Cluster Analysis of Acute Respiratory Distress Syndrome

Kumud Sharma 2867860S

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

Discussion of the context

Acute respiratory distress syndrome (ARDS) represents a significant cause of respiratory failure in critically ill patients, with an incidence rate of approximately 200,000 cases per year in the USA. Globally, ARDS is estimated to affect 10% of critically ill patients, with an associated mortality rate ranging from 30% to 40% (Hudson, et al., 2005) (G, et al., 2016). Acute respiratory distress syndrome (ARDS) is defined as acute hypoxic respiratory failure ($PaO_2/\ FiO_2 < 300\ mm\ Hg$), bilateral chest infiltrates, and the absence of cardiac failure as the primary diagnosis (Force, et al., 2012). Since the first consensus on defining ARDS in 1988, experts have continuously debated whether patients should be classified based on disease trajectory, clinical presentation, biological markers, or a combination of these factors. (Bernard, et al., 1994). Despite 50 years of research, there are no successful disease-altering pharmacological interventions in the treatment of ARDS. Biological heterogeneity subsumed within this clinical syndrome is considered one of the main causes for failure of pharmacological interventions in randomized controlled trials (RCTs) (School of Mathematics and Statistics, 2023-2024).

Aims of the proposed research

The primary aim of this analysis is to identify clusters in the biomedical markers data. Clustering analysis is used as an exploratory technique and to understand the inherent structure of the data. The secondary aim is to determine whether the clusters correspond to the outcome variables for survival. Additionally, the analysis further evaluates the impact of significant predictors identified through clustering on survival outcomes, utilizing clustering as a dimensionality reduction technique.

Description of the study and variables involved

The data consists of 450 patients. The data is available on the patient's biomarkers both before ECMO treatment and first day after ECMO treatment. However, this analysis only pertains to the data for first day after ECMO treatment (Day1ECMO). There are 29 biomedical markers. (School of Mathematics and Statistics, 2023-2024)

Variables under consideration:

- Identifiers: *Pt_ID* - A unique code for each patient.
- Outcome Variables :
  - *ECMO_Survival* - Survival indicator (Y = survivor, N = non-survivor).
  - *Hospital_Survival* - Secondary survival indicator (Y = survivor, N = non-survivor)
- Biomedical Markers (Day1ECMO):

  RR, *Vt* , FiO_2 (Inspired fraction of oxygen), *Ppeak* (Peak airway pressure), *Pmean* (Mean airway pressure), *PEEP* (Positive end expiratory pressure), *PF* (Arterial partial pressure of oxygen/inspired fraction of oxygen ratio), SpO_2 (Peripheral oxygen saturation), $PaCO_2$ (Arterial partial pressure of carbon dioxide), *pH* (Arterial pH1), *BE* (Arterial base excess), *Lactate* (Arterial lactate), *NAdose* (Noradrenaline dose), *MAP* (Mean arterial pressure), *Creatinine*, *Urea*, *CK* (Creatinine Kinase), *Bilirubin*, *Albumin*, *CRP* (C reactive protein), *Fibrinogen*, *Ddimer*, *ATIII* (Anti-thrombin III), *Leukocytes*, *Platelets*, *TNFa*, *IL6*, *IL8*, *silL2* (School of Mathematics and Statistics, 2023-2024)

CHAPTER 2: ANALYSIS OF THE DATA

DATA PREPROCESSING

The process below describes the detailed data wrangling techniques being used. After loading the relevant dataset, the PreECMO biomedical markers are removed as the analysis only pertains to Day1ECMO biomedical markers. The Patients are arranged in ascending order of patient's ID's. Then, the column names are shortened by removing the Day1ECMO prefix from the columns for easier readability. The *glimpse* function is used to check the class of each variable. The variables are converted to relevant class type. Since the *Pt_ID* is numeric, it is preferable to convert it to character type as it a unique identifier. The *Gender*, *Indication, Age* and *Duration_ECMO* columns are removed as they are not relevant to the analysis. The response variables *ECMO_Survival* and *Hospital_Survival* are converted to factor type, as these will be treated as categories. The rest of the data is quantitative and is of the correct class type. A check for missing values is conducted and it is found that there are 725 missing values. On further inspection, it is noted that the missing values are present in the 29 biomedical markers. Albumin is found to have 129 values missing and is omitted from the dataset. The rows containing 50% or more missing values are omitted from the dataset. These are 12 rows belonging to the patients with ID's: 623, 667, 709, 884, 951, 1260, 1284, 1344, 1488, 1712, 1870 and 1891. The remaining 314 missing values are imputed with using MICE, under the assumption that the data is missing at random (Wilson, 2021) . The new data frame has 438 rows and 35 columns.

Based on the summary statistics as well as the histograms, it is observed that some of the features have very large ranges and display high skewness. Skewness can have an impact on distance-based clustering algorithms. From the measure of asymmetry, it is noted that eighteen variables show high skewness, four show moderate skewness and six variables are symmetrical. The eighteen variables displaying high skewness are adjusted by using logarithmic transformation. Multiple outliers are observed for each variable, few of them quite extreme. To mitigate the impact of these extreme values on scaling, the lower extreme values of *Ppeak, Pmean, log_Vt, log_SpO2, log_Leukocytes, log_Platelets* and *log_siIL2* are bounded at 1% and the upper extreme values of *PEEP, PaCo2, BE, log_pH, log_NAdose, log_TNFa* and *log_siIL2* are bounded at 99%. There is no evidence to remove or apply winsorization to the other variables. Further, some variables like *Fibrinogen, CRP, MAP* and *ATIII* overpower the others. The variables are scaled and the centred around the mean. In Figure 1, the stacked boxplots show that the variables are now comparable. The summary statistics confirm that the mean is zero and standard deviation is one. There are points, highlighted in red, that are notably distant from rest of the observations. These will be further inspected using clustering algorithms. A few of the clustering algorithms are based on the Euclidean distance metric and will be impacted by these points.

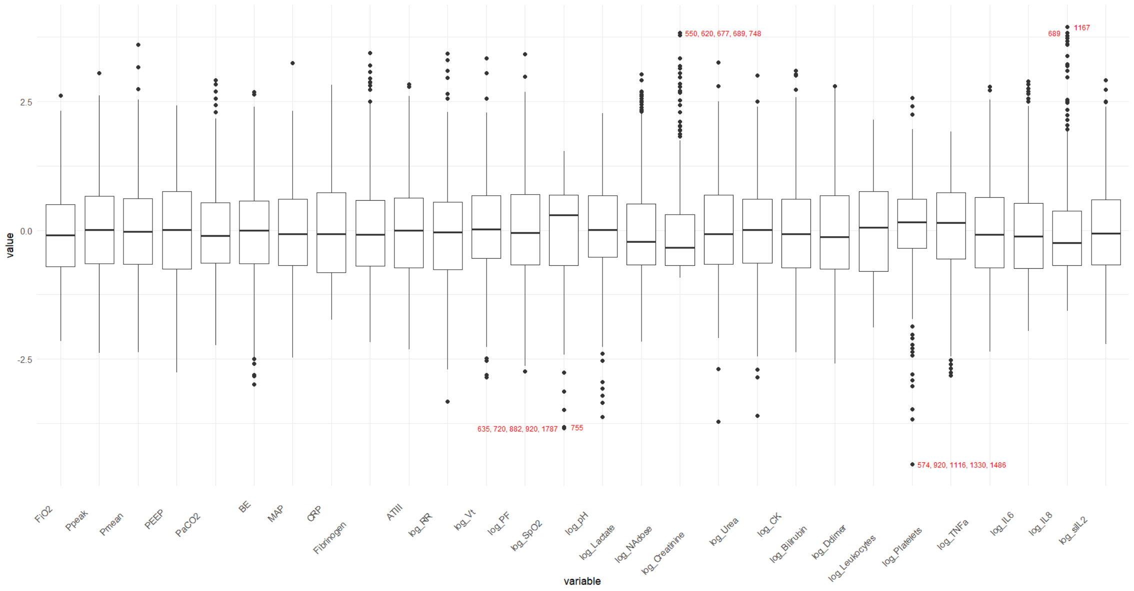


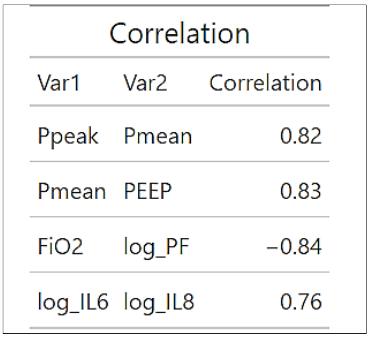
Figure 1: Boxplots of Biomedical Variables

EXPLORATORY DATA ANALYSIS

Correlation

Figure 2 shows that there are mostly weak to moderation correlations between the variables. However, high correlation exists between four pairs of variables as summarised in Table 1. In the variable descriptions, FiO_2 represents the inspired fraction of oxygen, while PF is the ratio of arterial oxygen partial pressure to the inspired oxygen fraction. Similarly, $Ppeak$ represent the Peak airway pressure and $Pmean$ represents Mean airway pressure. It would be sufficient to select one of these variables.

Table 1: Biomedical markers with High Correlation



Correlation

Var1	Var2	Correlation
Ppeak	Pmean	0.82
Pmean	PEEP	0.83
FiO2	log_PF	−0.84
log_IL6	log_IL8	0.76

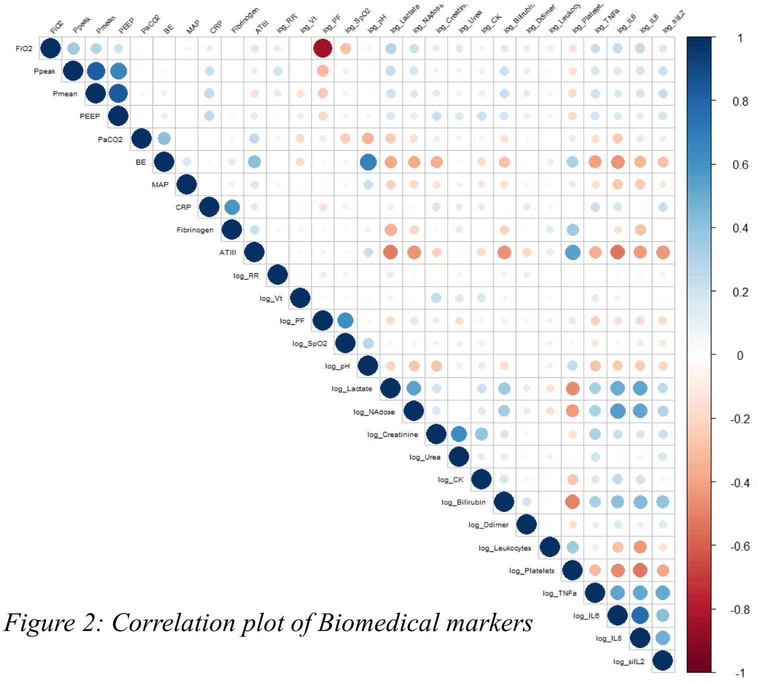


Figure 2: Correlation plot of Biomedical markers

Principal Component Analysis (PCA)

The dataset exhibits high correlations, suggesting that dimension reduction is possible. Using Kaiser's Method, nine components are retained, with PC1 to PC9 explaining 69% of the data's variability. The biplot in Figure 3 indicates three potential clusters, while the pair plot in Figure 4 with densities suggests evidence of two clusters and possible outliers. However, the clusters are not distinctly clear.

