Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210241

Identification of Similar Patients Through Medical Concept Embedding from Electronic Health Records: A Feasibility Study for Rare Disease Diagnosis

Xiaoyi CHEN[a,1], Carole FAVIEZ[a], Marc VINCENT[b], Nicolas GARCELON[b], Sophie SAUNIER[c], and Anita BURGUN[a,d,f]
[a] *Centre de Recherche des Cordeliers, Sorbonne Université, INSERM, Université de Paris, F-75006, Paris, France*
[b] *Université de Paris, Imagine Institute, Data Science Platform, INSERM UMR 1163, F-75015, Paris, France*
[c] *Université de Paris, Imagine Institute, Laboratory of Renal Hereditary Diseases, INSERM UMR 1163, F-75015, Paris, France*
[d] *Hôpital Necker-Enfants Malades, Département d'informatique médicale, Assistance Publique-Hôpitaux de Paris (AP-HP), F-75015, Paris, France*
[f] *PaRis Artificial Intelligence Research InstitutE (PRAIRIE), France*

Abstract. To identify patients with similar clinical profiles and derive insights from the records and outcomes of similar patients can help fast and precise diagnosis and other clinical decisions for rare diseases. Similarity methods are required to take into account the semantic relations between medical concepts and also the different relevance of all medical concepts presented in patients' medical records. In this paper, we introduce the methods developed in the context of rare disease screening/diagnosis from clinical data warehouse using medical concept embedding and adjusted aggregations. Our methods provided better preliminary results than baseline methods, with a significant improvement of precision among the top ranked similar patients, which is encouraging for further fine-tuning and application on a large-scale dataset for new/candidate patient identification.

Keywords. Patient similarity, Electronic Health Records, word embedding, rare disease diagnosis

1. Introduction

The rapidly growing of clinical data, mainly electronic health records (EHR), presents unprecedented opportunities for precise and comprehensive patient characterization using high throughput phenotyping. In the context of rare disease, historical EHR data are inestimable sources for patient phenotyping, especially for patients with complex rare diseases. However, it is challenging for clinicians to examine and derive insight from multidimensional, large-scale EHR data [1]. Patient similarity, as a central topic

[1] Corresponding Author: Xiaoyi Chen, Centre de Recherche des Cordeliers, INSERM, 15 rue de l'école de médecine, Paris, France; E-mail: xiaoyi.chen@inserm.fr

in precision medicine to stratify patients into clinically meaningful subgroups, has also been considered in rare disease domains, aiming to identify patients with similar medical histories, diagnoses and outcomes, and derive insights from the records of similar patients to help personalized predictions [2].

The basic similarity model is vector space model (VSM), considering cosine-type similarity between patients represented as vectors of all his/her phenotypes. However, it does not take into account the dependence between phenotypes. To address this issue, the information content (IC)-based semantic similarities, such as Resnik [3] and Lin [4] similarity, are the most considered measures for concept similarity, which require topological structure of concepts in an ontological base. Another developing idea is word embedding, considering a vector representation of medical concept that encodes its meaning such that concepts with similar meaning are closer in the vector space, such as cui2vec learned from large collection of clinical notes [5], and HPO2vec learned from heterogeneous knowledge sources [6]. To obtain patient similarity, concept similarity should be aggregated, either through the average best match method as suggested in [7], or using average embeddings as suggested in [6].

In the context of C'IL-LICO program, we aim to develop transformative diagnostic, prognostic and therapeutic approaches for patients suffering from ciliopathies, a group of rare, severe and multi-systemic diseases caused by ciliary dysfunction. In this study, we describe the methods that we have developed for identifying similar patients from clinical data warehouse using medical concept embeddings and adjusted aggregations.

2. Materials and methods

2.1. NPH patients and Dr. Warehouse UMLS phenotyping

The Necker Children's Hospital is a national reference center for rare and undiagnosed diseases, hosting the Imagine research institute. The research data repository contains more than 600 patients with genetic diagnosis of nephronophthisis (NPH)-related diseases or syndromes in the NPH cohort, which is one of the major ciliopathy cohorts. The clinical data warehouse (Dr. Warehouse) [8] of Necker/Imagine Institute contains EHR data from more than 700,000 patients. The high throughput phenotyping module within Dr. Warehouse is based on the extraction of phenotypes encoded with the Unified Medical Language System (UMLS), because of its large coverage of medical concepts in French language.

As we considered the task of identifying NPH-related ciliopathies from other diseases, we selected NPH patients with sufficient information in EHRs, and re used the control patients with overlapping phenotypes described previously [9]. As renal involvement is one of the most frequent manifestations in NPH-related ciliopathies, this control cohort included patients with any kidney disease (C0022658 and all its descendants) excluding ciliopathies.

2.2. Concept similarity with clinical concept embeddings

To obtain UMLS concept embedding, first, a skip-gram fastText model [10] was trained with default parameters on a collection of 2.5 million clinical narratives from Dr. Warehouse to compute word embeddings of dimension 300 that reflect linguistic

contexts of each word. Then all labels for each UMLS concept (e.g. C0278511 with two French synonymous terms/labels "sarcome ostéogène localisé" and "ostéosarcome localisé") were tokenized, the token embeddings were averaged to obtain label embeddings, and label embeddings were averaged to obtain the final concept embeddings. The semantic similarity between two concepts p_1 and p_2 was calculated using cosine similarity between two embeddings, denoted as $sim_{concept}\ (p_1, p_2)$.

2.3. Patient-patient similarity

To aggregate concept similarity into patient similarity, we first considered the average best match methods, i.e. for each concept in patient P_1, looking for the best matching concept in patient P_2, and averaging the best match similarity over all concepts in P_1. More precisely, for two sets of concepts (i.e., two patients), P_1 and P_2, the asymmetric set similarity is obtained with the formula:

$$sim_{set}(P_1 \rightarrow P_2) = \frac{1}{|P_1|} \sum_{p_{1i} \in P_1} \max_{p_{2j} \in P_2} sim_{concept}(p_{1i}, p_{2j}), \quad (1)$$

and the symmetric set similarity is the average of the two asymmetric set similarities: $sim_{set}(P_1, P_2) = \frac{1}{2}\left(sim_{set}(P_1 \rightarrow P_2) + sim_{set}(P_2 \rightarrow P_1)\right)$. Here each concept in the set contributes equally to the set similarity. However, with the EHR data, both the number of patients per concept and the number of concepts per patient can be very heterogeneous. The relevant and specific concepts should be aggregated differently with higher weights. We thus developed the adjusted average best match methods:

$$sim_{set}(P_1 \rightarrow P_2) = \frac{1}{|\boldsymbol{a_1} P_1|} \sum_{p_{1i} \in P_1} a_{1i} * \max_{p_{2j} \in P_2} sim_{concept}(p_{1i}, p_{2j}), \quad (2)$$

where $\boldsymbol{a_1}$ can present different weights of concepts in the set P_1. Three weighting methods were considered: (i) by distribution of phenotypes over patients (normalization 1), (ii) by distribution of patients over phenotypes (normalization 2), and (iii) by term frequency-inverse document frequency (tf-idf), as suggested in [9].

2.4. Performance evaluation

The EHR data of diagnosed ciliopathies and random controls were pooled. Patient-patient similarity was calculated for all patients. The patients are ranked by computing the average similarity with diagnosed ciliopathy patients, except his/herself, i.e. the diagonal similarity was removed to avoid the bias induced by the high similarity between the diagnosed ciliopathy patient and him/herself. The top k patients who are most similar with ciliopathies were predicted as highly suspected ciliopathies. The performance was assessed using metrics such as precision at k, i.e. the proportion of similar patients in the top-k set that are ciliopathies.

3. Results

3.1. Patient selection

The ciliopathy data set was composed of 79 patients with at least 6 distinct UMLS concepts extracted in EHRs. A set of 200 patients was randomly sampled from the

control cohort for this preliminary study, with the same restriction of minimum 6 concepts. The average number of distinct UMLS concepts was 64 for ciliopathy patients (max: 230, interquartile range: 15-90), and 55 for controls (max: 246, interquartile range: 17-76). The pooled dataset consisted thus of 279 patients, presenting 2593 distinct UMLS concepts.

3.2. Evaluation results

The precisions at k for k=30, k=50 and k=100 are shown in Table 1 for the baseline VSM methods and our methods using embeddings and different adjusted aggregations. The area under the receiver operating characteristic (ROC) curves (AUC) is provided as well, although it is less informative for diagnosis task. The partial ROC curves were plotted in Figure 1 for $k \leq 30$. The results showed that with a similar level of AUC, our developed methods with embeddings and adjusted aggregations could largely improve the performance for small values of k.

