**Project Proposal**

*Ted Liu, Parvati Jayakumar*

# **1. Introduction**

This project proposal outlines an important investigation into Ventilator-Associated Pneumonia (VAP) and its association with biomarkers, pathogen-specific profiles, and clinical outcomes. The collaboration with Dr. Eric Morrell at the University of Washington Department of Pulmonary, Critical Care, and Sleep Medicine addresses key questions regarding the potential existence of subphenotypes in VAP patients. This project aims to create a robust data pipeline, perform statistical analyses, and provide actionable insights that can influence clinical practices and patient outcomes. A more detailed explanation is provided below.

# **2. Problem Statement**

Ventilator-Associated Pneumonia (VAP) is a prevalent infection in critically ill patients, commonly linked to prolonged mechanical ventilation, increased hospital stays, and higher mortality rates. While certain pathogens associated with VAP have been identified, it remains unclear if they each exhibit unique biomarker profiles or distinctive pathophysiological patterns. Our project, in collaboration with Dr. Eric Morrell, a pulmonologist and critical care physician-scientist at the University of Washington Department of Pulmonary, Critical Care, and Sleep Medicine, seeks to explore these differences by (1) identifying if specific pathogens show unique biomarker signatures, (2) determining if subphenotypes exist within the VAP cohort, and (3) examining how these subphenotypes correlate with clinical outcomes, including ventilator-free days, severity of respiratory failure, and mortality.

Dr. Morrell has expressed challenges in processing clinical data into an analyzable format and implementing the rigorous statistical analyses required to answer these questions. Although he is experienced in pulmonary and critical care medicine, Dr. Morrell has limited training in computational and statistical methods, which are essential for addressing the aims of this project. The difficulty lies in ensuring that the project outcomes are both accurate and reliable, particularly since this research involves critically ill patients. Achieving a good level of accuracy and reliability is challenging due to the complexity of the data and the need for precise analysis to inform clinical decision-making effectively.

To meet these challenges, Dr. Morrell has requested assistance from our team with performing the necessary data engineering and statistical analysis to answer the questions identified in the project aims. Specifically, Dr. Morrell has asked for our help in developing an efficient, reproducible pipeline to generate datasets for analysis and performing rigorous statistical analysis of clinical data to address the project's key questions. To ensure accuracy and reliability, we will use validated statistical methods and advanced computational techniques, implementing rigorous data quality checks at every stage to eliminate errors and inconsistencies. Clear documentation and version control will be maintained throughout the process to ensure reproducibility and transparency. This collaborative effort will support Dr. Morrell's objectives, ensuring a robust and reliable investigation into pathogen-specific biomarkers, the potential for VAP subphenotypes, and their clinical impact.

# **3. Team Bios**

## **3.1 Ted Liu**

Ted holds a Bachelor’s degree in Computer Science with a minor in Biology. In the team, Ted is the dedicated communicator with Dr. Morrell - the stakeholder. His experience is primarily working in data analytics in the biomedical space. Currently, Ted works as a Research Data Analyst at the University of Washington Department of Medicine, where his work encompasses developing ETL pipelines to perform data analysis for publications in the pulmonary and critical care research space. Ted is skilled in SQL as well as R and Python tools more commonly used in the biomedical research space. He has an interest in applying computational and statistical methods within the context of medical research to further knowledge.

## **3.2 Parvati Jayakumar**

Parvati holds a Bachelor’s degree in Electronics and Communication, where she developed a strong foundation in signal processing and data analysis. Her early experiences working on speech processing models solidified her passion for applying technical skills to impactful fields, particularly healthcare. She subsequently joined a health tech company, where she analyzed patient health data to identify behavioral trends, enabling personalized care. Currently, she is a Data Science graduate student at the University of Washington, where her work focuses on designing data-driven solutions to enhance workflows and decision-making. Skilled in Python/frameworks and SQL, Parvati also has a strong command of data visualization tools such as Power BI and cloud solutions like Microsoft Azure, which she leverages to automate complex analyses and create scalable insights. She has a keen interest in medical data applications and is dedicated to furthering her expertise in statistics, machine learning, and data analysis.

# **4. Data Pipeline**

The data pipeline for our project focusing on electronic health records (EHR) is organized into two primary sections: ***Assembly*** and ***Analysis-Ready.*** This structure ensures efficient data handling among team members, particularly since one member works directly with the capstone project sponsor.

## **4.1 Pipeline Overview**

### **4.1.1 Assembly Section**

The Assembly section serves as the foundation of our data pipeline, functioning as an ETL process to extract, transform, and load necessary EHR data for analysis. This segment is exclusively managed by Ted and Dr. Eric Morrell due to the sensitive nature of personal health information (PHI) in EHR data. The primary goal is to extract, clean, assemble, and de-identify relevant data to ensure patient privacy is protected.

