DETECTION OF CHEMOTHERAPY TOXICITY AND RESPONSE FROM EHR

Alice Rogier, Bastien Rance, Adrien Coulet
PhD project
Équipe HeKa, Inserm, Inria, Université de Paris et Hopital Européen Georges Pompidou AP-HP

Abstract

To improve and adapt chemotherapy regimen, toxicity and treatment response must be monitored and analysed. Both toxicity and treatment response are complex events that are difficult to detect automatically from EHRs. However their automatic detection would be of great help for the development of clinical decision support systems. Indeed, our aim is to develop an automatic detection tool of these events. This tool will leverage structured data such as biological results or diagnostic codes (ICD10), and unstructured data, such as free text medical reports. The latest contain richer information, but difficult to extract in an automatic manner. We propose to leverage NLP approaches to detect such event and quantify their magnitude. Once detected, we aim at training machine learning models for toxicity prediction. This challenges will be tackled in my starting PhD project, under the supervision of AC and BR, and jointly funded by Inserm and Inria.

Context

Nowadays, Electronics Health Records (EHRs) offer unprecedented opportunities for using patient data to study variable patient outcomes, including drug response. Information about chemotherapy response and adverse drug reaction (ADR) occurrences is not directly available in CDWs. For example, an ADR due to irinotecan administration can be found through a value of bilirubin exceeding threshold in biological results, or through the mention of a cholecystitis in a narrative report. An automatic tool to detect toxicity and chemotherapy response would thus simplify research. Furthermore, this tool is required to elaborate predictive models.

Method

We aim to develop an automatic toxicity and response to chemotherapy detection tool from structured and unstructured EHR data, of the HEGP CDW. To reach this goal, we will extract biological data, chemotherapy prescription data and narratives associated with every patients receiving chemotherapy.
In a first approach, we will detect toxicity in narratives with rule-based system (RBS) and Named Entity Recognition (NER) approaches. We will base our algorithm on existing ones [2] [3]. We will adapt these algorithms to chemotherapy toxicity by building dictionaries from ADR french terminologies as WHOARTFRE (World Health Organisation Adverse Event French), MedDRA (Medical Dictionary Activities) and CTCAE (Common Terminology Criteria For Adverse Events). An overview of NER model that will be developed is detailed in the next section.

Model

A cohort of patients diagnosed with cancers will be defined, by querying the HEGP CDW. We will use diagnostic codes of International Classification of Diseases 10 (ICD10) to find out these patients using cancer specific ICD10 diagnostic codes that all start with "C". All narrative reports of these patients will be extracted to train, validate and test our toxicity detection model.
First of all, we will make a dictionary gathering all toxicities that can be found in French medical terminologies about ADRs. By exact match, we hope to find enough toxicity-related terms inside narrative reports.

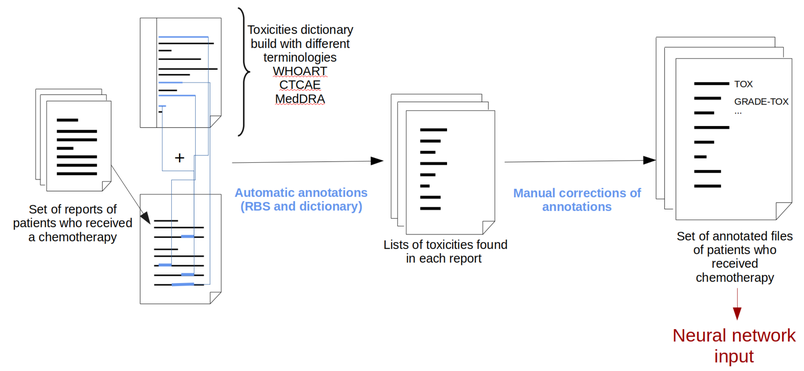

Fig. 1: Pre-processing

We will correct manually the term list found in each reports. In parallel, we will fine-tune a French word embedding algorithm (as CamemBERT or FlauBERT) on a set of medical reports. We will than obtain a french clinical language model. A word embedding of patients who received chemotherapy obtained thanks to this clinical language will also be an input of our neural network.

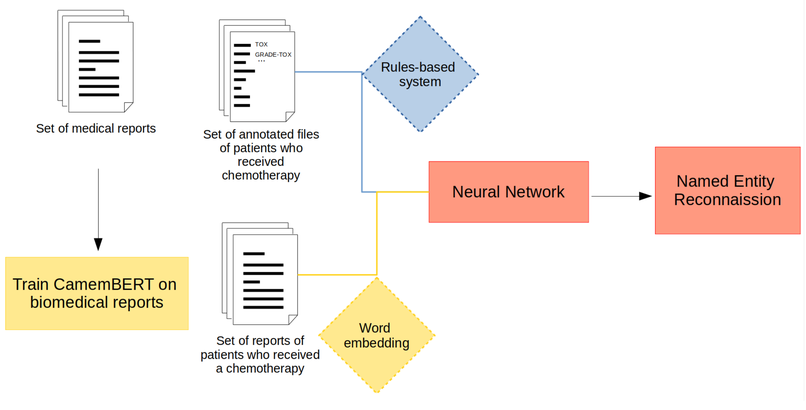

Fig. 2: Overview of the different steps of our NER approach.

Model evaluation

Our model will be evaluated with the test dataset. We will calculate the different performance metrics (recall, precision, F1-score). To facilitate the evaluation, Doccano will be used [4]. Doccano is a user-friendly interface to visualise and validate annotations.

Fig. 3: Aperçu des annotations sous Doccano.

Excepted results

If our model performance scores are high enough, it will be integrated among the set of annotation tools developed by our team [1] and installed at the hospital . It will be integrated to a decision-making tool used when administrating chemotherapy treatment. This will enable feeding oncolocy research with precious information.

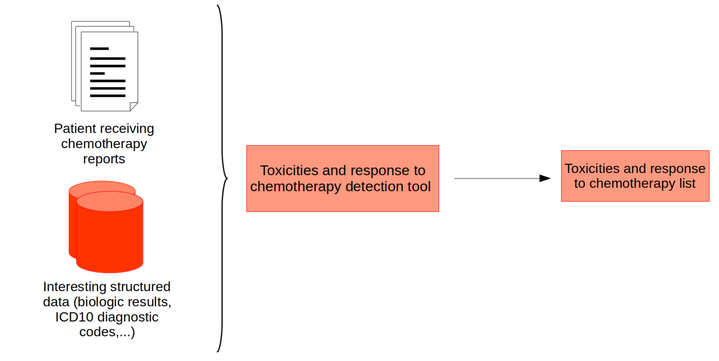

Fig. 4: Toxicity and response to chemotherapy detection tool

References

[1] Nicolas Garcelon et al. “A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse”. In: *Journal of biomedical informatics* 80 (2018), pp. 52–63.

[2] Jordan Jouffroy et al. “MedExt: combining expert knowledge and deep learning for medication extraction from French clinical texts”. In: *JMIR preprint* (2020).

[3] Ivan Lerner et al. “Learning the grammar of prescription: recurrent neural network grammars for medication information extraction in clinical texts”. In: *arXiv preprint arXiv:2004.11622* (2020).

[4] Hiroki Nakayama et al. *doccano: Text Annotation Tool for Human*. Software available from https://github.com/doccano/doccano. 2018. URL: https://github.com/doccano/doccano.