BMC Medical Research Methodology

RESEARCH

Open Access

Improving patient clustering by incorporating structured variable label relationships in similarity measures

Judith Lambert[1,2,3*], Anne-Louise Leutenegger[4], Anaïs Baudot[3,5,6†] and Anne-Sophie Jannot[2,7,8†]

Abstract

Background Patient stratification is the cornerstone of numerous health investigations, serving to enhance the estimation of treatment efficacy and facilitating patient matching. To stratify patients, similarity measures between patients can be computed from clinical variables contained in medical health records. These variables have both values and labels structured in ontologies or other classification systems. The relevance of considering variable label relationships in the computation of patient similarity measures has been poorly studied.

Objective We adapt and evaluate several weighted versions of the Cosine similarity in order to consider structured label relationships to compute patient similarities from a medico-administrative database.

Materials and methods As a use case, we clustered patients aged 60 years from their annual medicine reimbursements contained in the *Échantillon Généraliste des Bénéficiaires*, a random sample of a French medico-administrative database. We used four patient similarity measures: the standard Cosine similarity, a weighted Cosine similarity measure that includes variable frequencies and two weighted Cosine similarity measures that consider variable label relationships. We construct patient networks from each similarity measure and identify clusters of patients using the Markov Cluster algorithm. We evaluate the performance of the different similarity measures with enrichment tests based on patient diagnoses.

Results The weighted similarity measures that include structured variable label relationships perform better to identify similar patients. Indeed, using these weighted measures, we identify more clusters associated with different diagnose enrichment. Importantly, the enrichment tests provide clinically interpretable insights into these patient clusters.

Conclusion Considering label relationships when computing patient similarities improves stratification of patients regarding their health status.

Keywords Prior expert knowledge, Structured variable labels, Patient stratification, Patient clustering, Patient networks, Similarity measures

[†]Anaïs Baudot and Anne-Sophie Jannot contributed equally to this work.

*Correspondence:
Judith Lambert
judith.lambert@inserm.fr

Full list of author information is available at the end of the article

Background

Identifying similar patients can serve multiple purposes in healthcare. In routine practice, finding patients similar to a given patient can help elucidate undiagnosed cases, particularly in the case of rare diseases [1]. Similar patients can also provide prognostic guidance. In research, identifying groups of similar patients is useful to stratify the population. This enables, for instance, a more precise estimation of medicine efficacy for a given patient profile or matching similar patients in case-control studies [2, 3].

The similarity between patients can be computed from medical health records such as medico-administrative databases. Medico-administrative databases contain data used for health care reimbursement purposes, including information about hospitalization, medicine and medical device consumption. They therefore provide a comprehensive perspective on the entire healthcare pathway of a given patient.

The variables contained in these databases are labeled by terms that can be related to each other. For instance, medicines are labeled by codes organized into classification trees such as the Anatomical Therapeutic Chemical (ATC) classification. Within this classification, all antidiabetic medicines belong to the same class. Thus, the labels of two medicines used to treat diabetes are related according to the classification, and patients treated with these medicines are expected to be more similar than patients treated with medicines from different classes. Other classification systems are available for various types of medical data. For instance, the International Classification of Diseases (ICD) is used for diagnoses [4] and SNOMED-CT is used for clinical information [5].

The objective of similarity measures is to identify patients sharing characteristics, such as taking the same medicine or having the same diagnoses at a specific age. The most commonly used measures to compute similarities between patients are the Euclidean distance, the Jaccard index, and the Cosine similarity [6]. Weighted measures can also be incorporated to these similarity measures to account for the frequency of the variables. For instance, the Inverse Document Frequency is a weighted measure that assigns greater importance to rare variables [7, 8]. However, these classical similarity measures rely only on the variable values and do not consider other information associated with the variables, such as their labels. For instance, when considering medications, two patients are similar if they take similar dosage of similar medicine. In this case, similarity is defined at both the dosage level (i.e., variable value) and at the medicine level (i.e., variable label relationships). Hence, analyzing the relationships between variable labels, by leveraging a label classification system, is expected to provide pertinent information for computing patient similarity from a clinical perspective.

