

111 Biostatistics Final Project

In-hospital-mortality prediction

ICU-admitted HF PATIENTs

Team Members:

311554012 梁家瑤 / 10501125 陳祺侑



目錄

Topic One Introduction and Preparation Work	3
Topic Two Survival Analysis	8
Topic Three Categorical Data Analysis	19
Topic Four Conclusion	27
Topic Five References & Work Assignment	29

Topic One Introduction and Preparation Work

I. Main Goal

In our final project, we plan to figure out the **associating factor of mortality in heart failure patients in intensive care units**.

II. About Dataset

MIMIC-III (Medical Information Mart for Intensive Care):

a large, freely available database, developed by Alistair Johnson, Tom Pollard, Roger Mark et al., affiliated with the Laboratory for Computational Physiology at MIT, comprising de-identified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center (at Boston) between 2001 and 2012.

The dataset (from the Kaggle website, Zhou, Jingmin et al.) that we used was extracted from the **MIMIC-III dataset for intensive care units (ICU)-admitted HF(heart failure) patients**, and contained information on 1177 patients about demographic characteristics, vital signs, and laboratory values data. The source of this extracted dataset is Kaggle.¹

III. Variables

In our dataset, there are **49 independent variables and 1 dependent variable**, which is the in-hospital mortality. The independent variables are listed and categorized as following:

	Variables	Type	Note
Demographic features (人口特徵)	Age	int	
	Gender	int (binary)	1: Male, 2: Female
	BMI	float	
Cardiovascular features (心血管)	Hypertensive	int (binary)	0: No, 1: Yes
	Atrial fibrillation	int (binary)	0: No, 1: Yes
	CHD with no MI	int (binary)	0: No, 1: Yes
	NT-proBNP	int	
	Creatine kinase	float	
	EF	int	Ejection fraction
Vital signs (生命徵象)	Systolic blood pressure	float	
	Diastolic blood pressure	float	

¹ Kaggle / In-hospital-mortality-prediction (<https://bit.ly/3fENy9I>); Kaggle / Clean DD (<https://bit.ly/3zRNOsX>)



	Heart rate	float	
	Respiratory rate	float	
	Temperature	float	
	SPO2	float	
Hematologic features (血液學)	Hematocrit	float	
	RBC	float	
	MCH	float	$10\text{Hb(g/dL)}/\text{RBC count}$
	MCHC	float	$100\text{Hb(g/dL)}/\text{Hct}$
	MCV	float	$10\text{HCT} / \text{RBC count}$
	RDW	float	$\text{Std}_{\text{RBC volume}} * 100 / (\text{Mean MCV})$
	Leukocyte	float	
	Platelets	float	
	Neutrophils	float	
	Basophils	float	
	Lymphocytes	float	
	PT	float	
	INR	float	$(\text{PT}_{\text{patient}}/\text{PT}_{\text{meannormal}})^{\wedge \text{ISI}}$
Renal function (腎臟功能)	Urine output	int	Unit: cc
	Creatinine	float	
	Urea nitrogen	float	
Serum ions (血清離子)	Blood potassium	float	
	Blood sodium	float	
	Blood calcium	float	
	Chloride	float	
	Anion gap	float	$\text{Na}^+ - \text{Cl}^- + \text{HCO}_3^-$
	Magnesium	float	
	pH	float	
	Bicarbonate	float	
Comorbidities (共病症)	Diabetes	int (binary)	0: No, 1: Yes
	Depression	int (binary)	0: No, 1: Yes
	Hyperlipidemia	int (binary)	0: No, 1: Yes
	COPD	int (binary)	0: No, 1: Yes

	deficiencyanemias	int (binary)	(iron deficiency anemia) 0: No, 1: Yes
	Renal failure	int (binary)	0: No, 1: Yes
Others	Glucose	float	
	Lactic acid	float	
	PCO2	float	

As for the dependent variable, **0 represents a patient has been successfully discharged**, and **1 represents a patient expired during hospitalization**.

IV. Feature Extraction

Since MIMIC-III is not fully public and requires credential, the extracted dataset released on Kaggle does not include the entire attribute columns. **Two of the missing columns are "admission time" and "discharge time"**, which are important for survival analysis. We found two partial datasets² on Kaggle and combined them to obtain the admission time and discharge time of each patient.

It is worth to mention that in both admission time and discharge time, the year attribute is encrypted and replaced by a random number for privacy issues. Nonetheless, the month and date are sufficient for us to calculate survival time (admission time-discharge time). **Seven patients are dropped** since we could not find their corresponding admission time and discharge time.

² Same as in 1



	ID	Admission Date	Discharge Date	survival time
0	125047	2108/6/6	2108/6/12	6
1	139812	2186/8/3	2186/8/18	15
2	109787	2194/8/24	2194/8/28	4
3	130587	2123/2/19	2123/2/27	8
4	138290	2187/7/5	2187/7/13	8
...
1165	171130	2159/4/20	2159/4/27	7
1166	101659	2137/2/27	2137/3/19	20
1167	162069	2190/3/26	2190/3/31	5
1168	120967	2154/5/23	2154/6/12	20
1169	107636	2109/7/17	2109/7/22	5

1170 rows × 4 columns

Figure 1: Seven patients are dropped (1077->1070) and the dates are encrypted.

V. Missing value

In our dataset, most of the attribute columns have missing values. Possibly there are two ways to deal with missing values:

- A. **Predicting the values:** A straightforward idea is performing interpolation. However, we need to search for the most similar subject without missing value. Moreover, if there are missing values, that often means the doctor consider the data as normal, which might contradict the results of interpolation.
- B. **Simply drop the subject with at least on missing value:** Dropping subjects with missing values largely **decreases the number of subjects in our dataset to 425 patients**. In our final project, **we choose the second way** to tackle with missing values since we do not think interpolation is reasonable for medical data.

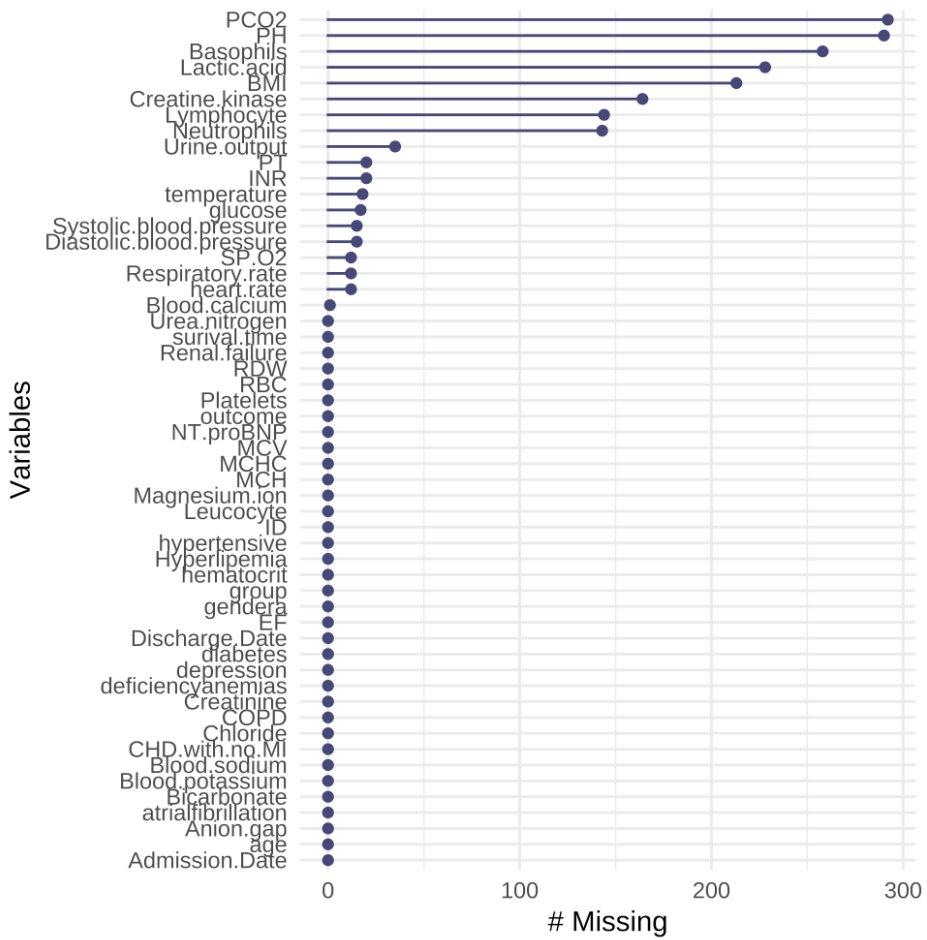


Figure 2: Numbers of subjects having missing value in each variable.

