UNIVERSITY of WASHINGTON

Project Report

(DATA 591: Data Science Capstone II - Project Implementation)

Ventilator Associated Pneumonia (VAP) Subphenotype Analysis

Parvati Jayakumar, Ted Liu

Table of Contents

Introduction

Pneumonia is the second most common hospital-acquired infections (HAI) among critically ill patients, with Ventilator-Associated Pneumonia (VAP) accounting for approximately 86% of cases [1]. VAP occurs in patients who require mechanical ventilation for at least 48 hours, making it a major concern in intensive care units (ICUs) worldwide. It is associated with prolonged hospital stays, increased healthcare costs, and high mortality rates, particularly in patients with underlying conditions. Despite extensive research and improvements in critical care protocols, the factors influencing VAP outcomes remain poorly understood.

The diagnosis of VAP is complex, requiring a combination of bedside clinical assessment, radiographic imaging, and microbiological analysis. However, heterogeneity among VAP patients, variations in causative pathogens, and differing host immune responses complicate disease characterization and treatment decisions. As a result, clinicians often face challenges in differentiating VAP from other forms of pneumonia and determining the most effective therapeutic strategies for individual patients. While traditional clinical criteria guide VAP management, recent advances in biomarker research and computational modeling suggest that molecular subphenotypes (or "endotypes") may exist, each associated with distinct disease trajectories and treatment responses.

This project aims to leverage electronic health record (EHR) data and biomarker analysis to classify VAP patients into clinically meaningful subphenotypes and investigate their association with mortality, ventilator-free days, and severity of respiratory failure. By applying advanced statistical and machine learning techniques, we seek to uncover patterns that may improve risk stratification, guide personalized treatment decisions, and enhance clinical management of VAP. Understanding whether distinct molecular and clinical signatures exist within VAP patients could enable more precise interventions, ultimately reducing complications and improving patient outcomes.

Through this project, we aim to contribute to the broader field of precision medicine in critical care, integrating data-driven methodologies with clinical expertise to refine diagnostic and therapeutic strategies for critically ill patients with VAP. By identifying biomarker-driven subphenotypes, this study has the potential to bridge the gap between traditional diagnostic models and personalized medicine, leading to more targeted and effective interventions in ICU settings.

Background

Latent Class Analysis and Its Relevance to Pulmonary and Critical Care Medicine

In the field of pulmonary and critical care medicine, clustering methods play a crucial role in identifying underlying patient subgroups that may have distinct clinical characteristics and outcomes. One widely used clustering method in this domain is Latent Class Analysis (LCA) / Latent Profile Analysis (LPA), which is particularly valuable for identifying hidden structures within heterogeneous patient populations. For our capstone project with Dr. Eric Morrell, we were tasked with applying LCA-based clustering to analyze biomarker profiles in Ventilator-Associated Pneumonia (VAP) patients, aiming to uncover potential subphenotypes that could inform risk stratification and clinical decision-making.

[2] provides a comprehensive overview of LCA, detailing its applications, methodological considerations, and key challenges. LCA is defined as a finite mixture model, where latent groups (or clusters) are inferred from observed variables (indicators). The paper distinguishes between Latent Class Analysis (LCA), which uses categorical indicators, and Latent Profile Analysis (LPA), which applies to continuous indicators. Importantly, the authors note that LCA is often used as a blanket term for both approaches, a convention also seen in key studies in pulmonary medicine, such as [3] and [4].

Methodological Framework for LCA

To ensure a rigorous application of LCA, the paper outlines a five-step approach that serves as a best-practice guide for performing LCA-based clustering:

1. **Generate Hypothesis:** Define the expected latent subgroups and their clinical significance.
2. **Data Set-Up:** Select relevant indicators (biomarkers, clinical measures) for the model.
3. **Estimate Models:** Fit LCA models with varying numbers of classes.
4. **Evaluate Models:** Compare models using fit statistics such as Bayesian Information Criterion (BIC), entropy, and likelihood-ratio tests.
5. **Interpret Optimal Models:** Assess the clinical relevance of identified subgroups and validate findings.

For our project, determining the optimal number of latent classes is a crucial step. The paper emphasizes that when using a mix of categorical and continuous indicators, the Bayesian Information Criterion (BIC) "elbow method" is the most effective technique for selecting the optimal number of clusters [2]. Additional tests, such as the Lo-Mendell-Rubin likelihood ratio test (VLMR) and its bootstrapped variant, are also discussed. However, the paper warns that the bootstrapped VLMR test often favors models with more classes than necessary, limiting its usefulness. Given that our dataset contains a mix of categorical and continuous variables, we

will primarily rely on the BIC elbow method to determine the number of subphenotypes in our VAP cohort.

Comparing Latent Class Analysis with Other Clustering Methods

Beyond LCA, clustering methodologies are fundamental to unsupervised learning, allowing for pattern discovery in large medical datasets. [5] explores various clustering techniques, highlighting their strengths, weaknesses, and application areas. The authors emphasize that no single clustering algorithm is universally applicable, reinforcing the importance of careful selection based on data characteristics. A critical distinction between LCA and traditional clustering algorithms (e.g., K-Means, Gaussian Mixture Models, Hierarchical Clustering) is the way group membership is assigned. Unlike distance-based clustering methods, which separate data points into discrete clusters based on similarity measures, LCA estimates the probability that each observation belongs to a given latent class [2]. This probabilistic nature provides greater flexibility and robustness, particularly in medical datasets where variability among patients is high.

Furthermore, empirical comparisons between LCA and K-Means clustering suggest that LCA often yields more accurate subgroup classifications. A cited study in the paper evaluated LCA and K-Means on a simulated dataset where the true subgroups were known but censored. The findings demonstrated that K-Means had a misclassification rate nearly four times higher than LCA. Additionally, LCA can accommodate both categorical and continuous variables within the same model, a significant advantage over K-Means, which typically requires numerical data. However, the primary drawback of LCA is its high computational cost, making it less practical for extremely large datasets.

Use of Modified Poisson and Logistic Regression for Risk Estimation

In addition to clustering techniques, our project requires an appropriate method for estimating the relationship between subphenotypes and clinical outcomes (e.g., mortality, ventilator-free days). We employ both Modified Poisson Regression and Logistic Regression, two commonly used approaches in epidemiological research.

Modified Poisson Regression for Relative Risk (RR) Estimation

We use Modified Poisson Regression with robust standard errors, as proposed in [6]. This method allows us to directly estimate RR while controlling for covariates, avoiding the common pitfall of interpreting Odds Ratios (OR) as RR in logistic regression models. The paper highlights that logistic regression is often misused in prospective studies, as OR can significantly overestimate RR when the event is common. Modified Poisson Regression solves this issue by applying a Poisson regression model with a robust sandwich estimator, ensuring valid confidence intervals and reliable RR estimates, even in small sample sizes [6]. Given our

interest in examining VAP subphenotypes and their association with mortality and ventilator-free days, this method will be crucial in our analysis.

