DIGITAL HEALTH
Volume 10: 1–12
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241240910
journals.sagepub.com/home/dhj

Improved interpretable machine learning emergency department triage tool addressing class imbalance

Clarisse SJ Look[1], Salinelat Teixayavong[1], Therese Djärv[2], Andrew FW Ho[1,3], Kenneth BK Tan[3] and Marcus EH Ong[1,3]

Abstract

Objective: The Score for Emergency Risk Prediction (SERP) is a novel mortality risk prediction score which leverages machine learning in supporting triage decisions. In its derivation study, SERP-2d, SERP-7d and SERP-30d demonstrated good predictive performance for 2-day, 7-day and 30-day mortality. However, the dataset used had significant class imbalance. This study aimed to determine if addressing class imbalance can improve SERP's performance, ultimately improving triage accuracy.

Methods: The Singapore General Hospital (SGH) emergency department (ED) dataset was used, which contains 1,833,908 ED records between 2008 and 2020. Records between 2008 and 2017 were randomly split into a training set (80%) and validation set (20%). The 2019 and 2020 records were used as test sets. To address class imbalance, we used random oversampling and random undersampling in the AutoScore-Imbalance framework to develop SERP+-2d, SERP+-7d, and SERP+-30d scores. The performance of SERP+, SERP, and the commonly used triage risk scores was compared.

Results: The developed SERP+ scores had five to six variables. The AUC of SERP+ scores (0.874 to 0.905) was higher than that of the corresponding SERP scores (0.859 to 0.894) on both test sets. This superior performance was statistically significant for SERP+-7d (2019: $Z = -5.843$, $p < 0.001$, 2020: $Z = -4.548$, $p < 0.001$) and SERP+-30d (2019: $Z = -3.063$, $p = 0.002$, 2020: $Z = -3.256$, $p = 0.001$). SERP+ outperformed SERP marginally on sensitivity, specificity, balanced accuracy, and positive predictive value measures. Negative predictive value was the same for SERP+ and SERP. Additionally, SERP+ showed better performance compared to the commonly used triage risk scores.

Conclusions: Accounting for class imbalance during training improved score performance for SERP+. Better stratification of even a small number of patients can be meaningful in the context of the ED triage. Our findings reiterate the potential of machine learning-based scores like SERP+ in supporting accurate, data-driven triage decisions at the ED.

Keywords

Machine learning, emergency department, triage, interpretable

Submission date: 6 August 2023; Acceptance date: 5 March 2024

Introduction

Emergency department (ED) triage is a preliminary clinical assessment that stratifies patients according to the severity of medical condition and urgency of care.[1] Trained nurses usually make triage decisions based on information such as the patient's vital signs and primary complaints.[2] Although conceptually simple, triage decisions in real-world ED

[1]Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore
[2]Department of Medicine Solna, Karolinska Institute, Stockholm, Sweden
[3]Department of Emergency Medicine, Singapore General Hospital, Singapore, Singapore

Corresponding author:
Clarisse Shern Jia Look, Duke-NUS Medical School, 8 College Road, 169857, Singapore, Singapore.
Email: clrslook@gmail.com

settings are complex. There is no consensus on the factors for determining triage category.[2] Additionally, a patient's future clinical course is often not apparent during triage due to the diversity of medical conditions and limited information available.[3] Nurses are also under time pressure to make triage decisions, as they need to allocate patients to the appropriate downstream resources quickly.[3]

One approach to ED triage is to use standardized triage systems such as the Emergency Severity Index (ESI)[4] or the Manchester Triage Scale (MTS).[5] This approach requires nurses to rely heavily on clinical acumen and subjective judgment when making triage decisions.[2] As a result, studies have shown that standardized triage systems are associated with reduced accuracy, poor inter-rater reliability, and high rates of mis-triage.[6–8] Another approach to ED triage is to use clinical risk scores. However, some scores only apply to specific subpopulations of patients, such as the PREDICT score for elderly patients[9] or the Mortality in Emergency Department Sepsis (MEDS) score for patients with suspected infection.[10] Several scores require variables that are not available during ED triage. For example, the Acute Physiology and Chronic Health Evaluation II (APACHE-II)[11] includes laboratory values. Moreover, widely used risk scores like the National Early Warning Score (NEWS) and Modified Early Warning Score (MEWS) have shown only fair or inconsistent predictive abilities for patients in the ED.[12,13]

In recent years, there has been growing interest in applying machine learning (ML) models to ED triage.[14–16] As depicted in Table 1, studies have investigated the use of various models in predicting key outcomes such as triage level, mortality, and critical care admission, yielding promising results.[17–23] The Score for Emergency Risk Prediction (SERP) is a machine learning-based mortality risk prediction score recently developed to address the limitations of existing standardized triage systems and clinical risk scores.[22] SERP was developed using the AutoScore framework, which provided steps for automatically generating simple, point-based clinical scores.[24] Specifically, SERP-2d, SERP-7d, and SERP-30d were three novel risk scores that used routinely available triage data to predict a patient's risk of death at two days, seven days, and 30 days, respectively.[22] SERP-30d achieved an area under the receiver operating characteristics curve (AUC) of 0.821 for two-day mortality, 0.826 for seven-day mortality, and 0.823 for 30-day mortality, outperforming other commonly used risk scores.[22] Besides demonstrating high predictive performance, SERP scores were interpretable, parsimonious, and contained variables already routinely collected during the ED triage.[22] However, the training dataset used in SERP had considerable class imbalance, which was not addressed during the model training. As ML models require many observations to learn patterns for prediction, models trained on imbalanced datasets tend to favor the majority class.[25] Consequently, these models may perform poorly on predictions of the minority class.[25]

The original AutoScore framework used for SERP did not address class imbalance. A recently published extension to the framework, AutoScore-Imbalance, has incorporated adjustments specifically for imbalanced medical datasets.[26] Yuan et al. demonstrated that a mortality risk score generated using the AutoScore-Imbalance achieved better predictive performance with fewer variables compared to a score derived from the AutoScore.[26] These findings suggest that SERP's predictive performance could be improved by incorporating methods to account for the class imbalance. Hence, this study aimed to determine if applying the AutoScore-Imbalance framework to address class imbalance can improve SERP's performance. To do this, we used the AutoScore-Imbalance to develop and evaluate new models, which we called SERP+ scores and compared them with the previously developed SERP scores. We hypothesized that addressing class imbalance would improve SERP's predictive ability. The "+" in the SERP+ scores was a reference to the addition of data rebalancing techniques in score development compared to the original SERP scores which did not include data rebalancing techniques. We reported score development and evaluation details in the Methods and Results sections. We have also highlighted key findings and the potential for clinical application of this tool in the Discussion section.

