Covid Data Analysis

Supplementary Data:

Section 1: Data Overview and Descriptive Statistics

This section provides a statistical summary of the dataset used in the COVID-19 prediction analysis.

- Basic Statistics
  - Row Counts
  - Percentage
  - Data Structure
  - Missing Data Profile
- Univariate Distribution
  - Histogram
  - Bar Chart (with frequency)
  - QQ Plot

Section 2: Supplementary Figures and Model Evaluation

This section includes additional visualizations and model diagnostics to support the main analysis.

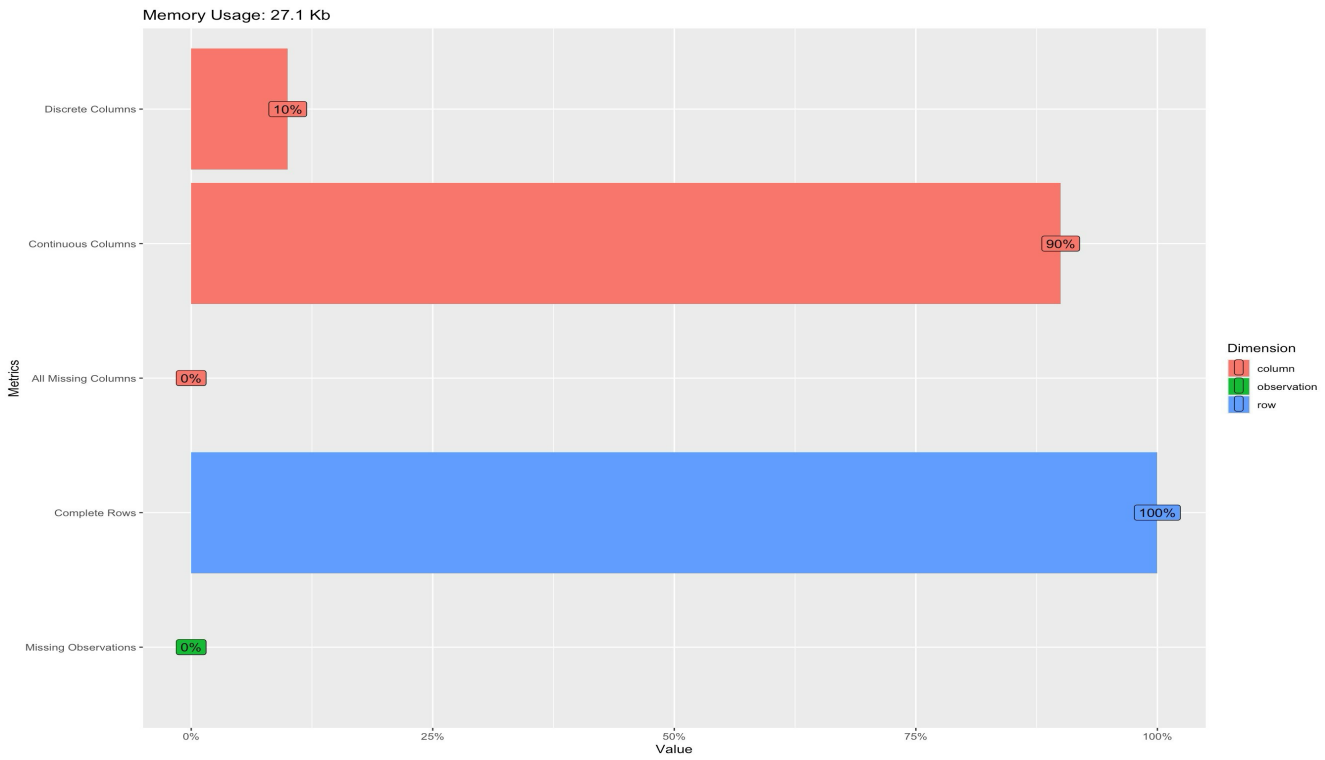
- Correlation Analysis
  - Figure S.1: Correlation matrix of cytokines (before and after filtration)
- Principal Component Analysis
  - Figure S.2: The scree plot displays the proportion of total variance for each PC
  - Figure S.3: Principal Component Analysis plot
- Statistical Modeling
  - Regularized Logistic Regression
    - Figure S.4: Regularized Logistic Regression Tunning via Cross-Validation
    - Table S.1: Regularized Logistic Regression Model parameters
    - Figure S.5: Top Predictors identified by Regularized Logistic Regression Model
  - Random forest
    - Figure S.6: Random Forest Tunning via Cross-Validation
    - Table S.2: Random Forest parameters
    - Figure S.7: Top Predictors identified by Random Forest Model
  - Generalized Linear Mixed Models with LASSO Regularization (GLMMLasso)
    - Figure S.8: GLMMLasso Tunning Curve
    - Table S.3: GLMMLasso Model parameter
    - Figure S.9: Top Predictors identified by GLMMLasso

Section1: Basic Statistics

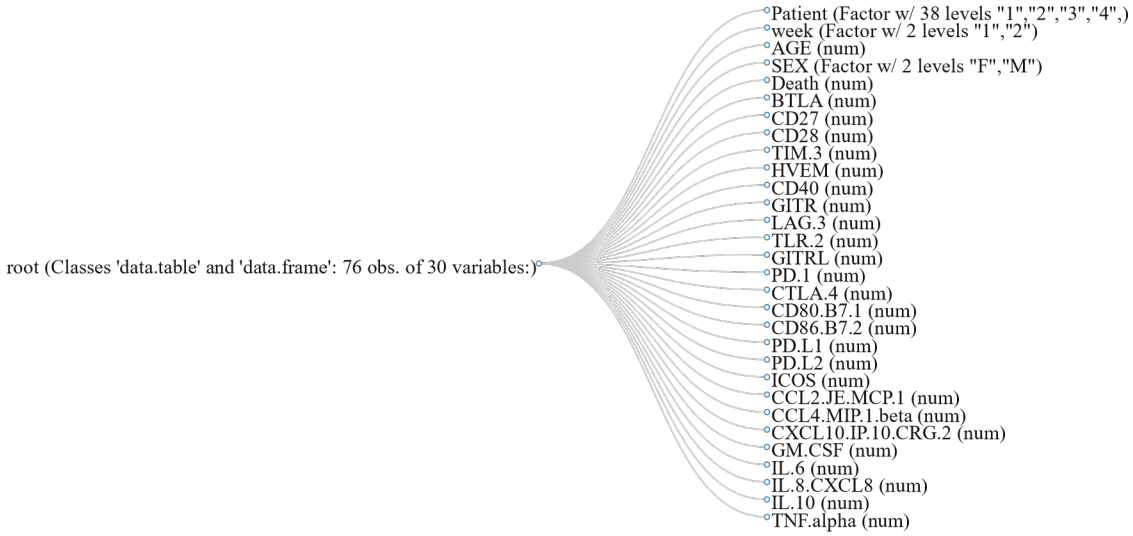
Raw Counts

Name	Value
Rows	76
Columns	30
Discrete columns	3
Continuous columns	27
All missing columns	0
Missing observations	0
Complete Rows	76
Total observations	2,280
Memory allocation	27.1 Kb