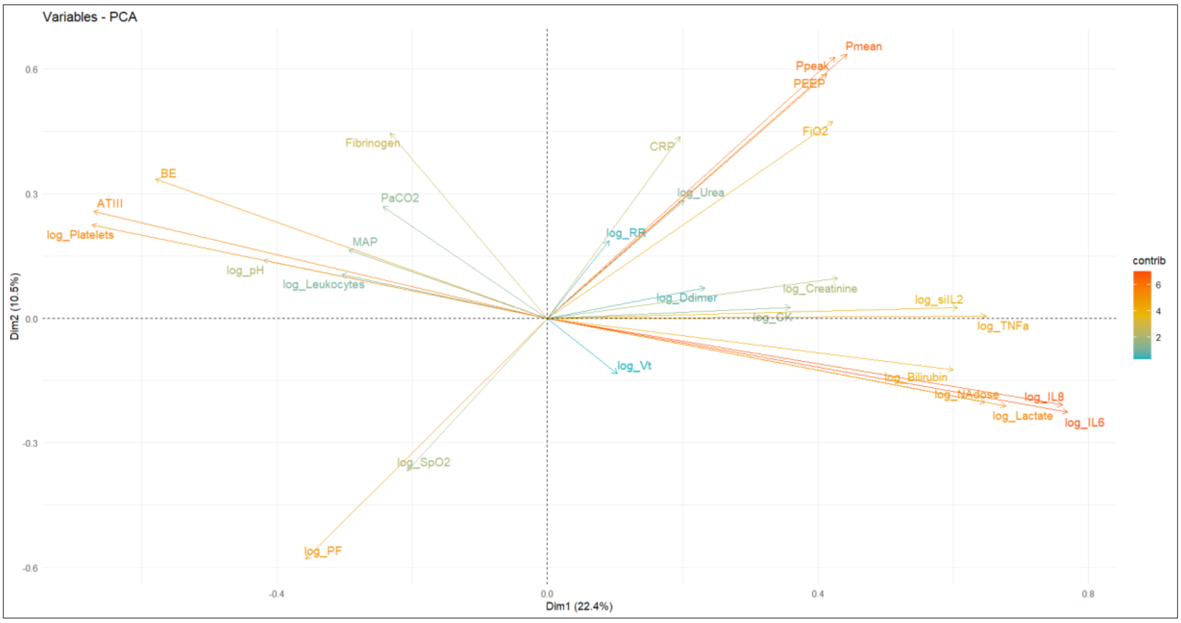


Figure 3: Biplot of the Biomedical markers after PCA

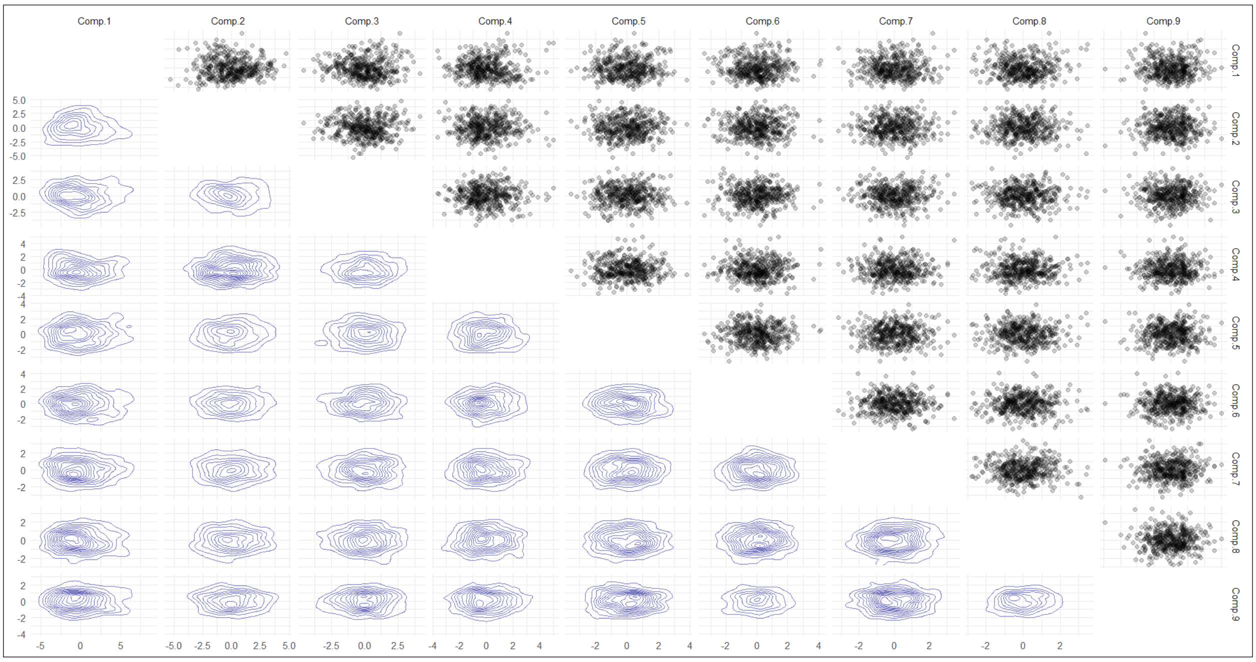


Figure 4: Pair plot with densities for Biomedical data after PCA

<u>Visual Assessment of Clustering Tendency</u>

Before performing clustering analysis, it is crucial to assess the data for clustering tendency. This assessment is carried out using a statistical method, the Hopkins statistic, as well as a visual method employing the VAT algorithm. (Kassambara, 2017)

1. <u>Hopkins statistic</u>: If the value of Hopkins statistic is close to 1 (far above 0.5), it concludes that the dataset is significantly clusterable (Kassambara, 2017).
2. <u>VAT (Visual Assessment of cluster Tendency):</u> The VAT detects the clustering tendency in a visual form by counting the number of square shaped dark (or coloured) blocks along the diagonal in a VAT image. (Kassambara, 2017).

Here, $H = 0.6548$, which implies that data is moderately clusterable. The ordered dissimilarity image in Figure 5 and the heatmap in Figure 6 suggest two or three potential clusters, based on observed patterns and rectangular formations, though further analysis is needed for confirmation.

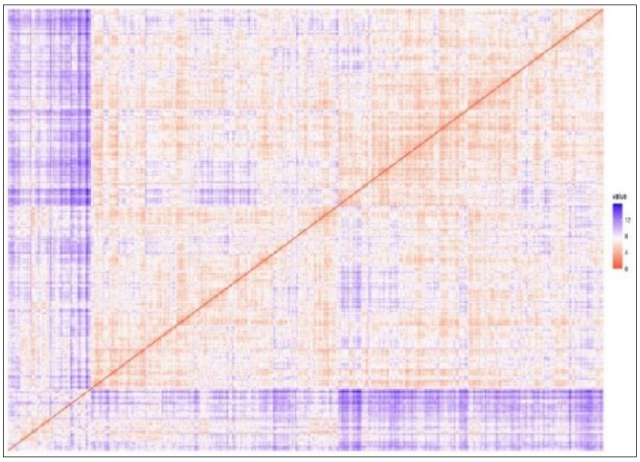


Figure 5: Ordered Dissimilarity Image (VAT)

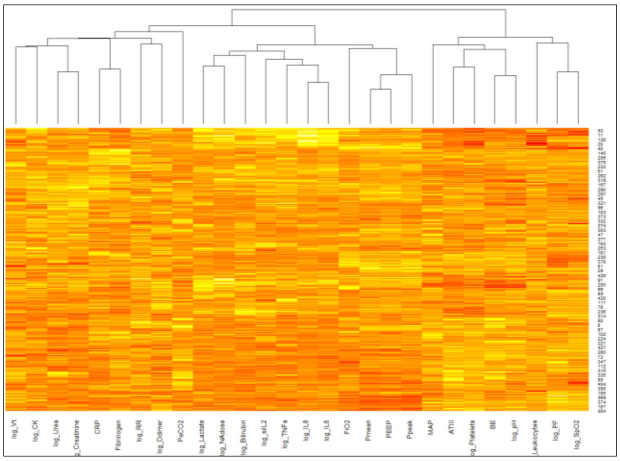


Figure 6: Heatmap of the Biomedical markers

findCorrelation function is used to determine which variables need to be removed to reduce pairwise correlation between the variables. In this function, the absolute values of pair-wise correlations are considered. If two variables have a high correlation, the function looks at the mean absolute correlation of each variable and removes the variable with the largest mean absolute correlation. (Li & Kuhn, n.d.) Using this, *log_IL6*,

$Pmean$, and FiO_2 are removed. In this new dataset, there are pairs of variables with only weak to moderate correlations. A PCA is conducted and the results are similar as before - the biplot indicates three potential clusters, while the pair plot with densities suggests evidence of two clusters and possible outliers. There is a mild improvement in the Hopkins statistic from 0.6548 to 0.6559.

CLUSTERING ANALYSIS

To determine optimal number of clusters, the following methods are used :

1. Elbow Plot : This is a line plot with a number of clusters on the horizontal axis and the total within sum of squares value for the model on the vertical axis. The number of k is usually chosen to be where there is a bend in the lines, hence called an "elbow plot". (Yang, Partitioning Cluster Analysis, Week 9, Lecture handout, 2023-2024).
2. Average Silhouette Plot : The silhouette is the measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). (Yang, Hierarchical Cluster Analysis, Week 8, Lecture handout, 2023-24)

$$silhouette\ width\ s_i = \frac{(b_i - a_i)}{\max\{a_i, b_i\}}, -1 \leq s_i \leq 1$$

where,

- a_i : is the average distance between the i^{th} observation and all other observations in the cluster containing the i^{th} observation
- b_i : is the minimum average distance of the i^{th} observation to the observation in other clusters

If s_i is close to 1, then the observation is appropriately clustered. A negative silhouette width means that the observation would be better suited to the neighbouring cluster. (Yang, Hierarchical Cluster Analysis, Week 8, Lecture handout, 2023-24). Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman & Rousseeuw, 1990).

3. Gap statistic: The gap statistic evaluates the total within-cluster variation for different values of k against expected values under a null reference distribution. The optimal number of clusters is the value that maximizes the gap statistic, indicating that the clustering structure is distinct from a random uniform distribution. (Kassambara, 2017)
4. NbClust: The *NbClust* package in R evaluates about 30 methods to determine the optimal number of clusters. It provides various indices and suggests the best clustering scheme by analyzing all combinations of cluster numbers, distance measures, and methods, all in one function call (Charrad, Ghazzali, Boiteau, & Niknafs, 2015).

The Elbow and average silhouette methods have the disadvantage of measuring only a global clustering characteristic and the elbow can be subjective. A more sophisticated approach is to use the gap statistic and *NbClust*, which formalizes the heuristic and statistically estimate the optimal number of clusters (Kassambara, 2017). It's advisable to compare optimal number of clusters using different methods to ensure consistency.

PARTITIONING CLUSTER METHODS

Partitioning cluster methods start by randomly partitioning observations into a set of disjoint clusters and then relocate observations by moving them from one cluster to another. Two widely uses partitioning clustering algorithms are k-means clustering and k-medoids clustering or Partitioning around Medoids (Yang, Partitioning Cluster Analysis, Week 9, Lecture handout, 2023-2024).

i. *k*-means clustering

This is a Hard clustering method and partitions the data set into k distinct, non-overlapping clusters. This method uses means as centroids. The assumptions for this method are that the clusters are spherical when using Euclidean distance metric, the cluster are of similar size (implying area covered and not the cardinality) and the number of clusters is fixed. Lloyd's algorithm (Yang, Partitioning Cluster Analysis, Week 9, Lecture handout, 2023-2024) is applied and the particular cluster is achieved by minimizing total within-cluster sum of squares given by $W(C)$ defined as:

$$W(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} d_E(x_i, \bar{x}_{C_k})^2 = \sum_{k=1}^{K} \frac{1}{2|C_k|} \sum_{x_i \in C_k} \sum_{x_j \in C_k} d_E(x_i, x_j)^2$$

where,

- x_i is an observation belonging to the cluster C_k;
- $\bar{x}_{C_k}$ is the average of all the points belonging to the cluster C_k known as cluster centroid;
- $|C_k|$ is the number of observations in the k^{th} cluster ;
- $d_E(x_i, x_j)^2$ is the squared Euclidean distance between p-dimensional continuous variables x_i and x_j

To perform *k*-means algorithm, the value of *k* needs to be specified beforehand. The optimal number of clusters, using the methods defined in the previous section, are two, three or six. Based on Average Silhouette Width, *k*-means with two clusters in Figure 7 is seen to be the most appropriate. Since this method is sensitive to the initial random selection of cluster centres, different seeds and initial configurations are used. To assess clustering similarity, Adjusted Rand Index is computed between different runs. The results are near one, indicating robust clustering.

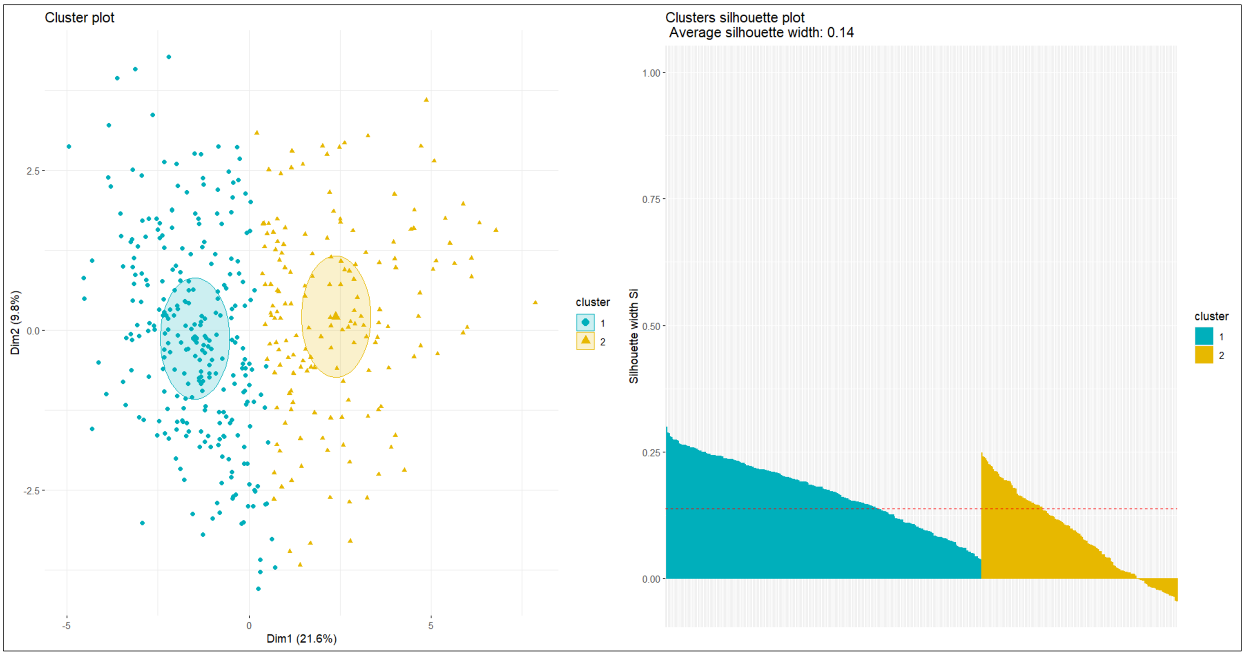


Figure 7: Cluster plot and Average silhouette width of k-means with two clusters

k-means clustering is a very simple and fast algorithm and can efficiently deal with large data sets. However, it has some drawbacks; it requires prior knowledge of the data including the pre-selection of appropriate number of clusters. Additionally, it is severely affected by to outliers and different data orderings can result in varying outcomes (Yang, Partitioning Cluster Analysis, Week 9, Lecture handout, 2023-2024).

ii. Partitioning around Medoids (PAM)

A robust alternative to *k*-means is PAM. The algorithm is less sensitive to noise and outliers. This algorithm accepts a dissimilarity matrix and minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances defined as

$$argmin \sum_{x_j \in C_i} d(x_i, x_j)$$

where, C_i is cluster containing point i and $d(x_i, x_j)$ is the dissimilarity between x_i *and* x_j (Yang, Partitioning Cluster Analysis, Week 9, Lecture handout, 2023-2024). The Elbow method, Silhouette method and Gap Statistic suggested two optimal clusters shown in Figure 8.

