Computers in Biology and Medicine 107 (2019) 145–152

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Outcome prediction with serial neuron-specific enolase and machine learning in anoxic-ischaemic disorders of consciousness

Emily Muller[a,b,*], Jonathan P. Shock[c], Andreas Bender[d,e], Julian Kleeberger[d], Tobias Högen[d], Martin Rosenfelder[e,f], Bubacarr Bah[a,b], Alex Lopez-Rolon[e]

[a] Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa
[b] African Institute Or Mathematical Sciences, Cape Town, South Africa
[c] Department of Mathematics and Applied Mathematics, University of Cape Town, Cape Town, South Africa
[d] Department of Neurology, University of Munich, Munich, Germany
[e] Department of Neurology, Therapiezentrum Burgau, Burgau, Germany
[f] Clinical and Biological Psychology, Institute of Psychology and Education, Ulm University, Ulm, Germany

ARTICLE INFO

Keywords:
Machine learning
Neurological outcome
Clinical prediction modelling
Imputation

ABSTRACT

Background: The continuation of life-sustaining therapy in critical care patients with anoxic-ischemic disorders of consciousness (AI-DOC) depends on prognostic tests such as serum neuron-specific enolase (NSE) concentration levels.

Objectives: To apply predictive models using machine learning methods to examine, one year after onset, the prognostic power of serial measurements of NSE in patients with AI-DOC. To compare the discriminative accuracy of this method to both standard single-day, absolute, and difference-between-days, relative NSE levels.

Methods: Classification algorithms were implemented and K-nearest neighbours (KNN) imputation was used to avoid complete case elimination of patients with missing NSE values. Non-imputed measurements from Day 0 to Day 6 were used for single day and difference-between-days.

Results: The naive Bayes classifier on imputed serial NSE measurements returned an AUC of (0.81 ± 0.07) for $n = 126$ patients (100 poor outcome). This was greater than logistic regression (0.73 ± 0.08) and all other classifiers. Naive Bayes gave a specificity and sensitivity of 96% and 49%, respectively, for an (uncalibrated) probability decision threshold of 90%. The maximum AUC for a single day was Day 3 (0.75) for a subset of $n = 79$ (61 poor outcome) patients, and for differences between Day 1 and Day 4 (0.81) for a subset of $n = 46$ (39 poor outcome) patients.

Conclusion: Imputation avoided the elimination of patients with missing data and naive Bayes outperformed all other classifiers. Machine learning algorithms could detect automatically discriminatory features and the overall predictive power increased from standard methods due to the larger data set.

Code availability: Data analysis code is available under GNU at: https://github.com/emilymuller1991/outcome_prediction_nse.

1. Introduction

The global[1] incidence of cardiac arrests (CA) is estimated to be 60/100,000 person-years [1]. Although immediate cardiopulmonary resuscitation (CPR) can be life-saving in CA cases [2], CA survivors may present an anoxic-ischemic encephalopathy (AIE) after successful CPR [3]. AIE may lead to a syndrome that falls within the spectrum of severe disorders of consciousness (DOC) [4]. DOC can be considered to form a continuum of altered states of consciousness that consists of coma, vegetative state (VS; also known as unresponsive wakefulness syndrome; UWS), and minimally conscious state (MCS) [5]. Current guidelines recommend running prognostic tests as early as possible after CA in order to predict clinical outcome when confronted with unresponsive hypoxic patients that can be classified as AIE/DOC. Clinical outcome is usually assessed with brief single-item clinical scales such as the Cerebral Performance Category (CPC) [4,6]. The objective of prognostic tests is primarily to determine the likelihood of a poor global clinical outcome [7]. According to the American Academy

[*] Corresponding author. Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa.
E-mail address: emily@aims.ac.za (E. Muller).

[1] Out-of-hospital CAs from 67 studies from Europe, North America, Asia and Australia.

https://doi.org/10.1016/j.compbiomed.2019.02.006
Received 3 November 2018; Received in revised form 11 February 2019; Accepted 12 February 2019

of Neurology (AAN) a poor outcome can be defined as death, persistent unresponsiveness or a severe disability that requires full-time nursing care [7]. Based on the results of prognostic tests, clinicians, especially critical care specialists, decide whether to continue or discontinue life-sustaining therapy. However, the process leading to such a decision may be based on misdiagnosis or inaccurate prognosis [8]. Prognosis can be inaccurate due to a biased interpretation of clinical data that produces falsely pessimistic predictions. Therefore, current research is directing its attention to increasing the predictive value of tests based on biological rather than behavioural parameters in order to reduce human error in prognostication and diagnosis. The present study is part of this ongoing clinical effort.

Neuron-specific enolase (NSE) is one important biological marker of severe brain injury. NSE is the dominant enolase-isoenzyme found almost exclusively in neurons and neuroendocrine cells [9]. Serum NSE concentration levels can be considered to be directly proportional to the extent of brain damage caused by conditions such as cerebral ischeamia [10]. Due to this fact and also to its associated low cost, serum NSE concentration level is often used as a prognostic parameter [11]. In 2006, the AAN recommended a cut-off value of more than $33\mu g/L$ of serum NSE, 1–3 days after CA, to predict poor outcome. Multiple studies have proved the prognostic value of this enzyme biomarker [9,12–14] and on-going research continues to report NSE cut-off values that aim to minimise inaccurate prognosis (false positives) [15–17]. Reported values range from as low as $8.8\mu g/L$ [18] to $90\mu g/L$ [15]. This wide range can be attributed to variability in measurement equipment and sample handling [15,19,20], differences in the definition of outcome and moment of outcome [21] and methodological issues regarding the heterogeneity of patient samples [22]. Such variability for a single day cut-off can be interpreted as indicating that research on NSE as a prognostic parameter should: (i) examine an alternative to cut-off values that consider the relative changes in NSE values over time [13,16,23], and (ii) attempt to build data-driven predictive models that provide probabilistic outputs of poor outcome whilst accounting for confounding variables [24].

The present study examines the use of machine learning classification algorithms in building predictive models of poor outcome given serial NSE measurements.

1.1. Clinical prediction models

Prediction models may assist clinical decision making by providing estimates of the individuals probability of risk [24]. For example, predictive models that include attributes collected at admission to multiple study centres have been built for traumatic brain injury from a cohort of over 8,000 patients, with external validation on 6,681 patients [25]. Despite their potential impact and increasing interest [24], clinical prediction models are rarely implemented by clinicians. This is due in part to the highly demanding statistical analysis required, but also due to issues related to accuracy and generalisability, as well as to internal and external model validity [26,27]. For instance, although the Framingham coronary heart disease (CHD) risk score is a widely used clinical prediction model, it was found that it overestimates risk in Asian populations five-fold [26]. Thus, the CHD risk score was recalibrated to account for risk prevalence in Asian populations [28]. The present study is limited to only internal validation.

1.2. Classification algorithms

The detection of patterns in serial NSE levels for the prediction of poor neurological outcome is a supervised classification task [29]. Within supervised learning the true class label of the dataset used for training is known, and parameter estimation is performed iteratively, subject to minimising an error function; a function of the actual and predicted class. Classification methods include, but are not limited to, logistic regression, support vector machines, nearest neighbours clustering and naive Bayes classification [29]. Artificial neural networks and decision trees have not been included due to lack of parameter convergence.

A calibrated classifier provides the likelihood of risk, $P(\text{outcome}|\text{variables})$, and the decision cut-off can be clinically assessed depending on the cost-benefit of true and false positives. Extracting relevant information from the variables and assessing which variables to include in the model is called feature extraction and selection. For large patient datasets, including many of the potential confounding variables may increase the predictive power. However, with small datasets, this partitions the dataset and reduces predictive power. Therefore, the average likelihood over all patient variables, $P(\text{outcome}|\text{serial NSE}) = \sum_i P(\text{outcome}|\text{variables}_i,\text{serial NSE})$, for each patient i is considered.

1.2.1. Machine learning classifiers

1.2.1.1. Logistic regression. Logistic regression allows for a binary classification of data using a logistic function. Over-fitting may be problematic when there are extreme outliers within a dataset which can result in poor predictive generalisation for previously unseen patient data [30]. Logistic regression has previously been implemented for multivariate analysis in predicting poor neurological outcome [13,17,31]. These studies typically include an NSE variable as well as demographic and clinical variables, such as out-of-hospital (OHCA) or in-hospital cardiac arrest (IHCA). A previous study reported that a change of $>2mg/L$ in serum NSE concentration occurring between 24 and 48 h [16] had a strong correlation with poor outcome (odds ratio 9.8, CI 3.5–27.7). Although logistic regression does provide useful information regarding the effects of each variable in the model, the task in predictive modelling is primarily an estimation problem [24]. The predictive power of a set of classification algorithms are thus examined here.

1.2.1.2. Support vector machines. Linear support vector machines (SVMs) define a straight line decision boundary in feature space. The decision boundary separates each class, subject to minimising the misclassification rate and maximising the margin. The margin is the distance from the decision boundary to the closest point. SVMs exploit good generalisability and global convergence [29]. Kernel SVMs allow for non-linear decision boundaries and further include pairwise similarities between variables [30].

1.2.1.3. k nearest neighbours. The k nearest neighbours (KNN) algorithm classifies each patient based on the majority class membership of their k nearest neighbours within a feature space, k typically being a small integer. Euclidean distance is a commonly used distance metric though this can vary depending on the feature space [29].

1.2.1.4. Naive bayes classifiers. A naive Bayes classifier aims to maximise the likelihood of the posterior distribution, $P(\text{NSE}|\text{outcome})$, assumed to be independent for each NSE variable.

Discriminative accuracy of multiple biomarkers in children with and without traumatic brain injury have been reported with a sensitivity and specificity of 87% and 90%, respectively, using linear, nearest neighbours and tree classification [32]. SVMs have been widely used to identify imaging biomarkers (mostly structural and functional MRI) for diagnosis and prognosis of neurological disorders [33], with 100% accuracy reported in two studies. Another study has analysed survival after OHCA, using six different machine learning algorithms and has reported arrival time, witnessed arrest, bystander CPR, initial Et CO_2 and final Et CO_2 to be predictive (>80% accuracy) [34]. In a study for the prediction of vegetative state coma, classifiers such as support vector machines, nearest neighbours and naive bayes have returned accuracies greater than 80% using multiple variable inputs such as age, sex and the need for a feeding tube [35].

1.3. Missing data

Missing data are a challenge in making clinical predictions. Common practice in statistical analysis is to remove patients with missing data: *complete case* analysis. This can lead to inefficient analysis and selection bias [24]. For example, if obtaining data from patients closer to brain death becomes more difficult, then data is *missing not at random* (MNAR); it is dependent on observed and unobserved variables. Simply removing these patients (complete case analysis) leads to selection bias. If missing data appears to be unrelated to unobserved and observed variables, then the data is *missing completely at random* (MCAR), for example, due to administrative errors. This assumption can be tested, by way of measuring significant differences in other variables for patients with missing and non-missing data. Performing a complete case analysis under MCAR assumes statistical inference to the population. Missing data dependent on observed variables only, is defined as missing at random (MAR). This may be for example, if a patient had a low recording one day then sequential days were not measured again. Under this assumption, the sample can no longer be assumed to be representative of the population and performing a complete case analysis will introduce bias. Various imputation methods are appropriate to reduce this bias. Imputation methods include *conditional mean (CM) imputation*: replacing a missing value with a mean conditioned on other variables, *single imputation* (SI): replacing a missing value with a draw from a predictive distribution, *multiple imputation* (MI): replacing a missing value with a random draw from predictive distribution multiple times [36]. Whilst MCAR can be tested, MAR cannot, and therefore MNAR remains a possibility [24]. This is because there is no way of testing if the data is dependent on unobserved variables. Given the assumption of MAR, it is important to test if the distribution of the imputed data is reflective of the available data. Several approaches exist, however, there is no gold standard to handling missing data [37]. MI methods, including joint modelling (e.g. multivariate normal distribution) and fully conditional specification (e.g. assume univariate distribution) [38], are often preferred since they account for variance in the data. However, there is no gold standard, and the data discussed within this work does not parameterise well and should not be considered as univariate (given the time series nature). Thus the conditional mean KNN imputation[2] is explored.

1.3.1. KNN CM imputation

KNN imputation accounts for the correlation amongst patients by applying mean imputation for patients with a similar average distribution [39]. Similarity between patients is defined by the average difference in value over each variable. For a given patient, *i*, a predetermined number, K,[3] of patients, creates a subset which minimises the similarity measure. The mean of the subset for each variable is then imputed for each missing value of patient *i*. In this study a KNN imputation method was adopted, since this method incorporates the distribution of each class without having access to the class labels (as is required in single/multiple imputation [24]). This allows the imputation of testing data, which is blind to class labels, and provides more clinically relevant results, given that within a clinical setting there may often be missing data.

1.4. Present study objective

This study analyses neuron specific-enolase (NSE) data obtained from a multi-centre observational prospective cohort study [3] in order to propose a novel method to examine its value as a predictor[4] of poor neurological outcome in CA survivors with AIE/DOC. Unlike most NSE studies to date, in an attempt to increase the ecological validity of the findings, the serial NSE data analysed is not limited to the first three days after onset. As explained above, missing data are a challenge and to avoid elimination of patients with missing data, conditional mean KNN imputation was implemented. Imputation allows for all of the available data to be used and therefore the most informed predictive model to be built, subject to eliminating bias. The predictive power of serial NSE measurements is assessed using machine learning classification algorithms and the results are compared to the standard single day and difference-between day methods. To the authors' knowledge, this is the first study to implement classification algorithms on serial NSE measurements to predict the probability of poor neurological outcome.

2. Method

2.1. Selection of participant sample

Patient data in the present study was collected prospectively within the framework of research project "Hypoxia and Outcome Prediction in Early-Stage Coma (HOPE)", an ongoing longitudinal, multi-centre, observational study [3]. After obtaining approval of the ethics committee of the University of Munich, unresponsive patients (Glasgow Coma Scale total score < 9) meeting inclusion and exclusion criteria were selected at seven different intensive care units (ICU) in Southern Germany, 3–14 days after IHCA or OHCA between November 2014 and December 2017. After discharge from the ICU, data collection continued during inpatient neurorehabilitation and ended 1 year after onset. For more information on project HOPE, the reader is referred to a detailed study protocol published elsewhere [3].

2.2. Measurement of serum NSE

Blood sampling for the measurement of NSE was carried out at the discretion of treating physicians. Samples were drawn after admission (day 0) and on subsequent mornings at approximately 6am as part of the daily routine. Therefore, values from day 1 were drawn 6 h–30 h after CA, for day 2, 30 h–54 h after CA and so on. NSE measurements were performed at irregular intervals for each patient, with a median of 3 measurements within the first 18 days. NSE values were not blinded to treating physicians and were used as part of a multimodal prognostic approach.

2.3. Assessment of outcome

Neurological outcome was assessed 1 year after onset with the Clinical Performance Category (CPC). The scale classifies patients into the following categories: dead (CPC 5), vegetative state (CPC 4), severe disability (CPC 3), moderate disability (CPC 2) and good recovery (CPC 1). Furthermore, patients with a CPC value of 1 or 2 can be classified as having a good outcome, whereas patients with CPC 3 or 5 can be classified as having a poor outcome. This dichotomisation has been employed in numerous NSE prognostic studies [13,15,17,18,23]. For patients with a missing CPC value 1 year after onset, the method of last observation carried forward (LOCF) was employed, from month 9, month 6 or month 3 [37].

2.4. Additional patient variables

Demographic and clinical variables were also collected in addition

[2] Principally, KNN method the same as for classification with a different objective: imputation.

[3] Note here that K is the parameter for imputation which is different from k which is a parameter for classification.

[4] A note on terminology: predictor and feature are equivalent, the former used in the context of statistics and the latter in machine learning.

to NSE predictor variables and outcome variables.

2.5. Statistical analysis

The baseline characteristics of patients with missing and non-missing data were tested to determine whether data was MCAR [24]. Baseline characteristic variables also contained missing values. Patients with missing values in baseline characteristic variables were removed from tests under the assumption of MCAR. χ^2 and Mann Whitney univariate tests were used to test the difference in distribution between missing and non-missing patient data for categorical and continuous baseline characteristics. Complete case and conditional mean imputation were applied to outcome and predictor variables respectively. CM KNN imputation was performed using the *fancyimputer.KNN* function in python, a dependency of the *cvxpy* module. The following prediction models were implemented: Gaussian naive Bayes, linear support vector machine, kNN and logistic regression using *scikit.learn* v2.0 in *Python* v3.4. An ensemble method was implemented using all four algorithms. Model parameter optimisation was implemented using *GridSearchCV*. The distance measure used for kNN is the standard euclidean distance which gives an equal weight to relative differences. Other free parameters are K, the number of neighbours to include when performing KNN imputation, and the days to include. After KNN imputation, days which had a significant difference in class distribution to the original data were removed. K was tested iteratively and was selected based on maximising the AUC. Discriminatory accuracy of each model is indicated by the area under the receiver operator curve (AUC). The AUC was calculated using a 5-fold cross validation method. This method splits the good and poor patient predictor data proportionally into 5 equal sized sets. One set is used for validation and the others are used for training. This is repeated for each of the other disjoint sets. The average AUC and standard deviation of all 5 folds is reported. Jouden sensitivity and specificity are reported and sensitivity is reported for specificity >95%. Isotonic and sigmoid calibration were applied to improve the reliability of the probabilistic output as a risk score. The Brier Loss Score, ranging from 0 to 1, is reported as a measure of calibration, with lower values indicating better calibration.

3. Results

3.1. Sample description

The data of a total of 143 patients (mean age at onset = 63 ± 13 years; 107 males, 36 females) were analysed for the present study. As shown in Table 1, a good outcome was observed in 21% of all patients (n = 26), and a poor outcome in 79% of the cases at follow-up. As shown in Table 2, the good and poor outcome subgroups did not differ significantly in their demographic and clinical characteristics.

Table 1
Patient characteristics.

Patient Characteristics	CPC-Based Clinical Outcome		
	Good n = 26	Poor n = 100	No Outcome n = 17
Onset Age in years	61.4 ± 12.8	63.5 ± 13.4	60.7 ± 15.3
Gender Male/Fem	22/4	75/25	10/7
Number of Deaths	0	72	0
MTH No/Yes	4/15 (7)	28/47 (25)	10/3 (4)
IHCA No/Yes	16/7 (3)	67/25 (8)	11/4 (2)
Shockable No/Yes	7/11 (8)	51/30 (19)	11/2 (4)
Rehabilitation in days	64.3 ± 47.5 (15)	76.3 ± 59.1 (63)	(17)
Acute Care in days	28.6 ± 20.1 (12)	26.3 ± 19.1 (55)	26.2 ± 20.9 (12)
CPR Duration in mins	19.5 ± 14.2 (13)	25.6 ± 18.3 (52)	7.4 ± 8.4 (11)

MTH = mild therapeutic hypothermia.
IHCA = In-hospital cardiac arrest.
CPR = Cardiopulmonary resuscitation.
Number of patients with missing variables in parenthesis.

