Explainable Automated Non-linear Computation scoring system for Health (EACH) score : a Machine Learning based Explainable Automated Nonlinear Computation scoring system for Health and an application for prediction of perioperative stroke

Mi-Young Oh, Hee-Soo Kim, Young Mi Jung, Hyung-Chul Lee, Seung-Bo Lee, Seung Mi Lee

Submitted to: Journal of Medical Internet Research
on: March 03, 2024

Table of Contents

Explainable Automated Non-linear Computation scoring system for Health (EACH) score
: a Machine Learning based Explainable Automated Nonlinear Computation scoring system for Health and an application for prediction of perioperative stroke

Mi-Young Oh[1*]; Hee-Soo Kim[2*]; Young Mi Jung[3]; Hyung-Chul Lee[4]; Seung-Bo Lee[2*]; Seung Mi Lee[4*]

[1]Bucheon Sejong Hospital Bucheon-si, Gyeonggi-do KR
[2]Keimyung University School of Medicine Daegu KR
[3]Korea University Guro Hospital Seoul KR
[4]Seoul National University College of Medicine, Seoul National University Hospital Seoul KR
[*]these authors contributed equally

Corresponding Author:
Seung Mi Lee
Seoul National University College of Medicine, Seoul National University Hospital
101 Daehak-ro, Jongno-gu
Seoul
KR

Abstract

Background: Machine learning (ML) has the potential to enhance performance by capturing nonlinear interactions. However, ML-based models have some limitations in terms of interpretability.

Objective: To address this, we developed and validated a more comprehensible and efficient ML-based scoring system using SHapley Additive exPlanations (SHAP) values.

Methods: We developed and validated the Explainable Automated nonlinear Computation for Health (EACH) framework score. We developed CatBoost based prediction model, identified key features, and automatically detected the top five steepest slope change points based on SHAP plots. Subsequently, we developed a scoring system (EACH) and normalized the score. Finally, the EACH score was used to predict perioperative stroke.

Results: When applied for perioperative stroke prediction among 44,901 patients undergoing noncardiac surgery, the EACH score achieved an area under the curve (AUC) of 0.829 [95% CI, 0.753-0.892]. In the external validation, the EACH score demonstrated superior predictive performance with an AUC of 0.784 [95% CI, 0.694-0.871] compared to a traditional score (AUC of 0.528 [95% CI, 0.457-0.619]) and another ML-based scoring generator (AUC of 0.784 [95% CI, 0.694-0.871]).

Conclusions: The EACH score is a more precise, explainable ML-based risk tool, proven effective in real-world data, outperforming traditional scoring system.

(JMIR Preprints 03/03/2024:58021)
DOI: https://doi.org/10.2196/preprints.58021

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original paper

Title:

Explainable Automated Non-linear Computation scoring system for Health (EACH) score

: a Machine Learning based Explainable Automated Nonlinear Computation scoring system for Health and an application for prediction of perioperative stroke

Keywords: machine learning, explainability, score

Author's information

Authors:

Mi-Young Oh, MD, Ph.D[a], Hee-Soo Kim, MS[b], Young Mi Jung, MD[c,d], Hyung-Chul Lee, MD, Ph.D[e], Seung-Bo Lee, Ph.D[b], Seung Mi Lee MD, Ph.D [c,e,f,g,h]

Affiliations:

[a]Department of Neurology, Bucheon Sejong Hospital, Bucheon-si, Gyeonggi-do, Korea

[b] Department of Medical Informatics, Keimyung University School of Medicine, Daegu, Korea

[c]Department of Obstetrics and Gynecology, Korea University Guro Hospital, College of Medicine, Korea University, Seoul, Korea

[d]Department of Anesthesiology and Pain Medicine, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Korea [d]

[e]Department of Obstetrics and Gynecology, Seoul National University College of Medicine, [f]Seoul National University Hospital, and [g]Innovative Medical Technology Research Institute, Seoul National University Hospital, [h]Institute of Reproductive Medicine and Population, Medical Research Center, Seoul National University, Seoul, Korea

MY Oh and HS Kim contributed equally as the first authors of this study.

SB Lee and SM Lee contributed equally as co-corresponding authors.

Author's contribution:

Mi-Young Oh, Hee-Soo Kim, Seung-Bo Lee, and Seung Mi Lee was contributed to

conceptualisation, data curation, formal analysis, methodology and investigation of study.

Seung Mi Lee was also contributed to funding acquisition of study.

Young Mi Jung, Hyung-Chul Lee was contributed to data collection and curation.

Correspondence to:

Seung Mi Lee, MD, PhD

[e]Department of Obstetrics and Gynecology, Seoul National University College of Medicine, [f]Seoul National University Hospital, and [g]Innovative Medical Technology Research Institute, Seoul National University Hospital, [h]Institute of Reproductive Medicine and Population, Medical Research Center, Seoul National University, Seoul, Korea

101 Daehak-ro, Jongno-gu, Seoul 03080, Korea

Tel: 82-2-2072-4857

Fax: 82-2-762-3599

E-mail: lbsm@snu.ac.kr

Corresponding Author's Email: lbsm@snu.ac.kr

Abstract

Background: Machine learning (ML) has the potential to enhance performance by capturing nonlinear interactions. However, ML-based models have some limitations in terms of interpretability.

Objective: To address this, we developed and validated a more comprehensible and efficient ML-based scoring system using SHapley Additive exPlanations (SHAP) values.

Methods: We developed and validated the Explainable Automated nonlinear Computation for Health (EACH) framework score. We developed CatBoost based prediction model, identified key features, and automatically detected the top five steepest slope change points based on SHAP plots. Subsequently, we developed a scoring system (EACH) and normalized the score. Finally, the EACH score was used to predict perioperative stroke.

Results: When applied for perioperative stroke prediction among 44,901 patients undergoing noncardiac surgery, the EACH score achieved an area under the curve (AUC) of 0.829 [95% CI, 0.753-0.892]. In the external validation, the EACH score demonstrated superior predictive performance with an AUC of 0.784 [95% CI, 0.694-0.871] compared to a traditional score (AUC of 0.528 [95% CI, 0.457-0.619]) and another ML-based scoring generator (AUC of 0.784 [95% CI, 0.694-0.871]).

Conclusion: The EACH score is a more precise, explainable ML-based risk tool, proven effective in real-world data, outperforming traditional scoring system.