Topic Two Survival Analysis

- I. Main goal: In this part, we would try to construct a **cox regression model** for predicting hazard ratio of a patient.
- II. Data description: As mentioned in the last section, we only utilize 425 of 1070 subjects in this analysis due to missing values.
- III. We would follow the steps to perform survival analysis on our data:
 - A. **Survival curve analysis** and **pick discrete variable** that significantly affects mortality.
 - B. **Construct a cox regression model** with the entire continuous variable and the chosen discrete variable in step 1.
 - C. Perform **backward selection** to discard redundant variables that increases AIC (Akaike information criterion).
 - D. Construct a cox regression model again with the variables chosen from step 3, and **check if the assumption of proportional hazard holds**.
 - E. **Add time interaction** into our cox regression model if needed.
 - F. Get our final cox regression model.
- IV. Step 1: Survival curve analysis.

In our data, we have 8 discrete variables: *Hypertension, Atrial Fibrillation, Anemia, Depression, Coronary heart disease without MI, Diabetes, Hyperlipidemia, and Renal failure*. Therefore, we can draw 8 groups of **Kaplan-Meier survival curve**, where each group contains survival curves for patients with a single variable and without that variable.

We perform **Log-rank test** on each group of Kaplan-Meier survival curves.

- A. **H_0 : There is no difference between two populations in the probability of death.**
- B. **H_1 : There is difference between two populations in the probability of death.**

Figure 3 shows the survival curves of the 8 groups of populations. The green line represents the survival curve for the population with the discrete variable, and the red line stands for the survival curve for the population without the discrete variable. The p-value shown in each subfigure is the log-rank p value. We may find that *AF(Atrial Fibrillation), Anemia, RF(Renal Failure)* have a p-value lower than 0.05. However, we can also there are crossings between survival curves in each group. Since **the crossing often happen in the late period, we may perform Gehan-Breslow**, a variant of log-rank test which values more in early period. The hypothesis is the same, and the testing results are shown in Figure 4. **By Gehan-Breslow, one more variable, DB(Diabetes), showed its significance in mortality**. Notice the crossing of survival curve may also **imply contradiction to the assumption of proportional hazard**, and we would test it further

in our analysis.

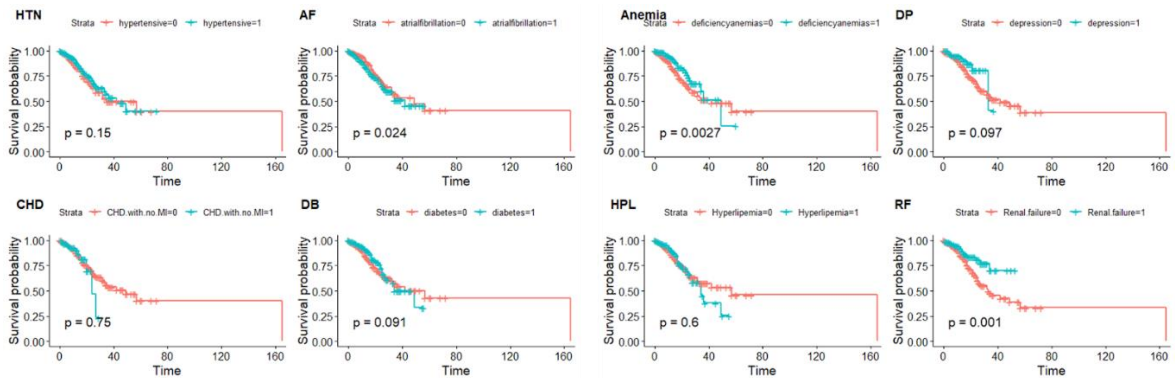


Figure 3. Survival curves across patients with 8 different discrete variables and log-rank test.

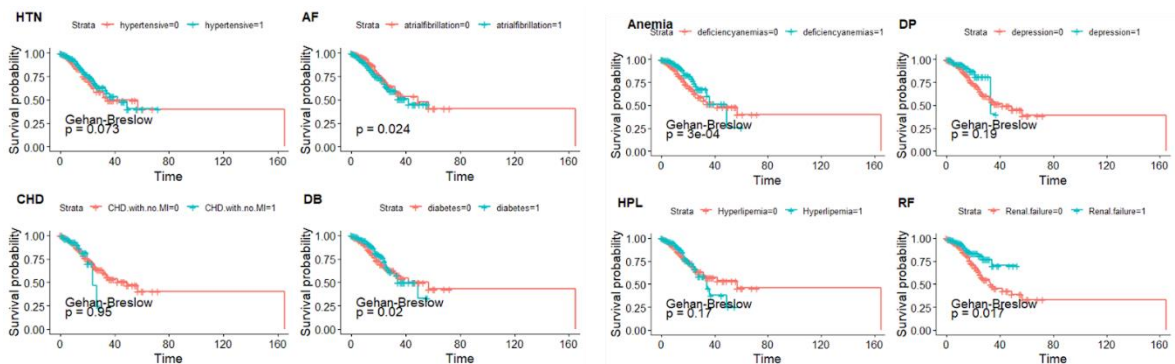


Figure 4. Survival curves across patients with 8 different discrete variables and Gehan-Breslow test.

V. Step 2: Construct a cox regression model with the entire continuous variable and the chosen discrete variable in step 1.

A. The hypothes of Cox regression is:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

And we would like to further explain two metrics in cox regression:

1. Concordance

a. The concordance of a cox regression model is defined as the probability of the prediction going toward the same direction as the actual data.

b. The higher the concordance, the better the fitness of our model.

2. AIC (Akaike information criterion)

a. $AIC = 2k - 2\ln(L)$, where L is the maximized value of likelihood function of the model, and k is the number of variables in the model.

b. The lower the AIC, the better the fitness of our model.

