**Context**

The thesis will be held in the HEKA Inria—Inserm team in Paris, in collaboration with the Georges-Pompidou European Hospital (Hôpital Européen des Maladies Infectieuses et Tropicales - HEMSAP). Clinical data will be mainly and firstly those of the HEMSAP clinical data warehouse.

The thesis will be co-directed by Adrien Coulet, researcher at Inria Paris, and Bastien Rance, Assistant Professor at the University of Paris.

The candidate will be co-located at the HEMSAP-Hôpital AP-HP (75015), in particular for data access, and secondly at Paris-Sainte Campus, Paris (75015).

**Assignment**

**Precision medicine** aims at improving clinical care using available information on individuals such as genetics, lifestyle, and environment.

Considering such information can help in prescribing the right drug at the right dose, and in the process reduce adverse drug reactions (ADRs), which are estimated to account for one-third of hospital adverse events and approximately 280,000 hospital admissions annually in the US.

The inter-individual variability in drug responses, including ADRs, may have diverse causes such as: patient condition, age, a renal dysfunction impacts response to renally excreted drugs; drug interactions; drug-food interactions; genetics. The variety of factors, both known and suspected or unknown, makes it challenging for health institutions to take proper precautions.

**Electronic Health Records (EHRs)** offer unprecedented opportunities for using patient data to study variable patient outcomes, including drug response.

In particular, EHR recorded phenotypes may be used as surrogate markers of individual variations that lead to differential drug responses. An advantage of phenotypes available in EHRs is to provide an 'integrated view' of both genetic and environmental factors impacting patients. We obtained initial results while using Stanford EHR data, in absence of genetic (or other ‘omic’) information, to predict response to drug exposure.

Indeed for 23 drugs out of 34, we successfully used phenotypic data of individuals, recorded in EHRs prior to the drug exposure, to predict a reduced drug-dosing event.

The European Hospital Georges-Pompidou (HEGEP) has adopted an EHR since its opening in the year 2000. The data collected during the every day care are integrated and stored in a clinical data warehouse, allowing for the secondary use of these data.

The aim of this computer science thesis is to study how drug toxicity can be systematically and precisely detected in French EHR data, and then how the whole available data on patients can be leveraged to learn efficient representations for the prediction of such toxicity. We will focus on two clinical cases: drug toxicities caused by cancer therapies and those observed in patients with renal dysfunction.

**Main activities**

This thesis concerns the development of usable methods and tools for the detection and prediction of drug toxicity in cases of cancer treatment or renal dysfunction.

We propose to focus on two main axes that are necessary for the prediction of such events: (i) the detection of drug responses; (ii) the learning of patient representations for the prediction of such events.

(i) Drug responses are complex phenotypes that are usually not encoded in one simple field, or characterised by a value over a threshold. Its capture may necessitates to extract features from clinical text, to tune thresholds depending on patient profiles, etc. In particular, it requires to be evaluated by experts and on data of various hospital to enable its improvement and transferability.

(ii) There are many ways to encode historical patient data to train a machine learning model, and it is an open challenge to find the best representation for a particular learning task. In this thesis, we would like to focus on the task of drug toxicity prediction and experiment empirically to identify the pros and cons of various, potentially learned, representations of patients’ data.

Regarding temporal dimension, we propose to investigate language models and in particular, we propose to investigate language models and in particular scenarios that have a worldwide impact.
particular representations that embed sequences of words either in one or two directions such as LSTM or BERT \cite{sundermeyer2012, devlin2018}. These models could be adapted to consider sequences of patients’ events, instead of sequences of words. Another advantage of these approaches is that they have been proved to better adapt to small-size samples \cite{steinberg2020}.

(ii-b) Regarding the embedding of domain knowledge, we will consider graph representations such as those generated by Graph Convolutional Networks (GCN) \cite{schlichtkrull2018}. Those generate a sparse representation from data available in the form of graphs \cite{kipf2016}, and when ported to knowledge graphs may embed elements of formal semantics associated with knowledge graphs \cite{gutierrez2018, monnin2019}. Since many patients’ clinical features are associated with domain knowledge (e.g., drug or phenotype ontologies), we would like to investigate how the GCN ability to embed such knowledge may improve the performances of patient representations.

Case studies

Oncology and response to treatments. The care of cancer patient is adapted according to the patient personal characteristics (their health status, but also their personal aspirations and desires), their biological and genetics make-up, their exposition to pollution... These PhD will aim at finding methods to identify and predict toxicities to cancer treatments.

The HEGP is a excellence center for the care of several cancer pathologies, and has access to a large cohort along with data ranging from text reports to omics data.

Chronic kidney disease. The same methods will be explored to study chronic kidney diseases and more specifically glomerular disease to search for subgroups of patients presenting common trajectories in the evolution of their disease.

Skills

Technical skills and level required:

Python, R

SQL

Languages:

English and French

Relational skills:

excellent professional oral and written communications

Benefits package

- Subsidised meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities