MLHC Final Project - Mortality Prediction and Endotype Classification for Sepsis Patients: Insights from the PROVIDE Clinical Trial

Keren Danan [*1] **Gal Bodek** [*1]

1. Introduction

Sepsis is known as a life-threatening organ dysfunction that is caused by a dysregulated host response to bacterial, fungal, or viral infection and is regarded as an imbalanced immune response leading to organ failure. Nowadays, it is the most frequent and leading cause of morbidity and mortality in hospitalized patients. Early diagnosis is necessary to properly manage sepsis, as the initiation of rapid therapy is key to reducing deaths from severe sepsis. Recently, machine learning (ML) is revolutionizing the medical field by offering innovative solutions to various challenges. Concerning sepsis, two major aims remain subjects of interest. The first involves the prediction of patient outcomes in cases of infection, including the assessment of their risk of developing sepsis and mortality. (Chao et al., 2022; Kong et al., 2020). The second is the stratification of sepsis patients into distinct endotypes for the application of precision medicine (Baghela et al., 2022; Scicluna et al., 2017). In this study, we focused on developing a new model to predict mortality in sepsis patients by leveraging a high quality data sourced from a recently performed clinical trial called PROVIDE (a Personalized Randomized trial Of Validation and restoration of Immune Dysfunction in severE infections and Sepsis) (Leventogiannis et al., 2022). During that clinical trial, clinical parameters of patients were captured and proteomics measurements were obtained using the Olink Proteomics platform. Employing the key features extracted from the prediction model, which represent the factors that affected the most the risk for mortality, we aimed to get a deeper understanding to classify sepsis patients, thus identifying sepsis endotypes. Due to the complex composition of the data, incorporating both clinical and in-depth biological information, we anticipated that this study could yield novel insights with relevance to sepsis research.

2. Cohort Description

The data used in this study was provided by the PROVIDE clinical trial, which is a double-blind, double-dummy randomized clinical trial that took place between November 2017 and December 2019 in 14 study sites in Greece (9 Intensive Care Units and 5 departments of Internal Medicine). The data contains 240 patients and two stages: (i) A screening/observation stage where various variables were recorded daily for a duration of 28 days, including the co-morbidities, blood cell count, coagulation tests, renal and hepatic biochemistry, arterial blood gases, culture results, and overall survival and (ii) An intervention stage where 36 patients diagnosed with septic shock and overactivation of immune cells called MALS (macrophage activation-like syndrome) or suppression of the immune system called immunoparalysis were randomized into adjunctive treatment either with personalized immunotherapy or placebo. The data consists of 240 patients and includes their clinical information along with Olink proteomics measurements. The proteomics data comprises protein expression levels for a specific set of proteins at three time points: baseline, day 4, and day 7 post-treatment. The unprocessed clinical data encompasses 240 patients and a total of 1,182 features. Meanwhile, the Olink proteomics data consists of 240 patients and 829 features.

3. Methods

3.1. Inclusion and exclusion criteria

We decided to include patients who got the treatment due to data size considerations and the similarity between the characteristic distributions of the two groups (Appendix A). In addition, to ensure the availability of all predictor variables in the prediction model development, we decided to include only measurements from day 1 to 15 to minimize missing values, since from day 15 less measurements were being recorded.

3.2. Data Preprocessing

Clinical data was divide into day-dependent measurements and "fixed" clinical information such as previous diseases, intake of drugs and medicine, and demographic information on each patient. We excluded features with data missing rate higher than 40% and dropped records that had more than 50% missing values. In addition, we noticed that variables such as hla-dr, ferritin and quantibrite were measured only at the first two days after admission, and since they have a clinical significance to sepsis state immune classification,

[0] GitHub link

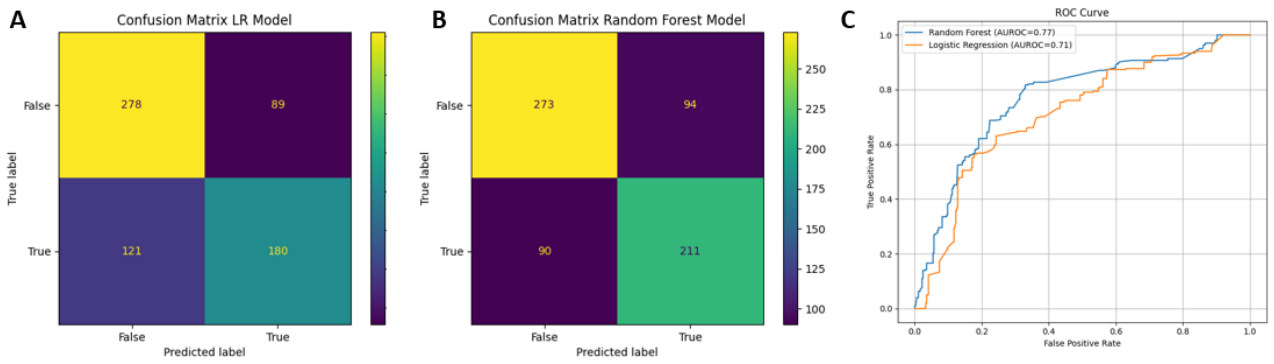

Figure 1. (A) Confusion matrix of the LR model (B) Confusion matrix of the RF model (C) ROC curves of both models

we decided to take the mean of each of them along the two days and consider them as "fixed" features. Regarding the Olink data, we decided to include only the Baseline day values since there were more than 87% missing values at day 4 and 7 in that data-set. Finally, a total of 25 clinical measurements, 77 clinical characteristics and 276 Olink proteomics measurements were used as independent predictor variables for the development of the model on a filtered cohort of 223 patients. On this filtered data, 47.5% of the records were associated with positive labels, with the remainder assigned to negative labels. Day-dependent features were imputed using the fill-forward method for each patient separately. Then, we handled numeric and categorical features separately. For the numeric features, including clinical measurements and Olink measurements, we first imputed the remaining missing values using the K-nearest-neighbors method. Then, we scaled and standardized the values. For the categorical features we performed One Hot Encoding. Then we merged the numeric processed data and the categorical processed data together.