Methods

Study design and setting

This study was a retrospective cohort study of Singapore General Hospital (SGH) patients. Singapore is a multiracial southeast Asian nation-state with a population of approximately 5.6 million.[27] SGH is Singapore's oldest and largest tertiary hospital, and its ED manages approximately 130,000 attendances annually. This study was approved by the Singapore Health Services' Centralised Institutional Review Board (CIRB Reference 2021/2122), which determined that the study was exempted from full board review due to less than minimal risk on subjects using deidentified data. Due to the study's retrospective nature, a waiver of consent was granted.

Study population

The dataset consisted of all adult SGH ED visits between 1 January 2008 and 31 December 2020.[28] Deidentified data was extracted from the SingHealth Electronic Health Intelligence System (eHints). eHints is an analytics platform that integrates and consolidates data from several information systems within the SingHealth healthcare cluster in Singapore. Inclusion criteria were ED records for adult patients, defined as persons 21 years old and above. We used 21 years of age as the threshold because this is the age of majority stipulated by common law in

Table 1. Summary of a selection of studies investigating the use of machine learning for emergency department triage in the past five years.

Author, year	Outcome measure(s)	Machine learning model(s)	Key findings
Choi et al., 2019[17]	Korean Triage and Acuity Scale level	Logistic regression Random forest XGBoost	The best performing models were the random forest and XGBoost models trained on the full dataset.
Jiang et al., 2021[18]	Triage level	Logistic regression Random forest XGBoost Gradient-boosted decision tree	All models showed good discriminative ability, achieving an AUC of greater than 0.90. XGBoost performed slightly better compared to other models.
Klug et al., 2019[19]	Mortality	XGBoost	The XGBoost model was highly predictive of patients at risk of early mortality, with an AUC of 0.962.
Raita et al., 2019[20]	Hospitalisation Critical care admission	Lasso regression Random forest Gradient-boosted decision tree Deep neural network	Machine learning models outperformed the ESI in the prediction of critical care outcomes and hospitalisation.
Tschoellitsch et al., 2023[21]	Ward admission ICU admission 30-day mortality	Neural network Random forest	The models for ward admission, ICU admission, and 30-day mortality showed good performance and had an AUC of 0.842, 0.819, and 0.925 respectively.
Xie et al., 2021[22]	two-day, seven-day, 30-day mortalities	AutoScore framework which is based on random forest and logistic regression methods	SERP-30d achieved an AUC greater than 0.82 for mortality outcomes and demonstrated superior performance compared to commonly used risk scores.
Yu et al., 2020[23]	Mortality Critical care admission	Logistic regression Deep learning	The machine learning and initial nursing assessment-based system showed higher predictive ability than the Korean Triage and Acuity Scale and Sequential Organ Failure Assessment.

Abbreviations: AUC: area under the receiver operating characteristics curve; ESI: Emergency Severity Index; ICU: intensive care unit; SERP-30d: Score for Emergency Risk Prediction 30-day mortality; XGBoost: eXtreme gradient boosting.

Singapore.[29] We excluded records associated with patients younger than 21 and any duplicate records. There were a total of 1,833,908 records representing information from 813,535 unique patients.

There were 138 variables in the dataset; the complete list of variables can be found in Table S1 of the Supplemental Material. Vital signs and consciousness level (Glasgow Coma Scale,[30] Alert Voice Pain Unresponsive scale[31]) values were taken from the earliest entry in the EHR for the ED visit. These values are usually collected at triage. Comorbidity information was extracted from the International Classification of Diseases, Ninth Revision (ICD-9)[32] codes and the International Classification of Diseases, Tenth Revision (ICD-10)[33] codes in hospital diagnosis and discharge records within five years before the index ED visit. The definition of the comorbidity variables follows the Charlson Comorbidity Index (CCI),[34] and the algorithms by Quan et al.[35] were used to link the ICD-9 and ICD-10 codes to the corresponding comorbidities.

Candidate variables

Candidate variables for score development were preselected based on several criteria to ensure that the derived mortality scores would be valid and useful for triage. These variables had to be available at the point of triage and be relevant to mortality risk prediction based on scientific literature or clinical expertise. We also favored the selection of objective variables that could be calculated rather than those that had to be determined through subjective judgment. This was to minimize the influence of varying levels of clinical expertise and experience on the performance of the derived scoring tool. Sensitive variables were excluded, such as a diagnosis of Human Immunodeficiency Virus (HIV). A visualization of the selection process is shown in Figure S2. The final candidate variables are described in Table S2 of the Supplemental Material.

Outcomes

The two-, seven- and 30-day mortalities were the primary outcome measures used to develop SERP+-2d, SERP +-7d, and SERP+-30d, respectively. These data were obtained from the national death registry records and matched with specific EHR records.

Data analysis

Data analysis was conducted using R software (version 4.2.2)[36] between February and May 2023.

Data pre-processing. Duplications and records relating to patients younger than 21 years were excluded. Missing values were imputed with the median values for continuous variables and the mode for categorical variables. Variables related to vital signs were also inspected for outliers. Vital sign values beyond plausible physiological ranges based on clinical knowledge were considered outliers and set to missing. Outliers included any values below zero, systolic blood pressure (BP) above 300 mmHg, diastolic BP above 180 mmHg, heart rate above 300 beats per minute, respiration rate above 50 breaths per minute, and oxygen saturation from pulse oximetry above 100%. All missing values were subsequently imputed using the median value. Very low vital sign values (e.g. systolic BP of 10) were deemed to be plausible upon admission to the ED as patients may be critically ill or dying.

ED visits between 2008 and 2018 were randomly split into a training set (80%) and a validation set (20%). There were two test sets consisting of ED visits from 2019 and 2020 respectively. This sequential testing design was adopted to be more consistent with future real-world application of the clinical risk scores. Test sets from 2019 to 2020 were separated to delineate any potential impact of the coronavirus disease of 2019 (COVID-19) pandemic on score performance. Baseline characteristics of the training, validation, and test sets were analyzed. Means and standard deviations were reported for continuous variables, and counts with percentages were reported for categorical variables.

SERP+ development. SERP+-2d, SERP+-7d, and SERP +-30d were developed using the AutoScore-Imbalance framework.[26] Following the systematic steps in this framework, model training was completed on the training dataset, and a hold-out validation approach was used by using an independent validation set for parameter tuning. In Block A of the AutoScore-Imbalance framework,[26] the raw imbalanced training dataset was first modified to produce a relatively balanced dataset. From the original dataset, five additional datasets were generated using random oversampling and random undersampling at ratios of 0.3, 0.35, 0.4, 0.45, and 0.5. The ratio represents the proportion of minority class samples relative to the entire dataset. For example, a ratio of 0.3 implies that 30% of the dataset has the minority class label. We used these five datasets to generate five sets of preliminary SERP+ scores and then evaluated the performance of these scores on the validation dataset by calculating the AUC on mortality outcomes. The dataset which produced the best performing score was chosen for the final score derivation. This systematic approach enabled us to choose the ratio that achieved an optimal balance between addressing class imbalance and preventing overfitting. Scoring for the original minority samples remains reliable, as the chosen ratio was validated based on its performance on the independent validation set.