Table 1. Evaluation results

	Precision@k			AUC
	k=30	**k=50**	**k=100**	
VSM_norm1	0.43	0.46	0.36	0.62
VSM_norm2	0.53	0.48	**0.45**	**0.70**
VSM_tf-idf	0.50	0.48	0.43	0.69
Embedding	0.53	0.42	0.39	0.64
Embedding_norm1	0.50	0.46	0.38	0.61
Embedding_norm2	**0.70**	**0.60**	0.43	0.69
Embedding_tf-idf	0.67	0.58	0.43	0.69

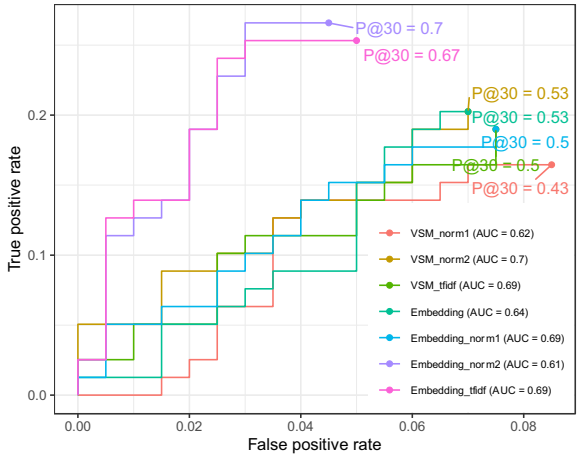

Figure 1. Partial ROC curves for k≤30. The precision@30 (P@30) is given at the end of partial ROC curves.

4. Discussion

In this article, we described the approaches that we have developed for identifying similar patients from clinical data warehouse using medical concepts embeddings and adjusted aggregations, which have taken into account the semantic relations between

medical concepts and the different importance of concepts. Our methods provided better preliminary results than the baseline methods, especially for small value of k, which is of particular interest as we aim to identify suspected ciliopathy patients at top ranked list who would benefit from genetic testing.

Our study had several limitations. As mentioned previously, the IC-based semantic similarity was often considered using for example Human Phenotype Ontology (HPO) as an ontological base [2] [7]. We did not use HPO for phenotyping because (i) it does not exist yet a reliable French version, and (ii) as a versatile thesaurus, the UMLS provides larger coverage of medical concepts. For the same reason as the latter, the UMLS network is less formally structured for semantic similarity. It would be interesting to compare IC-based methods and embedding-based methods. The embeddings could be improved with fine-tuning on more specific clinical records.

The next steps will be using fine-tuned medical concept embeddings to further improve the performance, and applying the patient similarity methods on a larger scale dataset for new/candidate patient identification.

Acknowledgement This work was supported by state funding from The French National Research Agency under the C'IL-LICO project (Reference: ANR-17-RHUS-0002), which was approved by the French National Ethics and Scientific Committee for Research, Studies and Evaluations in the field of Health (CESREES) under the number #2201437.

References

[1] Sharafoddini A, Dubin JA, Lee J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. JMIR Med Inform. 2017 Mar 3;5(1):e7.

[2] Faviez C, Chen X, Garcelon N, Neuraz A, Knebelmann B, Salomon R, Lyonnet S, Saunier S, Burgun A. Diagnosis support systems for rare diseases: a scoping review. Orphanet J Rare Dis. 2020 Apr 16;15(1):94.

[3] Resnik P, Using information content to evaluate semantic similarity in a taxonomy, Proceedings of the 14th International Joint Conference on Artificial intelligence. Volume 1. Aug. 1995; San Francisco, CA, USA; p. 448–53.

[4] Lin D, An Information-Theoretic Definition of Similarity, Proceedings of the Fifteenth International Conference on Machine Learning, Jul. 1998; San Francisco, CA, USA; p. 296-304.

[5] Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer N, Shi X, Cai T, Kohane IS. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. Pac Symp Biocomput. 2020;25:295-306.

[6] Shen F, Peng S, Fan Y, Wen A, Liu S, Wang Y, Wang L, Liu H. HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the Human Phenotype Ontology. J Biomed Inform. 2019 Aug;96:103246.

[7] P Peng J, Li Q, Shang X. Investigations on factors influencing HPO-based semantic similarity calculation. J Biomed Semantics. 2017 Sep 20;8(Suppl 1):34.

[8] Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, Munnich A, Burgun A, Rance B. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform. 2018 Apr;80:52-63.

[9] Chen X et al., Phenotypic similarity for rare disease: Ciliopathy diagnoses and subtyping. J Biomed Inform. 2019 Dec;100:103308.

[10] Joulin A, Grave E, Bojanowski P, Mikolov T, Bag of Tricks for Efficient Text Classification, ArXiv160701759 Cs, Aug. 2016. Available: http://arxiv.org/abs/1607.01759.