Predictive Modeling of In-Hospital Mortality in ICU-Admitted Heart Failure Patients

Peter Dunson[1], Ephrata Getachew[2], Nate Stevenson[3]

[1]Kenyon College, [2]Amherst College, [3]University of Virginia

Introduction

- Approximately 6.5 million Americans have heart failure, with nearly 1 million new cases annually, we aim to predict and identify patient risk profiles.

Data

- Mimic-III database is a freely accessible single-center database containing information about adult patients admitted to intensive care between 2001 and 2012 coming from one medical center in Boston, Massachusetts.
- The heart failure dataset in Mimic III database has 1,177 observations with 69 variables covering demographics, vital signs, comorbidities, and lab results.

Research Questions

- Which **clinical** and **demographic** variables are the most **significant predictors** of in-hospital **mortality**?
- How can we balance between model **interpretability** and predictive **accuracy**?
- How are varying **patient profiles**, based on their **features**, associated with the outcome?

Model Selection

- **Elastic net:** Combines the penalties from **LASSO** and **Ridge,** allowing for a more flexible model.

$$(\hat{\beta}_0, \hat{\beta}) = \arg \min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[\overbrace{\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2}^{\text{OLS}} + \lambda \overbrace{\left(\frac{1-\alpha}{2}\underbrace{\|\beta\|_2^2}_{\text{Ridge}} + \alpha\underbrace{\|\beta\|_1}_{\text{LASSO}} \right)}^{\text{Elastic Net}} \right]$$

- **Random Forest:** Improves accuracy by reducing overfitting through **bagging** and **random feature selection**.