With the optimal dataset, sample weight optimization was performed in Block B of the AutoScore-Imbalance framework,[26] followed by SERP+ derivation in Block C. A random forest was performed on the training dataset for each score to rank the predictor variables. Then, a parsimony plot was generated with the validation dataset, and this was used to evaluate several candidate models and ascertain the variables that should be included in the final model. For the scoring table, cut-off values of continuous variables were adjusted to ensure that they were reasonable and in line with standard clinical values. When the optimized sample weights, variables, and score cut-off points were determined, the final SERP+score was generated with a weighted logistic regression model.

Score evaluation. SERP+-2d, SERP+-7d, and SERP+-30d were evaluated on the 2019 and 2020 test sets to predict two-day, seven-day, and 30-day mortality outcomes. SERP-2d, SERP-7d, and SERP-30d were calculated for each record in the 2019 and 2020 test sets by mapping the value of each predictor variable to the points assigned by the SERP scoring table. The performance of SERP scores was evaluated for the prediction of two-day, seven-day, and 30-day mortalities. Additionally, we assessed the performance of MEWS,[37] NEWS,[38] Rapid Acute Physiology Score (RAPS),[39] and Rapid Emergency Medicine Score (REMS)[40] on both test sets to facilitate comparison of SERP+ and SERP with other commonly used triage scores.

The primary evaluation metric used was the AUC. The 95% confidence interval (CI) measures were calculated for AUC based on bootstrapped samples. The DeLong test was performed to compare the AUC of SERP+ and SERP. Additionally, sensitivity, specificity, balanced accuracy, positive predictive value (PPV), and negative predictive value (NPV) of SERP+ and SERP were determined at the optimal threshold. The optimal threshold was defined

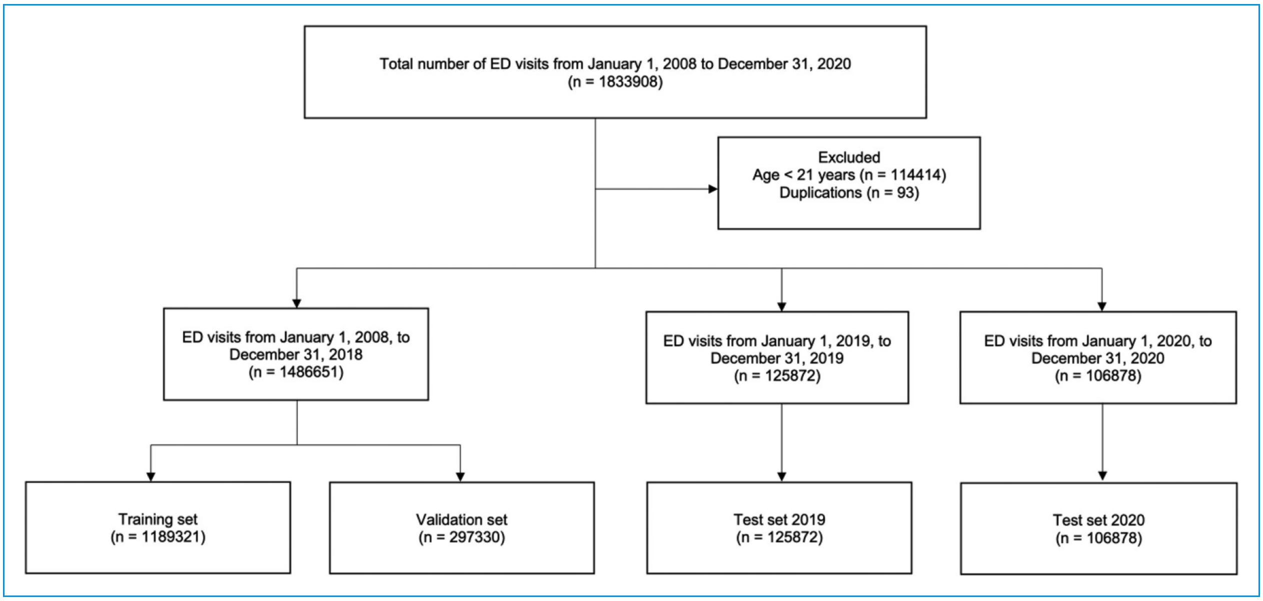
Total number of ED visits from January 1, 2008 to December 31, 2020
(n = 1833908)
Excluded
Age < 21 years (n = 114414)
Duplications (n = 93)
ED visits from January 1, 2008, to December 31, 2018
(n = 1486651)
ED visits from January 1, 2019, to December 31, 2019
(n = 125872)
ED visits from January 1, 2020, to December 31, 2020
(n = 106878)
Training set
(n = 1189321)
Validation set
(n = 297330)
Test set 2019
(n = 125872)
Test set 2020
(n = 106878)

Figure 1. Flowchart showing the formation of the training set, validation set and test sets. ED: emergency department.

by taking the point closest to the upper left corner of the receiver operating characteristics (ROC) curve.

Results

Formation of study cohorts

From the initial SGH ED dataset of 1,833,908 records, we excluded 114,507 records because their age was less than 21 or they were duplicates. The remaining ED visits were split into the training set (n = 1,189,321, 69.2%), validation set (n = 297,330, 17.3%), test set 2019 (n = 125872, 7.3%), and test set 2020 (n = 106,878, 6.2%). Figure 1 presents a flowchart showing the formation of these study cohorts.

Baseline characteristics of study cohorts

Table S3 in the Supplemental Material presents the baseline characteristics of the study cohorts. Overall, the cohorts were similar in gender, race, registration data, vital signs, consciousness level, and historical utilization of healthcare services. Patients in the test sets were, on average, older than those in the training and validation sets, which likely reflected demographic shifts due to the aging population in Singapore.[41] In terms of comorbidities, patients in the test sets are more likely to present with dementia, diabetes, and kidney disease, and they are less likely to have chronic pulmonary disease and peptic ulcer disease. The two-day, seven-day, and 30-day mortality rates in the training set were 0.46%, 0.92%, and 2.21% respectively. This proportion of class imbalance was similar in the other study cohorts.

SERP+ development

To derive SERP+, the original imbalanced training dataset was used to produce relatively balanced datasets at different ratios. The best performing dataset was at a ratio of 0.3 for SERP+-2d, 0.35 for SERP+-7d, and 0.5 for SERP+-30d, as seen in Table 2. The training sets were then used to derive optimal sample weights and the scores for SERP+. Five variables were chosen for SERP+-2d. These were age, heart rate, respiration rate, systolic BP, and diastolic BP. SERP+-7d and SERP+-30d selected these same

Table 2. Performance of different dataset ratios on the validation dataset, AUC (95% CI). The italicized value represents the AUC (95% CI) for the best performing dataset.

Ratio	SERP+-2d	SERP+-7d	SERP+-30d
0.3	*0.896* *(0.887–0.905)*	0.895 (0.889–0.901)	0.896 (0.892–0.899)
0.35	0.895 (0.886–0.904)	*0.897* *(0.891–0.903)*	0.897 (0.893–0.900)
0.4	0.886 (0.876–0.895)	0.895 (0.889–0.901)	0.897 (0.893–0.900)
0.45	0.889 (0.879–0.900)	0.894 (0.888–0.900)	0.897 (0.893–0.900)
0.5	0.891 (0.882–0.900)	0.893 (0.887–0.899)	*0.898* *(0.894–0.901)*

Abbreviations: AUC: area under the receiver operating characteristics curve; SERP+: Score for Emergency Risk Prediction+.

variables with the addition of cancer history. The number of variables was chosen based on the parsimony plots generated in the AutoScore-Imbalance (refer to Figures S3, S4, and S5 in Supplemental Material). Adding more variables to the scoring model after the selected number did not improve model performance significantly. SERP+ variables and scoring cut-off points are shown in Table 3.

Score evaluation

On test set 2019, AUC was 0.874 (95% CI, 0.856–0.890) for SERP+-2d, 0.883 (95% CI, 0.873–0.893) for SERP+-7d, and 0.888 (95% CI, 0.882–0.894) for SERP+-30d on each score's primary mortality outcome. For test set 2020, AUC for the primary mortality outcome was 0.905 (95% CI, 0.893–0.918) for SERP+-2d, 0.893 (95% CI, 0.883–0.902) for SERP+-7d, and 0.890 (95% CI, 0.884–0.895) for SERP+-30d. Score performance remained high across both test cohorts and did not appear to be affected by COVID-19. AUC values of SERP+ were higher than that of the corresponding SERP scores by 0.006 to 0.024. This superior performance was statistically significant for SERP+-7d (test set 2019: $Z = -5.843$, $p < 0.001$, test set 2020: $Z = -4.548$, $p < 0.001$) and SERP+-30d (test set 2019: $Z = -3.063$, $p = 0.002$, test set 2020: $Z = -3.256$, $p = 0.001$). SERP+-2d's higher AUC was not statistically significant compared to SERP-2d (test set 2019: $Z = -1.281$, $p = 0.200$, test set 2020: $Z = -1.821$, $p = 0.069$). In general, SERP+ also showed better sensitivity, specificity, balanced accuracy, and PPV measures than SERP, as seen in Table 4. NPV was identical when comparing SERP+ and SERP for the same mortality outcome (Table 4).

Figures 2 and 3 compare the ROC curves for SERP+, SERP, MEWS, NEWS, RAPS, and REMS on test set 2019 and test set 2020 respectively. These figures show that AUC measures for SERP+ were higher than those of SERP, MEWS, NEWS, RAPS, and REM on both test sets.

Discussion

This study aimed to evaluate if addressing class imbalance during model training can improve the performance of SERP, a mortality risk prediction score for triage. SERP was derived with the AutoScore framework, while SERP+ was developed using the AutoScore-Imbalance on the same mortality outcomes. Comparing SERP+ and SERP, we found that both were similar in the number and type of variables selected as predictors for mortality. SERP+-7d and SERP+-30d achieved significantly better performance than the corresponding SERP scores, showing that accounting for class imbalance during model training can improve score performance. SERP+ also outperformed existing triage risk scores, including the MEWS, NEWS, RAPS, and REMS.

Table 3. Variables and cut-off values for SERP+-2d, SERP+-7d, and SERP+-30d scores.

Variable	SERP +-2d	SERP +-7d	SERP +-30d
Age, year			
<40	0	0	0
40–59	17	15	14
60–74	25	22	23
≥75	31	30	30
Heart rate, /min			
<70	2	1	0
70–79	0	0	0
80–94	8	6	4
≥95	17	12	10
Respiration rate, /min			
<16	26	20	15
16–17	0	0	0
≥18	8	6	3
Systolic BP, mmHg			
<110	19	13	10
110–129	9	6	4
130–139	1	1	0
≥140	0	0	0
Diastolic BP, mmHg			
<60	7	6	5
60–69	0	0	0
70–79	6	3	1
≥80	4	3	1
Cancer history			
None	NA	0	0
Local tumor, leukemia, and lymphoma	NA	8	10
Metastatic solid tumor	NA	18	23

Abbreviations: SERP+: Score for Emergency Risk Prediction+; BP: blood pressure; NA: not applicable.

Table 4. Performance metrics of SERP+ and SERP on the test sets based on primary mortality outcome.

Model	Threshold	AUC	Sensitivity	Specificity	Balanced accuracy	PPV	NPV
Test set 2019							
SERP+-2d	≥54	0.874 (0.856–0.890)	0.796 (0.760–0.832)	0.813 (0.810–0.815)	0.805 (0.785–0.824)	0.017 (0.016–0.017)	0.999 (0.999–0999)
SERP+-7d	≥43	0.883 (0.873–0.893)	0.848 (0.825–0.870)	0.767 (0.765–0.766)	0.808 (0.795–0.818)	0.030 (0.029–0.030)	0.998 (0.998–0.999)
SERP+-30d	≥42	0.888 (0.882–0.894)	0.801 (0.786–0.816)	0.819 (0.817–0.821)	0.810 (0.802–0.819)	0.091 (0.089–0.093)	0.995 (0.994–0.995)
SERP-2d	≥21	0.864 (0.846–0.882)	0.804 (0.766–0.838)	0.768 (0.766–0.770)	0.786 (0.766–0.804)	0.014 (0.013–0.014)	0.999 (0.999–0.999)
SERP-7d	≥26	0.859 (0.848–0.871)	0.826 (0.803–0.848)	0.732 (0.730–0.735)	0.779 (0.767–0.792)	0.025 (0.024–0.026)	0.998 (0.998–0.998)
SERP-30d	≥23	0.882 (0.876–0.889)	0.822 (0.806–0.836)	0.785 (0.783–0.787)	0.804 (0.795–0.812)	0.080 (0.078–0.081)	0.995 (0.995–0.995)
Test set 2020							
SERP+-2d	≥54	0.905 (0.893–0.918)	0.837 (0.806–0.866)	0.828 (0.826–0.830)	0.833 (0.816–0.848)	0.025 (0.024–0.026)	0.999 (0.999–0.999)
SERP+-7d	≥48	0.893 (0.883–0.902)	0.776 (0.751–0.803)	0.856 (0.856–0.861)	0.816 (0.804–0.832)	0.050 (0.049–0.052)	0.998 (0.997–0.998)
SERP+-30d	≥42	0.890 (0.884–0.895)	0.815 (0.800–0.830)	0.816 (0.814–0.818)	0.816 (0.807–0.824)	0.097 (0.095–0.098)	0.995 (0.994–0.995)
SERP-2d	≥25	0.894 (0.878–0.910)	0.730 (0.696–0.768)	0.903 (0.901–0.905)	0.817 (0.799–0.837)	0.038 (0.036–0.040)	0.999 (0.998–0.999)
SERP-7d	≥26	0.874 (0.862–0.885)	0.837 (0.814–0.860)	0.733 (0.731–0.736)	0.785 (0.773–0.798)	0.029 (0.029–0.030)	0.998 (0.998–0.998)
SERP-30d	≥23	0.883 (0.876–0.890)	0.835 (0.821–0.849)	0.781 (0.779–0.784)	0.808 (0.800–0.817)	0.084 (0.083–0.086)	0.995 (0.995–0.995)

Abbreviations: SERP+: Score for Emergency Risk Prediction+; SERP: Score for Emergency Risk Prediction; AUC: area under the receiver operating characteristics curve; PPV: positive predictive value; NPV: negative predictive value.

