Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230342

Improving Patient Similarity Using Different Modalities of Phenotypes Extracted from Clinical Narratives

Xiaoyi CHEN[a,b,c,1], Carole FAVIEZ[b,c], Marc VINCENT[a], Sophie, SAUNIER[d], Nicolas GARCELON[a,b,c] and Anita BURGUN[b,c,e,f]
[a] *Data Science Platform, Imagine Institute, Université de Paris Cité, Inserm UMR 1163, Paris, France*
[b] *Inserm, Centre de Recherche des Cordeliers, Sorbonne Université, Université de Paris Cité, Paris, France*
[c] *HeKA, Inria Paris, Paris, France*
[d] *Laboratory of Renal Hereditary Diseases, Imagine Institute, Université de Paris Cité, Inserm UMR 1163, Paris, France*
[e] *Hôpital Necker-Enfants Malades, Département d'informatique médicale, Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France*
[f] *PaRis Artificial Intelligence Research InstitutE (PRAIRIE), France*
ORCiD ID: Xiaoyi Chen https://orcid.org/0000-0002-7378-5158

Abstract. In the context of medical concept extraction, it is critical to determine if clinical signs or symptoms mentioned in the text were present or absent, experienced by the patient or their relatives. Previous studies have focused on the NLP aspect but not on how to leverage this supplemental information for clinical applications. In this paper, we aim to use the patient similarity networks framework to aggregate different phenotyping modalities. NLP techniques were applied to extract phenotypes and predict their modalities from 5470 narrative reports of 148 patients with ciliopathies (a group of rare diseases). Patient similarities were computed using each modality separately for aggregation and clustering. We found that aggregating negated phenotypes improved patient similarity, but further aggregating relatives' phenotypes worsened the result. We suggest that different modalities of phenotypes can contribute to patient similarity, but they should be aggregated carefully and with appropriate similarity metrics and aggregation models.

Keywords. patient similarity, deep phenotyping, negated phenotype, experiencer

1. Introduction

Medical concept extraction from clinical narratives is an important subdomain of biomedical natural language processing (NLP), enabling applications ranging from clinical decision support to care quality improvement [1]. In addition to concept detection, normalization and disambiguation, extraction of context information, such as negation and experiencer (referred to as "modalities"), is critical for determining whether mentioned clinical signs were present or absent, experienced by the patient or by their

[1] Corresponding Author: Xiaoyi Chen, E-mail: xiaoyi.chen@institutimagine.org

relatives, which can have a high impact in clinical applications. Although recent efforts have improved modality prediction accuracy [2], little attention has been given to using this supplemental information in subsequent work. Previously, we used this metadata to keep only the patient's own present phenotypes when computing patient similarity for rare disease diagnosis. Recently, Slater et al. evaluated this choice in classifying primary diagnosis over MIMIC-III patient visits and confirmed its effectiveness [3].

Similar clinical characteristics are believed to be indicative of similar clinical outcomes [4]. This study hypothesizes that negated phenotypes and family histories can also contribute to patient similarity, as the negation may reflect the assumptions of clinicians, and family history may also be useful in predicting outcomes for health conditions related to genetic inheritance. Therefore, rather than simply removing negated phenotypes and relatives' phenotypes, we aim to aggregate different phenotyping modalities into similarity models. Patient similarity network (PSN) framework has been often considered for heterogeneous multi-omics data aggregation (such as mRNA expression, DNA methylation, etc.) [4], where all types of data are converted to a single type of input (similarity networks), integration is straightforward [5].

In this study, we explored the feasibility of using PSNs framework to improve patient similarity using different phenotyping modalities, namely negated phenotypes and phenotypes experienced by patient's family members. The study was conducted as part of the C'IL-LICO program, aiming to develop transformative diagnostic, prognostic and therapeutic approaches for patients suffering from ciliopathies, a group of rare diseases caused by ciliary dysfunction. The proposed method was evaluated in the context of stratifying ciliopathy patients into subgroups using deep phenotyping in their unstructured electronic health records (EHRs).

2. Materials and methods

2.1. Patient selection and ciliopathies subtypes

The joint data warehouse of Necker Children's Hospital and Imagine Institute, called Dr Warehouse, holds over 9 million documents of 800,000 patients, and structured data (gene, diagnosis, manually curated phenotypes) for more than 1200 patients with ciliopathies. This study focused on the 148 diagnosed patients at Necker Children's Hospital with sufficient documents, involving 47 ciliary genes and 26 Orphanet encoded diagnoses. The gene-diagnosis combination created 64 classes, with 54 classes having less than 3 patients. We thus grouped the genes based on their function and localization within the cilium based on [9]. Overlapping diagnoses were also grouped. The final class assignment was validated by a ciliopathy expert (SS), which resulted in five classes.

2.2. Clinical concept extraction and modality prediction

For phenotype extraction, a hybrid strategy combining a dictionary-based approach and a deep-learning approach using bidirectional Gated Recurrent Units and Conditional Random Fields (biGRU-CRF) model was adopted, representing extracted mentions as concepts in the Human Phenotype Ontology (HPO). For modality prediction, a deep learning pipeline was developed using fastText and contextual Bidirectional Encoder Representations from Transformers (BERT)-type embeddings, combined with GRU or

Long Short Term Memory (LSTM) recurrent neural networks, which were shown outperforming the rule-based approaches in negation and subject prediction tasks [6].