Introduction

The risk scoring systems have been proposed to prognosticate critical medical conditions, aiming to identify high-risk patients who are likely to experience adverse outcomes.(1-6) Traditionally, risk-scoring systems have been developed using conventional statistical approaches, based on the assumption of linearity between variables and outcomes.(7, 8) This method has provided a comprehensive understanding of patient risk profiles, but it may not fully capture complex, non-linear interactions, potentially leading to less accurate risk assessments. Moreover, the clinical variables incorporated into these systems were typically ascertained through univariate analysis, enriched by insights from expert opinions, or selected from a range of risk factors established in previous literature.(8) Consequently, these systems faced limitations in rapidly integrating state-of-the-art medical knowledge alongside medical advancements.(9-11) With the rapid expansion and increasing diversity of medical data, these limitations have become more pronounced.(10)

To address these limitations, machine learning (ML) techniques have emerged as promising avenues for creating new and diverse risk models by leveraging extensive electronic medical records.(12, 13) Despite their exceptional predictive performance, the lack of interpretability has limited their adoption in real-world medical practice.(14) Recent efforts in ML have addressed the 'black box' issue of existing models by presenting data in a more understandable fashion.(15) For instance, an ML-based automated scoring model was developed by integrating the Random Forest algorithm and logistic regression methods.(16) This model was designed to be easy to understand and apply to various clinical situations, such as in-hospital mortality and out-of-hospital cardiac arrest.(16, 17) However, these approaches still rely on scoring methods rooted in traditional statistics, using variables selected by the ML models.(16) Furthermore, the generalizability and scalability of ML

models are hampered by limited external validation. Therefore, there is a compelling need to develop and validate a fully automated ML-based scoring system that can address nonlinearity assumptions.

In response to these challenges, this research aimed to develop and validate an Explainable Automated nonlinear Computation scoring system for Health (EACH) score. This system addresses nonlinearity assumptions and enhances explainability. Additionally, we applied the EACH score to predict perioperative stroke to assess its performance in real-world clinical practice and examined its performance compared with traditional scores and other ML-based scoring systems.

Methods

We developed and validated the Explainable Automated nonlinear Computation scoring system for Health (EACH), designed to automate the development of clinical scoring models for predefined health outcomes. The development and application of EACH were segmented into several steps, as illustrated in Figure 1.

Step 1: Model training and hyperparameter optimization

When implementing the EACH score, the initial step involved training the CatBoost model with hyperparameter optimization for each set of clinical data, followed by determination of important features using SHapley Additive exPlanations (SHAP) values(18)

Step 2: The detection of slope change point in plots of SHAP values and selection of the top 5 steepest slope change points

In this step, we harnessed the power of automation to identify the critical features that influence the accuracy of the prediction model. The process began with the generation of SHAP value plots for each feature, which visually represented the impact of these features on the model predictions. The key advantage of our approach is the automated detection of the slope change points within these plots. The slope change points are significant shifts in the importance of each variable, indicating where the influence of the variable undergoes a notable change.

The algorithm identified the five most pronounced slope change points for each feature. These critical points indicate where the model's sensitivity to feature values is dramatically altered, thus playing a pivotal role in subsequent feature scoring. Figure 2 illustrates this method by contrasting linear and nonlinear cases, where the latter demonstrates the effectiveness of the method in recognizing complex nonlinear relationships often present in clinical data.

By isolating these key intervals, we delineated the feature value ranges that were most

influential on the model's predictions, thus allowing for a more accurate SHAP-based scoring system

Step 3: SHAP-based scoring system development

The third step shows how to generate scores using the SHAP values for each variable. The score was calculated by adding the SHAP values of the data points obtained from each section to accurately quantify the impact of individual intervals for each feature. This was for the model's sensitive interpretation of the different ranges within the variable and was reflected in the final scoring system.

Step 4: Data transformation and handling missing data

Normalization of scores across features

In this step, we focused on standardizing the scoring system across all features. This normalization process ensured that the scores from different features were comparable and appropriately weighted within the overall scoring model. Normalization is based on the total range of scores across all features, thereby aligning them on a unified scale.

Handling missing data with interval averaging

If the actual values for a particular interval were unavailable, our model adopted a fallback strategy that utilizes the average of values in adjacent intervals. If there is no data in the interval immediately next to the missing value, the entire range of the interval was used to preserve the integrity of the model's risk assessment in the face of data sparsity.

Clinical study design

In applying the EACH score to assess stroke risk during the perioperative period in noncardiac surgeries, we approached our clinical study design with a focus on real-world applicability. This study was structured using three datasets: training, internal validation, and external validation. The training and internal validation datasets were derived from patient

records at Seoul National University Hospital (SNUH) from 2016 to 2019. For a broader perspective and to assess the performance of the model in different settings, we included a geotemporal external validation using data from surgeries performed at the Boramae Medical Center (BMC) between 2020 and 2021.

A key aspect of our data preparation involved addressing missing values in preoperative variables. Recognizing the potential impact of incomplete data on the model's accuracy, we employed a methodological imputation approach. For continuous variables, missing values were imputed using their mean, whereas categorical variables were imputed using the mode. (19).

Patients were excluded from the study if the surgery lasted less than 20 minutes, if they were younger than 18 years or weighed less than 30 kg or more than 140 kg, if their height was outside the 135 cm to 200 cm range, or if they had a prior history of stroke. The developmental set consisted of 36,502 patients from Seoul National University Hospital. After applying the exclusion criteria, 404 patients from Boramae Medical Center were included in the external validation set (Figure 3).

Perioperative stroke was defined as an ischemic brain infarction occurring within 30 days postoperatively and was identified through new ischemic lesions on diffusion-weighted imaging.(20, 21) An experienced neurologist (M-Y Oh) confirmed the stroke diagnosis by reviewing the imaging outcomes.

Data collection

1) Demographics and comorbidities: We recorded the age, sex, physical metrics (height, weight, and body mass index), and a range of preexisting conditions, namely hypertension, diabetes with or without insulin medication, previous cardiovascular events, asthma, chronic obstructive pulmonary disease, liver and kidney diseases, and

tuberculosis.

