2022 COMP90089 Semester: Project Research Proposal – Group 10

Kevin Yuxuan Yang – 815565
Liang Min – 981061
Wei Leng – 636181
Yidan Hou – 1109120

1. Clinical Question

Research Topic: Predictions of **ICU length of stay** for adult patients who have shown **diabetic signs**.

Diabetes' high prevalence and the significant impact it has on hospital resources is a concern for healthcare planners. One of such is the long stay in the Intensive Care Unit (ICU), which is associated with poor patient outcomes, and increases the financial burden of care. The aim of this study was to examine the effectiveness of predicting Length of Stay (LoS) in ICU by comparing characteristics of patient admission to ICU with diabetic signs. In resource-constrained settings, these predictions may assist in reducing diabetic burdens through health policies and strategies.

2. Digital Phenotype

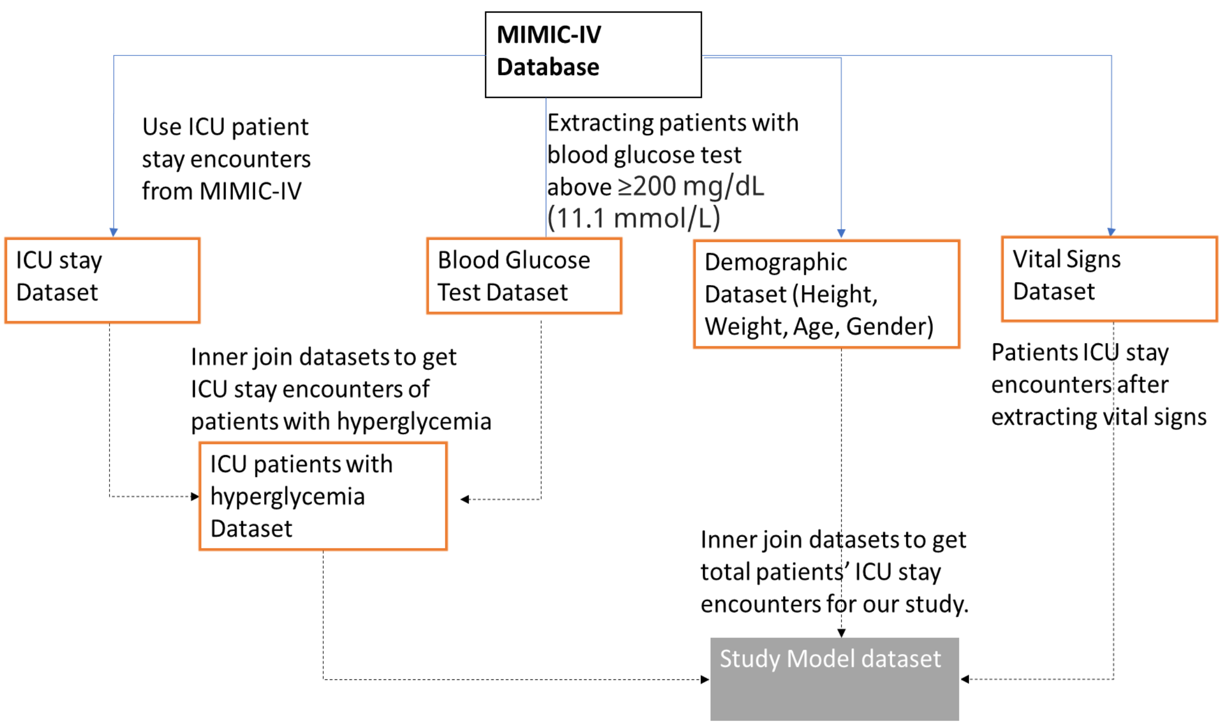

Figure 1: Digital phenotype from MIMIC-IV database (ICU: incentive care unit)

Our study consisted of extracting ICU stay data for adult patients from the Medical Information Mart for Intensive Care (MIMIC-IV) database before establishing Machine Learning (ML) Models that predict the ICU length-of-stay for diabetic patients from the data.

As illustrated in Figure 1, the phenotypic processes are described as follow:

1) We identify a patient's sign of diabetes characterized by ***hyperglycemia***[1], i.e. if a patient is admitted from ICU with a blood glucose test above ≥200 mg/dL (11.1 mmol/L) via lab tests.
2) The experiment will utilize ***vital signs*** that are later identified upon discovery (include but not limited to body temperature, heart rate, respiration rate, systolic blood pressure, diastolic blood pressure, and oxygen saturation [SpO2])
3) The model will include patient's ***demographic features*** (weight, height, age, and gender)

To ensure consistency of the data, we remove the following attribute(s):

1) Records with invalid vital signs
2) Patients' age less than 18

These data will undergo pre-processing stages i.e. one-hot encoding for multi-categorical variables, and normalization/standardizations between continuous variables to improve stability of ML model computations.

3. Method: Machine Learning Approaches

Whilst ICU length-of-stay can be expressed as a continuous variable in the unit of hours, we are also exploring unsupervised learning methods to divide outputs into specific numbers of bins, namely categorical variables. Both standard and state-of-arts machine learning algorithms are considered to establish effective predictions pertaining ICU patient's Hour-of-stays for those exhibiting diabetes mellitus.

For **continuous variable predictions**, regression models can be utilized to predict ICU length-of-stay. We plan to initiate a Linear regression model as a baseline, and followed by advanced tree-based models such as *Random Forest (RF) regression* and *XGboost*. Whilst *RF Regression* can generalize data regardless of the variety of feature sets, XGboost performs well when diabetic vital signs exhibit less noise.

If the experiment was deployed **with categorical variable predictions** (with categorical variables generated using unsupervised methods), we would then opt-in *logistic regression* as a benchmarking model, followed by RF and XGBoost classifier. The team may attempt to structure deep learning models, namely *Forward Neural-network (FNN)* and *Convolutional Neural-network (CNN)* to capture stable accuracies within the case study if time allows.

Depending on the size of the resultant queries, the experiment will utilize 90/5/5-split for **training, cross-validation, and testing** respectively for dataset more than 300k instances.

Train different models on the remaining data with *K-fold cross validation*, then evaluate each model on test data.

We are expecting Deep Learning models generally will outperform the rest given their ability to capture non-linearity and high-dimensional patterns, whilst the baseline model will be the least accurate in terms of accuracy. A range of evaluation metrics will be such as *Residual Sum-of-Square (RSS)*, utilized to distinct quality of the regression models, and Areas-Under-Curve (AUC), *F-scores* and – on the other hand – will be utilized for classification models and data imbalance will be checked against testing data.

[1] **Hyperglycemia**: *Classic symptoms of hyperglycemia (thirst, polyuria, weight loss, blurry vision) and has a random blood glucose value of 200 mg/dL (11.1 mmol/L) or higher (Inzucchi & Nathan, 2022).*

Reference:

Inzucchi, S., & Nathan, D. (2022). *Clinical presentation, diagnosis, and initial evaluation of diabetes mellitus in adults*. Uptodate.com. Retrieved 22 September 2022, from https://www.uptodate.com/contents/clinical-presentation-diagnosis-and-initial-evaluation-of-diabetes-mellitus-in-adults?search=diabetes&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=.