2.3. Patient similarity networks using different modalities of phenotyping

For each phenotype modality, i.e., patient's positive (pt_pos) and negative (pt_neg), family's positive (fm_pos) and negative (fm_neg), patient similarities were computed using the average best match method as described in [7]. More precisely, for two patients represented by two sets of concepts $P_1$and P_2, the similarity from patient P_1 to P_2 is the weighted average of the best-match concept similarities over all concepts in P_1:

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

where $\boldsymbol{a_1}$ is the weight vector indicating the relevance of each phenotype to the patient. The symmetric similarity is $sim_{set}(P_1, P_2) = \frac{1}{2}\left(sim_{set}(P_1 \to P_2) + sim_{set}(P_2 \to P_1)\right)$. Regarding the similarity between concepts, we considered the Lin's semantic similarity [8], which is based on the information content (IC) of the two phenotypes and the IC of their lowest common subsumer (LCS) in the HPO hierarchy:

$$sim_{concept}(p_i, p_j) = \frac{2 \times \mathrm{IC}\left(\mathrm{LCS}(p_i, p_j)\right)}{\mathrm{IC}(p_i) + \mathrm{IC}(p_j)}. \quad (2)$$

Four PSNs were built using the computed patient similarity matrix, with patients as nodes and similarities as weighted edges. The $\mathrm{PSN_{pt\_pos}}$ was considered as a baseline. Then the other PSNs were aggregated successively by considering a weighted average combination of the current network and previously aggregated networks: $\mathrm{PSN_{agg2}} = \alpha\mathrm{PSN_{pt\_pos}} + (1-\alpha)\mathrm{PSN_{pt\_neg}}$, $\mathrm{PSN_{agg3}} = \beta\mathrm{PSN_{agg2}} + (1-\beta)\mathrm{PSN_{fm\_pos}}$, and so on.

2.4. Clustering and evaluation for patient stratification

Hierarchical agglomerative clustering with complete-linkage was applied on the baseline PSN and on each aggregated PSN. The number of clusters was fixed to 5, which equals to the number of grouped gene-diagnosis classes. Rand index (RI) and adjusted RI (ARI) were used to measure the concordance with the ground truth class assignment. More precisely, $\mathrm{RI} = (\mathrm{a} + \mathrm{b})/C_2^N$, where a is the number of pairs of patients that are in the same gene-diagnosis class and in the same cluster, b is the number of pairs of patients that are in different classes and in different clusters, and C_2^N is the number of all possible pairs of patients. The ARI is the RI discounted by the expected RI of random labelings:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}(\mathrm{RI})}{\max(\mathrm{RI}) - \mathbb{E}(\mathrm{RI})}. \quad (3)$$

Therefore, the range be definition of RI and ARI are [0,1] and [-1,1], respectively.

3. Results

3.1. Different phenotype modalities and PSNs

After applying the NLP techniques, we identified 2157 distinct pt_pos phenotypes (extracted from 5470 narrative reports from 148 ciliopathy patients), 879 distinct pt_neg phenotypes (3366 reports of 131 patients), 275 distinct fm_pos phenotypes (575 reports

of 66 patients), and 49 distinct fm_neg phenotypes (124 reports of 21 patients). Due to the small number of patients with fm_neg phenotypes, patient similarities were computed only for the first three modalities, and the aggregation was limited to these three PSNs. Figure 1 showed the PSNs with the Fruchterman-Reingold layout. To facilitate the visualization, only the top 5% of the weighted edges for each node are displayed. Nodes were colored by the ground truth class assignment.

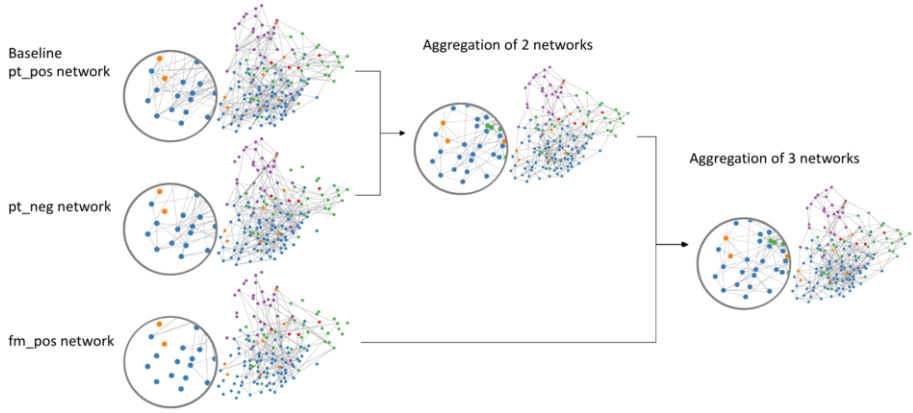

Figure 1. Individual and aggregated networks with a zoom on the bottom left.

