# Project : Healthcare - Persistency of a drug

## Week 8 Deliverable

**Batch**: LISUM38

| Team: Salus AI Data Avengers | | | | |
|---|---|---|---|---|
| **Name** | **Email** | **Country** | **Company** | **Specialisation** |
| Amrin Shaikh | amrin02@gmail.com | United Kingdom | Freelance | Data Science |
| Keilor Fallas Prado | kfallasprado@gmail.com | Costa Rica | Immune Technology Institute Madrid | Data Science |
| Hyejoon Lee | candy0543@gmail.com | United Kingdom | -- | Data Science |
| Raina Singh | rainasinghh1@gmail.com | United States of America | The Ohio State University | Data Science |

## Problem description:

The primary objective of the project is to analyse patient persistency data and identify the factors influencing adherence to medication. The data contains various demographic, clinical, and risk-related attributes, which will be leveraged to build a robust classification model. The ultimate goal is to provide insights into persistency patterns and address data quality challenges, such as handling missing values, outliers, and skewness, to ensure accurate predictions.

## Data Understanding

### Dataset Overview

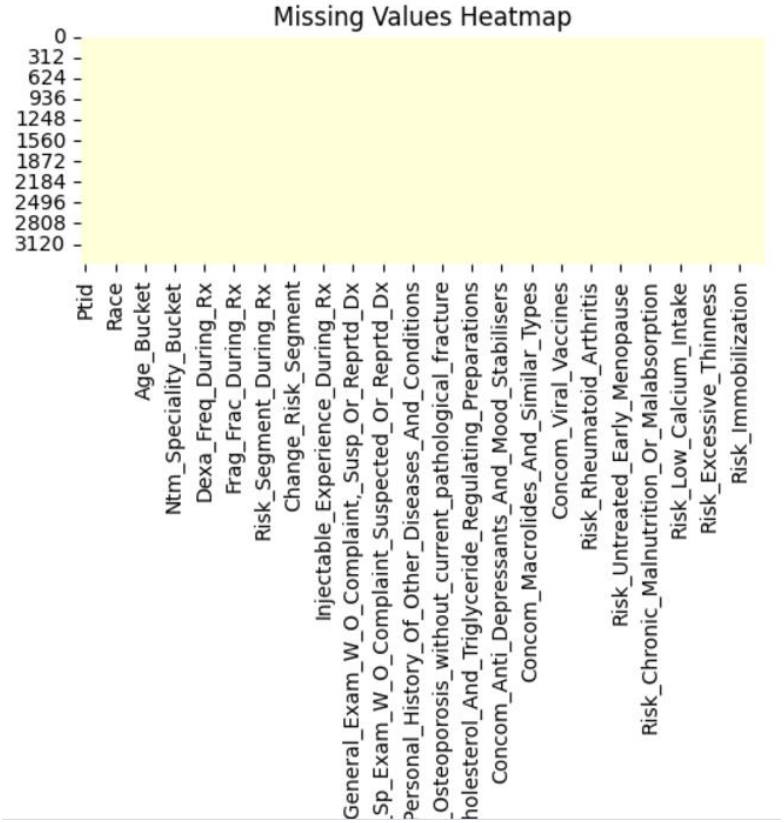
- **Dataset Size:** 3424 rows × 69 columns
- **Type of Data:** The dataset contains a mix of numerical and categorical features:
  - **Numerical Columns:** Dexa_Freq_During_Rx, Count_Of_Risks
  - **Categorical Columns:** Gender, Race, Ethnicity, Age_Bucket, Region, and binary risk/comorbidity indicators.
- **Key Columns:**
  - **Persistency_Flag:** Target variable indicating whether a patient is adherent (Yes/No).
  - **Dexa_Freq_During_Rx:** Frequency of DEXA scans during treatment, providing an indicator of patient monitoring.
  - **Count_Of_Risks:** Count of associated health risks, which may influence medication adherence.

- **Age_Bucket, Gender, Race, Ethnicity, Region:** Demographic data to assess adherence trends across populations.
- **Risk and Comorbidity Features:** Various binary indicators representing health conditions and risks associated with the patient.

## Issues in the Data

1. **Null Values:**

   There are no missing values in the dataset. All columns have 0 missing values, which is a positive aspect as it reduces the need for imputation or removal of data.




2. **Outliers:**

- **Dexa_Freq_During_Rx**: This numerical column exhibits outliers, as evidenced by a wide range (max value = 58). Extreme values are present, which may distort statistical models.
- **Count_Of_Risks:** While there are outliers, their impact is less severe compared to Dexa_Freq_During_Rx.

