# Discussion 2017-11-16

A presentation was given based on four parts that were discussed. Discussions are briefly mentioned in these minutes

**Data Analysis Report**

Logistic regression, as used by Deneer, should be mentioned in here.

**Deneer’s thesis discussion**

The discussion of my proposal to do after Deneer’s thesis had a couple of major points all around:

* Important in preprocessing was watching subpopulations. What should be included and what should be excluded? When should extrapolation be used? What effect should the distribution type have for such decisions? Does a nearest neighbor approach work in these instances? Which variables are relevant and which are not? This could be explained more thoroughly
* When making bold statements, a good reasoning should be behind it, why. An example would be using ‘R’ as a tool.
* When criticizing certain parts, such as the nomogram and the ordinal system, better proposals should be made how to improve it.
* A good way to find combinations of variables that are related to a model could be found with principal component analysis (PCA).
* A good way to go from classification to regression is looking into Mixed Linear Effects (MLE). This should definitely be looked into.

**Research proposal**

One big remark is that the focus would be better to do on data analysis. Not everything in a biomedical research is important to include in here.

* The final idea is to create a workflow that explains what to do best in what situation. This includes all previously discussed parts.
* A better focus is advised for the framework. At the start this focus should mainly be on how to preprocess data within the data driven aspect. Also the focus should be more on data mining instead of on modelling, when looking at target driven aspects.
* An important inclusion to the data driven aspect is handling populations. Looking at sub-populations or between sub-populations is something that should be considered.
* An additional research question should be added to also add small tutorials or tests that give the biomedical engineer a good idea of whether both the technique and data they are using are applicable in that case or not.
* The program driven aspects should be called tool driven, to avoid confusion
* The program (tool) driven aspect is split in two parts, using existing tools and creating possible extensions. These two should be split so possible holes in the framework can be filled with those extensions.
* At the knowledge part, knowledge of third parties should be taken into account as well

**Future work**

* Natal and Ruben Deneer will be contacted to gain access to the data and motivations for the report. This way his thesis can be used as a case study for the framework. -> This is also a good idea to use for the Data Science seminar. Which will be discussed with Joaquin.
* The literature behind the framework should be extended with all parts discussed.
* A possible skype meeting should be done somewhere at the start of the January and a big meeting should be held at the start of February with all four of us together. (Peter, Dragan, Joaquin and Tim)