# Project Plan for Research Questions Based on MIMIC-III Data

## Overview

The project aims to investigate two critical research questions using the MIMIC-III dataset, a comprehensive clinical database of ICU admissions:

1. Is it possible to accurately predict mortality based on data from the first 24 hours in ICU?

2. Does admission to ICU over the weekend increase the risk of mortality?

The first question focuses on developing predictive models to identify mortality risk early using data collected within the first 24 hours of ICU admission. The second question examines whether weekend admissions contribute to a higher mortality risk, potentially indicating differences in care quality or resource availability.

The data set is MIMIC- III (Medical Information Mart for Intensive Care III) dataset, a large, freely available clinical database comprising deidentified health-related data from over 40,000 patients who stayed in critical care units at the Beth Israel Deaconess Medical Center between 2001 and 2012. It provides a comprehensive view of ICU admissions, including demographic information, vital sign measurements, laboratory test results, medications, caregiver notes, imaging reports, and patient outcomes, including post-hospital discharge mortality.

MIMIC-III is notable for its diversity and granularity, containing detailed hourly bedside measurements, medication data, and structured clinical observations, making it a valuable resource for research in critical care, epidemiology, and machine learning. The dataset was collected using two different electronic medical record systems (Philips CareVue and iMDsoft MetaVision), which introduces variations in data representation.

The MIMIC-III dataset has several benchmarks and research papers that can help to optimise the project plan. For mortality prediction, models like Random Forest have been applied to this dataset. Additionally, models like GraFITi have been used for multivariate time series forecasting. These previous studies suggest that combining machine learning techniques (e.g., Random Forest) with feature engineering and time-series analysis could improve the accuracy and relevance of the predictions.

machine learning (ML) models can effectively predict ICU mortality using early physiological and laboratory data. MIMIC-III and other datasets have applied models like Random Forests, Gradient Boosting Machines, and deep learning (e.g., LSTM). Similarly, studies have investigated "weekend effects" using logistic regression and survival analyses, often highlighting care disparities or staffing challenges on weekends.

## Methods

### 1. Data Preprocessing

* Data Extraction: Extract relevant variables from MIMIC-III, focusing on patient demographics, vital signs, lab results, and interventions from the first 24 hours in ICU.
  + For multiple daily measurements, summary statistics (mean, max, min) from the first 24 hours will be used to ensure consistency while capturing variability.
  + `vitals\_hourly`: Vital signs like heart rate, blood pressure, respiratory rate, temperature, and mean arterial pressure.
  + `labs\_hourly`: Important lab results like blood glucose, creatinine, lactate, and hemoglobin levels.
  + `gcs\_hourly`: Glasgow Coma Score components (gcseyes, gcsmotor, gcsverbal), which are crucial for assessing consciousness.
  + `pt\_weight`: Weight measurements (e.g., admissionweight, dailyweight) that can indicate changes in fluid status.
  + `output\_hourly`: Urine output, which is an essential indicator of kidney function.
  + `vasopressors` and `pv\_mechvent`: Information on vasopressor usage and mechanical ventilation settings to capture patient acuity
* Data Cleaning: Handle missing values, outliers, and inconsistencies, ensuring a clean dataset for analysis.
  + Due to the sparse nature of ICU data, advanced imputation techniques could be considered, like forward filling (for continuous monitoring data) or multiple imputation for lab results to minimize data loss while retaining the integrity of the dataset.
* Feature Engineering: Generate meaningful features from raw data, such as aggregate statistics (e.g., mean blood pressure) and categorical transformations (e.g., admission type).
  + Time-Series Feature Aggregation: Since the project focuses on the first 24 hours, calculate summary statistics (mean, median, maximum, and standard deviation) for each variable over this period.
  + Dynamic Features: Include changes or trends in vital signs, lab results, and other key metrics (e.g., slope of heart rate, glucose levels) over time to capture clinical deterioration or improvement.
  + Binary Indicators: Create binary features to indicate the presence or absence of certain interventions, such as vasopressor administration or mechanical ventilation, as they can be crucial predictors of mortality.
* Outcome Definition
  + Research Question 1: "Mortality will be defined as in-hospital death (`hospital\_expire\_flag`).
  + Research Question 2: "Mortality will include all hospital stays and not be limited to the first 24 hours, aligning with clinical outcome relevance.

