determine the significance of the association between the predictor and outcome variables. For the continuous variables, the Student's *t*-test was conducted to analyse normally distributed data, whereas the Kolmogorov-Smirnov test or Mann-Whitney

U test was performed to compare non-normally distributed data. Results were presented as mean±SD. A P value <0.05 was considered statistically significant.

## RESULTS

### Demographic information and injury characteristics of the patients

Patients with head and neck injury had a higher AIS score. However, patients with injury in the extremities had a lower AIS score compared with those who survived (table 1 and online supplementary figure 1). The 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 with those who survived (table 1 and online supplementary 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 online supplementary 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 (ie, 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, Na level, age, GCS, temperature, number of AIS locations, ISS and 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=6306)

$$Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = 4.71648 - 0.00846 \times \text{platelet} \\ + 0.01189 \times \text{glucose} + 0.03459 \times \text{BUN} + 0.10667 \times \text{Cr} \\ + 0.00195 \times \text{AST} + 0.09513 \times \text{Na} + 0.02533 \times \text{age} \\ - 0.39968 \times \text{GCS} - 0.56396 \times \text{temperature} \\ - 0.93232 \times \text{number of AIS locations} + 0.14098 \times \text{ISS} \\ - 0.95726 \times \text{HTN}$$

Reduced samples (n=1510)

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

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  $\gamma$ ) of the SVM model must be determined. The accuracy was highly robust to small changes in the hyperparameters. 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  $\gamma$ . 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 specificity of 99.48%) and 94.37% (sensitivity of 59.31% and specificity of 98.10%) for all samples and reduced samples, respectively (table 3). The AUCs for all samples and reduced samples were 0.9534 and 0.9526, respectively (figure 1). With the use of the selected features in the model, the SVM had an accuracy of 98.62% (sensitivity of 64.14% and specificity of 99.43%) and 93.84% (sensitivity of 62.76% and specificity of 97.14%) (table 3), and AUC values of 0.9517 and 0.9518 for all samples and reduced samples, respectively (figure 1).

#### Decision tree

As shown in figure 2, in the DT model, GCS was identified as the variable of the initial split with an optimal cut-off value of >3. Among the patients with a GCS higher than 3, glucose level was selected as the variable of the second split at a discrimination level of 180 mg/dL and 177 mg/dL for all samples and reduced samples, respectively. Glucose level below 180 mg/dL or 177 mg/dL for all samples and reduced samples, respectively, was the best predictor of mortality; the next best predictor was platelet count, with an optimal cut-off value of  $201 \times 10^3/\mu\text{L}$ . For the node, in patients with a GCS not greater than 3, ISS below 24 and glucose level below 218 mg/dL, these predictors were considered as significant variables for all samples and reduced samples along with a GCS >8 and glucose level below 198 mg/dL, and the number of AIS locations  $\geq 3$  was considered as an additional predictor for the splitting of the reduced samples. With all the variables used in the model, the DT had an accuracy of 98.92% (sensitivity of 62.76% and specificity of 99.77%) and 95.83% (sensitivity of 68.97% and specificity of 98.68%) for all samples and reduced samples, respectively. The AUC values for all samples and reduced samples were 0.8872 and 0.9289, respectively. With the selected features used in the model, the DT had an accuracy of 98.92% (sensitivity of 64.14% and specificity of 99.74%) and 95.83% (sensitivity of 70.34% and specificity of 98.53%) for all samples and reduced samples, respectively. The AUC values for all samples and reduced

**Table 1** Demographics and injury characteristics of the patients regarding gender, helmet-wearing status, comorbidities, 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%)	

Continued

**Table 1** Continued

Variables		Total (N=7252)	Survival (n=7084)	Mortality (n=168)	P value
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%)	

AIS, Abbreviated Injury Scale; CAD, coronary artery disease; CHF, congestive heart failure; CVA, cerebral vascular accident; DM, diabetes mellitus; ESRD, end-stage renal disease; HTN, hypertension.