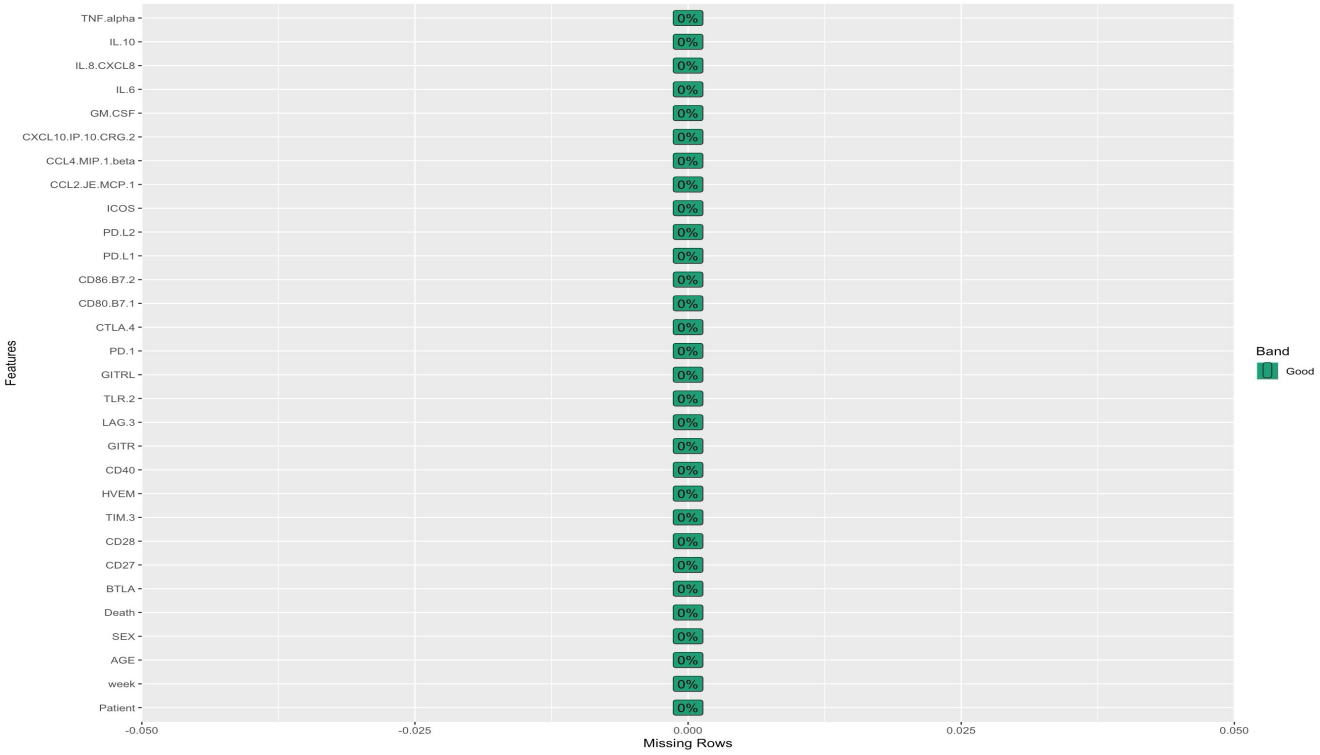
Percentages



Data Structure

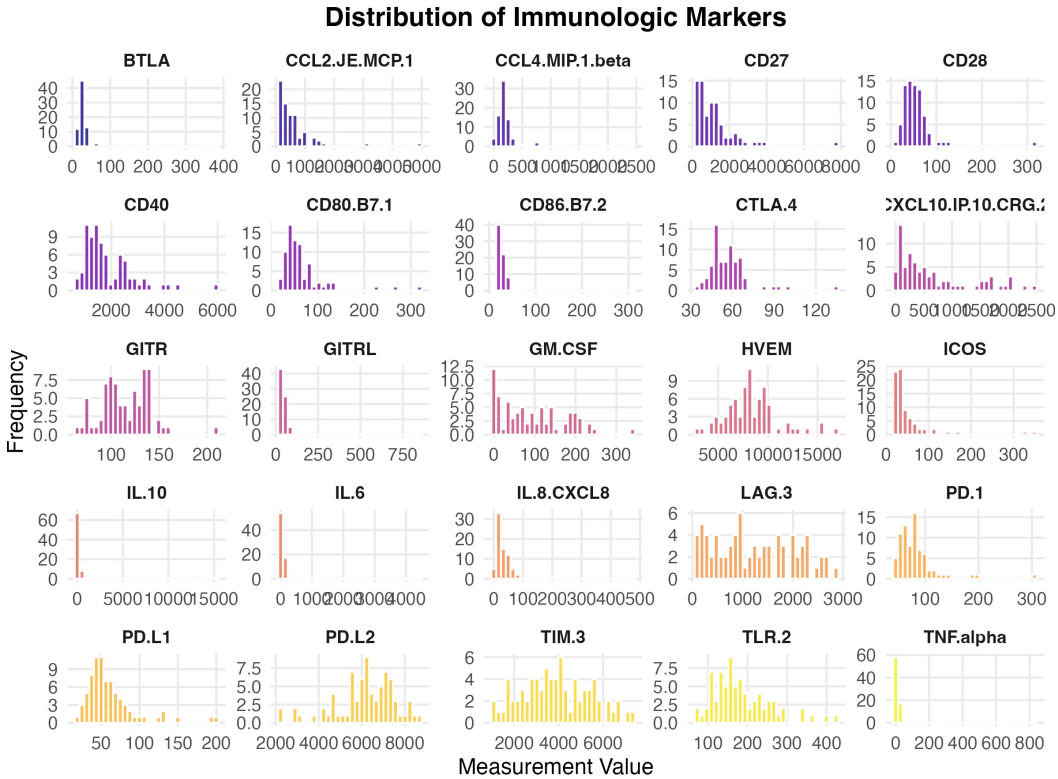


Missing Data Profile

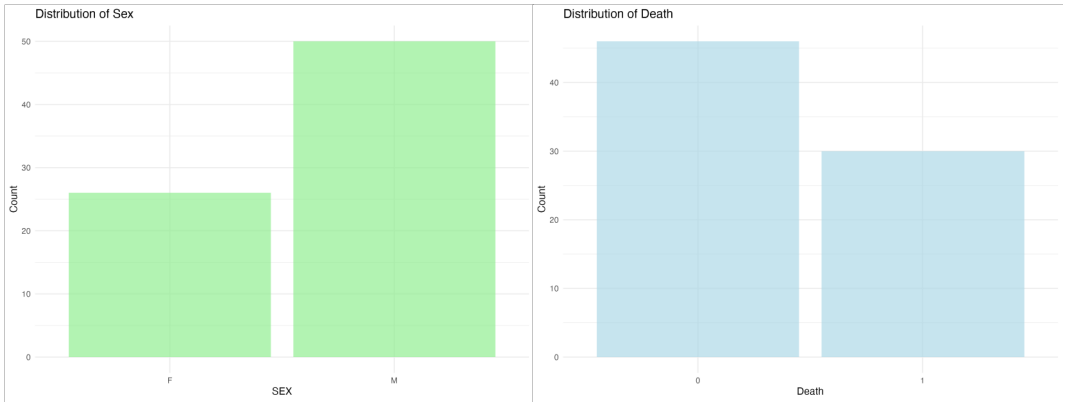