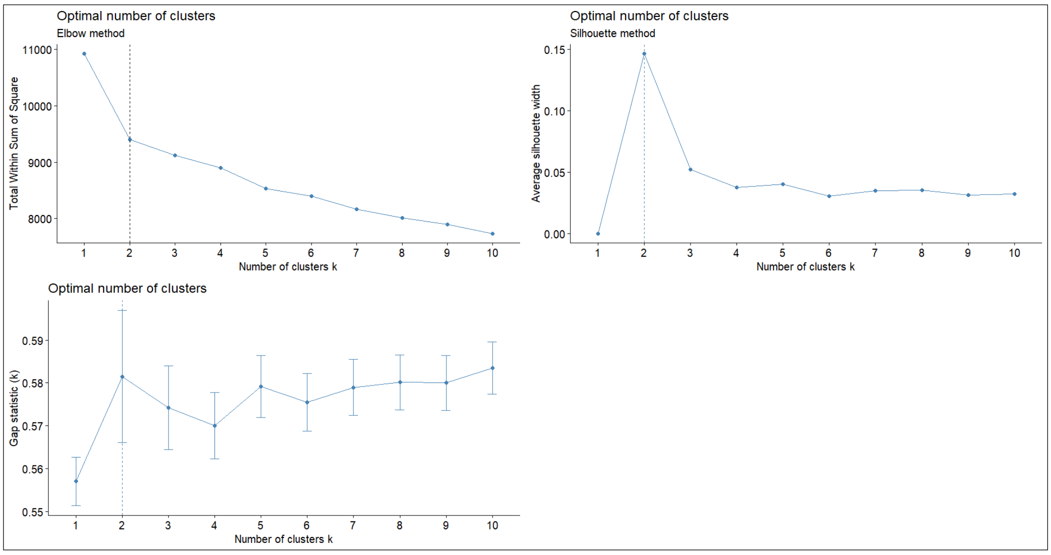


Figure 8: Optimal number of clusters for PAM

In *k*-medoids clustering, each cluster is represented by cluster medoids which corresponds to the most centrally located point in the cluster (one per cluster). In Figure 9
medoid for Cluster 1 : Patient 430 with patient ID 1939
medoid for Cluster 2 : Patient 164 with patient ID 974

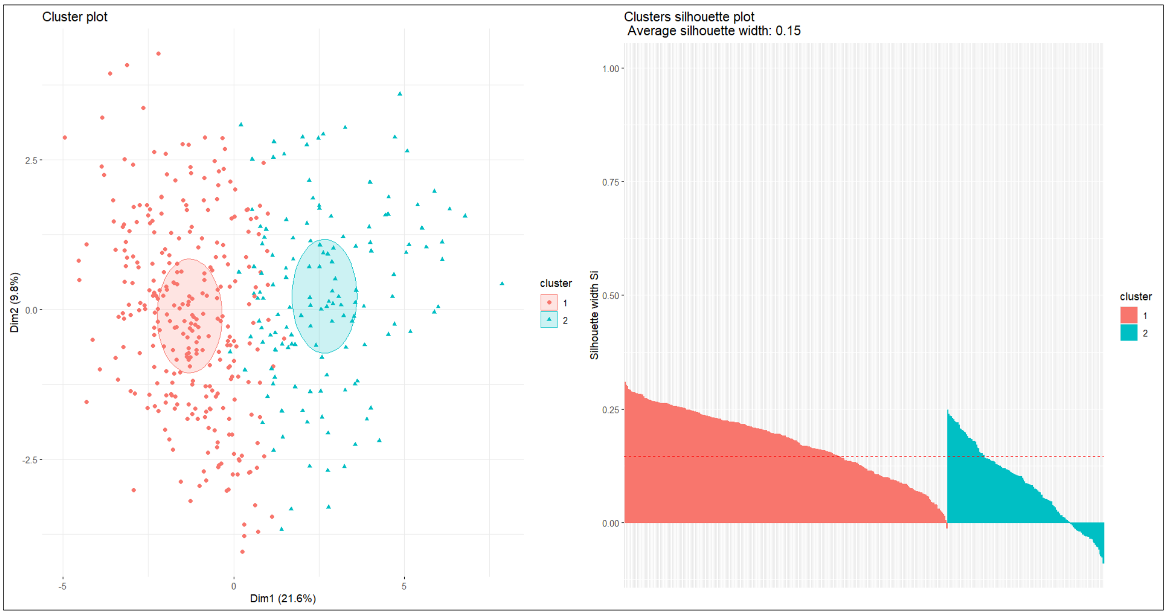


Figure 9: Cluster plot and Average silhouette width of PAM with two clusters

The summary statistics as well the cluster allocation of PAM is quite similar to that to *k*-means clustering as computed in Table 5 .

iii. Hierarchical clustering (Agglomerative)

Hierarchical clustering is an alternative approach to partitioning clustering for grouping observations based on their similarity. The algorithm of agglomerative clustering works in a “bottom-up” manner. The dissimilarity between observations is measured by Euclidean distance and the dissimilarity between clusters is measured by linkage criteria - Complete, Ward, Average, Single. This algorithm does not require to pre-specify the number of clusters (Yang, Hierarchical Cluster Analysis, Week 8, Lecture handout, 2023-24). However, the Elbow Method, Silhouette method, Gap statistic and *NbClust* methods are still applied to estimate the appropriate number of clusters which suggested two clusters. Interestingly, the Gap statistic suggested only one cluster which corresponds to the result from principal component analysis Figure 4. The optimal number of clusters on the basis of linkages: two clusters for Complete Linkage, three clusters for Ward

Linkage, four clusters for Average Linkage and two clusters for Single Linkage. The cophenetic correlation was computed and there is a moderate correlation between average and single linkage (Kassambara, 2017). Instead of plotting traditional dendrograms, a phylogenetic-like tree is used to plot for easier visual comprehension of the memberships (Kassambara, 2017).

Complete Linkage: This linkage $L(A, B)$ between two clusters A and B is defined as the maximum of the distances between all pairs of points with one point from cluster A and one point from cluster B (Yang, Hierarchical Cluster Analysis, Week 8, Lecture handout, 2023-24). This is shown in Figure 10.

$$L(A,B) = \max\{d(x_a, x_b) : x_a \in A, x_b \in B\} \text{ where } d(x_a, x_b) \text{ is the Euclidean distance between } x_a \text{ and } x_b$$

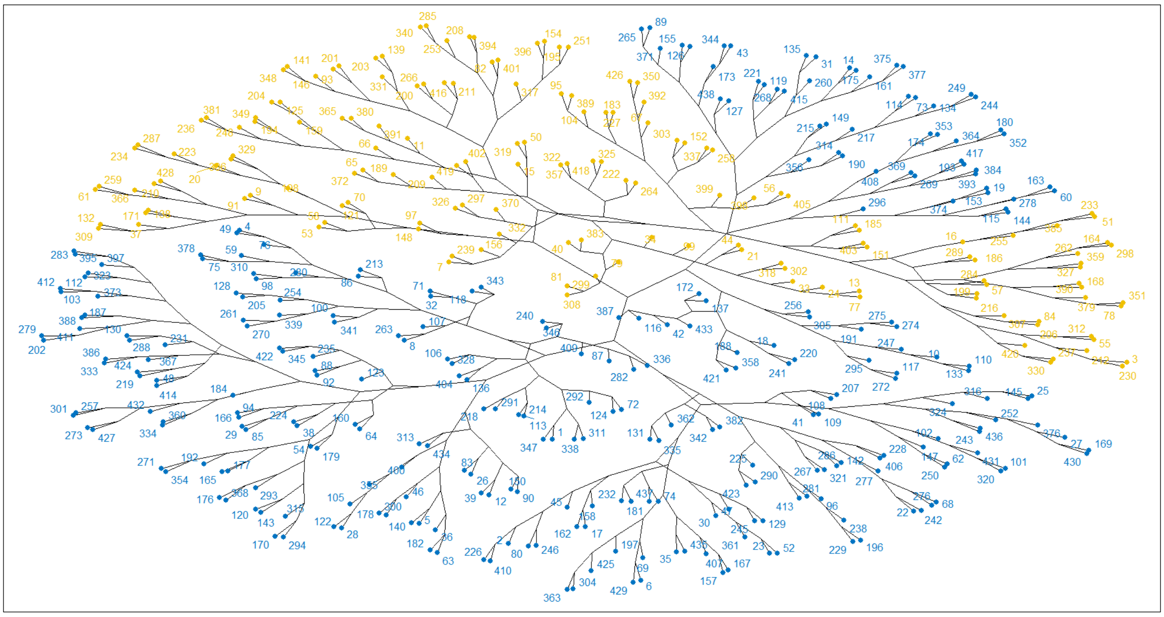


Figure 10: Complete linkage with two clusters

Ward Linkage: This linkage $L(A, B)$ between two clusters A and B is defined as the difference in the sum of squares for the combined cluster resulting from merging A and B, and the sum of the squares for the two clusters separately (Yang, Hierarchical Cluster Analysis, Week 8, Lecture handout, 2023-24). This is shown in Figure 11.

$$L(A,B) = SS(A,B) - \big(SS(A) + SS(B)\big) \text{where } SS(A) = \sum_{a:\, x_a \in A} (x_a - \bar{x}_A)^2$$

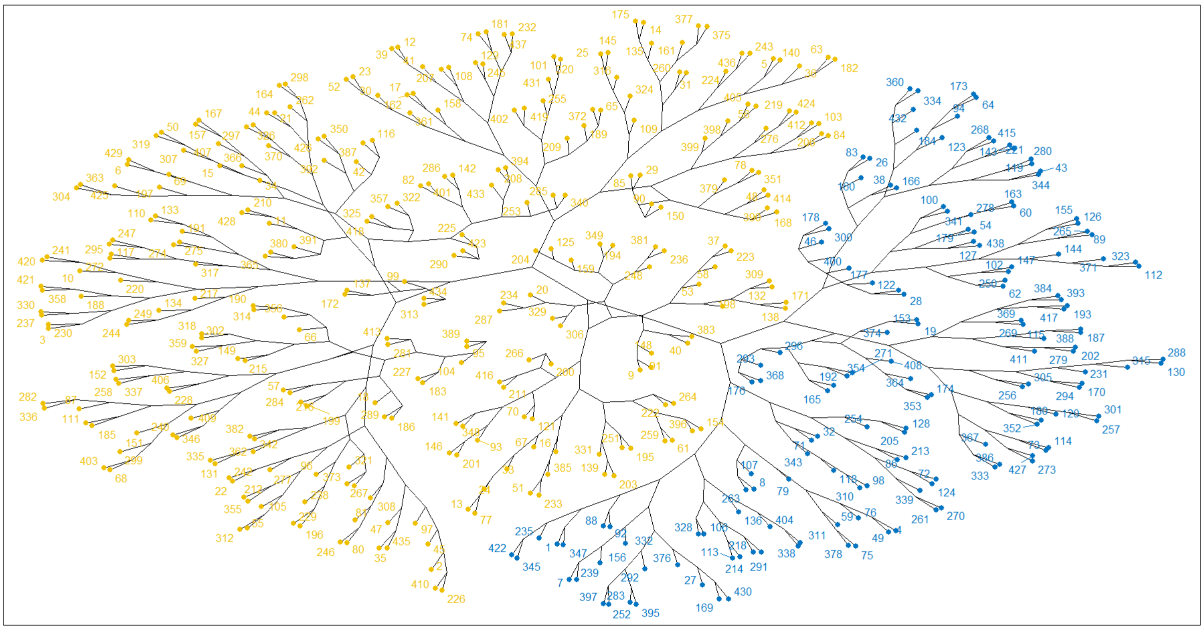


Figure 11: Ward linkage with two clusters

Average Linkage: This linkage $L(A,B)$ between two clusters A and B is defined as the average of all the distances between all pairs of observations with one point from cluster A and one point from cluster B (Yang, Hierarchical Cluster Analysis, Week 8, Lecture handout, 2023-24).

