Clinical DS – Final Assignment – Report

In this assignment, I reproduced a study published by Zhao et. al, that showed that platelet counts in sepsis patients can be a predictor for one-year survival. However, in their paper they used the MIMIC-III dataset, and in this exercise, I used the MIMIC-IV dataset, therefore, there can be differences in the results. This report describes my final analysis results and discusses the differences with the original paper.

Results

Participants

First, I used the MIMIC-IV dataset and selected the patients according to the original paper. I included only patients with the following criterions:

1. Patients with sepsis – I defined sepsis according to Martin criteria. According to their paper Martin criteria defines sepsis following ICD-9-CM codes that referring to infections and organ failures. The codes are: 038, 020.0, 790.7, 117.9, 112.5, 112.81. I used the diagnoses_icd table to collect all patients that were diagnosed with these codes. For all these patients with these codes I added a column which indicates if they have sepsis (=1) or not (=0).
2. Patients that their age >= 18 and <= 89 – I used the age table in mimic derived. It calculates for each patient its age on the given admission time. This age is more accurate than the anchor_age in the admission table because it calculates for each admission the current age based on the anchor age.
3. Patients that their admission time in the ICU > 24 hours – I used the icustay table to get only patients that were in the ICU for more than 24 hours (based on intime and out time).
4. Patients with SOFA score >= 2 – According to sepsis 3.0 definition. I used the first_day_sofa table to get patients with sofa score >=2.
5. Patients with vital signs – I used the first_day_vitalsign table to get patients with the required vital signs: HR, SBP, MBP, DBP, RR, T.
6. Patients with blood platelets data - I used the first_day_lab table to get patients that have blood platelets values.

Figure 1 shows the flow of the patients through all these inclusion exclusion criterions.

