Identification of effective biomarkers in predicting the survival of patients with severe sepsis and septic shock

December 16, 2022

Abstract

Sepsis and septic shock are life-threatening medical conditions caused by blood infection, resulting in unwanted reactions from the human immune system. As the outcome of these conditions benefits from early intervention, there is increasing interest in the prognostic value of various biomarkers, especially lactate level, established to effectively categorize sepsis severity and predict patient survival. To explore and verify these predictive values, we applied random forest and logistic regression models, with bagging and cross-validation for minimal overfitting, to data on patients at the National Hospital of Tropical Diseases, Vietnam. Out-of-sample accuracy and area under the curves (AUC) of receiver operating character (ROC) curves were used to assess these models. Our results reinforced the merits of regularly monitoring lactate levels (AUC = 0.77, accuracy = 76.12%), and identified predictive potential for procalcitonin level, along with other infection biomarkers, in blood (AUC = 0.66, accuracy = 71.64%). Future studies should account for time of death and focus on procalcitonin level in blood as a predictor in combination with other infection biomarkers.

1) Introduction

Sepsis is a life-threatening medical condition, in which the immune system of the body reacts to an infection typically caused by bacteria in blood (Gyawali et al., 2019). In contrast to other localized infections, sepsis is a multifaceted disruption between the pro-inflammatory and anti-inflammatory pathways that induces a cascade of cytokine activations, or also known as cytokine storm (Jarczak et al., 2021), and results in a wide range of possible symptoms, including low temperature, low blood pressure, rapid breathing, and low urine output (Vincent, 2016). Severe sepsis can eventually progress to septic shock, a condition characterized by low blood pressure and organ dysfunction (Singer et al., 2016). Despite advancement in the understanding of the pathophysiology as well as in the monitoring tools and resuscitation measures, sepsis and septic shock remain among the most prominent immediate causes of death with extremely high morbidity and mortality rate, especially in critically ill patients (Kaukonen et al., 2014; Rhee et al., 2019). It has been estimated that these conditions affect approximately 1.7 million adults with over 250,000 deaths in the United States each year, causing a significant burden on both human and financial resources (Rhee et al., 2017).

Because the outcome of sepsis and septic shock has been shown to benefit from early intervention (Kumar et al., 2006), there is an increasing interest in the prognostic values of various biomarkers – medical signs that objectively indicate the clinical state of the patients (Strimbu & Tavel, 2010). Among these biomarkers, lactate level has been established as an important measurement capable of categorizing the severity of sepsis and prognosing the survivability of patients with septic shock (Filho et al., 2016; Marty et al., 2013; Wacharasint et al., 2012). Lactate is the product of glycolysis under anaerobic condition with the catalysis effect of lactate dehydrogenase during the tricarboxylic acid cycle. The amount of lactate produced increases (hyperlactatemia) when patients suffer from septic shock as their blood circulation and respiration rate decreases, reducing the oxygen level of cells (Semler & Singer, 2019). Despite lactate being regarded as an effective prognostic biomarker, there is contradicting evidence on which time post-diagnosis of septic shock for lactate measurements would be the most appropriate biomarker, in addition to a lack of exploration on other potentially effective biomarkers. Therefore, we performed our study on patients at the National Hospital of Tropical Diseases in Vietnam with the following goals: (i) examine the optimal post-diagnosis time for measuring lactate level to serve as a good predictor and (ii) identify other potential biomarkers for sepsis and septic shock prognosis.