$L(A,B) = \frac{1}{|A||B|}\sum_{\{a,b\}:\, x_a \in A,\, x_b \in B} d(x_a\, , x_b)$ where $d(x_a, x_b)$ is the Euclidean distance between x_a and x_b

For $k = 2$, the second cluster includes patients with ID's 635, 755, 920, and 1787. For $k = 3$, the second and third clusters contain patients with ID's 620, 635, 755, 920, and 1787. For $k = 4$, the third and fourth clusters had the same patients.

Single Linkage: This linkage $L(A,B)$ between two clusters A and B is defined as the minimum of the distances between all pairs of points with one point from cluster A and one point from cluster B (Yang, Hierarchical Cluster Analysis, Week 8, Lecture handout, 2023-24).

$L(A,B) = \min \{d(x_a, x_b): x_a \in A, x_b \in B\}$ where $d(x_a, x_b)$ is the Euclidean distance between x_a and x_b

In the tree for two clusters, the second cluster only comprised of a single observation with patient ID 920. For $k = 3$, the second and third cluster had patients 920 and 677 respectively .

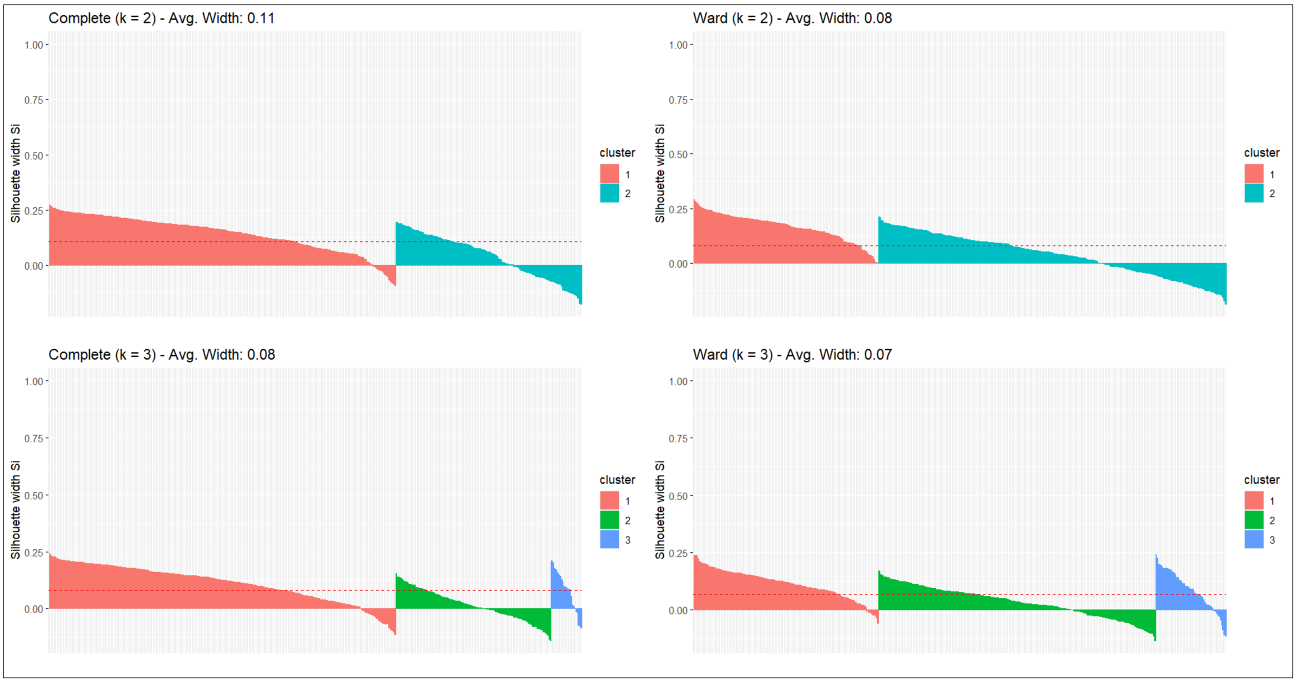


Figure 12: Average Silhouette Width of Complete and Ward Linkage with Two and Three Clusters

The average and single linkages provide an insight in recognising the outliers that are spotted in the boxplot Figure 1. However, they are unable to create clusters of reasonable sizes. In comparison, the complete and ward linkage are more stable. On the basis of average silhouette width Figure 12, hierarchical clustering with complete and ward linkage with two clusters seem to be the most appropriate. The Fowlkes-Mallows index is computed, which measures the similarity between two clustering under the null hypothesis of no relation (Galili, 2023). The null hypothesis is rejected and it is concluded that complete and ward linkage clustering have moderate level of agreement. The Bk plot helps in identifying the similarity between two dendrograms in different levels of k (number of clusters) which further agrees with the result from the Fowlkes-Mallows Index (Galili, 2023). The Dunn Index is computed for comparison and the value is higher for complete linkage. In conclusion, hierarchical clustering with complete linkage with two clusters is most appropriate since it has a higher value for average silhouette width and Dunn Index which implies that the clusters are much compact and well-separated in comparison to ward linkage.

Hierarchical clustering offers an intuitive dendrogram that visually shows connections between observations and doesn't require specifying the number of clusters in advance. It effectively identifies clusters of various shapes and sizes, providing more flexibility than partitioning methods. It is an easy to implement and interpret.

However, it is a computationally expensive technique, sensitive to outliers and noise, and determining the optimal number of clusters can be subjective and challenging (Kuchciak, 2024).

iv. Model-based clustering

This is a soft clustering method that takes a probabilistic approach to clustering. The model-based algorithm of clustering assumes that the observations come from a finite mixture of normal distributions and each component of that mixture corresponds to a different cluster (Fraley & Raftery, 2002) . Under this assumption, a Gaussian mixture model is fit to the data. The Gaussian mixture model is defined as:

$$x \sim \sum_{g=1}^{G} \pi_g N(x \mid \mu_g, \Sigma_g)\,, where\ 0 < \pi_g \leq 1\ \forall\ g, \sum_{g=1}^{G} \pi_g = 1$$

The expectation-maximisation (EM) algorithm is used to estimate the parameters and discover the different cluster groupings (i.e. the allocations of the *n* observations in the *G* clusters) (Yang, Hierarchical Cluster Analysis, Week 8, Lecture handout, 2023-24). The model suggests as three as optimal clusters. In the top three models based on Bayesian Information criterion (BIC), the algorithm proposed two and four clusters.

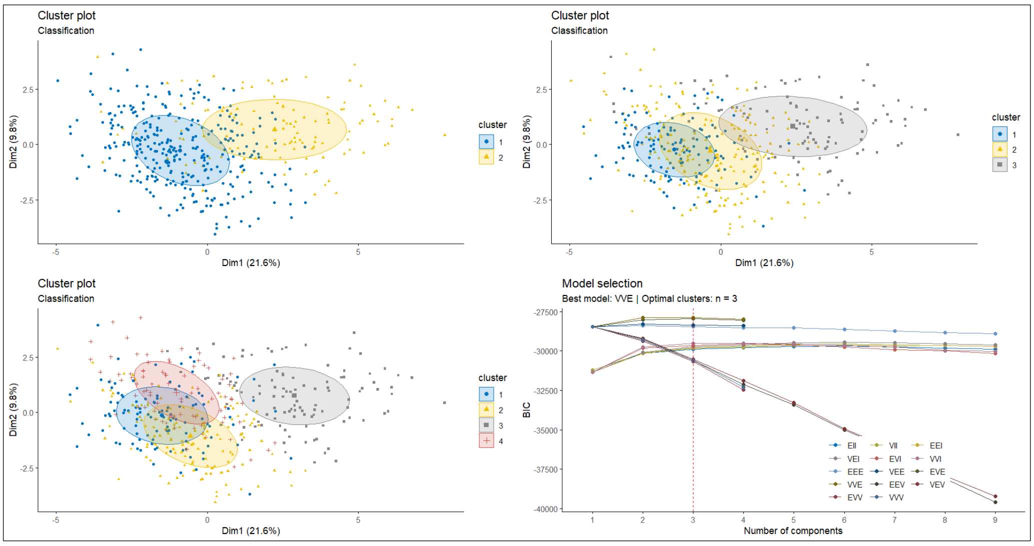


Figure 13: Cluster plot and BIC plot of model-based Clustering of two, three and four clusters

However, this algorithm resulted in low average silhouette widths for three and four clusters as depicted in Figure 14. In addition, the clusters have negative silhouette width implying that the observations are better suited to the neighbouring cluster. Hence, the model with two clusters seems to be the most appropriate.

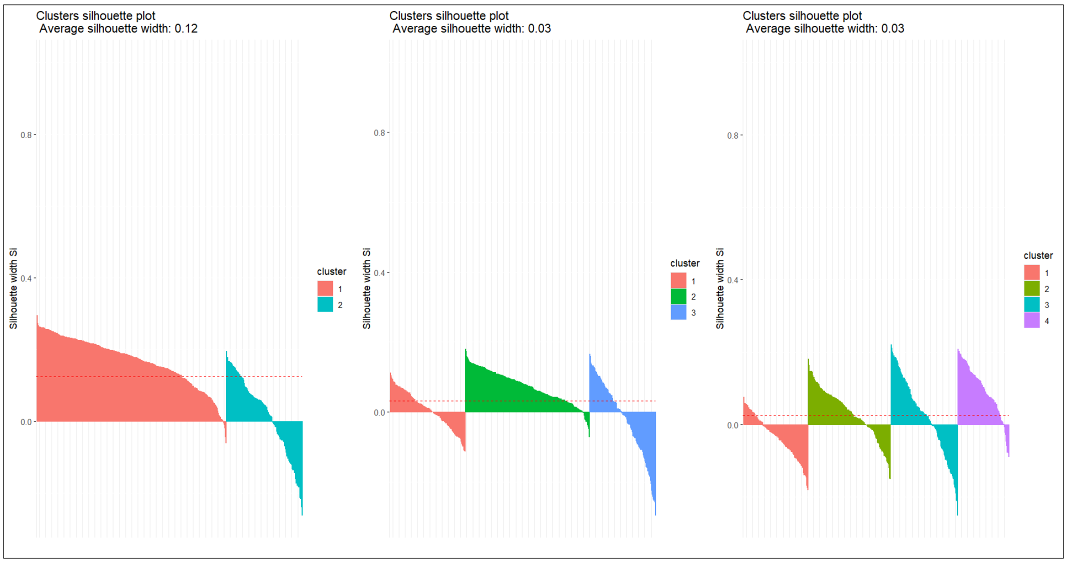


Figure 14: Average Silhouette width for model-based clustering for two, three and four clusters

v. Density-Based Clustering (DBSCAN)

DBSCAN is used to identify clusters of any shape in a data set containing noise and outliers. It requires no assumptions for data. Two parameters are required: epsilon (*eps*) and minimum points (*minPts*). The original DBSCAN paper (Ester, Kriegel, Sander, & Xu, 1996) suggests

- $minPts \geq d + 1,$ where *d* is the data dimensionality
- *eps*: The value for epsilon can then be chosen by using a k-distance graph, plotting the distance to the $k = minPts$ nearest neighbor. The value for epsilon is chosen where the plot shows a good bend (Hahsler, 2019) as shown in Figure 15.

