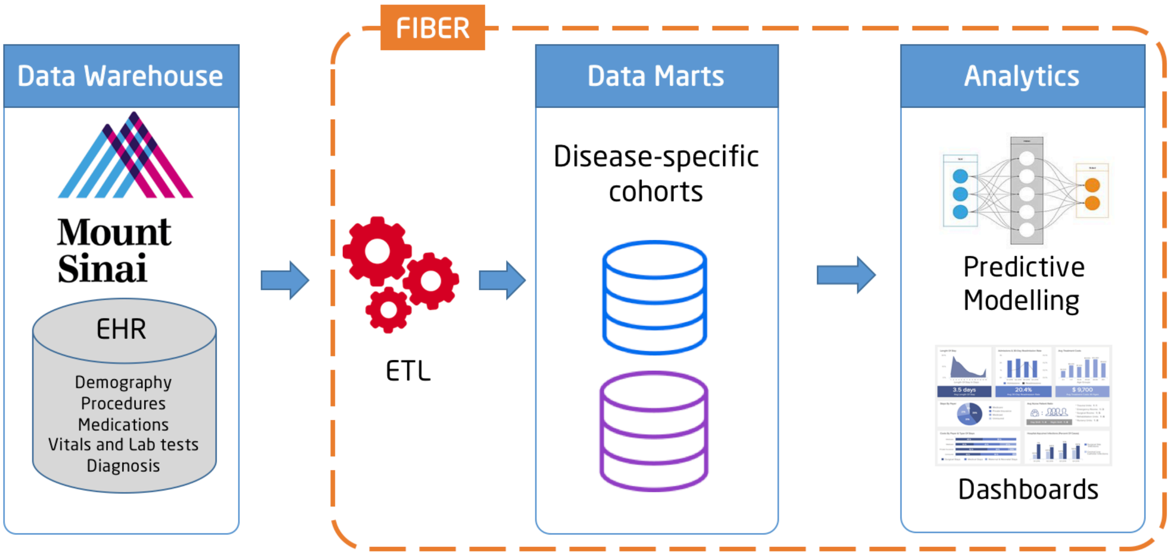
HPI Digital Health Center:
Research Group of Prof. Dr. Erwin Böttinger

Digital Health FIBER

A Flexible Data Mart for Clinical Predictive Modelling



Motivation

Electronic health records (EHR) data holds considerable promises for new research insights, improvement of health care delivery and more personalised patient treatment. However, in order for this data to be useful, it must be made available in a format that can be easily utilised by computational methods, such as machine learning. Current approaches, such as clinical data warehouses (CDW) or clinical data repositories (CDR), are typically geared towards the needs of financial reports and/or quality assurance, with little to offer with regards to predictive modelling of patient outcomes.

Therefore, whenever a clinical research wants to develop a new predictive model, a time-consuming process of data acquisition and preparation (including imputation) must be carried out before any modelling can take place. In this master project, you will design and implement a flexible data mart targeted specifically at the needs of clinical modellers in order to significantly reduce the time needed to develop a predictive model. The data mart thus developed will enable clinical researchers to visually specify patient cohorts, i.a., in terms of 1) underlying disease, 2) stratification factors (e.g. age, gender, disease severity, etc.) and 3) data required for modelling. Challenges pertaining particularly to clinical predictive modelling refer to, among others, the need to define observational periods, specify flexibly, i.a., time-to-event, medications, comorbidities, vital signs and the like.

Project Goals

The main goal of this project is to develop a tool which enables flexible data extraction targeted at clinical modelling for digital health researchers. Using the created data marts, you will implement a clinical predictive model using an existing frameworks (e. g. RapidMiner or Scikit-learn) to predict the onset of complications related to selected chronic diseases. For example, hypertensive patients are at increased risk of developing serious health related events, such as heart failure, coronary artery disease, stroke, renal disease, and peripheral arterial disease. Therefore, models targeted at preventing those outcomes are extremely important. The data marts will be based on Mt. Sinai data warehouse (snapshot from 2017). It comprises data from nearly 8 million patients, 120 million encounters, 18,718 different diagnosis based on ICD-9 codes and roughly 500 GB in size.

The main tasks of the project will be:

1. Interview digital health researchers on what their requirements are for a flexible data mart.
2. Carry out comparative analysis of different existing design approaches for a flexible data mart (flexibility, query complexity, etc.).
3. Create a set of "foundational queries" targeted at the needs of digital health researchers interested in development predictive models for patient outcomes.
4. Develop a tool (in Python) for data extraction, handling / wrangling (e.g. longitudinal data must be 'pivoted' to allow use of ML algorithms).
5. Based on the data mart created, we will perform the following tasks:
   - 5.1. Create a set of dashboards using either Apache Superset.
   - 5.2. Implement at least one clinical prediction model using data provided by the using a ML tool (RapidMiner, h20 flow, Scikit-learn, Orange etc.)

What you will learn

In this project, you will learn about:

1. Principles of (clinical) predictive modelling
2. Data visualisation and exploration tools (e.g. SAP Lumira or Apache Superset)
3. Analysis of query performance with large databases (complexity, runtime, etc.)
4. Design principles of medical databases (how to structure medical information?)
5. Fundamentals on medical terminologies, such as ICD-9, etc.
6. Ethical issues involved in dealing with observational healthcare data (data privacy perspectives when handling human subject data)

What you should bring with you

To carry out this project successfully, you will need solid foundations in:

1. Database principles, including normalisation, modelling notation and SQL
2. Basic programming skills (Python scripting)
3. Deep interest in learning more about data in the clinical domain
4. Fundamentals of machine learning (supervised learning, performance evaluation of predictive models)
5. Data visualisation principles

Contact

Get in touch for questions and ideas. We are located at the Digital Health Center on campus III, Rudolf-Breitscheid-Str. 187, 14482 Potsdam Building: G, Floor: 2nd.

Suparno Datta, M.Sc.
E: suparno.datta@hpi.de
T: 0331/5509-4817

Ariane Morassi Sasso, M.Sc.
E: ariane.morassi-sasso@hpi.de
T: 0331/ 5509-4827

Jan Philipp Sachs, MD, M.Sc.
E: jan-philipp.sachs@hpi.de
T: 0331/ 5509-4815

Harry Freitas da Cruz, MBA
E: harry.freitasdacruz@hpi.de
T: 0331/ 5509-1313

Prof. Dr. Erwin Böttinger
E: erwin.boettinger@hpi.de
T: 0331/5509-163