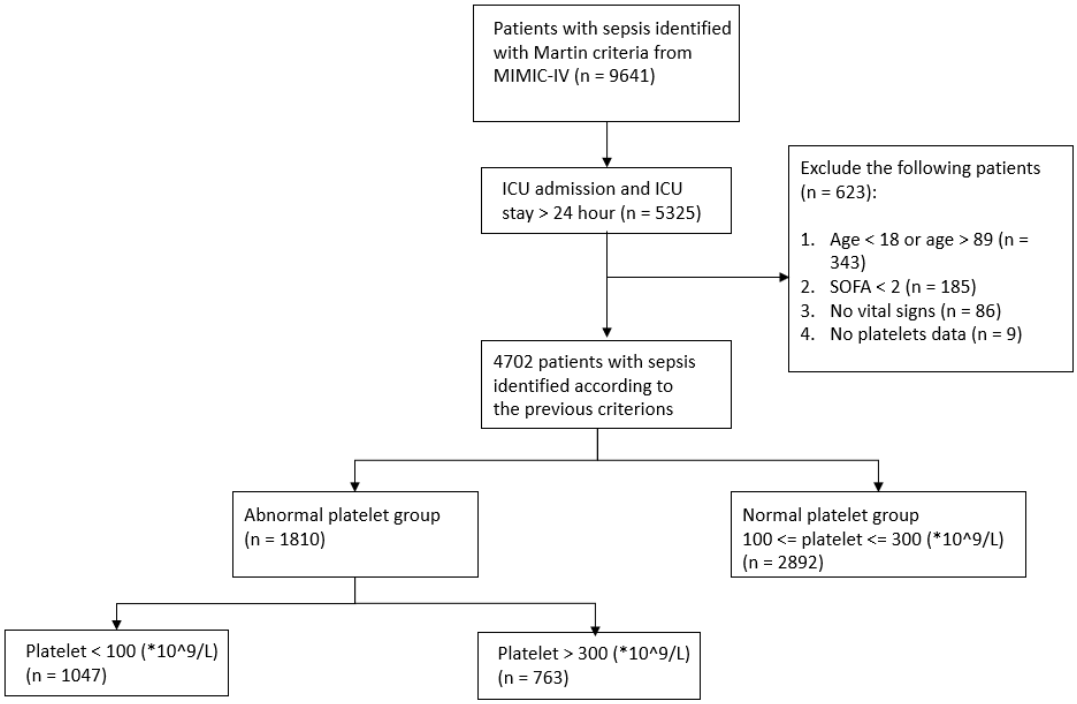

Figure 1: Flow chart of patient selection

After the selection, 4702 patients fulfilled these criterions. For each patient I extracted the following features, based on table 1 in the original paper, from its **first** ICU stay where he fulfilled all the criterions. The worst scores (min/max) of laboratory parameters (Except sodium, potassium and glucose where their maximum and minimum values were retrieved), as well as the mean value of vital signs during the first 24 hours of ICU admission were used:

1. Demographics - Age, Gender, Ethnicity, Admission type.
2. Vital signs - Heart Rate (HR), Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Mean blood pressure (MBP), Respiratory rate (RR) and Temperature (T), SpO2.
3. Laboratory test results – Creatinine (max), Glucose (min, max), Hemoglobin (min), Platelets (min), PTT (max), INR (max), PT (max), BUN (max), WBC (max), Potassium (min, max), Sodium (min, max).
4. Risk scores - SASPII, SOFA, GCS (min).
5. Outcomes – Hospital's LOS, One year survival binary (bin), One year survival continuous (con), In hospital mortality (binary).

- The worst scores of laboratory test results selected according to this supplemental material of the original paper.
- Vital signs retrieved from the first_day_vitalsign table (mimic-derived).
- Laboratory test results retrieved from the first_day_lab table (mimic-derived).
- Risk scores retrieved from the first day of the ICU admission (first_day_gcs, first_day_sofa and sapsii tables in mimic-derived).
- Hospital's LOS taken from icustay_detail table.

- One year survival continuous calculated using the difference between dod time and admission time.
- One year survival binary set to 1 if the difference between dod and admission time is less than 365 days.
- In hospital mortality is equals to hospital_expire_flag.

After the feature extraction, I found out that 359 patients had missing values, so I replaced these missing values with the mean value of each feature according to their paper. The percentage of missing values for PTT, INR and PT is between 6-7 percent, and the other features (SpO2, Hemoglobin, Creatinine, BUN, WBC, Potassium, Sodium, Glucose) missing less than 1%.

In addition, 3348 patients did not die in this one-year interval and 1354 patients did die.

Next, I divided the patients into three groups based on their platelets level: thrombocytopenia group (platelet < 100 * 10^9/L) has 1047 patients, normal platelets group (100 <= Platelet <= 300) has 2892 patients and thrombocytosis group (Platelet > 300) has 763 patients.

The patient characteristics and outcomes are summarized in the following table. For categorical features I used frequency and percentage, and for non-normal features I used the median and IQR. A non-parametric test (Kruskal Wallis test) was applied for data with non-normal distribution, and Categorical data were compared using the Pearson Chi-squared test:

		Grouped by platelet_group				
		normal	thrombocytopenia	thrombocytosis	P-Value	Test
n		2892	1047	763		
Age, median [Q1,Q3]		68.0 [57.0,79.0]	61.0 [52.0,73.0]	67.0 [56.0,77.0]	<0.001	Kruskal-Wallis
Gender, n (%)	F	1236 (42.7)	453 (43.3)	361 (47.3)	0.074	Chi-squared
	M	1656 (57.3)	594 (56.7)	402 (52.7)		
Ethnicity, n (%)	AMERICAN INDIAN/ALASKA NATIVE	5 (0.2)	6 (0.6)		0.196	Chi-squared (warning: expected count < 5)
	ASIAN	100 (3.5)	37 (3.5)	19 (2.5)		
	BLACK/AFRICAN AMERICAN	304 (10.5)	98 (9.4)	79 (10.4)		
	HISPANIC/LATINO	108 (3.7)	49 (4.7)	21 (2.8)		
	OTHER	128 (4.4)	52 (5.0)	33 (4.3)		
	UNABLE TO OBTAIN	42 (1.5)	20 (1.9)	14 (1.8)		
	UNKNOWN	199 (6.9)	78 (7.4)	61 (8.0)		
	WHITE	2006 (69.4)	707 (67.5)	536 (70.2)		
Admission_type, n (%)	DIRECT EMER.	76 (2.6)	83 (7.9)	19 (2.5)	<0.001	Chi-squared (warning: expected count < 5)
	DIRECT OBSERVATION	1 (0.0)				
	ELECTIVE	24 (0.8)	16 (1.5)	5 (0.7)		
	EU OBSERVATION	1 (0.0)				
	EW EMER.	2127 (73.5)	671 (64.1)	540 (70.8)		
	OBSERVATION ADMIT	7 (0.2)	3 (0.3)	1 (0.1)		
	SURGICAL SAME DAY ADMISSION	56 (1.9)	17 (1.6)	16 (2.1)		
	URGENT	600 (20.7)	257 (24.5)	182 (23.9)		
HR, median [Q1,Q3]		90.6 [78.5,102.7]	93.6 [81.4,106.0]	93.4 [81.8,105.0]	<0.001	Kruskal-Wallis
SBP, median [Q1,Q3]		109.3 [102.1,119.8]	107.3 [99.9,117.5]	109.4 [102.2,120.7]	<0.001	Kruskal-Wallis
DBP, median [Q1,Q3]		59.2 [53.6,65.7]	59.0 [52.7,65.8]	58.7 [53.0,65.4]	0.472	Kruskal-Wallis
MBP, median [Q1,Q3]		72.5 [67.1,79.1]	71.7 [66.2,78.6]	72.2 [67.3,79.4]	0.059	Kruskal-Wallis
SpO2, median [Q1,Q3]		96.9 [95.6,98.3]	96.9 [95.5,98.2]	97.2 [95.7,98.4]	0.089	Kruskal-Wallis
RR, median [Q1,Q3]		20.6 [17.8,23.4]	20.6 [17.4,24.1]	20.8 [17.8,23.8]	0.636	Kruskal-Wallis
T, median [Q1,Q3]		36.9 [36.6,37.4]	36.8 [36.4,37.3]	36.9 [36.6,37.3]	<0.001	Kruskal-Wallis
Cr, median [Q1,Q3]		1.4 [1.0,2.4]	1.7 [1.0,2.8]	1.3 [0.8,2.1]	<0.001	Kruskal-Wallis
Glucose_min, median [Q1,Q3]		108.0 [90.0,132.0]	106.0 [85.0,129.0]	108.0 [91.0,136.5]	0.001	Kruskal-Wallis
Hemoglobin, median [Q1,Q3]		9.9 [8.8,11.3]	8.7 [7.5,10.2]	9.2 [8.2,10.4]	<0.001	Kruskal-Wallis