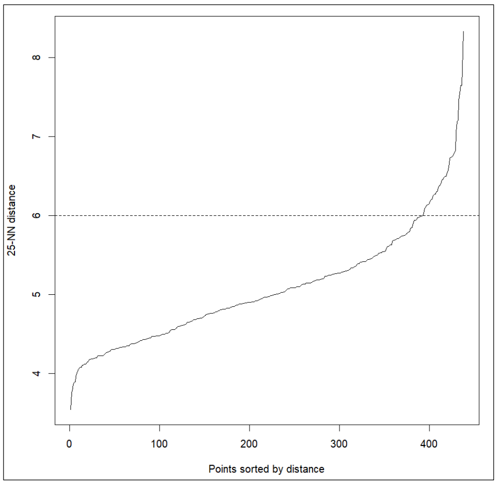


Figure 15: kNN Distance plot

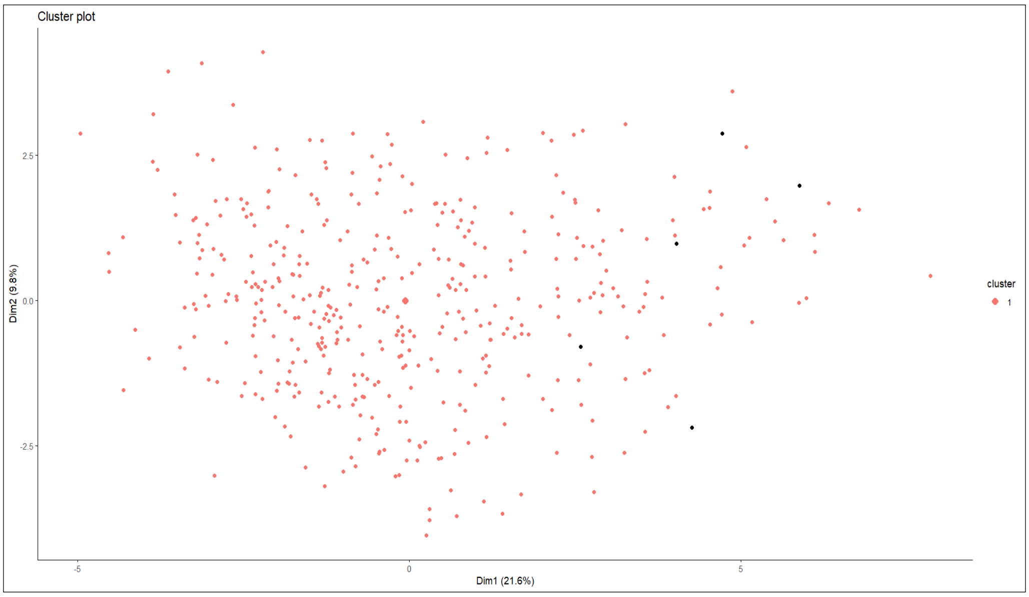


Figure 16: DBSCAN Cluster plot

After running the algorithm, it recognizes five noise points with patient Id's 574, 620, 677, 920 and 1116 as depicted in Figure 16. This result corresponds to the patients 620, 677 and 920 found in Hierarchical clustering using average and single linkage. Additionally, these points correspond to extreme points Figure 1.When DBSCAN is run on a reduced dataset, removing all the above-mentioned patients, it reveals that the data has only one cluster. The advantages of this method are that firstly it can find any shape of clusters. Secondly it doesn't need the number of clusters to be specified beforehand. Lastly, it can identify outliers and noise points. However, it is quite sensitive to the hyperparameters *minPts* and *eps*. This technique faces challenges with clusters of varying densities as can be observed in this data set (Kassambara, 2017).

CHAPTER 3 : METHODOLOGY

COMPARISON OF CLUSTERING ALGORITHMS

A well-performing clustering algorithm creates stable, statistically meaningful clusters that maintain strong internal quality. The package *clValid* is used to compute Internal and Stability measures (Brock, Pihur, Datta, & Datta, 2008).

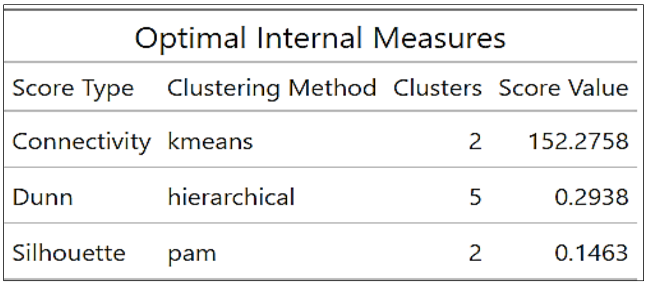
Internal measures (Brock, Pihur, Datta, & Datta, 2008)

1. **Connectivity** is an internal validation method that assesses how well neighboring observations are grouped within the same cluster. It should be minimized.
2. **Dunn Index** is another measure to evaluate the compactness and separation of the clusters. The Dunn Index is calculated as the ratio of the smallest distance between observations in different clusters to the largest distance within the same cluster.
3. **Silhouette** coefficient is the average silhouette width. This value should be maximized.

Stability Measures

The stability metrics are a particular type of internal metrics that assess the stability of a clustering outcome by comparing it with the clusters generated when one column is removed at a time. These are average proportion of non-overlap (**APN**) , average distance (**AD**), average distance between means (**ADM**) and figure of merit (**FOM**). These measures should be maximised (Brock, Pihur, Datta, & Datta, 2008). The stability measure calculation excluded model-based clustering due to its excessively long run time.

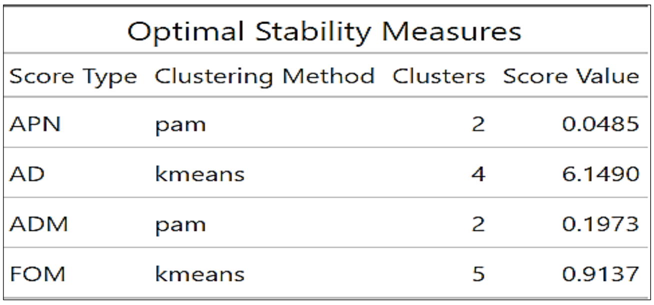
Table 2: Optimal Internal Measures for clustering methods



Optimal Internal Measures

Score Type	Clustering Method	Clusters	Score Value
Connectivity	kmeans	2	152.2758
Dunn	hierarchical	5	0.2938
Silhouette	pam	2	0.1463

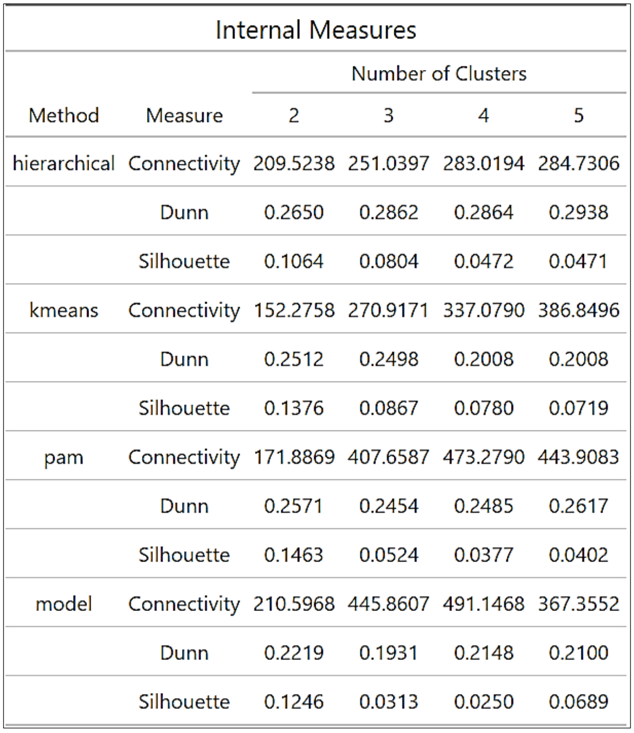
Table 4: Optimal Stability Measures for Clustering



Optimal Stability Measures

Score Type	Clustering Method	Clusters	Score Value
APN	pam	2	0.0485
AD	kmeans	4	6.1490
ADM	pam	2	0.1973
FOM	kmeans	5	0.9137

Table 3: Internal Measures for clustering methods



Internal Measures

		Number of Clusters			
Method	Measure	2	3	4	5
hierarchical	Connectivity	209.5238	251.0397	283.0194	284.7306
	Dunn	0.2650	0.2862	0.2864	0.2938
	Silhouette	0.1064	0.0804	0.0472	0.0471
kmeans	Connectivity	152.2758	270.9171	337.0790	386.8496
	Dunn	0.2512	0.2498	0.2008	0.2008
	Silhouette	0.1376	0.0867	0.0780	0.0719
pam	Connectivity	171.8869	407.6587	473.2790	443.9083
	Dunn	0.2571	0.2454	0.2485	0.2617
	Silhouette	0.1463	0.0524	0.0377	0.0402
model	Connectivity	210.5968	445.8607	491.1468	367.3552
	Dunn	0.2219	0.1931	0.2148	0.2100
	Silhouette	0.1246	0.0313	0.0250	0.0689

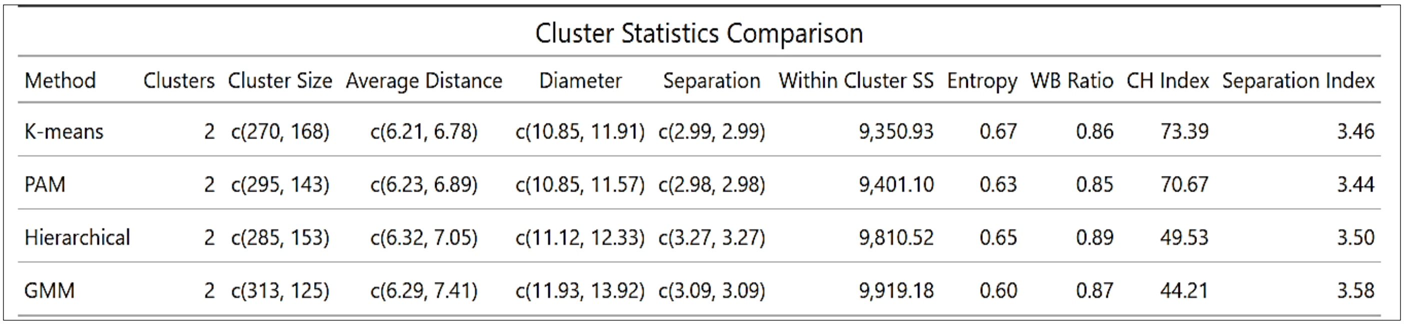
The Internal measures in Table 3 indicate that while PAM achieves the highest average silhouette width, *k*-means follows closely. Hierarchical clustering (complete linkage), shows higher Dunn scores across all cluster sizes, indicating better compactness and separation. However, its high connectivity and low silhouette width suggest poorer overall clustering quality. Thus, it appears *k*-means and PAM generally outperform other methods in most validation measures. Although stability measures in Table 4 suggest using more clusters for both *k*-means and PAM, the declining Dunn and increasing connectivity scores imply that clustering quality deteriorates as more clusters are added. Across all methods, two clusters appear optimal, given their consistent performance across measures. Model-based clustering, however, shows poor performance across all validation metrics, with high connectivity and low silhouette and Dunn scores.

Cluster Statistics

The *cluster.stats* function is used to calculate additional distance-based statistics, which are utilized for cluster validation and comparison between different clustering algorithms (Hennig, 2024). These metrics includes

1. **Average Distance**: cluster wise within cluster average distances
2. **Diameter**: cluster diameters (maximum within cluster distances)
3. **Separation**: cluster wise minimum distances of a point in the cluster to a point of another cluster
4. **Within Cluster SS**: within clusters sum of squares
5. **Entropy**: entropy of the distribution of cluster memberships
6. **WB Ratio** $= \frac{average\ distance\ within\ clusters}{average\ distance\ between\ clusters}$
7. **CH index** (Calinski and Harabasz index): The CH Index (also known as **Variance ratio criterion**) is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). CH index $= \frac{a \times Separation}{b \times Cohesion}$, where *a* and *b* are weights.
8. **Separation Index** : A separation index is calculated based on the distances from each point to the closest point not in the same cluster. The separation index represents the mean of the smallest proportion of these distances

Table 5: Cluster Statistics of Clustering methods



Cluster Statistics Comparison

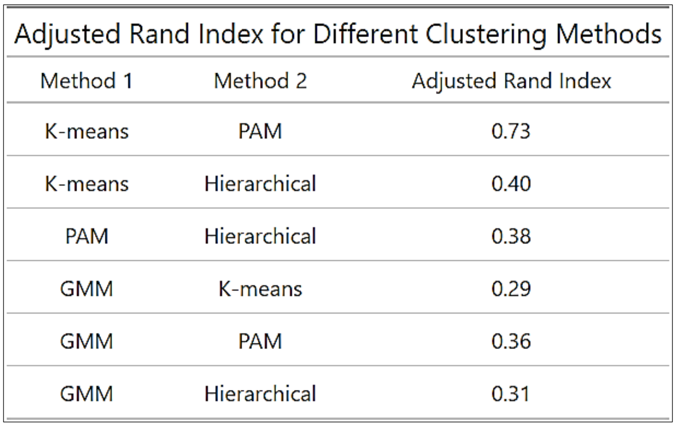
Method	Clusters	Cluster Size	Average Distance	Diameter	Separation	Within Cluster SS	Entropy	WB Ratio	CH Index	Separation Index
K-means	2	c(270, 168)	c(6.21, 6.78)	c(10.85, 11.91)	c(2.99, 2.99)	9,350.93	0.67	0.86	73.39	3.46
PAM	2	c(295, 143)	c(6.23, 6.89)	c(10.85, 11.57)	c(2.98, 2.98)	9,401.10	0.63	0.85	70.67	3.44
Hierarchical	2	c(285, 153)	c(6.32, 7.05)	c(11.12, 12.33)	c(3.27, 3.27)	9,810.52	0.65	0.89	49.53	3.50
GMM	2	c(313, 125)	c(6.29, 7.41)	c(11.93, 13.92)	c(3.09, 3.09)	9,919.18	0.60	0.87	44.21	3.58

A comparison is made between the two clusters for all the clustering algorithms using additional distance-based statistics. From Table 5, *k*-means has with the highest CH Index and lowest Within Cluster SS, making it the best at forming compact, well-separated clusters. *k*-means and PAM have almost identical statistics across all measures. Hierarchical Clustering has the highest Separation Index and Within Cluster SS, which implies it that the clusters are well-separated clusters but have more internal variability. Model-based clustering does well with the lowest entropy and highest Separation Index, though its clusters are larger and less compact.