Table 2
p-values from significant difference χ^2 and Mann-Whitney tests for patient characteristics.

Patient Characteristics	p-values	
	Good vs. Poor	Missing vs. Non-missing
Onset Age in years	0.18	0.29
Gender Male/Fem	0.44	0.32
Number of Deaths	NA	NA
MTH No/Yes	0.29	0.10
In-hospital CA No/Yes	0.96	0.83
Shockable No/Yes	0.11	0.13
Rehabilitation in days	0.26	NA
Acute Care in days	0.50	0.36
CPR Duration	0.41	0.07

3.2. Outcome missing data

24% of the patients have a missing outcome at one year. Of which 0.7% are also missing at 9 months, 1.4% at 6 months and 10% at month 3. The LOCF approach produces a similar result as the approach of choosing the best outcome over the entire assessment period as used in other studies [12,13,18] (2 patients change from poor to good). χ^2 and Mann Whitney tests for gender, mild therapeutic hypothermia, cardiac arrest location (IHCA/OHCA), shockable rhythm and acute care duration did not return a significant difference between patients with non-missing and missing outcome (see Table 2). All patients with missing outcome have no recorded date of death and no recorded rehab duration (as is the case with 43% and 62%, respectively of non-missing outcome patients). CPR duration trends towards significance (p = 0.07) with missing data patients having a longer CPR time. Duration of CPR has been shown to be associated with good patient outcome [40]. A complete case analysis was made by removing all patients with missing outcome class to prevent imputation compromising the prediction model [24]. This approach reduces the power as a result of a smaller sample size, however, it avoids bias from imputation, under the assumption of MCAR. Therefore, 12% of patients with no outcome assessment have been excluded from the analysis.

3.3. Predictor missing data

Patient NSE measurements increase in sparsity over 219 days. All days with less than 2 patients having NSE concentration measured were removed. The resulting data includes serial days from 0 to 18. Eighty percent of the serial NSE data was found to be missing. The number of non-missing data entries for each predictor is shown in Table 3.

The Mann Whitney test for the number of missing entries against outcome class did not return a significant difference (p = 0.26) indicating that missing NSE values are not dependent on the patient outcome. The KNN imputed distribution for poor outcome was significantly different from the original data for only day 17 and 18 (p = 0.03, 0.04, respectively), see Fig. 1 (for K = 10). Good outcome had no recorded observation for day 11,12,13,16 (Table 3) and otherwise had no significant difference in distribution for predictors. This is true for all K. Imputed predictor data from day 1 to day 16 were included, with K a free parameter of the predictive model.

3.4. Predictive modelling

Each prediction model had input data imputed using K, from K = 1 to K = 26 (size of good outcome class) number of neighbours, and the resulting AUC was recorded. K = 9 neighbours maximised the AUC for

Table 3
Number of non-missing data for predictors and percentage difference in NSE distribution for each outcome class.

Predictor	Number of Non-Missing Values		% difference between
	Poor Outcome	Good Outcome	Good and Poor outcome
NSE Day 0	25	5	8.6
NSE Day 1	64	14	18.8
NSE Day 2	65	16	66.9
NSE Day 3	61	18	79.3
NSE Day 4	45	12	72.9
NSE Day 5	30	4	66.2
NSE Day 6	18	5	71.0
NSE Day 7	15	5	39.6
NSE Day 8	8	1	20.7
NSE Day 9	11	5	35.0
NSE Day 10	12	3	30.9
NSE Day 11	9	0	–
NSE Day 12	4	0	–
NSE Day 13	4	0	–
NSE Day 14	5	1	12.5
NSE Day 15	5	1	11.9
NSE Day 16	3	0	–
NSE Day 17	3	1	28.7
NSE Day 18	1	1	45.5

all algorithms (all being above AUC= 0.7) (see Fig. 2).

The naive Bayes classifier had the best performance (AUC 0.81 ± 0.07), followed by a nearest neighbours classifier (k = 12, AUC 0.79 ± 0.12), a linear SVM (AUC 0.74 ± 0.07) and logistic regression (AUC 0.73 ± 0.08) (see Table 4). The AUC for nearest neighbours ranged from 0.62 to 0.73 for $k = 1$ to $k = 16$ with a global maximum of 0.79 for $k = 12$ and a local maximum of 0.77 at $k = 6$. An ensemble method using all four algorithms, with double weight for kNN and naive Bayes, returned an AUC of 0.83 ± 0.10.

For comparison, results for discriminatory accuracy of day 0 to day 6 absolute and differences in NSE values were calculated (see Table 5) for all patients with non-missing values. Day 3 had the greatest AUC for a single day (0.75) and the greatest AUC for differences between days occurs from day 5 to day 6 (0.91) and day 1 to day 4 (0.81). Each AUC metric analysed a subset of $n = 79$, $n = 46$ and $n = 13$ patients, respectively.

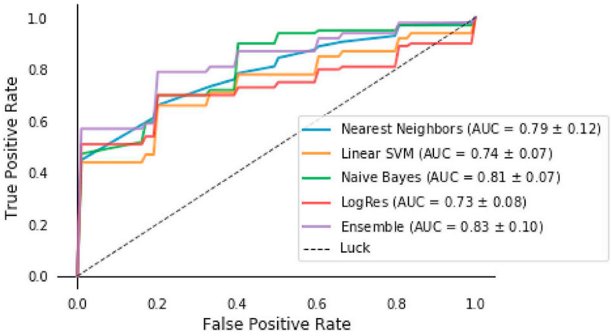

Fig. 2. ROC for each prediction model.

The predicted probability output from the naive Bayes classifier versus the NSE day 3 serum concentration has been plotted (for those patients who had recordings) along with the corresponding 95% specificity thresholds, as shown in Fig. 3. Both thresholds successfully avoid falsely pessimistic predictions.

All uncalibrated models had a sensitivity >0.4 for a specificity >0.95. The associated thresholds for each of these values are uncalibrated probability estimates. The calibration plots in Fig. 4 show how well each model is calibrated.

A well calibrated model should predict probabilities close to the true probability observed in the data. While naive Bayes had the highest AUC it was also least calibrated with the highest brier loss score of 0.39. Isotonic and sigmoid calibration improved the calibration score to 0.15 and 0.14, respectively. The corresponding AUCs for each naive Bayes calibration are (0.77 ± 0.07) and (0.79 ± 0.07) (see Figs. 5 and 6).