2) Materials & Methods

2.1) Study Population

The data contains information on all patients over 18 years old diagnoses with septic shock at the National Hospital of Tropical Diseases, Vietnam from June 2018 to July 2022. The criteria for septic shock diagnosis are proposed by the guidelines of the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) (Singer et al., 2016) and Surviving Sepsis Campaign (Rhodes et al., 2017), including: two or more Systemic Inflammatory Response Syndromes (temperature > 38°C or < 36 °C; heart rate > 90/min; respiratory rate > 20/min or $PaCO_2$ < 32mm Hg (4.3 kPa); white blood cell count > 12.000/mm3 or < 4000/mm3 or > 10% immature bands) and a Sequential Organ Failure Assessment (SOFA) score for the assessment of clinical condition over 2. The patients were also evaluated on the use of vasopressors required to maintain the mean arterial pressure ≥ 65 and lactate level ≥ 2 mmol/L for septic shock. Any patients that either were transferred from other hospitals, had a history of liver/kidney failure, or experienced circulatory/respiratory arrest before admittance were excluded from the study.

2.2) Data Collection & Processing

On admission to the intensive care unit (ICU), demographic information and various biomarkers related to respiration, blood circulation, kidney/liver/heart failure, and infections at the time of admittance and septic shock were recorded, with the exception of lactate level (Appendix 7.1 – Table 1.). The lactate level was measured at time of diagnosed septic shock (T0), 24 hours after the diagnosis (T1), 48 hours after the diagnosis (T2), and 72 hours after the diagnosis (T3). The outcome of the treatment, either death or survival, were also documented at the end.

R (version 4.2.1) was used to process and analyze the data upon retrieval. Exploratory data analysis was performed to provide a better understanding of the study population, identify collinear variables, and pinpoint potentially biased or problematic ones. Among each group of collinear variables identified, one representative variable was chosen to be included in data analysis. Two new variables were derived from the lactate levels at different time points, namely the peak lactate level for each patient and the associated time point. These new variables would facilitate a better assessment of lactate as an overall predictor while accounting for the discrepancies among the time points at which each patient's lactate level peaks. To account for missing entries in the data, k-nearest neighbors was used as a method to impute missing values for predictors other than those related to lactate. Any missing entries for lactate levels were extrapolated by carrying the last-available value forward.

2.3) Data Analysis

Once the data had been processed, a random forest model (Breiman, 2001) – collections of classification trees fit to differently bootstrapped data sets whose majority voting decides the classification result of the model – was applied to determine the variables that could best predict the treatment outcome of septic shock patients. To ensure the robustness of the model, a grid search was performed to find the optimal tuning paraments, including the number of decision trees in the forest, the number of variables randomly sampled to be candidates at each split in a decision tree, and the maximum number of nodes in each tree. Three iterations of the grid search were performed, starting with a wide span of values for the parameters, with each iteration closing in towards the most optimal values.

After a set of optimal tuning parameters was found, the random forest model was applied to the data and the importance of each variable was examined using mean decreases in Gini Index. High-importance variables – those with large mean decreases in Gini Index – would subsequently be used for logistic regression in hope of discovering simple models that could predict the treatment outcome with similar levels of accuracy as the complicated random forest model. To determine the optimal post-diagnosis time to measure lactate levels, logistic regression with six-fold cross-validation was performed using each of the time points in addition to the peak lactate levels as predictor variables. The efficacy of each model was assessed using out-of-sample accuracy rate and the area under the curves (AUC) of the receiver operating characteristic curve (ROC).

3) Results

The descriptive analysis reveals that the study population includes 134 patients, with an average age of 60.74 and a higher percentage of patients over 65, leaning towards the older end of the spectrum. In addition, the mean number of days in hospital is 13 while in ICU is 8, which indicates that the conditions of patients progress rapidly (Appendix 7.1 – Table 1.). The data consisting of 33 biomarkers had approximately 16% missing entries, all of which were imputed through either k-nearest neighbors or extrapolation.

Our grid search suggested the optimal random forest use 2000 decision trees, each containing a maximum of 9 nodes, and each fit using a randomly selected set of 15 candidate variables. Upon examining the importance of the variables using the mean Gini Index decreases (GID) (Appendix 7.2 – Table 2.), we identified the variables with high GID. This consisted of the total numbers of days hospitalized (GID = 7.25), lactate level in blood 72 hours post-diagnosis (GID = 6.04), lactate level in blood 48 hours post-diagnosis (GID = 3.56), procalcitonin level in blood at time of diagnosis (GID = 2.98), and peak lactate (GID = 2.62).