Adjusted Rand Index

In the absence of ground truth, the Adjusted Rand Index is used to assess the similarity of cluster label allocations across different clustering algorithms. From the Table 6, it can be observed that *k*-means and PAM have the highest agreement, indicating similar groupings. Both *k*-means and PAM show moderate agreement with Hierarchical clustering. Model-based clustering shows the least agreement, producing different groupings in comparison to others.

Table 6: Adjusted Rand Index of Clustering methods



Adjusted Rand Index for Different Clustering Methods

Method 1	Method 2	Adjusted Rand Index
K-means	PAM	0.73
K-means	Hierarchical	0.40
PAM	Hierarchical	0.38
GMM	K-means	0.29
GMM	PAM	0.36
GMM	Hierarchical	0.31

CORRESPONDENCE ANALYSIS

Cluster Profiles

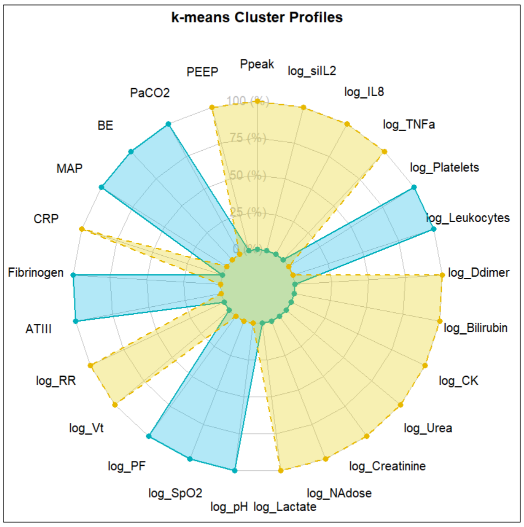


Figure 17: k-means Cluster Profiles

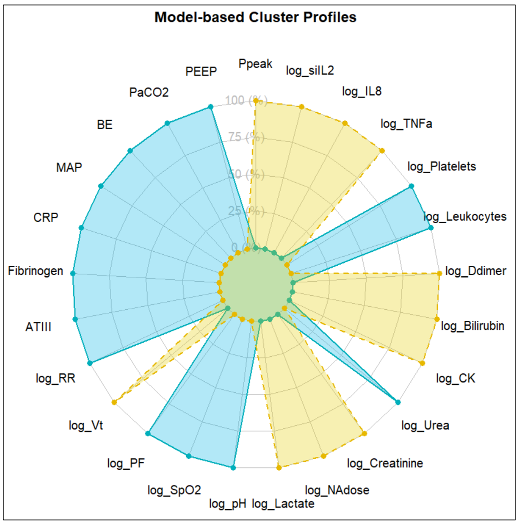


Figure 18: Model-based Cluster Profiles

Radar charts are used to visualize the cluster profiles for each clustering algorithm. For each algorithm, the mean values of the biomedical markers are calculated for each cluster. In the case of *k*-means in Figure 17, Cluster 1 is represented by the color green and Cluster 2 by yellow. Cluster 1 shows high mean values for ten variables namely *log_pH, log_SpO2, log_PF, ATIII, Fibrinogen, MAP, BE, PaCO2, log_Platelets* and *log_Leukocytes*. Cluster 2 shows high mean values for fifteen variables namely *PEEP, Ppeak, log_silL2, log_IL8, log_TNFa, log_Ddimer, log_Bilirubin, log_CK, log_Urea, log_Creatinine, log_NAdose, log_Lactate, log_Vt, log_RR* and *CRP*. Similarly, in model-based clustering Figure 18, Cluster 1 (green) and Cluster 2 (yellow) each show high mean values for distinct sets of biomedical markers.

Bar charts

The Figure 19 shows the distribution of ECMO survival and Hospital Survival rates across the clusters identified by *k*-means, PAM, Hierarchical Clustering using Complete Linkage and model-based clustering. In Figure 19 it can be observed that Cluster 1 performs better in terms of both ECMO and Hospital survival outcomes compared to Cluster 2.

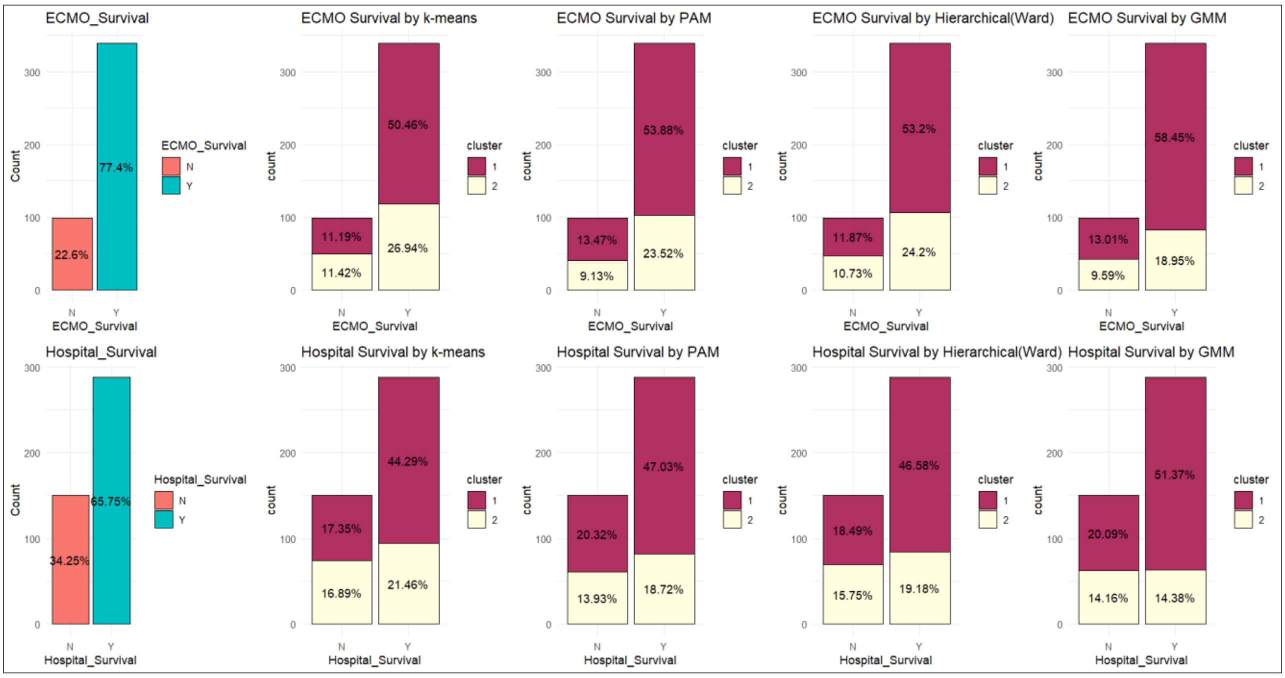


Figure 19: Bar charts of Survival outcomes

The higher proportions of survivors in Cluster 1 across both categories, in both survival outcomes, suggest that this cluster may represent a subgroup of patients with more favourable biomedical markers that lead to better survival rates. Conversely, Cluster 2, with lower survival proportions, may include patients with less favourable biomedical markers.

Chi-Square Tests and Cramér's V

These tests aim to measure the association between two categorical variables, the response variable ECMO Survival and Hospital Survival with respect to the cluster labels attained from applying various clustering mechanisms. The chi-square test is used to check of there is any relationship between the variables under the null hypothesis (Soetewey, 2020),

$H0$: The response variables (*ECMO_Survival* or *Hospital_Survival*) have no relationship with the clusters

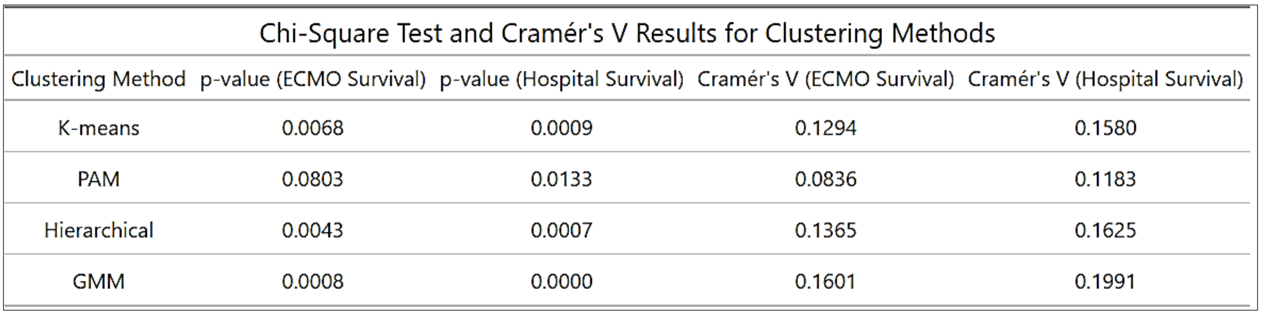
$H1$: There is a relationship between the response variables (*ECMO_Survival* or *Hospital_Survival*) and the clusters

The Cramer's V test further quantifies the relationship between the variables. Its value varies from 0 to1, with one indicting a perfect association (Mangiafico, 2016). The Table 7 shows:

Chi-Square Tests: For most of the clustering methods, the *p-values* < 0.05 indicate that the null hypothesis can be rejected. However, the relationship between PAM and ECMO Survival is not significant as suggested by the *p-value*. Therefore, clustering methods show significant relationship with the response variables ECMO Survival and Hospital Survival, indicating that the distribution of survival outcomes varies significantly across clusters.

Cramér's V: All values for Cramér's V are in the range typically interpreted as showing weak associations (0.08 to 0.20). This suggests that while there is a significant association between the groupings and the survival outcomes, the strength of this association is relatively weak.

Table 7: Chi-Square (p-value) and Cramer's V Test for clustering methods



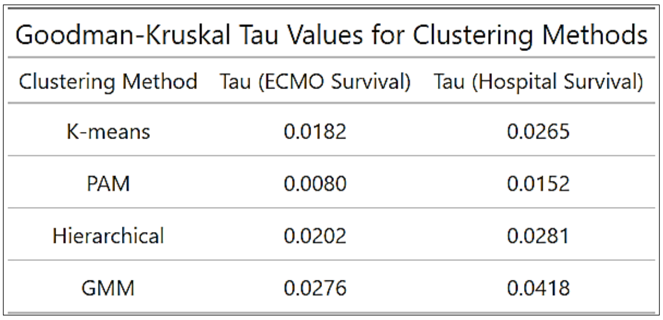
Chi-Square Test and Cramér's V Results for Clustering Methods

Clustering Method	p-value (ECMO Survival)	p-value (Hospital Survival)	Cramér's V (ECMO Survival)	Cramér's V (Hospital Survival)
K-means	0.0068	0.0009	0.1294	0.1580
PAM	0.0803	0.0133	0.0836	0.1183
Hierarchical	0.0043	0.0007	0.1365	0.1625
GMM	0.0008	0.0000	0.1601	0.1991

Goodman-Kruskal's Tau

Goodman-Kruskal's *lambda* test and Goodman-Kruskal's *tau* tests are also used to determine the strength of the association between two nominal variables. Goodman-Kruskal *tau* assesses the improvement in predicting the dependent variables (ECMO Survival and Hospital Survival) when considering the independent variable (clusters), compared to random category assignment. It varies from 0 to 1 (Signorell & Arppe, Goodman Kruskal's Tau, n.d.). The values across all clustering methods and survival outcomes in Table 8 indicate very weak associations. This suggests that the clustering methods (*k*-means, PAM, Hierarchical, and GMM) do not effectively predict whether patients will survive based on ECMO or Hospital survival outcomes. The Goodman-Kruskal's *lambda* test (Signorell & Arppe, Goodman Kruskal Lambda, n.d.) is also conducted and this test returns the value zero (indicating no association) for all the clustering algorithms. This happens since the modal category (Y) is consistent across all values of the independent variable (clusters 1 and 2), even if the frequencies or percentages of this category differ. Another thing to consider is that the values for Goodman-Kruskal's *tau* marginally improve when the cluster sizes increase. However, two clusters work well for this dataset as concluded from the cluster statistics.

Table 8: Goodman-Kruskal's Tau scores for clustering methods



Goodman-Kruskal Tau Values for Clustering Methods		
Clustering Method	Tau (ECMO Survival)	Tau (Hospital Survival)
K-means	0.0182	0.0265
PAM	0.0080	0.0152
Hierarchical	0.0202	0.0281
GMM	0.0276	0.0418

INSIGHTS FROM CLUSTERING ANALYSIS

As seen previously, the clusters are weakly associated with the survival outcomes. However, insights from clusters could shed a light on their influence on survival outcomes. Given that *k*-means appears to be the most stable clustering method, further analysis is conducted to identify the biomedical factors driving these cluster assignments and determine if these factors have an impact on survival outcomes. The *t*-tests show that the variables *CRP*, *log_RR* and *log_Vt* are not significant to cluster formation of, as there is no difference in means of the two clusters for these variables. Feature importance is assessed using *FeatureImpCluster*, which measures feature importance in *k*-means clustering by iterating through the permutation misclassification rate for each variable in the dataset. The average misclassification rate across all iterations is used to assess variable importance (Pfaffel, 2021) . In Figure 20, it can be observed that *log_IL8* has the highest misclassification rate followed *log_Platelets, log_Lactate, log_Bilirubin, ATIII, log_TNFa, log_siIL2, log_Creatinine, BE* and *log_NAdose*. These results are further validated by using Boruta algorithm for feature selection. The algorithm is a novel feature selection method wrapped around a Random Forest classification approach, designed to identify all relevant variables (Kursa & Rudnicki, 2010). The Boruta algorithm is made robust by using is using five-fold cross validation and varying seeds. Thus, these features are selected as primary features for clustering.

