Journal of Biomedical Informatics 73 (2017) 51–61

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack

Nicolas Garcelon [a,b,c,*], Antoine Neuraz [c,d], Vincent Benoit [a,b], Rémi Salomon [a,b,e], Sven Kracker [a,b,f], Felipe Suarez [a,b,g], Nadia Bahi-Buisson [a,b,h], Smail Hadj-Rabia [a,b,i], Alain Fischer [a,b,j,k,l], Arnold Munnich [a,b,m,n], Anita Burgun [c,d,o]

[a] Institut Imagine, Paris Descartes Université Paris Descartes-Sorbonne Paris Cité, Paris, France
[b] INSERM, Institut Imagine, UMR 1163, Université Paris Descartes, Sorbonne Paris Cité, Paris, France
[c] INSERM, Centre de Recherche des Cordeliers, UMR 1138 Equipe 22, Université Paris Descartes, Sorbonne Paris Cité, Paris, France
[d] Département d'informatique médicale, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[e] Service de Néphrologie Pédiatrique, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[f] Laboratory of Human Lymphohematopoiesis, INSERM UMR 1163, Paris, France
[g] Service de Hématologie, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[h] Service de neurologie pédiatrique, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[i] Service de Dermatologie, Centre de Références maladies Génétiques à Expression Cutanée (MAGEC), Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[j] Centre de Référence Déficits Immunitaires Héréditaires, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[k] Unité d'Immunologie-Hématologie et Rhumatologie Pédiatrique, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[l] College de France, Paris, France
[m] Département de génétique médicale, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[n] Centre de Référence des Maladies Osseuses Constitutionnelles, INSERM UMR 1163, Laboratoire de bases moléculaires et physiopathologiques de l'ostéochondrodysplasie, Paris Descartes-Sorbonne Paris Cité University, AP-HP, Institut Imagine, 75015 Paris, France
[o] Hôpital Européen Georges Pompidou, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France

ARTICLE INFO

Article history:
Received 24 February 2017
Revised 5 July 2017
Accepted 24 July 2017
Available online 25 July 2017

Keywords:
Data warehouse
Vector space model
Similarity measures
Rare diseases
Electronic health records

ABSTRACT

Objective: In the context of rare diseases, it may be helpful to detect patients with similar medical histories, diagnoses and outcomes from a large number of cases with automated methods. To reduce the time to find new cases, we developed a method to find similar patients given an index case leveraging data from the electronic health records.

Materials and methods: We used the clinical data warehouse of a children academic hospital in Paris, France (Necker-Enfants Malades), containing about 400,000 patients. Our model was based on a vector space model (VSM) to compute the similarity distance between an index patient and all the patients of the data warehouse. The dimensions of the VSM were built upon Unified Medical Language System concepts extracted from clinical narratives stored in the clinical data warehouse. The VSM was enhanced using three parameters: a pertinence score (TF-IDF of the concepts), the polarity of the concept (negated/not negated) and the minimum number of concepts in common. We evaluated this model by displaying the most similar patients for five different rare diseases: Lowe Syndrome (LOWE), Dystrophic Epidermolysis Bullosa (DEB), Activated PI3K delta Syndrome (APDS), Rett Syndrome (RETT) and Dowling Meara (EBS-DM), from the clinical data warehouse representing 18, 103, 21, 84 and 7 patients respectively.

* Corresponding author at: Imagine – Institute for genetic diseases, 24 boulevard du Montparnasse, 75015 Paris, France.
E-mail address: nicolas.garcelon@institutimagine.org (N. Garcelon).

http://dx.doi.org/10.1016/j.jbi.2017.07.016
1532-0464/

Results: The percentages of index patients returning at least one true positive similar patient in the Top30 similar patients were 94% for *LOWE*, 97% for *DEB*, 86% for *APDS*, 71% for *EBS-DM* and 99% for *RETT*. The mean number of patients with the exact same genetic diseases among the 30 returned patients was 51%.
Conclusion: This tool offers new perspectives in a translational context to identify patients for genetic research. Moreover, when new molecular bases are discovered, our strategy will help to identify additional eligible patients for genetic screening.