Due to the question of the study, we did not focus on the age of the patients and their number of days in hospital/ICU. Other high-importance variables from the random forest model were grouped by categories, with the exception of procalcitonin level and peak lactate level across all time points. These combinations are infection (procalcitonin level + C reactive protein level), immunity (platelet level + white blood cell level + neutrophil level + lymphocyte level), liver function (albumin level + aspartate aminotransferase level + alanine transaminase level), blood function (red blood cell level + D-dimer + hemoglobin level), and metabolism (urea level + sodium level + potassium level). Logistic regression models were subsequently applied to these combinations of variables along with peak lactate level and procalcitonin level. The resulting ROC curves show that while random forest model predicted the treatment outcome most effectively (AUC = 0.81, accuracy = 81.34%), both the models using peak lactate level (AUC = 0.77, accuracy = 76.12%) and the infection combination (AUC = 0.66, accuracy = 71.64%) demonstrate adequate effectiveness (Appendix 7.3 – Figure 3. & Table 3.). In addition, we also performed logistic regression on each post-diagnosis time point of lactate levels and compared them to logistic regression model of peak lactate level and the random forest model to identify the better time points. These comparisons indicate that later time points (T2 and T3) would be better in predicting the outcome of the patients (AUC = 0.79, accuracy = 76.12%; AUC = 0.82, accuracy = 76.87%) (Appendix 7.3 – Figure 4. & Table 3.).

4) Discussion

Measuring post-diagnosis lactate levels has been empirically established to be an effective method for predicting treatment outcome of patients with severe sepsis and septic shock (Filho et al., 2016; Marty et al., 2013; Wacharasint et al., 2012) and our findings continue to support this notion. Further examination of the cumulative results from random forest model and the comparisons between logistic regression models of lactate levels at different time points suggest that lactate levels can be a strong predictor when measured closer to a patient's eventual treatment outcome. This would normally correspond to later time points (T2 and T3) rather than earlier ones (T0 and T1). In addition to lactate, our analysis also reveals that procalcitonin level in blood can be a good predictor, especially when combined other biomarkers of the same category of infection indication such as C-reactive protein.

Despite our findings, it is crucially important to recognize that our study possess shortcomings that can be improved and revised by future work. To begin with, as mentioned previously, there is individual variation in the way lactate variables were measured: patients might die within 72 hours of their diagnosis, resulting in missing values for lactate levels at latter time points (Appendix 7.1 – Figure 2.). We tried to reduce this variation by transferring the closest available lactate values to latter time points for such patients but this transformation makes our findings less conclusive. In future research, we would propose that regularly monitoring of the lactate levels in patients with severe sepsis or septic shock can be useful in predicting their outcomes, rather than focusing on a specific time point. We would also suggest that the time of death should be recorded in the future to perform Cox regression with time-dependent covariates for more robust conclusions. Moreover, the data we retrieved from the National Hospital of Tropical Disease was missing a large portion of its data across all variables. This required us to perform imputation method, which may have affected our final conclusion to a certain extent. Lastly, having found that procalcitonin level may be a good predictor for treatment outcome, we believe that more focus should be placed on studying this biomarker, potentially in combination with other biomarkers related to the indication of infection, such as receptor expressed on myeloid cells-1 (sTREM-1) and immunoglobulin-Fc fragment receptor I (FcyRI) (Gibot et al., 2012).