Platelets, median [Q1,Q3]		175.0 [136.0,223.0]	59.0 [36.0,78.0]	369.0 [330.0,441.0]	<0.001	Kruskal-Wallis
PTT, median [Q1,Q3]		36.2 [30.2,46.3]	41.3 [33.1,57.0]	35.1 [29.4,46.3]	<0.001	Kruskal-Wallis
INR, median [Q1,Q3]		1.4 [1.2,1.9]	1.7 [1.3,2.3]	1.4 [1.2,1.9]	<0.001	Kruskal-Wallis
PT, median [Q1,Q3]		15.9 [13.7,20.7]	18.5 [14.8,24.4]	15.6 [13.8,20.7]	<0.001	Kruskal-Wallis
BUN, median [Q1,Q3]		29.0 [18.0,48.0]	34.0 [21.0,54.0]	26.0 [16.5,45.0]	<0.001	Kruskal-Wallis
WBC, median [Q1,Q3]		14.5 [10.0,20.2]	10.6 [5.7,17.1]	18.2 [13.6,24.8]	<0.001	Kruskal-Wallis
Potassium_min, median [Q1,Q3]		3.7 [3.4,4.2]	3.6 [3.3,4.2]	3.9 [3.5,4.3]	<0.001	Kruskal-Wallis
Potassium_max, median [Q1,Q3]		4.4 [4.0,5.0]	4.3 [3.9,5.0]	4.5 [4.1,5.0]	0.001	Kruskal-Wallis
Sodium_min, median [Q1,Q3]		136.5 [133.0,139.0]	136.0 [132.0,139.0]	136.0 [133.0,139.0]	<0.001	Kruskal-Wallis
Sodium_max, median [Q1,Q3]		140.0 [137.0,142.0]	139.0 [136.0,142.0]	139.0 [136.0,142.0]	<0.001	Kruskal-Wallis
SAPSII, median [Q1,Q3]		41.0 [32.0,51.0]	46.0 [35.0,57.0]	39.0 [31.0,50.0]	<0.001	Kruskal-Wallis
SOFA, median [Q1,Q3]		7.0 [4.0,10.0]	10.0 [7.0,14.0]	6.0 [4.0,8.0]	<0.001	Kruskal-Wallis
GCS, median [Q1,Q3]		13.0 [9.0,14.0]	13.0 [7.0,15.0]	13.0 [9.0,14.0]	0.438	Kruskal-Wallis
Los_hospital, median [Q1,Q3]		10.0 [6.0,19.0]	12.0 [6.0,23.0]	12.0 [7.0,21.0]	<0.001	Kruskal-Wallis
One_year_survival_con, median [Q1,Q3]		13.0 [5.0,27.0]	12.0 [4.0,33.0]	12.0 [5.0,28.0]	0.846	Kruskal-Wallis
One_year_survival_bin, n (%)	0	2191 (75.8)	608 (58.1)	549 (72.0)	<0.001	Chi-squared
	1	701 (24.2)	439 (41.9)	214 (28.0)		
In_hospital_mortality, n (%)	0	2312 (79.9)	666 (63.6)	585 (76.7)	<0.001	Chi-squared
	1	580 (20.1)	381 (36.4)	178 (23.3)		

- The table shows glucose_min and **not** glucose_max because according to their paper and the values they used they referred to Glucose as Glucose_min.

Table 1: Baseline characteristics, vitalsigns, laboratory parameters and outcomes of patients.

According to the table, differences in age, admission_type, HR, SBP, T, Cr, Glucose, Hemoglobin, Platelets, PTT, INR, PT, BUN, WBC, Potassium min and max, Sodium min and max between the 3 groups were statistically significant. In addition, the score of SAPSII and SOFA were significantly higher in thrombocytopenia group than patients in the normal or thrombocytosis groups. Moreover, hospital mortality and 1 year mortality in thrombocytopenia group (36.4%, 41.9%) and thrombocytosis group (23.3%, 28%) were higher than patients in the normal platelet group (20.1%, 24.2%).

Survival Analysis

I compared the survival rate of two groups: sepsis patients with normal platelets vs. sepsis patients with abnormal (low or high) platelets count. The following Kaplan-Meier curve shows the survival probabilities from observed survival times. The duration is the number of days until the patient die in one-year interval and the event that observed is one year survival (1 if the patient died else 0). The graph shows that patients in the normal platelet group had better long-term survival rates. The probability of a patient to survive longer than 1 year in the normal platelet group is approximately 75%, but in the abnormal group the probability is 63%. However, in order to investigate if the difference between the groups is significant, I used a log-rank hypothesis test. The null hypothesis indicates that the groups are identical, and the alternative hypothesis indicates that they are not. The p-value were less than 0.05 hence I rejected the null hypothesis that the survival curves of patients from the normal and abnormal groups are identical.