### 2. Predictive Modeling (for Question 1)

* Model Selection: Build logistic regression, random forest, and gradient boosting models to predict in-hospital mortality based on 24-hour data.
  + (optional): XGBoost which has ability to handle complex interactions and missing data.
* Training and Validation: Split data into training and validation sets, using cross-validation to assess model performance.
* Evaluation Metrics: Use metrics such as accuracy, precision, recall, F1 score, and ROC-AUC to evaluate predictive accuracy.
* (optional) Incorporate deep learning models like GraFITi, LSTM/GRU Neural Networks or Transformer networks, which have shown success in time-series analysis for ICU data. These models can capture temporal patterns from the first 24 hours more effectively than traditional models.

### 3. Statistical Analysis (for Question 2)

* Data Segmentation: Identify admissions occurring on weekends versus weekdays.
* Comparison Analysis: Use logistic regression to assess the impact of weekend admission on mortality, adjusting for potential confounders like age, severity, and comorbidities.
* Statistical Tests (Logistic Regression with Adjustment): Apply chi-square tests for categorical variables and t-tests for continuous variables compare mortality rates between weekend and weekday admissions.
* Survival Analysis: Implement propensity score matching combined with Cox proportional hazards models (adjusted for relevant covariates). This approach controls for confounding variables and offers a robust time-to-death comparison between weekend and weekday admissions.
* Propensity score matching will not be applied due to concerns about covariate overlap and balance. Instead, logistic regression and Cox proportional hazards models will be used, adjusted for confounders such as age, comorbidities, and ICU severity.

### 4. Sensitivity Analysis

Assess how different factors (e.g., patient demographics, comorbidities) influence the models' predictions and weekend admission effects on mortality by varying the inclusion of different features (e.g., excluding certain vital signs or lab results).

## Timelines

### Step 1-2: Data Understanding and Preprocessing

- Extract relevant variables from the pre-processed MIMIC-III dataset.

- Understand the structure of key tables (`vitals\_hourly`, `labs\_hourly`, `gcs\_hourly`, `pt\_weight`, `output\_hourly`, `vasopressors`, `pv\_mechvent`).

- Address missing values and inconsistencies.

### Step 3: Exploratory Data Analysis (EDA)

- Conduct EDA to identify trends, distributions, and potential data quality issues.

- Visualize important variables and their relationships with mortality.

### Step 4: Feature Engineering

- Create new features based on the first 24-hour data (e.g., statistical summaries, dynamic features, binary indicators).

- Construct time-series features to capture trends over time.

### Step 5: Model Selection and Baseline Model Building

- Select suitable models (e.g., XGBoost, LSTM/GRU, Random Forest).

- Build and evaluate a baseline model using Random Forest with initial feature sets.

- Implement time-based cross-validation.

### Step 6: Model Training and Hyperparameter Tuning

- Train the selected models on training data.

- Conduct hyperparameter tuning using GridSearchCV or RandomSearchCV.

- Evaluate models using metrics such as ROC-AUC, precision, recall, and F1 score.

### Step 7: Weekend Effect Analysis

- Perform statistical tests to compare mortality rates between weekend and weekday admissions.

- Implement logistic regression models adjusted for confounders.

- Initiate Cox proportional hazards modeling for time-to-death analysis.

### Step 8: Model Evaluation and Refinement

- Evaluate all predictive models on the test set.

- Refine models based on performance metrics and adjust feature sets as needed.

- Validate the logistic regression and Cox models for the weekend effect.

### Step 9: Sensitivity Analysis and Robustness Checks

- Conduct sensitivity analysis to assess the impact of different feature sets.

- Ensure robustness by testing various scenarios and handling missing data differently.

### Step 10: Finalizing Results and Report Writing

- Finalize all analyses, visualizations, and interpretations.

- Start writing the final report, integrating findings, insights, and visualizations.

- Review and edit the report for clarity and coherence.

### Step 11: Flexible/Backup Week

- Use this week for any unexpected delays, additional refinement, or final touches before submission.