5) Reference

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Filho, R. R., Rocha, L. L., Corrêa, T. D., Pessoa, C. M., Colombo, G., & Assuncao, M. S. (2016). Blood lactate levels cutoff and mortality prediction in sepsis-time for a reappraisal? A retrospective cohort study. *Shock*, *46*(5), 480-485. https://doi.org/10.1097/shk.0000000000000667

Gyawali, B., Ramakrishna, K., & Dhamoon, A. S. (2019). Sepsis: The evolution in definition, pathophysiology, and management. *SAGE Open Medicine*, *7*, 205031211983504. https://doi.org/10.1177/2050312119835043

Gibot, S., Béné, M. C., Noel, R., Massin, F., Guy, J., Cravoisy, A., Barraud, D., De Carvalho Bittencourt, M., Quenot, J.-P., Bollaert, P.E., Faure, G., & Charles, P.-E. (2012). Combination biomarkers to diagnose sepsis in the critically ill patient. *American Journal of Respiratory and Critical Care Medicine*, *186*(1), 65-71. https://doi.org/10.1164/rccm.201201-0037oc

Jarczak, D., Kluge, S., & Nierhaus, A. (2021). Sepsis-pathophysiology and therapeutic concepts. *Frontiers in Medicine*, *8*. https://doi.org/10.3389/fmed.2021.628302

Kaukonen, K.-M., Bailey, M., Suzuki, S., Pilcher, D., & Bellomo, R. (2014). Mortality related to severe sepsis and septic shock among critically ill patients in Australia and New Zealand, 2000-2012. *JAMA*, *311*(13), 1308. https://doi.org/10.1001/jama.2014.2637

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., Gurka, D., Kumar, A., & Cheang, M. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock*. *Critical Care Medicine*, *34*(6), 1589-1596. https://doi.org/10.1097/01.ccm.0000217961.75225.e9

Marty, P., Roquilly, A., Vallée, F., Luzi, A., Ferré, F., Fourcade, O., Asehnoune, K., & Minville, V. (2013). Lactate clearance for death prediction in severe sepsis or septic shock patients during the first 24 hours in Intensive Care Unit: An observational study. *Annals of Intensive Care*, *3*(1), 3. https://doi.org/10.1186/2110-5820-3-3

Rhee, C., Dantes, R., Epstein, L., Murphy, D. J., Seymour, C. W., Iwashyna, T. J., Kadri, S. S., Angus, D. C., Danner, R. L., Fiore, A. E., Jernigan, J. A., Martin, et al. (2017). Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. *JAMA*, *318*(13), 1241. https://doi.org/10.1001/jama.2017.13836

Rhee, C., Jones, T. M., Hamad, Y., Pande, A., Varon, J., O'Brien, C., Anderson, D. J., Warren, D. K., Dantes, R. B., Epstein, L., & Klompas, M. (2019). Prevalence, underlying causes, and preventability of sepsis associated mortality in US Acute Care Hospitals. *JAMA Network Open*, *2*(2). https://doi.org/10.1001/jamanetworkopen.2018.7571

Rhodes, A., Evans, L. E., Alhazzani, W., Levy, M. M., Antonelli, M., Ferrer, R., Kumar, A., Sevransky, J. E., Sprung, C. L., Nunnally, M. E., Rochwerg, B., Rubenfeld, G. D., Angus, D. C., Annane, D., Beale, R. J., Bellinghan, G. J., et al. (2017). Surviving sepsis campaign. *Critical Care Medicine*, *45*(3), 486-552. https://doi.org/10.1097/ccm.0000000000002255

Semler, M. W., & Singer, M. (2019). Deconstructing hyperlactatemia in sepsis using central venous oxygen saturation and base deficit. *American Journal of Respiratory and Critical Care Medicine*, *200*(5), 526-527. https://doi.org/10.1164/rccm.201904-0899ed

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis 3). *JAMA*, *315*(8), 801. https://doi.org/10.1001/jama.2016.0287

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, *5*(6), 463- 466. https://doi.org/10.1097/coh.0b013e32833ed177

Vincent, J.L. (2016). The clinical challenge of sepsis identification and monitoring. *PLOS Medicine*, *13*(5). https://doi.org/10.1371/journal.pmed.1002022