Univariate Distribution

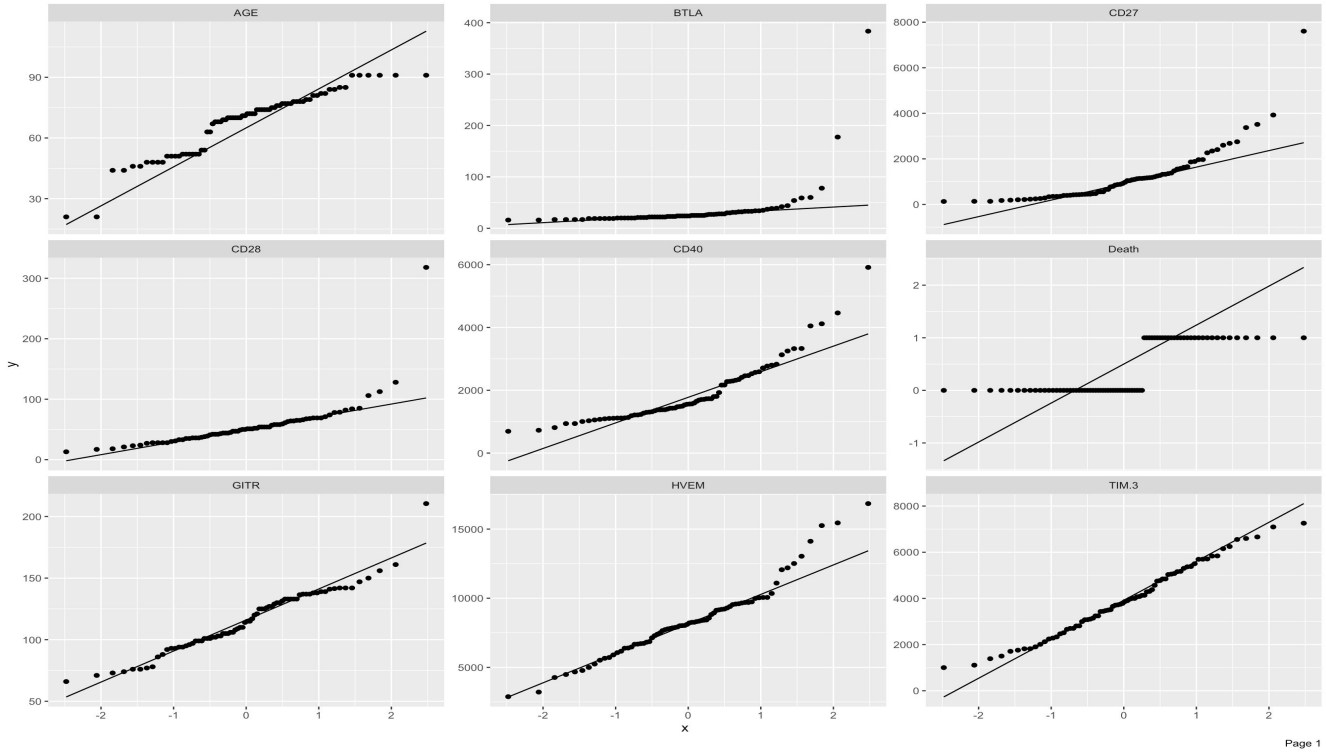
Histogram



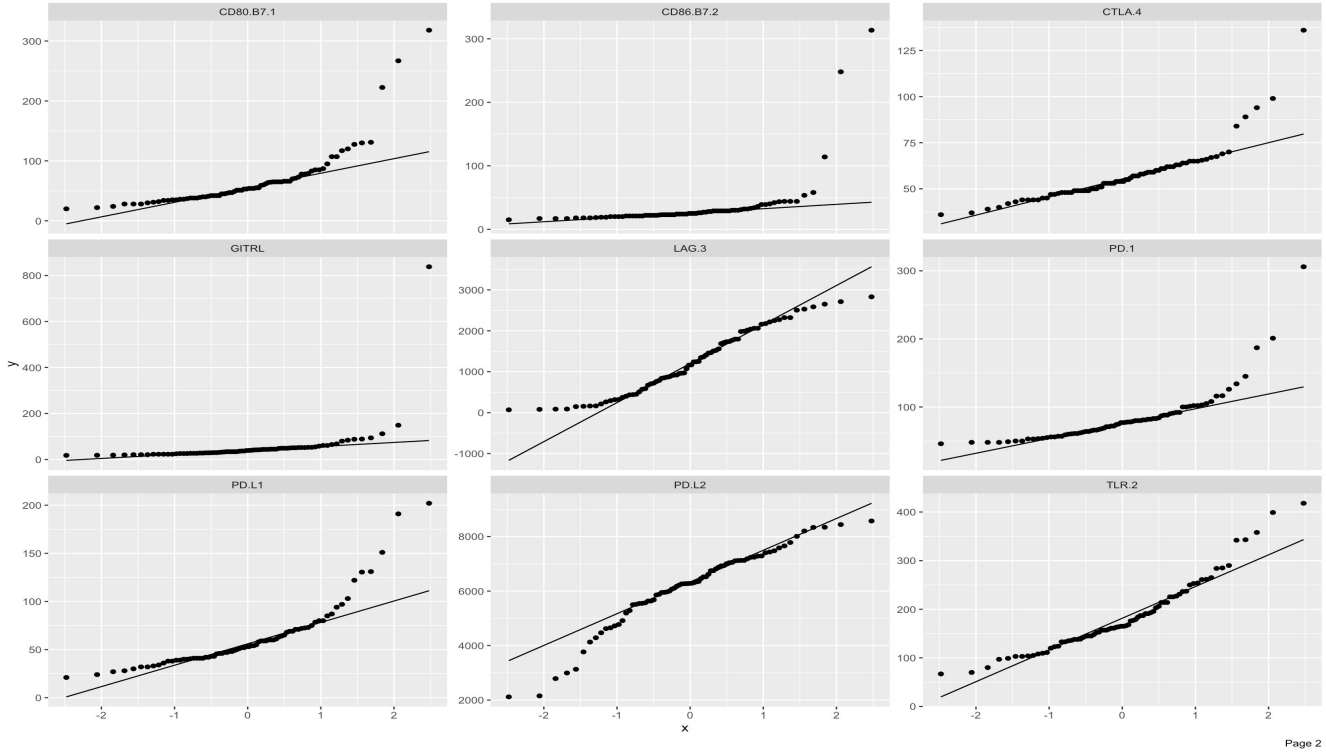
Bar Chart (with frequency)


