Extending the OMOP CDM to store the output of natural language processing pipelines

Mónica Arrúe, Sandra Pulido, Alvaro Abella, Gabriel Maeztu, Alberto Labarga

Background

There are huge amounts of data being generated at hospitals every day. Up to 80% of this data is collected in an unstructured format and a large portion of it as free text[1]. In order to extract value from this data it is necessary to turn it into structured data. This is specially relevant for clinical studies that benefit from all the data capture performed during the clinical practice. Natural language processing (NLP) solutions can structure clinical texts written by physicians, extracting and encoding relevant medical concepts and taking into account complex context such as negations, family/personal background, past events among others[2].

While OMOP CDM is a great schema to store structured data, NLP results can get messy and complex. Although OMOP CDM v6 provides a *note_nlp* table to store these results, queries to this table can become clumsy and slow, so we designed and extended the OMOP CDM with our own NLP schema to store the results generated in the annotation process of NLP.

In the following section we present this extension of the OMOP CDM able to store the output of NLP solutions while integrating with the vocabulary normalization process of the OMOP CDM.

Methods

Existing NLP solutions generate huge amounts of data that must be stored together with existing structured data. Although the OMOP CDM[3] is capable of storing structured and normalized data, the existing *note_nlp* table is too limiting for complex NLP solutions. The NLP schema we propose is a schema made up of nine different tables that not only allow us to store NLP results but also the metadata about the machine learning models and pipeline components of the NLP systems used to process the notes. This schema has already been successfully tested in several production environments for over a year and has already been presented to the OHDSI NLP Working Group for review.

A centralized fact table, the ***note*** table, is connected to several dimensions in a snowflake schema fashion (Figure 1). These dimensions represent the data captured by the NLP system about the content of the notes. Given the amount of data extracted, we decided to divide the extracted information in two different dimensions, for logical sharding reasons. One dimension represents all the clinical entities (*e.g.* “headache”, “Marfan’s syndrome”) captured from the text. The second one represents different parts of the note, such as paragraphs, sections, spans of text or the whole note. Normalized in other tables is the corresponding metadata about the entities or the text parts. An example of the metadata about an entity would be their context (e.g. “Present”, “Past”), experiencer (e.g. “Patient”, “Mother”) or relationship with another entity (e.g. “cause of”, “location of”) among others. An example of the metadata about a note part would be the start and end characters of the sections in the clinical note about the “Family history” section, the “Plan” or the “Physical examination”, or more complex classifications like “First diagnosis” or “Treatment change”.