3.3. Machine Learning-Based Models

We divided our work into two tasks: (i) Mortality Prediction Model and (ii) Endotype Classification Model. For the first task, we decided to investigate the use of Random-Forest model (RF) (Breiman, 2001) and Logistic Regression model (LR). Optimization of the models hyperparamaters was conducted using cross-validation (Appendix D.4).Then, we trained the selected models on the training dataset and evaluated their performance using the test dataset. For the second task, we used the k-means algorithm (Lloyd, 1982; MacQueen et al., 1967) for unsupervised machine learning classification. We evaluated the optimal number of clusters (n) using Elbow plot. After selecting n=5, we applied the k-means algorithm to the complete dataset, utilizing only the significant features identified through SHapley Additive Explanations (Lundberg & Lee, 2017) analysis, conducted on the RF model, as well as through Cox proportional hazards regression analysis (Cox, 1972). Finally, we used t-SNE for dimensionality reduction and data visualization of the clustered data (Van der Maaten & Hinton, 2008).

3.4. Performance Evaluation

We examined calibration of the predictor models through calibration plots and the computation of Brier scores (Brier, 1950). Following this, we implemented isotonic regression to calibrate the models' predictions and measured the impact on the Brier score. Next, we used the calibrated predictions on the test set for model performance evaluation. In addition to fundamental evaluation metrics such as model accuracy, positive predictive value (PPV) and F1 score, we also employed the area under the receiver operating characteristic curve (AUROC) and precision-recall curves (PR). Furthermore, we utilized SHapley Additive Explanations (Lundberg & Lee, 2017) to explore the individual contributions of each feature to the predictive outcomes.

4. Results

4.1. Mortality Prediction Model Performance and Evaluation

For the RF model, we used 100 trees in the forest, set the maximum tree depth to 12, and a minimum number of 4 samples required to split an internal node. For the LR model, we selected a penalty of L1 and regularization strengths 'C' as 1. The models were trained on the training set and evaluated on the test set. We used calibrated predictions for the evaluation of the models (calibration curve before and after calibration are provided in Fig, 7). We employed 6 different metrics for the evaluation of each model as can be seen in Table 1. Confusion metrics and ROC curves are shown in Fig. 1. After analyzing the validation metrics, we determined that the RF classifier outperformed the LR model, and thus, we chose to proceed with it.

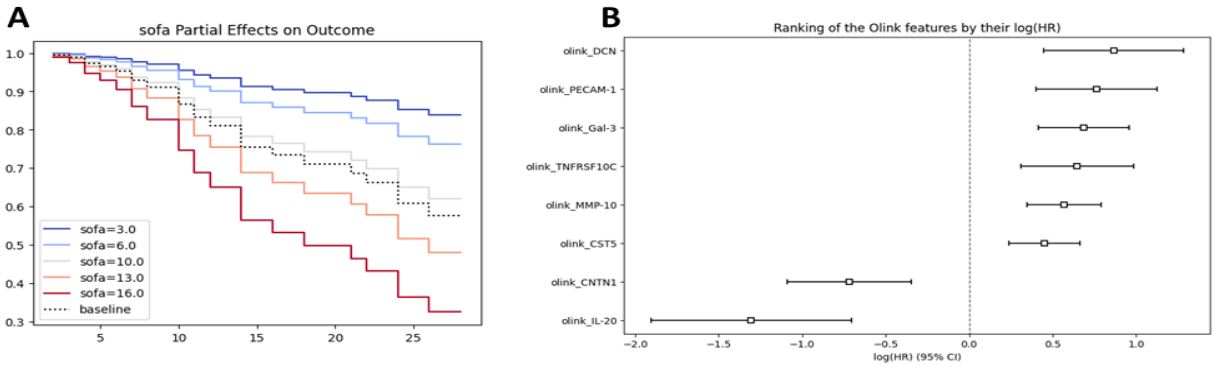

Figure 2. (A) Partial Effects on Outcome of SOFA features using Cox proportional hazards analysis (B) Ranking of significant Olink features by their log(Hazards-Ratio)

Table 1. Model Evaluation

Metric	Random Forest Classifier	Logistic Regression
Accuracy	0.72	0.68
F1 Score	0.69	0.63
PPV	0.69	0.67
Recall	0.7	0.6
AUROC	0.77	0.71
AUPR	0.71	0.62

4.2. Mortality Prediction Model Interpretation

In Fig.8A we show the importance of the independent features ranked by their MDI value in a descending order in the calibrated RF Model. These features have a greater impact on reducing impurity during the decision making process. Fig. 8B shows the features that most contribute to the predictive outcomes of the RF model, as determined by SHapley Additive Explanations. Of the 20 most important features, these two methods have 15 common significant features, and they differ mostly by Olink features and quantibrite-mean feature which has a high contribution according to SHAP score. For the top important features in the calibrated LR model, extracted by their absolute coefficients, LR and RF models had only 7 features in common.