Calibration improved the brier loss score by a maximum of 0.04 for logistic regression (from 0.19) and 0.01 for linear SVM (from 0.16) and

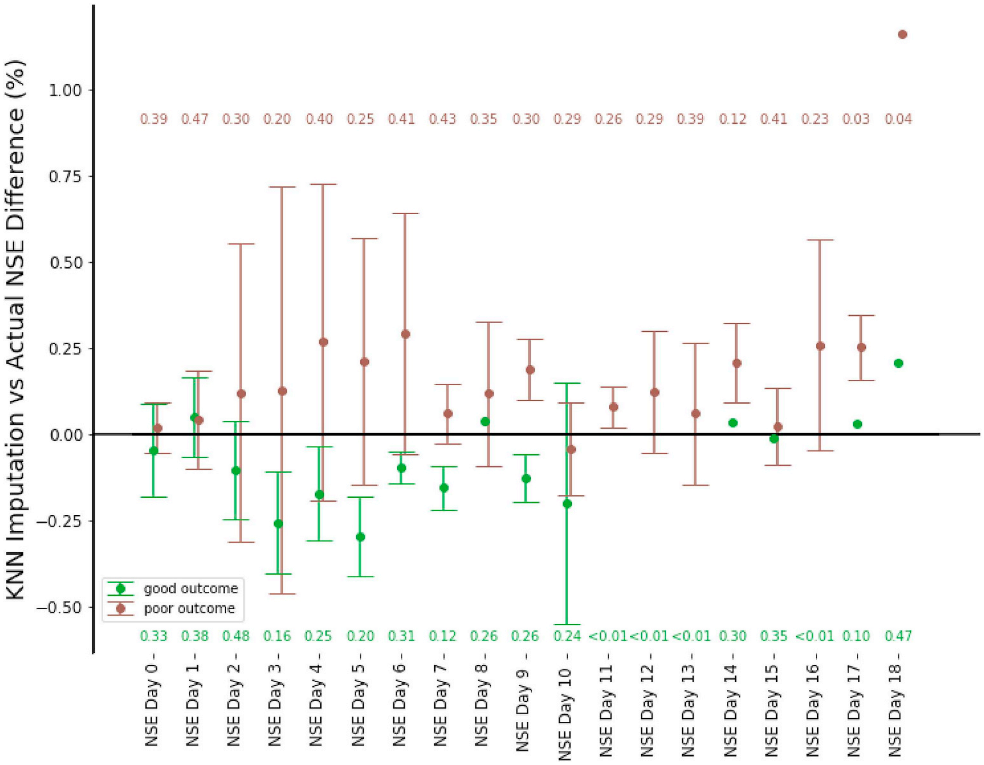

Fig. 1. 95% confidence intervals for difference in distribution of KNN imputed variables and observed data.

Table 4
Performance measures for predictive models. Area under the receiver operator curve (AUC). Youden index is *sensitivity* + *specificity* − 1. >95% specificity for poor outcome.

Model	AUC	Max Youden		>95% Specificity	
		Sensitivity	Specificity	Sensitivity	Specificity
Nearest Neighbours	0.79 ± 0.12	0.57	0.88	0.45	0.96
Linear SVM	0.74 ± 0.07	0.51	0.92	0.45	0.96
Naive Bayes	0.81 ± 0.07	0.56	0.92	0.49	0.96
Isotonic	0.77 ± 0.07	0.54	0.92	0.00	1.00
Sigmoid	0.79 ± 0.07	0.57	0.88	0.40	1.00
LogRes	0.73 ± 0.08	0.56	0.96	0.56	0.96
Ensemble	0.83 ± 0.10	0.73	0.88	0.42	0.96

Table 5
AUCs for absolute NSE values and difference between days for Day 0 to Day 6.

	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Day 0	.54						
Day 1	.60	.50					
Day 2	.60	.71	.70				
Day 3	.74	.72	.65	.75			
Day 4	.36	.81	.60	.60	.73		
Day 5	.71	.66	.57	.61	.65	.65	
Day 6	.00	.71	.50	.58	.74	.91	.74

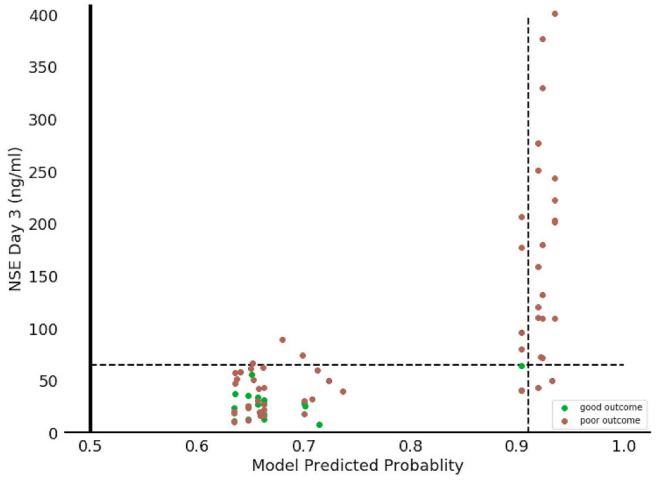

Fig. 3. Predicted probability from naive Bayes + Sigmoid calibration model versus NSE Day 3 concentration values. The dotted lines are thresholds for >95% specificity. There are 4 patients with NSE values exceeding $400 ng/ml$ which have not been included in the plot.

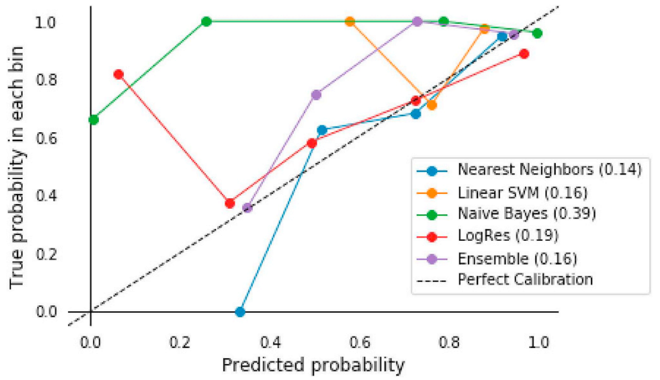

Fig. 4. Calibration Plots for each predictive model. Brier Loss score reported in brackets.