Wacharasint, P., Nakada, T., Boyd, J. H., Russell, J. A., & Walley, K. R. (2012). Normal-range blood lactate concentration in septic shock is prognostic and predictive. *Shock*, *38*(1), 4-10. https://doi.org/10.1097/shk.0b013e318254d41a

6) Acknowledgements

The authors are grateful for the guidance and assistance of professor Ryan Miller from Grinnell College on this project. We also appreciate the permission to use the data from doctor Nguyen Thi Huyen and Ta Thi Dieu Ngan from the National Hospital of Tropical Diseases and Hanoi Medical University in Vietnam. The study was approved by the ethical committee of Hanoi Medical University in Vietnam.

7) Appendix

7.1) Exploratory Data Analysis

Type of information	Variables/Biomarkers	Mean	Standard deviation	Range
Demographic information & Evaluation	Patient Age	60.74	14.75	22-98
	Patient ID	~	~	~
	The total number of days in hospital	13.04	11.67	1-70
	The total number of days in ICU	8.45	7.3	1-36
	The result of the treatment	~	~	~
	Sequential Organ Failure Assessment	12.9	3.61	3-21
Clinical Measurements (All variables in analysis were recorded at the time of diagnosed septic shock, with the exception of lactate level)	Mean arterial pressure	56.35	10.84	0-86.7
	Respiratory rate	29.55	7.16	14-50
	Pulse rate	116.7	26.59	60-280
	Amount of urine	1.26	0.44	1-2
	Site of initial infection	~	~	~
	Albumin level	26.83	5.31	15-44
	Procalcitonin level	39.03	31.65	0.39-100
	C Reactive Protein level	151.5	112.48	5.1-652
	Pro b-type natriuretic peptide level	10,602.16	12,173.25	50.94-36,000
	Creatinine level	170.36	125.85	42-726
	Platelet level	130.68	110.08	2-564
	White blood cell level	12.37	7.13	0.3-31.1
	Neutrophil level	80.44	19.36	2.74-97.5
	Lymphocyte level	11.23	14.93	1.46-6.62
	Red blood cell level	3.47	0.85	1.46-6.62
	Hemoglobin level	103.2	22.09	62-186
	Urea level	14.83	9.69	3.2-55.1
	Sodium level	134.23	15.57	3.09-163
	Potassium level	4.05	0.85	2.8-7.93
	Bilirubin level	30.88	35.01	3.2-294.8
	Aspartate aminotransferase level	394.3	925.57	20.7-6933
	Alanine transaminase level	263.1	918.74	2-7170
	Percent prothrombin	51.93	21.09	10-101
	Fibrinogen level	3.56	2.04	0.18-10.93
	D-dimer level	15,184	22375.88	122-128,000
	The number of failed organs	3.4	1.1	1-5
	The type of bacteria responsible for septic shock	~	~	~
	Lactate levels at the time of diagnosed septic shock (T0), 24h after the diagnosis (T1), 48h after the diagnosis (T2), and 72h after the diagnosis (T3)	T0: 5.52 T1: 5.56 T2: 4.67 T3: 3.72	T0: 3.97 T1: 4.43 T2: 3.98 T3: 3.2	T0: 1.18-22.49 T1: 0.81-24.33 T2: 0.95-20.04 T3: 0.92-16.53

Table 1. Measurements of patients at time of admittance to ICU and at time of diagnosed septic shock

