1. MIMIC Version Used

For this project, I worked with the MIMIC-IV dataset, which is available through Google BigQuery. Access to the dataset was granted through PhysioNet after successfully completing the required CITI certification in human subjects research. I used the mimiciv_derived and mimiciv_icu schemas to query the data and create cohorts.

2. Conceptualization Based on Liu et al. (2022)

Following the approach in Liu et al. (2022), I extracted clinical and chemical variables from the 24-hour window prior to extubation ("weaning") in mechanically ventilated patients. I created a dynamic cohort using adult patients with available data from ICU stays, and derived the worst values of the day for variables such as arterial pH, PaO_2, $PaCO_2$, base excess, WBC, hemoglobin, platelets, creatinine, anion gap, heart rate, respiratory rate, MAP, SpO_2, body temperature, FiO_2, and the oxygenation index (PaO_2/FiO_2). I also calculated total urine output, IMV duration, and whether vasopressors or CRRT were used in the final 24 hours before weaning.

Furthermore, I derived additional variables mentioned in the study, such as the GCS score (minimum value), PEEP, tidal volume, and durations of antibiotic and CRRT therapy, computed as the count of distinct administration days. I used the variable weaning_success to define outcomes, labeling as **failure** any patient who was reintubated, died, or required non-invasive ventilation for 48 hours within the post-extubation period. All other cases were labeled as **success**. This allowed me to prepare the dataset for binary classification.

3. Differences in Extracted Data vs. Liu et al. (2022)

My final dataset differs from that of Liu et al. (2022) in several ways. First, although I included the Charlson Comorbidity Index and individual comorbidities, they were obtained from the *charlson* derived table rather than computed from raw ICD codes. Second, durations for antibiotic and CRRT therapies were computed based on the number of distinct days with recorded administration, rather than exact hours or complete therapy cycles.

Additionally, my cohort was restricted to patients with both laboratory and charted data available in *BigQuery*, potentially excluding some individuals from the original population due to missing values or table availability. The specific data extraction from *BigQuery* may also differ in patient composition and time coverage compared to the local MIMIC-IV version used by Liu et al. (2022), due to dataset updates or versioning.

All 35 variables used in Liu et al. (2022) were successfully extracted or approximated using the available *derived* tables in MIMIC-IV v3.1. However, the exact methodology

and data granularity may differ, potentially affecting comparability and model performance (see Appendix-Table-Comparison of variables.pdf).
These differences may impact generalizability, although the variables and clinical time window were conceptually aligned with the original study.

4. Reflection on CITI Certification and Ethical Data Use

Completing the CITI certification was a critical step in understanding the ethical principles required to work with sensitive patient data. The training reinforced the importance of privacy, anonymity, and responsible data handling. Throughout this assignment, I remained mindful of avoiding any attempts to re-identify patients or misuse their information. All analyses were conducted within the secure BigQuery environment using de-identified records. This project not only strengthened my technical skills in SQL and machine learning but also deepened my awareness of the ethical responsibilities involved in clinical data research.

5. Machine Learning Model Development and Interpretation

To complement the SQL-based data extraction, I developed an XGBoost classifier in Python using scikit-learn and the xgboost library. The model was trained on the extracted variables using an 80/20 train-test split with stratification by outcome. I handled missing values, converted the categorical sex variable to numeric, and adjusted the decision threshold to increase sensitivity to the minority class. Feature standardization was not applied, as XGBoost does not require it due to its tree-based structure.
However, despite the promising AUROC of 0.71, the model shows **very poor performance predicting the class 0 (weaning failure)**, with an F1-score of only 0.02. This could be due to the strong class imbalance in the dataset, where class 1 represents over 83% of the samples. In contrast, **Liu et al. (2022)** also used an XGBoost model, but applied **propensity score matching (PSM), missing value imputation, and more comprehensive preprocessing** techniques to balance their dataset before training. Their internal AUROC reached 0.80 and external 0.86, showing not only better predictive power but also better generalization. In my case, the model was trained on raw data with minimal preprocessing, which likely contributed to the bias toward class 1 and the reduced performance.