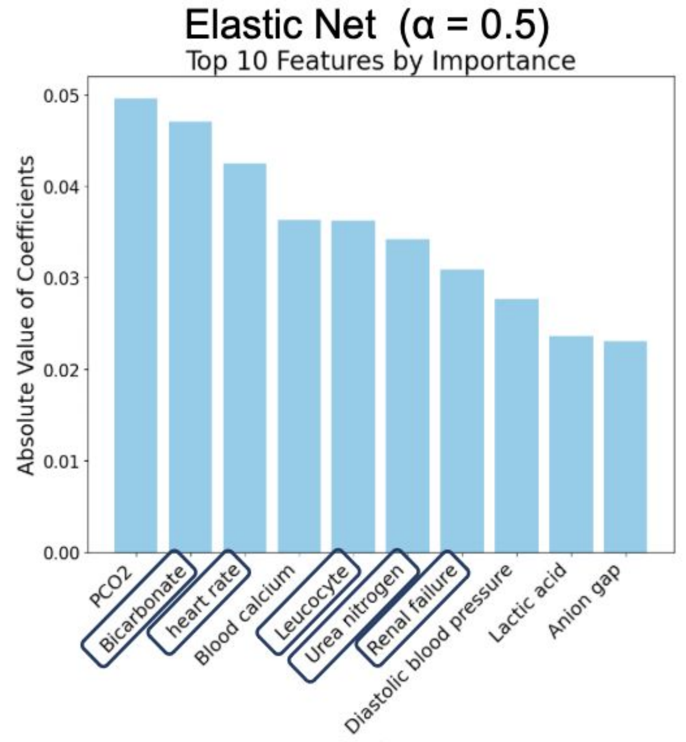
Results

Comparing the AUC of bootstrapped samples

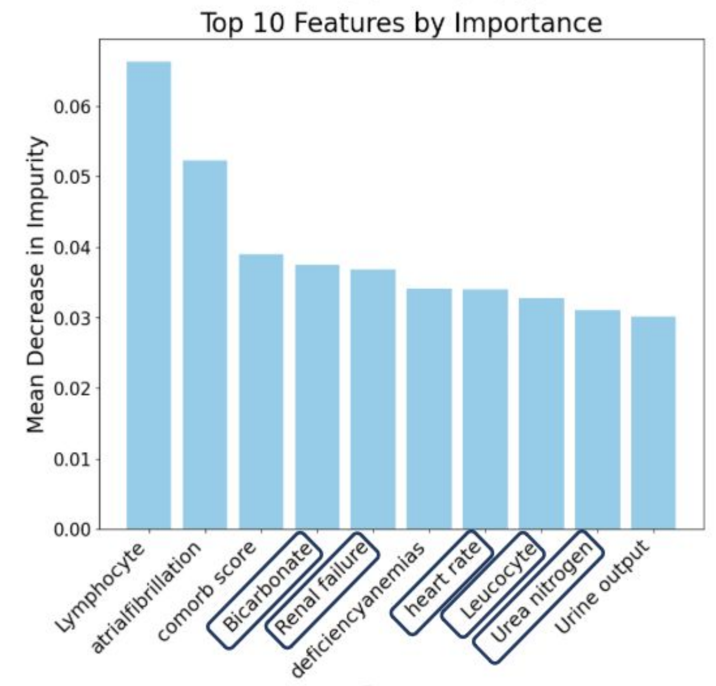
Method	Mean	Std. Deviation
Elastic Net	0.810	0.016
Random Forest	0.838	0.034

Variable Importance

Elastic Net (α = 0.5)
Top 10 Features by Importance



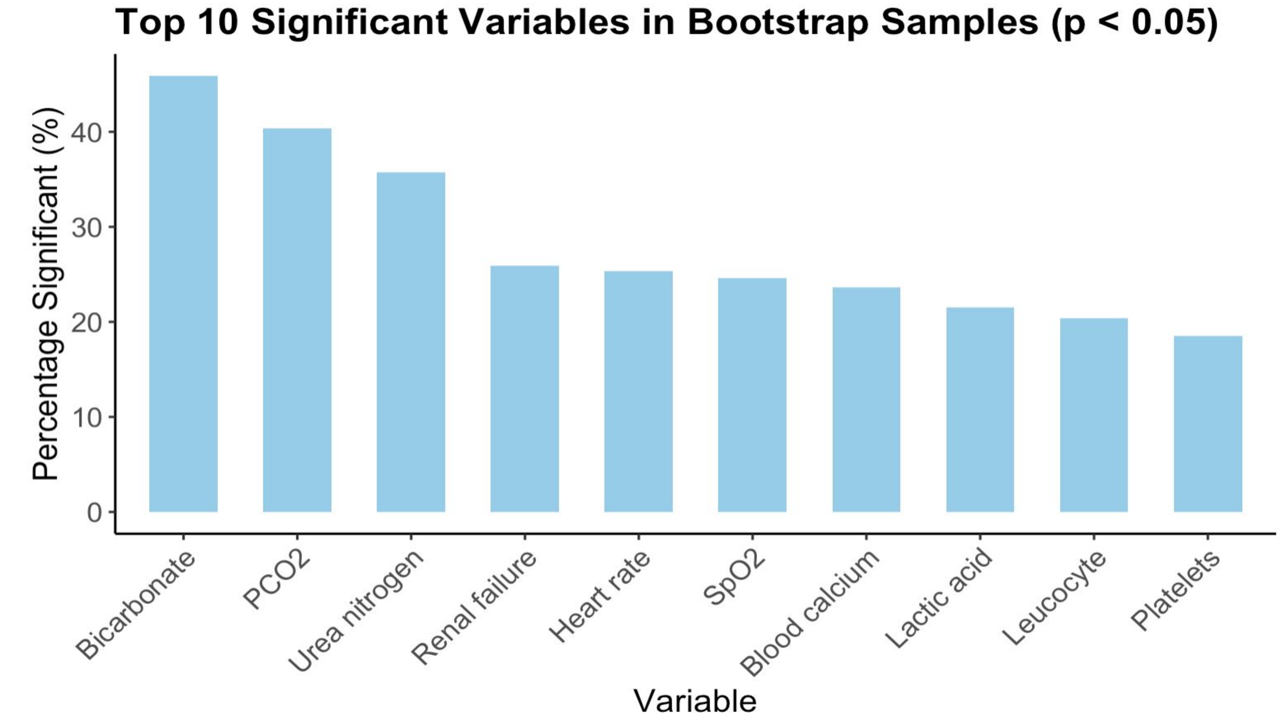
Random Forest
Top 10 Features by Importance