Logistic Regression for Odds Ratio (OR) Estimation

In parallel, we apply Logistic Regression to estimate Odds Ratios (OR), which measure the association between VAP subphenotypes and binary outcomes. Logistic regression remains a widely used statistical approach, particularly when controlling for multiple covariates. While OR can be misleading when interpreted as RR, it is still valuable for evaluating relative likelihoods of clinical outcomes across patient clusters.

Objective

Our project aims to identify and characterize subphenotypes of Ventilator-Associated Pneumonia (VAP) patients using Latent Class Analysis (LCA) and other clustering techniques. By utilizing Electronic Health Record (EHR) data and biomarker analysis, we seek to enhance risk stratification, improve clinical decision-making, and explore personalized treatment strategies for critically ill patients.

1. **Identify Subphenotypes of VAP:**
    - Utilize Latent Class Analysis (LCA) to classify VAP patients based on biomarker signatures and clinical variables.
    - Compare LCA with traditional clustering methods (K-Means, GMM) to assess which approach best differentiates patient subgroups.
        - Determine the optimal number of clusters using evaluation metrics such as Bayesian Information Criterion (BIC) elbow method, Silhouette Scores, Gap Statistics etc,.
2. **Assess Clinical Outcomes Across Subphenotypes:**
    - Evaluate whether distinct VAP subphenotypes exhibit differences in clinical outcomes such as Mortality, Ventilator-Free Days (VFDs), and Severity of Respiratory Failure
    - Determine whether these subphenotypes have prognostic value in guiding treatment strategies.
3. **Identify Key Biomarkers Using LASSO Regression:**
    - Perform Least Absolute Shrinkage and Selection Operator (LASSO) regression to identify which biomarkers are most predictive of VAP severity and patient outcomes.
4. **Perform Statistical Tests to Evaluate Associations with Clinical Outcomes:**
    - Utilize additional statistical hypothesis tests (e.g., chi-square tests, Kruskal-Wallis tests, etc.) to compare demographic, clinical, and biomarker distributions across subgroups.
    - Implement Modified Poisson Regression to estimate Relative Risk (RR) while adjusting for covariates.

    - Conduct Logistic Regression for comparative analysis, identify factors significantly associated with adverse outcomes in VAP patients.

By fulfilling these objectives, our project aims to contribute to precision medicine in critical care, improving the ability to predict which VAP patients are at highest risk for poor outcomes and tailoring interventions accordingly.

Data Source

The primary data source for this project is the EPIC Electronic Health Record (EHR) system, which provides a comprehensive set of patient health data. This dataset consists of 466 unique subjects and includes key clinical metrics such as vital signs, laboratory results, microbiology cultures, and clinical outcomes (e.g., Ventilator-associated pneumonia (VAP), Acute respiratory distress syndrome (ARDS)).

Given the sensitive nature of patient information, all data handling is performed in compliance with institutional privacy and security protocols. The dataset is de-identified and securely stored on a dedicated server at UW Medicine, managed by Dr. Eric Morrell. Regular updates to the datasets will be coordinated by Dr. Morrell to ensure that the most current data remains available for our analysis. Due to the sensitive nature of the data, in the team, only Ted had access to the server. He assisted Dr. Morrell in retrieving the data. Dr. Morrell finally shared the updated datasets with Parvati via email in Excel or CSV format.

Methodology

Data Acquisition

The data acquisition process follows a structured ETL (Extract, Transform, Load) workflow, ensuring the integrity and completeness of extracted data:

- **Data Extraction**:
    - The dataset is retrieved using SQL queries on the EPIC EHR system, focusing on key healthcare indicators such as vitals, laboratory tests, and microbiology cultures.
    - Extraction is handled exclusively by Ted and Dr. Eric Morrell, maintaining compliance with privacy regulations and patient data security.
- **Data Storage and Access**:
    - The extracted dataset is stored securely on a UW Medicine-managed server.
    - Ted assisted Dr. Morrell in retrieving and updating the dataset, ensuring data availability and quality.
    - The analysis-ready dataset is shared with Parvati via email in Excel or CSV format for further processing.

Data Processing

The data processing workflow transitions extracted data into an analysis-ready format, optimizing it for statistical modeling and machine learning applications.

Data Cleaning & Preprocessing

- The dataset underwent rigorous cleaning and preprocessing using Python and R, ensuring consistency and accuracy.
- Key cleaning steps include:
    - Handling missing data (imputation or removal based on data integrity checks).
    - Standardizing variable formats to align with protocols:
        - First, applying $\log_2$ transformation to stabilize variance and handle skewness.
        - Then, performing Z-score normalization to scale variables, ensuring they have a mean of 0 and standard deviation of 1, making them suitable for statistical analysis and machine learning models.
    - Ensuring correct data types for statistical modeling.

Data Formatting

- The final wide-format dataset presented each subject's data in a single-row structure, where individual biomarkers, clinical measures, and outcomes are represented as columns.
- This format allowed for efficient statistical analysis and predictive modeling while maintaining longitudinal insights into patient health trajectories..

Clustering

To identify potential VAP subphenotypes, we apply multiple unsupervised clustering methods to classify patients based on biomarker signatures and clinical characteristics.

We applied these unsupervised clustering methods to a panel of 25 biomarkers consisting of chemokines and proinflammatory cell markers. These biomarkers were measured in bronchoalveolar fluid in all 466 patients in the study.

Latent Class Analysis (LCA)

Latent Class Analysis or sometimes known as Latent Profile Analysis is widely used in the pulmonary and critical care research area. Latent Class Analysis is a probabilistic modelling algorithm. LCA is a finite mixture model whose results are obtained using maximum likelihood estimates. LCA functions on the assumption that the observed distribution of a variable is the result of a mixture of smaller distributions belonging to different latent classes.

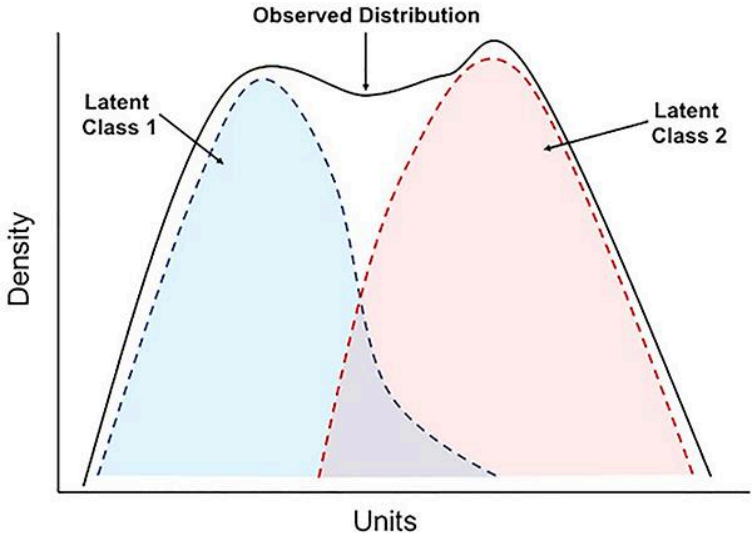

Figure 1: Latent Class Analysis

K-Means Clustering

K-Means Clustering is a widely used unsupervised machine learning algorithm that partitions a dataset into a predefined number of clusters ("k") by grouping data points that are most similar to each other based on their proximity to a central point (centroid) within each cluster. The algorithm iteratively adjusts cluster assignments to minimize the distance between data points and their assigned cluster center, aiming to uncover patterns and groupings within the data by ensuring that similar data points are placed together.

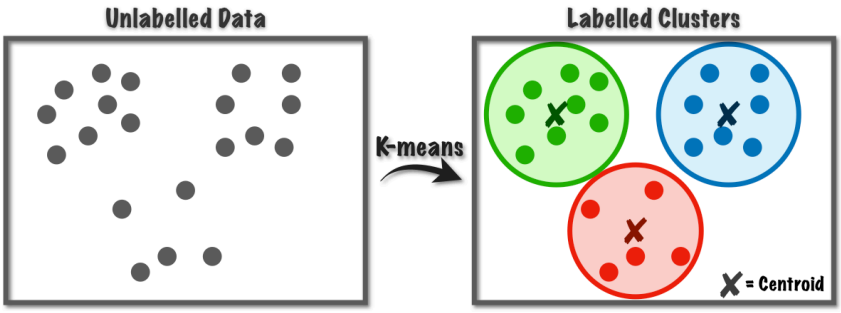

Figure 2: K-Means Clustering

K-Means will be applied to the same biomarker and clinical data used for LCA, allowing us to assess whether the patterns in patient subgroups remain consistent across different clustering methodologies. We will determine the optimal number of clusters ("k") using the Elbow Method to identify the point where adding more clusters provides minimal improvement. Additionally, we will evaluate cluster quality using Silhouette Scores, which measure how well-separated the clusters are.

Gaussian Mixture Modeling

GMMs extend the principles of K-Means by allowing clusters to have different shapes and probabilistic assignment rather than the hard-clustering approach used in K-Means. Unlike K-Means, which assumes spherical clusters, GMMs model clusters as elliptical Gaussian distributions, making them more flexible for real-world biomedical datasets where data distributions are rarely uniform. Using Expectation-Maximization (E-M), GMMs iteratively estimate the likelihood of data points belonging to each cluster, providing a probabilistic measure of cluster membership, which is particularly useful for analyzing uncertainty in subgroup classifications.

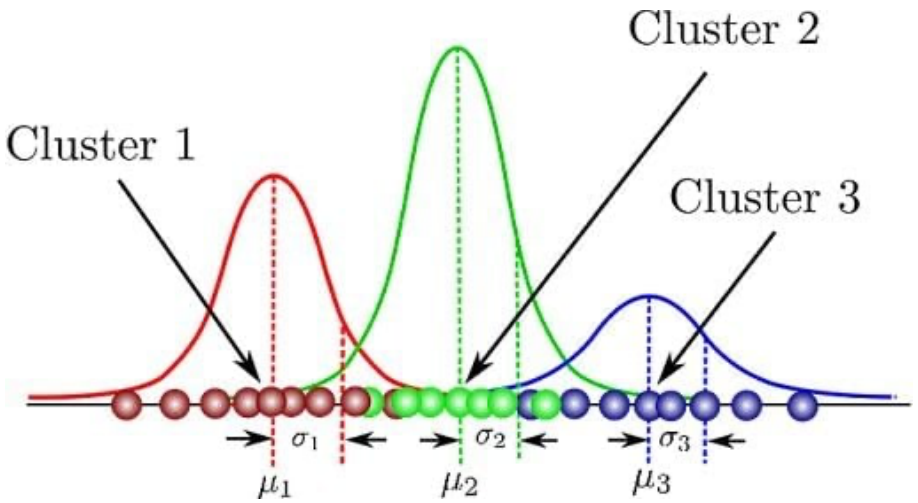

Figure 3: Gaussian Mixture Modeling (GMM)

GMMs will be applied to the same biomarker and clinical data used for LCA and K-Means. To determine the optimal number of clusters in GMMs, we use both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These metrics evaluate the trade-off between model complexity and goodness of fit, with lower AIC/BIC values indicating a better-fitting model. BIC tends to favor simpler models, while AIC is more flexible, allowing for more complex cluster structures when needed. By comparing AIC/BIC values across different numbers of clusters, we identify the best model for our data while minimizing overfitting.

After clustering patients with K-Means and GMMs, we compare the cluster assignments with those obtained from LCA to determine how well the methods align.

Statistical Analysis

Following clustering, we perform statistical analysis to assess the relationship between identified VAP subphenotypes and key clinical outcomes, ensuring that the clusters are not only statistically distinct but also clinically meaningful. Our approach includes feature selection, hypothesis testing, and regression modeling to extract insights from the data.

Feature Selection Using LASSO Regression

To determine which biomarkers contribute the most to VAP subphenotypes and clinical outcomes, we employ Least Absolute Shrinkage and Selection Operator (LASSO) regression. This method applies L1 regularization, which shrinks the coefficients of less important features to zero, effectively performing automated variable selection. LASSO helps in identifying the most predictive biomarkers while reducing overfitting, ensuring that only the most clinically relevant variables are included in the subsequent models.

Hypothesis Testing for Group Comparisons

To compare clinical characteristics and outcomes across the identified subphenotypes, we apply appropriate statistical tests based on data type and distribution:

- **Continuous Variables:** We use Mann-Whitney test (for two groups), and Kruskal-Wallis tests if the data is non-parametric.
- **Categorical Variables:** We perform Fisher's exact tests (for 2 groups), and Chi-square tests to evaluate differences in categorical features such as sex, comorbidities, or treatment responses across clusters.

Regression Models for Risk Estimation

To quantify the association between VAP subphenotypes and clinical outcomes, we apply both Modified Poisson Regression and Logistic Regression:

- **Modified Poisson Regression:** Used to estimate Relative Risk (RR) while adjusting for covariates, ensuring valid confidence intervals and accurate effect size estimation. This approach, proposed in [6], is particularly useful in cohort studies where odds ratios tend to overestimate risk. RR compares the probability of an outcome occurring in one group relative to another. It tells us how much more (or less) likely an outcome is in an exposed group compared to a reference group.

  $$RR = \frac{P(\text{Outcome} \mid \text{Exposed})}{P(\text{Outcome} \mid \text{Unexposed})}$$

  , where P(Outcome | Exposed) is the probability of the outcome in the exposed group, and P(Outcome | Unexposed) is the probability of the outcome in the reference (control) group.
- **Logistic Regression:** Used to compute Odds Ratios (OR), assessing the likelihood of adverse outcomes in each cluster. OR is useful for binary outcomes (e.g., mortality: yes/no), but we must interpret it cautiously in cases where the outcome is common to avoid misrepresenting risk. OR compares the odds of an outcome occurring in one group versus another, rather than the absolute probability. Unlike RR, it does not directly measure risk but rather the likelihood of the event occurring relative to its non-occurrence.

$$OR = \frac{P(\text{Outcome} \mid \text{Exposed}) / P(\text{No Outcome} \mid \text{Exposed})}{P(\text{Outcome} \mid \text{Unexposed}) / P(\text{No Outcome} \mid \text{Unexposed})}$$, where P(Outcome | Exposed) is the probability of the outcome in the exposed group, and P(Outcome | Unexposed) is the probability of the outcome in the reference (control) group.

By integrating feature selection, hypothesis testing, and regression modeling, our statistical analysis aims to provide clinically actionable insights into VAP subphenotypes, ensuring that our findings are both statistically sound and relevant for patient care.

Results

Before performing clustering and statistical analysis, we applied a series of data cleaning and normalization steps to ensure data quality and consistency. The initial dataset contained 814 rows and 2677 features, including biomarker values and clinical parameters. First, we filtered the dataset to include only VAP patients, removing repeat measurements to retain only the first recorded observation per subject. Additionally, we handled missing values by applying median imputation for biomarkers with moderate missingness, while those with more than 20% missing values were removed to maintain data reliability.

For selecting the biomarkers used in our analysis, we removed IL-8 chemokine and IL-10 proinflammatory biomarkers as they did not meet the quality control criteria defined by Dr. Morrell.

Next, we applied $\log_2$ transformation to stabilize variance across biomarker values, followed by Z-score normalization to ensure all features were on a comparable scale with a mean of 0 and a standard deviation of 1. These transformations help prevent certain biomarkers with large numerical ranges from disproportionately influencing clustering and regression models.

After preprocessing, the final dataset contained 466 unique patients and 25 biomarkers, with subject IDs retained to allow subject-specific tracking in downstream analyses.

Clustering Analysis

K-Means Clustering

To identify potential VAP subphenotypes, we applied K-Means Clustering to the standardized biomarker dataset. The Elbow Method (Figure 4) and Silhouette Score (Figure 5) analysis were used to determine the optimal number of clusters. Our analysis suggested that k = 2 was the most appropriate number of clusters, as it produced the highest silhouette score and the most distinct separation of patient groups.

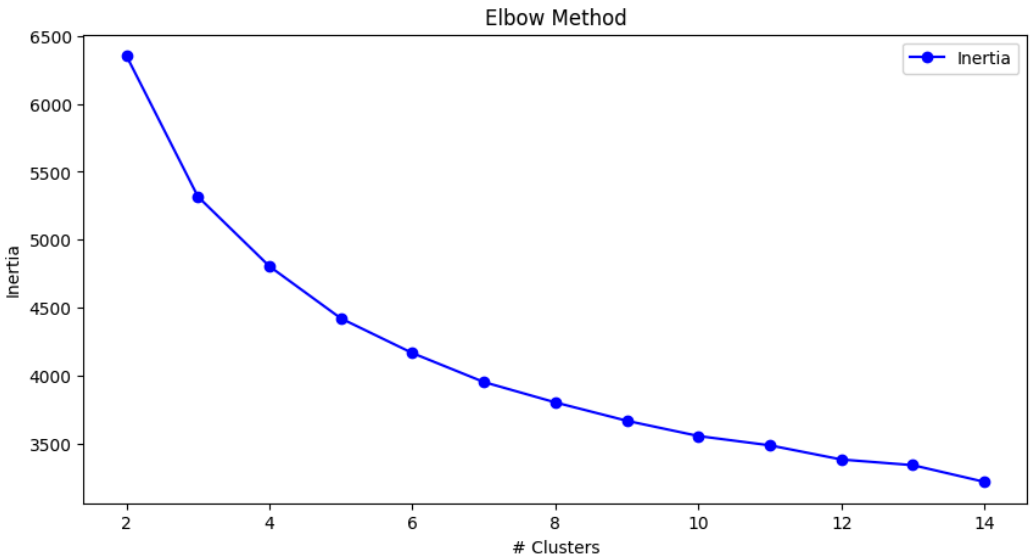

Figure 4: K-Means Elbow Method Results

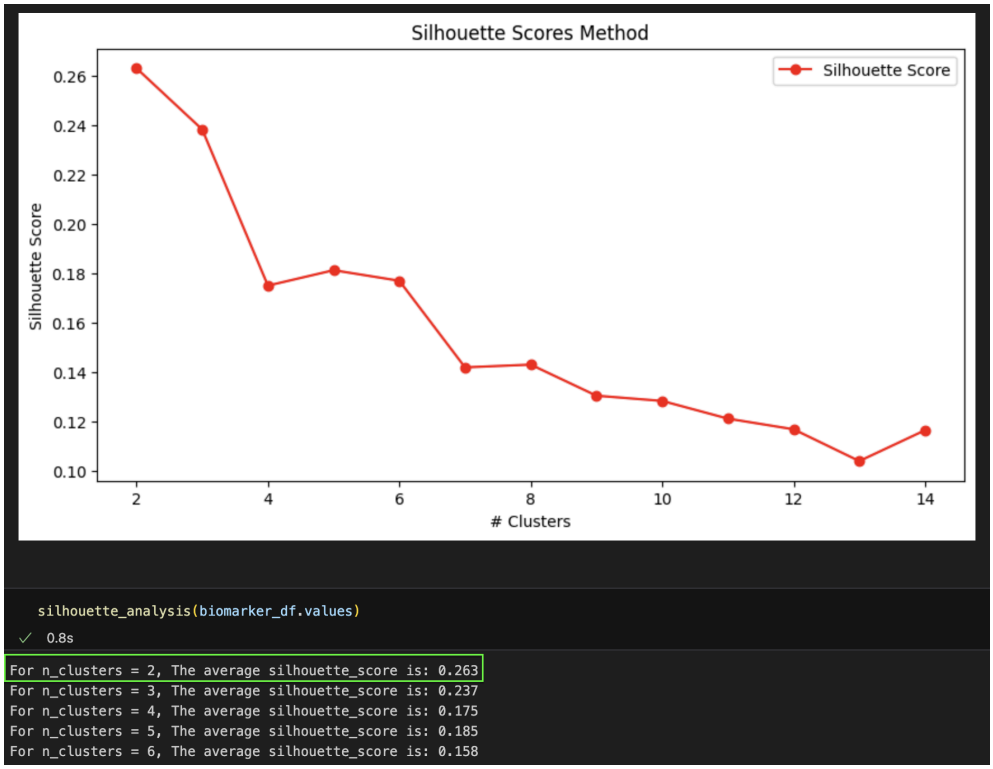

Figure 5: K-Means Silhouette scores Method Results

After performing K-Means clustering on the full biomarker dataset, we assigned each patient to one of two clusters, representing distinct biomarker-driven subgroups. To visually assess clustering quality, we generated 2D (Figure 6) and 3D PCA visualizations, which showed a clear

separation between the two clusters. Additionally, we plotted standardized biomarker values across clusters (Figure 7), revealing significant differences in biomarker expression between the identified subgroups.

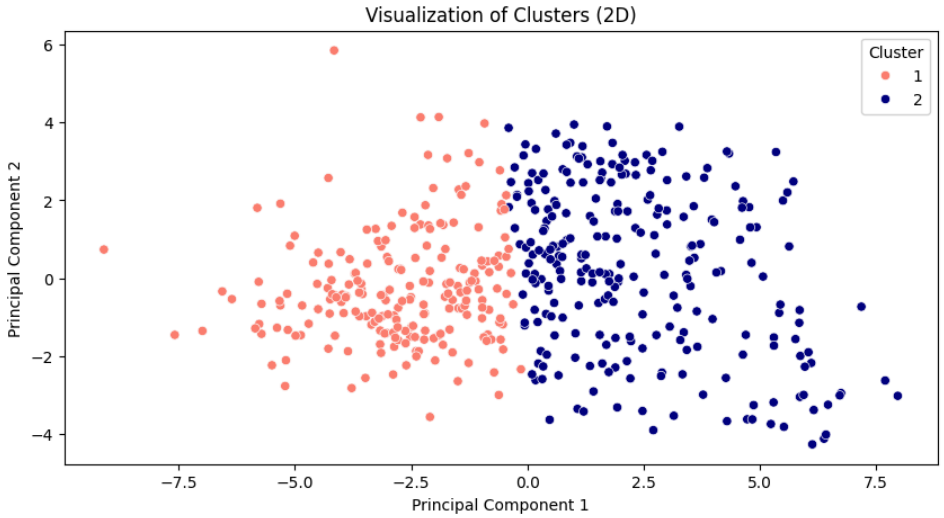

Figure 6: K-Means PCA Visualization

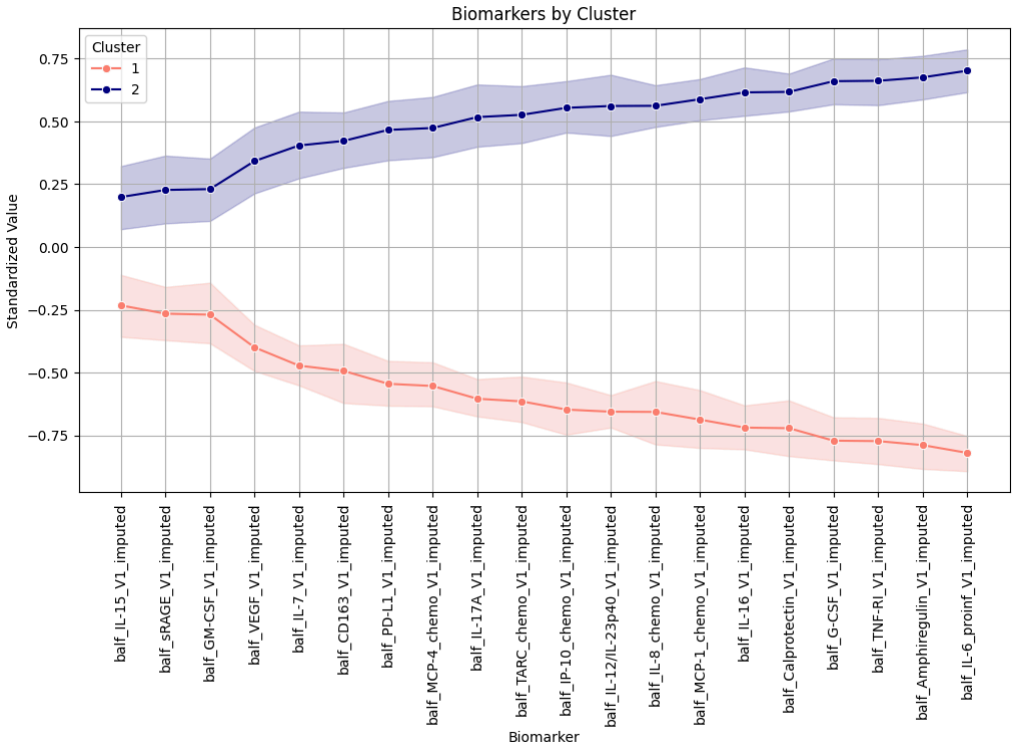

Figure 7: Standardized biomarker values across K-Means clusters

To examine the impact of highly correlated biomarkers, we repeated the K-Means clustering on a dataset where highly correlated features were removed. Despite the reduction in redundant features, the optimal number of clusters remained at k = 2, and the clustering results remained largely consistent.

To identify the key biomarkers driving cluster differentiation, we applied LASSO logistic regression on the K-Means cluster assignments. LASSO effectively performed feature selection by shrinking coefficients of less relevant biomarkers to zero, retaining only the most predictive variables. Our analysis revealed that CD163, Amphiregulin, and MCP-1 were the strongest biomarkers differentiating the two clusters. However, the AUC score of 1.0 on the validation set indicated perfect classification, which strongly suggests overfitting. This is particularly concerning given that K-Means clustering resulted in only two clusters (k=2), potentially oversimplifying patient heterogeneity. The limited number of clusters may have forced the model to find overly distinct separation, reducing generalizability to unseen data. To enhance robustness and biological relevance, we decided to explore alternative clustering methods, such as Gaussian Mixture Models (GMMs).

Gaussian Mixture Modeling

We applied Gaussian Mixture Models (GMMs) to allow for elliptical clusters and probabilistic cluster assignments, which better capture the complex distributions in biomarker data. Unlike K-Means, which forces rigid, spherical clusters, GMM is more flexible and enables soft clustering, meaning each patient is assigned a probability of belonging to each cluster rather than a strict binary label.

To find the best number of clusters (k), we used Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), which evaluate the balance between model fit and complexity (Figure 8).

Our analysis showed:

- BIC reached its lowest value at k = 4 before increasing, suggesting that adding more clusters beyond this point led to overfitting rather than meaningful separation.
- AIC continued to decrease, which suggests that additional clusters improved the likelihood, but since AIC often favors complex models, it can lead to unnecessary cluster fragmentation.

Given these results, we selected k = 4 as the optimal number of clusters.

We performed GMM clustering and assigned each patient to one of four clusters based on their biomarker profiles. To validate the separation of clusters, we generated 2D PCA visualizations (Figure 9). However, despite improved separation compared to K-Means, the clusters still showed some overlap, indicating that while GMM provided more flexibility, it still struggled to create clinically meaningful VAP subphenotypes.

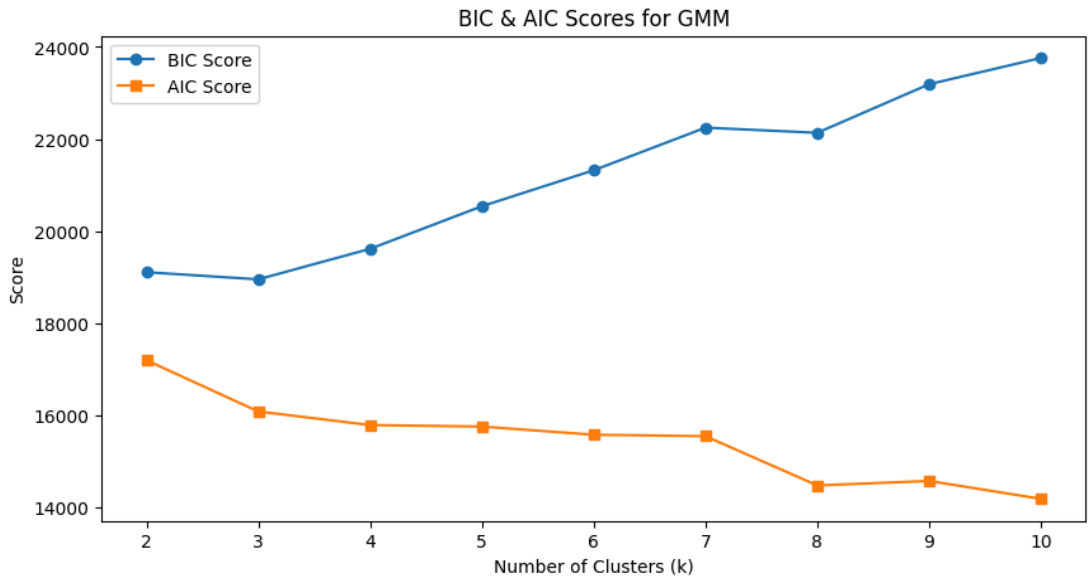

Figure 8: BIC & AIC Scores for GMM

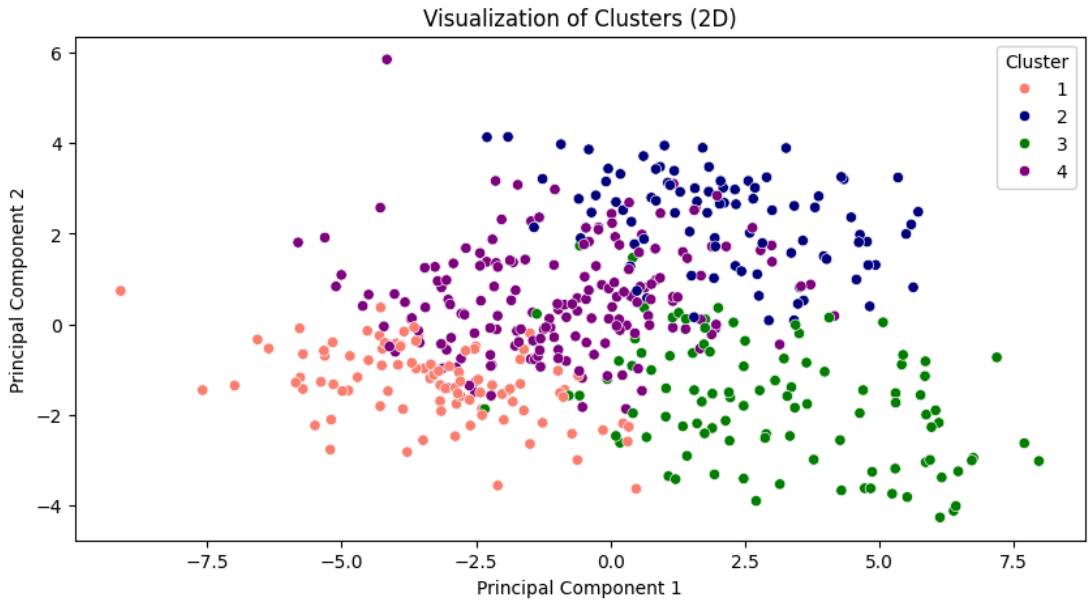

Figure 9: GMM PCA Visualization

Additionally, the BIC and AIC trends did not strongly indicate a single best k, making cluster selection somewhat ambiguous. Also, some clusters lacked strong clinical or biological differentiation, raising concerns about the relevance of the GMM-generated subphenotypes. Another key limitation was that soft clustering did not always align with real-world biological distinctions, as probabilistic assignments resulted in some patients having nearly equal probabilities of belonging to multiple clusters, leading to uncertainty in subgroup classification. Given these issues, we decided that GMM may not be the best approach for defining VAP subphenotypes. We then decided to try out Latent Class Analysis (LCA).

Latent Class Analysis (LCA)

We applied Latent Class Analysis in Mplus version 8.11 onto our dataset of 25 biomarkers.For selecting the optimal number of latent classes, we utilized the BIC, Entropy, and the Vuong-Lo-Mendall-Rubin (VLMR) Test (Table 1).

Mplus			
Classes	BIC	Entropy	VLMR p-value
2	28308.9	0.973	0
3	26917.3	0.959	0.024
4	25614.6	0.97	0.0105
5	24973.8	0.968	0.6997

Table 1: Mplus LCA results

The VLMR test uses Log-Likelihood and measures whether the model of *k* latent classes is a better fit compared to the model of *k-1* classes. We observed that while BIC continuously decreases as k increases, Entropy seems to reach a local maximum at *k=4* and the VLMR test is not significant at *k=5*. This suggests that the optimal latent classes are *k=4* due to the fact that BIC is decreasing, Entropy increases and the VLMR test is significant at *k=4*.

Feature Selection Using LASSO Regression

To identify the most predictive biomarkers associated with VAP subphenotypes, we applied LASSO logistic regression with stability selection. This method ensures that only the most important biomarkers are retained by applying L1 regularization, which shrinks the coefficients of less relevant features to zero. Additionally, we used bootstrap resampling (50 iterations) to assess the stability of biomarker selection, to make sure that the most frequently chosen features are truly significant rather than artifacts of a specific dataset split.

After merging the standardized biomarker dataset with LCA-based cluster assignments, we performed an 80-20 train-validation split and trained the LASSO model. The analysis identified five key biomarkers, with balf_TNF-RI_V1_imputed being the most stable predictor, appearing in 100% of bootstrap runs and showing a strong inverse relationship with mortality (coefficient = -0.43). Other frequently selected biomarkers included balf_IP-10_chemo_V1_imputed (76%), balf_TARC_chemo_V1_imputed (50%), balf_GM-CSF_V1_imputed (50%), and balf_IL-15_V1_imputed (60%), indicating their potential clinical relevance. Meanwhile, biomarkers such as balf_Amphiregulin_V1_imputed and balf_CD163_V1_imputed were rarely or never selected, suggesting limited predictive value in distinguishing patient subphenotypes.

To assess the effectiveness of our biomarker-based classification, we computed the Area Under the Curve (AUC) score on the validation set. The resulting AUC of 0.96 demonstrates high discriminative ability, effectively distinguishing clusters based on biomarker expression.

In upcoming statistical analyses, we will further investigate how these biomarkers correlate with hospital mortality and other key clinical outcomes.

Statistical Analysis

To evaluate whether our LCA-defined endotypes are associated with clinical outcomes, we performed statistical analyses on key variables, including hospital mortality, ventilator-free days (VFDs), and baseline severity scores.

Below is a heatmap of each of the latent classes identified using Mplus (Figure 10). We can observe that each cluster has a unique biomarke profile. Cluster 2 has a unique increase in chemokines and PD_L1 expression. Cluster 3 is notable for its decreased expression of cell markers across all 25 measured biomarkers. Cluster 4 is defined by its increased proinflammatory signature and cluster 1 is defined by its relatively average expression of cell markers.