Distribution of Sex
Count
SEX
F
M
0
10
20
30
40
50
Distribution of Death
Count
Death
0
1
0
10
20
30
40

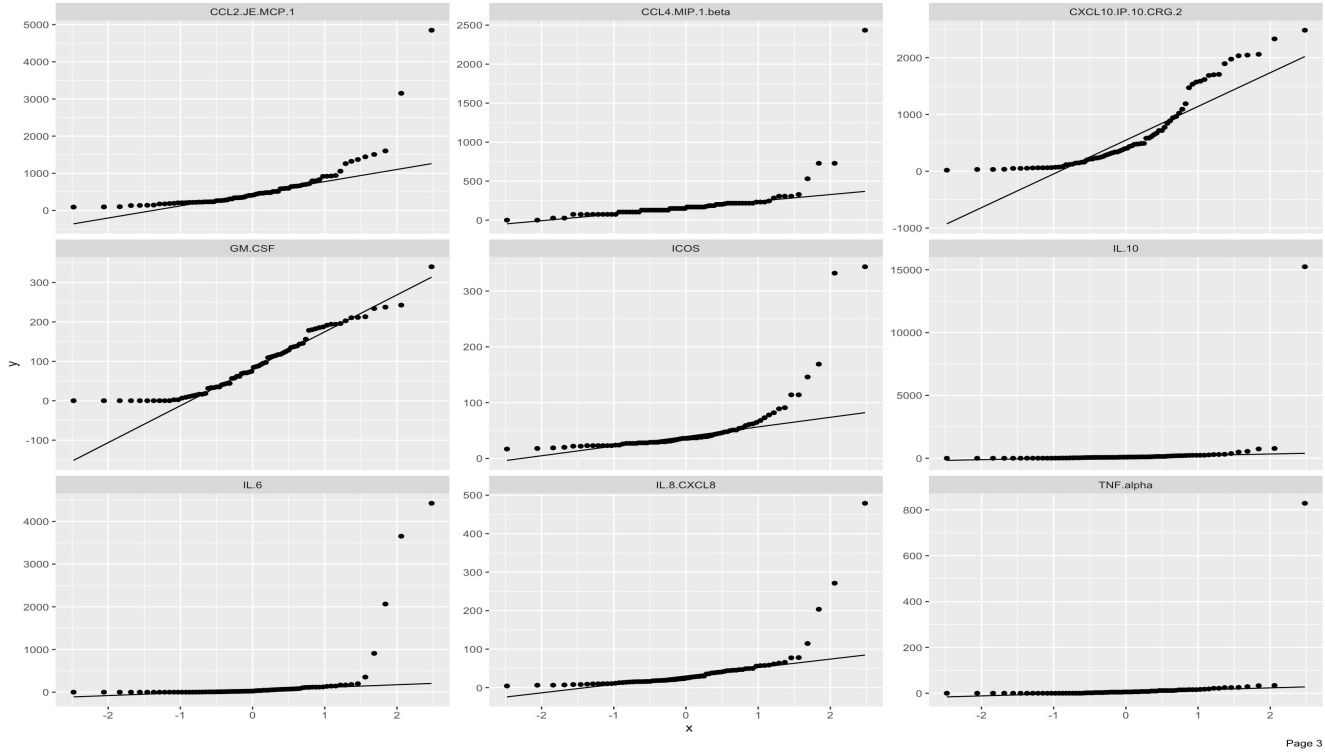
QQ Plot


AGE
BTLA
CD27
CD28
CD40
Death
GITR
HVEM
TIM.3
x
y

Page 1


CD80.B7.1
CD86.B7.2
CTLA.4
GITRL
LAG.3
PD.1
PD.L1
PD.L2
TLR.2
x
y

Page 2



Page 3

Section 2: Correlation Analysis

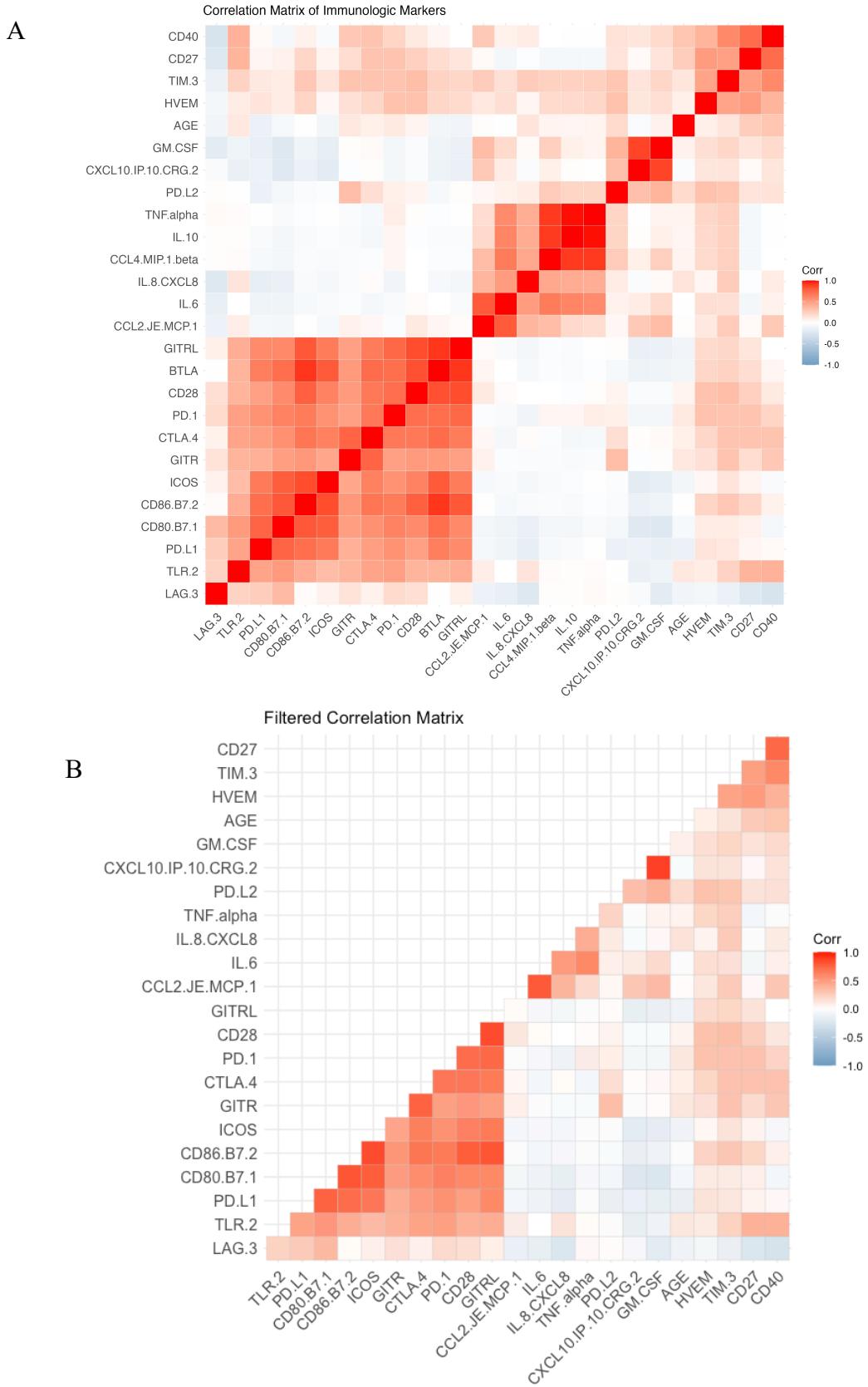


Figure S.1: The heatmap displays pairwise Pearson correlation coefficients between cytokine levels. Red shades indicate positive correlations, while blue shades indicate negative correlations. Strongly correlated variables (|r| > 0.85) were identified (A) and considered for removal as shown in (B) to reduce redundancy and improve model stability.

Principal Component Analysis

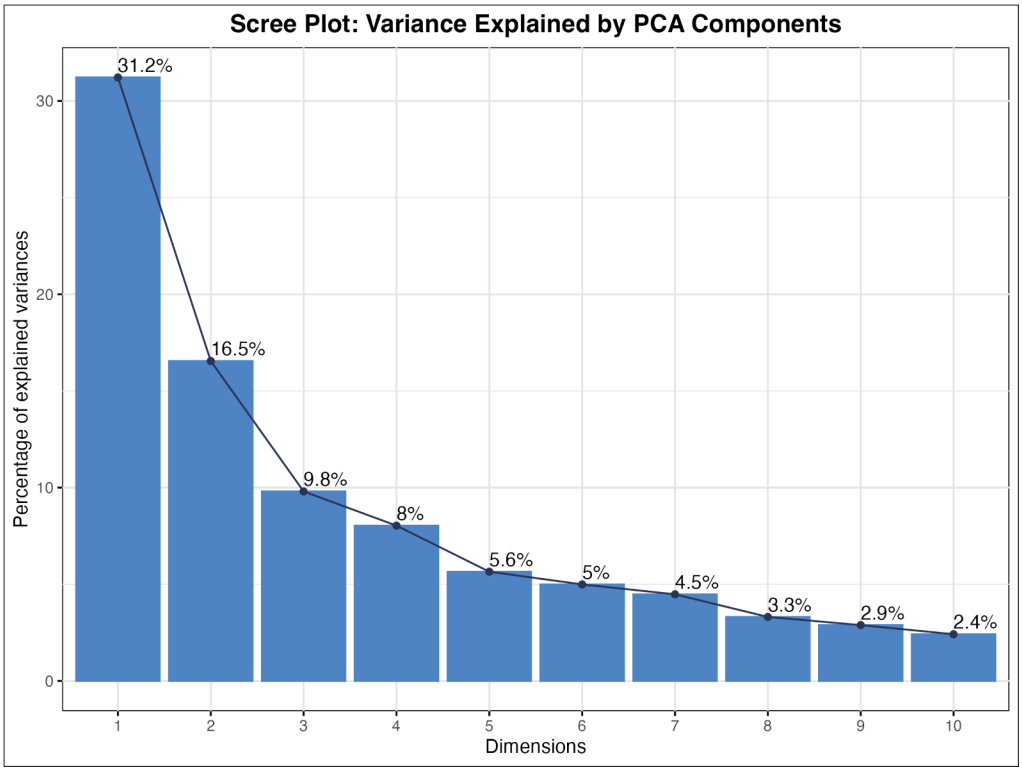


Figure S.2: The scree plot displays the proportion of total variance explained by each principal component.

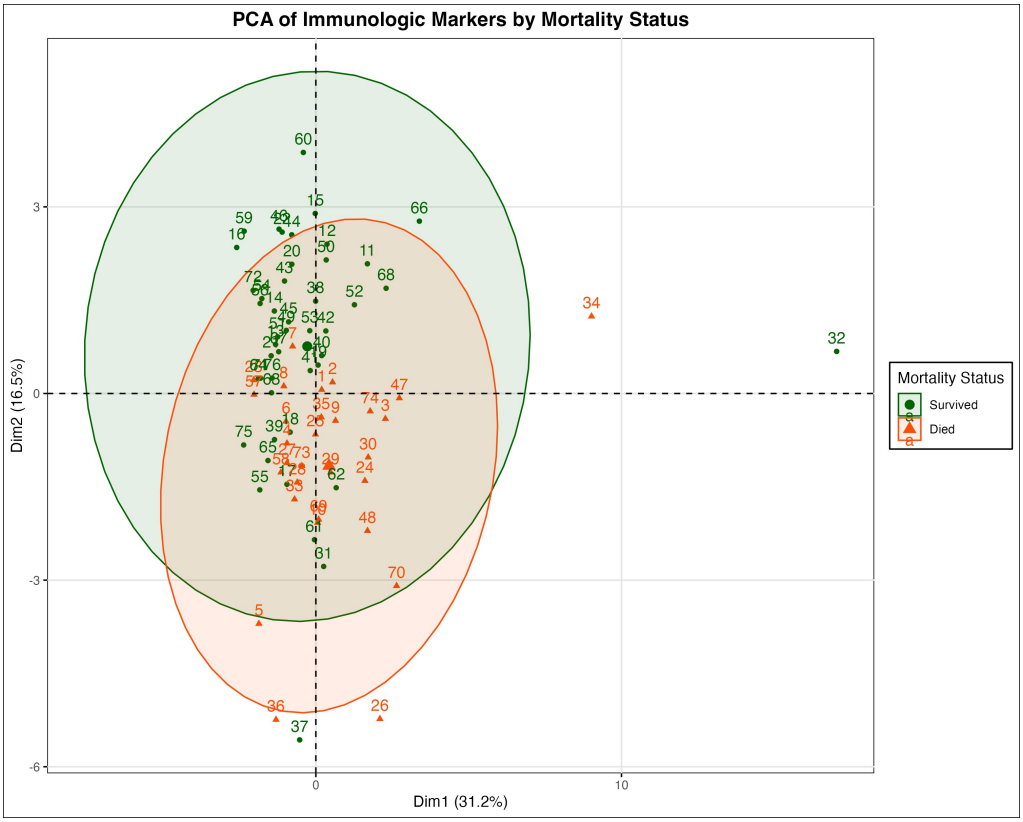


Figure S.3: Principal Components 1 and 2 (Dim1 and Dim2) together account for 47.7% of the total variation in the dataset. The PCA plot shows individual observations projected onto the first two principal components, highlighting the underlying structure and patterns within the cytokine profiles.