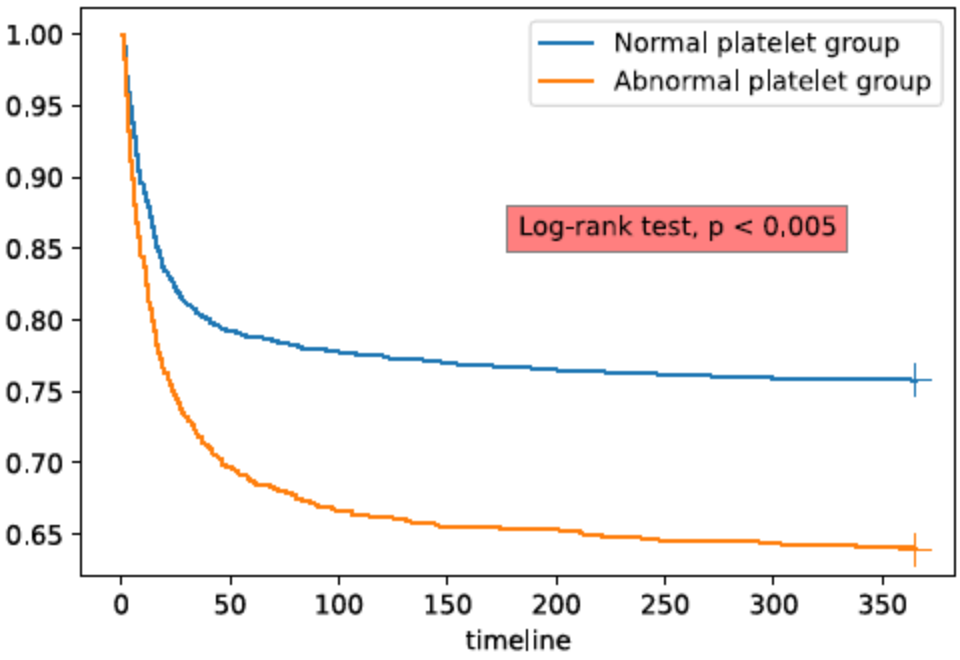

Figure 2: Kaplan-Meier Curve.

The Cox Proportional-Hazards Model was used to analyze the independent effects of various parameters on mortality. I selected to use the features they used in their paper according to table 2 (Except organ failure features) – all baseline and laboratory features, i.e., Age, sex, Cr, BUN, glucose, hemoglobin, platelets, PTT, INR, PT, WBC, sodium, potassium. But I also added the risk scores of SAPSII, SOFA and GCS because in my results their values were significantly different between the three groups (except GCS) and it improved the performance of the model. The results of the multivariate analysis can be seen in the following table:

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Age	0.01	1.01	0.00	0.01	0.02	1.01	1.02	0.00	5.49	<0.005	24.59
Gender	-0.09	0.91	0.06	-0.20	0.02	0.82	1.02	0.00	-1.67	0.10	3.39
Cr	-0.06	0.94	0.02	-0.10	-0.02	0.90	0.98	0.00	-3.16	<0.005	9.31
BUN	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.00	1.71	0.09	3.53
Glucose_min	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.00	-1.54	0.12	3.02
Hemoglobin	-0.04	0.96	0.01	-0.07	-0.01	0.93	0.99	0.00	-2.76	0.01	7.42
plat_thrombocytopenia	0.39	1.47	0.07	0.25	0.52	1.28	1.69	0.00	5.42	<0.005	24.00
plat_thrombocytosis	0.22	1.25	0.08	0.07	0.38	1.07	1.46	0.00	2.82	<0.005	7.71
INR	0.02	1.02	0.05	-0.07	0.11	0.93	1.12	0.00	0.39	0.70	0.52
PTT	0.00	1.00	0.00	0.00	0.01	1.00	1.01	0.00	5.42	<0.005	24.02
PT	0.01	1.01	0.00	-0.00	0.01	1.00	1.01	0.00	1.14	0.25	1.99
WBC	0.00	1.00	0.00	-0.00	0.01	1.00	1.01	0.00	1.23	0.22	2.21
Sodium_min	0.02	1.02	0.01	0.00	0.04	1.00	1.04	0.00	2.05	0.04	4.62
Sodium_max	-0.03	0.97	0.01	-0.05	-0.01	0.95	0.99	0.00	-3.20	<0.005	9.48
Potassium_min	0.24	1.27	0.06	0.13	0.35	1.14	1.42	0.00	4.28	<0.005	15.68
Potassium_max	0.00	1.00	0.04	-0.08	0.08	0.93	1.09	0.00	0.07	0.94	0.09
SAPSII	0.02	1.02	0.00	0.02	0.03	1.02	1.03	0.00	8.72	<0.005	58.32
SOFA	0.06	1.06	0.01	0.04	0.08	1.04	1.09	0.00	6.07	<0.005	29.52
GCS	-0.04	0.96	0.01	-0.06	-0.03	0.94	0.97	0.00	-5.09	<0.005	21.45