4.3. Cox Proportional Hazards Regression Analysis

In Fig.2A we show the partial effects of the SOFA feature on the mortality outcome. As expected, patients with higher SOFA scores have lower survival probabilities. In addition, only SOFA, p02, neutrophils, and hladr-mean have shown significance among the clinical parameters. (Fig 9). Fig.2B shows the ranking of the significant Olink features by their log(HR) values and Fig.10 shows the whole results analysis.

4.4. Endotype Classification Model Interpretation

In Fig. 3A , the separation between the groups in the t-SNE plot is not entirely distinct; however, we can see that clusters 1 and 2 are well-defined and homogeneous. On the other hand, clusters 3 and 4 appear to merge into a single cluster. Additionally, the densely crowded points within each cluster correspond to the same patients. This observation is based on the fact that the data is time-dependent and some records belong to the same patient. When combined with the imputation process, these factors have likely caused the data points for individual patients to cluster together. In Fig.3B we see the distribution of the outcome of the trial (0 for survived, 1 for deceased), where cluster 2 has much more survivors compared to cluster 1 which has the higher deceased mean rate. The heatmap in Fig.3C represents the differences between the medians of each Olink feature across the clusters. Some features don't show much variation, while few, such as CA5A, DECR1, MMP-10 and IGFBP-1 proteins vary the most between clusters 1 and 2. In addition, we can see that clusters 0 and 2 are very similar in their median values, strengthening their short distance in the t-SNE projection. Furthermore, it's worth noting that we examined the distribution of patients across clusters (Table3) and we see that most of the records from the same patient were clustered together. We suspect that the clusters may represent a timeline progress of their condition.

5. Discussion

In this study, we used the PROVIDE clinical trial dataset to gain new insights into two primary objectives: 1) predicting mortality in sepsis patients following diagnosis, and 2) enhancing precision medicine for sepsis patients by uncovering sepsis endotypes. We started by creating a 28-day mortality prediction model, comparing between LR and RF models. With a total of 378 selected features and a filtered cohort of 223 patients, we showed that the RF model had

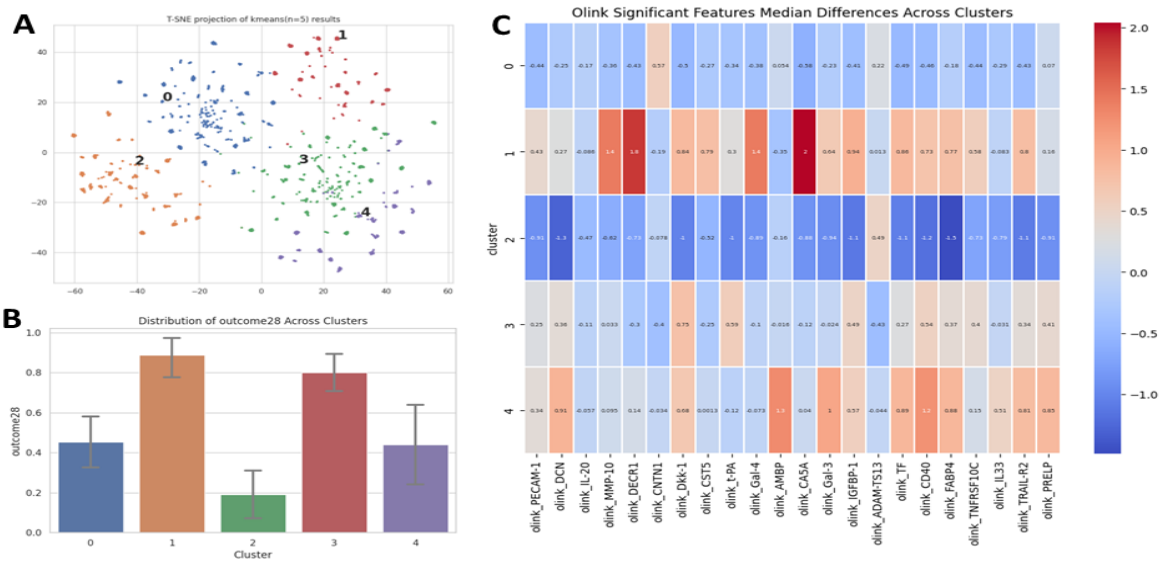

Figure 3. (A) t-sne Projection of patients clustered to n=5 endotypes using k-means. (B) Distribution of the label "outcome28" (1-deceased, 0- survived) among the endotypes identifiifed. (C) Heatmap of the median of each Olink feature used for the endotypes classification.

