Unified Auto Clinical Scoring (Uni-ACS) with Interpretable ML models

Firstname Lastname NAME@EMAIL.EDU
Department
University
City, State, Country

Firstname Lastname NAME@EMAIL.EDU
Department
University
City, State, Country

Abstract

Despite significant progress in explainable Machine Learning (ML) tools (such as LIME, SHAP and explainable boosting machines) in explaining ML models' risk predictions in clinical problems (such as heart failure, acute kidney injury, sepsis and hypoxaemia during surgery), the interpretations generated remain to be an unfamiliar language to most clinicians. Clinical scores continue to be the preferred tool for risk stratification as they are concise, clinically correlatable and can be used at patient's bedside without a machine. In this work, we reproduce the classical clinical scoring development approach to uncover its limitations in determining categorical features and using logistic regression coefficients to derive additive integer scoring systems. Subsequently, we propose the Unified Automatic Clinical Scoring (Uni-ACS) development framework, which overcomes these limitations to translating ML models into clinical scores by leveraging on explainable outputs from SHAP compatible ML models. We hypothesize that this approach is model agnostic, can be automated and can retain the complex predictive power of the underlying ML model, while relating key model insights to clinicians in a clinical risk scoring format. In our experiments, we applied Uni-ACS to a variety of ML models trained on the MIMIC III and MIMIC IV sepsis cohorts to predict mortality and ICU admission. We showed that Uni-ACS derived clinical score retained a greater proportion of the underlying ML models' predictive performance (lowest AUROC drop of 2.44%), compared against the baseline clinical score (lowest AUROC drop of 5.79%). We further verified Uni-ACS derived clinical score's insights against the current literature to show its clinical applicability. Uni-ACS and datasets used for method validation are open-sourced for the community to use and verify[1].

1. Introduction

Recent advances in Machine Learning (ML) interpretability, in the form of Local Interpretable Model Agnostic Explanations (LIME) by (Ribeiro et al., 2016), SHapley Additive exPlanations (SHAP) by (Lundberg and Lee, 2017) and explainable boosting machines by (Nori et al., 2019), have held great promise in exposing the "black box" ML models, thus

1. https://anonymous.4open.science/r/Uni-ACS-2758/

reducing the barrier for widespread adoption of ML models for clinical risk prediction. However, clinical scores remain to be the default and most popular means of estimating patient outcomes. Online medical calculator website MDCalc estimates that "millions of medical professionals" over "200+ countries" and "65% of US physicians" use its services monthly to calculate risk with clinical scores (Walker and Habboushe, 2022). Additionally, they are commonly applied and widely validated for medical conditions such as ischaemic heart disease (TIMI by (Morrow et al., 2000), GRACE by (Fox et al., 2006)), atrial fibrillation (CHA2DS2-VASc by (Lip et al., 2010), HAS-BLED by (Pisters et al., 2010)), sepsis (APACHE by (Knaus et al., 1985), qSOFA by (Singer et al., 2016)) and many other diseases (Well's criteria for deep venous thrombosis by (Wells et al., 1995), MELD for end stage liver disease by (Kamath et al., 2001)).

Clinical scores are defined as additive integer scoring systems designed to stratify risk for specific patient outcomes. They are conventionally developed based on logistic regression models with manually selected clinical features. Individual features are assigned integer values for: (a) specific feature value ranges for continuous variables and, (b) specific categories for nominal or binary feature variables. These integer values are added up to a final clinical score. Different cumulative scores correlate with different risks associated with an outcome.

Plausible reasons for why clinical scores remain the preferred interface for risk models amongst clinicians are as follows. Firstly, clinical scores are concise, which allows clinicians to remember them easily. Secondly, clinical scores are easy to interpret and correlate clinically, as they provide a quantifiable integer weightage for each predictor's impact on adverse outcome. Finally, clinical scores can be used without a machine. This is critical when clinicians require a bedside estimation of risk while treating infectious diseases like SARS-CoV-2 infections, where clinicians would be in full personal protective equipment and might not have access to digital devices for infection control purposes.

While explainable ML has been applied to predicting clinical outcomes for a non exhaustive list of conditions such as heart failure (Lu et al., 2021), acute kidney injury (Tseng et al., 2020) and hypoxaemia in surgery (Lundberg et al., 2018), the interpretations do not offer an equivalent interface to clinical scores. Therefore, this paper proposes the implementation of Unified Automatic Clinical Scoring (Uni-ACS) to translate clinical ML models into clinical scores. We highlight the following generalizable insights regarding our approach.

Generalizable Insights about Machine Learning in the Context of Healthcare

- Uni-ACS is an automated and model agnostic methodology for translating SHAP compatible ML models into clinical scores, an interface currently most preferred by clinicians for estimating clinical risk.

- Uni-ACS clinical scores retain global and local SHAP interpretations of ML models, thus providing clinicians with consistent interpretable risk estimation methods from bedside to machine.

- Uni-ACS clinical scores would preserve a reasonable portion of the original ML models' predictive performance. In this paper, we also showed that Uni-ACS clinical scores,

generated from ML models, had superior predictive performance compared to baseline, classical clinical scores.

2. Related Work

As suggested in the introduction, there are a number of scores, such as TIMI, GRACE, CHADS2VASC, HAS-BLED, APACHE, qSOFA, Well's and MELD, which are currently in active clinical use. Our literature review revealed the following groups of work, where clinical scoring were done in conjunction with ML modelling: (a) Conventional clinical scores were developed in parallel to ML models. Examples include prediction of arterial hypertension in primary hyperaldosteronism (Buffolo et al., 2021) and severity of disease in COVID-19 ISARIC score (Knight et al., 2020). (b) ML models were built on clinical scores as input features. Examples include prediction of obstructive coronary artery disease in coronary computed tomography angiography (Al'Aref et al., 2020). (c) A framework was defined for automatic clinical scoring development, where ML was used to choose feature subset and logistic regression was used to develop the final score (Xie et al., 2020). (d) An approach was proposed to learn clinical score directly from data with a mixed integer non linear program augmented by a cutting plane algorithm, Risk SLIM (Ustun and Rudin, 2019). Examples of Risk SLIM's applications included clinical problems of sleep apnea (Ustun and Rudin, 2016), seizure (Ustun and Rudin, 2017), appendicitis (Aparicio et al., 2021), kidney transplant (Profitlich and Sonntag, 2019) and the non-clinical problem of criminal law (Wang et al., 2022).

The classical clinical scoring development approach, applied in several variants as described non-exhaustively in related works (a), (b) and (c), was a post logistic regression modelling approach, supplemented by heuristics such as scaling and rounding regression coefficients to integer scores (Cole, 1993). In contrast to this post hoc approach, related work (d), Risk SLIM, derived its advantage in being a highly predictive scoring system by optimizing the clinical score directly from data, with integer coefficient and operational constraints. Without a post ML modelling approach to developing clinical scores in the current literature, highly predictive ML models and clinical scores would have to be developed in separate silos. This denies the clinician an opportunity to apply a clinical score, which is consistent in its understanding of risk with the best ML model.

3. Methods

We describe the classical clinical score development methodology here for 2 reasons: (1) Uni-ACS was built on its definition and algorithmic heuristics. (2) We aim to use it as a baseline. Hence, we systematically reproduce it as a series of steps and discuss the variations at each step.

3.1. Classical clinical score development

Clinical scores have the following components:

- **Component A, clinical scoring table:** This table consists of risk factors and their associated integer score values. Each risk factor is a predictive feature with a well

defined range of values. If a patient's feature value falls within this range of values, the feature will be assigned the integer score adjacent to it. Individual scores would be added up to give a final aggregated score.

- **Component B, Score to risk mapping table:** This table maps the aggregate scores to their respective actual risk percentages. An alternate representation to risk percentages would be the odds ratio.

TIMI, a clinical score by (Morrow et al., 2000) for predicting mortality risk in patients with acute coronary syndrome, is used to illustrate these components. Appendix A's Supplementary Table 1 show's TIMI's clinical scoring table and Appendix A's Supplementary Table 2 show's TIMI's score to risk mapping table. The developmental process for such a clinical score can be outlined with the following 5 steps.

Step 1: Feature selection

Clinical features for inclusion into the clinical score can be manually handpicked by clinicians or chosen automatically with a search algorithm. Clinicians might opt to choose a specific set of features for practical reasons: (1) Not all features might be obtainable at point of consult. (2) Specific features might have close physiological relation to the outcome of interest. Any algorithmic feature selection method can be classified as either a wrapper, filter or embedded method, as described by (Guyon and Elisseeff, 2003). Features of traditionally established clinical scores were selected with a filter based method, in which a statistical measure would be used to choose the most parsimonious subset of features. Features of contemporary clinical scores, such as the ISARIC COVID severity score (Knight et al., 2020), were selected with a wrapper based method, in which ensemble random forest decision trees would be commonly used as the learning algorithm of choice.