samples were 0.8872 and 0.9289, respectively (figure 1). In the condition wherein reduced samples but not all samples were used in the DT model, the number of AIS locations would be added in the split of the node, thus slightly increasing the sensitivity from 62.76% to 68.97% and from 64.14% to 70.34% with the input composed of all variables and selected variables, respectively. In addition, in the condition wherein selected features but not all variables were used in the DT model, the level of K was not used in the splitting of the node and substituted

by the cut-off value of AST ( $\geq 104$  IU/L), therefore slightly increasing the sensitivity from 62.76% to 64.14% and from 68.97% to 70.34% with input composed of all samples and reduced samples, respectively. The AUC values for all samples and reduced samples were 0.8875 and 0.9292, respectively (figure 1).

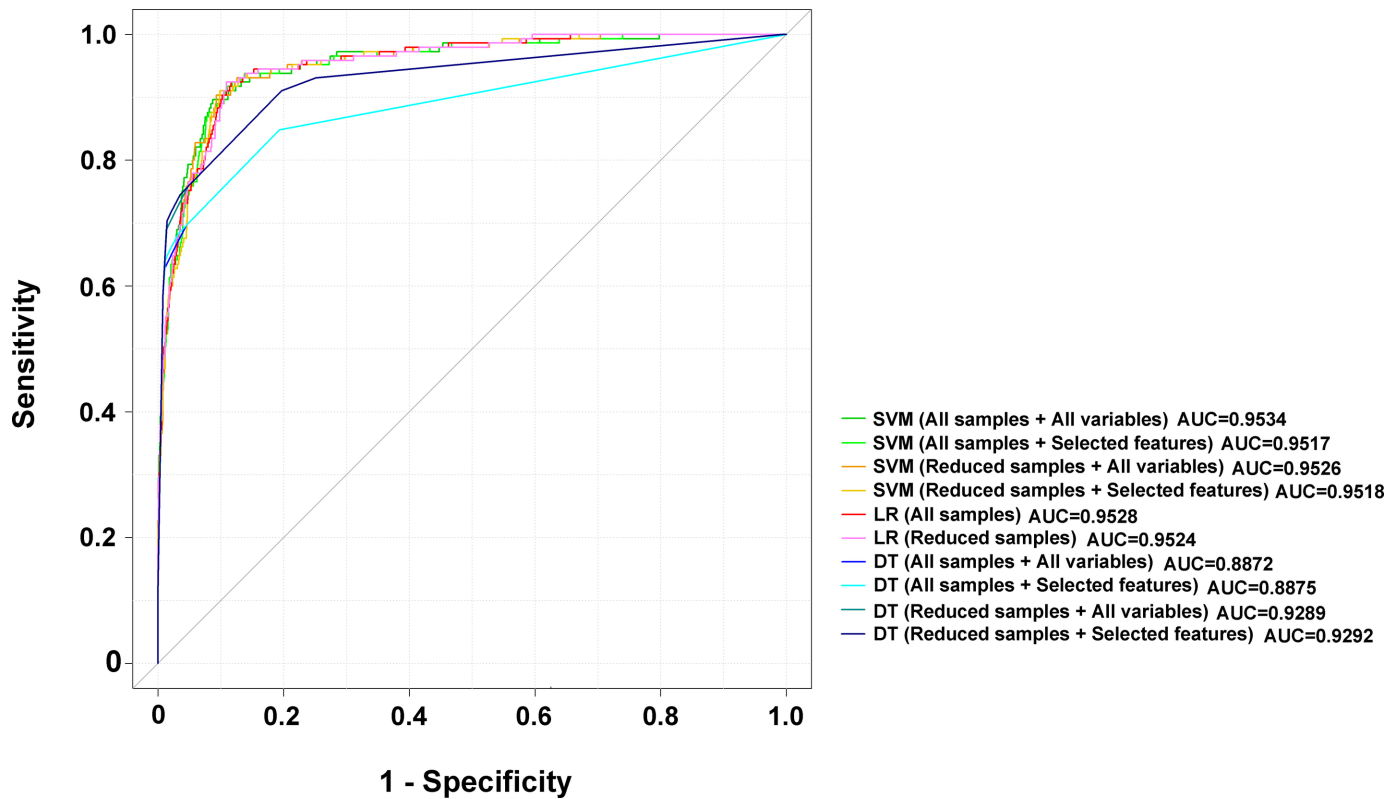
#### Comparison of the results of AUC analysis