Statistical Modeling & Model Evaluation

Regularized Logistic Regression

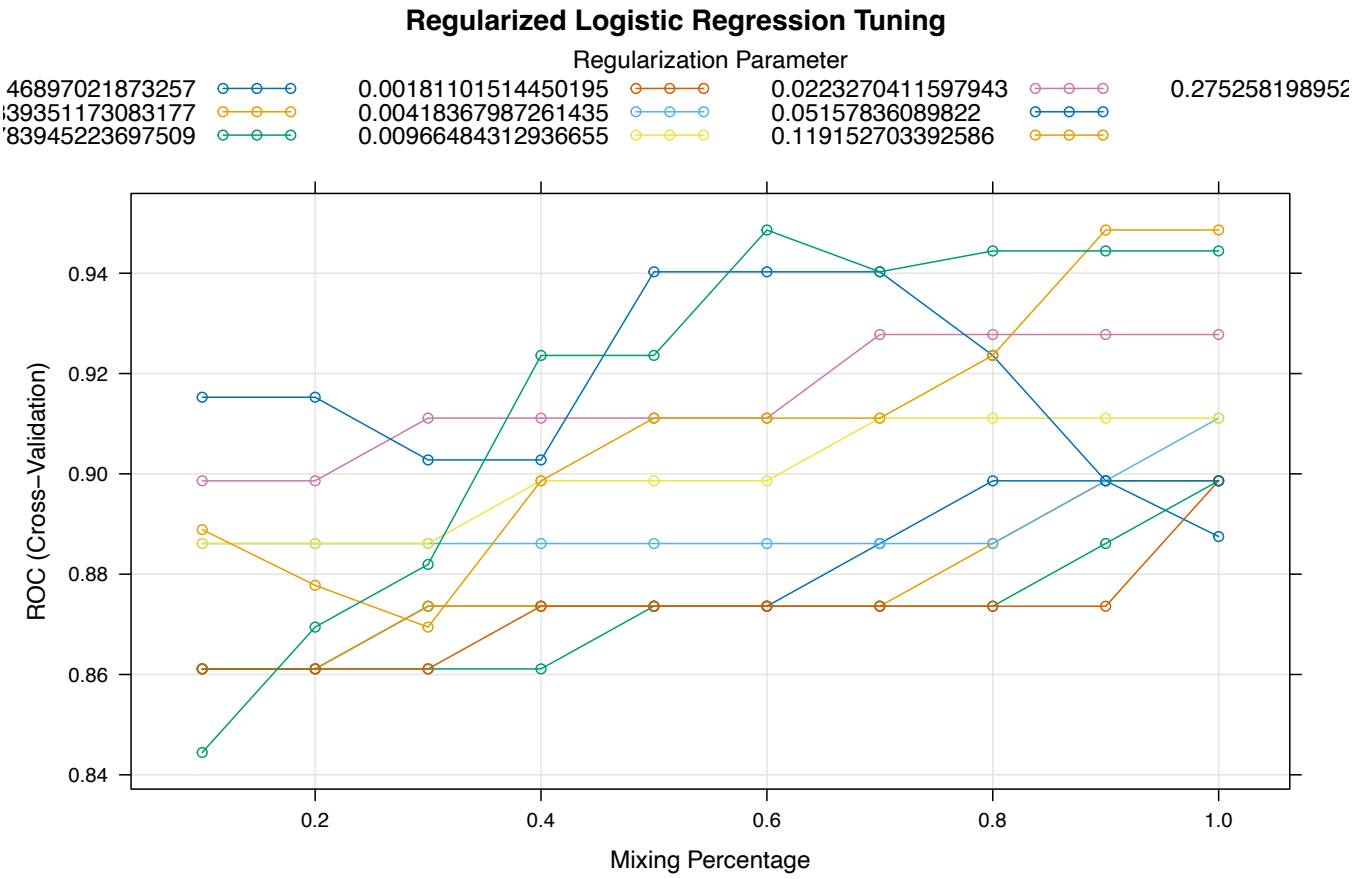


Figure S.4: Regularized Logistic Regression via Cross-Validation. This plot shows ROC AUC values obtained from 10-fold cross-validation for Elastic Net logistic regression models across a range of alpha (α) and lambda (λ) values. The y-axis represents the cross-validated ROC AUC score, while the x-axis reflects the mixing percentage (α), which balances LASSO and Ridge penalties. The optimal performance was achieved with α = 0.6 and λ = 0.275, indicating a balanced contribution from both LASSO and Ridge penalties. This combination yielded the best model in terms of discrimination ability.

Table S.1: Regularized Logistic Regression Model parameters

alpha	lambda	ROC	Sensitivity	Specificity	ROCSD	SensSD
0.6	0.2752581990	0.9486111	1.0000000	0.3833333	0.08587932	0.00000000

Confusion matrix: Regularized Logistic Regression Model:

Actual: Survived	Actual: Survived	Actual: Died
Predicted: Survived	10 (True Negatives, TN)	5 (False Negatives, FN)
Predicted: Died	1 (False Positives, FP)	2 (True Positives, TP)