an AUROC of up to 0.77 and an accuracy of 0.72 on the testing set. Considering that our data is not imbalanced, the RF model performs impressively. Given that the LR model yielded less favorable performance metrics, we opted for the RF model as the most suitable choice for our task. In addition, due to the RF's ability to generate multiple decision trees to fit our dataset, we were able to interpret the ranking of the important features. As expected, both SOFA score and age parameters arose as significant factors in predicting mortality outcomes. Interestingly, the subsequent most important features in the prediction task were derived from Olink proteomics measurements, notably IGFBP-1 (Ashare et al., 2008; Mavrommatis et al., 2001) and DECR1 (Long et al., 2022; Miao et al., 2021) proteins, which have been previously mentioned in sepsis research. In addition, the application of the Shapley values method provided a validation for the significance of most of the features identified in the RF methodology. Therefore we can conclude that the incorporation of Olink measurements, which are not conventionally part of routine ICU assessments, could be beneficial for treatment. Apart from selecting the best predicting features using the RF model, we also performed a Cox proportional hazards regression to understand the effect of quantitative variables on trial patients' survival. Our rationale for conducting this analysis was to complement the set of important features by adding those that lead to the increase of the HR value and therefore increase the likelihood of the event occurring. The results of this analysis indicated that certain clinical parameters such as hla-dr, pO2 and neutrophils measurements and various Olink parameters (some of which were already highlighted in the RF feature importance ranking), could potentially play a significant role in patient survival and therefore we believed thatthis analysis could improve the endotypes classification model. In this unsupervised learning task, finding the optimal number of clusters (n) for k-means was a challenge. When we applied t-SNE with n=5 to the k-means solution, we discovered possibly multiple endotypes. However, the solution also revealed some cluster impurity. On one hand, it is clear that some clusters differ by their survival rate, and also by their mean SOFA which is linearly related to survival. On the other hand, the proteomics heatmap between the clusters highlights that some proteins characterize more some clusters than others, and therefore some other optional endotypes can be suggested. Further analysis and remodeling of that unsupervised task is necessary to address the classification to endotypes properly. We would suggest statistical testing on the results of the heatmap, such as ANOVA. Finally we believe that the predictive model can assist physicians and clinical trials by collecting the proper clinical features and especially since the proteomics data used for the model on was based only on the first day of admission since the trial. Our hope is that it would facilitate the stratification of sepsis patients for precision medicine based on these interesting insights.

References

Chao, H.-Y., Wu, C.-C., Singh, A., Shedd, A., Wolfshohl, J., Chou, E. H., Huang, Y.-C., & Chen, K.-F. (2022). Using Machine Learning to Develop and Validate an In-Hospital Mortality Prediction Model for Patients with Suspected Sepsis [Number: 4 Publisher: Multidisciplinary Digital Publishing Institute]. *Biomedicines*, *10*(4), 802. https://doi.org/10.3390/biomedicines10040802

Kong, G., Lin, K., & Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC Medical Informatics and Decision Making*, *20*(1), 251. https://doi.org/10.1186/s12911-020-01271-2

Baghela, A., Pena, O. M., Lee, A. H., Baquir, B., Falsafi, R., An, A., Farmer, S. W., Hurlburt, A., Mondragon-Cardona, A., Rivera, J. D., et al. (2022). Predicting sepsis severity at first clinical presentation: The role of endotypes and mechanistic signatures. *EBioMedicine*, *75*.

Scicluna, B. P., Van Vught, L. A., Zwinderman, A. H., Wiewel, M. A., Davenport, E. E., Burnham, K. L., Nürnberg, P., Schultz, M. J., Horn, J., Cremer, O. L., et al. (2017). Classification of patients with sepsis according to blood genomic endotype: A prospective cohort study. *The Lancet Respiratory Medicine*, *5*(10), 816–826.

Leventogiannis, K., Kyriazopoulou, E., Antonakos, N., Kotsaki, A., Tsangaris, I., Markopoulou, D., Grondman, I., Rovina, N., Theodorou, V., Antoniadou, E., et al. (2022). Toward personalized immunotherapy in sepsis: The PROVIDE randomized clinical trial. *Cell Reports Medicine*, *3*(11).

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

MacQueen, J., et al. Some methods for classification and analysis of multivariate observations. In: In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability*. *1*. (14). Oakland, CA, USA. 1967, 281–297.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Ashare, A., Nymon, A. B., Doerschug, K. C., Morrison, J. M., Monick, M. M., & Hunninghake, G. W. (2008). Insulin-like growth factor-1 improves survival in sepsis via enhanced hepatic bacterial clearance. *American journal of respiratory and critical care medicine*, *178*(2), 149–157.

Mavrommatis, A, Papanicolaou, S, Kostadelou, E, Kotanidou, A, Malefaki, A, Katsaris, G, & Zakynthinos, S. (2001). Sepsis progression is associated with a gradual depletion of both insulin-like growth factor i (igf-i) and insulin-like growth factor binding protein-3 (igfbp3) and a progressive elevation of growth hormone (gh) serum levels. *Critical Care*, *5*(Suppl 1), P130.

Long, Q., Li, G., Dong, Q., Wang, M., Li, J., & Wang, L. (2022). Exploration of the shared gene signatures between myocardium and blood in sepsis: Evidence from bioinformatics analysis. *BioMed Research International*, *2022*.

Miao, H., Chen, S., & Ding, R. (2021). Evaluation of the molecular mechanisms of sepsis using proteomics. *Frontiers in Immunology*, *12*, 733537.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

Mantel, N., et al. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, *50*(3), 163–170.