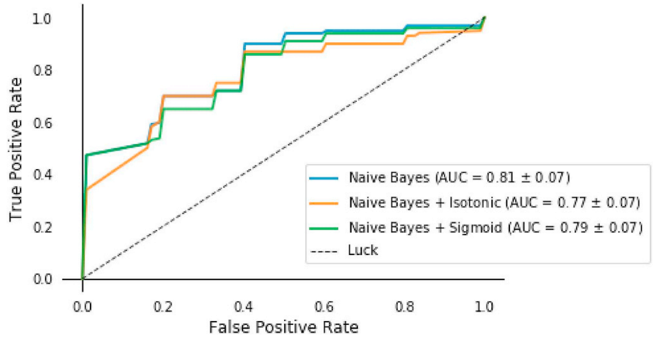

Fig. 5. ROC for Naive Bayes classifier with isotonic and sigmoidal calibration.

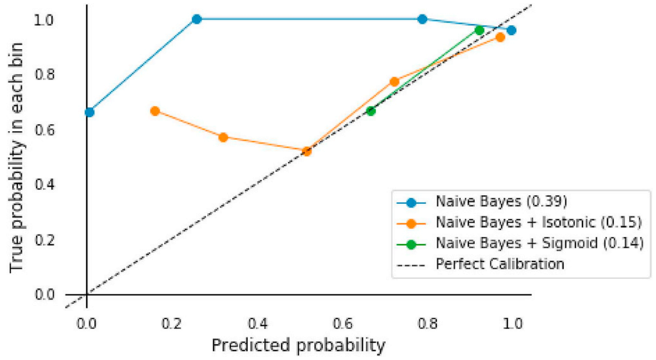

Fig. 6. Calibration Plots for Naive Bayes classifier with isotonic and sigmoidal calibration. Brier Loss score reported in brackets.

did not change the brier loss score for nearest neighbours (0.14). For the naive Bayes classifier with sigmoid calibration, a cut-off probability of 90% had a specificity of 96% and sensitivity of 49%. This is comparable to NSE Day 3 as a predictor.

4. Discussion

This work compared a number of different machine learning methods to classify patients based on KNN-imputed serial NSE measurements and further compared these to single day and relative days AUCs. The naive Bayes classifier learned the best model for discriminating between good and poor patient outcomes. It outperformed logistic regression, the statistical gold standard for modelling categorical outcomes.

A previous study, based on a different set of features, also found that naive Bayes outperformed other classification algorithms [35]. A Gaussian naive Bayes classifier assumes each predictor is independent and parameterised by a Gaussian distribution with class dependent mean and variance. The optimal performance of a naive Bayes classifier, despite highly correlated features, has been proposed to arise from dependencies cancelling each other out amongst features [41]. In contrast, naive Bayes has been shown to consistently perform worse than SVMs, logistic regression and kNNs over a set of 11 binary classification problems. In this work, the test sets were between 2 and 7 times larger than the training set, suggesting the optimality of naive Bayes is reduced for larger external validation sets. This raises the important concern of model over-fitting. Model over-fitting has been addressed where possible; by implementing 5-fold cross-validation and optimising *k* neighbours in kNN classification. However, the ultimate test will be on larger external patient datasets.

The same study has shown SVMs to outperform logistic regression, kNNs and naive Bayes. The superior performance of SVMs are a direct result of reduced sensitivity to outliers and better generalisability [30]. In this work, the proportion of training data in our work is greater than the previous study, and therefore, the SVMs have not enough testing data to exhibit the superior generalisability. In addition, logistic regression and linear SVMs returned a similar accuracy, due to both algorithmic models including a linear combination of serial NSE measurements (albeit with different optimisation constraints, see Section 1.2). Since the predictors are highly correlated, this linear function does not take into account the interaction of different days and performs worse than naive Bayes and kNN. The good performance of kNN is not unexpected, since KNN imputation retains class distribution. Whilst nearest neighbours performs close to naive Bayes, the latter is preferable since there are no parameters that require tuning.

Aside from these four algorithms, an ensemble method was implemented. The ensemble method achieved a greater AUC than all algorithms, leveraging the predictive power of each at the cost of computation. Artificial neural networks and decision trees were attempted, however, failed to converge to stable prediction parameters. In addition, kNN was implemented using both brute force algorithm and kd tree based indexing. The former compares the test patient to each and every other patient in the dataset and the latter recursively partitions the patient set in each dimension using the median and then compares the test patient to a subset of patients. Kd tree based indexing returned an AUC 3% lower than the brute force method and has not been included in the results. The computational efficiency of kd tree based indexing should be considered for larger patient datasets (see Table 6).

Compared to the optimal machine learning classifier using serial NSE measurements, the optimal single day discriminatory accuracy is 6% lower and the optimal accuracy for relative days matches the performance of naive Bayes. There is an AUC between day 5 and day 6 which exceeds 0.81, however, there are only 13 patients, of which 2 are good outcome (see Table 5). These two patients have the second and third lowest NSE differences and hence the high (but not 100%) accuracy. On such a small patient sample it is difficult to determine if this is a feature which can be expected for good outcome patients, and therefore this AUC is treated as an outlier. For an AUC of 0.81 between day 1 and day 4 there are 7 good outcome patients and 39 poor outcome patients. All 7 good outcome patients have a decrease between day 1 and day 4, whilst 15 out of 39 patients have a decrease from the poor outcome class. The naive Bayes classifier has been able to automatically detect the most relevant feature for classification without having to search exhaustively for each combination of possible differences. In addition, since it makes use of the imputed dataset, the result is valid for the entire patient dataset. In practice, this would allow predictions to be made for patients with sparse NSE recordings, since sparse serial measurements can be imputed.

Table 6
Computational complexity of each algorithm with n samples and m features [43].

Algorithm	Training	Classification
Logistic Regression	$O(mn^2 + n^3)$	$O(m)$
SVM	$O(m^2n)$	$O(m)$
Naive Bayes	$O(mn + nc)$	$O(m)$
kNN	$O(1)$	$O(n)$ - brute force $O(log(n))$ - kd tree

In contrast to single day and relative days cut-off values, the machine learning methods provide a probabilistic risk score. The naive Bayes model outputs a risk score close to the true population probability after sigmoid calibration. A perfectly calibrated model would, for each probability threshold, report the same proportion of incidences in the training set. This can be interpreted as follows. For any new patient with an output risk score of 90%, there is a 90% chance of this patient having a poor outcome, given all patient data. Machine learning models will require further calibration on large and external datasets in order to be practically useful [24]. A fully validated and calibrated model will propose a score that can support a clinician in the assessment of the patient. There will always be a possibility of outliers. The naive Bayes + sigmoid calibrated model produces one good outcome patient who has a risk score close to 90%, the threshold for poor outcome based on 95% specificity. This patient has the greatest increase from day 1 to day 2 of all the good outcome patients, and has the highest NSE value for day 2 and day 3 of the good outcome class. An increase on day 1 to day 2, and higher values within the first 3 days are consistent with the poor outcome class and hence this patient has come close to misclassification. A potential explanation for falsely high NSE-levels could be due to hemolysis [42]. In practice, such a case would require a higher level of scrutiny, taking into account additional biological and clinical assessments.