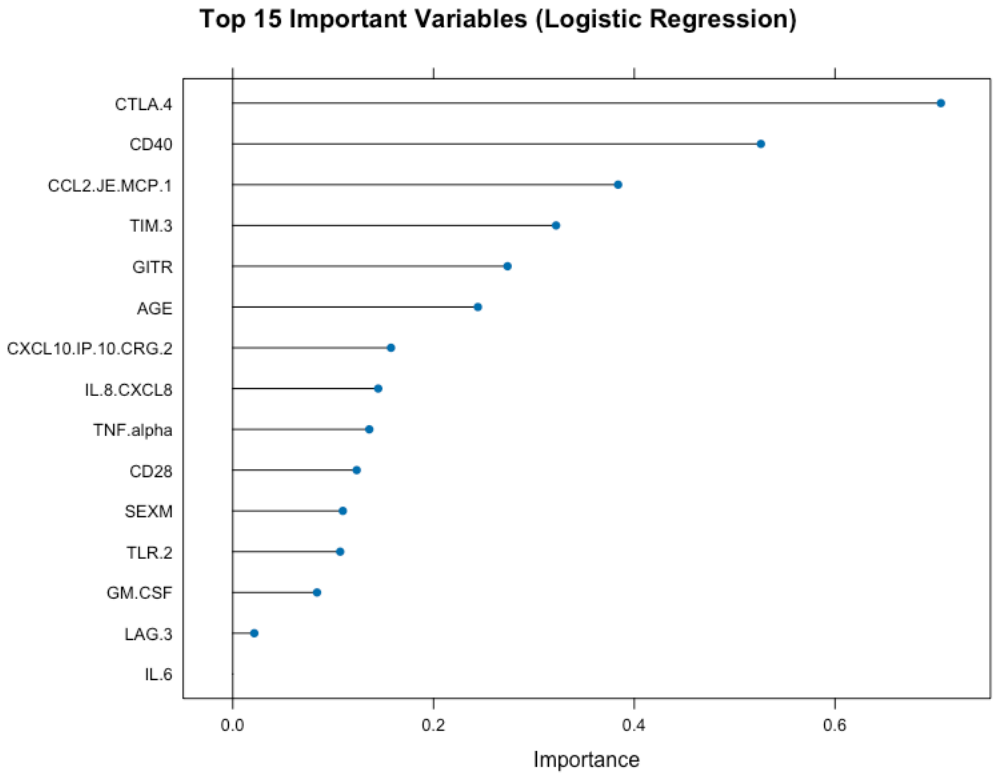


Figure S.5: Top Predictors identified by Regularized Logistic Regression Model

Random Forest

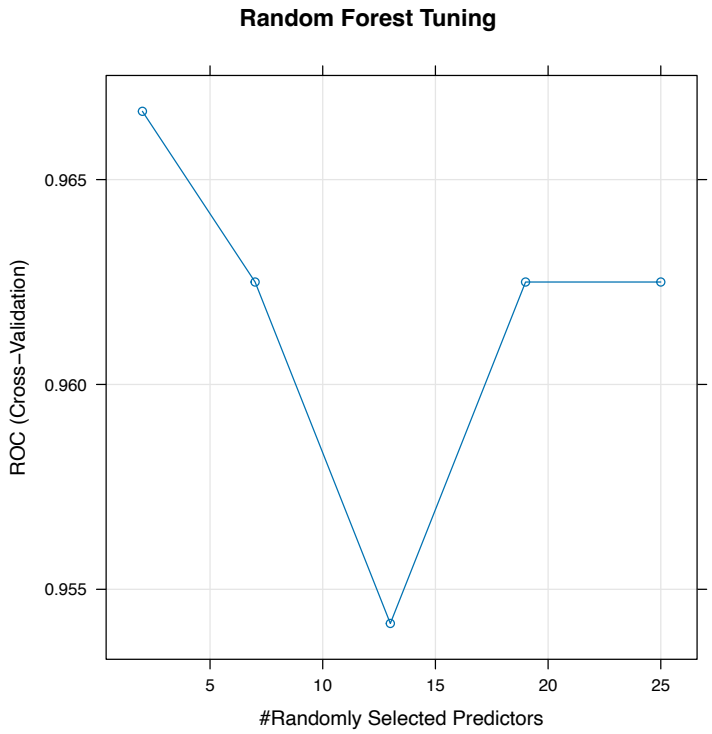


Figure S.6: Random Forest Tunning via Cross-Validation. This plot displays the performance of the Random Forest classifier across different values of "mtry", representing the number of randomly selected predictors at each split. The y-axis shows the mean ROC AUC from 10-fold cross-validation, while the x-axis indicates the "mtry" values tested. The highest performance was achieved at "mtry" = 2, with a cross-validated AUC of 0.97, indicating strong predictive accuracy.

Table S.3: Random Forest Model Performance

mtry	AUC (ROC)	Sensitivity	Specificity
2	**0.967**	**0.917**	0.733
7	0.963	0.858	0.717
13	0.954	0.858	0.800
19	0.963	0.850	0.850
25	0.963	0.850	0.850

Confusion matrix: Random Forest

Actual: Survived	Actual: Survived	Actual: Died
Predicted: Survived	10 (True Negatives, TN)	3 (False Negatives, FN)
Predicted: Died	1 (False Positives, FP)	4 (True Positives, TP)