Variable Selection Using Elastic Net and Logistic Regression with Bootstrapping

On each bootstrapped dataset, run elastic net → Select set of variables on the **training data** → Refit logistic regression using the response and the selected set of variables on the **test data**

Top 10 Significant Variables in Bootstrap Samples (p < 0.05)

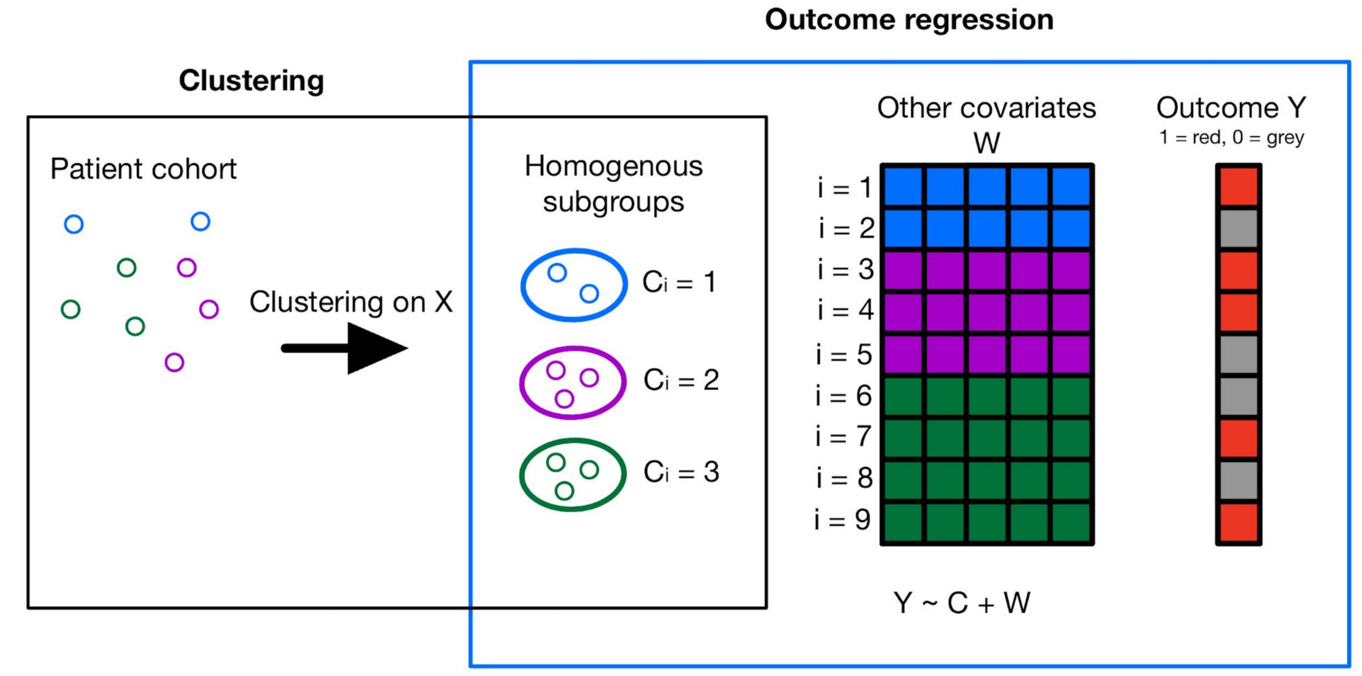


- The top 10 significant predictors of in-hospital mortality in the bootstrapped logistic regression are
  - Vital signs: Heart rate, SpO2
  - Lab results: Bicarbonate, PCO2, Urea nitrogen, Renal failure, Blood calcium, Lactic acid, Leucocyte, and Platelets.

