## Introduction

The project aimed to support doctors in identifying which factors are most predictive of developing disease and discovering patients at high risk, enabling earlier intervention. The analysis involved data cleaning, exploration, and developing a predictive model using key clinical variables.

## Analysis Summary

### Data cleaning

The data was cleaned analysis. Column names were corrected for consistency. Nineteen fully duplicated rows were removed. Multiple entries for the same patient ID were found where all columns matched except for 'Blood_Chemistry_III'; these were retained, as they may reflect repeated measurements important to the analysis.
Missing values in clinical features were replaced with median values.

### Exploratory analysis

The outcome was imbalanced, with more patients without disease than with disease. Bivariate analysis (comparing each feature to outcome) and correlation analysis revealed that Blood_Chemistry_I, BMI and Pregnancies were the strongest predictors of disease. The exact nature of the variable Blood_Chemistry_I is not specified in the dataset, but it may represent a key laboratory marker such as blood glucose, enzyme, or other routine blood test. Further clarification of this variable's clinical meaning would be valuable for medical interpretation. These features were used for model building, with the others considered supportive.

### Modelling

A supervised learning approach using a Random Forest classifier was employed. The model used only the most predictive clinical features and was evaluated on a holdout test set using accuracy, recall and ROC-AUC.
The model achieved high performance with an accuracy of 99.9%, recall of 100% and ROC-AUC of 0.999.
For best practice, the workflow was checked for data leakage to ensure only clinical features were included.

## Recommendation

Based on the analysis, Blood_Chemistry_I is the most predictive marker for disease risk in your patients. However, BMI and Pregnancies also contribute to disease prediction. These factors can be used in early screening and further testing to support early intervention and improve patient care.