2) Surgical information: Surgical risk was classified according to the American Society of Anesthesiologists (ASA) standards. Information on whether the surgery was emergent and the type was collected.
3) Preoperative laboratory findings: A comprehensive set of laboratory results were collected, including hemoglobin levels, renal function indicators (blood urea nitrogen, creatinine, and estimated glomerular filtration rate), nutritional markers (albumin), electrolytes (sodium and potassium), glucose levels, liver enzymes (aspartate aminotransferase and alanine aminotransferase), platelet count, and coagulation status assessed by partial thromboplastin time (PTT).
4) Revised Cardiac Risk Index (RCRI))(13): This traditional scoring system is used to assess the risk of major adverse cardiac events including stroke in noncardiac surgery. The RCRI is composed of the factors mentioned above, such as type of surgery, history of ischemic heart disease, congestive heart failure, cerebrovascular disease, preoperative treatment with insulin, and a preoperative creatinine level greater than 2 mg/dL. The RCRI was used as the comparative score.

Prediction model based on other ML techniques

We developed prediction model based on several ML algorithms, including Support Vector Machine (SVM)(22), Decision Tree Classifier(23), Random Forest (RF)(24) , and CatBoost (25) to compare the performance with the EACH score. The developmental dataset was systematically divided into a training set comprising 70% of the data for model development, and the remaining 30% formed the test set used to assess model performance. Hyperparameter optimization for each model was performed using grid-search cross-validation to identify the optimal settings that maximize the discriminative power of the

model.(26)

Statistical analysis

Continuous variables were analyzed using Student's t-test and the Mann-Whitney U test to determine the significance of differences between the two groups.

Categorical variables in both datasets were evaluated using the chi-square test to investigate the presence of significant associations or discrepancies between the categories. All tests were two-sided, and statistical significance was set at $P < 0.05$. ML modeling was performed in python 3.8 using the Scikit-Learn package.(27) The area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, positive predictive value (PPV), and negative predictive value (NPV) were calculated to evaluate the performance of the prediction model(28)

Ethical approval and patients consent

This study was approved by the Institutional Review Board (IRB) of Seoul National University Hospital and Boramae Medical Center. The IRB determined that participant consent was waived in this retrospective study. We tried to follow the tripod guideline(29)

Results

Baseline characteristics of the study cohorts

Our retrospective observational cohort study utilized data from SNUH comprising the training set (25,551 cases) and the internal validation set (10,951 cases) and the data of 411 patients from BMC to externally validate the prediction model in a geographically and temporally different population. The baseline characteristics of patients in the SNUH and BMC cohorts are summarized in Table 1. Patients in the SNUH cohort were younger than those in the BMC cohort. The prevalence of diabetes was slightly higher in the BMC cohort. Hemoglobin, creatinine, and albumin levels were higher in patients in the SNUH cohort, whereas glucose levels were slightly higher in patients in the BMC cohort.

Table 1. Baseline Characteristics

Characteristics	SNUH cohort	BMC cohort	P-value
Number of stroke			
Age (years)	56.91 ± 15.23	62.44 ± 14.57	< 0.001
Sex, male (%)	16,265 (55.44)	204 (50.50)	0.479
Height (cm)	161.97 ± 8.70	160.70 ± 8.61	0.006
Weight (kg)	63.46 ± 12.01	63.36 ± 11.91	0.998
BMI (kg/m^2)	24.12 ± 3.73	24.49 ± 3.93	0.038
*Preoperative ASA (%)			0.707
1	10,778 (29.53)	25 (6.19)	
2	21,283 (58.31)	298 (73.76)	
3	4,172 (11.43)	74 (18.32)	
4	252 (0.69)	7 (1.73)	
5	17 (0.05)	0 (0.00)	
Emergency of surgery (%)	2,132 (5.84)	19 (4.70)	0.532
Preoperative Hypertension	11,395 (31.22)	168 (41.58)	0.059

(%)			
Preoperative Diabetes (%)	5,543 (15.19)	92 (22.77)	0.900
Preoperative Cardiovascular accident (%)	729 (2.00)	26 (6.44)	1.000
Preoperative Asthma (%)	230 (0.63)	8 (1.98)	1.000
Preoperative COPD (%)	234 (0.64)	7 (1.73)	0.154
Preoperative Liver disease (%)	1,437 (3.94)	42 (10.40)	0.686
Preoperative Kidney disease (%)	1,179 (3.23)	39 (9.65)	0.969
Preoperative Tuberculosis (%)	366 (1.00)	26 (6.44)	0.913
Surgery type (%)	19,787 (54.21)	212 (52.48)	0.514
Hemoglobin(g/dL)	12.90 ± 1.86	12.08 ± 1.93	< 0.001
Platelet ($x10^3$/ µL)	245.18 ± 79.48	247.54 ± 87.81	0.862
Blood Urea Nitrogen(mg/dL)	15.82 ± 9.18	16.88 ± 11.59	0.341
Creatinine(mg/dL)	1.01 ± 1.34	0.43 ± 1.55	< 0.001
Albumin(g/dL)	4.05 ± 0.47	3.75 ± 0.60	< 0.001
Sodium(mmol/L)	140.17 ± 2.66	138.85 ± 3.07	< 0.001
Potassium(mmol/L)	4.24 ± 0.41	3.81 ± 0.54	< 0.001
Glucose(mg/dL)	114.88 ± 39.92	121.52 ± 45.85	0.005
Prothrombin Time (%)	103.58 ± 14.86	11.96 ± 1.12	< 0.001
Partial thromboplastin time(sec)	31.45 ± 5.56	27.66 ± 3.90	< 0.001
Aspartate aminotransferase (IU/L)	25.33 ± 69.93	25.32 ± 52.62	0.968
Alanine aminotransferase (IU/L)	25.44 ± 52.62	28.68 ± 32.74	0.005
Estimated Glomerular Filtration Rate (mL/min/1.73 m^2)	83.79 ± 27.25	83.83 ± 24.82	0.843

Abbreviations: ASA, American Society of Anesthesiologists classification; BMI, body mass index; COPD, chronic obstructive pulmonary disease.
Data are presented as number (%) or mean ± SD
* Surgery types included intrathoracic, intra-abdominal, and supra-inguinal vascular surgery.