Step 2: Feature value categorisation

Component A of a clinical score, as exemplified in Appendix A's Supplementary Table 1, requires risk factors to have distinct categories. Dichotomous clinical features such as history of Diabetes Mellitus (DM) or Hypertension (HTN) need no further processing since they are already a binary feature. However, continuous clinical features such as Systolic Blood Pressure (SBP) or Heart Rate (HR) would have to be categorised. The categorisation can be manually decided based on clinical heuristics such as the age cutoff for elderly patients (> 65 years old) for demographic features or abnormal levels of white blood cells ($> 10 \times 10^3$ cells per mL of blood) for laboratory features. Alternatively, the categorisation of continuous clinical features can be done with Generalised Additive Models (GAM) (Barrio et al., 2013). Another approach would be to categorise the feature according to it's probability distribution quantiles (Xie et al., 2020).

Regardless of the approach chosen for feature value range categorisation, the outcome is to map input features x_{ij} for $i = N$ rows of data and $j = M$ columns of features to input categories c_{ijk}, where c_{ijk} is the kth categorical assignment for x_{ij}. As the number of categories vary across $j = M$ columns, maximum number of categories is $k = K_j$.

Step 3: Modelling with logistic regression

Logistic regression modelling is preferred in Step 3 for the following reasons. The logistic model is given by

$$f(c) = \ln\left(\frac{p(c)}{1-p(c)}\right) = B_0 + \sum_{j=1}^{M}\sum_{k=1}^{K_j} B_{jk}c_{jk}, \tag{1}$$

where $f(c)$ is the logistic regression model of categorised input features, c. c_{jk} represents the vector of categorised input features and B_{jk} represents the log of odds of each category, at the jth feature and kth category. Firstly, the beta values of features, i.e., the coefficients B_{jk}, in such a model, as shown in (1), can be interpreted as an additive log of odds effect for every unit change in feature value. As the the input features, x, have been transformed into input categories c in Step 2, a change in category from 0 to 1 of an input category, c_{jk} will represent an additive increase in the log of odds by a magnitude equivalent to the category's beta value, B_{jk}. These beta values can thus be used to estimate the integer score values as in Component A of the clinical score. Secondly, the probability of a class having a specific adverse outcome is directly modelled. Thus, it allows for the calculation of risk percentages as in Component B of the clinical score.

Step 4: Create clinical scoring table with beta values

While it is established that beta values of regression model can be used to estimate clinical scores, beta values are not necessarily integers and non-negative numbers. This violates the requirements, as set out in Table 1, where scores are non-negative integers. Thus, Component A of the clinical score is completed when beta values are converted to integer scores through the following process:

1. Map a unit integer score in a scoring table to the lowest absolute beta value of the logistic regression model.
2. Beta values are divided by this lowest absolute beta value and rounded to the nearest whole integer number.
3. To remove negative scoring, the absolute value of the most negative score is added to all scores.

Step 5: Create clinical score to risk mapping table

Since $f(c)$ is modelled as the log of odds of c, we can take the expit (i.e., logistic sigmoid) of the product of the unit beta value and incremental integer values between 0 and maximum possible score to create Component B of the clinical score, the score to risk mapping table. Alternatively, we can go back to the data and directly measure the ratio of patients with positive labels amongst groups of patients with different scoring.

3.2. Uni-ACS clinial score development

Classical clinical score development cannot be applied directly to ML models because, unlike logistic regression models, ML models have the following limitations:

- **Limitation I:** They do not model output as a linear function of features, which is a necessary requirement for generating scores that can be summed.
- **Limitation II:** They do not have an equivalent beta coefficient to directly quantify feature category's impact on overall risk, which is a necessary requirement for generating an integer score for individual feature categories.

In this section, we outline changes to the order of the steps in classical clinical score development and propose specific algorithmic heuristics to translate ML models into clinical scores. We hypothesize that our proposed method, Uni-ACS, will have the following attributes:

1. Offers an automatic and model agnostic approach to translating ML models into clinical scores with the essential Components A and B.
2. Has consistent frame of local and global interpretations across ML models and their derived clinical scores.
3. Translated clinical scores will preserve most of the underlying ML models' original predictive performance.

Step 1: ML Modeling and SHAP application

A desired feature of Uni-ACS is the retainment of the original ML model with the translated clinical score as an explainable interface of the underlying model. Therefore, instead of performing the modeling as described in Step 3 of the classical clinical score development process, we propose ML modeling to be completed in Step 1.

Since clinical scores are developed primarily on top of tabular clinical data, we shall use the best known existing risk prediction models for tabular clinical data in our experiments: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB) and Neural Networks (NN). Data is split into training dataset and held out test dataset, in a 70% to 30% ratio. 5 fold cross validation repeated 10 times is applied on the training dataset to search for the best set of hyperparameters for each model, based on target model performance of interest. The best model of each modeling method is applied on the held out test set.

We then apply the well known feature attribution method known as SHAP, which was described by (Lundberg and Lee, 2017). The SHAP model is given by

$$f(x) = g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j, \tag{2}$$

where the original model, denoted by $f(x)$, can be matched by the explanation model $g(x')$, and x' is the simplified representation of the original input x and ϕ_j, also known as SHAP value, is the model output with all simplified inputs x' turned off.

SHAP is applied to best models for global and local interpretations because SHAP has the following key characteristics necessary for subsequent clinical score development:

(a) SHAP is an additive feature attribution method, as shown by (2). This characteristic resolves issues raised in **Limitation I**.

(b) We can draw parallels between (1) and (2), where B_{jk} is analogous to but not equal to ϕ_j. This characteristic resolves issues raised in **Limitation II**.

(c) Explanations of Kernel SHAP, Tree SHAP and linear SHAP are applicable to most ML models used for tabular clinical data. Deriving clinical scores from SHAP would therefore allow a unified approach to translating ML models into clinical scores.

Step 2: Feature selection

Similar to Step 1 of the classical clinical development process, features can be handpicked by clinicians or chosen with a search algorithm. However, as SHAP provides a consistent approach across all ML methods to rank features, top features from SHAP's global feature explanations will be selected for clinical score development.

Step 3: Feature value categorisation

Similar to Step 2 of the classical clinical score development process, we can perform feature value categorisation based on clinical heuristics, GAM or quantiles. However, with the application of SHAP, we instead propose to derive these categories c_{jk} from SHAP values ϕ_j and feature values x_j. We choose this approach because ϕ_j measures the impact to model output $f(x)$ with respect to various feature values x_j. Thus, by creating categories with ϕ_j, the clinical score will have more accurate representations of the underlying model's belief. To illustrate how feature value categorisation can be achieved with this proposed approach, we plot the partial dependence plot of ϕ_j against x_j and steps 3(i) to 3(iii) in Figure 1. In Step 3(i), a smoothing function such as a cubic spline S (blue line) can be fitted over the SHAP scatter points (black dots). This was characterised as an explainability curve by (Ong, 2021). In Step 3(ii), intersections Z can be determined by intersecting the line $\phi_j = 0$ (dash line) with the spline S. Finally in Step 3(iii), feature value categories C can thus be determined from intersections Z. Hence, x_j can be transformed to c_{jk}. As positive SHAP values imply increased risk and vice versa, this form of categorisation classifies ranges of feature values broadly according to their propensity to increase or decrease risk based on signs of the SHAP values. In the example shown in Figure 1, three such categories, represented by c_{j1}, c_{j2} and c_{j3}, were created. The above procedure is summarized in Algorithm 1.

A more complex approach of determining intersections Z and categories C by considering the concentration of SHAP scatter points or the gradient of the fitted spline can be adopted. However, this comes at the cost of increased number of Z and C, thus compromising on the conciseness tenet of a clinical score.

Step 4: Create clinical scoring table with feature aggregate SHAP values

As the beta coefficients, B_{jk}, were instrumental in the determination of integer scores in the classical clinical score development process, we discuss how the analogous variable ϕ_j can be used to determine integer clinical scores in this step. We have stated previously in SHAP's key characteristic (b) that ϕ_j is analogous to but not equivalent to B_{jk}. This is because:

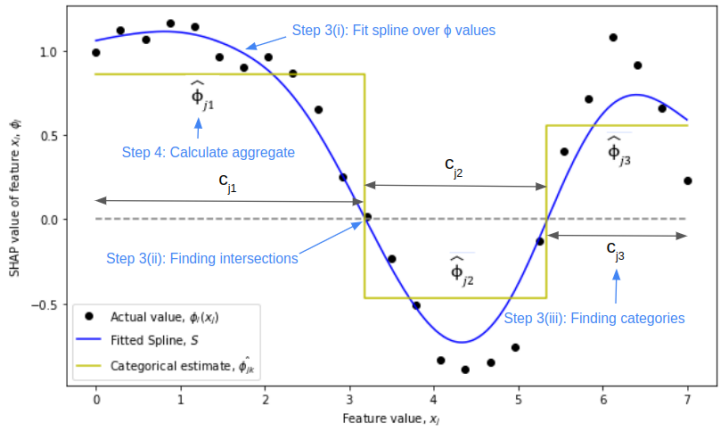


Figure 1: Partial dependence plot of SHAP value, ϕ_j, against feature value, x_j, annotated with Steps 3(i) to 3(iii) and Step 4(i).