Variables selected

SERP+ scores selected five to six variables, similar to the number of variables presented in SERP. Having five to six variables constitutes a parsimonious score, as the number of variables required in existing mortality risk scores and ML algorithms can be significantly higher. For instance, the APACHE II used 14 variables,[11] while a deep learning mortality algorithm required 66 variables.[42] Having many variables in a score hinders clinical implementation because there are real-world costs involved in gathering and mapping a high number of variables. In contrast, parsimonious scores like SERP+ minimize the burden of data collection, increasing the likelihood that such scores will be incorporated into clinical practice.

We also found that SERP+ selected similar types of variables compared to SERP. All SERP+ and SERP scores chose age, respiration rate, systolic BP, diastolic BP, and heart rate. It is evident that age and vital sign measurements feature prominently in SERP+ and SERP, a finding supported by existing research studies. Previous research has shown that older patients exhibited higher seven-day and 30-day mortalities across all triage priority levels when compared with younger patients.[43] Old age was also an

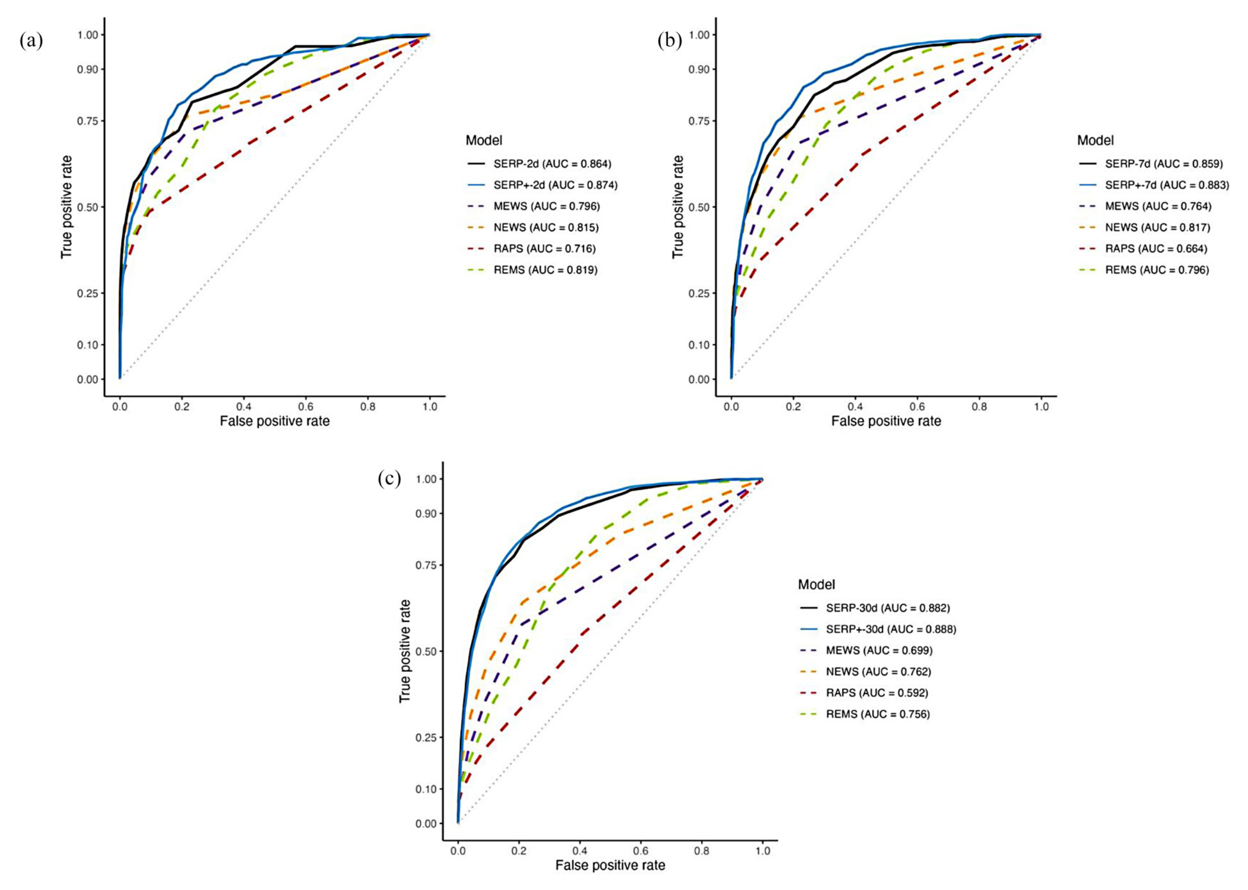

Figure 2. Receiver operating characteristic curves for SERP+, SERP, and other existing triage risk scores for (a) two-day mortality, (b) seven-day mortality, and (c) 30-day mortality on test set 2019. SERP+: Score for Emergency Risk Prediction+; SERP: Score for Emergency Risk Prediction; MEWS: Modified Early Warning Score; NEWS: National Early Warning Score; RAPS: Rapid Acute Physiology Score; REMS: Rapid Emergency Medicine Score; AUC: area under the receiver operating characteristics curve.

independent risk factor for the deterioration of vital signs at the ED despite a stable appearance on arrival.[44] Changes in vital signs have been shown to precede a serious adverse event by several hours, with studies demonstrating that the presence of abnormal vital signs during ED triage is strongly predictive of in-hospital mortality and ICU admission.[44,45] The inclusion of age and vital signs in SERP+ and SERP further reinforces the significance of these variables in mortality prediction.

Additionally, SERP+-7d and SERP+-30d selected cancer history as a variable for mortality prediction. SERP-30d also included this variable.[22] Unlike age and vital signs, existing mortality risk prediction scores do not commonly include cancer history. However, variable ranking in AutoScore-Imbalance identified cancer history as the second and fourth most important variable during the development of SERP+-30d and SERP+-7d, respectively. Studies outside the ED that have investigated cancer history found that it is associated with a higher mortality risk in groups such as hospitalized COVID-19 patients[46] and patients who have undergone primary percutaneous coronary intervention.[47] More research is needed to ascertain the importance and specific contribution of cancer history to the prediction of mortality outcomes at the ED.

