Natural Language Processing in Clinical and Translational Research: Of the Importance of Section Headers for Accurate Modifiers Assignments

Alexandre Yahi, MS, Ning Shang, PhD, Nicholas P. Tatonetti, PhD, Noémie Elhadad, PhD, George Hripcsak, MD, MS
Department of Biomedical Informatics, Columbia University, New York, NY, USA

COLUMBIA UNIVERSITY MEDICAL CENTER

OHDSI

Background

The secondary use of the Electronic Health Records (EHR) has enabled researchers to use data science in unprecedented way for clinical and translational studies. However, the medical characterization, or phenotyping, of patients or cohort of patients to support these studies needs to be accurate and reliable:

- Phenotyping is a central question and various methods coexists, from clinical algorithms, to classification models and deep learning methods such as word embeddings
- The Common Data Model (CDM) is a unique opportunity to share data and methods to advance research in cohort stratifications
- As the CDM is evolving to integrate the output of Natural Language Processing in its schema, it is important to evaluate the challenges of the integration of NLP outputs into research workflow and identify the methods it can empower

To do so, we evaluated if the modifiers provided by NLP tools are enough to identify medical terms that represent the truth for the patient at the time of the note. More specifically, we studied the impact of the additional semantic layer provided by note sections on diseases and symptoms modifiers.