Algorithm 1 Pseudocode for Steps 3 and 4(i)

```
/* i and j subscripts represent row and column numbers of input x          */
for j ← 1 to M do
    S ← Fit spline for φ_j against x_j over all i = 1 to i = N inputs;      // Step 3(i)
    Z ← {z_jk| K − 1 intersections between spline S and line φ_j = 0};     // Step 3(ii)
    C ← {c_jk| K categories derived from Z};                               // Step 3(iii)

    /* Step 4(i): φ value aggregation in K categories for jth feature     */
    for k ← 1 to K do
        φ_jk ← {φ_ij| lower φ bound of c_jk < φ_ij ≤ upper φ bound of c_jk};
        I ← total number of φ_ij elements in φ_jk;
        i_lower ← element number of smallest element in φ_jk
        i_upper ← element number of largest element in φ_jk
        φ̂_jk ← (Σ_{i_lower}^{i_upper} φ_ij) / I;                           // Mean applied.
    end
end
```

- B_j are constant coefficients of the logistic regression model. On the other hand, ϕ_j is calculated as a function of the original ML model $f(x)$ and the input features x as described by (Lundberg and Lee, 2017).
- The logistic regression model, in Step 3 of classical clinical score development, is trained on categorical features c_{jk} generated from the original data. On the other hand, ϕ_j are interpretations of the model trained on the original input features x.

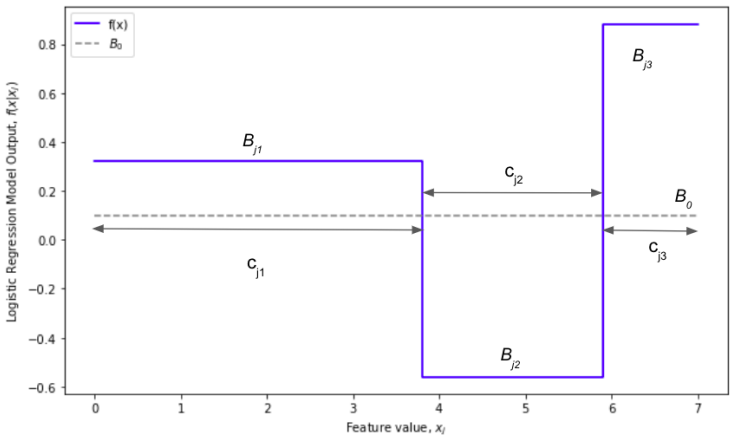


Figure 2: Partial dependence plot of logistic regression model, $f(x_j)$, against feature value, x_j, annotated with beta values, B_{jk}, for categories, c_{jk}.

We plot Figure 2 to illustrate the classical clinical score development's logistic regression model's log odds output $f(c)$ with respect to a single feature's categories c_{jk}. The model's log odds output response consist of step functions, with widths equivalent to the feature categorisation value ranges and heights equivalent to the beta coefficient values B_{jk}. Comparing ϕ_j plotted as a blue line in Figure 1 to B_{jk} plotted as a blue line in Figure 2, we can observe the differences between ϕ_j and B_{jk} as listed above. Hence, we can infer that a key requirement to building a clinical score is the calculation of a constant explanatory variable for each feature category c_{jk}. Therefore, in Step 4(i), we propose aggregating ϕ_j over the categories previously established in Step 3, as $\hat{\phi_{jk}}$. The method for aggregation can be the mean, mode, median or any statistical representation. In Uni-ACS, mean was chosen as the method of aggregation. For example shown in Figure 1, $\hat{\phi_{j1}}$, $\hat{\phi_{j2}}$, $\hat{\phi_{j3}}$ were calculated for categories c_{j1}, c_{j2} and c_{j3} and represented as light yellow line. Step 4(i) is also described in Algorithm 1.

With Steps 3 and 4(i), we can rewrite the SHAP model in (2) as (3) to show a higher level of equivalency with (1) as follows: Like B coefficients, $\hat{\phi}$, set of all $\hat{\phi_{jk}}$, are now constant coefficients of derived categories C, set of all c_{jk}. This give us

$$f(x) = g(x') \approx h(c) = \phi_0 + \sum_{j=1}^{M} \sum_{k=1}^{K_j} \hat{\phi_{jk}} c_{jk}, \tag{3}$$

where $h(c)$ is an approximation of the explanation model $g(x')$, c_{jk} represents the categorized input values x and $\hat{\phi_{jk}}$ is an aggregated SHAP value representative of the respective categories.

Therefore, we can perform Step 4(ii), which is the same approach to developing integer scores as described in Step 4 of the classical clinical score development. The method is described in Algorithm 2 and the following descriptive sub-steps:

- Map a unit integer score in the scoring table to the lowest absolute $\hat{\phi}$ value.
- $\hat{\phi}$ values are divided by this lowest absolute $\hat{\phi}$ value and rounded to the nearest whole integer number.
- To remove negative scoring, the absolute value of the most negative score is added to all scores.

Algorithm 2 Pseudocode for Steps 4(ii)

```
/* Step 4(ii):  Calculation of integer scores                    */
```
$\hat{\phi} = \{\hat{\phi_{jk}} | 1 \leq j \leq M, 1 \leq k_j \leq K_j\}$ // jth feature has K_j categories

$\hat{\phi_{unit}} \leftarrow min(|\hat{\phi}|)$ // Find min non-negative ϕ

$\psi \leftarrow round(\frac{\hat{\phi}}{\hat{\phi_{unit}}})$ // Round scores to integer

$\psi \leftarrow \psi - min(\psi)$ // Ensure non-negative scores

Step 5: Create clinical score to risk mapping table

Contrary to logistic regression models, ML models' $f(x)$ outputs are not modelled to directly estimate class probabilities. However, component B of the clinical score requires the generation of risk estimates for final aggregated integer scores. Hence, we suggest plotting calibration plots of the original underlying ML model to evaluate the clinical score's probabilistic outputs. This plot can be seen in the Appendix's Supplementary Figure 1. Additionally, the plot can be used to determine if further calibration is required. Appropriate calibrators to choose from include sigmoid or isotonic regressors as described in (Niculescu-Mizil and Caruana, 2005). After calibration is completed as necessary, a post calibration curve can be plotted, as shown in the Appendix's Supplementary Figure 2, to ascertain that model probabilities roughly match fraction of positive labels.

Once we are satisfied that $f(x)$ models have well calibrated probabilities, we calculate the risk for the discrete integer scores between 0 and maximum possible score. Component B of the clinical score would thus be completed.

4. Study Design

4.1. Experiments

To test the hypothesis of the Uni-ACS method (refer to Section 3.2), we implemented the Uni-ACS method and compared it against state of the art clinical scoring methodology, Risk SLIM, and the classical clinical score development method (refer to section 3.1) as a baseline. The following key areas of comparisons were made in our experiments:

1. Quantitative comparisons of model predictive performance, such as Area Under Receiver Operating Curve (AUROC), Area Under Precision Recall Curve (AUPRC) and accuracy between Uni-ACS and baseline method.

2. Quantitative comparisons of model predictive performance, such as AUROC and AUPRC, of ML models prior to and post application of Uni-ACS.

3. Qualitative comparisons of the interpretations of clinical scores derived from Uni-ACS against current clinical literature. This is to ensure that derived scores have practical clinical application.

4.2. Cohort

We evaluated Uni-ACS and classical clinical score development on the clinical problem of predicting mortality and morbidity in sepsis because it is a health problem of global concern, given its high worldwide incidence of 48.9 million and mortality of 11 million accounting for 20% of the world's death in 2017 (Rudd et al., 2020).

Selection criteria

Using landmark sepsis studies (Seymour et al., 2016) as reference, 2 cohorts of patients treated at Beth Israel Deaconess Medical Center from 2001 to 2012 and from 2008 to 2019 for sepsis based on ICD-9 diagnosis codes, who were older than 18 years old at point of admission, were included in the study. Outcomes of interest were: death within 28 days from date of admission (1st cohort); death within 28 days from date of admission and transfer to Intensive Care Unit (ICU) within 2 days from date of admission (2nd cohort).

Feature Choices

Chosen features included demographics, past background medical history, laboratory data such as full blood count, inflammatory markers, renal function test, liver function test, iron panel, cardiac markers, thyroid function tests, arterial blood gas, microbiological investigations, Electro-Cardiogram (ECG) and Chest X-Ray (CXR) findings.

Data Extraction and Post Processing

Data for the 2 cohorts were extracted from the Medical Information Mart for Intensive Care (MIMIC) III (Johnson et al., 2016) and MIMIC IV (Johnson et al., 2020) respectively. Non numerical features indicating feature presence such as Gender (Male or Female) and CXR finding of pneumonia (Yes or No) were converted into binary features. Missing data were imputed with Multiple Imputation with Chained Equations (MICE).

5. Results

Descriptive statistics for the 2 cohorts can be found in the Appendix B's Supplementary Tables 1, 2 and 3. The best set of hyperparameters for all models, including classical clinical score and Uni-ACS, found from 5 fold cross validation repeated 10 times applied on the training dataset can be found in Appendix B's Supplementary Table 4. Details of how the top clinical features were selected for construction of the clinical score in Appendix C.

