Feature Discovery in Small-Sized Experiments in Early Drug Development

Steven Wallaert – Promotor: Prof. Dr. Ir. Olivier Thas

Context

- Pre-clinical pharmacological research
- Biomarker discovery
- (Multi-) Omics

Problem

- Increasing popularity machine learning
- High hopes for better, more, easier discoveries

However

- High dimensionality: up to 30.000 and more features
- Extremely small sample sizes (10 to 50)

And

- Little is known about the performance of methods in these extreme situations

Therefore

- Need for a realistic simulation study

Research questions

- How should these kinds of data (not) be analyzed?
  - How do different statistical methods perform in these situations?
  - Identification of pitfalls

Simulation study

- Variety of scenarios
  - Data generating mechanism
  - Sample size
  - Number of features
  - Number of predictive features
  - Degree of discriminativeness
- Evaluation criteria
  - Number of true/false detections
  - Chance of true/false detection
  - Discriminative ability

Included methods

- Welch's t-test with FDR control
- Welch's t-test with empirical Bayes based selection bias correction using Tweedie's formula
- Logistic regression with L1 regularization
- Random forests based: Boruta
- Support vector machine based RFE

Challenges

- Selection of methods
- Computational demands
- Visualization and interpretation of results

Example analysis