

# ROCL Group - Data Science Specialization (Healthcare Drug Persistence)

## Members:

Reem Aboelsoud, Germany, niologic GmbH  
reem.abulsoud@gmail.com

Oana Cucu, United Kingdom, University of Bristol  
omcucu@gmail.com

Leslie Hadaway, Saint Vincent and the Grenadines, Trinity School of Medicine  
lihadaway@outlook.com

Cindy Huynh, United States of America, University of Texas at Austin  
cindytrinhhuynh@gmail.com

## Problem description

The ABC pharmaceutical company aims to conduct an investigation into the determinants that affect the longevity of a drug, specifically focusing on the persistence rates among patients who are prescribed the treatment for a particular medical condition. The data available for analysis encompasses patient demographics, risk factors, pre-existing conditions, and comorbidities, along with limited details about the attending physician (such as specialization in a specific field or being a general practitioner). No specific information pertaining to the drug under investigation is provided.

To build a model of persistence predictors based on the available data, a classification approach can be applied. The data can be divided into input features (predictors) and the target variable (persistence).

A possible classification of the data:

### 1. Input Features (Predictors):

- Patient demographics: Age, gender, ethnicity, etc.
- Risk factors: Smoking status, obesity, family history, etc.
- Underlying conditions and comorbidities: Diabetes, hypertension, cardiovascular diseases, etc.
- Physician information: Specialist or general practitioner, experience level, etc.
- Psychosocial factors: Patient beliefs, attitudes, social support, healthcare provider-patient interactions, etc.

### 2. Target Variable:

- Persistence: This can be represented as a binary variable, indicating whether a patient continued or discontinued the prescribed treatment. For example, "1" can represent continued use, and "0" can represent discontinuation.

Once the data is classified into predictors and the target variable, we can apply various classification algorithms to build a model – a commonly used algorithm is logistic regression.

The model-building process involves training the model using labelled data (where persistence outcomes are known) and then evaluating its performance using appropriate metrics such as accuracy, precision, recall, and F1 score. Additionally, techniques like feature selection, cross-validation, and hyperparameter tuning can be employed to optimize the model's performance.

By applying this classification approach to the data, we can develop a model that predicts drug persistence based on the provided predictors, which can provide valuable insights and assist in decision-making processes related to treatment adherence.

Github repo link:

<https://github.com/lihadaway/data-glacier-internship/tree/c191138ff6dae50782971f90893e0b8791ef70d1/drug-persistency-classification>