#### **Data Extraction:**

The process begins with SQL queries on the EPIC electronic health record system to retrieve key healthcare data, including:

* **Vitals**: Metrics like heart rate, blood pressure, and other physiological indicators.
* **Laboratory Results**: Comprehensive lab data essential for evaluating patient conditions.
* **Microbiology Cultures**: Data on infection controls and treatments, critical for clinical decision-making.

#### **Data Cleaning and De-identification:**

After the data is extracted, it undergoes a meticulous cleaning process using Python. This involves:

* Initial data scrubbing to eliminate inaccuracies or incomplete records.
* De-identification procedures that ensure all personally identifiable information (PII) is removed, thereby safeguarding patient confidentiality.

#### **Data Formats:**

The cleaned data is formatted into three distinct comma-separated formats:

1. **Long-Format Dataset**: This format captures daily data points for individuals, extending up to 28 days post-hospital admission. Each patient's data is organized chronologically, facilitating longitudinal studies.
2. **Wide-Format Dataset**: This format condenses the data for each patient into a single row, where various clinical measures are represented as columns (via date suffixes). It includes additional fields such as microbiology results and clinical outcomes such as Ventilator-associated pneumonia (VAP) and Acute respiratory distress syndrome (ARDS).

These datasets were developed based on the specific requirements outlined by Dr. Eric Morrell. The resulting files are securely stored on a server within UW Medicine, with access restricted to Dr. Morrell and Ted.

#### **Analysis-Ready Section**

The Analysis-Ready section marks the transition from semi-cleaned dataset to a fully prepared dataset ready for statistical analysis. This phase begins with the wide-format dataset, which has already undergone initial cleaning and de-identification in the Assembly section. At this stage, data features used in statistical analysis undergo ruther cleaning (e.g. ensuring correct data types, imputation) ensuring it is ready for advanced statistical and analytical workflows.

* **Final Cleaning:** Additional cleaning tasks are performed to meet the specific requirements of analysis protocols. This involves handling missing values, verifying data integrity, and ensuring consistent formatting across the dataset.
* **Data Structure:** The final dataset comprises many unique subjects, with each subject represented as a single row including their associated biomarkers, clinical measures, and outcomes, organized as distinct columns.
* **Statistical Analysis Preparation:** The cleaned dataset is prepared for statistical analysis using R and Python. Each software has specific strengths; R is particularly well-suited for statistical modeling, while Python excels in data manipulation and machine learning applications.

## **4.2 Data Streams**

As mentioned in the overview above, the primary data streams employed throughout this pipeline are sourced directly from the EPIC EHR system using SQL. Which is then loaded and transformed in Python into the long and wide datasets mentioned above. This structured querying method allows access to detailed and essential patient data,

This ETL process allows us to pull all the necessary and detailed data that allow us to achieve our goals in this capstone project.

## **4.3 Data Format**

In total, the analysis-ready dataset consists of **470 unique subjects**, creating a rich collection of individual data points for clinical analysis. Each subject’s data provides detailed insights into clinical measures and outcomes, offering a valuable resource for exploring health trajectories within the scope of our study. The dataset is structured in a **wide-format**, where each row corresponds to a single subject. Columns contain associated biomarkers, clinical measures, and outcomes.

## **4.4 Data Residency and Access**

The datasets are securely stored on a **dedicated server managed by UW Medicine and owned 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 has access to the server. He will assist Dr. Morrell in retrieving the data. Dr. Morrell will then share the updated datasets with both Parvati and Ted via email in Excel or CSV format.

## **4.5 Software Utilization**

The data workflow involves the following software tools for various processing stages:

* **SQL:** Used for querying the EPIC EHR system, facilitating effective data extraction.
* **Python:** Used for data cleaning, transformation, data visualization, and ML modeling; its libraries assist in achieving robust data manipulation.
* **R:** Used mainly for statistical analysis, allowing for advanced modeling and data visualization techniques.

So far, we have not encountered significant issues with data access or integrity. If challenges arise, Parvati, Ted, and Dr. Eric Morrell will work together and if needed, we will reach out to Professor Megan for help and find a solution.

# **5. Background Summary**

As background research, the group read articles on clustering methods. Parvati read a review on clustering algorithms while Ted read an article on a specific clustering methodology.

The review paper on clustering algorithms provides a general overview of the variety of clustering algorithms that exist and methods for evaluating said clustering algorithms. A key takeaway from the review is that there is no “one-size fits all” clustering method and that understanding the differences between each method is crucial. The review covers algorithms such as the well-known K-means clustering, which is a distance-based clustering methodology, to other clustering algorithms such as DBSCAN which are cluster based on density rather than distance. The review also discusses methods on evaluating the optimal number of clusters. Key to our project with Dr. Morrell are these evaluation metrics as we will need to be able to provide concrete evidence for the number of clusters. The review describes three evaluation methods - the Elbow Method, Silhouette Scores, and the Gap Statistic. Most likely we will implement all three metrics in our decision for the optimal number of clusters in our data.