As part of quantitative analysis, predictive performance of LR and the ML models applied on the held out test set can be found in Table 1. Predictive performance of the baseline, Risk SLIM, classical clinical score, and the respective Uni-ACS converted clinical

Table 1: Comparison of original model performance

Dataset	Method	AUROC	AUPRC	Accuracy
MIMICIII (Mortality)	Logistic Regression (LR)	0.816 (±0.007)	0.754 (±0.009)	0.769 (±0.007)
	Gradient Boosting (GB)	0.866 (±0.005)	0.823 (±0.007)	0.789 (±0.006)
	Random Forest (RF)	0.851 (±0.008)	0.800 (±0.010)	0.764 (±0.006)
	Neural Networks (NN)	0.787 (±0.010)	0.704 (±0.009)	0.735 (±0.007)
MIMICIV (Mortality)	Logistic Regression (LR)	0.878 (±0.013)	0.427 (±0.029)	0.930 (±0.003)
	Gradient Boosting (GB)	0.899 (±0.011)	0.479 (±0.029)	0.933 (±0.004)
	Random Forest (RF)	0.900 (±0.010)	0.477 (±0.032)	0.928 (±0.010)
	Neural Networks (NN)	0.803 (±0.012)	0.422 (±0.051)	0.889 (±0.025)
MIMICIV (ICU)	Logistic Regression (LR)	0.898 (±0.009)	0.818 (±0.014)	0.865 (±0.007)
	Gradient Boosting (GB)	0.910 (±0.008)	0.837 (±0.013)	0.870 (±0.007)
	Random Forest (RF)	0.902 (±0.008)	0.824 (±0.013)	0.856 (±0.006)
	Neural Networks (NN)	0.799 (± 0.009)	0.779 (±0.013)	0.801 (±0.008)

Table 2: Comparison of clinical score performance

Dataset	Method	AUROC	AUPRC	AUROC drop	AUPRC drop
MIMICIII (Mortality)	Baseline	0.670 (±0.009)	0.536 (±0.010)	17.9% (±1.8%)	28.9% (±2.9%)
	Risk SLIM	0.732 (±0.010)	0.645 (±0.013)	N.A.	N.A.
	Uni-ACS on LR	0.694 (±0.009)	0.563 (±0.012)	15.0% (±0.8%)	25.3% (±3.8%)
	Uni-ACS on GB	0.745 (±0.010)	0.646 (±0.017)	13.9% (±1.7%)	20.8% (±3.4%)
	Uni-ACS on RF	0.750 (±0.009)	0.658 (±0.012)	11.9% (±1.8%)	17.8% (±2.4%)
	Uni-ACS on NN	0.590 (± 0.009)	0.486 (±0.015)	25.0% (±3.6%)	31.0% (±4.6%)
MIMICIV (Mortality)	Baseline	0.785 (±0.016)	0.250 (±0.021)	10.6% (±3.3%)	41.5% (±11.7%)
	Risk SLIM	0.801 (±0.015)	0.228 (±0.021)	N.A.	N.A.
	Uni-ACS on LR	0.821 (±0.016)	0.283 (±0.024)	6.55% (±3.24%)	33.7% (±12.4%)
	Uni-ACS on GB	0.867 (±0.014)	0.364 (±0.026)	3.56% (±2.78%)	24.0% (±11.5%)
	Uni-ACS on RF	0.862 (±0.013)	0.373 (±0.033)	4.22% (±2.67%)	21.8% (±13.6%)
	Uni-ACS on NN	0.612 (±0.015)	0.288 (±0.042)	23.8% (±3.36%)	31.8% (±22.0%)
MIMICIV (ICU)	Baseline	0.846 (±0.007)	0.729 (±0.015)	5.79% (±1.78)	10.8% (±3.55)
	Risk SLIM	0.829 (±0.011)	0.690 (±0.016)	N.A.	N.A.
	Uni-ACS on LR	0.857 (±0.010)	0.747 (±0.017)	4.57% (±2.12%)	8.68% (±3.79%)
	Uni-ACS on GB	0.880 (±0.008)	0.775 (±0.014)	3.30% (±1.76%)	7.41% (±3.23%)
	Uni-ACS on RF	0.880 (±0.009)	0.772 (±0.013)	2.44% (±1.88%)	6.31% (±3.16%)
	Uni-ACS on NN	0.601 (±0.010)	0.536 (±0.021)	24.8% (±2.38%)	31.2% (±4.36%)

scores applied to the held out test set can be found in Table 2. As Risk SLIM was not built with any underlying LR or ML model, no performance drop value would be provided.

As part of qualitative analysis, we produce the output clinical scores for prediction of mortality in septic patients from MIMIC III: (a) Clinical score derived from Uni-ACS applied on GB in Table 3 and Table 4. (b) Clinical scores derived from baseline and other ML models applied on the MIMIC III dataset can be found in Appendix D. The output clinical scores from MIMIC IV analysis can be found on our Github repository.

6. Discussion

6.1. Advantages of Uni-ACS

Performance of Uni-ACS applied on ML models

Uni-ACS clinical scores applied to ensemble decision tree methods RF and GB consistently produced higher predictive performances compared to Risk SLIM and baseline classical clinical scores, across both MIMIC III and MIMIC IV cohorts, for the prediction of mortality and ICU admission, as shown in Table 2. This is possible because a complex ML model can be used as a base model for construction of the clinical score. Such a procedure is made possible by: (1) Step 1 of Uni-ACS, where we take a "model-first" approach to the clinical score development, with SHAP as the interpretation methodology to derive explainable

Table 3: Uni-ACS clinical scoring table

Risk factor	Score
Length Of Stay (LOS) $<$ 9 days	13
RDW $\geqslant$ 15.5 (%)	7
ICU length of stay $\geqslant$ 6 days	6
Age $\geqslant$ 75 years old	5
Inotropes prescribed	5
Lactate Dehydrogenase $\geqslant$ 208 (U/L)	3
Haptoglobin $<$ 158.8	3
Phosphate $\geqslant$ 3.45 (mmol/L)	2
ICU stays in 1 admission $>=$ 1	5
Albumin $<$ 2.73 (g/dL)	2

Table 4: Uni-ACS score to risk mapping

Score	Mortality Risk
23	10%
27	20%
31	30%
34	40%
39	50%
44	60%
48	70%
52	80%
54	90%

variables. (2) Steps 3 and 4 of Uni-ACS, where categorical conversion of input features x and their associated categorical aggregate SHAP values $\hat{\phi}$ can be determined, for the purposes of clinical score development.

Additionally, the baseline classical clinical score and the Uni-ACS clinical score derived from logistic regression model had marginal differences in predictive performance. This is a reassuring outcome as it shows that Uni-ACS applied on the logistic regression model can reproduce the results of the baseline method, despite the changes to algorithm heuristics as described in section 3.2.

Reduction in original ML model's performance

LR and ML models' performances dropped after conversion to clinical scores. This is expected as classical and Uni-ACS clinical score development reduces model complexity in the following ways: (1) For both classical and Uni-ACS clinical score development, only top features of models are selected for clinical score conversion. (2) In classical clinical score development Step 2 and Uni-ACS Step 3, input features x_{ij} are converted into categorical features c_{ijk}, albeit through different algorithmic heuristics. (3) In Uni-ACS Step 4(i), further complexity is lost through the conversion of ϕ_j into $\hat{\phi_{jk}}$ for categorised inputs c_{jk}. (4) In classical clinical score development Step 4 and Uni-ACS Step 4(ii), beta coefficients B_{jk} and SHAP value aggregates $\hat{\phi_{jk}}$ are converted into non-zero integer clinical scores. The process includes division and rounding of numbers to integers, which precipitates further loss of model information. (5) Possible scores of a clinical score take on a finite set number of integer values. On the other hand, the original model's output can take on a theoretically infinite number of possible values from 0 to 1 or from $-\infty$ to $+\infty$, depending on whether probability or log of odds is the output.

Excluding Uni-ACS applied to NN, we assert that Uni-ACS clinical scores' performance drops post conversion were reasonable, when compared with the baseline method. For instance, across all datasets and outcomes, Uni-ACS clinical scores' AUROC drop were in the range of 2.44% to 15.0%, compared to baseline of 5.79% to 17.9%. The AUPRC drop were in the range of 6.31% to 33.7%, compared to baseline of 10.8% to 41.5%. Furthermore, despite the drop, the performance of Uni-ACS on the various models were still comparable to the performance of direct optimization methods, such as Risk SLIM.

For the unique case of Uni-ACS applied on NN where the performance drop was the worst (AUROC drop of 23.8% to 25.0% and AUPRC drop of 31.0% to 31.8%), the SHAP values were calculated with Kernel SHAP. As this is an exact computation approach, the

number of data samples used to compute the SHAP explanations of the NN were reduced to save compute resource and decrease compute time. Nevertheless, this shows a proof of concept that Uni-ACS is model agnostic and can be applied to NN models.

