

SAP for Classification of ARDS Data

Robert Edwards

I. Population

- Patients with Acute Respiratory Disease Syndrome.

II. Primary Objective

- Can ECMO treatment survival be accurately predicted by PreECMO biomedical markers?

III. Secondary Objective

- What is the future expected performance of predictions?
- Which biomedical markers are needed for accurate prediction and which can be dropped?

IV. Data Collection

V. Variables Under Consideration

- **ECMO_Survival** (Categorical) - a survival indicator
 - Y = survival
 - N = non-survivor
- **Gender** (Categorical) - Patient gender
 - m = male
 - f = female
- **Indication** (Categorical) - A disease indicator
 - ALF = acute lung failure
 - 1 = viral pneumonia
 - 2 = bacterial pneumonia
 - 3 = aspiration pneumonia

- 4 = ARDS Trauma
- 5 = ARDS surgery
- 6 = Chemo
- 7 = other
- Age (years) - Age of patient
- Pre ECMO - biomarkers before ECMO treatment
 - RR (Continuous) - Respiratory Rate
 - Vt (Continuous) - Tidal volume
 - FiO2 (Continuous) - Inspire fraction of oxygen
 - Ppeak (Continuous) - Peak airway pressure
 - Pmean (Continuous) - Mean airway pressure
 - PEEP (Continuous) - Positive end expiratory pressure
 - PF (Continuous) - Arterial partial pressure of oxygen/inspired fraction of oxygen ratio
 - SpO2 (Continuous) - Periperal oxygen saturation
 - PaCO2 (Continuous) - Arterial pressure of carbon dioxide
 - pH (Continuous) - Arterial pH
 - BE (Continuous) - Arterial base excess
 - Lactate (Continuous) - Arterial lactate
 - NAdose (Continuous) - Noradrenaline dose
 - MAP (Continuous) - Mean arterial pressure
 - Creatinine (Continuous) - Serum Creatinine is an important indicator of renal health because it is an easily measured byproduct of muscle metabolism that is excreted unchanged by the kidneys.
 - Urea (Continuous) - Also known as carbamide Urea serves an important role in the metabolism of nitrogen-containing compounds by animals and is the main nitrogen-containing substance in the urine of mammals. High concentrations in the blood can be damaging.
 - CK (Continuous) - Creatine Kinase is assayed in blood tests as a marker of damage of CK-rich tissue such as in myocardial infarction (heart attack), rhabdomyolysis (severe muscle breakdown), muscular dystrophy, autoimmune myositides, and acute kidney injury.
 - Bilirubin (Continuous) - Bilirubin is excreted in bile and urine, and elevated levels may indicate certain diseases.
 - Albumin (Continuous) - Albumin Serum albumin is the main protein of human blood plasma. It binds water, cations (such as Ca^{2+} , Na^+ and K^+), fatty acids, hormones, bilirubin, thyroxine (T4) and pharmaceuticals (including barbiturates): its main function is to regulate the oncotic pressure of blood.
 - CRP (Continuous) - C reative protein
 - Fibrinogen (Continuous) -
 - Ddimer (Continuous) -
 - ATIII (Continuous) - Anti-thrombin III
 - HB (Continuous) - Haemaglobin
 - Leukocytes (Continuous) -
 - Platelets (Continuous) -
 - TNFa (Continuous) -
 - IL6 (Continuous) - Interleukin 6 is an interleukin that acts as both a pro-inflammatory cytokine and an anti-inflammatory myokine.
 - IL8 (Continuous) - Interleukin 8 is an important mediator of the immune reaction in the innate immune system response.
 - siL2 (Continuous) -

More information about protein pathways can be found here: www.uniprot.com

VI. Missing Data Procedures

- Cases without `ECMO_Survival`, `Gender`, or `Indication` are to be removed from the analysis
- Cases with less than 50% of other covariates missing to have missing data imputation performed

Imputation Method

Research into how the data are missing will need to be conducted to determine the most appropriate imputation method:

- Mean Imputation -
- Median Imputation - for skewed data
- KNN Imputation -

VII. Summaries to be Presented:

Missing Data

- Counts of number of missing observations to be given for each variable (or a table)
- If any patterns to the missing data are found an appropriate table will be included

Categorical Data

- Frequency table and relative frequency (proportions) for:
 - `ECMO_Survival`
 - `Gender`
 - `Indication`

Continuous Variables

- Boxplots, mean, standard deviation, median, IQR

VIII. Models to be Fitted

Models

- Logistic Regression (main dissertation)

Variable Selection

- Lasso Regression will be used for variable selection

IX. Advanced Models to be Fitted

Models

- **Support Vector Machine**
- Decision Tree
- Random Forest
- K-Nearest Neighbors
- Neural Net

Other Analyses

- Bayesian models (logistic regression, knn)

X. Model Performance

Model performance will be evaluated on:

- Accuracy
- Precision
- Sensitivity
- Specificity
- F1 Score (?)

In addition the following tables/plots will be reported:

- Confusion matrix
- ROC curve