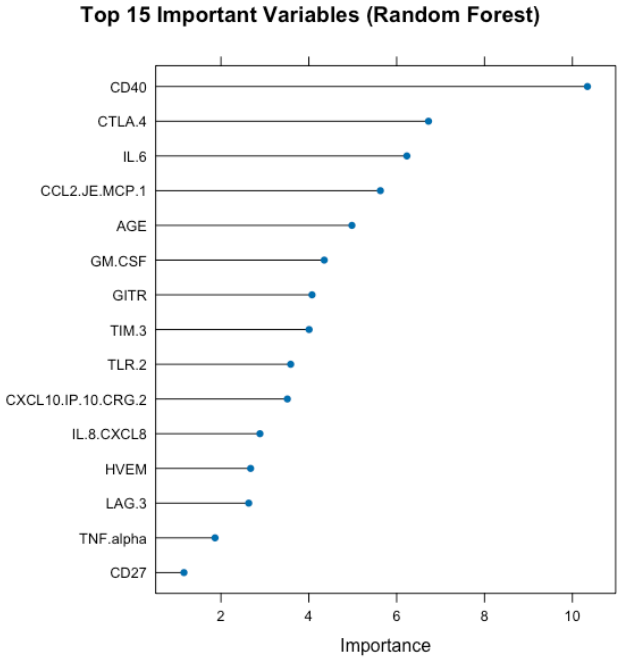


Figure S.7: Top Predictors identified by Random Forest Classifier

GLMMLasso

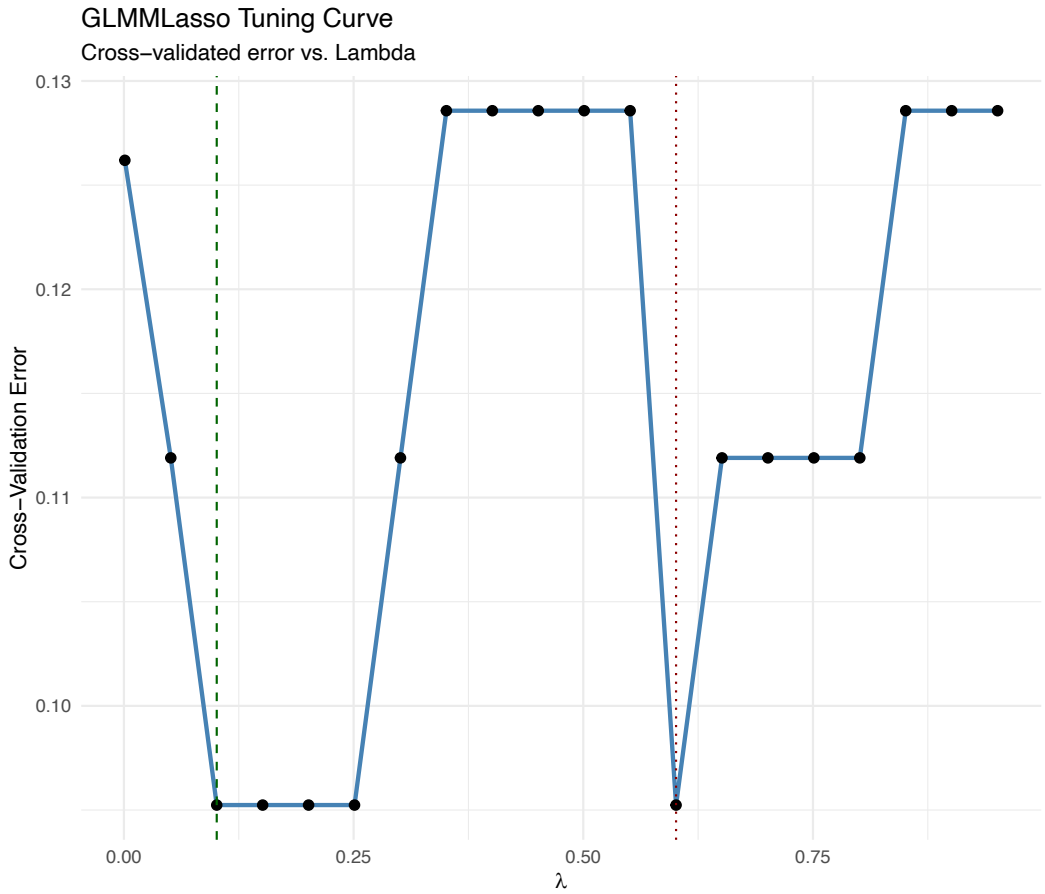


Figure S.8: GLMMLasso Tunning Curve. The plot displays cross-validation error (Y-axis) across a range of λ (lambda) values (X-axis). Each point represents the model's performance at a specific level of regularization. The

optimal λ value (λ = 0.101) was selected based on the minimum cross-validation error. This value balances model complexity and predictive accuracy by penalizing less informative variables, promoting sparsity while retaining key predictors.

Table S.4: GLMMLasso Model performance

Metric	Value
Accuracy	0.80 (95% CI: 0.3229 – 0.8366)
Kappa	0.25
Selected Lambda (λ)	0.101
Random Effect (Patient ID) Std. Dev	0.49
Sensitivity (Recall for class 0)	0.5000
Specificity	0.8000
Positive Predictive Value (Precision for class 0)	0.8333
Negative Predictive Value	0.4444
Balanced Accuracy	0.8200
No Information Rate (NIR)	0.6667
P-value [Acc > NIR]	0.7970
McNemar's Test P-value	0.2207

Confusion matrix: GLMMLasso

Actual: Survived	Actual: Survived	Actual: Died
Predicted: Survived	11 (True Negatives, TN)	2 (False Negatives, FN)
Predicted: Died	1 (False Positives, FP)	4 (True Positives, TP)

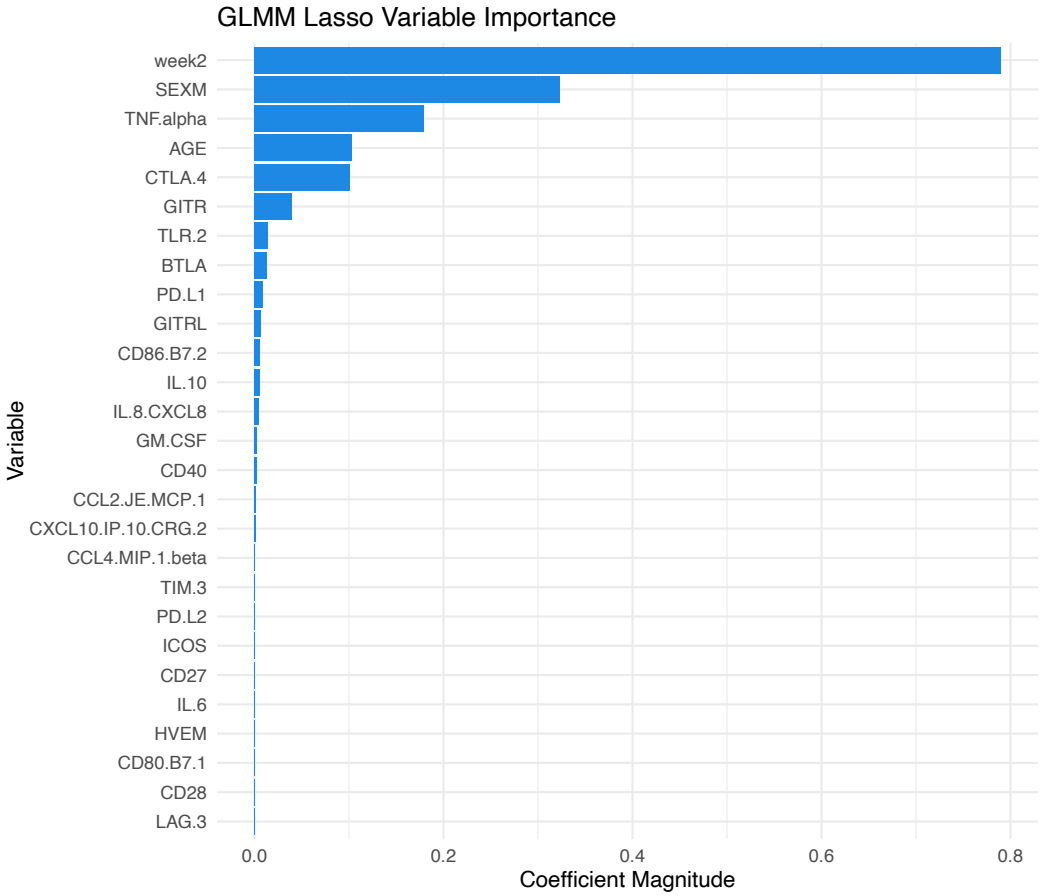


Figure S.9: Top Predictors that were identified by GLMMLaosso