Uni-ACS derived clinical score correlates well with clinical knowledge

We further made a qualitative comparison of Uni-ACS clinical score derived from the GB model derived from the MIMIC III dataset (shown in Tables 3 and 4), with respect to current clinical literature. Firstly, age, number of recurrent ICU admissions, prolonged ICU LOS and prolonged hospital LOS were known predictors of adverse outcomes in sepsis (Yang et al., 2010). Secondly, while heart rate and blood pressure (variables found in the APACHE II (Knaus et al., 1985) and SIRS (Bone et al., 1992) scores) were not seen in the scoring table, the number of inotropes (medications used to moderate heart rate and increase blood pressure) was found to be a strong indicator of adverse outcome. In other words, the clinical score concurred that patients with poor hemodynamic function have a higher risk of adverse outcome. Finally, although inflammatory markers such as white blood cells, C-reactive protein and procalcitonin were not seen in the scoring table, alternate blood investigation markers indicating severe sepsis such as Lactate Dehydrogenase (marker of organ hypoperfusion during sepsis, studied by (Lu et al., 2018)) and haptoglobin (protective in high levels during sepsis, studied by (Janz et al., 2013)) were found to have significant contributions toward adverse outcome.

Consistency between Uni-ACS score and underlying ML model

To prove empirically that Uni-ACS derived clinical scores' explanations were consistent with underlying ML models' explanations, we first plot the original ML model's SHAP response, ϕ_j, against feature values, x_j, as a blue line in Figure 3. We further overlay the plot of the clinical scores, $\hat{\phi_j}$, against feature values, x_j, as a yellow line. We make this plot for each feature of the Uni-ACS clinical score, shown in Table 3. We can now observe that the clinical scores' explanations were consistent with the models' explanations. For instance, the most important feature, LOS, represented by the first panel in Figure 3 and the first line in Table 3, the original ML model's belief of a drop in risk attribution, ϕ_j, at LOS of 9 to 10 days was captured as "LOS $<$ 9 days for 13 points" in the clinical score. Similar patterns of consistency could be observed for the other 9 panels in Figure 3, corresponding to the other 9 features in Table 3.

Understanding risk from Uni-ACS and deriving possible discrete treatments

With empirical proof of consistency, clinical scores retain the underlying ML models' understanding of pathology. We can therefore infer disease patterns and determine possible discrete treatment. For instance, Albumin, represented by the last panel in Figure 3, could be seen to have a transition from high to low risk from 2 to 3 g/dL based on the underlying ML model's SHAP values, ϕ_j. This was reflected in the Uni-ACS clinical score, in the last line of Table 3, as a score of 3 for Albumin values higher than 2.73g/dL. Clinicians may use this threshold to determine determine hypoalbuminemia as a driving cause for adverse outcome and initiate treatment to replenish albumin reserves in order to reduce risk.

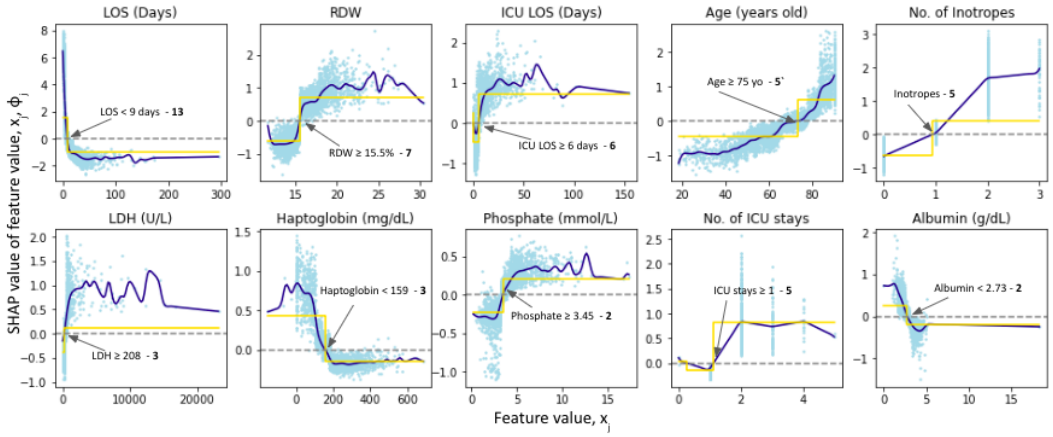


Figure 3: Plots of (A) SHAP outputs of original ML model, ϕ_j, against feature values, x_j (light blue dots for original SHAP points, blue line for best fit of SHAP points) and, (B) Clinical score, $\hat{\phi}_j$, against feature values, x_j (yellow line).

6.2. Limitations

Firstly, Uni-ACS was applied on ML models trained on 2 patient cohorts' data for the prediction of mortality and morbidity in sepsis. Ideally, we would like to show that Uni-ACS can be generalised by testing it with several different types of diseases. Secondly, although we strive to completely automate the clinical scoring conversion process, there are edge cases and few points of possible failure, which might require manual review of data and tuning of specific method parameters. For instance, default spline fitting parameters might overfit the ϕ_j values in Step 3. This might lead to unrealistic number of intersections Z and categories C. To circumvent the problem, the ϕ_j values and the fitted spline would have to be plotted and visually inspected to adjust spline parameters. Another instance of automation failure is when the minimum absolute aggregate ϕ_{jk} value in Step 4 is too small relative to the other aggregate ϕ values. This will result in a large clinical scoring range, which increases the complexity of the clinical score. While the method is by default opinionated and automatic, users can opt for a manual approach to fine tuning the clinical score.

6.3. Future work

While patient risk stratification clinical scores were developed on binary disease outcomes in this paper, clinical scores can also be developed to estimate patient survival over longitudinal periods with the proportional hazards model using Cox regression. (Kvamme et al., 2019) and (Spooner et al., 2020) showed ensemble tree based and neural network approaches of using ML to perform Cox Regression. By making modifications to Uni-ACS, we can theoretically extend Uni-ACS to estimate clinical scores for survival analysis. Finally, while we will make the software open source through Github, we hope to further refine the algorithm, resolve the aforementioned limitations and develop it into a full package.

7. Conclusion

We showed that Uni-ACS is a model agnostic approach to converting ML models into clinical scores, a tool that is well established within the medical community. Additionally, we demonstrated that Uni-ACS applied on ML models can retain a reasonable portion of the underlying model's predictive performance. Our experiments also showed that Uni-ACS applied on ML models have superior predictive performance compared to the classical clinical score development approach. Finally, Uni-ACS delivers a consistent frame of reference for clinicians when doing patient risk stratification both at bedside and at machine with clinical decision support.

References

Subhi J Al'Aref, Gabriel Maliakal, Gurpreet Singh, Alexander R van Rosendael, Xiaoyue Ma, Zhuoran Xu, Omar Al Hussein Alawamlh, Benjamin Lee, Mohit Pandey, Stephan Achenbach, et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the confirm registry. *European heart journal*, 41 (3):359–367, 2020.

Pedro Roig Aparicio, Ričards Marcinkevičs, Patricia Reis Wolfertstetter, Sven Wellmann, Christian Knorr, and Julia E Vogt. Learning medical risk scores for pediatric appendicitis. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1507–1512. IEEE, 2021.

Irantzu Barrio, Inmaculada Arostegui, José M Quintana, et al. Use of generalised additive models to categorise continuous variables in clinical prediction. *BMC medical research methodology*, 13(1):1–13, 2013.

Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, Roland MH Schein, and William J Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655, 1992.

Fabrizio Buffolo, Jacopo Burrello, Alessio Burrello, Daniel Heinrich, Christian Adolf, Lisa Marie Müller, Rusi Chen, Vittorio Forestiero, Elisa Sconfienza, Martina Tetti, et al. Clinical score and machine learning-based model to predict diagnosis of primary aldosteronism in arterial hypertension. *Hypertension*, 78(5):1595–1604, 2021.

TJ Cole. Algorithm as 281: scaling and rounding regression coefficients to integers. *Applied statistics*, pages 261–268, 1993.

Keith AA Fox, Omar H Dabbous, Robert J Goldberg, Karen S Pieper, Kim A Eagle, Frans Van de Werf, Álvaro Avezum, Shaun G Goodman, Marcus D Flather, Frederick A Anderson, et al. Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (grace). *bmj*, 333(7578):1091, 2006.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

David R Janz, Julie A Bastarache, Gillian Sills, Nancy Wickersham, Addison K May, Gordon R Bernard, and Lorraine B Ware. Association between haptoglobin, hemopexin and mortality in adults with sepsis. *Critical care*, 17(6):1–8, 2013.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and R Mark IV. Mimic-iv (version 0.4). *PhysioNet*, 2020.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Patrick S Kamath, Russell H Wiesner, Michael Malinchoc, Walter Kremers, Terry M Therneau, Catherine L Kosberg, Gennaro D'Amico, E Rolland Dickson, and W Ray Kim. A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2): 464–470, 2001.