Clustering Analysis

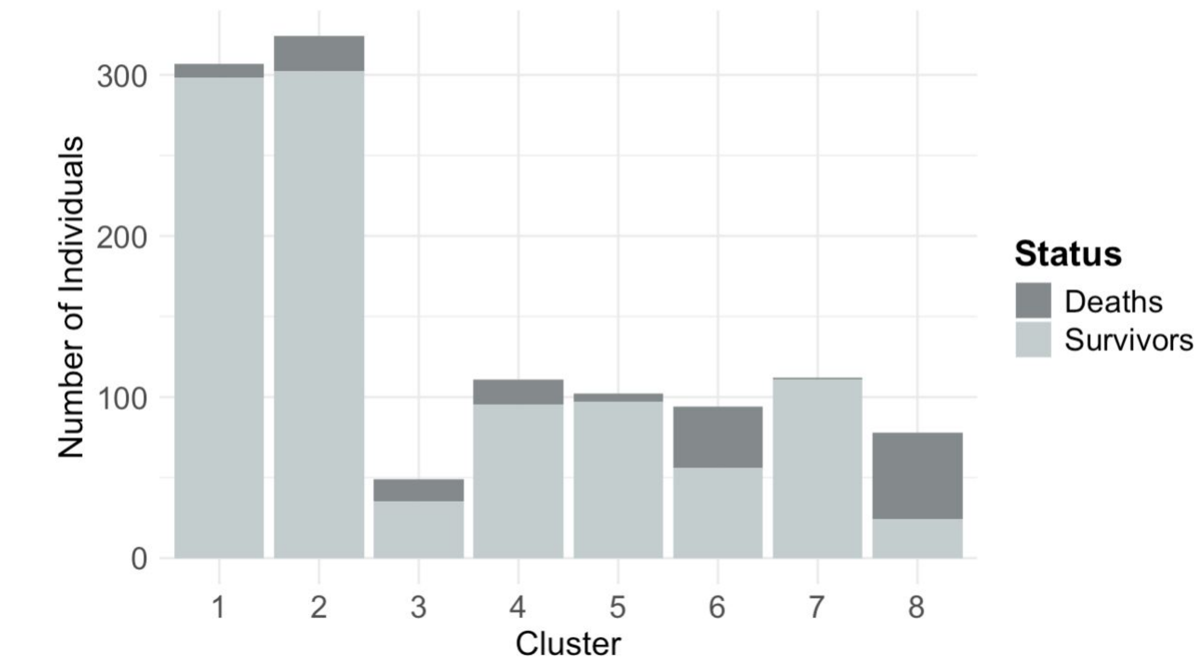
Bayesian Profile Regression

- **Model Overview**:
  - Dirichlet Process Prior
  - Semi supervised mixture model.
  - Model identifies clusters with similar risk profiles.
  - Features that vary the most across clusters are selected to illustrate variations.
- **Uncertainty Quantification**:
  - MCMC samples from the posterior characterize uncertainty.
  - Credible intervals for feature means are used to visualize uncertainty.