Table 2: Results of The Cox Proportional-Hazards Model

According to the results, age, Cr, Hemoglobin, platelets, PTT, Sodium_min, Sodium_max, Potassium_min, Potassium_max and risk scores remained independent prognostic factors for sepsis patients ($p< 0.05$).

Model Development

I performed univariate analysis using a logistic regression model for each variable from the list in Table 2 (except organ failure) and computed the p-value, the odds ratio and 95% confidence intervals (the exponent of the CI based on the paper) of the variable for 1-year survival. The results summarized in the following table:

	[0.025	0.975]	Odds Ratio	P-value
Age	0.987	0.989	0.988	0.0
Gender	0.360	0.426	0.391	0.0
Cr	0.779	0.820	0.799	0.0
BUN	0.987	0.990	0.988	0.0
Glucose_min	0.992	0.993	0.993	0.0
Hemoglobin	0.903	0.915	0.909	0.0
plat_normal	0.294	0.348	0.320	0.0
plat_thrombocytopenia	0.639	0.816	0.722	0.0
plat_thrombocytosis	0.333	0.456	0.390	0.0
INR	0.743	0.788	0.765	0.0
PTT	0.987	0.989	0.988	0.0
PT	0.971	0.977	0.974	0.0
WBC	0.959	0.966	0.963	0.0
Sodium_min	0.993	0.994	0.993	0.0
Sodium_max	0.993	0.994	0.994	0.0
Potassium_min	0.789	0.816	0.802	0.0
Potassium_max	0.824	0.846	0.835	0.0
SAPSII	0.985	0.988	0.987	0.0
SOFA	0.939	0.952	0.945	0.0
GCS	0.909	0.919	0.914	0.0

Table 3: Univariate analysis results

The p-value of all the features is less than 0.05 hence I used all the features in the multivariate logistic regression model.

Model Specification

After the univariate analysis, I performed a multivariate analysis using all the features that their p-value is less than 0.05. I used a Logistic Regression model and computed the 95% confidence interval, the odds ratio and the p-value of each variable. According to the results, age, Cr, BUN, Hemoglobin, INR, PTT, Sodium_max, Potassium_min and risk scores are significant. I assumed that the reason that the p-value of platelets is greater than 0.05 is because I included the risk

scores in my model. This is because these risk scores already include the platelets levels in their score, hence the platelet values are redundant. When I remove the risk scores from the model the p-value of platelet is significant, but the model calibration drops a little bit. Therefore, I decided to remove the risk scores from the model because I test the prognostic factor of platelets.

