Clustering of disease trajectories and prediction of outcomes from large real-world datasets

Introduction. Many diseases are complex and characterized by heterogeneous patient populations. This makes the design of successful treatments and clinical trials very difficult. Longitudinal healthcare data (e.g. from electronic health records) contain unique information on disease progression over time. Yet data are sparse, irregularly sampled and multimodal.

Objective. The aim of this project is to encode data from large population-wide registries and biobanks (learn low-dimensional representations) and model these data to identify clusters of disease trajectories and predict clinically relevant outcomes. Registry data from 95 million patients and HPC resources will be available. Additional data from biobanks are also available. Publication of the results and code release will be highly supported.

Requirements. Knowledge in machine learning and deep learning, with focus on unsupervised methods, autoencoders and attention models; programming in SQL (database query) and Python (analysis).