Addressing class imbalance

There was a statistically significant improvement in performance for SERP+-7d and SERP+-30d compared to the corresponding SERP scores. ML studies have reported contradictory findings on the effects of class rebalancing on model performance. Khushi et al. found that addressing dataset imbalance improved the predictive capability of a lung cancer prediction model over baseline models derived from imbalanced datasets.[48] Another paper also reported a positive relationship between the rebalancing of class ratios and an increase in AUC when using linear discriminant analysis.[49] The most significant improvement was achieved when the majority and minority classes were equal.[49] In contrast to these findings, Thabtah et al.

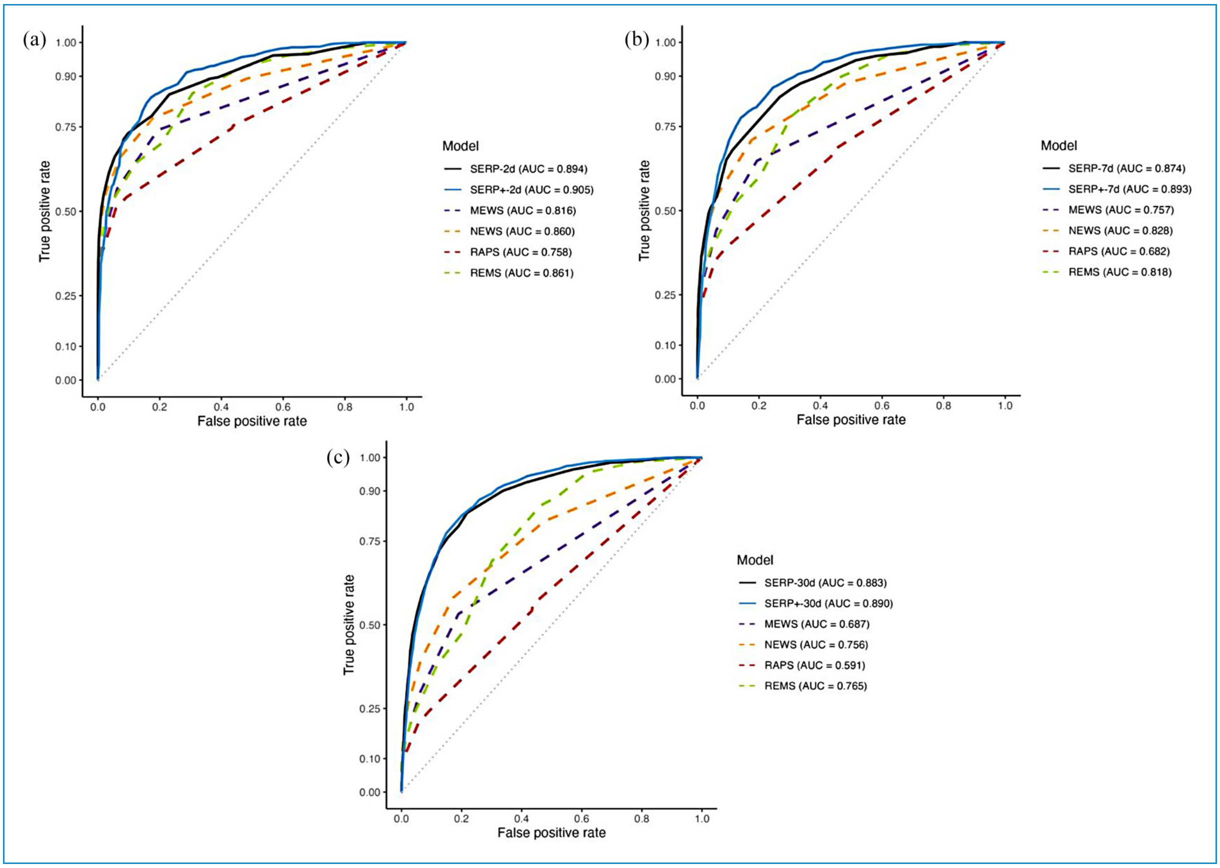

Figure 3. Receiver operating characteristic curves for SERP, SERP+, and other existing triage risk scores for (a) two-day mortality, (b) seven-day mortality, and (c) 30-day mortality on test set 2020. SERP+: Score for Emergency Risk Prediction+; SERP: Score for Emergency Risk Prediction; MEWS: Modified Early Warning Score; NEWS: National Early Warning Score; RAPS: Rapid Acute Physiology Score; REMS: Rapid Emergency Medicine Score; AUC: area under the receiver operating characteristics curve.

found that a Naive Bayes classifier exhibited the worst performance with perfectly balanced datasets.[50] In their study, the best performance was found when there was a relatively large imbalance between the majority class and the minority class at a ratio of 90:10 or 10:90. Similarly, in a study evaluating eight classifiers on 31 datasets with different degrees of imbalance, the model performance was more likely to worsen rather than improve after the application of resampling techniques.[51] Dataset rebalancing was discovered to be more beneficial with linear classifiers such as logistic regression over other types of models.[51] As SERP+ and SERP are based on logistic regression, this may explain why more balanced datasets produced better performing models in our study. Data rebalancing appears to have differing effects on model performance depending on factors such as class ratio and type of classifier used. Hence, it is imperative to validate the effectiveness of any rebalancing technique for the proposed use case before application to model development.

Although all SERP+ scores had higher AUCs than their SERP counterparts on both test sets, AUC improvement was modest, ranging from 0.6% to 2.4%. While this improvement appears marginal, better stratification of even a small number of patients can still be meaningful in healthcare, where decisions can impact human lives. Based on previous studies, we chose a hybrid approach of random oversampling and random undersampling to rebalance our dataset.[26] Adopting other data sampling approaches could further enhance model performance. For example, informed undersampling approaches such as EasyEnsemble[52] or one-sided selection[53] can be used. For oversampling, methods like Synthetic Minority Oversampling Technique (SMOTE) can be evaluated as it has shown to be helpful in several healthcare datasets.[54,55] There is currently no consensus on the best methods to address dataset imbalance, as effectiveness appears to be influenced by factors like the type of data, ratio of class imbalance, and type of classifier.[48,56] Computational cost and training efficiency are also important considerations. For instance, performing SMOTE is computationally more expensive than random oversampling. Further research comparing the different rebalancing approaches is needed to inform decisions regarding method choice for highly imbalanced medical datasets.