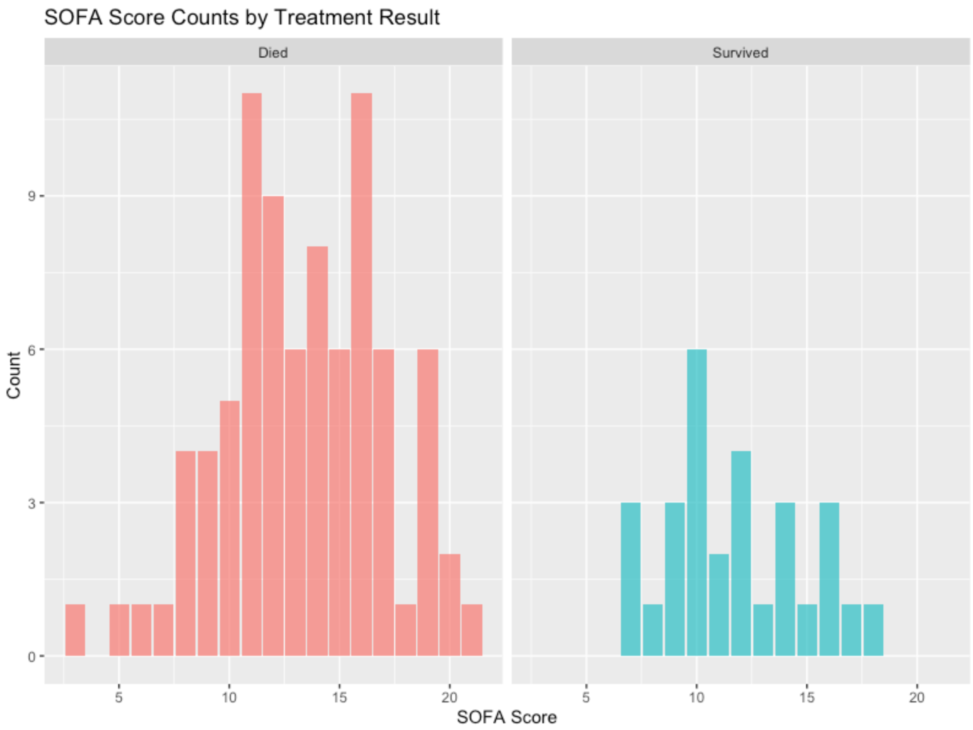

Figure 1. The distribution of Sequential Organ Failure Assessment Score in the study population (Higher scores indicate more severe conditions with more dysfunctional organs)