A. Statistical Analysis

The data consist of 240 patients, from 14 different clinical centers, with 59.2% of them being Males. The average age of the patients was 72.633 ± 14.439. A total of 144 patients (60%) did not survive beyond the 28-day cutoff point. Septic shock was diagnosed in 73.8% of the patients. 44 patients (18.3%) were diagnosed with MALS, 2 patients (0.8%) were diagnosed with Immunosuppression and the remaining were categorized as Unknown. Within the 36 patients enrolled in the treatment part of the trail, 2 were diagnosed with Immunosuppression and the rest with MALS. The average day 1 SOFA score for all patients was 10.98 ± 4.11 whereas for the enrolled patients it was 14.41 ± 2.93. The statistical analysis results for both the entire patient group and the enrolled group is provided in Table 2. Distribution plots are provided in Fig. 4.

B. Survival Analysis

We used the KM multivariate log rank test p-values for the survival analysis results of the different characteristic groups. We noticed that the sepsis state received a significant p-value ($p<0.01$), meaning different sepsis states experienced different survival outcomes, as expected. Additionally, age groups showed significant differences in survival ($p<0.01$), which is noteworthy but unsurprising. Notably, there was no significant difference in survival outcomes among different clinical centers, which indicates that clinical center location does not significantly impact sepsis survival. Gender groups did not exhibit a significant p-value. Regarding treatment type among enrolled patients, there was no significant difference in survival between those who received immunotherapy and those who received a placebo. This finding suggests that the examined treatment in the trial may not be very effective. Additionally, when examining the enrolled group to the remaining patients, we saw that the enrolled group experienced a significantly poorer survival rate ($p<0.01$). This observation aligns with our expectations, as the patients selected for treatment were already in a more critical condition, and it appears that the treatment did not provide significant benefit. These findings further support our decisions to include the enrolled patients in the predictor dataset.

C. Limitations

Our study has a few limitations that are important to take into account. One of them is the limited sample size. Smaller sample sizes can lead to unstable estimates and may not provide enough data to build robust prediction models. In addition, when integrating clinical trial data into a prediction model, it's important to acknowledge some limitations that can arise. The first is the impact of inclusion and exclusion criteria applied during the trial. These criteria are designed to ensure a controlled and specific study population, enhancing the internal validity of the trial. However, this controlled selection might lead to a dataset that doesn't fully represent the broader patient population encountered in real-world clinical practice. Additionally, clinical trials are usually conducted in a controlled setting, and may not represent the standard real-world conditions and care that patients in ICU usually receive. This may affect the feature values used for the model. Furthermore, it may be that not all the information used as modeling inputs in this study is routinely collected from the emergency department. In the context of the treatment offered in the trial, we can remove the treatment features and we believe there won't be a big difference, since they did not appear as significant features for the model's prediction. Moreover, the outcome that we adopted in this study, the 28-day mortality, might not be the best endpoint to improve the care of septic patients.

D. Methods

D.1. Data Exploratory

We explored the data in two aspects- through the distribution of the clinical information and through survival analysis. We conducted survival analysis using the Kaplan–Meier estimator (Kaplan & Meier, 1958) and the Multivariate log rank test (Mantel et al., 1966) to assess differences in survival among various characteristic groups (Fig. 5)

D.2. Feature Selection

Feature selection was considered in several steps of the study. For the feature selection of the mortality prediction task, we used correlations matrices between the numeric features. For a group of features that had a correlation value higher than 85%, only one representative feature was kept. As for the endotypes classification task, we used the most significant features based on Cox proportional hazards regression survival analysis on continuous features and based on the mortality prediction model top features.

D.3. Data Partitioning

Data was divided into 70% training set and 30% test set, stratified by outcome as well as by patients (meaning a patient cannot have records in both sets) to avoid leakage. We used a 5-fold cross-validation method on the training dataset to fine-tune our prediction model's hyperparameters, and made sure that the partitioning folds was also stratified by outcome and patients.

D.4. Models Hyperparamaters Optimization

To optimize the performance of both models we explored different hyperparameters through cross-validation to determine the combination of hyperparameters that yields the best Receiver Operating Characteristic Area Under the Curve (ROC AUC) result. For the RF model, we explored the following hyperparameters: maximum depth of the tree (6, 12), minimum number of samples required to split an internal node (2, 4) and the number of trees in the forest (100, 200). For the LR model we considered 2 penalty options ('L1', 'L2') and explored different regularization strengths, as defined by the C value (0.25, 0.5, 1).

E. Supplementary Tables

Table 2. Statistical Analysis of the Data Set

Feature	All (n=240)	Enrolled (n=36)
Age average	72.63± 14.43	70.08± 12.57
Gender (Male)	59.2%	61.1%
SOFA Day1 average	10.98± 4.11	14.41± 2.93
28Day Outcome (Not survived)	60%	89%