Clinical application

The unique contribution of this paper was developing an improved version of SERP scores through the application of the AutoScore-Imbalance framework to address class imbalance. By leveraging data rebalancing techniques, SERP+ has the potential to facilitate more accurate triage decisions compared to existing standardized triage systems and clinical risk scores. Several characteristics of SERP+ make the scores particularly well-suited for clinical application. SERP+ showed superior predictive accuracy on a general ED cohort consisting of patients with varying demographics, acuity levels, and medical presentations. Hence, it can be used as a highly predictive general-purpose mortality risk scoring tool at the ED, removing the need for different risk scores for different patient subgroups. In addition, SERP+'s predictions are interpretable, as the contribution of each variable to the final prediction is apparent from the assigned points. With this transparency, healthcare professionals can understand and explain the predictions made by SERP+, facilitating trust in the model. Reliability and interpretability are crucial for high-stake healthcare decisions, which can significantly affect human lives.[57] In contrast, black box ML models have complex reasoning processes beyond human comprehension, making it difficult for healthcare professionals to judge whether predictions are made on a sound basis.[57] SERP+ scores are also simple, parsimonious, and used variables already routinely collected in most ED settings. Using SERP+ reduces the burden of data collection and allows score use to be more easily incorporated into clinical ED workflows. Using a risk score format is also familiar to most healthcare professionals, which could further facilitate clinical adoption. With the SERP+ also incorporating deaths at two, seven, and 30 days, it can also be potentially used as a discharge tool as well. Future research may include implementation studies to better understand the facilitators and barriers to the clinical use of ML-based triage scores like SERP+.[58]

Limitations

There are several limitations in this study. First, as this was a single-site study, the working practices at SGH ED may influence the derived scores, and score performance may vary in other settings. Second, the list of candidate variables used in score development was not exhaustive. Other variables, such as intubation status, may be found to be a strong predictor if included in the list of candidate variables. Lastly, the PPVs of SERP+ and SERP were low when the score threshold was optimized for AUC. The low PPVs are related to the low prevalence of mortality events. Scores with a lower PPV may be acceptable in the ED, given that a certain level of over-triage is expected to prevent overlooking critically ill patients. There are always trade-offs in test performance metrics, and each healthcare institution can adjust score thresholds according to their requirements.

Threats to validation

The study's focus on a single site introduces a potential threat to external validity as the performance of SERP+ scores may not generalize to other ED departments. Replication of the study in different ED environments is necessary to establish the external validity of the scoring system. Furthermore, as ED working practices are dynamic, the temporal validity of SERP+ may be susceptible to change over time. It is imperative to periodically reassess the model's performance for practical clinical application while considering any evolving factors. Additionally, the omission of potentially relevant candidate variables could be a threat to content validity and may have implications for the predictive accuracy of SERP+. Despite this concern, we believe that our use of a systematic process for variable selection resulted in a comprehensive candidate variable list encompassing standard variables likely present in most ED settings globally.

Conclusions

Accounting for class imbalance during training improved score performance for SERP+. Better stratification of even a small number of patients can be meaningful in ED triage and could positively impact health outcomes. Our findings reiterate the potential of interpretable ML-based scores like SERP+ in supporting accurate, data-driven triage decisions at the ED. Future research can include implementation studies and explore the practical aspects of integrating SERP+ into clinical settings.[58] Furthermore, prospective studies may also be conducted to validate the predictive capabilities of SERP+ and to determine if they can contribute to better triage accuracy and patient outcomes.

Acknowledgments: We thank Dr Liu Nan for his invaluable data science expertise. We also thank Yuan Han for his advice on the application of the AutoScore-Imbalance framework.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Agency for Science, Technology and Research (A*STAR) [grant number H20C6a0032].

Ethical approval: This study was approved by the Singapore Health Services' Centralised Institutional Review Board (CIRB

Reference 2021/2122). Due to the retrospective nature of the study, a waiver of consent was granted.

Guarantor: MO.

Contributorship: CL and MO conceived the study. CL performed data analysis and wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

ORCID iD: Clarisse SJ Look https://orcid.org/0000-0001-9473-7721

Supplemental material: Supplemental material for this article is available online.

References

1. Hinson JS, Martinez DA, Cabral S, et al. Triage performance in emergency medicine: a systematic review. *Ann Emerg Med* 2019; 74: 140–152.
2. Farrohknia N, Castrén M, Ehrenberg A, et al. Emergency department triage scales and their components: a systematic review of the scientific evidence. *Scand J Trauma, Resusc Emerg Med* 2011; 19: 42.
3. Bijani M and Khaleghi AA. Challenges and barriers affecting the quality of triage in emergency departments: a qualitative study. *Galen Med J* 2019; 8: e1619.
4. Agency for Healthcare Research and Quality. Emergency Severity Index (ESI): A Triage Tool for Emergency Department, https://www.ahrq.gov/patient-safety/settings/emergency-dept/esi.html (2022, accessed 29 January 2023).
5. Mackway-Jones K, Marsden J and Windle J. *Emergency Triage: Manchester Triage Group*. 3rd ed. West Sussex, UK: Wiley Blackwell, 2014, https://www.wiley.com/en-us/Emergency+Triage&per;3A+Manchester+Triage+Group&per;2C+3rd+Edition-p-9781118299067 (accessed 21 March 2023).
6. Christ M, Grossmann F, Winter D, et al. Modern triage in the emergency department. *Dtsch Arztebl Int* 2010; 107: 892–898.
7. Mistry B, Stewart De Ramirez S, Kelen G, et al. Accuracy and reliability of emergency department triage using the emergency severity Index: an international multicenter assessment. *Ann Emerg Med* 2018; 71: 581–587.e3.
8. Hinson JS, Martinez DA, Schmitz PSK, et al. Accuracy of emergency department triage using the emergency severity Index and independent predictors of under-triage and over-triage in Brazil: a retrospective cohort analysis. *Int J Emerg Med* 2018; 11: 3.
9. Moman RN, Loprinzi Brauer CE, Kelsey KM, et al. PREDICTing mortality in the emergency department: external validation and derivation of a clinical prediction tool. *Acad Emerg Med* 2017; 24: 822–831.
10. Shapiro NI, Wolfe RE, Moore RB, et al. Mortality in emergency department sepsis (MEDS) score: a prospectively derived and validated clinical prediction rule. *Crit Care Med* 2003; 31: 670–675.
11. Naved SA, Siddiqui S and Khan FH. APACHE-II Score correlation with mortality and length of stay in an intensive care unit. *J Coll Physicians Surg Pak* 2011; 21: 4–8.
12. Mitsunaga T, Hasegawa I, Uzura M, et al. Comparison of the national early warning score (NEWS) and the modified early warning score (MEWS) for predicting admission and in-hospital mortality in elderly patients in the pre-hospital setting and in the emergency department. *PeerJ* 2019; 7: e6947.
13. Hamilton F, Arnold D, Baird A, et al. Early warning scores do not accurately predict mortality in sepsis: a meta-analysis and systematic review of the literature. *J Infect* 2018; 76: 241–248.
14. Chen Y, Chen H, Sun Q, et al. Machine learning model identification and prediction of patients' need for ICU admission: a systematic review. *Am J Emerg Med* 2023; 73: 166–170.
15. Xiao Y, Zhang J, Chi C, et al. Criticality and clinical department prediction of ED patients using machine learning based on heterogeneous medical data. *Comput Biol Med* 2023; 165: 107390.
16. Sánchez-Salmerón R, Gómez-Urquiza JL, Albendín-García L, et al. Machine learning methods applied to triage in emergency services: a systematic review. *Int Emerg Nurs* 2022; 60: 101109.
17. Choi SW, Ko T, Hong KJ, et al. Machine learning-based prediction of Korean triage and acuity scale level in emergency department patients. *Healthc Inform Res* 2019; 25: 305–312.
18. Jiang H, Mao H, Lu H, et al. Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. *Int J Med Inf* 2021; 145: 104326.
19. Klug M, Barash Y, Bechler S, et al. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score. *J GEN INTERN MED* 2020; 35: 220–227.
20. Raita Y, Goto T, Faridi MK, et al. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* 2019; 23: 64.
21. Tschoellitsch T, Seidl P, Böck C, et al. Using emergency department triage for machine learning-based admission and mortality prediction. *Eur J Emerg Med* 2023; 30: 408.
22. Xie F, Ong MEH, Liew JNMH, et al. Development and assessment of an interpretable machine learning triage tool for estimating mortality after emergency admissions. *JAMA Netw Open* 2021; 4: e2118467.
23. Yu JY, Jeong GY, Jeong OS, et al. Machine learning and initial nursing assessment-based triage system for emergency department. *Healthc Inform Res* 2020; 26: 13–19.
24. Xie F, Chakraborty B, Ong MEH, et al. Autoscore: a machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR Med Inform* 2020; 8: e21798.
25. Kumar P, Bhatnagar R, Gaur K, et al. Classification of imbalanced data: review of methods and applications. *IOP Conf Ser: Mater Sci Eng* 2021; 1099: 012077.
26. Yuan H, Xie F, Ong MEH, et al. AutoScore-Imbalance: An interpretable machine learning tool for development of clinical scores with rare events data. *J Biomed Inform* 2022; 129: 104072.

