# Bariatric dataset proposal

Tim Beishuizen – Computational Biology - Eindhoven University of Technology

## Background data

The number of bariatric surgeries is increasing worldwide. Although initially thought otherwise, this type of surgery has added benefits on top of losing weight. An example of those benefits the remission of metabolic co-morbidities can be named. Bariatric surgeries are done by the Catharina Hospital in Eindhoven and these surgeries are monitored by gathering data from both manual and blood testing. Ruben Deneer did research on these available data sets and tried to retrieve a severity score of co-morbidities with this data. There may however be many other relations that can be retrieved from the data set, which are not found before.

## Proposal

There might be relations in the data sets that are worth exploring. These relations may be unknown beforehand, however can become visible after exploring the data sets. With this data driven approach, new insights can be discovered that can aid the medical staff with diagnosing diseases. Several different approaches can be made to discover these insights in a new manner. Two of them will be focused on.

* Comparing Populations: Are there differences between populations with regards to both variables and co-morbidities?
  + Example: Women of age 20-30 may need other markers to diagnose hypertension than women of age 30-40
  + Example: Men without dyslipidemia could have a different success rate for tackling obesity than men with dyslipidemia.
* Feature exploration: Are there features that are highly related to each other that can be useful?
  + Example: A different combination of the blood markers of the lipid panel might give better prediction of dyslipidemia than just weighing them and adding them together.
  + Example: An improvement of diabetes may be related to a series of computations of the available variables instead of mainly focusing on the marker HbA1c.

The insights on those approaches are “data driven”, which means that results are generated with the data instead of with literature. These results afterwards can either be explained with literature or a new research can be made to further investigate them. Multiple techniques can be used to obtain those insights, such as:

* Principal Component Analysis (checks how relevant each feature is for use)
* Supervised learning (tries to predict an outcome with knowing previous examples)
* Unsupervised learning (clusters data that are closely related to each other)

## Exploration details

* The project has a very wide scope to find possible relations
* After the project recommendations can be done for further projects
* The project should take about 1 or 2 months, after the data is analysed by multiple analysis techniques.