Scoring system based on clinical feature intervals

The application of the EACH score to the clinical data is presented in Table 2. Scores ranging from 0 to 100 were assigned to the clinical features based on specific intervals determined by the slope change points. For example, albumin levels less than 4.1 g/dL were associated with a higher risk of perioperative stroke. Specifically, the highest score of 55.4 was assigned to the 2.4-3.5 g/dL range, indicating a particularly elevated risk for some low albumin levels. Similarly, hemoglobin levels were segmented, with the highest score of 51.5 assigned to the 5.0-11.0 g/dL range. Additionally, the 18.6-20.7 kg/m² interval of body mass index received a score of 51.8, suggesting a higher risk associated with specific BMI levels. However, age was divided into intervals, with scores increasing incrementally from 49.0 for the 18-49 years age group to 51.8 for those aged 64-97 years. Creatinine levels did not differ significantly between intervals. These results suggest that continuous variables could impact predictive performance in various nonlinear patterns and that this pattern is well reflected in the EACH scoring system. Categorical variables were scored on the basis of their relative importance in predicting perioperative stroke.

Table 2. SHAP value scoring by feature interval

Variables and interval	Scoring
Age (years)	
18 ~ 49	49.0
49 ~ 55	49.9
55 ~ 64	49.4
64 ~ 65	49.9
64 ~ 97	51.8
Height (cm)	
135.0 ~ 150.4	50.3
150.4 ~ 151.6	50.0
151.6 ~ 158.1	49.9
158.1 ~ 160.6	49.8
160.6 ~ 194.3	50.1

Weight (kg)	
30.0 ~ 56.6	53.0
56.6 ~ 65.2	48.6
65.2 ~ 84.2	49.0
84.2 ~ 135.0	50.2
Estimated Glomerular Filtration Rate	
(mL/min/1.73m^2)	
0.4 ~ 61.3	50.5
61.3 ~ 72.5	50.4
72.5 ~ 85.3	50.7
85.3 ~ 96.4	50.8
96.4 ~ 109.0	48.7
109.0 ~ 419.8	49.1
BMI (kg/m^2)	
11.7 ~ 18.6	51.1
18.6 ~ 20.7	51.8
20.7 ~ 21.2	50.5
21.2 ~ 26.6	48.9
26.6 ~ 27.5	49.4
27.5 ~ 49.7	48.5
Hemoglobin (g/dL)	
5.0 ~ 11.0	51.5
11.0 ~ 11.8	49.7
11.8 ~ 12.9	49.5
12.9 ~ 14.0	49.7
14.0 ~ 20.4	49.8
Platelet (x10^3/μl)	
8.0 ~ 163.0	50.7
163.0 ~ 165.0	50.0
165.0 ~ 179.0	49.9
179.0 ~ 233.0	49.7
233.0 ~ 245.0	50.0
245.0 ~ 1263.0	49.8
Albumin (g/dL)	
0.4 ~ 2.4	50.5
2.4 ~ 3.5	55.4
3.5 ~ 4.1	50.5
4.1 ~ 4.4	46.8
4.4 ~ 4.9	47.2
4.9 ~ 5.7	49.9
Blood Urea Nitrogen (mg/dL)	
2.0 ~ 14.0	49.3
14.0 ~ 17.0	49.5
17.0 ~ 20.0	49.8
20.0 ~ 127.0	51.1
Creatinine (mg/dL)	
0.15 ~ 0.59	50.4
0.59 ~ 0.91	49.5

0.91 ~ 7.33	50.0
7.33 ~ 25.62	50.4
Sodium (mmol/L)	
111.0 ~ 139.0	50.4
139.0 ~ 142.0	49.7
142.0 ~ 144.0	49.9
144.0 ~ 160.0	50.4
Potassium (mmol/L)	
2.3 ~ 3.5	50.8
3.5 ~ 4.0	53.5
4.0 ~ 4.1	50.0
4.1 ~ 4.2	49.7
4.2 ~ 4.5	48.6
4.5 ~ 8.4	47.5
Prothrombin Time (%)	
14.0 ~ 97.0	52.1
97.0 ~ 100.0	50.0
100.0 ~ 106.0	49.3
106.0 ~ 118.0	48.9
118.0 ~ 172.0	49.8
Partial thromboplastin time (sec)	
18.3 ~ 27.0	61.9
27.0 ~ 31.3	46.5
31.3 ~ 32.1	47.8
32.1 ~ 34.3	46.5
34.3 ~ 34.6	49.7
34.6 ~ 400.0	47.6
Glucose (mg/dL)	
1.0 ~ 77.0	49.9
77.0 ~ 95.0	47.4
95.0 ~ 97.0	49.5
97.0 ~ 110.0	49.1
110.0 ~ 114.0	50.1
114.0 ~ 755.0	54.1
Alanine aminotransferase (IU/L)	
1.0 ~ 11.0	49.9
11.0 ~ 21.0	47.3
21.0 ~ 28.0	49.2
28.0 ~ 41.0	51.3
41.0 ~ 5379.0	52.4
Aspartate aminotransferase (IU/L)	
2.0 ~ 20.0	50.6
20.0 ~ 24.0	49.6
24.0 ~ 26.0	49.8
26.0 ~ 34.0	49.7
34.0 ~ 3451.0	50.3
Preoperative ASA	
1 ~ 2	46.2
2 ~ 6	53.8

Preoperative Hypertension	
0	50.7
1	49.3
Preoperative Diabetes	
0	49.8
1	50.3
Preoperative Cardiovascular accident	
0	47.4
1	52.5
Preoperative Asthma	
0	50.0
1	50.0
Preoperative COPD	
0	49.6
1	50.4
Preoperative Liver disease	
0	49.8
1	50.2
Preoperative Kidney disease	
0	49.8
1	50.2
Surgery type	
0	100.0
1	0.100
Preoperative Tuberculosis	
0	49.8
1	50.3
Emergency of surgery	
0	47.1
1	53.0

Abbreviations: BMI, body mass index; COPD, chronic obstructive pulmonary disease; SHAP, SHapley Additive exPlanations.

*Surgery type included intrathoracic, intra-abdominal, and supra-inguinal vascular surgery.

Comparative analysis of the performance of EACH score relative to other ML models

The EACH score demonstrated superior performance compared to other ML-based models, such as Random Forest and CatBoost, as detailed in Table 3. The EACH score had the highest AUC (0.829) and sensitivity (0.881). The Support Vector Machine, Decision Tree Classifier, Random Forest, and CatBoost models exhibited competitive yet slightly lower AUCs of 0.500, 0.501, 0.802, 0.822, and 0.822, respectively. In the external validation, the EACH model consistently maintained high performance. It achieved an AUC of 0.784 and attained the highest sensitivity at 0.900 (Table 3).