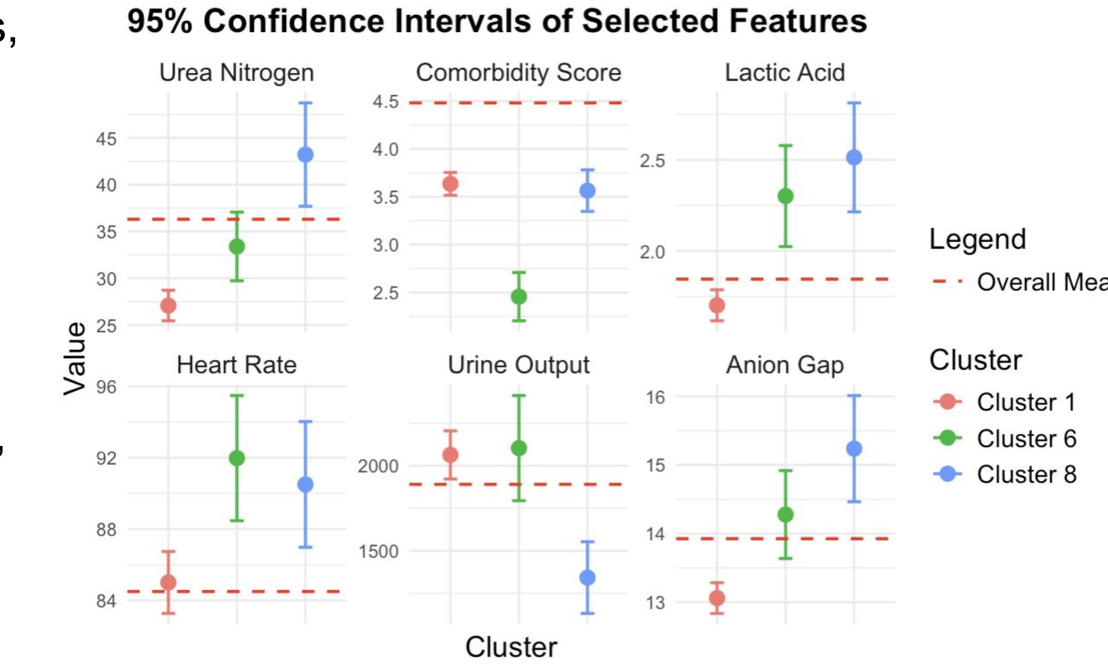
Results from BPR

Cluster Sizes and Outcome Distribution



- Cluster 1 - 307 patients, risk = 0.18
  - Deaths: 9 of 307, 2.9%
- Cluster 6 - 94 patients, risk = 0.18
  - Deaths: 38 of 94, 40.4%
- Cluster 8 - 78 patients, risk = 0.17
  - Deaths: 54 of 78, 69.2%

95% Confidence Intervals of Selected Features

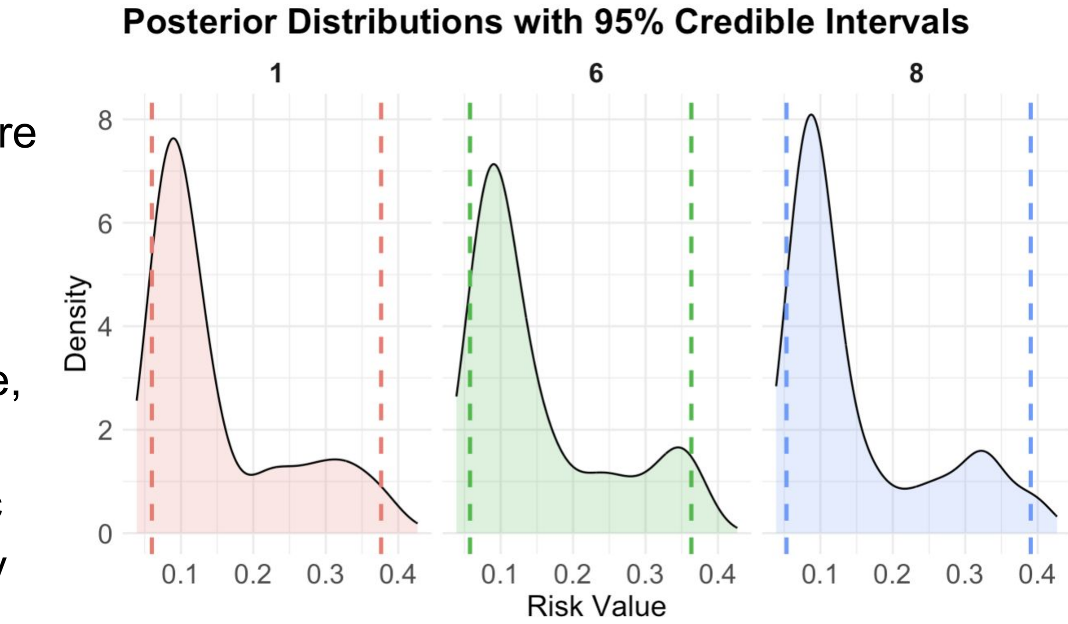


*Risk represents the probability of an adverse outcome occurring in each cluster based on the identified features.

Conclusions on BPR

- **Key Features**:
  - Cluster 1: High Urea Nitrogen, Renal Failure prevalence
  - Cluster 6: Low Comorbidity Score, minimal Renal Failure, high outcome risk
  - Cluster 8: High Lactic Acid, low COPD, very high outcome risk

Posterior Distributions with 95% Credible Intervals



Acknowledgment

We would like to thank our mentors, Rahul Ladhania, Ph. D., Snigdha Panigrahi, Ph. D., and co-mentor Mengbing Li, for their guidance and support and the BDSI program for the opportunity to conduct research.

Limitations and References