B. Figure 5 shows the cox regression with the entire 38 continuous variables and 4 discrete variables. The concordance of it is 0.914, and the AIC is 554.2759.

explanatory	HR	L95	U95	p
age	1.0256817	0.9925616	1.059907e+00	0.1299898
BMI	0.9773096	0.9316734	1.025181e+00	0.3468650
heart.rate	1.0073966	0.9793235	1.036274e+00	0.6093136
Systolic.blood.pressure	1.0178362	0.9897742	1.046694e+00	0.2152007
Diastolic.blood.pressure	0.9320014	0.8888291	9.772707e-01	0.0036135
Respiratory.rate	1.0802415	0.9920055	1.176326e+00	0.0758415
temperature	0.5970107	0.3332409	1.069562e+00	0.0829342
SP.O2	0.9946122	0.8843570	1.118613e+00	0.9281920
Urine.output	1.0000481	0.9997130	1.000383e+00	0.7784469
hematocrit	0.8914044	0.5096887	1.558994e+00	0.6869015
RBC	3.7093460	0.0279001	4.931610e+02	0.5993003
MCH	0.0212641	0.0002641	1.712340e+00	0.0854773
MCHC	22.9168601	0.4539708	1.156864e+03	0.1175194
MCV	3.8067233	0.8771946	1.651987e+01	0.0742609
RDW	0.9416686	0.7884560	1.124653e+00	0.5070995
Leucocyte	1.0200920	0.9753105	1.066930e+00	0.3851164
Platelets	0.9952290	0.9916759	9.987948e-01	0.0087718
Neutrophils	0.9851053	0.9238886	1.050378e+00	0.6466293
Basophils	1.3206108	0.3094518	5.635813e+00	0.7071930
Lymphocyte	0.9481462	0.8562097	1.049954e+00	0.3062053
PT	1.0183435	0.7428763	1.395957e+00	0.9100649
INR	1.3087774	0.0823880	2.079062e+01	0.8487465

NT.proBNP	1.0000159	0.9999976	1.000034e+00	0.0892213
Creatine.kinase	0.9999633	0.9998949	1.000032e+00	0.2937430
Creatinine	0.5519749	0.2730268	1.115921e+00	0.0980093
Urea.nitrogen	1.0121373	0.9923455	1.032324e+00	0.2311698
glucose	1.0025287	0.9955177	1.009589e+00	0.4806099
Blood.potassium	1.7426826	0.2883890	1.053071e+01	0.5450691
Blood.sodium	0.9354851	0.2080289	4.206783e+00	0.9307162
Blood.calcium	0.6283632	0.2888838	1.366779e+00	0.2412397
Chloride	1.0329662	0.2291570	4.656279e+00	0.9663253
Anion.gap	1.3567803	0.2982975	6.171197e+00	0.6930008
Magnesium.ion	0.7845830	0.2027105	3.036698e+00	0.7253340
PH	2380.6831063	0.2430016	2.332351e+07	0.0972678
Bicarbonate	0.8985014	0.1930126	4.182652e+00	0.8915103
Lactic.acid	0.9086535	0.6515959	1.267122e+00	0.5723541
PCO2	1.0911797	1.0188925	1.168595e+00	0.0125904
EF	0.9924866	0.9673019	1.018327e+00	0.5652250
atrialfibrillation	0.4485601	0.1997267	1.007407e+00	0.0521267
diabetes	0.9735081	0.4508222	2.102199e+00	0.9455016
deficiencyanemias	0.2727997	0.1070204	6.953788e-01	0.0065098
Renal.failure	0.2415436	0.0953598	6.118230e-01	0.0027348

Figure 5

VI. Step3: Perform backward selection

We utilize the step function in R to perform backward selection. In each iteration, the function randomly cancels a variable. **If AIC decreases after this cancellation, then this variable can be considered as a redundant variable** and be dropped from the model. After backward selection, **we choose 21 out of 42 variables**. The updated cox regression results are shown in figure 6. The **concordance dropped to 0.901** after dropping half of independent variables. On the other hand, its **AIC improved to 520.7752**, which should be the benefit of a more compact model.

explanatory	HR	L95	U95	p
age	1.0304689	1.0044575	1.057154e+00	0.0213958
Diastolic.blood.pressure	0.9574329	0.9266684	9.892189e-01	0.0090419
Respiratory.rate	1.0851172	1.0148678	1.160229e+00	0.0167502
temperature	0.6485776	0.3960662	1.062077e+00	0.0853190
MCH	0.0106396	0.0001557	7.269628e-01	0.0350382
MCHC	43.8859352	1.0163420	1.895007e+03	0.0490219
MCV	4.4705978	1.1374725	1.757075e+01	0.0319997
Platelets	0.9960298	0.9933307	9.987362e-01	0.0040608
INR	1.4371329	1.1284306	1.830286e+00	0.0032899
NT.proBNP	1.0000135	0.9999982	1.000029e+00	0.0838123
Creatinine	0.5515854	0.3132461	9.712697e-01	0.0393084
Urea.nitrogen	1.0140044	0.9981379	1.030123e+00	0.0839277
Blood.potassium	2.0690203	0.9645299	4.438271e+00	0.0618712
Blood.calcium	0.5928569	0.3205305	1.096555e+00	0.0956732
Anion.gap	1.3803921	1.1989805	1.589252e+00	0.0000073
PH	9519.1691846	3.9067332	2.319446e+07	0.0213097
Bicarbonate	0.8838551	0.7670957	1.018387e+00	0.0876506
PCO2	1.0870029	1.0272857	1.150191e+00	0.0038068
atrialfibrillation	0.4664923	0.2337871	9.308258e-01	0.0305153
deficiencyanemias	0.2566877	0.1173547	5.614483e-01	0.0006604
Renal.failure	0.1922472	0.0862441	4.285394e-01	0.0000553

Figure 6

I. Step 4: Check if the assumption of proportional hazard holds

A. How do we test the assumption of proportional hazard?

1. Scaled Schoenfeld residuals for each variable

a. It is an estimate of the coefficient of the covariate over time. Intuitively, we hope that the residuals do not vary across times.

b. Derivation

1. It is derived from the partial likelihood of survival information (As we have learned in class, including the whole likelihood makes the coefficient estimation intractable, so we discard some of the terms to form a partial likelihood).

$$\begin{aligned}
 & \Pr(B_1, A_1, \dots, B_D, A_D) = \\
 & \Pr(B_1) \Pr(A_1 | B_1) \Pr(B_2 | B_1, A_1) \Pr(A_2 | B_1, A_1, B_2) \dots \Pr(B_D | \dots, B_{D-1}, A_{D-1}) \Pr(A_D | \dots, A_{D-1}, B_{D-1}, B_D) \\
 & = \prod_{j=1}^D \Pr(B_j | B^{(j-1)}, A^{(j-1)}) \Pr(A_j | B^{(j)}, A^{(j-1)}) \\
 & = \left\{ \prod_{j=1}^D \Pr(B_j | B^{(j-1)}, A^{(j-1)}) \right\} \left\{ \prod_{j=1}^D \Pr(A_j | B^{(j)}, A^{(j-1)}) \right\} \quad \text{Partial likelihood}
 \end{aligned}$$

Full likelihood of survival information

2. Then we take log over partial likelihood and rewrite it as: $\ln p$

$$\ell_p = \sum_j \beta^T x_{ij} - \sum_j \log(\sum_{k \in R_j} \exp(\beta^T x_k))$$

Log partial likelihood of survival information

3. If we take the partial derivative of ℓ_p over β_1 :

$$\frac{d\ell_p}{d\beta_1} = \sum_j x_{1ij} - \sum_j \frac{\sum_{k \in R_j} x_{1k} \exp \beta^T x_k}{\sum_{k \in R_j} \exp \beta^T x_k}$$

4. The Schoenfeld residual can be defined as r_{s1j} at time j , where $\hat{\beta}$ is the coefficient estimate when setting log partial likelihood to zero:

$$r_{s1j} = x_{1ij} - \frac{\sum_{k \in R_j} x_{1k} \exp \hat{\beta}^T x_k}{\sum_{k \in R_j} \exp \hat{\beta}^T x_k}$$

5. Finally, we may scale r_{s1j} by the covariance of $\hat{\beta}$ to obtain the scaled Schoenfeld residual. **The expectation over the scaled Schoenfeld residual would simply be the log hazard ratio of variable 1.**
6. Qualitatively, for each variable X_k , we may calculate its scaled Schoenfeld residual at each time j . We **expect these residuals being time invariant if we plot them with respect to time.** Quantitatively, we may set:

H_0 : Scaled Schoenfeld residual does not change over time.

H_1 : Scaled Schoenfeld residual changes over time.

and get a chi-square statistic of scaled Schoenfeld residual. If $p < 0.05$, the H_0 hypothesis is rejected, which means the proportional hazard does not hold.

