**Data Pipeline**

**Team Lung Spelunkers:** *Ted Liu, Parvati Jayakumar*

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.

# **1. Pipeline Overview**

## **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.

### **1.1.1 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.

### **1.1.2 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.

### **1.1.3 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.

## **1. 2 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.

# **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.

# **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. 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.

# **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.