3.2. Clustering performance

The RI and ARI were computed for clustering using PSN_{pt_pos} and each aggregated PSN, the results obtained with the best aggregation weights that were determined via grid search for each aggregation are shown in Table 1. We can observe that aggregating patient's negated phenotypes improved the performance, but further aggregating family's phenotypes worsened the result. In order to better understand the performance improvement using PSN_{agg2}, we applied the same clustering method with the same similarity model to manually curated phenotypes in the structured research data. In contrast to the automated phenotype extraction from narrative reports, manually curated phenotypes are comprehensive, precise and relevant, thus can be considered as the best phenotypic representation of patient. The RI and ARI using the manually curated phenotypes was 0.784 and 0.514, respectively, which are not close to 1, due to the phenotypic and genetic heterogeneity and overlap of ciliopathies. Given these values as references, the relative increase in RI and ARI towards the references is 43% and 66%, respectively, showing an important improvement. An expert (SS) reviewed the discrepancies of the clustering results obtained from PSN_{pt_pos} and PSN_{agg2}, and confirmed the improvement by aggregating patient similarity on negated phenotypes. For example, a better stratification of isolate and syndromic Leber congenital amaurosis (LCA) caused by different genes was achieved using PSN_{agg2}.

Table 1. Clustering performance for the baseline PSN (pt_pos), aggregated PSNs, and the PSN using manual curated phenotypes

	Baseline pt_pos	**Agg2 +pt_neg**	**Agg3 +pt_neg+fm_pos**	**Reference manually curated phenotypes**
Rand index	0.735	**0.756**	0.721	0.784
adj. Rand index	0.393	**0.473**	0.349	0.514

4. Discussion

Gliozzo et al. [4] distinguished three aggregation approaches in the recent review: input data-fusion, PSN-fusion, and output-fusion. Input data-fusion is unsuitable for negations and family histories, since the same phenotype can be present and absent for the same patient at different times, and for the patient and also for the relatives. Output-fusion is also not possible, since data in some modalities can be too sparse for a reliable prediction. Therefore, an intermediate integration was considered to compute patient similarities using each modality independently, then aggregating them successively. While the lack of ground truth makes evaluating clustering performance challenging, expert grouped gene-diagnosis classes were established, and RI and ARI using manually curated phenotypes were provided as a reference. Our results showed significant improvement of ARI when aggregating negated phenotypes, but worse performance with family's phenotypes. We think that is because most of the ciliopathies in our dataset are recessive disorders, and the subject prediction can only distinguish whether the experiencer is the patient or not, without further precision (parents, siblings, or other family members).

Our study has some limitations, which will be addressed in future work. As the first attempt of aggregating different phenotyping modalities into patient similarity, we used the same similarity metric for negated phenotypes as for positive phenotypes, and a simple aggregation model, i.e., a weighted sum of individual similarities obtained on each modality. The evaluation was conducted on a small dataset. The next step will be to explore other similarity metrics that may suit better negation and family history, investigate more sophisticated aggregation models, and perform a broader evaluation involving also dominant disorders to assess the impact of family history.

This work was supported by State funding from The French National Research Agency (ANR) under "Investissements d'Avenir" programs (ANR-10-IAHU-01) and C'IL-LICO project (ANR-17-RHUS-0002).

References

[1] Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, Shen F, Wang L, Wang Y, Wen A, Zhao Y, Sohn S, Liu H. Clinical concept extraction: A methodology review. J Biomed Inform 2020 Sep;109:103526. PMID:32768446

[2] Slater LT, Bradlow W, Motti DF, Hoehndorf R, Ball S, Gkoutos GV. A fast, accurate, and generalisable heuristic-based negation detection algorithm for clinical text. Comput Biol Med 2021 Mar;130:104216. PMID:33484944

[3] Slater LT, Karwath A, Hoehndorf R, Gkoutos GV. Effects of Negation and Uncertainty Stratification on Text-Derived Patient Profile Similarity. Front Digit Health 2021 Dec 6;3:781227.

[4] Gliozzo J, Mesiti M, Notaro M, Petrini A, Patak A, Puertas-Gallardo A, Paccanaro A, Valentini G, Casiraghi E. Heterogeneous data integration methods for patient similarity networks. Brief Bioinform 2022 Jul 18;23(4):bbac207. PMID:35679533

[5] Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. netDx: interpretable patient classification using integrated patient similarity networks. Mol Syst Biol 2019 Mar 14;15(3):e8497. PMID:30872331

[6] 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 2022 Jun 6;290:282–286.

[7] Chen X, Faviez C, Vincent M, Garcelon N, Saunier S, Burgun A. Identification of Similar Patients Through Medical Concept Embedding from Electronic Health Records: A Feasibility Study for Rare Disease Diagnosis. Stud Health Technol Inform 2021 May 27;281:600–604. PMID:34042646

[8] Lin D. An Information-Theoretic Definition of Similarity. Proc Fifteenth Int Conf Mach Learn San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1998. p. 296–304.

[9] Reiter JF, Leroux MR. Genes and molecular pathways underpinning ciliopathies. Nat Rev Mol Cell Biol 2017 Sep;18(9):533–547. PMID:28698599