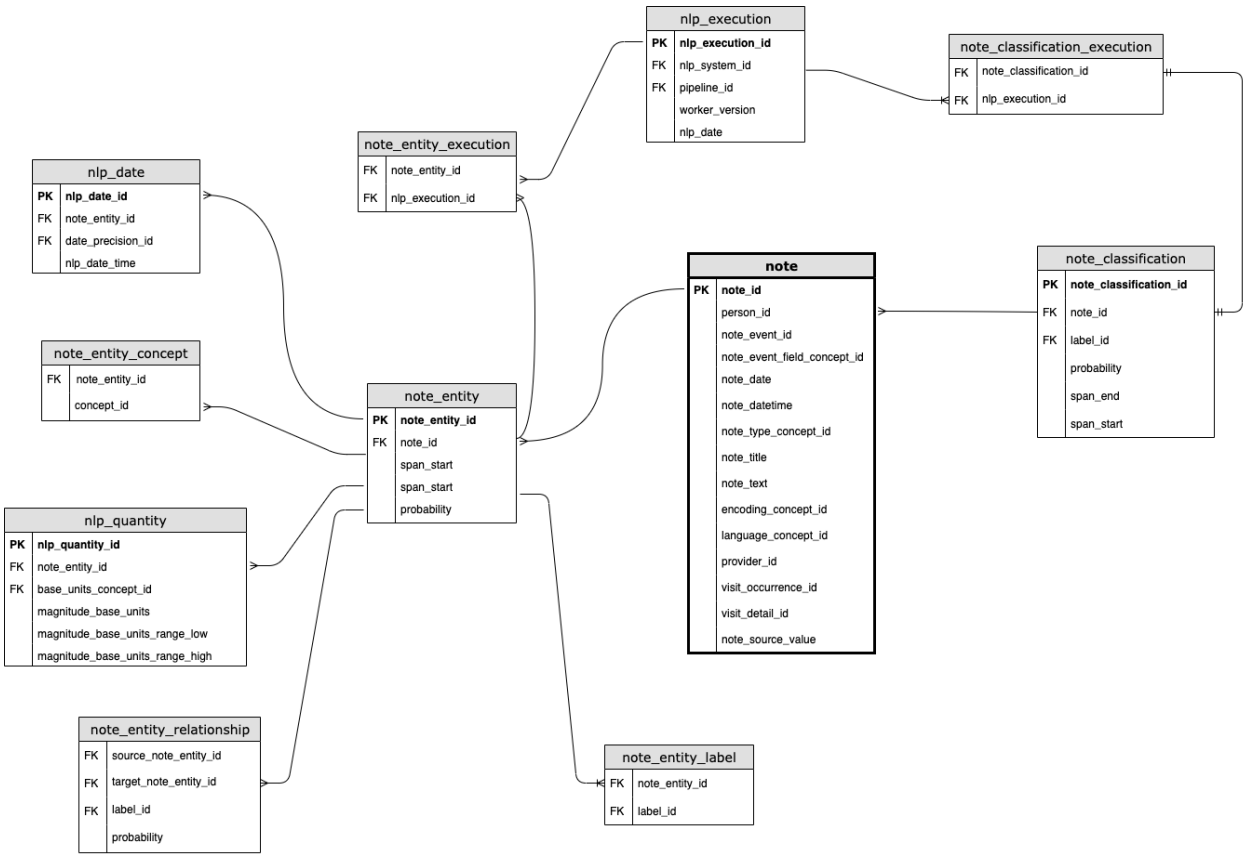

Figure 1. Schema used to store the results produced by the NLP annotation process.

With this schema, in order to retrieve all events with a note where a condition of psoriasis was positively stated would involve a joint query to the note, note_entity, note_entity_concept and note_entity_label, filtering by the concept_id of psoriasis, and the labels corresponding to Certain, Present and Positive. If we would like to know those events where psoriasis was first diagnosed we would add note_classification to the query, and filter by the corresponding label 'First diagnosis'.

Results

We have installed our platform in 5 hospitals in Spain and, to date, our tools have processed more than 33 million clinical notes of more than 4 million patients in Spain (Figure 2). Thanks to this, we not only make it easier for hospitals to participate in clinical studies, but also, by having their data transformed to the OMOP CDM, they can participate in OHDSI studies including NLP results. An example of this is the study "Characterizing Health Associated Risks, and Your Baseline Disease In SARS-COV-2 (CHARYBDIS)"[4], in which Hospital del Mar in Barcelona took part.

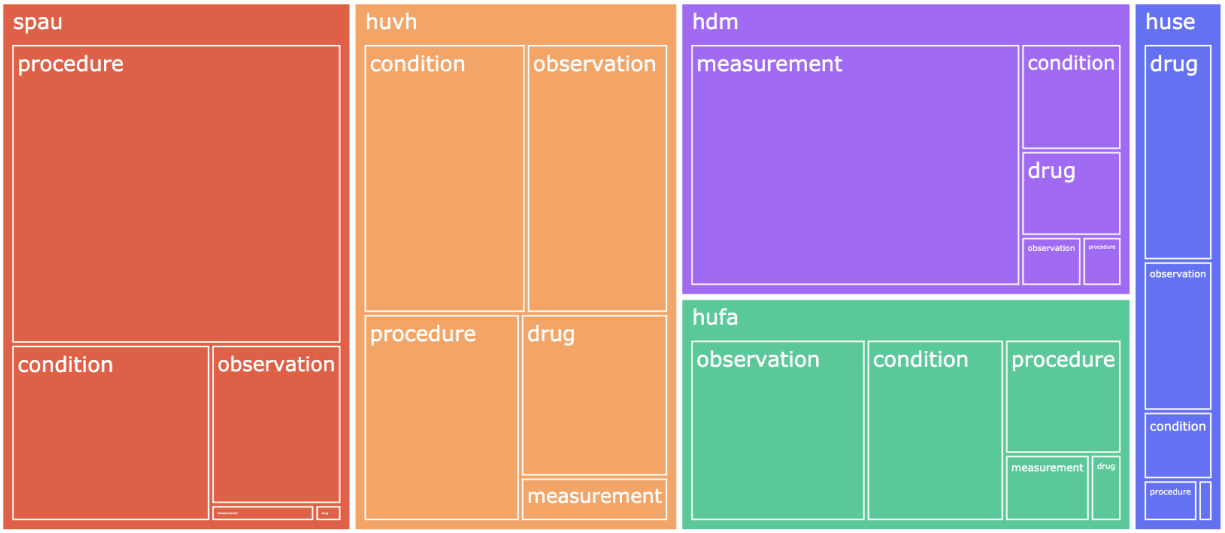

Figure 2. Number of events retrieved from structured data and clinical texts by hospital and domain

As an example of the joint power of the ETL + NLP approach we analysed the prevalence of four dermatological diseases. The following are the OMOP concept identifiers for the diseases:

- Hidradenitis Suppurativa: 4241223
- Chronic Urticaria: 4198855
- Atopic Dermatitis: 133834
- Psoriasis: 3468820

The analysis performed corresponds to patients above 18 years old, and who have at least one visit to any department in the last 6 years, which were easily retrieved using person and visit_occurrence tables. Patients found via NLP were searched for in notes from the dermatology department only. Figure 3 shows the enrichment gained by using NLP processing on the OMOP CDM NOTE table compared to using only already structured data.