Table 3. Patients Distribution in clusters

# of Matching Clusters	# of Patients
1	201
2	19
3	3

F. Supplementary Figures

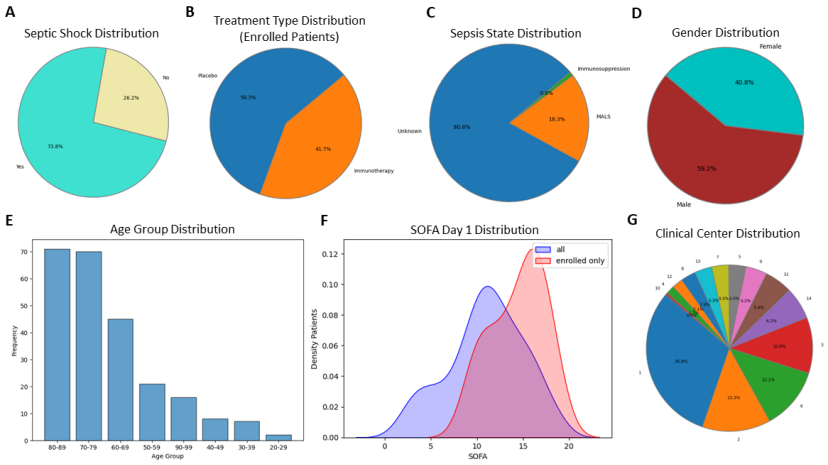

Figure 4. (A) Pie chart of Septic Shock distribution among patients (B) Pie chart of treatment type distribution among patients enrolled in the treatment stage (C) Pie chart of Sepsis state distribution among patients (D) Pie chart of gender distribution among patients (E) Bar plot of age-group distribution among patients (F) Density plot comparing Day 1 SOFA scores for all patients with those specifically enrolled (G) Pie chart of Clinical Center distribution among patients

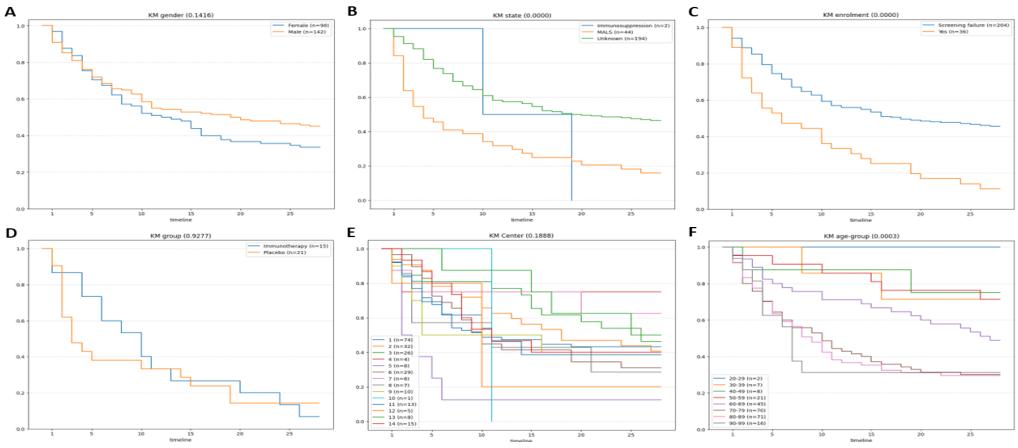

Figure 5. (A) Kaplan Meier curves comparing survival of Males and Females (B) Kaplan Meier curves comparing survival of patients with different sepsis state (C) Kaplan Meier curves comparing survival of total patients and the enrolled patients (D) Kaplan Meier curves comparing survival of of enrolled patients who received immunotherapy and those who received placebo (E) Kaplan Meier curves comparing survival of patients from different clinical centers (F) Kaplan Meier curves comparing survival of total patients from different age groups

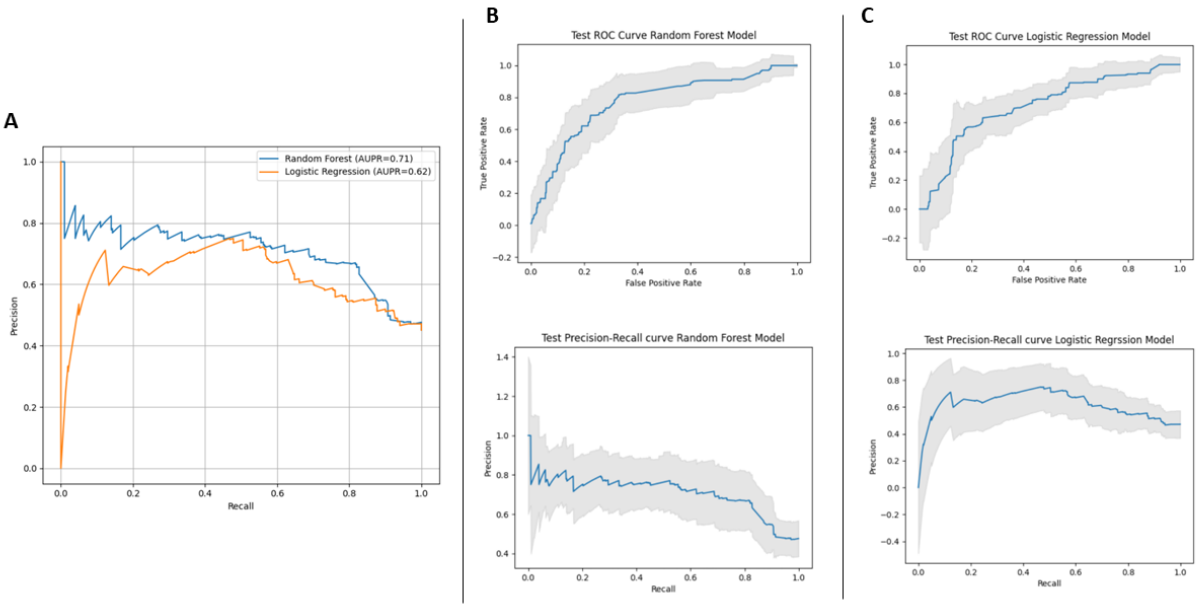

Figure 6. (A) ROC curves of the Random Forest model and the Logistic Regression model (B-C) ROC curve (top) and PR curve (bottom) of the Random Forest model (left) and the Logistic Regression model (right) with the shade indicating the standard deviation across 20 random test set sub samples, each containing 20% of the test set with replacements.