27. Department of Statistics Singapore. Population Trends 2022, https://www.singstat.gov.sg/-/media/files/publications/population/population2022.ashx (2022, accessed 22 March 2023). 2022.
28. Liu N, Xie F, Siddiqui FJ, et al. Leveraging large-scale electronic health records and interpretable machine learning for clinical decision making at the emergency department: protocol for system development and validation. *JMIR Res Protoc* 2022; 11: e34201.
29. Baker McKenzie. Minors | Global Data Privcy and Security Handbook, Singapore, https://resourcehub.bakermckenzie.com/en/resources/data-privacy-security/asia-pacific/singapore/topics/minors (2022, accessed 19 October 2023). 2022.
30. Teasdale G and Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet* 1974; 2: 81–84.
31. American College of Surgeons' Committee on Trauma. *Advanced trauma life support for doctors*. 6, 1977.
32. World Health Organization. *International Classification of Diseases, Ninth Revision (ICD-9)*. Geneva, Switzerland: World Health Organization, 1977, https://iris.who.int/handle/10665/42980 (1977).
33. World Health Organization. *ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision, 2nd ed*. Geneva, Switzerland: World Health Organization, 2003, https://iris.who.int/handle/10665/42980 (2003).
34. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987; 40: 373–383.
35. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005; 43: 1130–1139.
36. R Core Team. R: A language and environment for statistical computing, https://www.R-project.org/ (2022). 2022.
37. Subbe CP, Kruger M, Rutherford P, et al. Validation of a modified early warning score in medical admissions. *QJM* 2001; 94: 521–526.
38. Royal College of Physicians. *National Early Warning Score (NEWS) 2*. London: Royal College of Physicians, 2017, https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2 (19 December 2017, accessed 30 June 2023).
39. Rhee KJ, Fisher CJ and Willitis NH. The rapid acute physiology score. *Am J Emerg Med* 1987; 5: 278–282.
40. Olsson T, Terent A and Lind L. Rapid emergency medicine score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients. *J Intern Med* 2004; 255: 579–587.
41. Tan CC, Lam CSP, Matchar DB, et al. Singapore's healthcare system: key features, challenges, and shifts. *Lancet* 2021; 398: 1091–1104.
42. Li D, Fu J, Zhao J, et al. A deep learning system for heart failure mortality prediction. *PLoS One* 2023; 18: e0276835.
43. Ruge T, Malmer G, Wachtler C, et al. Age is associated with increased mortality in the RETTS-A triage scale. *BMC Geriatr* 2019; 19: 139.
44. Henriksen DP, Brabrand M and Lassen AT. Prognosis and risk factors for deterioration in patients admitted to a medical emergency department. *PLoS One* 2014; 9: e94649.
45. Barfod C, Lauritzen MMP, Danker JK, et al. Abnormal vital signs are strong predictors for intensive care unit admission and in-hospital mortality in adults triaged in the emergency department - a prospective cohort study. *Scand J Trauma Resusc Emerg Med* 2012; 20: 28.
46. Meng Y, Lu W, Guo E, et al. Cancer history is an independent risk factor for mortality in hospitalized COVID-19 patients: a propensity score-matched analysis. *J Hematol Oncol* 2020; 13: 75.
47. Wang F, Gulati R, Lennon RJ, et al. Cancer history portends worse acute and long-term noncardiac (but not cardiac) mortality after primary percutaneous coronary intervention for acute ST-segment elevation myocardial infarction. *Mayo Clin Proc* 2016; 91: 1680–1692.
48. Khushi M, Shaukat Dar K, Mahboob Alam T, et al. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access* 2021; 9: 109960–109975.
49. Xue J-H and Hall P. Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis? *IEEE Trans Pattern Anal Mach Intell* 2015; 37: 1109–1112.
50. Thabtah F, Hammoud S, Kamalov F, et al. Data imbalance in classification: experimental evaluation. *Inf* 2020; 513: 429–441.
51. Kim M and Hwang K-B. An empirical evaluation of sampling methods for the classification of imbalanced data. *PLoS One* 2022; 17: e0271260.
52. Liu X-Y, Wu J and Zhou Z-H. Exploratory undersampling for class-imbalance learning. *IEEE Trans Syst Man Cybern Syst* 2009; 39: 539–550.
53. Kubat M and Matwin S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In: *Proceedings of the 14th International Conference on Machine Learning*. 1997, pp. 179–186.
54. Li J, Liu L, Fong S, et al. Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *PLoS One* 2017; 12: e0180830.
55. Kishor A and Chakraborty C. Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE. *Int J Syst Assur Eng Manag* 2021. Epub ahead of print 23 June 2021. DOI: 10.1007/s13198-021-01174-z.
56. Wongvorachan T, He S and Bulut O. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information* 2023; 14: 54.
57. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; 1: 206–215.
58. Chan SL, Lee JW, Ong MEH, et al. Implementation of prediction models in the emergency department from an implementation science perspective—determinants, outcomes, and real-world impact: a scoping review. *Ann Emerg Med* 2023; 82: 22–36.