William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

Stephen R Knight, Antonia Ho, Riinu Pius, Iain Buchan, Gail Carson, Thomas M Drake, Jake Dunning, Cameron J Fairfield, Carrol Gamble, Christopher A Green, et al. Risk stratification of patients admitted to hospital with covid-19 using the isaric who clinical characterisation protocol: development and validation of the 4c mortality score. *bmj*, 370, 2020.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *arXiv preprint arXiv:1907.00825*, 2019.

Gregory YH Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A Lane, and Harry JGM Crijns. 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*, 137(2):263–272, 2010.

Jun Lu, Zhonghong Wei, Hua Jiang, Lu Cheng, Qiuhua Chen, Mingqi Chen, Jing Yan, and Zhiguang Sun. Lactate dehydrogenase is associated with 28-day mortality in patients with sepsis: a retrospective observational study. *Journal of Surgical Research*, 228:314–321, 2018.

Shuyu Lu, Ruoyu Chen, Wei Wei, and Xinghua Lu. Understanding heart-failure patients ehr clinical features via shap interpretation of tree-based machine learning model predictions. *arXiv preprint arXiv:2103.11254*, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

David A Morrow, Elliott M Antman, Andrew Charlesworth, Richard Cairns, Sabina A Murphy, James A de Lemos, Robert P Giugliano, Carolyn H McCabe, and Eugene Braunwald. Timi risk score for st-elevation myocardial infarction: a convenient, bedside, clinical score for risk assessment at presentation: an intravenous npa for treatment of infarcting myocardium early ii trial substudy. *Circulation*, 102(17):2031–2037, 2000.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

Ming Lun Ong. Developing a holistic explainable machine learning framework: Data science applications in healthcare. *Scholar Bank at National University of Singapore*, 2021.

Ron Pisters, Deirdre A Lane, Robby Nieuwlaat, Cees B De Vos, Harry JGM Crijns, and Gregory YH Lip. 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*, 138(5):1093–1100, 2010.

Hans-Jürgen Profitlich and Daniel Sonntag. Interactivity and transparency in medical risk assessment with supersparse linear integer models. *arXiv preprint arXiv:1911.12119*, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211, 2020.

Christopher W Seymour, Vincent X Liu, Theodore J Iwashyna, Frank M Brunkhorst, Thomas D Rea, André Scherag, Gordon Rubenfeld, Jeremy M Kahn, Manu Shankar-Hari, Mervyn Singer, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):762–774, 2016.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

Annette Spooner, Emily Chen, Arcot Sowmya, Perminder Sachdev, Nicole A Kochan, Julian Trollor, and Henry Brodaty. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific reports*, 10 (1):1–10, 2020.

Po-Yu Tseng, Yi-Ting Chen, Chuen-Heng Wang, Kuan-Ming Chiu, Yu-Sen Peng, Shih-Ping Hsu, Kang-Lung Chen, Chih-Yu Yang, and Oscar Kuang-Sheng Lee. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Critical Care*, 24(1):1–13, 2020.

Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.

Berk Ustun and Cynthia Rudin. Optimized risk scores. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1125–1134, 2017.

Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *J. Mach. Learn. Res.*, 20 (150):1–75, 2019.

Graham Walker and Joe Habboushe. About us, 2022. URL https://www.mdcalc.com/about-us.

Caroline Wang, Bin Han, Bhrij Patel, and Cynthia Rudin. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, pages 1–63, 2022.

PhilipS Wells, Jack Hirsh, DavidR Anderson, AnthonyW A Lensing, Gary Foster, Clive Kearon, Jeffrey Weitz, Robert D'Ovidio, Alberto Cogo, Paolo Prandoni, et al. Accuracy of clinical assessment of deep-vein thrombosis. *The Lancet*, 345(8961):1326–1330, 1995.

Feng Xie, Bibhas Chakraborty, Marcus Eng Hock Ong, Benjamin Alan Goldstein, Nan Liu, et al. Autoscore: A machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR medical informatics*, 8(10):e21798, 2020.

Yong Yang, Kok Soong Yang, Yin Maw Hsann, Vincent Lim, and Biauw Chi Ong. The effect of comorbidity and age on hospital mortality and length of stay in patients with sepsis. *Journal of critical care*, 25(3):398–405, 2010.

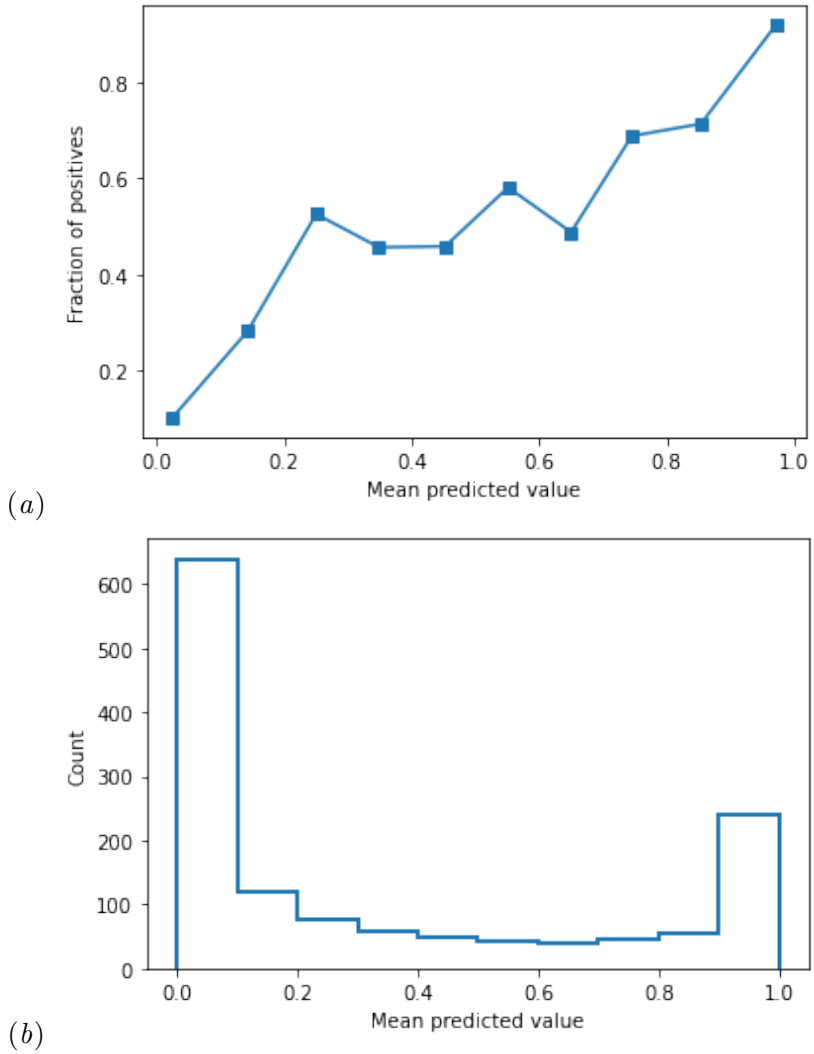
Appendix A. Supplementary material for Methods

S. Table 1: TIMI's clinical scoring table

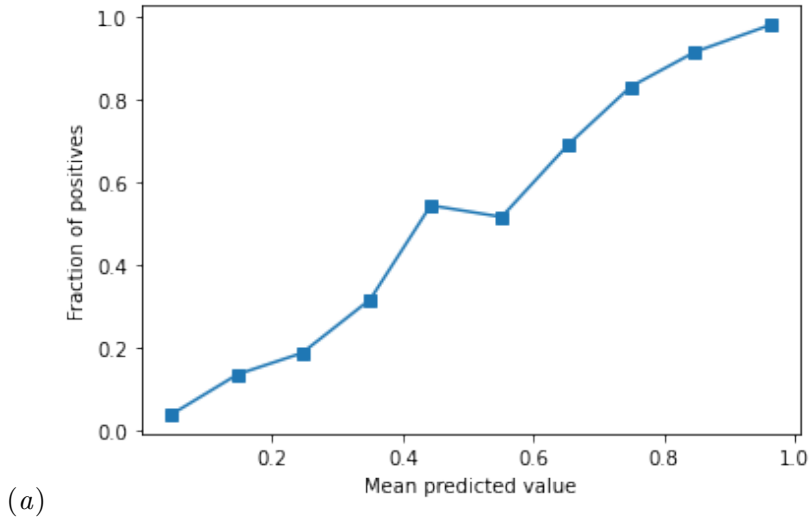
Risk factor	**Score**
Age 65 to 74yo	2
Age $\geqslant$ 75yo	3
DM, HTN or Angina	1
SBP $<$ 100	3
HR $>$ 100	2
Killip II-IV	2
Weigh $<$ 67kg	1
Anterior STE or LBBB	1
Time to Rx	1

S. Table 2: TIMI's score to risk mapping

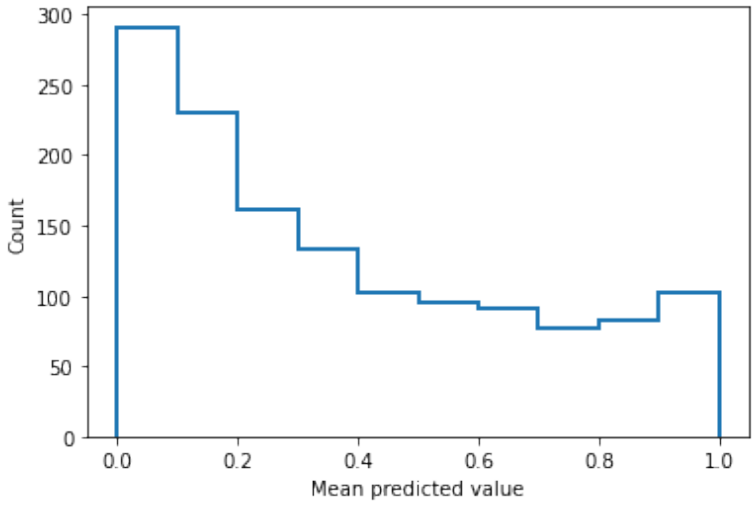
Score	**Mortality Risk**
0	0.8%
1	1.6%
2	2.2%
3	4.4%
4	7.3%
5	12.4%
6	16.1%
7	23.4%
8	26.8%
>8	34.9%



S. Fig. 1: Pre-calibration plots: (a) Fraction of positive label against predicted probability, (b) Counts of positive label against predicted probability.



