Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220604

Enriching UMLS-Based Phenotyping of Rare Diseases Using Deep-Learning: Evaluation on Jeune Syndrome

Carole FAVIEZ[a,b,1], Marc VINCENT[c], Nicolas GARCELON[a,b,c], Caroline MICHOT[d,e], Genevieve BAUJAT[d,e], Valerie CORMIER-DAIRE[d,e], Sophie SAUNIER[f], Xiaoyi CHEN[a,b,c], and Anita BURGUN[a,b,g]

[a] *Centre de Recherche des Cordeliers, Sorbonne Université, INSERM, Université de Paris, Paris, France*

[b] *Inria Paris, France*

[c] *Université de Paris, Imagine Institute, Data Science Platform, INSERM UMR 1163, Paris, France*

[d] *Reference Centre for Constitutional Bone Diseases, laboratory of Osteochondrodysplasia, INSERM UMR 1163, Imagine Institute, Université de Paris, Paris, France*

[e] *Hôpital Necker-Enfants Malades, Service de génétique, AP-HP, Paris, France*

[f] *Laboratory of Renal Hereditary Diseases, INSERM UMR 1163, Imagine Institute, Université de Paris, Paris, France*

[g] *Hôpital Necker-Enfants Malades, Département d'informatique médicale, AP-HP, Paris, France*

Abstract. The wide adoption of Electronic Health Records (EHR) in hospitals provides unique opportunities for high throughput phenotyping of patients. The phenotype extraction from narrative reports can be performed by using either dictionary-based or data-driven methods. We developed a hybrid pipeline using deep learning to enrich the UMLS Metathesaurus for automatic detection of phenotypes from EHRs. The pipeline was evaluated on a French database of patients with a rare disease characterized by skeletal abnormalities, Jeune syndrome. The results showed a 2.5-fold improvement regarding the number of detected skeletal abnormalities compared to the baseline extraction using the standard release of UMLS. Our method can help enrich the coverage of the UMLS and improve phenotyping, especially for languages other than English.

Keywords. Named entity recognition, electronic health records, deep phenotyping, rare disease

1. Introduction

The increasing digitization of health-related data provides the unique opportunity to bring new insights into disease knowledge and patient care. For rare diseases, electronic health records (EHR) provide a precious source for patient high throughput

[1] Corresponding Author, Carole Faviez, Centre de Recherche des Cordeliers, INSERM, 15 rue de l'école de médecine, F-75006, Paris, France; E-mail: carole.faviez@inserm.fr.

phenotyping that can help reduce the risks of mis- and delayed diagnosis [1]. However, the amount of unstructured data in EHRs requires that named entity recognition (NER) be developed to automatically extract medical entities from text. NER approaches include solutions leveraging thesauri to search for mentions of the terms in the documents as well as machine or deep learning. QuickUMLS [2] and cTAKES [3] are thesaurus-based solutions that have been evaluated on clinical notes in English. These systems provide normalized terms, e.g., based on the Unified Medical Language System (UMLS) CUIs [4] which enables the easy reuse and comparison of the retrieved phenotypes and the conversion to other thesauri or languages. These methods generally reach better precision than recall [2], due to the presence of specific synonyms or variations of the terms in text. Moreover, the coverage of used thesauri can vary across languages, especially for rare phenotypes. To address these issues, non-dictionary-based methods leveraging deep learning have been used to detect medical entities, such as BiLSTM-CRF [5] and PhenoTagger [6], which have been evaluated on medical articles. These developments have been part of a wider move toward deep learning in healthcare, due to the large availability of biomedical data [7]. Regarding medical entity extraction, these methods do not provide the same normalization as thesaurus-based ones and their performance on clinical text must be analyzed. In this article, we present a hybrid method to extract phenotypes from EHR documents combining a dictionary-based method [8] with deep learning. The pipeline was evaluated in a French rare disease center, with focus on Jeune syndrome (Jeune asphyxiating thoracic dystrophy), a rare genetic disorder characterized by skeletal abnormalities such as small, narrow thorax, short ribs, shortened bones of the arms and legs, unusually shaped pelvis, and extra fingers and/or toes[2].

2. Methods

2.1. Patient selection and extraction of phenotypes

The Necker Children's Hospital is a French reference center for rare and undiagnosed diseases that hosts the Imagine Institute, a research center specialized in genetic diseases. Its clinical data warehouse (Dr. Warehouse [8]) contains EHRs of more than 800,000 patients. Dr. Warehouse enables the automatic extraction of medical entities from EHRs based on the UMLS [4]. In addition, a research database is used to store structured research information from rare disease patients of Imagine Institute, including patient diagnoses. We selected all patients diagnosed with Jeune syndrome from the Imagine research database and limited ourselves to the patients followed up at Necker hospital, i.e., with EHRs in Dr. Warehouse.