	[0.025	0.975]	Odds Ratio	P-value
Age	1.014	1.024	1.019	0.000
Gender	0.750	0.988	0.861	0.033
Cr	0.895	0.982	0.938	0.006
BUN	1.006	1.012	1.009	0.000
Glucose_min	0.998	1.001	0.999	0.415
Hemoglobin	0.903	0.969	0.936	0.000
plat_normal	0.014	0.644	0.096	0.016
plat_thrombocytopenia	0.033	1.456	0.220	0.116
plat_thrombocytosis	0.016	0.722	0.107	0.022
INR	0.929	1.229	1.068	0.354
PTT	1.008	1.011	1.009	0.000
PT	0.989	1.017	1.003	0.703
WBC	1.001	1.012	1.006	0.017
Sodium_min	0.981	1.027	1.004	0.760
Sodium_max	0.960	1.007	0.983	0.153
Potassium_min	1.197	1.582	1.376	0.000
Potassium_max	1.027	1.240	1.128	0.012

(a)

	[0.025	0.975]	Odds Ratio	P-value
Age	1.008	1.019	1.013	0.000
Gender	0.780	1.043	0.902	0.164
Cr	0.864	0.959	0.910	0.000
BUN	1.001	1.008	1.004	0.020
Glucose_min	0.998	1.001	0.999	0.357
Hemoglobin	0.888	0.957	0.922	0.000
plat_normal	0.094	5.872	0.742	0.777
plat_thrombocytopenia	0.161	9.934	1.264	0.824
plat_thrombocytosis	0.120	7.511	0.949	0.961
INR	0.897	1.196	1.036	0.633
PTT	1.004	1.009	1.006	0.000
PT	0.991	1.020	1.006	0.459
WBC	0.995	1.006	1.000	0.968
Sodium_min	0.995	1.046	1.020	0.124
Sodium_max	0.933	0.983	0.957	0.001
Potassium_min	1.187	1.594	1.376	0.000
Potassium_max	0.895	1.099	0.992	0.875
SAPSII	1.023	1.037	1.030	0.000
SOFA	1.035	1.093	1.063	0.000
GCS	0.907	0.945	0.926	0.000

(b)

Table 4: Results of multivariate Logistic Regression (a) without risk scores (b) with risk scores

Model Performance

The performance of the multivariate Logistic Regression model was assessed by discrimination and calibration. The discriminative ability of the model was determined by the area under the receiver operating characteristic curve, which ranged from 0.5 (no discrimination) to 1 (perfect discrimination). The calibration of the prediction model was performed by a visual calibration plot comparing the predicted and actual probability of prognosis of sepsis patients. To test the performance, I split the data into 80% train and 20% test in a stratified manner to keep the distribution of the target variable the same.

I tested the results one time using all the features (including the risk scores) and one time without risk scores.

The results with the risk scores show (Figure 3) that the multivariate Logistic Regression showed a robust discrimination, with an area under the receiver operating characteristic curve of 0.816, with confidence Interval of 95% [0.786, 0.846]. The calibration curves of the probability of the prognosis of sepsis patients showed a good agreement between the probability as predicted by the model and the actual probability.

In addition, the results without the risk scores show (Figure 4) that the multivariate Logistic Regression showed smaller discrimination, with an area under the receiver operating characteristic curve of 0.744, with confidence Interval of 95% [0.708, 0.778]. The calibration curves of the probability of the prognosis of sepsis patients showed a good agreement (except for the last bin) between the probability as predicted by the model and the actual probability.