Table 3. Comparison between the Performance of EACH score and that of other ML models

Models	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
Internal validation						
Support Vector Machine	0.500	0.954	0.888	0.268	0.003	0.999
Decision Tree Classifier	0.501	0.961	0.820	0.566	0.007	0.999
Random Forest	0.802	0.640	0.856	0.639	0.009	0.999
CatBoost	0.822	0.686	0.770	0.686	0.010	0.999
EACH score	0.829	0.655	0.881	0.654	0.010	0.999
External validation						
Support Vector Machine	0.500	0.955	0.700	0.655	0.049	0.989
Decision Tree Classifier	0.503	0.964	0.821	0.621	0.005	0.999
Random Forest	0.744	0.797	0.600	0.802	0.071	0.988
CatBoost	0.782	0.693	0.900	0.688	0.068	0.996
EACH score	0.784	0.688	0.900	0.683	0.067	0.996

Abbreviations: AUC, area under the receiver operating characteristic curve; EACH, Explainable Automated nonlinear Computation scoring system for Health; ML, machine learning; NPV, negative predictive value; PPV, positive predictive value.

Comparative analysis of the performance of the EACH score relative to traditional scores

We compared the effectiveness of the EACH score with that of a traditional scoring system based on classical statistical methods and ML algorithms (Figure 4). The EACH score

demonstrated superior performance compared to the RCRI, a scoring system used to predict perioperative stroke based on classical logistic regression analysis (AUC, 0.528, vs. AUC, 0.784).

Comparative analysis of the performance of the EACH score relative to other ML-based scoring systems

In addition, we compared the performance of the EACH score to that of the ML-based score generation system, AutoScore (e.g., Random Forest[n=12, n=24]). The performance of AutoScore increased with the number of features: with nine features, the AUC was 0.501; with 12 features, the AUC was 0.564; and with 24 features, the AUC was 0.773. However, the EACH score surpassed the performance of AutoScore, demonstrating a superior AUC of 0.784 (Figure 4).

Risk stratification using the EACH score: low- vs. high-risk patients for perioperative stroke

The two patients representing low and high perioperative stroke risk based on the EACH score are shown in Table 4. The table illustrates how the EACH score model assigned risk scores to individual patient characteristics, leading to a cumulative risk assessment. In assigning scores to patients, the low-risk case, characterized by younger age (42.77 years) and moderate BMI (24.02 kg/m^2), tended to have lower ASA scores, no emergency operation, and no premorbidities. The laboratory results included slightly higher hemoglobin levels (13 g/dL) and higher platelet counts (293 $\times 10^3$/μL), resulting in a lower total score (1379.16) indicative of a lower risk profile. Conversely, the high-risk patient was characterized by older age (71.07 years), lower BMI (22.84 kg/m^2), higher ASA scores, emergency operation, and surgery with a high risk for cardiovascular complications. The patient was had lower

hemoglobin levels (12 g/dL), lower platelet counts (92 $\times 10^3/\mu L$), lower albumin levels (3 g/dL), and higher glucose (115 mg/dL) levels. These factors contributed to the higher total score (1517.29) of this patient, indicating a higher-risk profile. The key distinguishing factors in risk assessment included age, ASA score, surgery type, and laboratory values. The model integrated a wide range of clinical variables into a single risk assessment metric. The ability of the EACH score to distill complex clinical data into a quantifiable stroke risk assessment provides a clear and actionable tool for perioperative stroke risk stratification.

Table 4. Comparative risk stratification using EACH score: low vs. high-risk patients for perioperative stroke

	Actual value	Low risk score	Actual value	High risk score
Age	42.77	49.02	71.07	51.82
Sex, male	1	48.52	1	48.52
Height	160.4	49.78	156.3	49.91
Weight	61.8	48.60	55.8	53.02
BMI	24.02	48.85	22.84	48.85
***Preoperative ASA**	1	46.17	4	53.76
Emergency operation	0	47.10	1	52.95
Preoperative Hypertension	0	49.30	0	49.30
Preoperative Diabetes	0	49.83	0	49.83
Preoperative Cardiovascular accident	0	47.39	0	47.39
Preoperative Asthma	0	50.00	0	50.00
Preoperative COPD	0	49.59	0	49.59
Preoperative Liver disease	0	49.81	1	50.17
Preoperative Kidney disease	0	49.80	0	49.80
Preoperative Tuberculosis	0	49.80	0	49.80
Surgery type *	0	0	1	100
Hemoglobin	13	49.66	12	49.49
Platelet	293	49.80	92	50.66
Blood urea nitrogen	14	49.27	8	49.27
Creatinine	0	50.40	0	50.40
Albumin	4	50.54	3	55.35
Sodium	137	50.41	137	50.41
Potassium	4	53.46	3	50.75
Glucose	81	47.39	115	54.08
Prothrombin time	11.9	52.13	11.6	52.13
Partial thromboplastin time	28	46.54	28	46.54

Aspartate aminotransferase	14	47.30	64	52.37
Alanine aminotransferase	22	49.64	78	50.30
Estimated Glomerular Filtration Rate	120	49.06	90	50.83
Total scoring		1379.16		1517.29

Abbreviations: BMI, body mass index; COPD, chronic obstructive pulmonary disease; EACH, Explainable Automated nonlinear Computation scoring system for Health.

*Surgery type included several type of surgeries Intrathoracic, intraabdominal, and supra-inguinal vascular surgery.

Discussion

Principal results

In this study, we developed and validated a comprehensible automated scoring system, the EACH score, using CatBoost. This system, when applied to predict perioperative stroke, showed superior performance compared with other ML models. It also outperformed the traditional risk score, the RCRI score, and other ML-based risk-scoring systems, such as AutoScore. Moreover, the EACH score effectively discriminated between patients with high- and low-risk perioperative stroke.