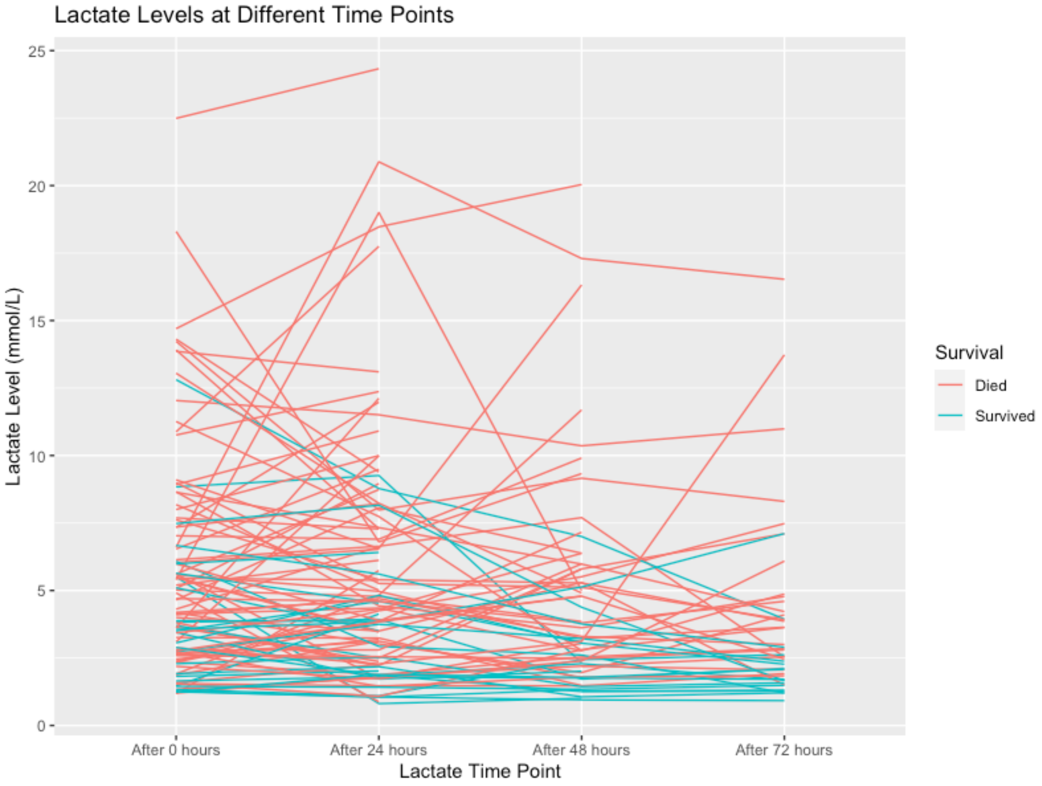

Figure 2. The lactate levels of patients at different post-diagnosis time points (Surviving patients are more likely to have lactate level measured at all time points)

7.2) Variable importance in random forest model

Variables/Biomarkers	Mean Gini Index Decreases
Patient age	**1.35**
The total number of days in hospital	**7.25**
The total number of days in ICU	**1.93**
Mean arterial pressure	0.88
Respiratory rate	1.08
Pulse rate	0.8
Amount of urine	0.21
Albumin level	**1.29**
Procalcitonin level	**2.98**
C-reactive protein level	1.03
Pro b-type natriuretic peptide level	0.79
Creatinine level	1.02
Platete level	0.94
White blood cell level	0.63
Neutrophil level	1.01
Lymphocyte level	0.77
Red blood cell level	0.9
Hemoglobin level	**1.13**
Urea level	**1.32**
Sodium level	0.8
Potassium level	**1.3**
Bilirubin level	0.77
Aspartate aminotransferase level	**1.47**
Alanine transaminase level	0.78
Percent prothrombin	0.82
Fibrinogen level	1.04
D-dimer level	**1.32**
The number of failed organs	0.85
Lactate level at peak	**2.62**
The time of lactate level at peak	0.25
Lactate levels at the time of diagnosed septic shock (T0), 24h after the diagnosis (T1), 48h after the diagnosis (T2), and 72h after the diagnosis (T3)	**T0: 1.18** **T1: 1.56** **T2: 3.56** **T3: 6.04**

Table 2. The mean Gini Index decreases in each variable in the random forest model with a set seed to ensure reproducibility (bolded variables indicate high importance)

