# Data Analysis Report: Diabetic Patient Dataset

## 1. Introduction

The dataset consists of medical information of diabetic patients, including various health metrics and demographic details. The main objective of this analysis is to explore the data, clean it, and extract insights that can potentially help in predicting readmission rates and understanding patient behavior.

## 2. Data Collection and Overview

The dataset is structured with columns representing patient demographics, medical history, test results, and hospital visit details. The key columns include:

* **Age**: The age of the patient.
* **Gender**: The gender of the patient.
* **Time in Hospital**: Number of days the patient stayed in the hospital.
* **Lab Procedures**: The number of lab procedures conducted.
* **Readmitted**: Whether the patient was readmitted to the hospital after discharge.

## 3. Data Preprocessing and Cleaning

Before diving into analysis, the following steps were performed to clean the data:

### 3.1 Handling Missing Values:

* **Max Glu Serum & A1Cresult**: These columns had missing values which were filled with the string 'No Test' to indicate that no test was performed.
* **Race, Payer Code, Medical Specialty, Diagnoses (diag\_1, diag\_2, diag\_3)**: Missing values in these categorical columns were filled with the most frequent value (mode) in each column.

### 3.2 Data Type Conversion:

* **Age**: The 'age' column, initially represented as ranges (e.g., '[0-10)'), was converted to numeric values by extracting the lower bound of each range.
* **Num Lab Procedures**: This column was converted to integers, and any non-numeric values were replaced by the median value of the column.
* **Gender & Race**: Both columns were converted to categorical types to optimize memory usage and ensure correct analysis.

### 3.3 Removing Outliers:

Outliers in the numerical columns such as num\_lab\_procedures were either handled through imputation or transformation, as needed, to reduce their influence on the analysis.

### 3.4 Dropping Irrelevant Columns:

The 'weight' column was removed from the dataset as it contained a large number of missing values and did not provide significant value to the analysis.

## 4. Exploratory Data Analysis (EDA)

### 4.1 Univariate Analysis

#### ****Distribution of Age****:

The age distribution primarily spans between 50-70 years, with a relatively uniform spread. A few patients were found to be younger or older, but they represent a small fraction of the dataset.

#### ****Readmission Status****:

The readmitted column contains three categories:

* NO (No readmission)
* >30 (Readmission after 30 days)
* <30 (Readmission within 30 days)

A large proportion of patients fall into the NO category, indicating that they were not readmitted within 30 days.

#### ****Distribution of Lab Procedures****:

The num\_lab\_procedures column reveals that most patients undergo fewer lab tests, while some patients require a higher number of procedures, potentially indicating more severe medical conditions.

### 4.2 Bivariate Analysis

#### ****Age vs Readmission****:

There doesn't appear to be a strong relationship between age and the likelihood of readmission. The age distribution remains fairly consistent across the different readmission categories.

#### ****Num Lab Procedures vs Time in Hospital****:

There is a noticeable positive correlation between the number of lab procedures and the length of hospital stays. This suggests that patients undergoing more tests tend to stay in the hospital longer, possibly due to the complexity of their conditions.

#### ****Correlation Between Numerical Features****:

The correlation matrix revealed a strong relationship between the num\_lab\_procedures and time\_in\_hospital, which supports the previous finding that more tests are linked to longer hospital stays.

## 5. Feature Engineering and Data Preparation

### 5.1 Feature Engineering

Several new features were created from the existing ones:

* **Age Groups**: The age column was divided into age groups (e.g., '0-20', '20-40', etc.) to capture potential age-related trends more effectively.
* **Binned Lab Procedures**: The num\_lab\_procedures column was categorized into three groups (Low, Medium, High) to make it more interpretable.

### 5.2 Data Transformation

* **Scaling**: Features like num\_lab\_procedures and time\_in\_hospital were normalized using **Min-Max scaling** to ensure that all numerical features are on a similar scale. This is crucial for machine learning models that require features to be within the same range.
* **One-Hot Encoding**: Categorical variables such as gender and race were one-hot encoded to convert them into numerical representations suitable for machine learning algorithms.

## 6. Key Insights and Findings

### 6.1 Age and Readmission:

While age does not show a strong correlation with readmission status, factors such as medical conditions, treatment procedures, and duration of hospital stay appear to play a more significant role.

### 6.2 Lab Procedures and Hospital Stay:

There is a clear positive correlation between the number of lab procedures and the time spent in the hospital. Patients with more procedures tend to stay longer in the hospital, which could indicate more complex health issues.

### 6.3 Data Quality:

The preprocessing steps ensured that missing data was handled appropriately, and all categorical variables were correctly encoded. After cleaning, the dataset is ready for further modeling.

## 7. Conclusion

This analysis provided valuable insights into the relationships between patient demographics, medical procedures, and the likelihood of readmission. By cleaning the dataset, transforming features, and performing exploratory analysis, we laid the foundation for future modeling and predictions.

### Recommendations:

* **Focus on Readmission Prediction**: Further modeling can be done to predict readmission likelihood, using features like the number of lab procedures, time in hospital, and possibly patient age groups.
* **Further Exploration**: Other potential features like patient history, medications, or insurance details could be valuable for improving predictions.

## 8. Next Steps

* **Predictive Modeling**: Use machine learning techniques (e.g., logistic regression, decision trees, or random forests) to predict patient readmission.
* **Advanced Feature Engineering**: Explore additional features or data sources (e.g., medical history) to enhance model accuracy.