MODELING OF BLOOD CELL COUNT IN SURVIVAL ANALYSIS FOR EMERGENCY COHORT

APPLIED STATISTICS 2024/2025

TEAM MEMBERS: LI KEJIA, MAO YANG HAO, YI JIAXIANG, NI GIOVANNI , GU XINYUE

31

"CRITICAL CELLS, CRUCIAL CARE"

INTRODUCTION

Emergency room blood analysis is an important part of medical emergency and plays a vital role in the diagnosis and treatment of patients. When a patient enters the emergency room, medical staff usually quickly perform a blood analysis to assess their health status. Blood analysis can provide information about red blood cells, white blood cells, platelets.
We employed multiple data analysis methods to construct the analytical dataset and **investigate the relationships between various blood cell parameters and patient survival outcomes.**
Specifically,

- to analyze BCDC surveillance data from ED patients
- to predict survival in adults who have visited the emergency department (ER) for illness or trauma for more than 1 year.

Eos	Eosinophils
Neu	Neutrophils
WBC	White blood cells
RDW	Red cell distribution width
MCV	Mean red cell volume
Hb	Hemoglobin
RBC	Red blood cells
Lym	Llymphocytes
Mon	Monocytes
PLT	Plateletes
PCT	Platelet hematocrit
PDW	Platelet distribution
Bas	Basophils

DATA SKETCH

The dataset includes data from **11,052** adult patients (>=18 years) admitted to the emergency room in 2020.
Each record contains demographic, clinical, and laboratory information collected **at first ER presentation**, including a blood cell differential count (BCDC), with **365-day follow-up**.

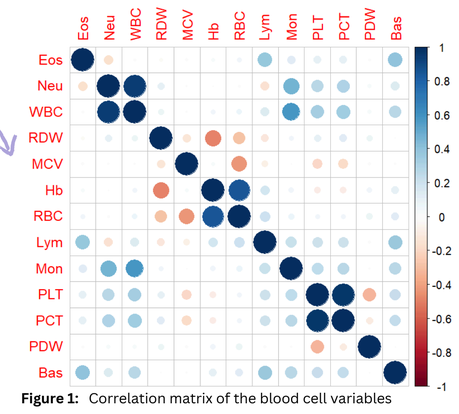
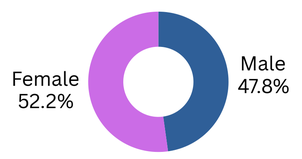
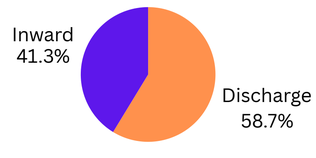
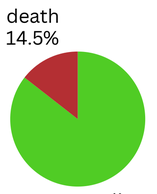
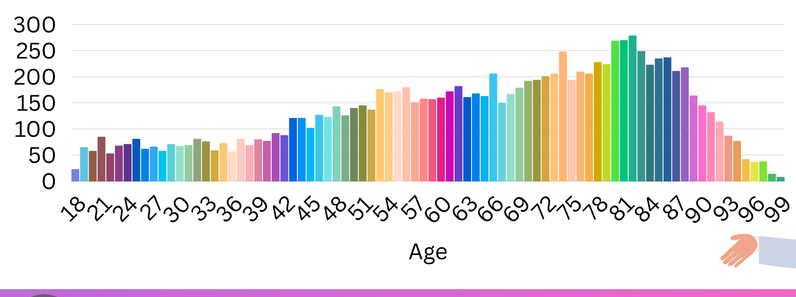


Figure 1: Correlation matrix of the blood cell variables



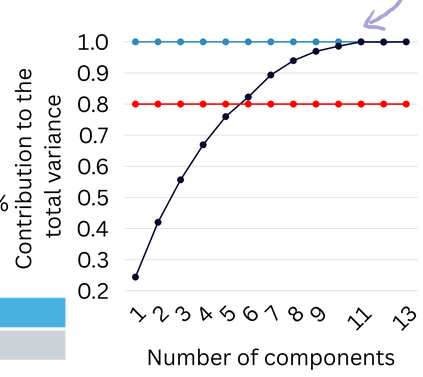






PRINCIPAL COMPONENT ANALYSIS

This **scree plot** shows the explained variance of each principal component and the cumulative variance. A red line marks the threshold of explained variance (usually 80%). The **first 6 components** together explain over 80% of the total variance.
The **first principal component (PC1)** alone explains 24.41% of the variance, while **the second component (PC2)** explains an additional 17.59%, as summarized in Table 2.



	PC1	PC2
Variance	0.2441	0.1759

Table 2: Proportion of variance explained by the first two principal components

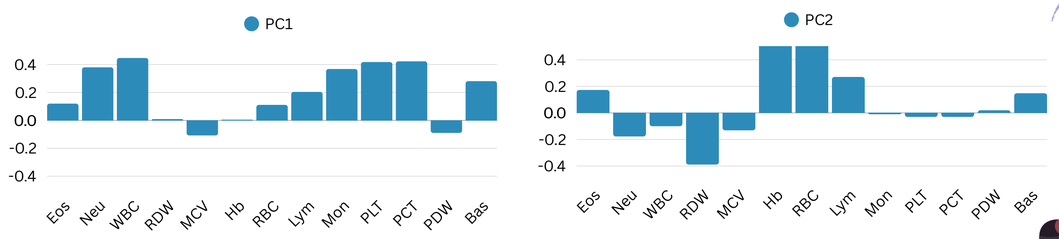


Figure 3: The loadings for the first two principal components

The figure 3 shows how each variable contributes to the first two principal components.
PC1 is mainly driven by WBC, Lym, PLT, PCT, and Bas with positive loadings, while MCV and PDW contribute negatively. **PC2** is positively influenced by Hb, RBC, and Lym, and negatively by MCV, Neu, and RDW.
This indicates that **PC1** captures patterns mainly related to immune/inflammatory markers and platelet indices, while **PC2** reflects variation in red blood cell characteristics (RBC, Hb, MCV).

PATIENTS CLUSTERING

Following the PCA results, K-Means clustering (K = 3) was applied to the blood cell count data for unsupervised grouping. This process divided patients into three distinct clusters, as shown in Figure 4. Each cluster exhibited **distinct differences in blood marker levels**, indicating the potential power of blood cell counts in clinical diagnosis.
Specifically:

- **Cluster 0** (Low risk): **Largest group** with the lowest mortality rate (8.0%), moderate hospitalization rate (36.1%) and relatively low in-hospital mortality (16.2%). This cluster possibly represents patients without significant abnormalities shown but relatively normal blood values.
- **Cluster 1** (High risk): **Smallest group** but with the highest mortality rate (36.1%), suggesting serious underlying health issues and hugely high risk that may require urgent and intensive medical intervention.
- **Cluster 2** (Moderate risk): **Intermediate size** and moderate mortality rate (11.5%), possibly indicating mild inflammation or chronic issues. While not as severe as Cluster 1, this group presents a higher clinical risk than Cluster 0 and likely represents patients who require ongoing monitoring and management.

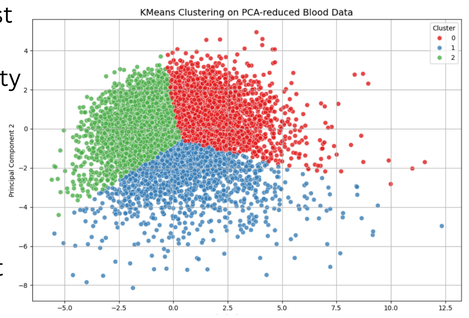


Figure 4: K-means clustering applied to blood cell data reduced by PCA

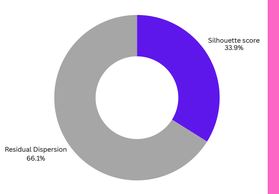


Figure 5: proportion of silhoutte score in total clustering evaluation

Cluster	total	Death_rate	Inward_rate	Death_rate_in_hospital
0	3677	0.080	0.361	0.162
1	1847	0.361	0.665	0.415
2	5527	0.115	0.363	0.208

Table 6: Summary of death and hospitalization (inward) rates by cluster

COX REGRESSION

Various approaches were explored during the feature selection process, but the initial results were not satisfactory.
Given the unique nature of the "follow-up time" variable (including censoring data), we referred to the estimation results of the Cox proportional hazards model. A set of features was selected for the final modeling, based on the statistical significance of each variable's association with mortality risk.

```
          coef  exp(coef)  Pr(>|z|)
Age  0.0627124  1.0647206   < 2e-16 ***
SexM 0.5373244  1.7114217  2.96e-06 ***
WBC  0.1154714  1.1224024  0.014524 *
Neu -0.0565788  0.9449921  0.257757
Lym -0.4861907  0.6149645  0.000134 ***
Mon -0.1453056  0.8647580  0.482408
Eos -0.4702717  0.6248325  0.428874
Bas -5.8726942  0.0028153  0.034203 *
PLT  0.0002467  1.0002468  0.948505
Hb  -0.0606417  0.9411604  0.027412 *
MCV  0.0229005  1.0231647  0.000253 ***
RDW  0.1780536  1.1948894  1.33e-12 ***
PCT -1.3413003  0.2615054  0.705507
PDW  0.0135629  1.0136553  0.678402
```

Figure 8: Cox regression model estimating the effect of clinical variables on survival

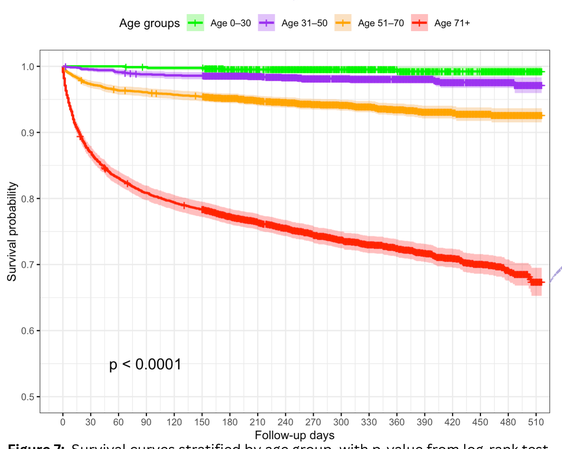


Figure 7: Survival curves stratified by age group, with p-value from log-rank test

	Lymphocytes(Lym)	Basophils(Bas)	Hemoglobin(Hb)	Age	SexM(Male)	WBC	MCV	RDW
Type	Protective	Protective	Protective	Risk	Risk	Risk	Risk	Risk
HR	0.6149	0.00215	0.9412	1.0647	1.7114	1.1224	1.0232	1.1949
P-value	0.000134	0.0342	0.0274	$<2e^{-16}$	$2.96\ e^{-6}$	0.0145	0.000253	$1.33e^{-12}$

Table9: Significant Cox regression results. Lymphocytes, basophils, and hemoglobin are associated with reduced risk (protective factors), whereas age, male sex, WBC, MCV, and RDW are associated with increased risk

Finally, we found **Age, male gender, WBC, Lym(lymphocytes), Bas, Hb(hemoglobin), MCV, and RDW** are significantly associated with mortality.

$$\left(\frac{h(t \mid X)}{h_0(t)}\right) = \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Sex} + \beta_3 \cdot \text{WBC} + \beta_4 \cdot \text{RDW} + \beta_5 \cdot \text{MCV} + \beta_6 \cdot \text{Hb} + \beta_7 \cdot \text{Lym} + \beta_8 \cdot \text{Bas}$$

LOGISTIC REGRESSION

The logistic regression model shows good accuracy and AUC, especially for predicting survival. However, it **struggles to correctly identify deaths**, likely due to class imbalance. Improving recall for the minority class (death) would require rebalancing techniques or more flexible models.

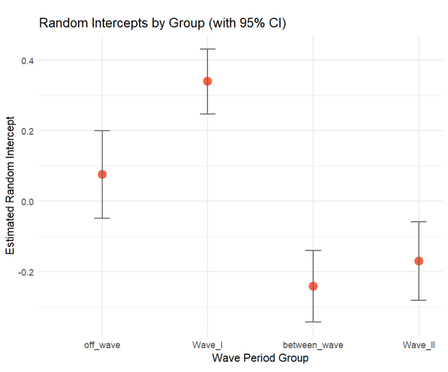
$$\text{logit}(P(\text{death}_{ij} = 1)) = X\beta + \varepsilon_{ij}$$

Model	Model_1	Model_2	Model_3	Model_4	Model_5
Accuracy	0.858	0.865	0.865	0.86	0.862
AUC	0.804	0.851	0.851	0.819	0.836

Table 10: Including random effects in logistic regression improves model performance

The variables were selected **based on prior information** from survival time. Specifically, we used the Cox proportional hazards model to incorporate survival duration in identifying variables **that may be more strongly associated with survival**, rather than relying solely on model outputs.

