Name: Wong Yi Ting
Student ID: S2152880
Course: WQD7005 | Semester 2 | 2024/2025
Title: AI-Assisted Summary Report and Visualization
All code, data generation scripts, preprocessing steps, and visualization notebooks are available in the following GitHub repository: ***https://github.com/LorraineWong/WQD7005-Data-Mining-S2152880***

0. Introduction

This project presents an AI-assisted pipeline for simulating and analyzing synthetic inpatient monitoring records. The objective is to construct a dataset that captures clinical variability commonly observed in real-world settings, including differences in hospitalization duration, missing values, and recovery progression. Each synthetic patient is assigned a clinical scenario and monitored over a variable period ranging from 10 to 30 days, reflecting outcomes such as early discharge, transfer, or deterioration. Each daily record consists of six vital signs — oxygen saturation, heart rate, temperature, blood pressure, weight, and blood glucose accompanied by a clinical note written in natural language.

The dataset was generated through prompt-based interaction with large language models (LLMs), embedding realistic medical context and controlled variability. This structured and semantically enriched dataset serves as a foundation for downstream tasks such as exploratory data analysis, intelligent preprocessing, and AI-driven insight extraction.

1. Dataset Simulation using GenAI

The synthetic dataset was generated through a GenAI-powered simulation pipeline, where a language model (GPT-4o) was prompted to simulate daily inpatient monitoring records for 500 patients. Each patient was randomly assigned to one of clinical scenarios such as post-surgery recovery or acute deterioration determining their condition progression and monitoring duration.

Instead of enforcing a fixed 30-day window, durations were flexibly simulated between 10 and 30 days to reflect real-world variability. Each record included six key vitals — oxygen saturation, heart rate, temperature, blood pressure, weight, and blood glucose along with a clinical note written in a realistic medical tone. To enhance realism, missingness was intentionally introduced by omitting one vital sign on up to three random days per patient.

All model outputs were returned in JSON format, then parsed, validated, and compiled into a structured DataFrame. The cleaned dataset, containing 6,501 daily records across all patients, was saved as **generate_patient_dataset.csv**, forming the foundation for all downstream preprocessing and analysis.

A sample of raw generated records for Patient P0001 is shown below. The records include six core vital signs (e.g., oxygen saturation, temperature, blood pressure) alongside structured clinical notes reflecting day-by-day progression.

patient_id	timestamp	oxygen_saturation	heart_rate	temperature	blood_pressure	weight	blood_glucose	clinical_note
P0001	2025-01-01	96	82	37.2	128/82	70.5	112.0	Patient stable post-surgery. Vitals within normal range. Monitoring for signs of infection or complications.
P0001	2025-01-02	95	85	37.4	130/84	70.3	115.0	Mildly elevated heart rate and temperature. Monitoring for early signs of infection. Patient reports mild incision site discomfort.
P0001	2025-01-03	94	88	37.8	132/86	70.2		Temperature trending upward. Possible low-grade fever. Encouraging fluid intake and assessing wound for signs of infection.
P0001	2025-01-04	95	84	37.6	130/85	70.1	120.0	Temperature stabilizing. Patient reports improved pain control. Wound site appears clean with no significant erythema.
P0001	2025-01-05	96	81	37.3	126/82	70.1	108.0	Patient showing signs of steady recovery. Vitals normal. Early ambulation initiated and tolerated well.
P0001	2025-01-06	96	80	37.1	124/80	70.0		Patient progressing well. No signs of complications. Encouraged to continue light physical activity.
P0001	2025-01-07	97	78		122/78	70.1	110.0	Oxygen saturation and heart rate within optimal range. Patient denies any new symptoms. Monitoring continues.
P0001	2025-01-08	96	79	37.2	125/81	70.0	105.0	Patient stable with normal vitals. Wound healing as expected. No concerns noted during rounds.
P0001	2025-01-09	95	83	37.5	128/83	70.1	118.0	Slight increase in temperature and glucose levels. Monitoring closely for any deviations from recovery trajectory.
P0001	2025-01-10	96	80		124/80	70.0	112.0	Patient remains stable. Gradual improvement in overall condition. Continuing routine post-surgical care.
P0001	2025-01-11	97	78	37.1	122/79	70.1	106.0	Normal vital signs. Patient reports improved energy levels. Cleared for additional physical therapy sessions.
P0001	2025-01-12	97	76	36.9	120/78	70.2	104.0	Recovery continues without complications. Patient tolerating increased activity well. Discharge planning initiated.
P0001	2025-01-13	98	75	36.8	118/76	70.3		Patient exhibits consistent improvement. Discharge scheduled for tomorrow if no issues arise overnight.
P0001	2025-01-14	98	74	36.7	116/75	70.4	102.0	Patient discharged in stable condition. Recovery progressing as expected. Follow-up appointment scheduled in one week.