7.3) The ROC curves of different models

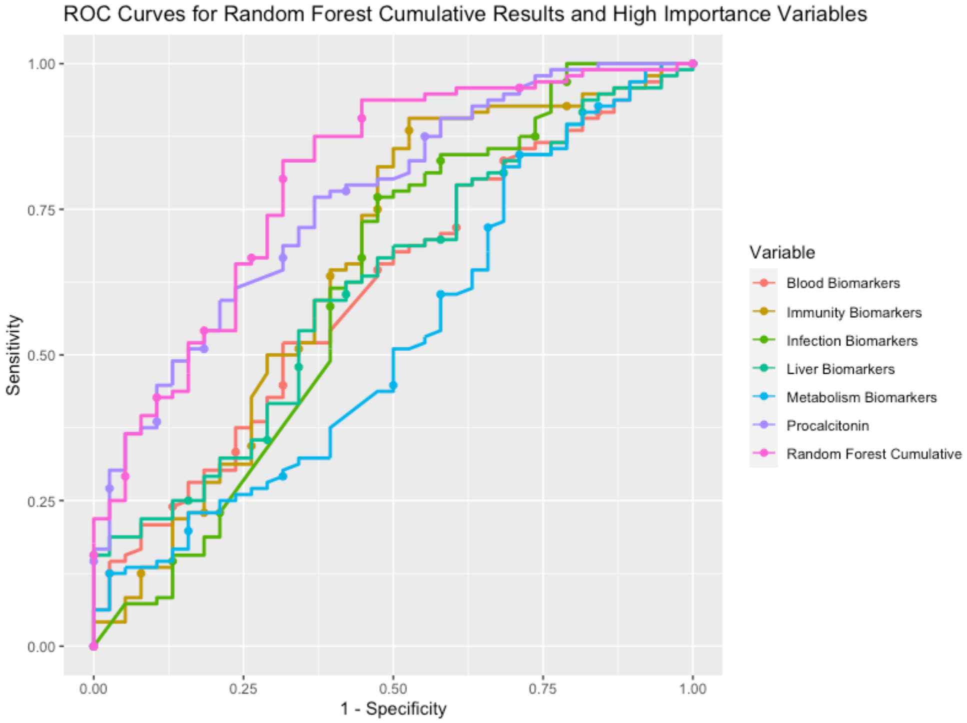

Figure 3. The ROC curves of logistic regression models of different variable combinations in comparison to the random forest model

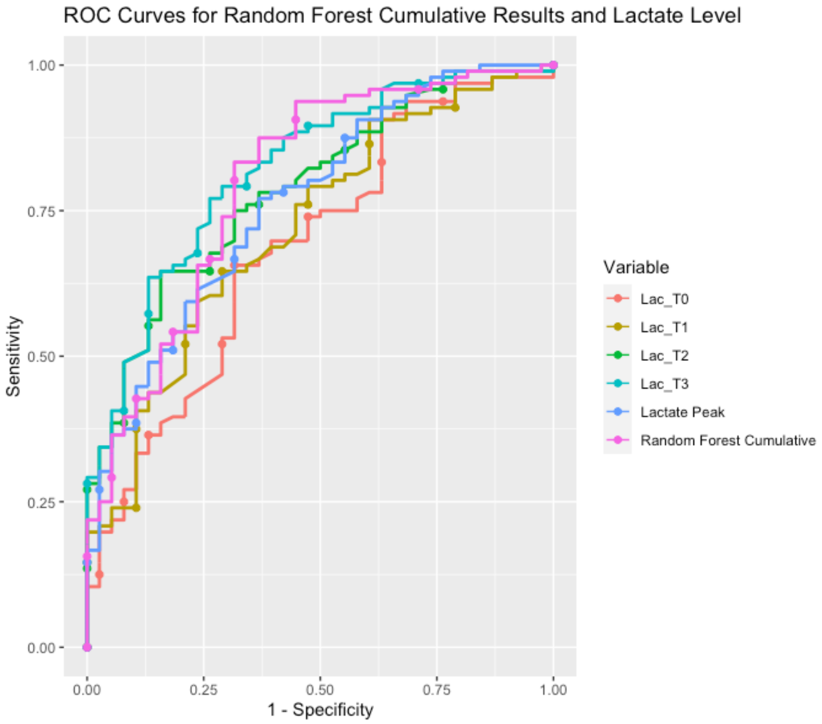

Figure 4. The ROC curves of logistic regression models of different lactate time points in comparison to the random forest model

Model	AUC	Accuracy
Random forest	0.8074	81.34%
Lactate peaks	0.7652	76.12%
Procalcitonin	0.6168	71.64%
Infection	0.6571	71.64%
Immunity	0.6168	69.4%
Liver	0.5276	70.9%
Blood	0.5738	71.64%
Metabolism	0.6121	70.9%
Lactate level at time of diagnosis (T0)	0.6882	70.15%
Lactate level 24h post-diagnosis (T1)	0.7211	71.64%
Lactate level 48h post-diagnosis (T2)	0.7880	76.12%
Lactate level 72h post-diagnosis (T3)	0.8211	76.87%

Table 3. The out-of-sample AUC value and the accuracy rate of each model