Recent advancements in ML-based models have shown promising performance by integrating diverse datasets, irrespective of their linearity or nonlinearity. These models have the potential to overcome the limitations inherent in traditional statistical methods. However, broader application of these models is often restricted because of their lack of explainability and insufficient external validation.(30, 31)

Comparison with prior work

To address these issues, Xie et al. developed AutoScore, a scoring generator that combines Random Forest-based approaches with traditional logistic regression. However, AutoScore may not fully capture complex clinical relationships or accurately represent biological contexts owing to methodological limitations and reliance on logistic regression.(16, 17)

The uniqueness of the present study lies in the development of an entirely ML-based scoring system. The EACH score is adaptable to a variety of data types without relying on predefined assumptions. In this system, the range of each continuous variable was determined by the inflection point on the plot using SHAP values. SHAP values demonstrate the key features and the absolute and relative predictive values of each feature within the model and are widely used to enhance the explainability of ML models. (14, 18) The inflection point on the plot of SHAP values could be assumed to be a significant change of its predictive impact on

outcome.(32) Using the above characteristics of SHAP values, we revealed that the importance of each feature did not increase linearly with its value, but rather showed various relationships. Consequently, the EACH score uniquely reflects the varying significance of identical features, thereby influencing the performance of the prediction model in a distinct manner. Another strength of the EACH score is its robust handling of missing data. If the data for an interval were missing, they were replaced with the average of the surrounding intervals' values, maintaining score completeness and accuracy.(19, 33)

When applied to the real-world clinical data, EACH outperformed the traditional score and other ML-based systems. This demonstrates that this approach could contribute to the construction of a novel risk assessment tool that achieves high predictive accuracy and is intuitive for clinicians to interpret and apply in real-world practice.

Limitations

In addition, we addressed the second common limitation in the application of ML-based prediction models by validating the EACH score in external cohorts from a secondary medical center with different geographical locations, patient volumes, and characteristics from the developmental cohorts. The superior performance of the EACH score in the external validation cohort added robustness to its reliability. Despite these advancements, it is important to acknowledge that the efficacy of the EACH score is contingent on the quality and comprehensiveness of the input data. Incomplete or biased data can result in skewed results. However, the EACH score showed excellent performance even with imbalanced data, such as perioperative stroke, which is a rare disease.(20, 34) Nevertheless, there is a need for ongoing updates and validation of the model across diverse clinical settings. Furthermore, because the current dataset comprises only Asian patients, it is imperative to prospectively validate the EACH score in multiracial cohorts to ensure its generalizability.

Conclusion

The EACH score represents a substantial advancement in the development of efficient, explainable, and automated ML-based risk assessment tools. The universal format of the EACH score is easily applicable to various medical scenarios, thereby reducing the need for labor-intensive data collection. Once integrated with the electronic medical records, it generates specific predictions of different adverse outcomes.(11, 35) Future developments should aim to enhance its user-friendliness for clinicians, potentially through the creation of intuitive interfaces or decision-support tools that simplify the interpretation of results.(36-38)

Acknowledgement: None

Conflicts of interest: None

Abbreviations

ML: machine learning

Funding:

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI22C1295). This research was also supported by the Seoul National University Hospital research fund (grant number: 0320222110).

References

1. Foote C, Woodward M, Jardine MJJAJoKD. Scoring risk scores: considerations before incorporating clinical risk prediction tools into your practice. 2017;69(5):555-7.
2. HJS JJBJS. Risk scoring in surgical patients. 1999;86:149-57.
3. Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. Neth Heart J. 2008;16(6):191-6.
4. Pisters R, Lane DA, Nieuwlaat R, de Vos CB, Crijns HJ, Lip GY. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. Chest. 2010;138(5):1093-100.
5. Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C, et al. Accuracy of clinical assessment of deep-vein thrombosis. Lancet. 1995;345(8961):1326-30.
6. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-72.
7. Royston P, Altman DGJSim. Visualizing and assessing discrimination in the logistic regression model. 2010;29(24):2508-20.
8. Steyerberg EW, Vergouwe YJEhj. Towards better clinical prediction models: seven steps for development and an ABCD for validation. 2014;35(29):1925-31.
9. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230-43.
10. Mohsen F, Ali H, El Hajj N, Shah Z. Artificial intelligence-based methods for fusion of electronic health records and imaging data. Sci Rep. 2022;12(1):17981.
11. Carrasco-Ribelles LA, Llanes-Jurado J, Gallego-Moll C, Cabrera-Bean M, Monteagudo-Zaragoza M, Violan C, et al. Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review. J Am Med Inform Assoc. 2023;30(12):2072-82.
12. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPJJotAMIAJ. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. 2017;24(1):198.
13. Goldman L, Caldera DL, Nussbaum SR, Southwick FS, Krogstad D, Murray B, et al. Multifactorial index of cardiac risk in noncardiac surgical procedures. 1977;297(16):845-50.

14. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, informatics PQCJBm, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. 2020;20:1-9.

15. Rudin CJNmi. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. 2019;1(5):206-15.

16. Xie F, Chakraborty B, Ong MEH, Goldstein BA, Liu NJJmi. AutoScore: a machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. 2020;8(10):e21798.

17. Yuan H, Xie F, Ong MEH, Ning Y, Chee ML, Saffari SE, et al. AutoScore-Imbalance: An interpretable machine learning tool for development of clinical scores with rare events data. 2022;129:104072.

18. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. 2018;2(10):749-60.

19. Little RJ, Rubin DB. Statistical analysis with missing data: John Wiley & Sons; 2019.

20. Ko SB. Perioperative stroke: pathophysiology and management. Korean journal of anesthesiology. 2018;71(1):3-11.

21. Leary MC, Varade P. Perioperative Stroke. Current neurology and neuroscience reports. 2020;20(5):12.

22. Cortes C, Vapnik VJMl. Support-vector networks. 1995;20:273-97.

23. Safavian SR, Landgrebe DJItos, man,, cybernetics. A survey of decision tree classifier methodology. 1991;21(3):660-74.

24. Breiman LJMl. Random forests. 2001;45:5-32.

25. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin AJAinips. CatBoost: unbiased boosting with categorical features. 2018;31.

26. Bergstra J, Yamins D, Cox D, editors. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. International conference on machine learning; 2013: PMLR.

27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. 2011;12:2825-30.

28. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.

29. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a

multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. Circulation. 2015;131(2):211-9.

30. Yin J, Ngiam KY, Teo HHJJomIr. Role of artificial intelligence applications in real-life clinical practice: systematic review. 2021;23(4):e25759.

31. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar LJWIRDM, et al. Interpretability of machine learning-based prediction models in healthcare. 2020;10(5):e1379.

32. Arin P, Minniti M, Murtinu S, Spagnolo NJORM. Inflection points, kinks, and jumps: A statistical approach to detecting nonlinearities. 2022;25(4):786-814.

33. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona OJJoBD. A survey on missing data in machine learning. 2021;8(1):1-37.

34. Lindberg A, Flexman AJBe. Perioperative stroke after non-cardiac, non-neurological surgery. 2021;21(2):59.

35. Wu X, Huang Y, Liu Z, Lai W, Long E, Zhang K, et al. Universal artificial intelligence platform for collaborative management of cataracts. Br J Ophthalmol. 2019;103(11):1553-60.

36. Fonseca A, Ferreira A, Ribeiro L, Moreira S, Duque C. Embracing the future-is artificial intelligence already better? A comparative study of artificial intelligence performance in diagnostic accuracy and decision-making. Eur J Neurol. 2024:e16195.

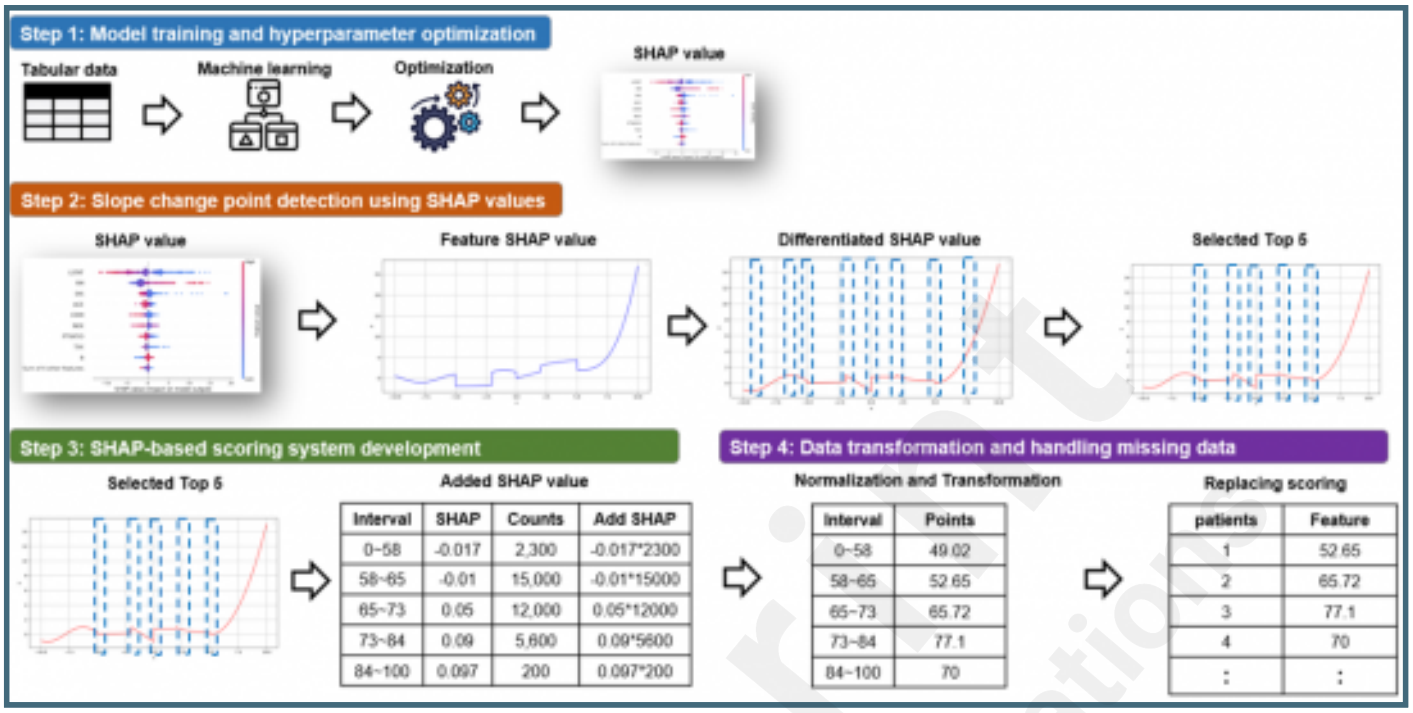
37. Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial Intelligence and Surgical Decision-making. JAMA Surg. 2020;155(2):148-58.

38. Loftus TJ, Upchurch GR, Jr., Bihorac A. Use of Artificial Intelligence to Represent Emergent Systems and Augment Surgical Decision-making. JAMA Surg. 2019;154(9):791-2.

Supplementary Files

Figures

Visual guide to the sequential steps and their detailed execution.