$$\text{logit}(P(\text{death}_{ij} = 1)) = X\beta + b_j + \varepsilon_{ij}$$



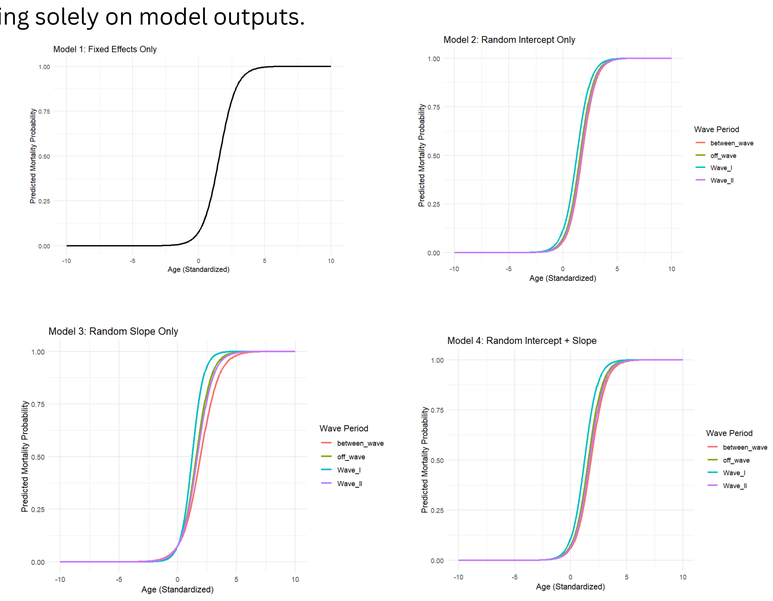


Figure 11 Predicted mortality probability by age under different mixed model structures and estimated random intercepts across wave periods

RANDOM FOREST

Random Forest models were trained to predict patient survival using different feature sets.

Multiple models were constructed using **(1) only blood biomarkers, (2) Blood biomarkers + demographic features, (3) Demographic features only (age and gender)**
The combination of blood biomarkers with age and gender yielded the highest AUC on both training (0.91) and test sets (0.86), outperforming models using only blood or demographic features.

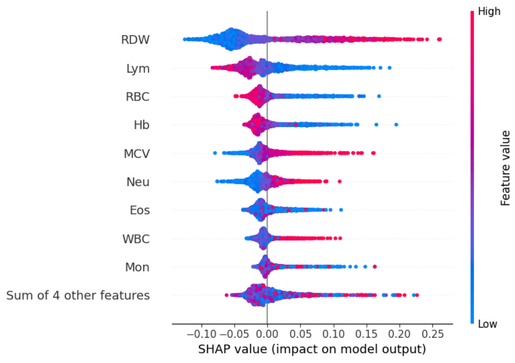


figure 12: SHAP values showing the contribution of blood-related features to survival prediction

Feature Set	Training AUC	Test AUC
Blood Only	0.8557	0.8057
Blood + AgeSex	0.9104	0.8604
Age + Sex Only	0.8619	0.8119

Table 13: Performance comparison of feature sets based on AUC values in training and test datasets

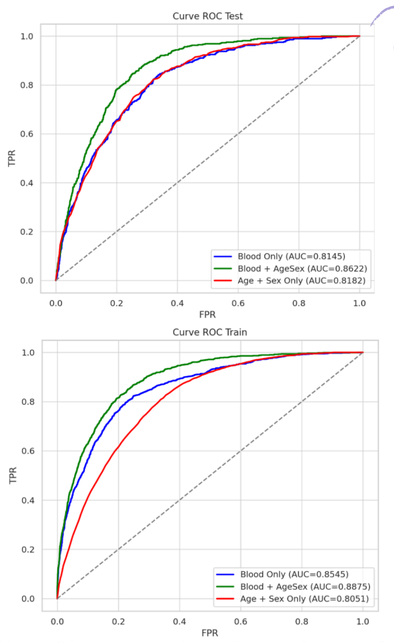


Table 14: ROC curves comparing predictive performance of different feature combinations in survival classification

CONCLUSION

This study analyzed data from over 11,000 emergency room patients, integrating routine blood test results and clinical information using **PCA, clustering, logistic regression, Cox regression, and Random Forest models**. The **key findings** are as follows:

- **Unsupervised clustering** based on blood biomarkers identified three clinically meaningful patient groups, each showing distinct levels of **immune response and risk of mortality or hospitalization**, suggesting potential for early risk stratification upon ER admission.
- **Multimodal models** that combined blood indicators with demographic features significantly outperformed models using a single data type. The best-performing **Random Forest model** achieved the highest AUC in both training and test sets, indicating strong generalizability.

While the models revealed meaningful patterns and strong predictive potential, several **limitations** remain:

- The dataset was derived from a single center and year;
- **Class imbalance** reduced the model's sensitivity to mortality;
- Despite the use of regularization, **overfitting** remains a potencial concern .

Future Work:

- **Expand to multi-center, multi-year data** to improve generalizability
- Improve class balance using resampling or weighting strategies
- Perform external validation to test robustness
- Incorporate **additional clinical variables** (e.g., comorbidities, vital signs) to enhance prediction accuracy