(*a*)



(*b*)

S. Fig. 2: Post-calibration plots: (a) Fraction of positive label against predicted probability, (b) Counts of positive lable against predicted probability.

Appendix B. Supplementary material for Results

S. Table 3: Descriptive Statistics of MIMIC III Sepsis Mortality Cohort

Feature	Total (n=4555)	Alive or Death after 28 days (n=2820)	Death within 28 days (n=1735)	P-value
Demographics				
Age	67.0 (16.2)	65.1 (16.7)	70.1 (15.0)	< 0.001
Gender	2526 (55.5%)	1530 (54.3%)	996 (57.4%)	0.041
Co-morbids				
Hypertension	1660 (37.5%)	1053 (37.6%)	607 (37.3%)	0.896
Hyperlipidemia	178 (4.0%)	137 (4.9%)	41 (2.5%)	< 0.001
Diabetes	190 (6.5%)	191 (6.8%)	99 (6.1%)	0.379
Hospital parameters				
LOS	15.3 (17.1)	15.9 (16.9)	14.3 (17.5)	0.002
ICU LOS	7.2 (9.8)	6.9 (9.9)	7.8 (9.7)	0.003
Full Blood Count				
Hemoglobin	11.3 (2.2)	11.4 (2.2)	11.0 (2.2)	< 0.001
Hematocrit	33.9 (6.4)	34.1 (6.2)	33.4 (6.6)	< 0.001
RDW	15.8 (2.4)	15.4 (2.2)	16.6 (2.6)	< 0.001
Inflammatory Markers				
White blood cells	14.2 (12.0)	14.1 (11.6)	14.4 (12.7)	0.527
Neutrophils (%)	77.6 (17.5)	78.1 (16.5)	76.7 (18.9)	0.009
Lymphocyte (%)	10.8 (11.8)	10.7 (11.6)	10.9 (12.2)	0.601
Basophils (%)	0.2 (0.4)	0.2 (0.5)	0.2 (0.4)	0.263
Eosinophils (%)	0.9 (2.4)	0.8 (2.0)	0.9 (2.9)	0.346
Monocytes (%)	4.6 (4.7)	4.5 (4.7)	4.7 (4.7)	0.128
C-Reactive Protein	108.4 (97.3)	102.4 (95.4)	124.9 (100.9)	0.009
Coagulopathy				
Platelet	242.5 (151.4)	251.4 (151.5)	227.3 (150.0)	< 0.001
PT	18.0 (11.1)	17.1 (9.9)	19.5 (12.7)	< 0.001
PTT	36.5 (19.7)	34.6 (17.4)	39.6 (22.6)	< 0.001
D-Dimer	4727.0 (4519.6)	4137.8 (4152.0)	5291.7 (4784.4)	0.001
Fibrinogen	443.4 (235.8)	481.1 (231.8)	383.6 (229.7)	< 0.001
Fibrin	47.0 (135.8)	43.8 (137.6)	49.6 (134.5)	0.538
Renal Panel				
Sodium	137.5 (6.6)	137.5 (6.0)	137.4 (7.4)	0.425
Potassium	4.4 (1.0)	4.3 (0.9)	4.5 (1.0)	< 0.001
Creatinine	2.2 (12.1)	1.9 (1.8)	2.7 (19.4)	0.103
Calcium	8.2 (1.0)	8.2 (1.0)	8.3 (1.1)	0.007
Magnesium	1.9 (0.5)	1.8 (0.4)	2.0 (0.5)	< 0.001
Phosphate	3.8 (1.7)	3.5 (1.5)	4.2 (1.8)	< 0.001
Arterial Blood Gas				
pH	6.3 (1.1)	6.3 (1.1)	6.4 (1.0)	0.001
pCO2	41.5 (14.4)	41.4 (13.5)	41.5 (15.5)	0.825
pO2	131.8 (103.7)	131.0 (101.2)	132.8 (107.1)	0.584
ECG				
STEMI	293 (6.9%)	170 (6.5%)	123 (7.8%)	0.126
AF	1154 (27.4%)	598 (22.7%)	556 (35.0%)	< 0.001
CXR				
Pneumonia findings	2507 (71.5)	1514 (69.0)	993 (75.7)	< 0.001

S. Table 4: Descriptive Statistics of MIMIC IV Sepsis Mortality Cohort

Feature	Total (n=4555)	Alive or Death after 28 days (n=7155)	Death within 28 days (n=535)	P-value
Demographics				
Age	69.3 (17.2)	68.9 (17.4)	74.7 (14.2)	< 0.001
Gender	3589 (46.7%)	3283 (45.9%)	306 (57.2%)	< 0.001
Caucasian	5360 (69.7%)	5006 (70.0%)	354 (66.2%)	0.073
African	1204 (15.7%)	1130 (15.8%)	74 (13.8%)	0.253
Hispanic	349 (4.5%)	332 (4.6%)	17 (3.2%)	0.144
Other races	623 (8.1%)	572 (8.0%)	51 (9.5%)	0.240
Unknown race	154 (2.0%)	115 (1.6%)	39 (7.3%)	< 0.001
Vital Signs				
Temperature, min	97.3 (6.7)	97.5 (6.1)	95.6 (12.2)	< 0.001
Temperature, max	99.3 (10.7)	99.4 (10.9)	98.2 (8.5)	0.002
RR, min	16.5 (3.0)	16.4 (2.8)	17.7 (4.6)	< 0.001
RR, max	22.3 (6.2)	21.9 (5.9)	26.6 (7.4)	< 0.001
HR, min	78.9 (16.8)	78.3 (16.4)	86.5 (20.9)	< 0.001
HR, max	97.1 (20.9)	96.3 (20.4)	107.4 (24.1)	< 0.001
SBP, min	111.2 (23.2)	112.3 (22.8)	96.5 (23.1)	< 0.001
SBP, max	140.6 (23.8)	141.3 (23.6)	131.1 (24.4)	< 0.001
DBP, min	57.8 (13.5)	58.3 (13.3)	50.7 (14.3)	< 0.001
DBP, max	86.5 (199.1)	87.3 (206.3)	76.1 (17.6)	< 0.001
Co-morbids				
DM (No Complications)	1816 (23.6%)	1668 (23.3%)	148 (27.7%)	0.026
DM (complications)	572 (7.4%)	540 (7.5%)	32 (6.0%)	0.213
IHD	719 (9.3%)	635 (8.9%)	84 (15.7%)	< 0.001
CHF	1760 (22.9%)	1571 (22.0%)	189 (35.3%)	< 0.001
Stroke	617 (8.0%)	550 (7.7%)	67 (12.5%)	< 0.001
AIDS	86 (1.1%)	82 (1.2%)	3 (0.6%)	0.290
Full Blood Count				
Hemoglobin	10.9 (2.0)	10.9 (2.0)	10.1 (2.1)	< 0.001
Hematocrit	33.3 (6.0)	33.4 (5.9)	31.8 (6.5)	< 0.001
RDW	15.0 (2.1)	14.9 (2.0)	16.5 (2.7)	< 0.001
White blood cells	10.4 (8.5)	10.0 (7.4)	14.7 (16.9)	< 0.001
Platelet	232 (123.4)	234 (121.4)	208.8 (145.8)	< 0.001
Biochemistry				
Urea Nitrogen	25.7 (20.4)	24.5 (19.0)	42.0 (29.6)	< 0.001
Glucose	128.6 (64.9)	128.0 (64.3)	137.5 (71.9)	0.003