This study has not controlled for confounding variables due to the extent of missing data (see Table 1). Previous work has suggested that differences in NSE concentrations in the blood can arise from discrepancies in blood measuring methods [20] and heterogeneity of patient samples [22].

It is clear that the size of the dataset is a limitation in this work. In part, since confounding variables were not included, and in part because there is not enough power in the calibrated probability statistics. In the future, not only larger internal datasets are required but also external datasets to validate the models [24]. Computational complexity of each algorithm should be considered when scaling to large patient datasets. For example, the classification cost of naive Bayes is lower than kNN, however, training is more expensive (see Table 6).

KNN imputation has been implemented as a means of filling the missing data without knowledge of the class labels (unlike MI methods). This allows new patient data, without labels assigned, to be imputed and classified. Whilst it maintains the class distributions (see Fig. 1), unlike MI imputation, it does not account for variability. Further research into non-parametric MI methods without class labels can be explored.

A machine learning approach in clinical prediction modelling has benefits and drawbacks. Firstly, it is fast and highly automated; it can provide a risk score in a matter of moments and continuously update its own model for new data it receives. However, it also requires statistical expertise in the clinical setting which may not be readily available.

5. Conclusion

Compared to the conventional single day cut-off based method, relative days method and logistic regression, a naive Bayes classifier predicted as good as or more accurately good and poor clinical outcome in patients with AIE/DOC after CA. Classification for a greater number of patients was possible after imputation of serial NSE measurements. Models applied to serial measurements provide risk score estimates, unlike the single and relative days cut-off approach. These results suggests that predictive models for prognosis with NSE could provide an additional tool for accurate outcome prediction for this challenging clinical population. Further research should focus on providing external validation with larger datasets as well as on confounding variables and additional objective biological parameters.

Acknowledgements

The present study was partially supported by grant 2014-A55 from the Else Kröne-Fresenius Stiftung that provided funding for Project HOPE.

References

[1] J. Berdowski, R.A. Berg, J.G.P. Tijssen, R.W. Koster, Global incidences of out-of-hospital cardiac arrest and survival rates: systematic review of 67 prospective studies, Resuscitation 81 (11) (2010) 1479–1487.

[2] G. Jürgen, C. Mühlhoff, G.S. Doig, S. Reinartz, K. Bode, R. Dujardin, K.C. Koch, E. Roeb, U. Janssens, Health care costs, long-term survival, and quality of life following intensive care unit admission after cardiac arrest, Crit. Care 12 (4) (2008) R92.

[3] A. Lopez-Rolon, A. Bender, Hypoxia and outcome prediction in early-stage coma (project hope): an observational prospective cohort study, BMC Neurol. 15 (82) (05 2015).

[4] T.J. Quinn, J. Dawson, M.R. Walters, K.R. Lees, Reliability of the modified rankin scale: a systematic review, Stroke 40 (10) (2009) 3393–3395.

[5] J. Giacino, J. Fins, S. Laureys, N.D. Schiff, Disorders of consciousness after acquired brain injury: the state of the science, Nat. Rev. Neurol. 10 (2) (2014) 99–114.

[6] C.H. Hsu, J. Li, M.J. Cinousis, K.R. Sheak, D.F. Gaieski, B.S. Abella, M. Leary, Cerebral performance category at hospital discharge predicts long-term survival of cardiac arrest survivors receiving targeted temperature management, Crit. Care Med. 42 (12) (2014) 2575–2581.

[7] E.F.M. Wijdicks, A. Hijdra, G.B. Young, C.L. Bassetti, S. Wiebe, Practice parameter: prediction of outcome in comatose survivors after cardiopulmonary resuscitation (an evidence-based review), Neurology 67 (2) (2006) 203–210.

[8] K. Howell, E. Grill, A.M. Klein, A. Straube, A. Bender, Rehabilitation outcome of anoxic-ischaemic encephalopathy survivors with prolonged disorders of consciousness, Resuscitation 84 (10) (2013) 1409–1415.

[9] W. Fogel, D. Krieger, M. Veith, H.P. Adams, E. Hund, B. Storch-Hagenlocher, F. Buggle, D. Mathias, W. Hacke, Serum neuron-specific enolase as early predictor of outcome after cardiac arrest, Crit. Care Med. 25 (7) (1997) 1133–1138.

[10] F.C. Barone, R.K.L. Clark, W.J. Price, R.F. White, G.Z. Feuerstein, B.L. Storer, E.H. Ohlstein, Neuron-specific enolase increases in cerebral and systemic circulation following focal ischemia, Brain Res. 623 (1) (1993) 77–82.

[11] C. Daubin, C. Quentin, S. Allouche, O. Etard, C. Gaillard, A. Seguin, X. Valette, J.J. Parienti, F. Prevost, M. Ramakers, N. Terzi, P. Charbonneau, D. du Cheyron, Serum neuron-specific enolase as predictor of outcome in comatose cardiac-arrest survivors: a prospective cohort study, BMC Cardiovasc. Disord. 11 (1) (2011) 48.

[12] H.N. Rosen, K.S. Sunnerhagen, J. Herlitz, C. Blomstrand, L. Rosengren, Serum levels of the brain-derived proteins s-100 and nse predict long-term outcome after cardiac arrest, Resuscitation 49 (2) (2001) 183–191.

[13] W. Schoerkhuber, H. Kittler, F. Sterz, W. Behringer, M. Holzer, M. Frossard, S. Spitzauer, A.N. Laggner, Time course of serum neuron-specific enolase, Stroke 30 (8) (1999) 1598–1603.

[14] P. Martens, A. Raabe, P. Johnsson, Serum s-100 and neuron-specific enolase for prediction of regaining consciousness after global cerebral ischemia, Stroke 29 (11) (1998) 2363–2366.