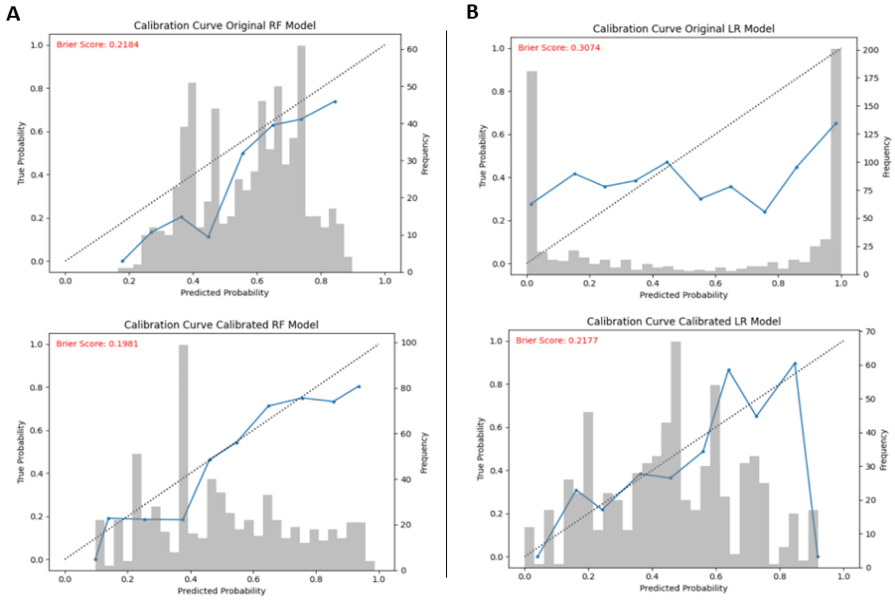

Figure 7. (A-B) Calibration curves of the Random Forest model (left) and the Logistic Regression model (right) before (top) and after (bottom) calibrating the models using isotonic regression. Brier Score is mentioned in red.

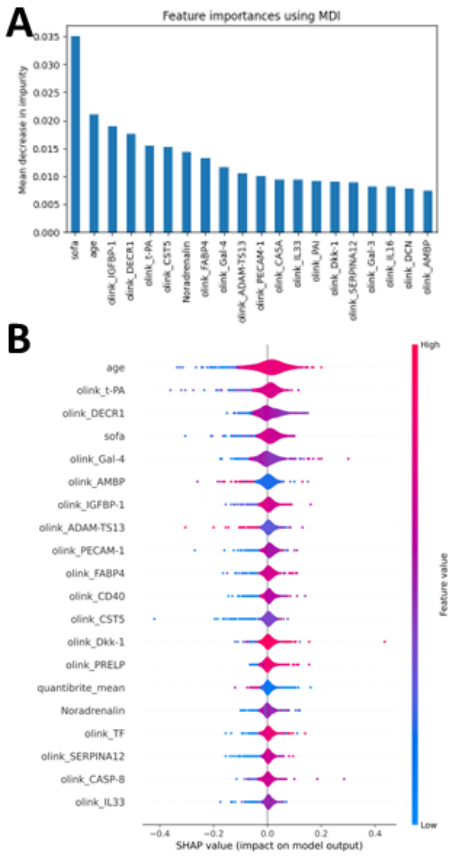

Figure 2. (A) Feature importance using MDI of RF model (B) SHAP values for RF model

Figure 8. (A) Feature importance using MDI of RF model (B) SHAP values for RF model

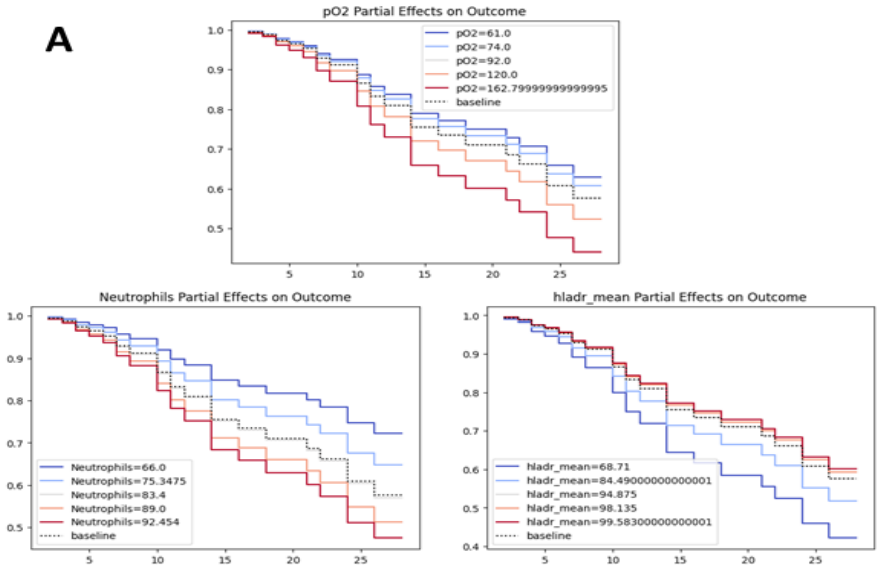

Figure 9. Partial Effects on Outcome of pO2,Neutrophils and hladr-mean clinical features using Cox proportional hazards analysis