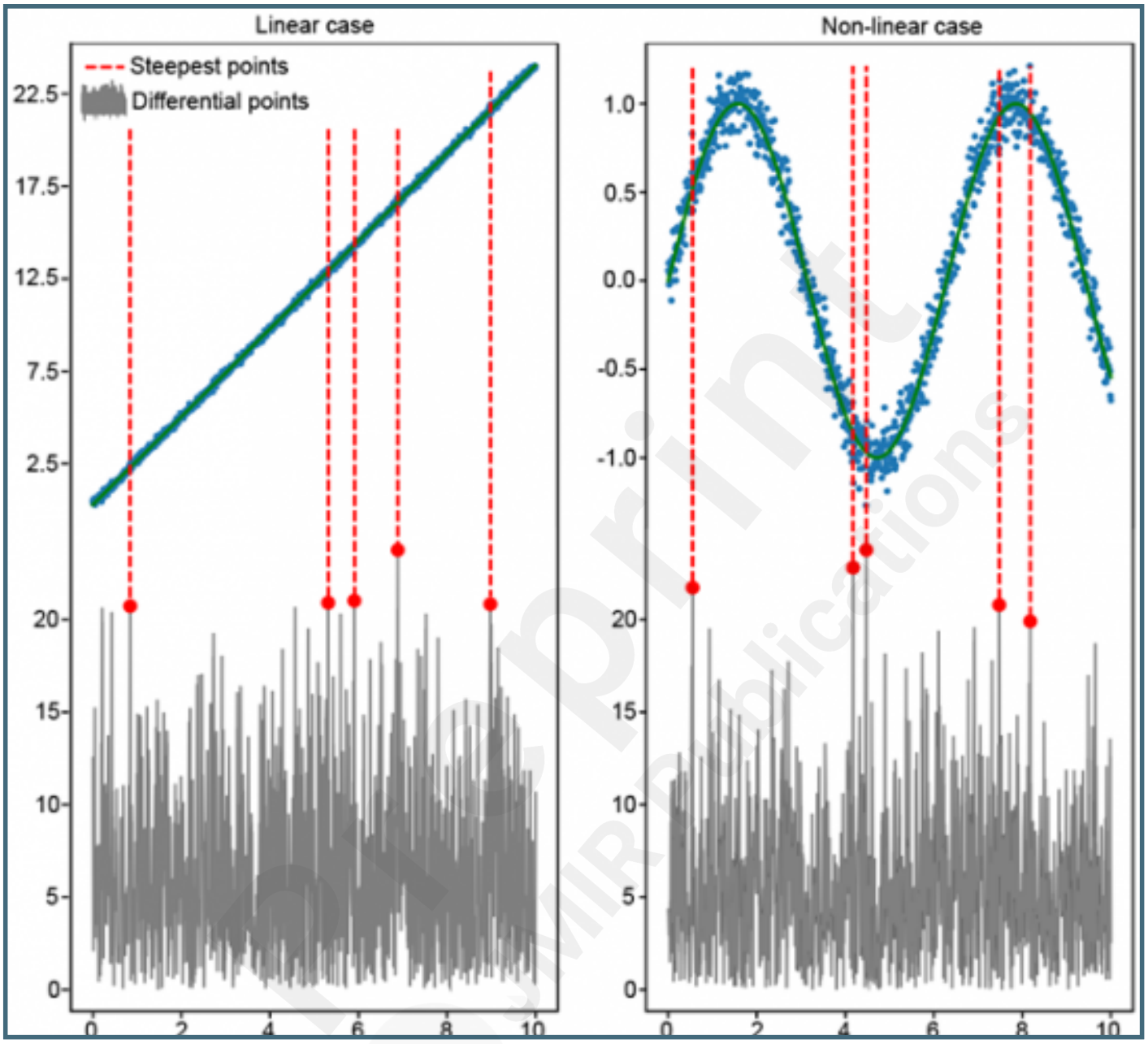


Interval	SHAP	Counts	Add SHAP
0~58	-0.017	2,300	-0.017*2300
58~65	-0.01	15,000	-0.01*15000
65~73	0.05	12,000	0.05*12000
73~84	0.09	5,600	0.09*5600
84~100	0.097	200	0.097*200

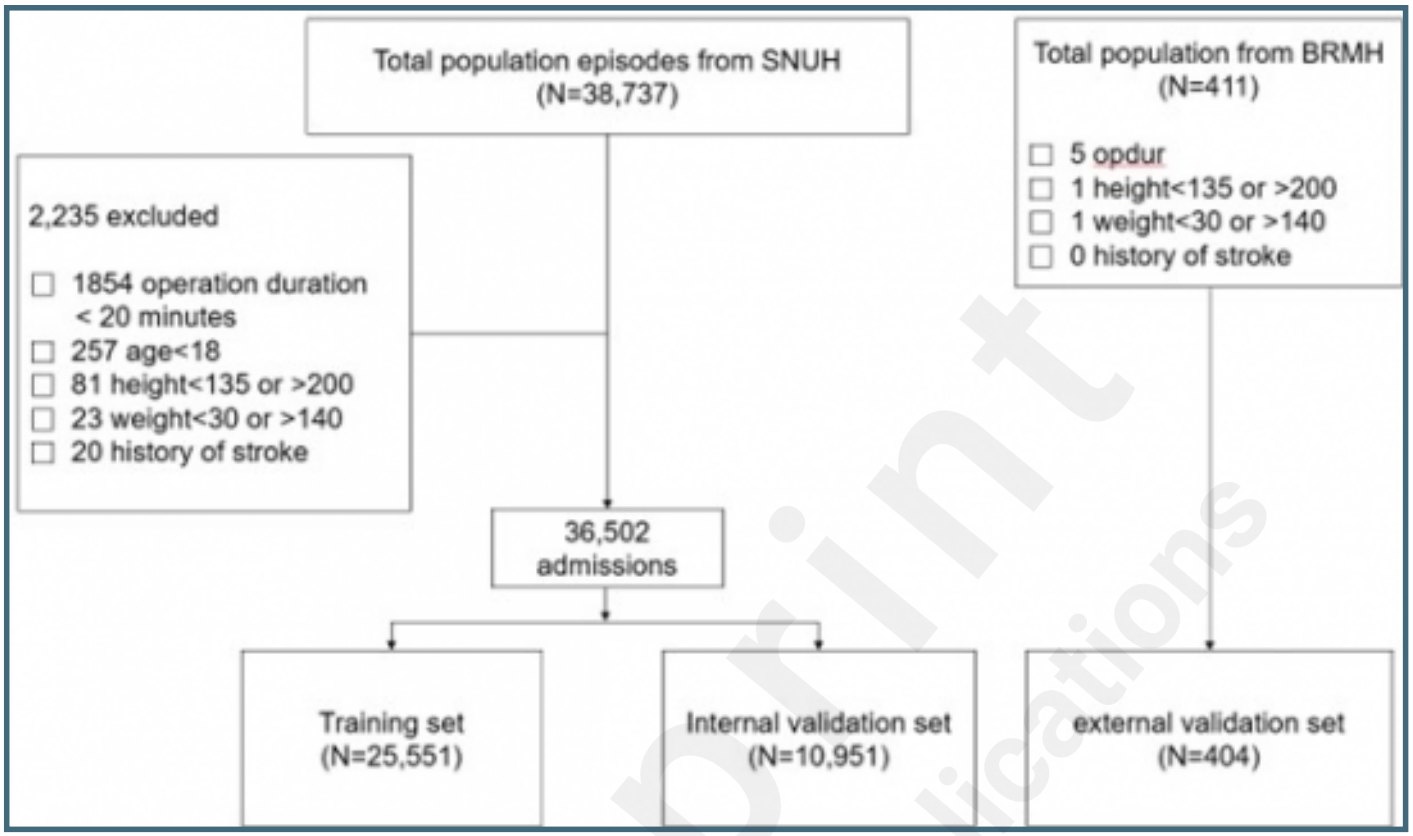
Interval	Points
0~58	49.02
58~65	52.65
65~73	65.72
73~84	77.1
84~100	70

patients	Feature
1	52.65
2	65.72
3	77.1
4	70
:	:

Differential slope change comparison: Linear vs Non-linear cases.



Flow chart of study population.



Comparing ROC curves.