1. Introduction

There are approximately 7000 different types of rare diseases and disorders affecting a large population worldwide [1]. Collecting and analyzing patient data is critical aspect of rare and undiagnosed diseases programs that has the potential to increase the opportunity of discovering new knowledge. Because of their individual rarity, identifying patients that share similar phenotypes can be particularly challenging. To quote E. Ashley, the co-chair of the steering committee for the Undiagnosed Diseases Program, the challenge with rare disease data "is not so much finding the needle in the haystack as finding the right needle in a whole pile of needles" [2]. It may be helpful to detect patients with similar medical histories, diagnoses and outcomes from a large number of cases with automated methods. For example, when a new causal mutation for a disease is discovered, being able to find potential cases in a retrospective database by looking for patients similar to the 2 or 3 patients already diagnosed and genotyped could reduce their diagnosing journey and confirm the causal association.

The ability to retrospectively mine all patient records has proven beneficial [3]; and it has been made possible with the widespread adoption of clinical data warehouses such as i2b2 [4] and STRIDE [5] in hospitals. In most cases, the users query the data warehouse using a Boolean combination of criteria e.g., a list of signs and phenotypic traits. But in the case of rare complex disorders, it may be difficult to query the data warehouse in the quest for similar patients, with a formal and precise list of symptoms. The search for similar cases in the data warehouse based on similarity metrics would be more powerful and versatile.

We have adapted the vector space approach to provide similarity metrics between rare disease patient medical records. In this paper, we present the method, its implementation and its evaluation on the Necker-*Enfants Malades*/*Imagine* data warehouse.

2. Background and significance

Both supervised and unsupervised machine learning methods [6] have been used to compute patient similarity, e.g., support vector machines. All these methods are based on learning models thus require training sets of a sufficient number of cases.

The Vector space model (VSM) was first proposed by Salton in 1968 [7]. Then it was implemented in SMART [8], an information retrieval system that computes similarity between documents represented as vectors of keywords. The matrix (documents by indexing terms) consists of binary values indicating the presence (1) or absence (0) of a term in a document. Salton demonstrated in 1968 the noticeable improvement in performance by using the term frequency weight [9] instead of binary values. Spärck Jones (1972) demonstrated the interest of the frequency of a term in a collection [10] and introduced the tf-idf (term frequency – inverse document frequency) weight.

Since SMART, the VSMs have been broadly used in information retrieval [11–17], in classification [18,19], and clustering [20]. The VSM has also been used in other applications. For example in social network analysis, Lee et al. used the VSM to find new network ties [21]. Santos et al. used Topic-based VSM approach to enhance a spam filter [22]. Castells et al. also used ontology in association with the VSM to improve the ranking of a search engine [23].

The VSM has also been used to compute similarity in the biomedical domain.

With the objective of identifying potentially related diseases based on genetic relationships, Sarkar et al. [24] proposed an adaptation of the VSM that bridges gene and disease knowledge inferred across three knowledge bases: Online Mendelian Inheritance in Man, GenBank, and Medline. In this study, the relatedness between diseases, via this network of gene-based relationships, was determined using a cosine similarity metric. The authors concluded that VSM was a potentially powerful method for exploring the complex landscape of polygenetic diseases. Lee et al. [25] used a VSM and applied a cosine-similarity-based patient similarity metrics to an intensive care unit database to identify patients who are most similar to each index patient and predict their outcomes. They applied a VSM to MIMIC II structured data, including ICD-9 codes and quantitative data, and showed that their approach outperformed the standard severity scores usually used in the intensive care units.

In a study published in 2013, we applied a VSM approach to identify surgical site infections after neurosurgical procedures in full-text reports [26]. The method applied to patient narrative documents achieved a high recall score (92%) and a precision of 40%, much higher than the same approach based on ICD-10 codes (85% and precision 5%). These results are consistent with several studies that demonstrated that information extraction from unstructured clinical narratives is essential to most clinical applications [27]. For example, structured data alone is insufficient in resolving eligibility criteria for recruiting patients onto clinical studies [28].

All of these studies have suggested that the VSM approach can be effective at representing and computing similarity between patient reports. In the next section, we describe the VSM-based system that we have developed to search for similar patients attending the *Imagine* Institute, a research and healthcare institute focusing on genetic and rare diseases associated with the *Necker-Enfants Malades* Hospital in Paris.

3. Methods

The goal of this study was to explore the potential of using a VSM approach to identify potentially similar patients in a rare disease data warehouse.

3.1. Data warehouse

Necker Enfants Malades Hospital is an Assistance Publique-Hôpitaux de Paris (AP-HP) children's hospital located in Paris, France. The hospital is specialized in rare diseases and is associated with a research institute, the *Imagine* Institute of Genetic Diseases. The hospital and the institute hold a joint clinical data warehouse,