UMLS extraction was performed by the high throughput phenotyping module of Dr. Warehouse [8]. The deep learning extraction was performed by applying a biGRU-CRF [9] extension to the method described in [10]. This model is well suited for NER from noisy natural language text of EHRs due to its ability to remember information across different ranges of contextual dependencies [11]. Each document was tokenized and fed as a string of fasttext word embeddings to the NER model. We reused an internal annotated dataset of 625 clinical reports in French from various departments at Necker hospital for model selection and training. These reports were manually

[2] https://rarediseases.info.nih.gov/diseases/3049/jeune-syndrome

annotated with spans corresponding to phenotypic entities (i.e., observable physical or mental characteristics of the individuals, including both pathological and normal characteristics). The annotation provided 6928 phenotypes stemming from 2.5×10^5 tokens.

2.2. UMLS and biGRU-CRF extraction and evaluation

The set of Jeune syndrome EHRs was randomly split into two datasets: training set (two thirds of the patients) and test set (the remaining third). Phenotypes were extracted from the training set using the UMLS extraction and the biGRU-CRF method. Phenotypes detected exclusively by the biGRU-CRF method were manually reviewed to verify whether they were skeletal abnormalities associated with Jeune syndrome and mapped to UMLS concepts. The list of skeletal abnormalities and their mappings to the UMLS were validated by three bone disease experts from Necker/Imagine (VCD, GB, CM). The enriched version of the UMLS integrating these new synonyms in French will be referred to as UMLS+ in the following. The last step consisted in extracting phenotypes from the test set with the dictionary-based method using standard UMLS and UMLS+ and comparing the results. The pipeline overview is given in Figure 1.

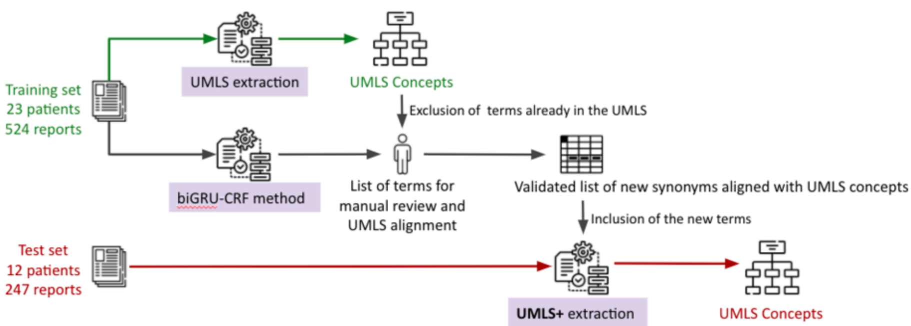

Figure 1. Overview of the pipeline.

3. Results

3.1. Training set and UMLS enrichment

Starting from the research database, we identified 142 patients with Jeune syndrome. Among them, 35 patients were followed up at Necker, with an average of 22 documents per patient in Dr Warehouse (median 10). 23 patients out of 35 were randomly selected for the training set. These patients were associated with a total of 524 documents on which UMLS extraction and biGRU-CRF techniques were applied. The standard UMLS extraction identified 592 distinct terms (3387 occurrences) in the training set. Compared with this method, the biGRU-CRF method was able to detect 1036 additional distinct terms (1922 occurrences). The review of these additional French terms by clinical experts led to the identification of 119 skeletal abnormalities related to Jeune syndrome, among which some key phenotypes such as *thoracic dystrophy*, *narrow thorax*, and *short ribs*. These 119 skeletal abnormalities were mapped to 76 UMLS CUIs. With the addition of lexical variants (e.g., plural forms), 152 different words or expressions were added to the UMLS+. This list is available upon request to the corresponding author. This process allowed us to add French translations of English terms (e.g., *thorax etroit* is equivalent to *narrow thorax* (C0426790)) but also to identify new synonyms that were absent from the Metathesaurus in English and in French (e.g., *duplication of the fifth finger* can be mapped to *post-axial polydactyly* (C0220697)).

3.2. Comparison of UMLS and UMLS+ based extraction on the test set

Phenotypes were extracted from our test set of 12 patients and 247 documents, using the UMLS and the UMLS+. The standard UMLS extraction provided 354 distinct terms (1168 occurrences, 328 distinct CUI), including 34 skeletal abnormalities associated with Jeune syndrome (148 occurrences, 33 CUI). With UMLS+, we detected 67 skeletal abnormalities, which is to say that after enrichment, the system was able to detect 33 additional skeletal abnormalities (222 occurrences, 25 CUI). In other terms, we obtained twice as many phenotypes and a 2.5-fold increase of phenotypic information. Of note some terms are of major interest for Jeune syndrome diagnosis, like *etroitesse thoracique* (*narrow thorax*), with 21 occurrences in the test set. The most frequent phenotypes related to skeletal abnormalities are displayed in Table 1.