- B. Here, we show the qualitative results of scaled Schoenfeld residual as in Figure 7. We can see that although most of the variables follow the proportional hazard ($p > 0.05$), but **the entire model itself does not ($p = 0.03$)**. Therefore, we would like to **figure out the hidden association of our model with time.**

	chisq	df	p
age	0.38	1.00	0.54
Diastolic.blood.pressure	1.48	1.00	0.22
Respiratory.rate	1.10	1.00	0.29
temperature	4.39	1.00	0.04
MCH	0.01	1.00	0.94
MCHC	0.06	1.00	0.80
MCV	0.01	1.00	0.91
Platelets	0.74	1.00	0.39
INR	0.78	1.00	0.38
NT.proBNP	1.48	1.00	0.22
Creatinine	1.82	1.00	0.18
Urea.nitrogen	0.20	1.00	0.66
Blood.potassium	0.01	1.00	0.93
Blood.calcium	0.78	1.00	0.38
Anion.gap	0.01	1.00	0.91
PH	2.68	1.00	0.10
Bicarbonate	2.93	1.00	0.09
PCO2	0.04	1.00	0.84
atrialfibrillation	3.11	1.00	0.08
deficiencyanemias	3.54	1.00	0.06
Renal.failure	3.30	1.00	0.07
GLOBAL	34.56	21.00	0.03


Figure 7

VII. Step 5: Add time interaction

By the result of Step 3, we would like to choose a variable that may correlate with time, and add its time interaction variable into our cox regression model.

To add time interaction variable, we need to obtain the data at each time point. Since our dataset only contains the total amount of hospitalization time, we may **split the data of each patient by a short time interval**. The data of each patient at different time points is exactly the same. (Figure 8).

id	BMI	Hospitalization time
100000	20.91	5



id	BMI	Start time	End time
100000	20.91	0	1
100000	20.91	1	2
100000	20.91	2	3
100000	20.91	3	4
100000	20.91	4	5

Figure 8

Now we can choose the possible variable that correlates with time. By the results from step 3, we found that **temperature has a p-value < 0.05 on scaled Schoenfeld test**. However, if we added the time interaction with temperature into our model, **our model**



still fails on scaled Schoenfeld test. (Figure 9)

	chisq	df	p
age	3.85	1.00	0.05
Diastolic.blood.pressure	1.96	1.00	0.16
Respiratory.rate	0.02	1.00	0.88
temperature	2.66	1.00	0.10
MCH	0.02	1.00	0.90
MCHC	-1.27	1.00	1.00
MCV	3.74	1.00	0.05
Platelets	1.97	1.00	0.16
INR	24.93	1.00	0.00
NT.proBNP	0.36	1.00	0.55
Creatinine	-13.80	1.00	1.00
Urea.nitrogen	4.41	1.00	0.04
Blood.potassium	-82.88	1.00	1.00
Blood.calcium	-0.53	1.00	1.00
Anion.gap	45.77	1.00	0.00
PH	54.88	1.00	0.00
Bicarbonate	19.99	1.00	0.00
PCO2	2.13	1.00	0.14
atrialfibrillation	1.23	1.00	0.27
deficiencyanemias	-4.34	1.00	1.00
Renal.failure	1.10	1.00	0.29
temperature:Start_time	13.64	1.00	0.00
GLOBAL	100.19	22.00	0.00

Figure 9

Therefore, we turn to *"iron deficiency anemia(deficiencyanemias)"*, which has the second lowest p-value. We add time interaction with iron deficiency anemia and the results in Figure 10 show that **our model finally follows the assumption of proportional hazard** ($p > 0.05$)

	chisq	df	p
age	1.13	1.00	0.29
Diastolic.blood.pressure	1.37	1.00	0.24
Respiratory.rate	0.14	1.00	0.71
temperature	3.96	1.00	0.05
MCH	0.02	1.00	0.90
MCHC	0.61	1.00	0.44
MCV	0.10	1.00	0.75
Platelets	0.71	1.00	0.40
INR	0.42	1.00	0.52
NT.proBNP	1.59	1.00	0.21
Creatinine	1.17	1.00	0.28
Urea.nitrogen	0.63	1.00	0.43
Blood.potassium	0.15	1.00	0.70
Blood.calcium	0.39	1.00	0.53
Anion.gap	0.01	1.00	0.94
PH	2.50	1.00	0.11
Bicarbonate	2.62	1.00	0.11
PCO2	0.07	1.00	0.79
atrialfibrillation	2.39	1.00	0.12
deficiencyanemias	0.35	1.00	0.55
Renal.failure	2.65	1.00	0.10
deficiencyanemias:Start_time	0.04	1.00	0.84
GLOBAL	33.06	22.00	0.06

Figure 10

VIII.Step 6: Get our final cox regression model

The results are shown in Figure 11. **The concordance and AIC are 0.899, 518.1894, respectively.** The concordance is slightly lower before time interaction is added. In contrast, AIC score is better than the model before time interaction added.



explanatory	HR	L95	U95	p
age	1.029226e+00	1.0029257	1.056215e+00	0.0291709
Diastolic.blood.pressure	9.571196e-01	0.9259820	9.893043e-01	0.0093984
Respiratory.rate	1.084960e+00	1.0137679	1.161152e+00	0.0185302
temperature	6.383285e-01	0.3886497	1.048407e+00	0.0761904
MCH	6.471500e-03	0.0000853	4.911078e-01	0.0224957
MCHC	6.888140e+01	1.4484336	3.275709e+03	0.0317147
MCV	5.231203e+00	1.2863535	2.127369e+01	0.0207897
Platelets	9.959863e-01	0.9932937	9.986862e-01	0.0035934
INR	1.401249e+00	1.0987874	1.786969e+00	0.0065416
NT.proBNP	1.000013e+00	0.9999972	1.000028e+00	0.1087319
Creatinine	5.420734e-01	0.3058813	9.606460e-01	0.0359505
Urea.nitrogen	1.019070e+00	1.0021037	1.036324e+00	0.0274338
Blood.potassium	1.867227e+00	0.8661962	4.025111e+00	0.1110645
Blood.calcium	5.828710e-01	0.3139367	1.082188e+00	0.0873068
Anion.gap	1.371214e+00	1.1919963	1.577377e+00	0.0000100
PH	1.012517e+04	4.1384430	2.477237e+07	0.0205176
Bicarbonate	8.836245e-01	0.7667028	1.018377e+00	0.0875430
PCO2	1.087214e+00	1.0273486	1.150568e+00	0.0038077
atrialfibrillation	4.820479e-01	0.2397144	9.693626e-01	0.0406322
deficiencyanemias	7.527300e-02	0.0158303	3.579219e-01	0.0011480
Renal.failure	1.883061e-01	0.0837634	4.233255e-01	0.0000535
deficiencyanemias:Start_time	1.100409e+00	0.9999022	1.211019e+00	0.0502343

Figure 11

To examine deeper the performance of our constructed cox regression model, we utilize two qualitative techniques:

1. **Dfbeta**: Delete an observation once at a time, and record the estimated coefficient change. If the dfbetas accumulate around zero, this means that there does not exist an obvious outlier that affect the estimated coefficient a lot.
2. **Deviance**: A normalized transformation of Martingale residual. Martingale residual r_{Mi} is defined as:

$$r_{Mi} = \delta_i - \hat{H}_0(t_i) \exp(\hat{\beta} x_i)$$

where δ_i is the true outcome of a patient (0=censored, 1=dead), t_i is the time point, x_i is variable, and $H_0(t_i)$ is the cumulative hazard at time t_i . If our model fits perfect on the data, then r_{Mi} is supposed to be 0. Deviance is an extension of Martingale residual, which can be defined as:

$$r_{Di} = \text{sign}(r_{Mi}) [-2\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}]^{\frac{1}{2}}$$

This transformation makes the average of all r_{Di} becomes zero. Simply put,

we want our deviance of each observation to be around zero. If the deviance is larger than 0, it means that the patient expires earlier than expected; conversely, if the deviance is less than 0, it means that the patient lives longer than expected.

