# Discussion 2017-11-27

A meeting with Ruben Deneer was held to discuss his master’s project.

**Yes/No translation from Diabetes criteria**

The follow-up labels were very unclear and subjective for different doctors. They were retrieved from texts from the doctors. Text mining might be possible if made anonymous.

**Extrapolation for DATO set**

Extrapolation is something that can be explored. More information can maybe be retrieved, about 10% is the estimation.

**3 months difference between lab/DATO set**

Doctors said that was fine

**Addition of 36, 48 and 60 months**

Is possible to look at, however not useful when splitting in population.

**fT4 and MMA**

Both are not combined with biomarker that made testing required, however can maybe be done

**Measurements only pre-surgery**

These were tested and gave proper results. Especially magnesium. However doctors don’t measure that one because of knowledge, so it is hard to have them measure it.

**Did you try some testing with LDL? It is a special combination**

Not tested, as multicollinearity is a thing.

**Looking at co-morbidities separately**

Because severity of the patient was needed, they were combined. Looking at all three separately though can be done and adding them afterwards.

**White-box reasons**

White boxes are used to explain the hospital. Black boxer are possible too, however it is hard to explain why they worked as they should.

**Combinations variables checked**

Mainly from known combinations “knowledge driven”. Not from the data, “data driven”

**Sub-populations other than time and co-morbidity**

Is it worth it to look at other variables (age, gender, etc.), Ruben agreed

**Nomogram**

The nomogram is a known method to make it clear for doctors how to interpret the outcome. It is shown just for clarity purposes.

**Feature selection**

More insight would be useful when looking at feature selection. A good balance should be made between number of variables and precision.

**Logistic regression**

Logistic regression was hard to use because of all the correlations. A Mixed Linear Effect model would seem better.

**Main challenge**

Find a good balance between knowledge driven research and data driven research. What is more important and why?