In the case of Chronic Urticaria (OMOP 4198855), there are no matches in the structured diagnoses from the hospital. This is due to the absence of a ICD-9 or ICD-10 code which specifically signifies “chronic urticaria”.

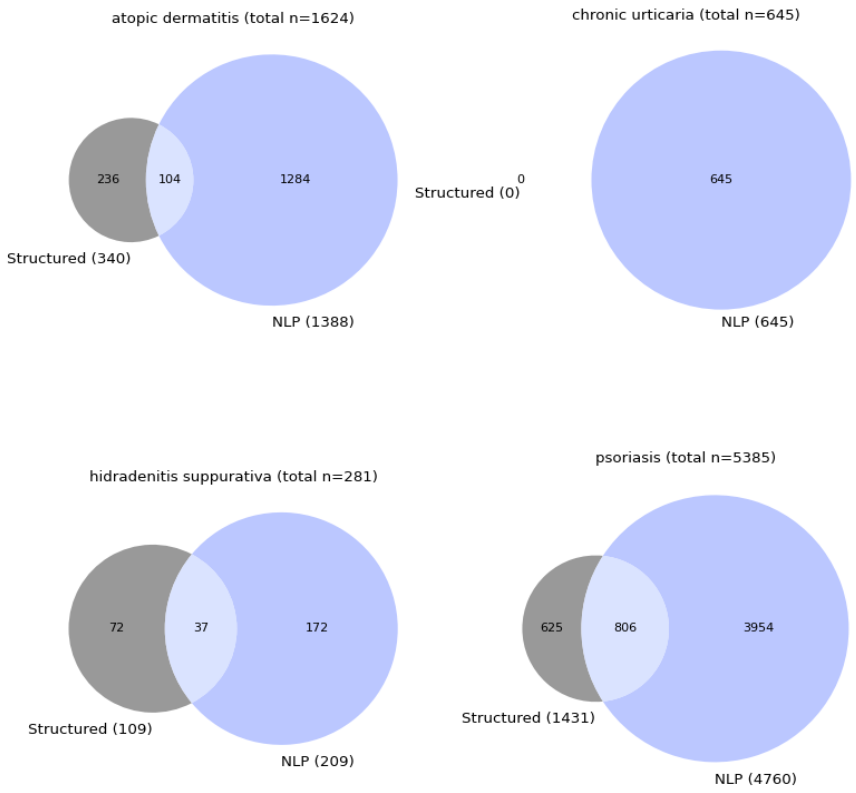

Figure 3. Number of patients per disease and extraction method. We divide each graph into patients found via hospital diagnosis ("Structured") and via NLP.

Conclusion

We have presented a solution to accelerate clinical research by structuring clinical text using advanced NLP algorithms. In this context, we have introduced a new extension of the OMOP CDM able to store the output of NLP solutions.

Currently, our work is mainly focused on improving our NLP models and adapting them to other languages, such as English. In addition, an increasing number of available hospitals is expected in the coming months so we will integrate these new hospitals by transforming their data into the OMOP CDM.

References/Citations

1. SyTrue. Why Unstructured Data Holds the Key to Intelligent Healthcare Systems [Internet]. Healthcare IT News. 2015 [cited 2021Jun18]. Available from: https://hitconsultant.net/2015/03/31/tapping-unstructured-data-healthcares-biggest-hurdle-realized/#.YMxPCjYzbJx

2. Chocrón P, Abella A, De Maeztu G. ContextMEL: Classifying Contextual Modifiers in Clinical Text. Procesamiento del Lenguaje Natural. 2020; 65.
3. OMOP CDM v6.0 [Internet]. cdm60.utf8. [cited 2021Jun18]. Available from: https://ohdsi.github.io/CommonDataModel/cdm60.html#OMOP_CDM_v60
4. Prats-Uribe A, Sena A G, Lai L Y H, Ahmed W, Alghoul H, Alser O et al. Use of repurposed and adjuvant drugs in hospital patients with covid-19: multinational network cohort study BMJ 2021; 373 :n1038 doi:10.1136/bmj.n1038