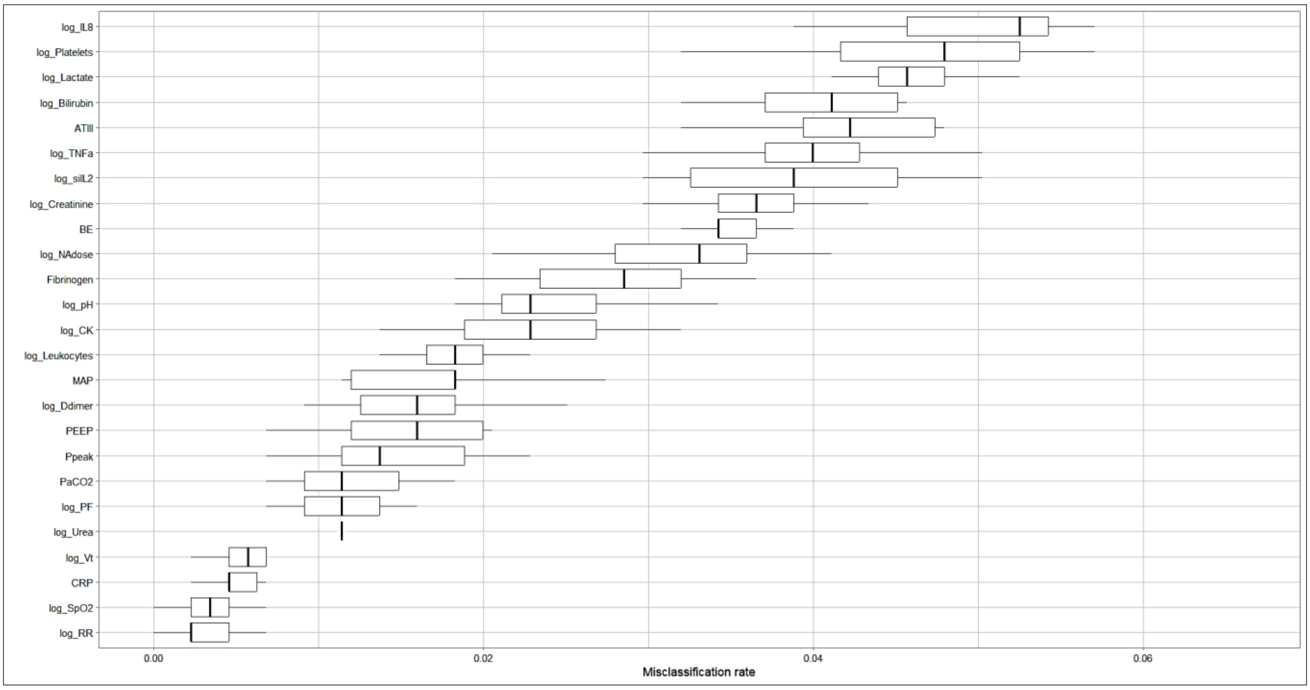


Figure 20: Variable Importance for k-means

k-means clustering is performed on the reduced feature set, yielding results comparable to those from the original clustering. The contingency table shows that the majority of the diagonal entries are high, indicating that the clusters in both results match well. Additionally, the high Adjusted Rand Index further supports the strong similarity between the two clustering outcomes.

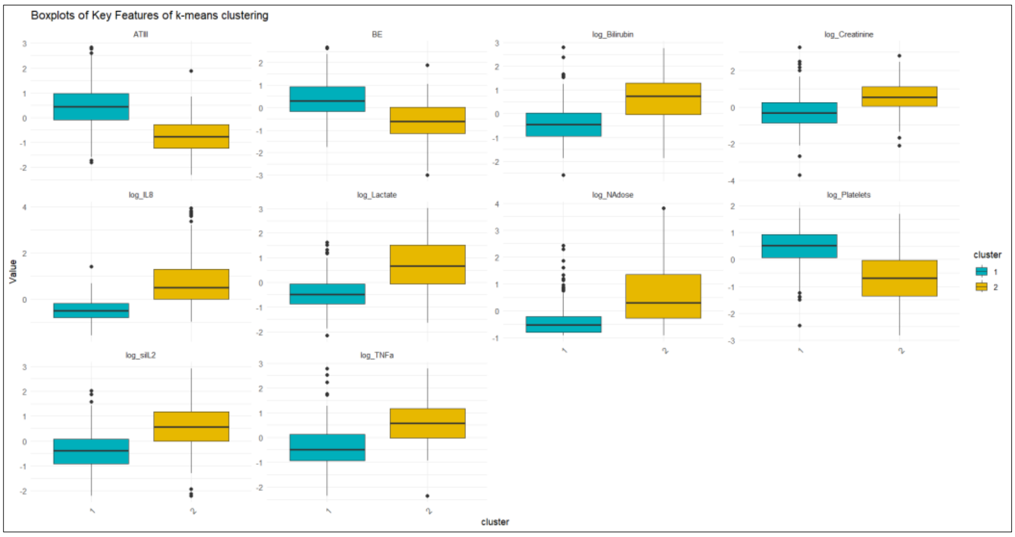


Figure 21: Boxplots of key Biomedical markers for k-means

Figure 21 shows the cluster profiles of the reduced features set which correlate with results from Figure 17. The cluster profiles are defined as

Cluster 1: high *ATIII*, high *BE*, low *log_Bilirubin*, low *log_Creatinine*, low *log_IL8*, low *log_Lactate*, low *log_NAdose*, high *log_Platelets*, low *log_siIL2*, low *log_TNFa*

Cluster 2: low *ATIII*, low *BE*, high *log_Bilirubin*, high *log_Creatinine*, high *log_IL8*, high *log_Lactate*, high *log_NAdose*, low *log_Platelets*, high *log_siIL2*, high *log_TNFa*.

A similar analysis, using *t*-tests and the robust Boruta Algorithm is implemented on PAM, Hierarchical clustering (complete linkage) and model-based clustering. The identified key features are then used to re-cluster the data, and this new clustering is compared with the original. The key features are used to check if they improve the prediction of the response variables. Two feature subsets are created: one based on important features from *k*-means and PAM (Subset A) and the other from Hierarchical and model-based clustering (Subset B). The main difference between the two subsets is Subset A contains *log_Creatinine* but Subset B does not, and Subset B includes *log_pH*, *Fibrinogen,* and *log_Leukocytes* which are absent in Subset A.

Subset A: *log_IL8*, *log_Platelets*, *log_Lactate*, *log_Bilirubin*, *ATIII*, *log_TNFa*, *log_siIL2*, *log_Creatinine*, *BE*, *log_NAdose*

Subset B: *log_NAdose*, *log_Lactate*, *log_IL8*, *log_siIL2*, *log_Platelets*, *log_Bilirubin*, *ATIII*, *BE*, *log_pH*, *log_TNFa*, *Fibrinogen*, *log_Leukocytes*

Support Vector Machine (SVM) classification models are fitted using the Radial Basis Function (RBF) kernel (Yang, Support Vector Machines and Kernelisation, Week 5, Lecture Handout, 2023-24, 2023-24). Cross-validation is used to find the optimal kernel and parameters. There are three models that are fitted and compared

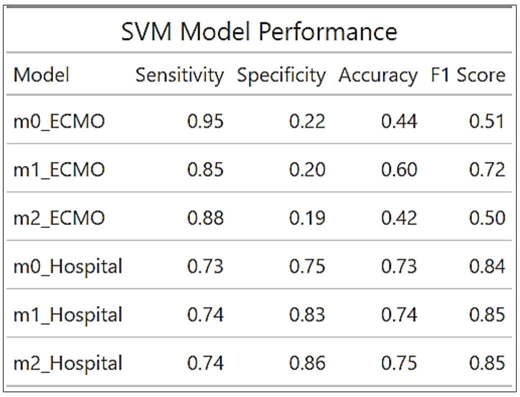
Base model : SVM model containing all the biomedical markers vs the response variables

m0 model: model containing the biomedical markers from subset **A** vs the response variables

m1 model: model containing the biomedical markers from subset **B** vs the response variables

From the Table 9, it can be observed, the models show an improved performance in the metrics Accuracy and F1 score. The model *m2* for ECMO shows a slight decrease in F1 score and Accuracy from the base model.

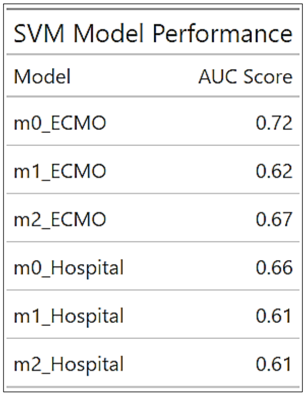
Table 9: Performance metrics of models



SVM Model Performance

Model	Sensitivity	Specificity	Accuracy	F1 Score
m0_ECMO	0.95	0.22	0.44	0.51
m1_ECMO	0.85	0.20	0.60	0.72
m2_ECMO	0.88	0.19	0.42	0.50
m0_Hospital	0.73	0.75	0.73	0.84
m1_Hospital	0.74	0.83	0.74	0.85
m2_Hospital	0.74	0.86	0.75	0.85

Table 10: AUC scores of



SVM Model Performance

Model	AUC Score
m0_ECMO	0.72
m1_ECMO	0.62
m2_ECMO	0.67
m0_Hospital	0.66
m1_Hospital	0.61
m2_Hospital	0.61

Compared to the base model in Table 9, the *m1* for ECMO improves the correct identification of survivors and reduces the Type II error rates (false negatives). The model successfully reduces Type II error, but this improvement is accompanied by an increase in Type I error. This trade-off is a common and often unavoidable aspect of model optimization (Yang, Support Vector Machines and Kernelisation, Week 5, Lecture Handout, 2023-24, 2023-24) . The *m2* for ECMO exhibits a slight decrease in performance relative to the base model by slightly increasing Type I error and decreasing the correct identification of non-survivors. However, the correct identification of survivors and Type II error rates remain unchanged.

Both models for Hospital Survival show improvements in terms of correctly identifying non-survivors. They exhibit a reduction in Type I error (false positives) compared to the base model, while maintaining equivalent true positive rates and Type II error rates (false negatives). The base model, by contrast, shows high sensitivity in identifying survivors with very low Type II error. Thus, the improvements in the Hospital Survival models are reflected in their enhanced specificity without compromising the sensitivity demonstrated by the base model.

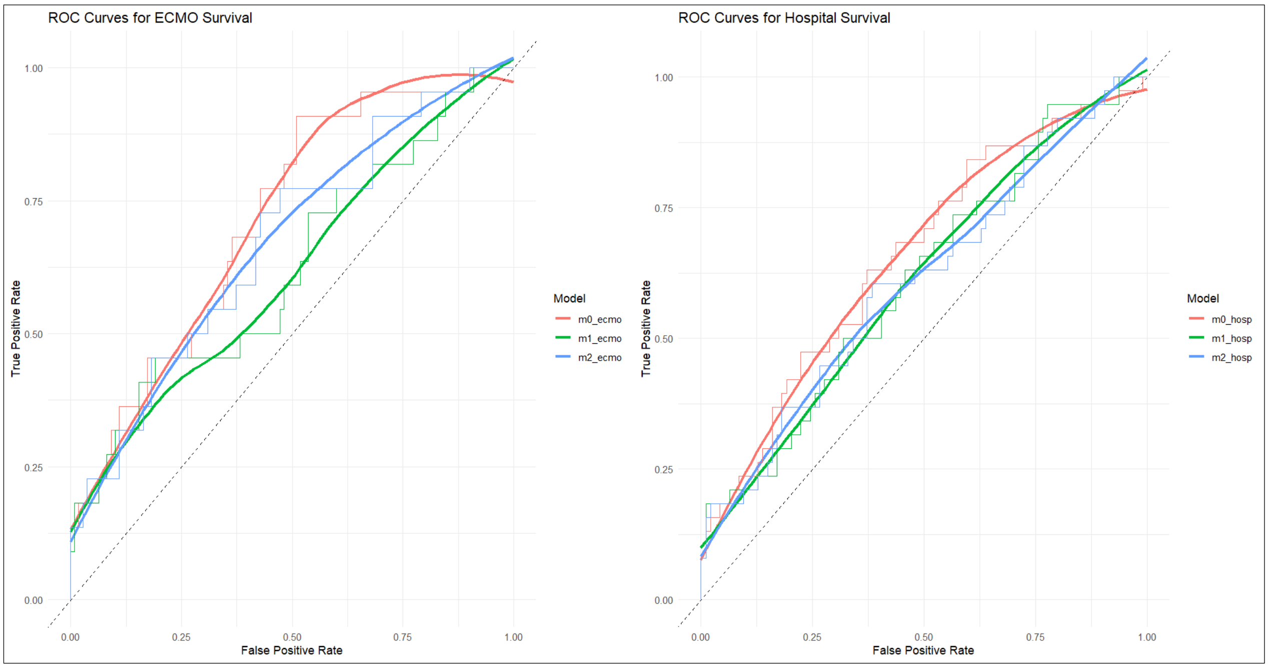


Figure 22: ROC of models

The comparative performance of the models is demonstrated by using the ROC curves in Figure 22 and AUC scores in Table 10. Although the base model exhibits a superior performance compared to the reduced models, the latter demonstrate effectiveness in reducing errors and improving the prediction of survivors and non-survivors.