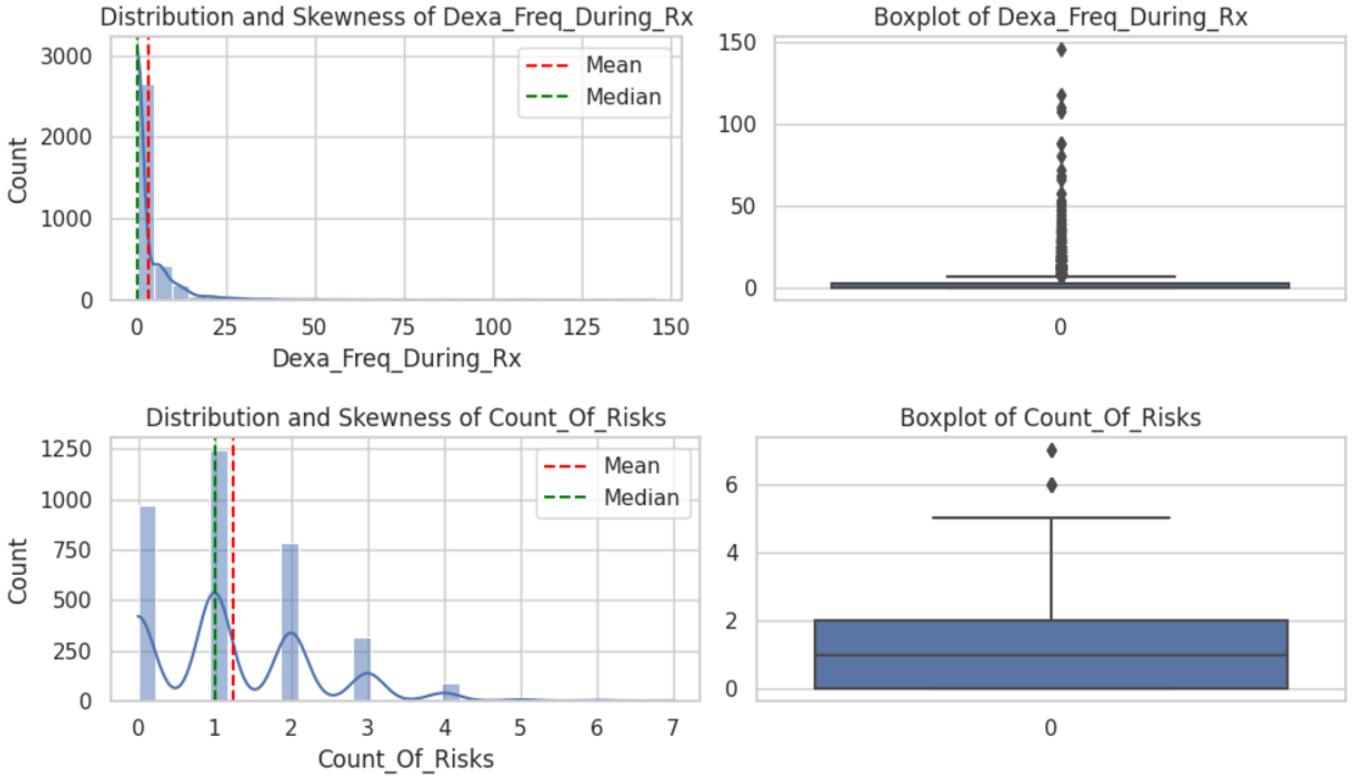
3. **Skewness:**

- **Dexa_Freq_During_Rx:** Highly right-skewed.
- **Count_Of_Risks**: Moderately right-skewed.

4. **Data Quality Challenges:**

- **Ptid:** High This column has high cardinality, with 3,424 unique patient IDs. It likely does not add predictive value and can be removed.

- **Ntm_Speciality**: There are 36 unique categories in this column. This may require binning or encoding strategies (e.g., one-hot encoding) to convert it into a usable format for modelling.
- **Persistency_Flag:** The target variable might have class imbalance (e.g., a higher number of "Yes" or "No" values), which will require techniques like resampling or adjustments to model algorithms to ensure a balanced outcome.




# Proposed Approaches

### 1. Handling Outliers

Method:

- For Dexa_Freq_During_Rx, apply the IQR (Interquartile Range) method to cap extreme values, ensuring outliers do not skew model predictions.
- For Count_Of_Risks, use Z-score transformation to handle any mild outliers, ensuring they do not disproportionately affect the model.

Reason: Outliers can skew statistical estimates and lead to model overfitting. Capping and transforming the data ensures that the model is not influenced by extreme values.

### 2. Addressing Skewness

Transformation:

- For Dexa_Freq_During_Rx, apply a logarithmic transformation to reduce right skewness, bringing the distribution closer to normality.
- For Count_Of_Risks, apply a square root transformation to moderate the right skewness.

Reason: Transforming skewed data helps improve model performance, as many algorithms assume data to be normally distributed. Reducing skewness allows the model to learn patterns more effectively.

3. **Data Encoding:**
    - For Categorical Variables like Gender, Race, and Ethnicity, apply one-hot encoding to convert them into a numerical format suitable for the model.
    - For Ntm_Speciality, apply binning or one-hot encoding, depending on the uniqueness and significance of categories.

4. **Dealing with Class Imbalance:**
    - If an imbalance exists in the Persistency_Flag target variable, resampling techniques (like SMOTE) or algorithm-level adjustments (e.g., class weights in models) can be used to ensure that the minority class is adequately represented.

## Summary

This project aims to predict medication persistency by analysing factors such as demographic characteristics, clinical attributes, and health risks. By addressing data quality issues such as outliers, skewness, and encoding challenges, we will build a robust classification model. The insights derived from this analysis will help healthcare providers understand the factors influencing medication adherence and improve patient outcomes