[15] K.J. Streitberger, C. Leithner, M. Wattenberg, P.H. Tonner, J. Hasslacher, M. Joannidis, T. Pellis, E. Di Luca, M.J. Foedisch, A. Krannich, C.J. Ploner, C. Storm, Neuron-specific enolase predicts poor outcome after cardiac arrest and targeted temperature management: a multicenter study on 1,053 patients, Crit. Care Med. 45 (7) (2017) 1145–1151.

[16] M. Rundgren, T. Karlsson, N. Nielsen, T. Cronberg, P. Johnsson, H. Friberg, Neuron specific enolase and s-100b as predictors of outcome after cardiac arrest and induced hypothermia, Resuscitation 80 (7) (2009) 784–789.

[17] I.G. Steffen, D. Hasper, C.J. Ploner, J.C. Schefold, E. Dietz, F. Martens, J. Nee, A. Krueger, A. Jörres, C. Storm, Mild therapeutic hypothermia alters neuron specific enolase as an outcome predictor after resuscitation: 97 prospective hypothermia patients compared to 133 historical non-hypothermia patients, Crit. Care 14 (2) (2010) R69.

[18] M. Tiainen, R.O. Roine, V. Pettilä, O. Takkunen, Serum neuron-specific enolase and s-100b protein in cardiac arrest patients treated with hypothermia, Stroke 34 (12) (2003) 2881–2886.

[19] M. Rundgren, T. Cronberg, H. Friberg, A. Isaksson, Serum neuron specific enolase–impact of storage and measuring method, BMC Res. Notes 7 (1) (2014) 726.

[20] M. Mlynash, M.S. Buckwalter, A. Okada, A.F. Caulfield, C. Venkatasubramanian, I. Eyngorn, M.M. Verbeek, C.A.C. Wijman, Serum neuron-specific enolase levels from the same patients differ between laboratories: assessment of a prospective post-cardiac arrest cohort, Neurocritical Care 19 (2) (2013) 161–166.

[21] P. Stammet, O. Collignon, C. Hassager, M.P. Wise, J. Hovdenes, A. Åneman, J. Horn, Y. Devaux, D. Erlinge, J. Kjaergaard, Y. Gasche, M. Wanscher, T. Cronberg, H. Friberg, J. Wetterslev, T. Pellis, M. Kuiper, G. Gilson, N. Nielsen, Neuron-specific enolase as a predictor of death or poor neurological outcome after out-of-hospital cardiac arrest and targeted temperature management at 33°c and 36°c, J. Am. Coll. Cardiol. 65 (19) (2015) 2104–2114.

[22] P. Stammet, Blood biomarkers of hypoxic-ischemic brain injury after cardiac arrest, Semin. Neurol. 37 (1) (2017) 75–80.

[23] C. Storm, J. Nee, A. Jörres, C. Leithner, D. Hasper, C.J. Ploner, Serial Measurement of Neuron Specific Enolase Improves Prognostication in Cardiac Arrest Patients Treated With Hypothermia: A Prospective Study vol. 20, (2012), p. 6.

[24] E.W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Statistics for Biology and Health, Springer, New York, 2008.

[25] E.W. Steyerberg, N. Mushkudiani, P. Perel, I. Butcher, J. Lu, G.S. McHugh, G.D. Murray, A. Marmarou, I. Roberts, J.D.F. Habbema, A.I.R. Maas, Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics, PLoS Med. 5 (8) (2008) 1–11.

[26] Y.H. Lee, H. Bang, D.J. Kim, How to establish clinical prediction models, Endocrinol. Metabol. 31 (1) (2016) 38–44.

[27] D.G. Altman, Y. Vergouwe, K.G.M. Royston Pand Moons, Prognosis and prognostic research: validating a prognostic model, BMJ 338 (2009) b605.

[28] R.B. D'Agostino, S. Grundy, L.M. Sullivan, P. Wilson, CHD Risk Prediction Group, Validation of the framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation, J. Am. Med. Assoc. 286 (2) (2001) 180–187.

[29] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[30] Pochet NL and Suykens JA. Support vector machines versus logistic regression: improving prospective performance in clinical decision-making. Ultrasound Obstet. Gynecol., 27(6):607–608.

[31] J. Reisinger, K. Höllinger, W. Lang, C. Steiner, T. Winter, E. Zeindlhofer, M. Mori, A. Schiller, A. Lindorfer, K. Wiesinge, P. Siostrzonek, Prediction of neurological outcome after cardiopulmonary resuscitation by serial determination of serum neuron-specific enolase, Eur. Heart J. 28 (1) (2007) 52–58.

[32] R.P. Berger, S. Ta'asan, A. Rand, A. Lokshin, P. Kochanek, Multiplex assessment of serum biomarker concentrations in well-appearing children with inflicted traumatic brain injury, Pediatr. Res. 65 (1) (2009) 97–102.

[33] G. Orrú, W. Pettersson-Yeo, A.F. Marquand, G. Sartori, A. Mechelli, Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review, Neurosci. Biobehav. Rev. 36 (4) (2012) 1140–1152.

[34] M. Krizmaric, M. Verlic, G. Stiglic, S. Grmec, P. Kokol, Intelligent analysis in predicting outcome of out-of-hospital cardiac arrest, Comput. Methods Progr. Biomed. 95 (2, Suppl) (2009) S22–S32.

[35] L. Pignolo, V. Lagan, Prediction of outcome in the vegetative state by machine learning algorithms: a model for clinicians? J. Software Eng. Appl. 4 (6) (2011) 388–390.

[36] D.B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.

[37] A.M. Wood, I.R. White, S.G. Thompson, Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals, Clin. Trials 1 (4) (2004) 368–376.

[38] J.S. Murray, Multiple imputation: a review of practical and theoretical findings, Stat. Sci. 33 (2) (2018) 142–159.

[39] L. Beretta, A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation, BMC Med. Inf. Decis. Mak. 16 (3) (2016) 74.

[40] A. Temple, R. Porter, Predicting neurological outcome and survival after cardiac arrest, Cont. Educ. Anaesth. Crit. Care Pain 12 (6) (2012) 283–287.

[41] Harry Zhang, The optimality of naive bayes, A A 1 (2) (2004) 3.

[42] L. Ramont, H. Thoannes, A. Volondat, F. Chastang, M.C. Millet, F.X. Maquart, Effects of hemolysis and storage condition on neuron-specific enolase (nse) in cerebrospinal fluid and serum: implications in clinical practice, Clin. Chem. Lab. Med. 43 (11) (2005) 1215–1217.

[43] C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, K. Olukotun, Map-reduce for machine learning on multicore, Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06, MIT Press, Cambridge, MA, USA, 2006, pp. 281–288.