graphs. representations that enable the consideration of (ii-a) temporal dimensions of representations of patients \cite{bengio2013}. In particular we are interested in learned and experiment empirically to identify the pros and cons of various, potentially learned, model, and it is an open challenge to find the best representation for a particular (ii)

trained with terms of standard clinical data schema and vocabularies such as FHIR, their evaluation in various clinical settings, these rules or models will be expressed and learning model that enables the automatic annotation of novel patients. To facilitate will learn (semi automatically or automatically) an explicit set of rules, or a machine that from a set of patient EHR, for which the outcome has been manually annotated, we development of efficient and transferable phenotyping algorithms for drug toxicity. We

enable its improvement and transferability \cite{banda2018}. We would like to study the particular, it requires to be evaluated by experts and on data of various hospital to

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 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 patient with renal dysfunction.

Principal axes

We propose to focus on two main axes that are necessary for the prediction of such events: (i) the detection of drug responses, i.e, of the outcome; (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 necessitate 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 \cite{banda2018}. We would like to study the development of efficient and transferable phenotyping algorithms for drug toxicity. We envisage here to use supervised approaches or semi-supervised approaches, meaning that from a set of patient EHR, for which the outcome has been manually annotated, we will learn (semi automatically or automatically) an explicit set of rules, or a machine learning model that enables the automatic annotation of novel patients. To facilitate their evaluation in various clinical settings, these rules or models will be expressed and trained with terms of standard clinical data schema and vocabularies such as FHIR, OMOP Common Data Model and vocabulary \cite{rupcsak2015, thao2016}.

(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 \cite{bengio2013}. In particular we are interested in learned representations that enable the consideration of the \texttt{x(t)}(i-a) temporal dimensions of patient trajectory and \texttt{x(t)}(i-b) domain knowledge represented in knowledge graphs.

(i-a) Regarding temporal dimension, we propose to investigate language models and in
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
representations are sparse 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.

Compétences
Technical skills and level required:
Python, R
SQL

Languages:
English and French

Relational skills:
Excellent professional oral and written communications

Avantages
- Subsidized 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