When the AUCs for LR, SVM and DT were used for the training set (table 4 and figure 1), both LR and SVM

**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 (mm Hg)	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 ( $10^{12}/L$ )	4.6 (0.8)	4.6 (0.8)	4.3 (1.1)	<0.001
WCC ( $10^9/L$ )	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 ( $10^3/\mu L$ )	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

ALT, alanine aminotransferase; AST, aspartate aminotransferase; BAC, blood alcohol concentration; BUN, blood urea nitrogen; Cr, creatinine; GCS, Glasgow Coma Scale; Hb, haemoglobin; Hct, haematocrit; HR, heart rate; ISS, Injury Severity Score; K, potassium; Na, sodium; RBC, red blood cell; RR, respiratory rate; SBP, systolic blood pressure; WCC, white cell count.



**Figure 1** ROC curves for LR, SVM and DT models in predicting mortality of motorcycle riders. AUC, area under the curve; DT, decision tree; LR, logistic regression; ROC, receiver operating characteristic; SVM, support vector machine.

had a significantly higher AUC than DT, regardless of whether all samples or reduced samples and whether all variables or selected features were used. However, no significant difference was observed in the AUC of LR and SVM, regardless of whether all samples or reduced samples, as well as all variables or selected features, were used. In addition, the DT sample reduction had a significantly higher AUC than that obtained using all samples. However, no significant difference was observed in the AUC of DT, regardless of whether all variables or selected features were used.

#### Performance of ML classifiers in test set

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

In addition, the SVM model for all samples with selected features had the highest sensitivity and geometric mean.

#### DISCUSSION

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

**Table 3** Summary of 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

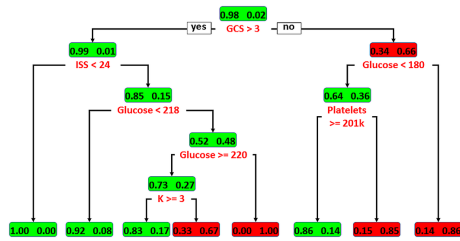
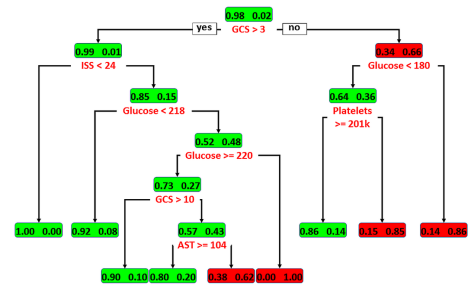
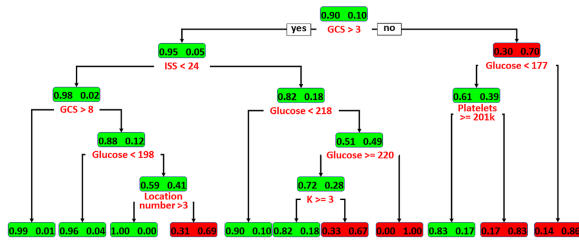
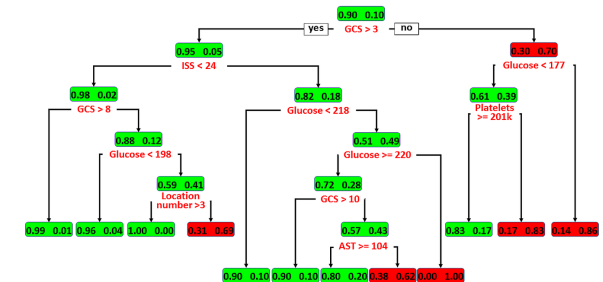
DT, decision tree; LR, logistic regression; SVM, support vector machine.

or approximately 70% in the training dataset. In this study, both LR and SVM had a significantly higher AUC than DT in the training set, regardless of whether all samples or reduced samples and whether all variables or selected features were used.

This study included the different variants of SVM, considering the sample size and feature selection, to show all possible improvements and conventional strategies, such as LR or DT. Although the sample reduction for SVM had been proposed to significantly improve the training speed of the SVM and save a lot of storage space,<sup>55,56</sup> kernel use is a more efficient technique for the representation between samples. Thus, the computational complexity of SVM is not wholly governed by the number of samples but by the number of features, which is advantageous for the analysis in high-dimensional settings.<sup>54</sup> In addition, feature selection in SVM may maximise the AUC.<sup>25</sup> When aided by feature selection, the proposed SVM method identifies the most discriminating indexes for mortality prediction. Although both LR and SVM did not have a

different AUC in the training procedure, the SVM model for all samples with selected features had a significantly higher sensitivity (86.96%) in predicting the mortality of motorcycle riders in the test set compared with the rest of the models. The higher sensitivity of SVM in the test set compared with that in the training set may be attributed to an improved quality of registered content and less missing data in our registered data after continuous quality assessment and years of working experience with the registers. Such increased sensitivity was also found in the LR model in the test set. With the addition of more data in the model, the SVM model may have an increased predictive power. In the present study, the feasibility of using SVM classification with feature selection can predict the mortality risk of motorcycle riders admitted in trauma care centres. However, the SVM model generally works like a black box, and it cannot identify the relationships between mortality and various explanatory variables. Therefore, this model cannot be directly used to validate our hypothesis on the increased sensitivity in the test set.



**All samples + All variables**

**All samples + Selective features**

**Reduced samples + All variables**

**Reduced samples + Selective features**


**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 colours, respectively. CART, classification and regression trees; DT, decision tree.

This study has several limitations. First, the patients who had incomplete records were excluded from the analysis. This could have caused result bias, and the results could have been different from the data acquired if the patients with incomplete records were included and the missing data on a variable were replaced by a value that is drawn from an estimate of the distribution of this variable.<sup>57–59</sup> Imputation can include patients who might have relevant features for analysis. However, these patients were excluded due to errors in data collection or recording.<sup>57–59</sup> Second, the exclusion of patients who were declared dead (either

on arriving at the hospital or at the accident area itself) and patients with injuries who were discharged against the advice of physicians in the emergency department may cause a potential bias. Third, important data regarding injury mechanism and circumstance, including motorcycle speed and type, helmet material and impact force during collision, were missing. In addition, the imputation of physiological and laboratory data collected from the time of arrival at the emergency department cannot reflect the dynamic changes in haemodynamic and metabolic variables of the patients with trauma when resuscitation is

**Table 4** Comparison of AUC between LR, SVM and DT models in the training set

		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	

\*P<0.05.

AS, all samples; AUC, area under the curve; AV, all variables; DT, decision tree; LR, Logistic regression; RS, reduced samples; SF, selected features; SVM, support vector machine.

possible. Furthermore, other DT-related methods, such as DT by C4.5,<sup>60</sup> combined classifiers of LR and DT by C4.5,<sup>48</sup> and random forest,<sup>61</sup> have extremely satisfying performance in dealing with the classification problem. However, these techniques were not investigated in this study. Lastly, the study population was limited to a single urban trauma centre in southern Taiwan, which may not be representative of other populations.