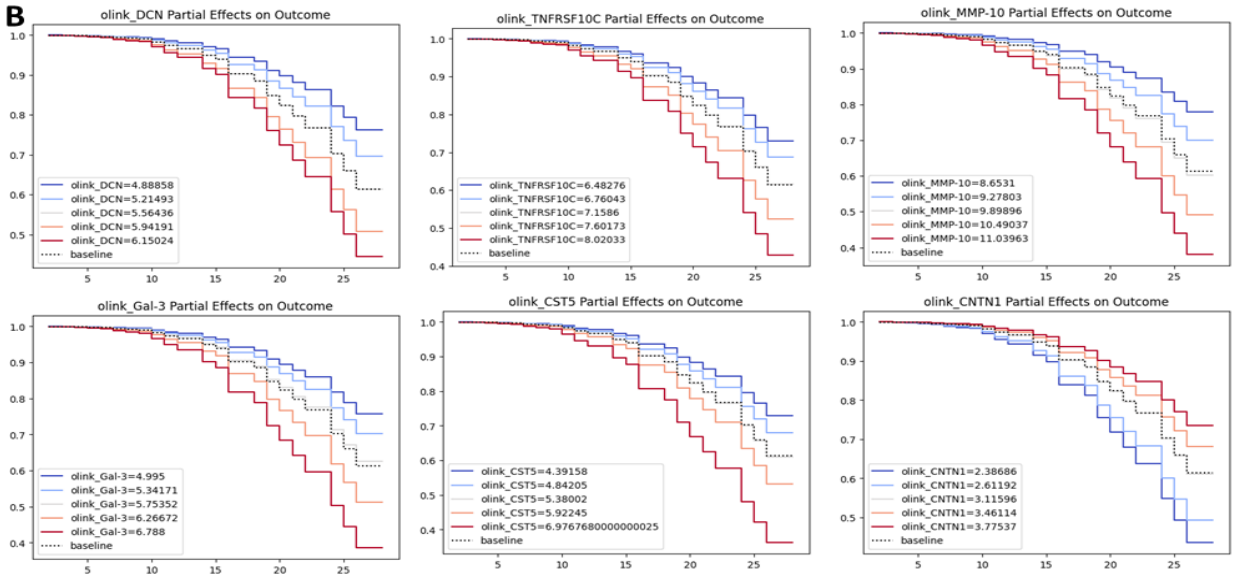

Figure 10. Partial Effects on Outcome of DCN, TNFRSF10C, MMP-10, Gal-3, CST5 and CNTN1 Olink proteomics features using Cox proportional hazards analysis

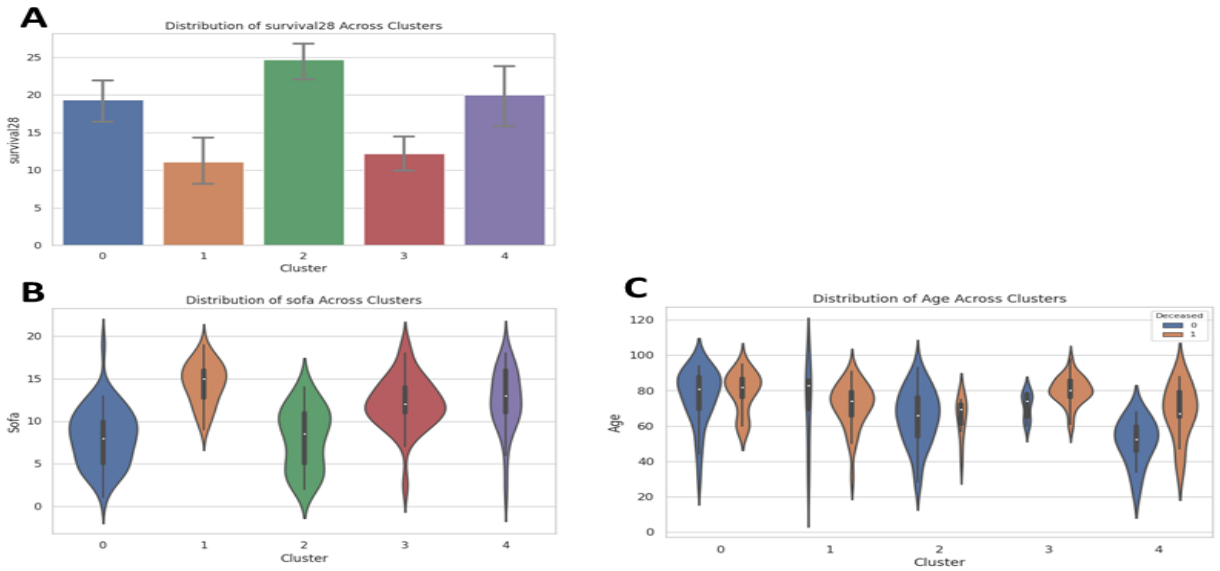

Figure 11. (A) Distribution of days of survival across the identified endotypes (B) Distribution of SOFA score across the identified endotypes (C) Distribution of age across the identified endotypes differed by ”deceased” label.