The article on the specific clustering methodology - Latent Class / Profile Analysis is an overview paper whose targeted audience are physician-scientists rather than statisticians. Despite this, it provides a good summary and introduction to the clustering methodology. Unlike K-means and DBSCAN mentioned prior, Latent Class Analysis (LCA) is a finite mixture method where the clusters are the groups that are found through patterns in features of the data. This is a notable difference as the other methodologies are somewhat rooted in distance-based clustering. Outside of describing LCA, the article also goes on to emphasize the importance of having a clear and quality study design - laying out a 5 stepwise approach:

1. Generate Hypothesis
2. Data set-up
3. Estimate Models
4. Evaluate Models
5. Interpret Optimal Models

The article also warns and advises to keep an eye out for the “Salsa effect” which simply means that clusters are simply divided by magnitude of features (i.e. high vs. low) and should be carefully interpreted.

The background research our group has done provides additional information on the various clustering methods available for our project. From our readings we now have a better understanding of the variety of tools as well as how to properly use and evaluate them. We will be focusing majorly on implementing the K-means and LCA algorithms in our capstone project.

# **6. Project Deliverables**

For this project, we will have multiple intermediary deliverables, with the final goal being a well-rounded **research paper**. The deliverables are the following:

* First, we’ll create a data file that assigns each subject to a cluster. This file will show which group every subject belongs to, using methods like k-means or LCA.
* Next, we’ll put together a “Table 1” for these subgroups. It’ll summarize key details about each group, like demographics and clinical data. This table will follow the standard format usually seen in research.
* The third step is all about digging deeper with statistical tests, like t-tests, to see if there are real differences between the subgroups. For example, are some biomarkers higher in one group? Or does one group have worse outcomes, like higher mortality? We’ll look at the numbers, report the p-values, and include simple charts.
* Finally, all of this will come together in a research paper. The paper will be polished in stages so we can get feedback and make it better.

# **7. Proposed Schedule**

The project focuses on exploring Biomarkers Associated with Ventilator-Associated Pneumonia (VAP) through collaboration with Dr. Eric Morrell (UW Pulmonary). ​To ensure a structured workflow and adequately address each component of the project, we have developed a detailed schedule that outlines tasks, their due dates, dependencies, and responsible team members.​

We use an Excel spreadsheet to manage our project schedule and status because it suits our small team of two. It allows us to communicate directly with each other first and update the sheet ourselves. This keeps things straightforward, helps us stay aligned, and ensures real-time tracking of task statuses, due dates, and milestones.

## **7.1 Work Done to Date**

We had an important meeting with Dr. Morell on 10/25/2024 to kick off the project. During this meeting, Dr. Eric Morell provided a clear explanation of the project’s scope, expectations, and deliverables. We also established regular check-ins every Monday to keep the communication consistent and stay aligned on the project’s goals.

We began by acquiring the necessary clinical dataset from Dr. Eric Morell on 10/28/2024, which was a critical step for our project. After receiving the dataset, both of us thoroughly explored it to understand the variables, their types, distributions, and possible values. This exploration was completed by 11/5/2024. The next phase involved data preprocessing, which we completed by 11/10/2024. We cleaned the data, handling missing values, normalizing field formats, and resolving inconsistencies, making it ready for analysis. In parallel, Ted worked on setting up the preliminary code for Latent Class Analysis (LCA) clustering, and Parvati focused on the K-Means clustering algorithm, both of which were finalized by 11/15/2024. During this time, we also worked on updating the data pipeline to include new EHR pulls for bronchoscopy notes, which Ted successfully modified by 11/22/2024.

## **7.2 Schedule Tracking System**

To keep our task management clear and organized, we’ve divided our tracking system into two main categories: **Project tracking** and **Event tracking**.

**Project tracking** focuses on the overall structure of the capstone project. It outlines all key tasks, timelines, deliverables, and responsibilities, such as acquiring datasets, conducting analyses, writing reports, and project presentation. This section gives us a big-picture view, keeping us aligned with our long-term goals and ensuring all tasks support the core research objectives.

**Event tracking**, on the other hand, zeroes in on specific occurrences with immediate importance, like sponsor meetings, check-ins, or task deadlines. This part helps us stay on top of day-to-day priorities, facilitating effective communication and quick responses to any changes or developments along the way.

A detailed schedule can be accessed here: [Schedule\_Lung-Spelunkers](https://docs.google.com/spreadsheets/d/1VE4ter8T9P7Dyw9IoaM-TGWxcX9u4XD06o6ftDjTkio/edit?usp=sharing)

An important intermediate milestone is scheduled for **February 10, 2025**. By this date, we plan to complete and submit the first draft of our final research paper. This is an important step where we’ll ensure all our core findings and methodologies are clearly documented and ready for review by Dr. Morrell. It’s a measurable goal, as we can confirm progress once the draft is completed and sent for feedback.

# 