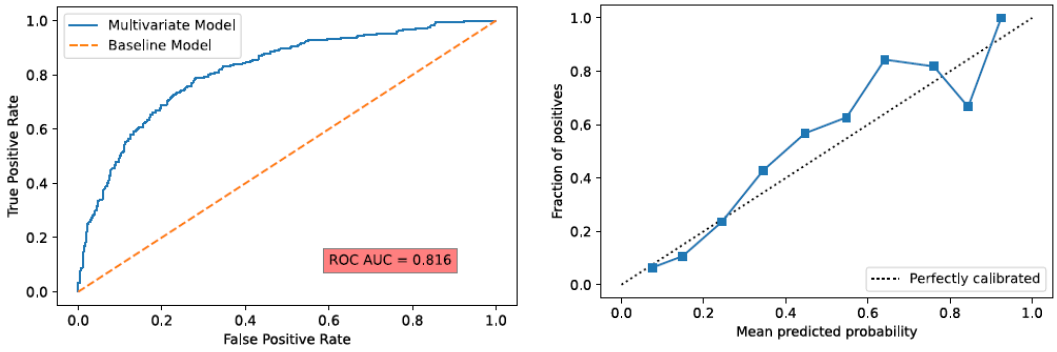

Figure 3: The ROC AUC and Calibration of multivariate Logistic Regression with risk scores

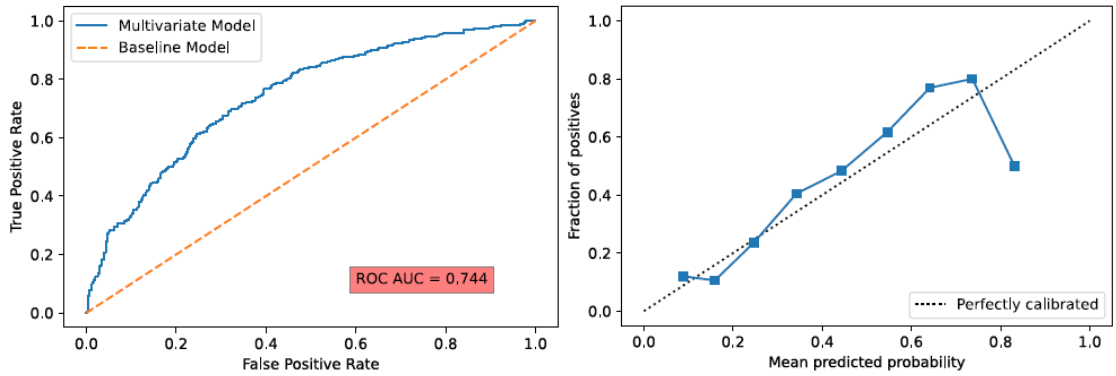

Figure 4: The ROC AUC and Calibration of multivariate Logistic Regression without risk scores

Discussion

This implementation has multiple differences compared to the original paper. First, the dataset is different, they used the MIMIC-III dataset which contains records from 2001 to 2012 and I used the MIMIC-IV dataset which its records are from 2008-2019. As a result, the number of patients used in this analysis is different and the results are different. Moreover, in their paper they did not mention exactly the procedure how they extracted patients with sepsis (the tables they used and the SQL queries) therefore it could add an additional difference. In addition, in the results section I had four differences with their table 1: the p-value of admission type and potassium_min is less than 0.05 in contrast to their paper, and the RR and GCS are greater than 0.05. According to the Kaplan-Meier graph the survival rates in my analysis are better than their analysis it could be because technology advancement (the MIMIC-IV data is newer than the

MIMIC-III) or because the new definition of sepsis 3.0 (2016). Another difference is the models that were used, I used a Logistic Regression model while they used a nomogram. One last thing that is important, on the multivariate Logistic Regression results I got that the p-value of thrombocytopenia group is not significant in contrast to their paper, so it may impact the conclusion that platelets can be a prognostic marker for sepsis.

All their limitations are suitable for this study, but it has also additional limitations. I used all the features that were used in the original paper even when part of them were not significant in my results. As I said, it was hard to set the same population according to their paper because they missed a lot of information (I have also tried to extract the same patients from the MIMIC-III, in order to get the same amount, but it was close and not exactly like their results). Lastly, the selected population contained only 4702 patients so it may be biased.

However, like the conclusion of their paper, my results show that patients with thrombocytopenia had a higher SAPSII and SOFA score and it strengthens the results of previous studies that the less platelets level, the more severe the disease.

Finally, This model can be used in early stages of sepsis when the platelet levels are low and as a result the patient can get an appropriate treatment as early as possible. However, it requires more investigation and expert suggestions before using this model.