Figure 12 shows the dfbeta and deviance of our model with respect to the dataset. The left subfigure displays the dfbeta value for each observation in each variable, and we can observe that the values did accumulate around zero. The right subfigure displays the deviance value for each observation, and the data points accumulate around zero as well. However, few data points have greater deviance value, meaning **few patients die before the expected survival time.**

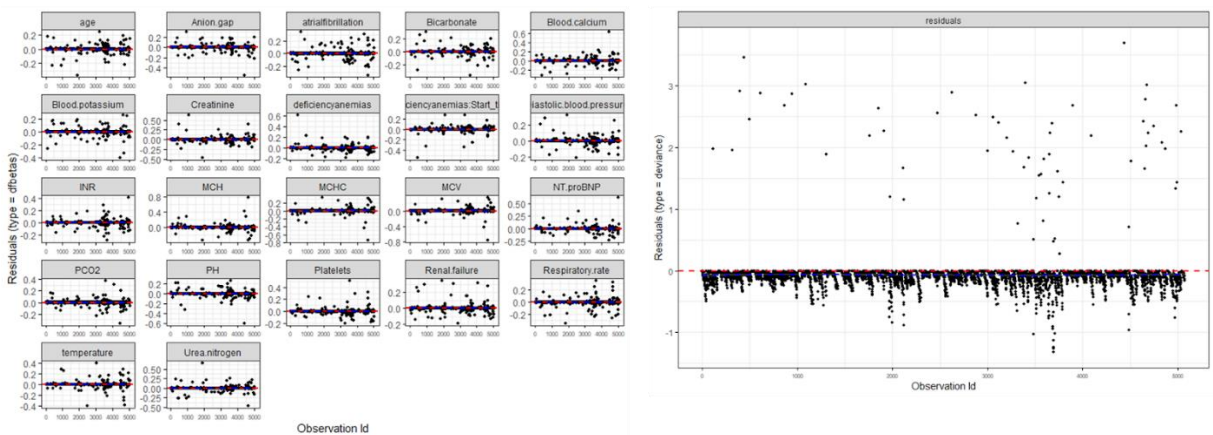


Figure 12

IX. Discussion

- A. Cox regression model:
$$h(t|X)/h_0(t) = \exp(\ln(1.029)\text{age} + \ln(0.975)\text{diastolic.blood.pressure} + \ln(1.085)\text{respiratory.rate} + \ln(0.638)\text{temperature} + \ln(0.006)\text{MCH} + \ln(68.814)\text{MCHC} + \ln(5.231)\text{MCV} + \ln(0.996)\text{Platelets} + \ln(1.401)\text{INR} + \ln(1.000)\text{NT-proBNP} + \ln(0.542)\text{Creatinine} + \ln(1.019)\text{BUN} + \ln(1.867)\text{K}^+ + \ln(0.583)\text{Ca}^{2+} + \ln(1.371)\text{Anion.gap} + \ln(1012.517)\text{pH} + \ln(0.884)\text{Bicarbonate} + \ln(1.087)\text{pCO}_2 + \ln(0.482)\text{AF} + \ln(0.075)\text{IDA} + \ln(0.188)\text{RF} + \ln(1.100)\text{IDA} * \text{time})$$
- B. In our cox regression model, we included *age*, *diastolic blood pressure*, *respiratory rate*, *temperature*, *MCH*, *MCHC*, *MCV*, *platelets*, *INR*, *NT-proBNP*, *creatinine*, *urea nitrogen(BUN)*, *blood potassium(K⁺)*, *blood calcium(Ca²⁺)*, *anion gap*, *pH*, *bicarbonate*, *pCO₂*, *atrial fibrillation(AF)*, *iron deficiency anemia(IDA)*, and *renal failure(RF)*.
- C. Among these chosen variables, *age*, *platelets*, *NT-proBNP*, *blood potassium*, *blood calcium* has reasonable results on mortality in our cox regression model.

- D. Interestingly, diseases included in our model (e.g. *atrial fibrillation*, *iron deficiency anemia*, and *renal failure*) decreases the hazard ratio. Moreover, *renal failure* has the lowest p-value among all variables, which means it significantly decreases hazard ratio. However, this does not sound reasonable. Further inspect our model, high creatinine level also decreases hazard ratio, which corresponds to the results of renal failure (Renal failure is partially defined by high creatinine level.)
- E. *Iron deficiency anemia (IDA)* is one of the possible time-correlated variables that we observed. Additionally, *IDA* alone decreases hazard ratio, but the interaction between time and *IDA* increases hazard ratio, which indicate that as time passes by, *IDA* turns into a possible risk factor of death.
- F. In our model, *MCV/MCHC* and *MCH* act conversely on mortality, though *MCV*, *MCHC*, *MCH* should be positively correlated in real world.
- G. Although high *pH*, *pCO₂* level increase hazard ratio, low level bicarbonate decreases hazard ratio in our dataset. The results seem to be contradicting since *pCO₂* indicates pH decrease, and bicarbonate indicates pH increase. We inferred that this might be due to the physiological compensation. Therefore, the concentration of *CO₂* or bicarbonate does not necessarily reflect to the pH value.
- H. Large coefficient value on pH value may be due to small physiological range of plasma pH in human body (7.35~7.45).

Topic Three Categorical Data Analysis

I. Precondition

Before we build the model, the following points need to be explained:

- A. We discard a row of data if it has missing values in any column, so our generalized linear model can only be trained on 425 rows.
 1. Deleting data hastily is dangerous in data analysis because it may corrupt the data structure. To solve this problem, before deleting the data, we tried to fill in the missing values (such as the estimated mean of BMI, PCA algorithm, etc.), but the result was not well.
 2. Later, it was speculated that this dataset is not suitable for filling missing values, because **some data is not missing, but does not exist in the first place**. For example, comorbidities cannot be arbitrarily added to patients who do not have a certain disease.
- B. We use the “probit link function” instead of the “logit link function”, because the distribution of the target variable (*outcome*) is steep, and the number of targets we focus on (death: 1) is small, so the **“probit link function” can better fit its distribution** (shown in Figure 13).

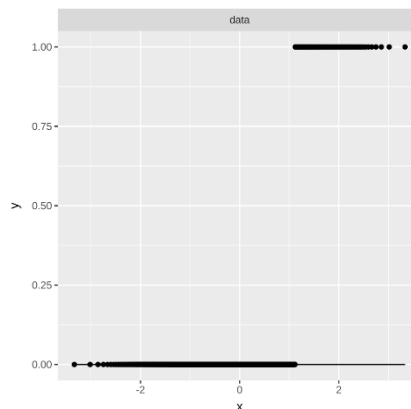


Figure 13. Distribution of target variable.

Then, we use the “probit link function” and “logit link function” to build the full model to verify our guess. Even there are only slight differences, it can still be seen from Figure 14 that the performance of the probit model is indeed higher than that of the logit model.



Null deviance: 353.17 on 424 degrees of freedom	
Residual deviance: 162.73 on 375 degrees of freedom	
AIC: 262.73	
Number of Fisher Scoring iterations: 7	logit

Null deviance: 353.17 on 424 degrees of freedom	
Residual deviance: 162.82 on 375 degrees of freedom	
AIC: 262.82	
Number of Fisher Scoring iterations: 9	probit

Figure 14. The performance difference of the full model using probit and logit.

According to the above results, in the next stage of model building, our generalized linear models are all based on probit link functions.

II. Model Construction Process

Figure 15 is a flowchart of our model construction.

First, we will build a full model, and then use forward selection and backward selection to filter out important variables, and the first stage is over.

In the second stage, we will consider the interaction between each variable, and then bring it into the model of the first stage. Next, we also filter the variables considered important by the model again as in the first stage.

Finally, we analyze variable importance and its performance in the final model.

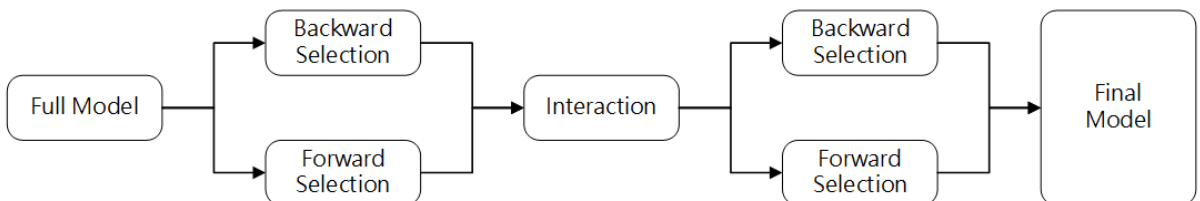


Figure 15. Flowchart of model construction.

III. Stage 1: GLMs (without interaction)

A. Full Model

First, we use all variables to build a full model as a baseline, and its performance is shown in Figure 16 and 17.

```
Call:
glmFormula = outcome ~ ., family = binomial(link = "probit"),
data = data_omt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.68524  -0.32749  -0.89471  -0.81197   2.91627

Coefficients:
(Intercept)          Estimate Std. Error z value Pr(>|z|)
age              -3.738e+01  4.387e+01  -0.852  0.39422
gender            -1.365e-02  2.859e+01  0.477  0.63081
BMI               -1.739e-02  1.765e-02  -0.985  0.32456
hypertensive      -1.214e-01  2.928e-01  -0.416  0.67762
arrhythmia        -3.675e-01  2.976e-01  -1.236  0.21664
CHF_with_no_MI    -7.692e-02  4.226e-01  -0.182  0.85558
diabetes          8.518e-02  3.149e-01  0.270  0.78997
deficiencyanemias -1.888e-01  3.284e-01  -0.575  0.56816 *
depression        -1.453e-01  4.248e-01  -0.342  0.73239
hyperlipidemia    5.402e-01  2.892e-01  1.869  0.05994
Renal_failure     -1.644e+00  3.941e-01  -4.171  3.83e-05 ***
COPD              -7.248e-01  4.608e-01  -1.555  0.11987
heart_rate        1.547e-02  1.816e-02  0.524  0.59703
Systolic.blood.pressure 5.349e-03  9.863e-03  0.542  0.58755
Diastolic.blood.pressure -2.635e-02  2.812e-02  -0.926  0.34996
Respiratory.rate   -5.137e-02  3.338e-02  -1.552  0.12071
temperature       -5.827e-01  2.254e-01  -2.580  0.01072 *
Sp_O2            -2.108e-02  6.166e-02  -0.340  0.73405
urine_output       7.934e-01  1.149e+00  0.689  0.49000
hematocrit        -2.177e-01  1.992e-01  -1.093  0.27433
BIC              -1.788e+01  1.739e+00  -1.031  0.30582
WCI               -1.082e+00  1.280e+00  -0.846  0.39745
MCHC              8.254e-01  1.871e+00  0.437  0.66474
MCH               1.155e-01  4.124e-01  0.279  0.78136
RDW               -0.567e-02  7.556e-02  -0.074  0.94288
Leucocyte         2.804e-02  3.185e-02  0.881  0.37638
Platelets         -4.853e-01  1.532e-01  -3.164  0.00114 **
Neutrophils       -5.614e-02  3.659e-02  -1.534  0.12497
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 353.17 on 424 degrees of freedom
Residual deviance: 162.82 on 375 degrees of freedom
AIC: 262.82

Number of Fisher Scoring iterations: 9
```

Figure 16. Feature saliency for the full model.

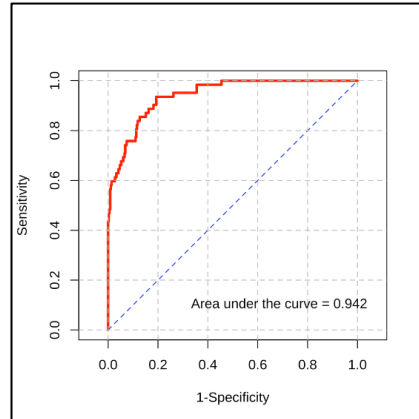


Figure 17. ROC & AUC for the full model.

B. Forward Selection

Then we use forward selection to filter the variables from the full model. From Figure 18, we can find that forward has screened out 15 important variables from the original 49 variables, most of which have a significant relationship with the target variable.

The model effect after forward selection is shown in Figure 19, it can be found that after the variable screening, the effectiveness of the model does not drop too much.

```
Coefficients:
(Intercept)          Estimate Std. Error z value Pr(>|z|)
Anion.gap            0.186779  0.069610  2.683  0.007292 **
Blood.calcium        -0.692428  0.238848  -2.899  0.003743 ***
Renal.failure        -1.370635  0.318395  -4.305  1.67e-05 ***
Urea.nitrogen         0.023276  0.006913  3.367  0.000759 ***
deficiencyanemias    -0.643052  0.268203  -2.398  0.016502 *
age                  0.020074  0.009058  2.216  0.026689 **
Platelets            -0.002925  0.001099  -2.661  0.007789 **
Respiratory.rate      0.061493  0.025455  2.416  0.015701 *
PCO2                 0.054650  0.014938  3.658  0.000254 ***
Creatinine           -0.427272  0.196323  -2.176  0.029527 *
Lymphocyte           -0.029831  0.016463  -1.812  0.069995 .
INR                  0.214993  0.127918  1.681  0.092818 .
Bicarbonate           -0.093193  0.041101  -2.267  0.023365 *
COPD                 -0.712364  0.406707  -1.752  0.079852 .
Hyperlipidemia       0.412363  0.232749  1.772  0.076443 .

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 18. Feature saliency for the full model after forward selection.

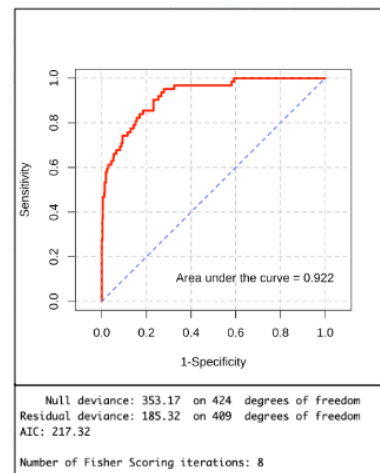


Figure 19. ROC & AUC for the full model after forward selection.

C. Backward Selection

Finally, we use backward selection to filter the variables from the full model. From Figure 20, we can find that backward has screened out 19 important variables from the original 49 variables, most of which have a significant relationship with the target variable.

The model effect after backward selection is shown in Figure 21, it can be found that after the variable screening, the effectiveness of the model does not drop too much.

Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	8.538725	8.166572	1.046	0.295760
age	0.014671	0.009881	1.485	0.137592
deficiencyanemias	-0.734042	0.283801	-2.586	0.009697 **
Hyperlipemia	0.568170	0.237699	2.390	0.016835 *
Renal.failure	-1.333556	0.321587	-4.147	3.37e-05 ***
COPD	-0.804121	0.427968	-1.879	0.060254 .
heart.rate	0.014606	0.007937	1.840	0.065729 .
Diastolic.blood.pressure	-0.027200	0.014392	-1.890	0.058768 .
Respiratory.rate	0.056264	0.027275	2.063	0.039126 *
temperature	-0.281379	0.180811	-1.556	0.119659
Platelets	-0.003208	0.001183	-2.711	0.006705 **
Lymphocyte	-0.027309	0.016730	-1.632	0.102610
Creatinine	-0.460429	0.208744	-2.206	0.027404 *
Urea.nitrogen	0.017751	0.007174	2.474	0.013355 *
Blood.sodium	-0.080671	0.046273	-1.743	0.081269 .
Blood.calcium	-0.723610	0.235811	-3.069	0.002151 **
Chloride	0.087143	0.040651	2.144	0.032058 *
Anion.gap	0.303127	0.071889	4.217	2.48e-05 ***
Magnesium.ion	0.814031	0.476407	1.709	0.087509 .
PCO2	0.051282	0.014445	3.550	0.000385 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Figure 20. Feature saliency for the full model after backward selection.

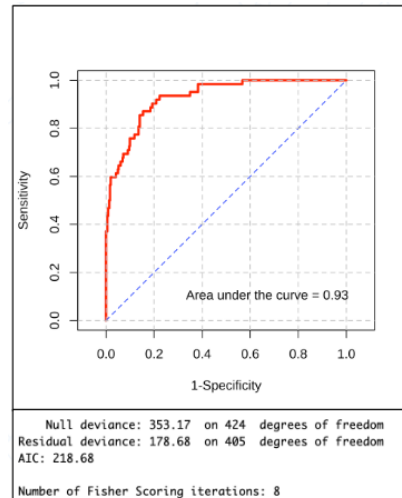


Figure 21. ROC & AUC for the full model after backward selection.

IV. Interaction

Therefore, from the comparison of Figure 16, Figure 18, and Figure 20, it can be found that among all the original variables, there is an interference relationship between variables.

Therefore, there is an interaction between the original variables, and we use the variance inflation factor (VIF) to find out the variables that may interfere with each other (shown in Figure 22).

vif(omit_m2)

age	gendera	BMI	hypertensive
2.228893	1.683049	1.991793	1.572973
atrialfibrillation	CHD.with.no.MI	diabetes	deficiencyanemias
1.807158	1.376718	2.015830	1.458585
depression	Hyperlipemia	Renal.failure	COPD
1.257162	1.671586	2.350590	1.356191
heart.rate	Systolic.blood.pressure	Diastolic.blood.pressure	Respiratory.rate
2.254568	1.984678	2.823797	1.611406
temperature	SP.O2	Urine.output	hematocrit
1.898029	1.724983	1.737795	78.581001
RBC	MCH	MCHC	MCV
97.115038	824.550347	180.609558	649.268884
RDW	Leucocyte	Platelets	Neutrophils
2.135057	2.336325	1.897213	9.563172
Basophils	Lymphocyte	PT	INR
2.470160	8.470205	140.442658	139.044563
NT.proBNP	Creatine.kinase	Creatinine	Urea.nitrogen
1.462701	1.473182	4.232363	3.409273
glucose	Blood.potassium	Blood.sodium	Blood.calcium
1.989068	4.244014	345.411523	2.018201
Chloride	Anion.gap	Magnesium.ion	PH
499.001704	129.448903	1.619526	4.630478
Bicarbonate	Lactic.acid	PCO2	EF
598.669686	1.999617	12.414826	1.666243
survival.time			
1.501068			

Figure 22. VIF of original variables in the full model.

From the original 49 variables, we found that there are 12 variables (marked in the red box in Figure 22) that may have interaction terms between them.

Comparing with Figure 18 and Figure 20, we found that only 4 variables remained after the screening, such as *Anion.gap* (陰離子間隙), *Blood.sodium* (血液中鈉離子濃度), *Chloride* (氯化物), and *PCO2* (血液中二氧化碳分壓) (comparison results are shown in Figure 23).

Forward Variables	Backward Variables
Anion.gap	Anion.gap
-	Blood.sodium
-	Chloride
PCO2	PCO2

Figure 23. Observe the interaction term retained in the filtered model.

Since the above steps only screen out potential variables, not all of them affect each other. Therefore, we observed the correlation between these four variables to confirm whether there was an interaction between them.

From Figure 24, we can find that the variables in the bottom right corner are highly correlated, and these four variables are also concentrated here, so there is a high possibility of existing interaction.

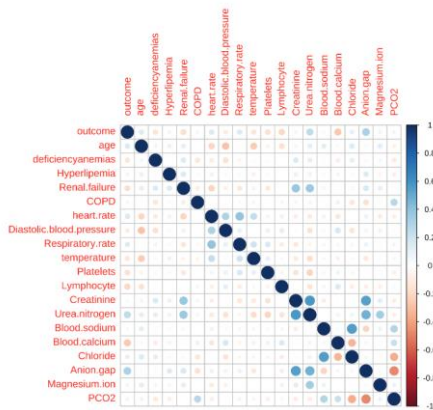


Figure 24. Correlation of all variables in the backward selection model.

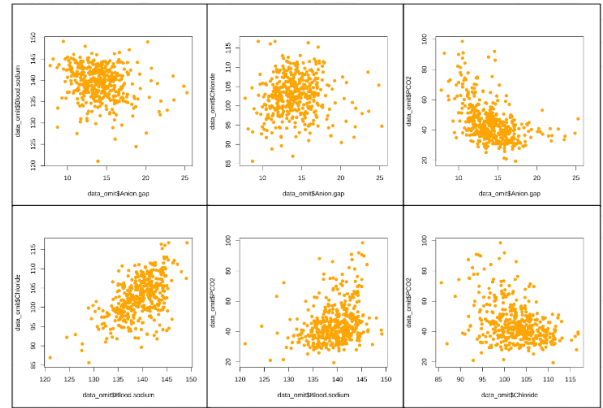


Figure 25. Distribution of four potential variables relative to each other.

V. Stage 2: GLMs (with interaction)

In the second stage, all possible combinations (11 kinds) of interactions between these four variables are included in the first-stage model and screened again.

Note that four methods (i.e. Forward-to-Forward, Forward-to-Backward, Backward-to-Forward, and Backward-to-Backward) should have been performed, but results based on the forward model were the same as the forward-only model, so we only illustrate the results based on the backward model.

A. Backward-to-Forward

After incorporating the 11 possibilities into the backward model, we used forward selection to screen variables.

We found that forward selection screened out all interaction items, but due to the consideration of interaction items, the importance of variables for forward selection has also changed, so in the end, 17 important features were retained, and most of them had a significant relationship with the target variable (shown in Figure 26).

Compared with the base forward model, the AUC of the backward-to-forward model is higher (shown in Figure 27).

Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	11.140704	7.115564	1.566	0.11742
Anion.gap	0.274415	0.067784	4.048	5.16e-05 ***
Blood.calcium	-0.912240	0.218348	-4.178	2.94e-05 ***
Renal.failure	-1.404052	0.315847	-4.445	8.77e-06 ***
Urea.nitrogen	0.017670	0.007234	2.443	0.01458 *
deficiencyanemias	-0.697299	0.279558	-2.494	0.01262 *
age	0.015250	0.009536	1.599	0.10978
Platelets	-0.003348	0.001182	-2.832	0.00463 **
Respiratory.rate	0.054187	0.026515	2.044	0.04099 *
PCO2	0.030591	0.009870	3.100	0.00194 **
Creatinine	-0.389107	0.206775	-1.882	0.05986 .
Lymphocyte	-0.026061	0.016175	-1.611	0.10714
Hyperlipemia	0.466830	0.228623	2.042	0.04116 *
COPD	-0.697566	0.420992	-1.657	0.09753 .
Diastolic.blood.pressure	-0.032721	0.014106	-2.320	0.02036 *
heart.rate	0.015105	0.007813	1.933	0.05322 .
temperature	-0.324860	0.179572	-1.809	0.07044 .
Magnesium.ion	0.770558	0.467094	1.650	0.09901 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

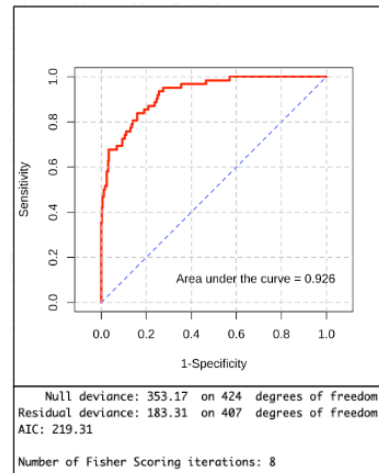


Figure 26. Feature saliency for the backward model after forward selection.

Figure 27. ROC & AUC for the backward model after forward selection.

B. Backward-to-Backward

Next, after incorporating the 11 possibilities into the backward model, we used backward selection to screen variables.

We found that backward selection retains 4 kinds of interactions (marked in the red box in Figure 28), none of which are included in *Blood.sodium*, so there does not exist interaction with the other three in it.

Finally, backward-to-backward retains 22 important features, and most of them are significantly correlated with the target variable (shown in Figure 28).

Compared with the base backward model, the AUC of the backward-to-backward model is higher (shown in Figure 29).

Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.155e+02	6.242e+01	1.851	0.06415 .
age	1.936e-02	9.695e-03	1.996	0.04588 *
deficiencyanemias	-7.415e-01	2.834e-01	-2.616	0.00889 **
Hyperlipemia	5.597e-01	2.424e-01	2.309	0.02094 *
Renal.failure	-1.417e+00	3.287e-01	-4.311	1.62e-05 ***
COPD	-7.941e-01	4.423e-01	-1.796	0.07256 .
heart.rate	1.544e-02	8.120e-03	1.902	0.05717 .
Diastolic.blood.pressure	-2.264e-02	1.418e-02	-1.597	0.11025
Respiratory.rate	4.847e-02	2.834e-02	1.710	0.08721 .
Platelets	-3.325e-03	1.189e-03	-2.796	0.00517 **
Lymphocyte	-2.451e-02	1.688e-02	-1.452	0.14657
Creatinine	-4.505e-01	2.072e-01	-2.174	0.02971 *
Urea.nitrogen	2.069e-02	7.408e-03	2.792	0.00523 **
Blood.sodium	-8.039e-02	4.846e-02	-1.659	0.09715 .
Blood.calcium	-7.531e-01	2.434e-01	-3.095	0.00197 **
Chloride	-1.090e+00	6.145e-01	-1.774	0.07614 .
Anion.gap	-7.161e+00	4.206e+00	-1.703	0.08862 .
Magnesium.ion	7.356e-01	4.799e-01	1.533	0.12538
PCO2	-2.692e+00	1.463e+00	-1.840	0.06570 .
Chloride:Anion.gap	7.478e-02	4.107e-02	1.821	0.06866 .
Anion.gap:PCO2	1.769e-01	1.018e-01	1.737	0.08238 .
Chloride:PCO2	2.747e-02	1.431e-02	1.920	0.05485 .
Chloride:Anion.gap:PCO2	-1.772e-03	9.946e-04	-1.782	0.07480 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

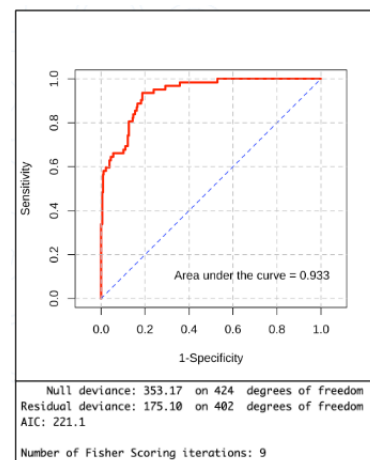


Figure 28. Feature saliency for the backward model after backward selection (Left).

Figure 29. ROC & AUC for the backward model after backward selection (Right).

VI. Final Models

Figure 30 shows the performance indicators of the above five models, and it can be found that:

- A. **Backward has better performance on “Residual deviance” and “AUC” .**
- B. **Forward performs better on the “Number of variables” and “AIC” .**

Finally, there is no significant difference in performance between the Backward-to-Forward model and the Backward-to-Backward model, but the former has the advantage of “requiring fewer variables” , while the latter has the advantage of “being more comprehensive” .

Therefore, we believe that both models are applicable to this research topic.

Model	#Variables↓	AIC↓	Residual deviance↓	AUC↑
Full Model	49	262.82	162.82	0.942
Forward Selection	15	217.32	185.32	0.922
Backward Selection	19	218.68	178.68	0.93
Backward Selection + Interaction + Forward Selection	17	219.31	183.31	0.926
Backward Selection + Interaction + Backward Selection	22	221.10	175.10	0.933

Figure 30. Summary of all models.

Topic Four Conclusion

I. Survival Analysis

- A. We identify 19 of 49 variables which may have association with mortality in heart failure patients in ICU units.
- B. *Age*, *PLT* (血小板), *NTproBNP* (利钠肽), *K+*, *Ca2+* has reasonable results on mortality in our model.

II. Categorical Data Analysis

- A. We got two GLMs that can have better adaptability on this dataset.
- B. Know which variables are important in the problem from Figure 14 and Figure 16.
- C. The coefficient of *Renal.failure* (肾衰竭), *COPD* (慢性阻塞性肺病), and *Creatinine* (肌酐) in our model, the result for mortality is unreasonable.

Therefore, we use a 2×2 table to analyze the relationship between *Renal.failure* (肾衰竭) and the target variable (as shown in Figure 31), and we can find that the relationship between these two data is consistent with the results of our model.

\$data	Renal_failure		
state	No	Yes	Total
Alive	218	52	270
Death	145	10	155
Total	363	62	425
\$measure			
odds ratio with 95% C.I.			
state	estimate	lower	upper
Alive	1.0000000	NA	NA
Death	0.2932953	0.1358982	0.5737689
\$p.value			
two-sided			
state	midp.exact	fisher.exact	chi.square
Alive	NA	NA	NA
Death	0.0001847945	0.0002961195	0.0003176104
\$correction			
[1] FALSE			

Figure 32. 2×2 table of *Renal.failure* and *outcome*.

Finally, we found that this unreasonable phenomenon comes from the source of this data, because this data is collected from ICU patients, it is possible that the patient was admitted to the ICU because of renal failure, so focus on the relationship between *Renal.failure* and target variables (*outcome*) is meaningless.

- D. For interaction, we found that they interact as shown in Figure 32 (i.e. Under the interaction of these variables, as long as any index increases, the patient's mortality rate will also increase).



Chloride:Anion.gap	7.478e-02	4.107e-02	1.821	0.06866	.
Anion.gap:PCO2	1.769e-01	1.018e-01	1.737	0.08238	.
Chloride:PCO2	2.747e-02	1.431e-02	1.920	0.05485	.
Chloride:Anion.gap:PCO2	-1.772e-03	9.946e-04	-1.782	0.07480	.

Figure 32. The interactions in the Backward-to-Backward model.

Topic Five Reference & Work Assignment

1. Reference

- (1) Li, F., Xin, H., Zhang, J., Fu, M., Zhou, J., & Lian, Z. (2021). Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. *BMJ open*, 11(7), e044779.

2. Work Assignment

Name	Work Assignment	Common Work
陳祺侑	Survival Analysis	EDA, Making Slides, Writing Report
梁家瑤	Feature Extraction, Categorical Data Analysis	