Chair Digital Health - Personalized Medicine &
The Hasso Plattner Institute for Digital Health at Mount Sinai

Contact Riccardo Miotto

Study of machine learning methods to enable patient stratification at scale

Deriving disease subtypes from electronic health records (EHRs) can guide next-generation personalized medicine. From a computational perspective, patient stratification is a data-driven, unsupervised learning task that groups patients according to their clinical characteristics and summarize these groups into clinically relevant sub-phenotypes. Ideally, when new patients enter the medical system, their health status progression can be tied to a specific subgroup, thereby informing the treating clinician of personalized prognosis and possible effective treatment strategies. This can be helpful in cases where a certain diagnosis is difficult and a more thorough examination is required, which sometimes might not come to mind to a busy clinician (e.g., specific genetic or lab tests). Moreover, the clinical characteristics of the different subtypes can potentially lead to intuitions for novel discoveries, such as comorbidities, side-effects or repositioned drugs, which can be further investigated analyzing the patient clinical trajectories.

We developed a model based on deep learning to derive disease subtypes at scale using EHRs, which was tested with different conditions, including T2D, Alzheimer's disease and Parkinson's disease. The paper is available here: https://arxiv.org/abs/2003.06516.

Multiple projects are available to further improve this work both on machine learning and clinical prospective. All projects will be performed using a fully de-identified dump of the Mount Sinai Data Warehouse, including EHRs for about 8M Mount Sinai patients. In particular, I am interested in exploring (among all):

(1) different machine learning models to derive patient representations (including BERT, GANs, variational autoencoders);

(2) how to best include clinical notes in the modeling;

(3) different clustering algorithms and strategies to identify clinically meaningful sub-types;

(4) different ways to better describe and represent the subtypes once identified with clustering analysis;

(5) how to operationalize the disease subtypes (i.e., how can we use them in the clinic);

(6) validate the idea and the models by deriving disease subtypes for other conditions.

Supervisors:
Erwin Boettinger
Riccardo Miotto

Additional contact persons:
Jan-Philipp Sachs
Suparno Datta