Figure 1. Sample of GenAI-generated monitoring records (Patient P0001)

2. Exploratory Data Analysis and Visualizations

Exploratory analysis was conducted to evaluate the dataset's structure, completeness, and clinical plausibility. An automated profiling report generated using the ydata-profiling package confirmed that all fields were correctly typed, with no structural issues. Among numerical features, missing values appeared primarily in blood_glucose (3.5%), weight (2.9%), and temperature (2.1%), consistent with the simulation's design.

A histogram of monitoring durations revealed that most patients were observed for 12 to 16 days, with a smaller number falling below 10 or above 25. This pattern aligns with the intended simulation logic while also demonstrating the generative model's natural variance. The presence of patients with fewer than 10 days supports the decision to allow flexible monitoring durations and enhances the dataset's realism.

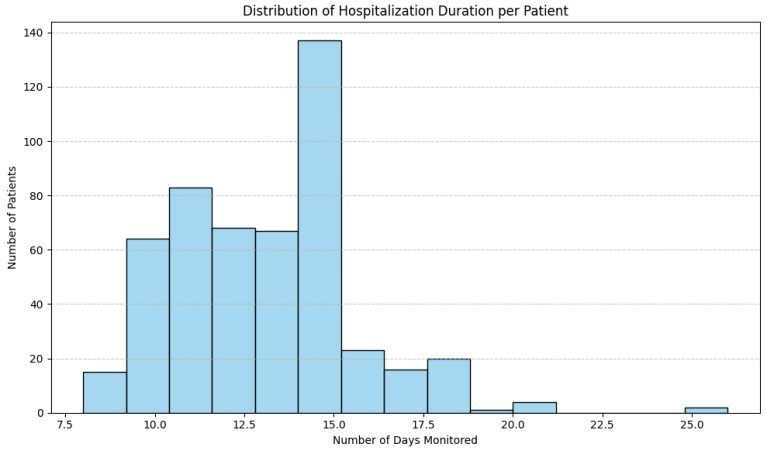


Figure 2. Distribution of hospitalization durations across 500 patients.

The distributions of vital signs such as heart rate, temperature, and blood glucose were unimodal and fell within clinically acceptable ranges. To further explore individual recovery patterns, time-series plots were created for a subset of patients (e.g., P0001–P0010), showing physiological changes over time and overlaid with shaded clinical reference zones. These visualizations helped distinguish between stable trends, recovery, or irregular fluctuations.

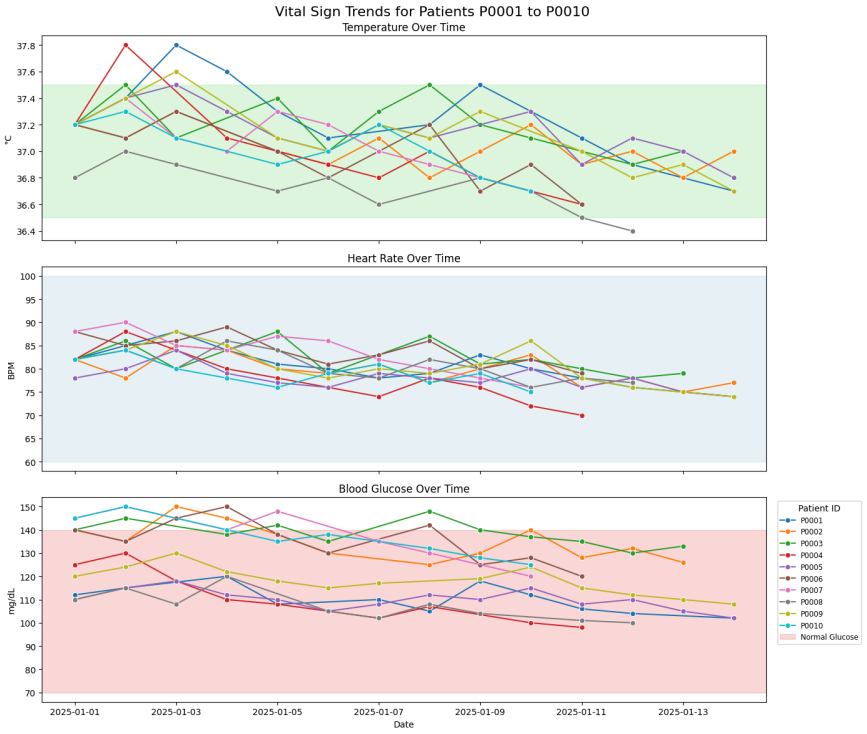


Figure 3. Time-series trend visualization for selected patients. Shaded areas indicate normal reference ranges.