## CONCLUSION

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

## Author affiliations

<sup>1</sup>Department of Plastic and Reconstructive Surgery, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Kaohsiung, Taiwan

<sup>2</sup>Department of Anesthesiology, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Kaohsiung, Taiwan

<sup>3</sup>Department of Neurosurgery, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Kaohsiung, Taiwan

**Contributors** P-JK wrote the manuscript. S-CW revised the manuscript. P-CC performed the statistical analyses and machine-learning programming. C-SR analysed the tables. Y-CC and H-YH collected the data and are responsible for the integrity of the registered data. C-HH designed the study and contributed to the analysis and interpretation of data. P-JK and S-CW contributed to this article equally. All authors have read and approved the final manuscript.

**Funding** This study was funded by Chang Gung Memorial Hospital with the grant CMRPG8F0891.

**Competing interests** None declared.

**Ethics approval** The Institutional Review Board of Chang Gung Memorial Hospital (reference no: 201600653B0).

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>

© Article author(s) (or their employer(s) unless otherwise stated in the text of the article) 2018. All rights reserved. No commercial use is permitted unless otherwise expressly granted.

## REFERENCES

- Weiss H, Agimi Y, Steiner C. Youth motorcycle-related brain injury by state helmet law type: United States, 2005-2007. *Pediatrics* 2010;126:1149-55.
- National Highway Traffic Safety Administration (NHTSA). 2011 motorcycles traffic safety fact sheet. May 2013. <http://www-nrd.nhtsa.dot.gov/Pubs/811765.pdf>.
- 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. *Am J Epidemiol* 2007;166:212-8.
- Chang HL, Lai CY. Using travel socialization and underlying motivations to better understand motorcycle usage in Taiwan. *Accid Anal Prev* 2015;79:212-20.
- Jou RC, Yeh TH, Chen RS. Risk factors in motorcyclist fatalities in Taiwan. *Traffic Inj Prev* 2012;13:155-62.
- 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 Pediatr* 2015;15:105.
- Liu HT, Liang CC, Rau CS, et al. Alcohol-related hospitalizations of adult motorcycle riders. *World J Emerg Surg* 2015;10:2.
- Hsieh CH, Hsu SY, Hsieh HY, et al. Differences between the sexes in motorcycle-related injuries and fatalities at a Taiwanese level I trauma center. *Biomed J* 2017;40:113-20.
- Hsieh CH, Liu HT, Hsu SY, et al. Motorcycle-related hospitalizations of the elderly. *Biomed J* 2017;40:121-8.
- Densmore JC, Lim HJ, Oldham KT, et al. Outcomes and delivery of care in pediatric injury. *J Pediatr Surg* 2006;41:92-8.
- Rogers SC, Campbell BT, Saleheen H, et al. Using trauma registry data to guide injury prevention program activities. *J Trauma* 2010;69(4 Suppl):S209-S213.
- Norrie J. Mortality prediction in ICU: a methodological advance. *Lancet Respir Med* 2015;3:5-6.
- Demetriades D, Chan L, Velmanos GV, et al. TRISS methodology: an inappropriate tool for comparing outcomes between trauma centers. *J Am Coll Surg* 2001;193:250-4.
- Jones JM, Skaga NO, Søvik S, et al. Norwegian survival prediction model in trauma: modelling effects of anatomic injury, acute physiology, age, and co-morbidity. *Acta Anaesthesiol Scand* 2014;58:303-15.
- Bergeron E, Rossignol M, Osler T, et al. Improving the TRISS methodology by restructuring age categories and adding comorbidities. *J Trauma* 2004;56:760-7.
- de Munter L, Polinder S, Lansink KW, et al. Mortality prediction models in the general trauma population: A systematic review. *Injury* 2017;48:221-9.
- Fueglistaler P, Amsler F, Schüepp M, et al. Prognostic value of sequential organ failure assessment and simplified acute physiology II Score compared with trauma scores in the outcome of multiple-trauma patients. *Am J Surg* 2010;200:204-14.
- 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;63:908-13.
- Stoica B, Paun S, Tanase I, et al. Probability of survival scores in different trauma registries: a systematic review. *Chirurgia* 2016;111:115-9.
- Cohen AM, Ambert K, McDonagh M. A Prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. *AMIA Annu Symp Proc* 2010;2010:121-5.
- 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;ehw302.
- 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* 2016;4:5.
- Oquendo MA, Baca-Garcia E, Artés-Rodríguez A, et al. Machine learning and data mining: strategies for hypothesis generation. *Mol Psychiatry* 2012;17:956-9.
- Kotoku J. An Introduction to Machine Learning. *Igaku Butsuri* 2016;36:18-22.
- 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;43:2040-52.
- 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.
- VV. *Statistical learning theory*. New York: John Wiley, 1998.
- de Boves Harrington P. Support vector machine classification trees. *Anal Chem* 2015;87:11065-71.
- Lee Y. Support vector machines for classification: a statistical portrait. *Methods Mol Biol* 2010;620:347-68.
- 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.
- 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;35:1531–42.
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;34:551–62.
36. Zintzaras E, Bai M, Douligeris C, *et al.* A tree-based decision rule for identifying profile groups of cases without predefined classes: application in diffuse large B-cell lymphomas. *Comput Biol Med* 2007;37:637–41.
37. Kasbekar PU, Goel P, Jadhav SP. A decision tree analysis of diabetic foot amputation risk in indian patients. *Front Endocrinol* 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 Intensive Care Med* 2017;32:333–8.
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. *J Viral Hepat* 2017;24:132–40.
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;16:503.
41. Amaratunga D, Cabrera J, Lee YS. Resampling-based similarity measures for high-dimensional data. *J Comput Biol* 2015;22:54–62.
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:e157.
43. Vapnik VN. *The Nature of Statistical Learning Theory*. 2nd ed New York, 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;21:315–25.
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;120:86–96.
46. Chang LY, Wang HW. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. *Accid Anal Prev* 2006;38:1019–27.
47. Ripley B. Tree: Classification and regression trees. R package version 1.0-34, 2013. <http://CRAN.R-project.org/package=tree>
48. Sanz J, Paternain D, Galar M, *et al.* A new survival status prediction system for severe trauma patients based on a multiple classifier system. *Comput Methods Programs Biomed* 2017;142:1–8.
49. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
50. Knol MJ, Vandenbroucke JP, Scott P, *et al.* What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *Am J Epidemiol* 2008;168:1073–81.
51. Gu W, Vieira AR, Hoekstra RM, *et al.* Use of random forest to estimate population attributable fractions from a case-control study of Salmonella enterica serotype Enteritidis infections. *Epidemiol Infect* 2015;143:2786–94.
52. Lemon SC, Roy J, Clark MA, *et al.* Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Ann Behav Med* 2003;26:172–81.
53. Chen S, Zhou S, Yin FF, *et al.* Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. *Med Phys* 2007;34:3808–14.
54. Orrù 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;36:1140–52.
55. Du Hongle LQ, Jing C. *Reduce the samples for svm based on euclidean distance*. 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization, 2013.
56. Laskar FAT RH, Paul B, Chakrabarty D. Sample reduction using recursive and segmented data structure analysis. *J Eng Comput Innov* 2011;59:67.
57. Donders AR, van der Heijden GJ, Stijnen T, *et al.* Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59:1087–91.
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 Med Res Methodol* 2006;6:57.
59. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. *J Clin Epidemiol* 2002;55:329–37.
60. Wiharto W, Kusnanto H, Herianto H. Interpretation of clinical data based on C4.5 algorithm for the diagnosis of coronary heart disease. *Health Inform Res* 2016;22:186–95.
61. Rigatti SJ. Random Forest. *J Insur Med* 2017;47:31–9.