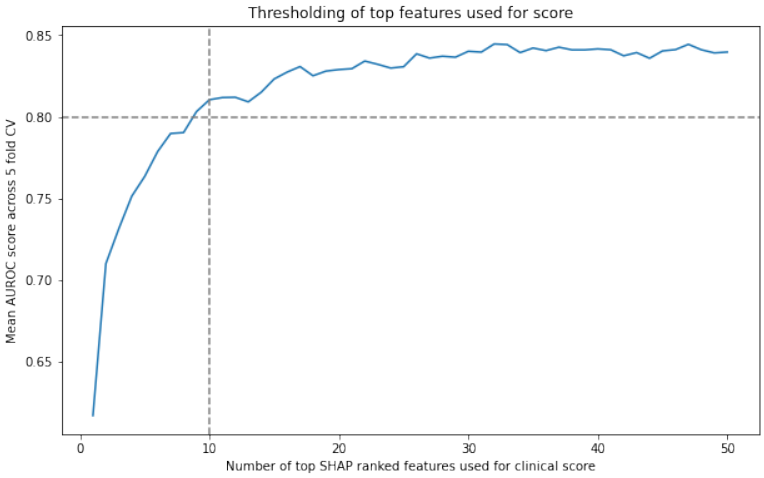
S. Table 5: Descriptive Statistics of MIMIC IV Sepsis ICU Cohort

Feature	Total (n=7690)	No ICU admission (n=5623)	ICU admission (n=2067)	P-value
Demographics				
Age	69.3 (17.2)	69.0 (17.6)	70.1 (16.1)	0.008
Gender	3589 (46.7%)	2530 (45.0%)	1059 (51.2%)	< 0.001
Caucasian	5360 (69.7%)	3951 (70.3%)	1409 (68.2%)	0.081
African	1204 (15.7%)	941 (16.7%)	263 (12.7%)	< 0.001
Hispanic	349 (4.5%)	257 (4.6%)	92 (4.5%)	0.872
Other races	623 (8.1%)	443 (7.9%)	180 (8.7%)	0.256
Unknown race	154 (2.0%)	31 (0.6%)	123 (6.0%)	< 0.001
Vital Signs				
Temperature, min	97.3 (6.7)	97.8 (4.0)	96.2 (11.1)	< 0.001
Temperature, max	99.3 (10.7)	99.4 (11.9)	99.2 (6.6)	0.346
RR, min	16.5 (3.0)	16.2 (2.3)	17.3 (4.3)	< 0.001
RR, max	22.3 (6.2)	20.8 (4.9)	26.2 (7.3)	< 0.001
HR, min	78.9 (16.8)	76.7 (15.1)	84.9 (19.6)	< 0.001
HR, max	97.1 (20.9)	93.5 (18.4)	106.8 (23.9)	< 0.001
SBP, min	111.2 (23.2)	116.2 (21.1)	97.7 (23.1)	< 0.001
SBP, max	140.6 (23.8)	143.3 (22.8)	133.2 (24.8)	< 0.001
DBP, min	57.8 (13.5)	60.1 (12.7)	51.6 (13.7)	< 0.001
DBP, max	86.5 (199.1)	89.2 (232.2)	79.3 (27.2)	0.002
Co-morbids				
DM (No complications)	1816 (23.6%)	1242 (22.1%)	574 (27.8%)	< 0.001
DM (Complications)	178 (4.0%)	137 (4.9%)	41 (2.5%)	< 0.001
IHD	719 (9.3%)	436 (7.8%)	283 (13.7%)	< 0.001
CHF	1760 (22.9%)	1024 (18.2%)	736 (35.6%)	< 0.001
Stroke	617 (8.0%)	387 (6.9%)	230 (11.1%)	< 0.001
AIDS	86 (1.1%)	59 (1.0%)	27 (1.3%)	0.048
Full Blood Count				
Hemoglobin	10.9 (2.0)	11.0 (2.0)	10.4 (2.2)	< 0.001
Hematocrit	33.3 (6.0)	33.7 (5.7)	32.3 (6.5)	< 0.001
RDW	15.0 (2.1)	14.7 (2.0)	15.6 (2.4)	< 0.001
White blood cells	10.4 (8.5)	9.4 (7.5)	13.1 (10.2)	< 0.001
Platelet	232 (123.4)	236.4 (118.8)	220.7 (134.5)	< 0.001
Biochemistry				
Urea Nitrogen	25.7 (20.4)	22.9 (17.2)	33.2 (25.8)	< 0.001
Glucose	128.6 (64.9)	122.8 (56.2)	144.4 (82.0)	< 0.001

S. Table 6: Hyperparameters of the best LR and ML models found from 5 fold cross validation repeated 10 times

Model	**Hyperparameters**	**Software**
Logistic Regression	solver="liblinear", penalty="L1", random_state=7	Scikit-Learn
Gradient Boosting	n_estimators=100, max_depth=5, subsample=1.0, min_samples_split=2, min_samples_leaf=1, criterion="friedman_mse", random_state=7	Scikit-Learn
Random Forest	n_estimators=100, max_depth=7, min_samples_split=2, min_samples_leaf=1, criterion="gini", random_state=7	Scikit-Learn
Neural Networks	Customised MLP NN of 2 layers: 1st layer 113 nodes, 2nd layer 100 nodes, criterion="Binary Cross Entropy Loss", optimizer="Stochastic Gradient Descent"	Pytorch
Classical Clinical Score (Baseline)	top_n_features=10, feature_categorisation_method="GAM" **Generalised Additive Method as proposed by Barrio et al. (2013) is the most automatic feature categorisation method for the classical approach**	N.A.
Uni-ACS applied on ML model	top_n_features=10, feature_categorisation_method="novel" **"novel" refers to the feature categorisation method as proposed in section 3.2**	N.A.

Appendix C. Supplementary material for feature selection results



S. Fig. 3: Feature selection for Uni-ACS clinical score construction. The threshold for most parsimonious set of clinical features was selected at n=10 by determining the miminumum required of fetaures required to create a model with AUROC of 0.8 or higher.

Appendix D. Supplementary material for clinical scores derived from MIMIC III sepsis mortality cohort

S. Table 7: Classical Clinical Scoring

Risk factor	Score
$15.8 \leqslant$ RDW < 30.0 (%)	4
$1034 \leqslant$ LDH < 20933 (U/L)	4
$93.4 \leqslant$ Urea Nitrogen < 204.2 (mg/dL)	3
Lactate $\geqslant 16.0$ (mmol/L)	3
Phosphate $\geqslant 0.674$ (mmol/L)	4
PTT $\geqslant 64$ (s)	3
Yeast present in blood cultures	3
Antifungals prescribed	3
Inotropes prescribed	3
Immunosuppressant prescribed	2

S. Table 8: Classical Clinical Score to Risk

Score	Mortality Risk
1	10%
3	20%
5	30%
7	40%
8	50%
10	60%
11	70%
12	80%
14	90%

S. Table 9: Risk SLIM Clinical Scoring

Risk factor	Score
LDH (U/L)	5
RDW (%)	4
Urea Nitrogen (mg/dL)	3
Lactate	3
Number of inotropes	4
PTT (s)	0
Yeast	0
Antifungals prescribed	0
Phosphate (mg/dL)	0
Immunosuppressant prescribed	0

S. Table 10: Risk SLIM Score to Risk

Score	Mortality Risk
1	12%
2	27%
3	50%
4	73%
5	88%
6	95%

S. Table 11: Uni-ACS LR Clinical Scoring

Risk factor	Score
LOS (days) ≤ 15.5	7
Inotropes > 0	4
Age > 67.3	4
Bicarbonate > 22.7	3
Chloride > 103	4
Albumin ≤ 2.96	2
Sodium ≤ 138	2
ICU LOS > 7.45	3
Bilirubin, Indirect > 1.33	4
RDW > 15.6 (%)	3

S. Table 12: Uni-ACS LR Score to Risk

Score	Mortality Risk
10	10%
16	30%
22	50%
28	70%
34	90%

S. Table 13: Uni-ACS RF Clinical Scoring

Risk factor	**Score**
LOS $\leq$ 7.7 (days)	5
RDW $>$ 15.5 (%)	6
Uric acid (mg/dL) $>$ 6.69	5
Vitamin B12 $>$ 1120 (pg/mL)	5
Haptoglobin $\leq$ 142 (mg/dL)	5
Fibrinogen, Functional $\leq$ 355 (mg/dL)	5
Inotropes $>$ 0	4
Phosphate $>$ 3.81 (mg/dL)	3
Potassium $>$ 4.29 (mmol/L)	3
Age $>$ 66 (years old)	2

S. Table 14: Uni-ACS RF Score to Risk

Score	**Mortality Risk**
8	10%
16	20%
24	30%
31	50%
39	70%
43	90%

S. Table 15: Uni-ACS NN Clinical Scoring

Risk factor	**Score**
Inotropes $\geq$ 2	2
LOS ¡ 5 (days)	1
Immunosuppresant prescribed	1
Hyperlipidemia	1
Nucleated Red Cells found	2
Bilirubin, Indirect $\geq$ 63 (mg/dL)	1
Creatinine $\geq$ 3.2 (mg/dL)	1
Calculated Bicarbonate (mEq/L) ¡ 14.6	1
Ferritin $\geq$ 310 (ng/mL)	1
No. of ICUs $\geq$ 3	1

S. Table 16: Uni-ACS NN Score to Risk

Score	**Mortality Risk**
1	10%
3	30%
6	50%
10	70%
12	90%