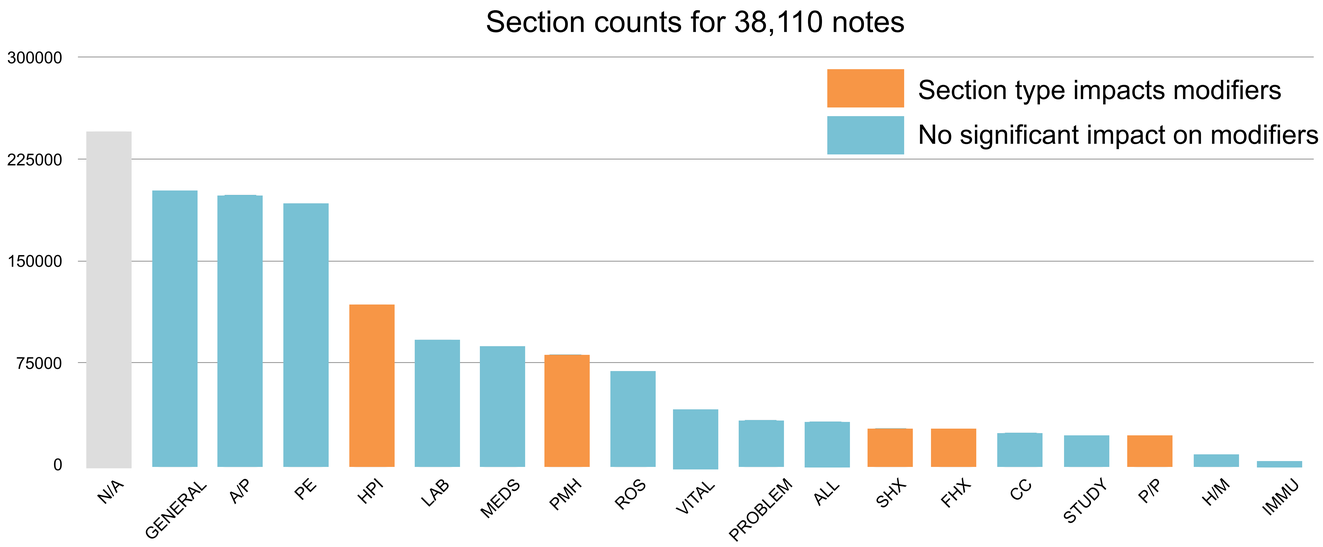
Methods

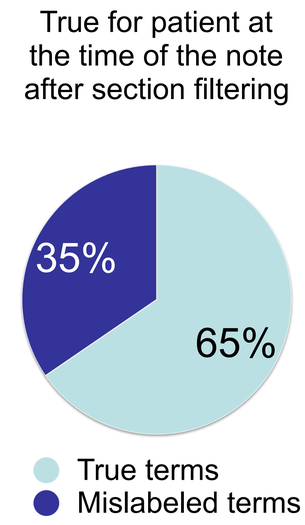
We focused on the clinical notes of 1,685 patients at Columbia University Medical Center/New York Presbyterian (CUMC/NYP) from August 1999 to July 2014. These patients belong to a pilot cohort consisting of individuals with available genotyping data. It represents 38,110 documents across 114 note types. We parsed these clinical notes using the clinical Text Analysis and Knowledge Extraction System (cTAKES). This tool's named entity recognition (NER) annotator implements a terminology- agnostic dictionary look-up algorithm and maps each concept to terminologies including SNOMED CT and RxNORM. The modifiers annotated by cTAKES were: polarity, uncertainty, conditional, generic, subject, historyOf. For the section identification, we manually curated note sections based on the paragraph headers and classified them into 20 categories. These 20 section header types were based on the most frequent section headers found in the 114 notes types studied. We grouped some headers together (e.g., "PMH psychology" in "PMH") in order to have sections non-specific to hospital services but reflecting semantic properties in terms of modifiers. We also grouped section together when the outcome would not have an important impact on the term modifiers, or is a synonym (e.g. "Diagnosis" grouped with "Assessment").

Contact: Alexandre Yahi (alexandre.yahi@columbia.edu)

Results

We were able to associate a note section to the following count of terms: past medical history and past surgical history (P/P) 19865, family history (FHX) 26317, problem list (PROBLEM) 32965, summary before the first section (GENERAL) 199151, social history (SHX) 27154, physical examination (PE) 192246, review of systems (ROS) 67937, past medical history (PMH) 81262, chief complain (CC) 23749, studies (STUDY) 20343, no section identified (N/A) 245289, laboratory tests (LABS) 91702, history of present illness (HPI) 117904, vital signs (VITAL) 40343, assessment and plan (A/P) 198810, allergies (ALL) 31909, health care maintenance (H/M) 7549, medications (MEDS) 86698, immunization (IMMU) 2047.





In terms of modifiers combinations captured by cTAKES, the most frequent tuple was "positive, non-conditional, certain, non-generic, for the patient, in the present" 1,247,997 times, followed by "negative, non-conditional, certain, non-generic, for the patient, in the present" 215,330 times.
However, these modifiers are not enough to interpret the complete semantic of the terms capture by the NLP engine. In the section relative to history (i.e., PSH, P/P, PMH, SHX, FHX), the modifier "historyOf" was triggered only 4.65% of the time. Conversely, for the family history section, the modifier "family_member" was only captured 15.35% of the time. By taking into account the note sections into the filtering process to keep what is true for the patients at the time of the note, we went from 1,247,997 to 816,836 terms, representing a 34.5% decrease.

Conclusions

In conclusion, we demonstrated that the combination of modifiers identified by NLP engines is not enough to classify parsed terms. They should be used along with note sections, in particular to modulate modifiers relative to past/present and patient/family. In this study, limitations came from both the NLP pipeline used for the term matching and form the manual curation of sections. This latter process could be performed in a supervised manner as demonstrated by Li et al. using Hidden Markov Model (HMM)[1] and using the SecTag terminology proposed by Denny et al.[2] to identify and classify section headers. Structuring NLP outputs into tables in the CDM along with note section identification and other parsing tools able to extract specific values in clinical notes will advance the dissemination of phenotyping algorithms. However, such tables can only marginally improve performances of machine learning approaches in phenotyping, mostly improving recall.[3] With approaches leveraging deep learning in biomedical informatics[4], and word embeddings demonstrating their power on free-text, further work needs to be done to understand how and if NLP and relational databases will be of any

References

1. Li, Y., Lipsky Gorman, S., & Elhadad, N. (2010). Section classification in clinical notes using supervised hidden markov model. the ACM international conference (pp. 744–750). New York, New York, USA: ACM. http://doi.org/10.1145/1882992.1883105
2. Denny, J. C., Johnson, K. B., & Spickard, A. (2008). Development and evaluation of a clinical note section header terminology., 2008, 156–160.
3. Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., & Karlson, E. W. (2015). Development of phenotype algorithms using electronic medical records and incorporating natural language processing. Bmj. http://doi.org/10.1136/bmj.h1885
4. Miotto, R., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports, 6, 26094. http://doi.org/10.1038/srep26094