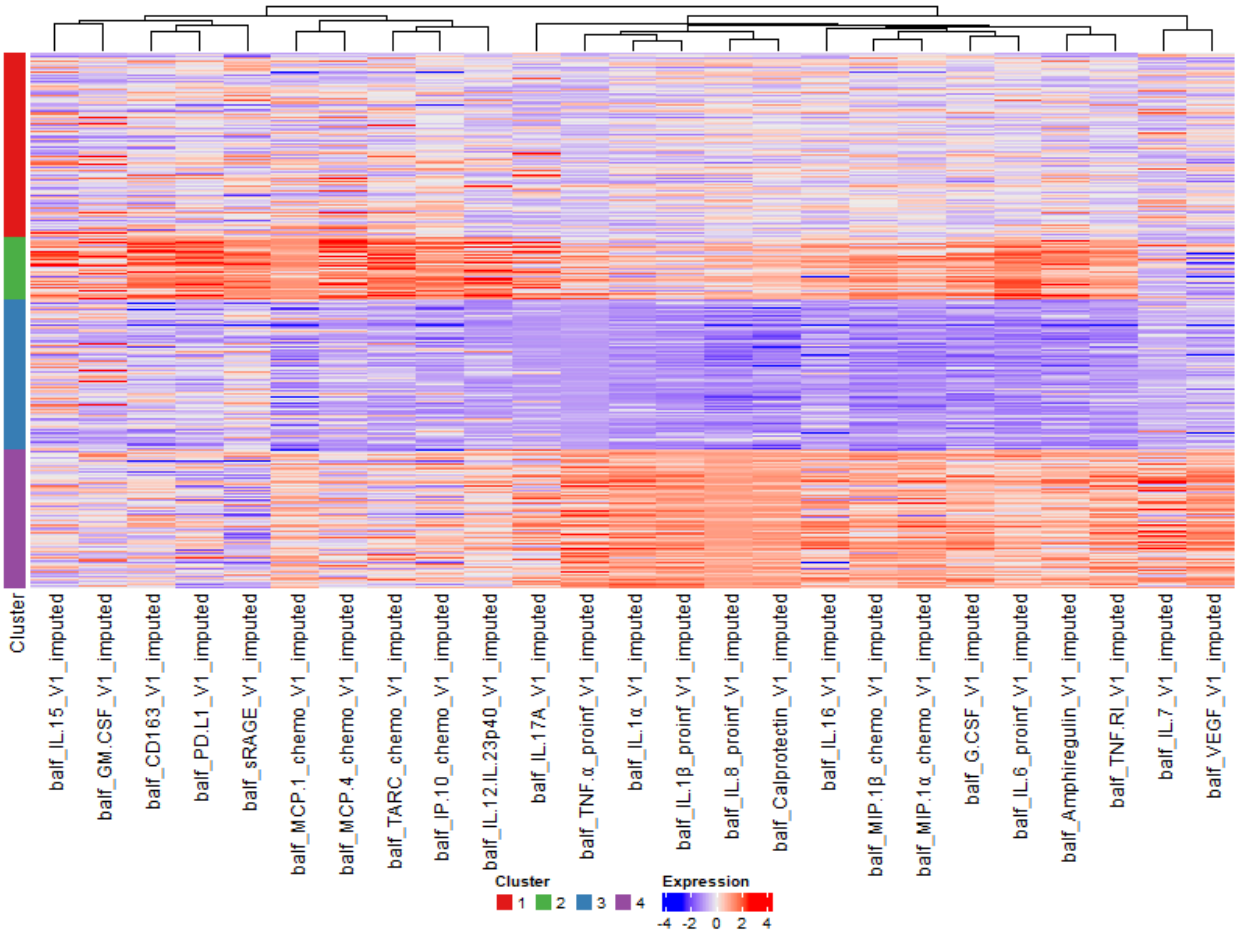

Figure 10: Cluster-Biomarker expression heatmap

Hypothesis Testing for Group Comparisons

We generated a summary table of key demographic and clinical variables, using the LCA-defined clusters, and applied statistical tests to assess differences across the groups. Continuous variables were analyzed using the Kruskal-Wallis test, while categorical variables were compared using the Chi-square test.

- Age differences were significant (p = 0.0028), with patients in Cluster 2 being the youngest (median age: 38.7 years), while those in Cluster 1 were the oldest (median age: 55.5 years). This suggests that age may play a role in how VAP progresses, potentially influencing recovery.
- Ventilator-free days (VFDs) varied across clusters (p = 0.048), with Cluster 4 having the lowest median (9 days). This means that patients in this group spent more time on mechanical ventilation and had fewer days without ventilatory support, suggesting more severe respiratory illness or slower recovery.
- Baseline SOFA scores also differed significantly (p = 0.003), with Cluster 2 having the highest median score (7.0),suggesting that these patients were more critically ill at ICU admission.
- Hospital mortality showed an upward trend across clusters, increasing from 16.2% in Cluster 1 to 27.3% in Cluster 4. However, the p-value (0.085) suggests that this difference was not strongly statistically significant, meaning that while a pattern was observed, we cannot confidently conclude that endotype classification alone predicts mortality risk.
- BMI differences were borderline significant (p = 0.061), with Cluster 1 having the highest median BMI (28.7) and Cluster 2 the lowest (26.3).
- Race and ethnicity distribution varied among clusters, with most patients being non-Hispanic (85.6%), while Cluster 3 had the highest proportion of Hispanic/Latino patients (11.5%).

Logistic Regression Analysis on Hospital Mortality

The results of the logistic regression analysis indicate that cluster membership alone does not significantly predict hospital mortality, as none of the cluster variables in the first model (Model 1: hospital_mortality ~ cluster) were statistically significant (p-values > 0.05), and the pseudo R-squared value was quite low (0.01406), suggesting a poor model fit. When Age and Sex were added in the second model (hospital_mortality ~ cluster + Age + Sex (adjusting for Age and Sex), the explanatory power improved slightly (pseudo R-squared = 0.07567), but cluster membership and Sex remained statistically insignificant. The only significant predictor in the second model was Age (p-value = 6.59e-07), with a positive coefficient (0.0386), indicating that an increase in age is associated with a higher likelihood of mortality. The intercept was also highly significant in both models, reinforcing the baseline odds of mortality. Overall, while the addition of Age slightly improved model performance, cluster and Sex did not have a meaningful impact on hospital mortality in this dataset.

Relative Risk Estimation Using Modified Poisson Regression

To assess whether LCA-defined clusters were associated with hospital mortality, we performed Modified Poisson Regression with robust standard errors to estimate Relative Risk (RR).

Model 1: hospital_mortality ~ cluster

- Cluster 2 was used as the reference group.
- Cluster 1 had nearly the same risk as Cluster 2 (RR = 1.12, p = 0.77), suggesting no meaningful difference.
- Cluster 3 had a 53% higher risk of hospital mortality (RR = 1.53), but it was not statistically significant (p = 0.24).
- Cluster 4 had the highest hospital mortality risk (RR = 1.88, p = 0.08), suggesting a potential but inconclusive association.

Model 2: hospital_mortality ~ cluster + Age + Sex (adjusting for Age and Sex)

- Each additional year of age increased the risk of hospital mortality by ~3% (RR = 1.03, p < 0.001), confirming that age is a strong predictor of hospital mortality.
- Male patients had an RR of 0.87, meaning their mortality risk was 13% lower than females, but this was not statistically significant (p = 0.47).
- Cluster 4 still had the highest mortality risk (RR = 1.59, p = 0.187), but it did not reach statistical significance.

Adding age and sex improved the model's Pseudo R^2 (from 0.011 to 0.059) and log-likelihood (-245.05 to -233.53), indicating a better fit.

Summary of VAP Patient Clusters

The analysis identified four distinct biomarker-driven clusters among Ventilator-Associated Pneumonia (VAP) patients, each with unique clinical and demographic characteristics. **Cluster 1** represented patients with relatively balanced biomarker expression, neither showing extreme inflammation nor immune suppression. This group had the oldest median age (55.5 years) and the highest proportion of non-Hispanic White patients. Clinically, they exhibited the lowest hospital mortality rate (16.2%), suggesting a more stable disease trajectory.

In contrast, **Cluster 2** consisted of the youngest patients (median age 38.7 years) and was characterized by elevated chemokines and PD-L1 expression, indicating an activated immune response. This group also had the highest SOFA scores at baseline, suggesting greater illness severity upon ICU admission. **Cluster 3**, on the other hand, displayed a blunted immune response with low expression across most biomarkers. This cluster had the highest proportion of Hispanic/Latino patients and a moderate hospital mortality rate of 22.3%, though the difference was not statistically significant.

Lastly, **Cluster 4** emerged as the most severe group, with elevated proinflammatory biomarkers suggesting a heightened systemic inflammatory response. Patients in this cluster had the lowest

ventilator-free days (9 days), indicating prolonged mechanical ventilation, and the highest hospital mortality rate (27.3%), though statistical significance was borderline. These findings highlight potential biological mechanisms influencing VAP progression, with Cluster 4 showing the worst clinical outcomes and Cluster 1 the most stable. Further research is needed to determine the clinical utility of these subphenotypes in guiding personalized treatment strategies.

Limitations

While this study provides valuable insights into Ventilator-Associated Pneumonia (VAP) subphenotypes, we do have several limitations:

1. **Biomarker Panel:**
    - The biomarker panel is limited to only 25 biomarkers. It is possible that a broader spectrum of biomarkers could identify additional latent classes that our panel could not.
2. **Small Sample Size**:
    - Certain subgroups had a limited number of patients, which may have reduced the power of statistical tests and increased variability in estimates.
3. **Potential Confounding Variables**:
    - Although we adjusted for age and sex in regression models, other unmeasured factors could still influence the results.
4. **Cluster Stability and Generalizability**
    - Clustering results may vary across datasets, and our findings need validation in external cohorts. The study was based on a single healthcare system, which may limit the applicability of results to other ICU populations.
5. **Limitations of LCA and Regression Models**
    - The probabilistic nature of Latent Class Analysis (LCA) introduces classification uncertainty, and the optimal number of clusters was based on statistical criteria rather than biological validation. Logistic regression estimates odds ratios, which may not accurately reflect risk, and while Modified Poisson Regression improves interpretation, it showed wide confidence intervals due to small sample sizes in some groups.
6. **Missing Data and Imputation**
    - Some biomarkers had missing values that required imputation, potentially introducing bias. Features with high missingness were excluded, which may have limited the scope of the analysis.
7. **Computational Complexity**
    - The methods used, including LCA and Gaussian Mixture Models, require significant computational resources, which may be challenging to scale for larger datasets or real-time clinical applications.

While we identified statistically significant differences between subphenotypes, further study should validate these findings in larger, multi-center datasets and explore additional biomarkers to determine whether these classifications improve patient management and outcomes.

Conclusion

Through our analysis, we identified four distinct molecular subphenotypes within the VAP patient population using unsupervised clustering methods. While not all clusters showed statistically significant differences in clinical outcomes, we observed meaningful variations in biomarker expression and relative risk for hospital mortality. The distinct biomarker profiles and borderline statistically significant varying outcomes warrant further investigation into these molecular endotypes to better understand the biological mechanisms behind VAP and how it affects the patient's hospital course.

Understanding these subphenotypes has the potential to bridge the gap between biomarker research and clinical application, offering a step toward more personalized treatment strategies in critical care. However, further validation in larger, multi-center cohorts is essential to confirm these findings and determine their clinical utility. By using data-driven approaches, our project highlights the promise of precision medicine in improving risk stratification and guiding more targeted interventions for critically ill patients.

Acknowledgments

We are deeply grateful to **Dr. Eric D. Morrell** for his invaluable mentorship and guidance throughout this project. His expertise in pulmonary and critical care medicine helped shape our research, and his recommendations on methodologies were instrumental in refining our approach. His feedback on our poster, along with his insights into the clinical aspects of our findings, greatly enhanced our ability to communicate our work effectively.

We also sincerely thank **Dr. Megan Hazen**, our capstone coordinator, for her continuous support and encouragement. Her regular check-ins ensured we stayed on track, provided clarity when needed, and helped us navigate challenges throughout the project.

Additionally, we appreciate the **members of Dr. Morrell's team** for their assistance in obtaining and processing the dataset, as well as their contributions to various aspects of the project. Their support played a key role in making this work possible.

A heartfelt thank you to our **friends and peers**, who took the time to review our work and provided valuable feedback. Their suggestions helped us refine both our poster and report, making them stronger and more effective.

Finally, we would like to thank **each other** for our dedication, teamwork, and persistence throughout this project. From tackling complex analyses to refining our presentation, we

supported and challenged each other to produce the best possible work. This project has been a collaborative effort, and we truly appreciate the learning experience we shared together.

References

[1] Kohbodi GNA, Rajasurya V, Noor A. Ventilator-Associated Pneumonia. [Updated 2023 Sep 4]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK507711/

[2] Sinha, P., Calfee, C. S., & Delucchi, K. L. (2021). Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls. *Critical Care Medicine, 49(1), e63–e79*. https://doi.org/10.1097/CCM.0000000000004710

[3] Calfee, C. S., Delucchi, K., Parsons, P. E., Thompson, B. T., Ware, L. B., Matthay, M. A., & NHLBI ARDS Network. (2014). Subphenotypes in acute respiratory distress syndrome: Latent class analysis of data from two randomised controlled trials. *The Lancet. Respiratory Medicine, 2(8), 611–620.* https://doi.org/10.1016/S2213-2600(14)70097-9

[4] Bhatraju, P. K., Zelnick, L. R., Herting, J., Katz, R., Mikacenic, C., Kosamo, S., Morrell, E. D., Robinson-Cohen, C., Calfee, C. S., Christie, J. D., Liu, K. D., Matthay, M. A., Hahn, W. O., Dmyterko, V., Slivinski, N. S. J., Russell, J. A., Walley, K. R., Christiani, D. C., Liles, W. C., … Wurfel, M. M. (2019). Identification of Acute Kidney Injury Subphenotypes with Differing Molecular Signatures and Responses to Vasopressin Therapy. *American Journal of Respiratory and Critical Care Medicine, 199(7), 863–872.* https://doi.org/10.1164/rccm.201807-1346OC

[5] Yin, H., Aryani, A., Petrie, S., Nambissan, A., Astudillo, A., & Cao, S. (Year). A Rapid Review of Clustering Algorithms. *Swinburne University of Technology and Australian National University*. https://doi.org/10.48550/arXiv.2401.07389

[6] Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004 Apr 1;159(7):702-6. doi: 10.1093/aje/kwh090. PMID: 15033648. 10.1093/aje/kwh090