CHAPTER 4 : CONCLUSION AND DISCUSSION

Conclusion

The first aim of this study is the identification of the clusters in biomedical markers data. The clustering analysis is applied as an exploratory tool to understand the inherent structure of the data. This analysis is conducted by applying *k*-means, Partitioning around mediods, Hierarchical clustering, model-based clustering and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Hard-partitioning methods, *k*-means and PAM, are found to be almost identical in terms of average distance, diameter as well as separation. This is further confirmed by the high adjusted rand index between them. Hierarchical clustering using average and single linkage is useful in identifying potential outliers. DBSCAN also deduces five noise points which are identified previously in the hierarchical clustering and exploratory boxplots. Hierarchical clustering using complete and ward linkage is moderately similar but based on a better average silhouette width and higher Dunn Index, complete linkage is selected. In the model-based clustering algorithm, the silhouette width of clusters is quite low and negative for one cluster implying that the observations are better suited in the neighbouring cluster. The comparison of clustering algorithms reveals that each method offers distinct advantages depending on the evaluation metric. While *k*-means and PAM generally perform well in forming compact and well-separated clusters, hierarchical clustering excels in separation but with more internal variability. Model-based clustering, though less compact, effectively identifies larger and more separated clusters. It is also observed that increasing the number of clusters resulted in poor performance on the cluster validation metrics. These results suggest that partitioning the data into two distinct clusters yield the most stable performance across various clustering algorithms, effectively capturing the inherent structure of the data.

The secondary aim of the analysis is to assess whether the clusters correspond to survival outcome variables. The radar charts show the differences in biomedical markers across clusters, highlighting the variables with the high values in each cluster. Bar charts showed that Cluster 1 generally has higher survival proportions than Cluster 2. Chi-square tests indicated significant associations between clusters generated by *k*-means, hierarchical clustering (complete linkage), and model-based clustering with both response variables, while PAM shows no significant association with ECMO survival. Cramér's V reveals weak associations between clusters and response variables, and Goodman-Kruskal’s tau indicates weak predictive strength of the clusters for survival outcomes. This suggest that the clusters alone may not be sufficient for predicting survival outcomes.

The cluster analysis identified two key feature subsets influencing cluster formation: one common to k-means and PAM (subset A), and another to Hierarchical and model-based clustering (subset B). Applying a Support Vector Machine to these subsets improved accuracy and F1 scores compared to the base model, despite a reduction in overall predictive measures (ROC and AUC). For ECMO survival, subset A reduces false negatives and increases true positives, while for hospital survival, both subsets A and B maintain false negative rates and true positives but reduce false positives and increase true negatives. While the base model remains superior overall, the reduced models effectively reduce Type I and Type II errors in different scenarios, balancing predictability and error reduction. Thus, the reduced models enhance specific aspects of performance, demonstrating their utility in optimizing predictions. Overall, the clustering analysis can be useful in reducing the dimensions to improve the prediction. This analysis highlights the potential use of clustering analysis as an exploratory tool for dimension reduction to identify patient biomedical subgroups that differ significantly in their survival rates, which can inform more personalized treatment approaches or further investigation into the characteristics defining these clusters.

Limitations

Cluster analysis is a powerful tool for recognising patterns in the data, but it presents several challenges and limitations. One key difficulty is determining the optimal number of clusters. Methods like the elbow method, silhouette score, gap statistic and Bayesian Information Criterion provide guidance by comparing cluster solutions based on criteria such as within-cluster variation, between-cluster separation, and model complexity.

However, these methods may yield conflicting results or may not align with domain knowledge. Another significant limitation is the choice of distance metric. Although Euclidean distance is used in this analysis, alternative metrics such as Mahalanobis distance, Manhattan distance etc. could also be explored. Initially, PAM was compared using both Euclidean and Manhattan distances, and the clustering results were found to be quite similar. However, this comparison was not analysed in depth, nor was it extended to other clustering algorithms.

The data analysis faces significant limitations due to the presence of numerous outliers, as indicated by the visualizations. Without input from a clinical expert, it is unclear whether these outliers represent erroneous values or are simply rare occurrences within the context. The lack of domain knowledge about the biomedical profiles prevents informed decisions on whether to remove these values, apply winsorization, or use imputation techniques. Initially, analysis is conducted by removing the outliers using the IQR method, but this does not significantly improve the average silhouette width, and there is no clear justification for excluding these values. After careful consideration, the most extreme values are capped, and winsorization is applied in an attempt to make these values more comparable to less extreme values, acknowledging their impact on the clustering algorithms. In particular, one of the clusters in *k*-means had a large average distance and diameter. Hierarchical clustering methods (complete, average and single linkage) are notably affected by extreme values, with some methods forming entire clusters around single outliers, such as the *log_PH* value for patient ID 1462.

It is important to note that cluster analysis will always partition the data into clusters, but this does not guarantee that the clusters are meaningful. The absence of distinct clusters was suggested by both the principal components analysis and the assessment of clustering tendency. This is further supported by the DBSCAN results, which, after noise removal, produced a single cluster. Additionally, the average silhouette width for the optimal number of clusters is quite low, with the highest value being only 0.15, indicating weak clustering structure. Given these findings, it is possible that the data may only support a single meaningful cluster.

Further Analysis

Given the significant impact of outliers on the clustering results, further analysis could benefit from the application of more robust statistical techniques. Additionally, considering a detailed sensitivity analysis where outliers are systematically excluded or down-weighted could provide insight into the stability and reliability of the clusters. The choice of distance metric is critical in clustering analysis, as it directly influences the resulting clusters. Beyond Euclidean distance, which is commonly used, exploring alternative metrics like Mahalanobis distance, which accounts for correlations between variables, or Manhattan distance, which may be more robust to outliers, could lead to different clustering outcomes. OPTICS (Ordering Points to Identify the Clustering Structure) is an extension of DBSCAN that addresses some of its limitations, particularly in handling clusters of varying density. By applying OPTICS, it would be possible to identify and visualize the hierarchical structure of clusters, including sub-clusters that may not be captured by DBSCAN. This method could also help in identifying meaningful clusters within noisy data. To validate the robustness and predictive power of the key features of the clusters, additional classification models like Random Forests, Gradient Boosting Machines, or Neural Networks could be utilized to assess how well the identified features generalize to predicting survival outcomes.

References

Bernard, G. R., Artigas, A., Brigham, K. L., Carlet, J., Falke, K., Hudson, L., . . . Spragg, R. (1994, March). The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. Retrieved from Am J Respir Crit Care Med: https://doi.org/10.1164/ajrccm.149.3.7509706

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). *clValid: An R Package for Cluster Validation*. Retrieved from Journal of Statistical Software: https://doi.org/10.18637/jss.v025.i04

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2015, April 13). *Package 'NbClust'*. Retrieved from The Comprehensive R Archive Network: https://cran.r-project.org/web/packages/NbClust/NbClust.pdf

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters. *Association for the Advancement of Artificial Intelligence (AAAI)*. Retrieved from AAAI (www.aaai.org): https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf

Force, T. A., VM, R., GD, R., BT, T., ND, F., E, C., . . . AS, S. (2012, June 20). Acute Respiratory Distress Syndrome: The Berlin Definition. *JAMA*. Retrieved from JAMA: https://doi.org/10.1001/jama.2012.5669

Fraley, C., & Raftery, A. E. (2002, June). Model-based clustering, discriminant analysis, and density estimation. Retrieved from Journal of the American Statistical Association: https://sites.stat.washington.edu/raftery/Research/PDF/fraley2002.pdf

G, B., JG, L., T, P., E, F., L, B., A, E., . . . Group, E. T. (2016, February 2023). Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. *JAMA*. Retrieved from JAMA: https://doi.org/10.1001/jama.2016.0291

Galili, T. (2023, March 24). *The Fowlkes-Mallows Index and the Bk plot*. Retrieved from Introduction to dendextend: https://cran.r-project.org/web/packages/dendextend/vignettes/dendextend.html#the-fowlkes-mallows-index-and-the-bk-plot

Hahsler, M. (2019, October 1). Density-based Spatial Clustering of Applications with Noise (DBSCAN). *Journal of Statistical Software*. Retrieved from R documentation: https://www.jstatsoft.org/article/view/v091i01

Hennig, C. (2024, April 30). *Cluster validation statistics*. Retrieved from The Comprehensive R Archive Network: https://search.r-project.org/CRAN/refmans/fpc/html/cluster.stats.html

Hudson, G. D., Caldwell, E., Peabody, E., Weaver, J., Martin, D. P., Neff, M., . . . D., L. (2005, October 20). *Incidence and Outcomes of Acute Lung Injury*. Retrieved from New England Journal of Medicine: https://www.nejm.org/doi/full/10.1056/NEJMoa050333

Kassambara, A. (2017). *Practical Guide To Cluster Analysis In R Unsupervised Machine Learning Sthda (2017)*. Retrieved from STHDA: http://www.sthda.com/english/articles/25-clusteranalysis-in-r-practical-guide/

Kaufman, L., & Rousseeuw, P. J. (1990). In P. J. Leonard Kaufman, *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley.

Kuchciak, M. (2024, January). *Hierarchical clustering – pros, cons, interpretation, application*. Retrieved from Rpubs by Rstudio: https://rpubs.com/TusVasMit/HierarchicalClusteringOverwiev

Kursa, M. B., & Rudnicki, W. R. (2010, September 11). *Feature Selection with the Boruta Package*. Retrieved from Journal of Statistical Software: https://www.jstatsoft.org/article/view/v036i11

Li, D., & Kuhn, m. b. (n.d.). *Determine highly correlated variables*. Retrieved from The Comprehensive R Archive Network: https://search.r-project.org/CRAN/refmans/caret/html/findCorrelation.html

Mangiafico, S. (2016). *Summary and Analysis of Extension Program Evaluation in R, version 1.20.07, revised 2024*. Retrieved from Rcompanion: http://rcompanion.org/handbook/

Pfaffel, O. (2021, October 20). *Package 'FeatureImpCluster' - Feature Importance for Partitional Clustering*. Retrieved from The Comprehensive R Archive Network: https://cran.r-project.org/web/packages/FeatureImpCluster/FeatureImpCluster.pdf

School of Mathematics and Statistics, U. o. (2023-2024, May 10). *MSc Projects in Statistics and Data Analytics 2023/24.* Retrieved from STATS5029P/5090P Statistics Project and Dissertation 2023-2024 (Summer 2024): https://moodle.gla.ac.uk/pluginfile.php/8096979/mod_resource/content/3/2024MScProjectsList.pdf

Signorell, A., & Arppe, A. (n.d.). *Goodman Kruskal Lambda*. Retrieved from The Comprehensive R Archive Network: https://search.r-project.org/CRAN/refmans/DescTools/html/Lambda.html

Signorell, A., & Arppe, A. (n.d.). *Goodman Kruskal's Tau*. Retrieved from Comprehensive R Archive Network: https://search.r-project.org/CRAN/refmans/DescTools/html/GoodmanKruskalTau.html

Soetewey, A. (2020, January 27). *Chi-square test of independence in R*. Retrieved from Stats and R: https://statsandr.com/blog/chi-square-test-of-independence-in-r/#chi-square-test-of-independence-in-r

Wilson, S. (2021, September 06). *The MICE Algorithm*. Retrieved from Comprehensive R Archive Network: https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html#:~:text=miceRanger%20can%20make%20use%20of,value%20of%20the%20missing%20sample.

Yang, X. (2023-2024). *Partitioning Cluster Analysis, Week 9, Lecture handout.* Retrieved from Moodle, University of Glasgow: https://moodle.gla.ac.uk/pluginfile.php/6887526/mod_resource/content/9/Week9-Partitioning%20cluster%20analysis.pdf

Yang, X. (2023-24). *Hierarchical Cluster Analysis, Week 8, Lecture handout.* Retrieved from Moodle, University of Glasgow: https://moodle.gla.ac.uk/pluginfile.php/6887509/mod_resource/content/7/Week8-Hierarchical%20cluster%20analysis.pdf

Yang, X. (2023-24). *Support Vector Machines and Kernelisation, Week 5, Lecture Handout, 2023-24.* Retrieved from Moodle, University of Glasgow: https://moodle.gla.ac.uk/pluginfile.php/6887458/mod_resource/content/8/Week5-Support%20vector%20machines.pdf