Table 1. Top 10 skeletal abnormalities detected in the test set with the UMLS+.

UMLS CUI	Found text (fr)	English translation	Frequency	Origin
C1406921	Dystrophie thoracique	Thoracic dystrophy	82	Added term
C0036439	Scoliose	Scoliosis	41	UMLS
C0302142	Deformation	Deformity	33	UMLS
C0426790	Etroitesse thoracique	Narrow thorax	21	Added term
C0410528	Maladie osseuse constitutionnelle	Constitutional bone disease	18	Added term
C0426817	Cotes courtes	Short ribs	17	Added term
C0022821	Gibbosite	Gibbosity	16	Added term
C0022821	Cyphose	Kyphosis	13	UMLS
C0426790	Thorax etroit	Narrow thorax	12	Added term
C1439256	Deformation thoracique	Thorax deformity	11	Added term

4. Discussion

Even though UMLS is a rich source for extracting and normalizing phenotypes from EHRs, we demonstrate in this study the need for enriching the UMLS with terms that are used in clinical documents, especially in languages other than English. We showed that deep learning techniques such as biGRU-CRF can be used as a complementary method to improve phenotyping especially in the context of rare disease. As for Jeune Syndrome, the enriched UMLS (UMLS+) enabled a 2.5-fold increase in phenotyping. Although we were able to map all the extracted phenotypes to existing CUIs, some key phenotype terms in French were totally absent from the Metathesaurus. Moreover, some refined terms were missing even in English, e.g., *duplication of the fifth finger*. These results are similar to those obtained by Vasilakes et al. [12], who compared a dietary supplement knowledge base (iDISK) with the UMLS and found that although 99% of the iDISK concepts were present in the UMLS, only 30% of the terms/synonyms were included. Our proposed enrichment process can be easily reproduced for other rare diseases. The limitation is that a manual review of new extracted phenotypes and normalization to UMLS concepts can be time-consuming for large corpora. To address this issue, it would be interesting to develop a module for the automatic enrichment of UMLS terms. Such a system could rely on term embeddings to group terms detected from text with UMLS terms that convey similar meanings [13], as proposed by Sarker [14]. Regarding the main contributions to the field, the enrichment process that we propose provides more accurate and comprehensive phenotyping and can improve the performances of EHR-based screening of patients. This is crucial for rare diseases, as underdiagnosis and delayed diagnosis are frequent, leading to a high social and psychological burden. Future work will consist in generalizing the method to other diseases.

Acknowledgement This work was supported by the French National Research Agency (ANR) under the C'IL-LICO project (ANR-17-RHUS-0002), approved by the French ethics and scientific committee for health research (CESREES) (#2201437), and as part of the "Investissements d'avenir" (ANR-19-P3IA-0001) (PRAIRIE 3IA Institute).

References

[1] Faviez C, Chen X, Garcelon N, Neuraz A, Knebelmann B, Salomon R, et al. Diagnosis support systems for rare diseases: a scoping review. Orphanet Journal of Rare Diseases. 2020 Apr 16;15(1):94.
[2] Soldaini L, Goharian N. QuickUMLS: a fast, unsupervised approach for medical concept extraction. 2016;4.
[3] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Oct;17(5):507–13.
[4] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-270.
[5] Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017 Jul 15;33(14):i37–48.
[6] Luo L, Yan S, Lai P-T, Veltri D, Oler A, Xirasagar S, et al. PhenoTagger: A Hybrid Method for Phenotype Concept Recognition using Human Phenotype Ontology. Bioinformatics. 2021 Jan 20;btab019.
[7] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018 Nov 27;19(6):1236–46.
[8] Garcelon N, Neuraz A, Salomon R, Bahi-Buisson N, Amiel J, Picard C, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet Journal of Rare Diseases. 2018 May 31;13(1):85.
[9] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:14123555 [cs] [Internet]. 2014 Dec 11 [cited 2022 Jan 14]; Available from: http://arxiv.org/abs/1412.3555
[10] Vincent M, Douillet M, Lerner I, Neuraz A, Burgun A, Garcelon N. Using deep learning to improve phenotyping from clinical reports. Stud Health Technol Inform. 2021 In press;
[11] Jagannatha AN, Yu H. Bidirectional RNN for Medical Event Detection in Electronic Health Records. Proc Conf. 2016 Jun;2016:473–82.
[12] Vasilakes J, Bompelli A, Bishop JR, Adam TJ, Bodenreider O, Zhang R. Assessing the enrichment of dietary supplement coverage in the Unified Medical Language System. Journal of the American Medical Informatics Association. 2020 Oct 1;27(10):1547–55.
[13] Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. J Biomed Inform. 2022 Jan 4;126:103983.
[14] Sarker A. LexExp: a system for automatically expanding concept lexicons for noisy biomedical texts. Bioinformatics. 2021 Aug 25;37(16):2499–501.