In addition to statistical visualizations, Large Language Models (LLMs) were leveraged to generate concise narrative summaries of patient progression. By grouping daily clinical notes per patient and prompting the LLMs with structured instructions, each summary distilled the overall trend, key turning points, and final health status into 2–3 medically coherent sentences. This automated process yielded 500 summaries, each offering a semantically rich interpretation that paralleled observed numerical trends.

These model-generated narratives served as a complementary layer of insight, translating fluctuating values into interpretable clinical trajectories such as recovery stabilization or signs of deterioration without requiring manual annotation. The results were stored in **patient_summary.csv**, providing both human-readable context and additional features for downstream modeling or visualization.

For example, patient P0001 was assigned a post-surgery scenario, with corresponding recovery signals captured as follows:

"The patient demonstrated a steady recovery post-surgery, with vitals largely within normal ranges and no major complications. Early concerns of mildly elevated temperature and heart rate, potentially indicative of infection, were effectively managed, leading to stabilization and progressive improvement in pain control and physical activity tolerance. The patient was discharged in stable condition with recovery progressing as expected and a follow-up appointment scheduled."

Such summaries not only validated the simulation's realism but also created a semantically rich layer of interpretation, supporting downstream classification and visualization tasks.

3. Advanced Data Preprocessing with SLMs / LLMs

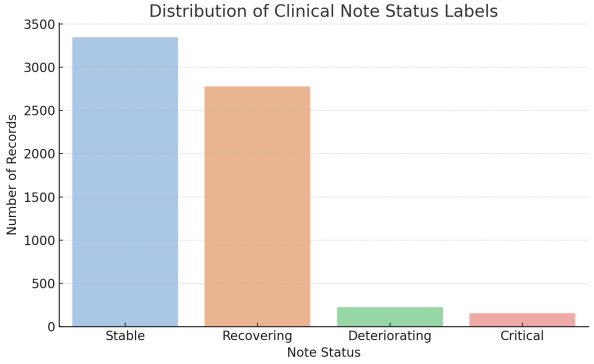
To prepare the dataset for downstream modeling and analysis, a structured pipeline was developed incorporating both statistical preprocessing and LLM/SLM-powered semantic enrichment. The process involved four main stages:

First, the composite blood_pressure field (e.g., "130/85") was split into two separate numerical features: systolic_bp and diastolic_bp. This enabled finer control in analysis and allowed these values to be normalized independently.

Next, to handle missing values resulting from GenAI simulation, patient-level median imputation was applied to five vital sign features: temperature, weight, blood_glucose, systolic_bp, and diastolic_bp. This patient-specific approach preserved intra-individual variation and avoided distortion from global aggregation.

Subsequently, Z-score normalization was applied to key numeric fields (temperature, heart_rate, blood_glucose, oxygen_saturation, systolic_bp, diastolic_bp, and weight) to ensure all features operated on a standardized scale. Original columns were dropped to reduce redundancy, with normalized versions retained (e.g., temperature_zscore).

Lastly, semantic classification of clinical notes was conducted using a zero-shot transformer model (facebook/bart-large-mnli). Each daily clinical note was mapped to one of four health statuses: Stable, Recovering, Deteriorating, or Critical. A total of 6,501 notes were processed without manual labeling, generating a new categorical feature note_status. These status labels were subsequently encoded numerically (via note_status_encoded) to support supervised learning tasks.



note_status	count
Stable	3346
Recovering	2775
Deteriorating	224
Critical	156

Figure 4. Distribution of Clinical Note Status Labels

The fully cleaned and enriched dataset was saved as **preprocessing_generate_patient_dataset.csv**, now containing 12 standardized columns across 6,501 entries. This dataset is fully ready for visualization, modeling, and further medical AI experimentation.

4. Conclusion

This project successfully demonstrates an end-to-end AI-assisted pipeline for generating, analyzing, and interpreting synthetic inpatient monitoring data. Through prompt-based simulation using GenAI, a semantically rich and structurally diverse dataset was created, reflecting real-world clinical variations in hospitalization durations, physiological trends, and data missingness. Instead of assuming uniform records, flexible monitoring periods were implemented to better mirror real discharge scenarios, enhancing realism and relevance.

Subsequent preprocessing using statistical methods and SLMs/LLMs-powered techniques enabled intelligent handling of missing values, categorical encoding, and clinical note classification. Exploratory visualizations and profiling confirmed dataset validity and illustrated physiological recovery trends over time. The integration of LLMs-generated summaries and zero-shot classifications added a semantic layer of interpretability, transforming raw records into high-level insights without requiring manual annotation.

By combining structured data engineering, visual analysis, and AI-driven narrative generation, this project not only mimics real EMR workflows but also offers a reusable framework for educational, research, and prototyping purposes in healthcare AI.