Several measures have been proposed to analyze variable label relationships. The Wu and Palmer measure examines the relationships between two variable labels by considering their depth in the classification tree [9] while the Lin measure considers both their depth in the classification tree and their frequency [10]. These measures have been incorporated into the computation of patient similarities in various studies. For instance, Ni et al. compute patient similarities based on their ICD-10 diagnosis, using a weighted similarity measure that considers the classification depth of ICD-10 codes [11]. Girardi et al. adapted the Jaccard distance to include diagnose relationships in patient similarity computation based on the depth of their ICD-10 codes [12]. However, these weighted patient similarity measures are limited to variables associated with binary values (i.e., Boolean variables) and cannot be applied to quantitative variables. To the best of our knowledge, similarity measures able to simultaneously consider variable label relationships from classifications and quantitative values of variables are lacking in the field.

The efficiencies of the similarity measures are usually estimated by assessing the quality of the clusters obtained using the measures. The clustering performance is evaluated with metrics such as silhouette score or accuracy. However, interpreting these performance metrics from a clinical perspective can be challenging. An alternative method to assess the performance of clustering involves using external variables that were not used initially to compute patient similarities. These external variables can be related to prognosis [13] or tumor characteristics [14], for instance.

In this paper, we propose to weight the Cosine similarity to include variable label relationships in order to identify similar patients. We further aim to assess the added value of incorporating this information thanks to an evaluation protocol that can be interpreted clinically. Our study focuses on a specific use case related to medicine reimbursement in a national French medico-administrative database. We first compute several weighted similarity measures and employ them to cluster patients, thereby revealing groups of similar patients. We then assess the performance of the different similarity measures in identifying clusters of patients thanks to enrichment tests based on external variables (i.e., diagnoses). Notably, our study represents a pioneering effort in applying and evaluating these weighted measures within medico-administrative databases. We observed that taking into account the relationships between the variable labels in the computation of patient similarities improves the quality of the identified patient clusters.

Materials and methods

Let I be the set of variables (i.e., medicines). Let X and Y be the vectors of variables from I for two patients. For instance, X and Y could represent the amount of reimbursed medicine at a given age for each type of medicine. We compute the similarity between patients using four different measures. The first two measures (Cosine similarity and Cosine similarity weighted by the Inverse Document Frequency (IDF)) rely on quantitative variables, while the remaining two (Cosine similarity weighted by the Wu and Palmer measure and Cosine similarity weighted by the Lin measure) rely on both quantitative variables and label relationships of the variables. We chose to weight the Cosine similarity by (i) IDF in order to consider variable frequency, (ii) the Wu and Palmer measure because this measure has been demonstrated as effective in the biomedical domain to account for label variable depth in the classification tree [15, 16], and (iii) the Lin measure because this measure demonstrated superior performance compared to other measures considering both variable frequency (through information content) and label variable depth [17].

Defining patient similarity measures

Cosine similarity

The Cosine similarity between two patient vectors X and Y is defined as the cosine of the angle (θ) between the two vectors [18]:

$$\cos_\theta(X,Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} \quad (1)$$

The Cosine similarity values range from −1 to 1, with value equal to -1 when vectors are opposite, 0 when vectors are different (i.e., orthogonal) and 1 when they are identical.

Cosine similarity weighted by the inverse document frequency

The Cosine similarity weighted by the Inverse Document Frequency (IDF) is defined as follows for two patient vectors X and Y [7]:

$$\cos_{\theta,IDF}(X,Y) = \frac{\sum_{i\in I} IDF(i)X_iY_i}{\sqrt{\sum_{i\in I} IDF(i)X_i^2}\sqrt{\sum_{i\in I} IDF(i)Y_i^2}} \quad (2)$$

with $IDF(i) = log\frac{N_I}{N_i}$, where N_I is the total number of observations of all the variables in the set I and N_i is the total number of observations of the variable i.

IDF is commonly associated with Term Frequency (TF) in the TF-IDF measure. Term Frequency denotes the number of occurrences of variables, which is represented by the two patient vectors X and Y in formula (2). Here, we used the IDF to weight the Cosine similarity and not the TF for the sake of comparison of the different weighted and unweighted measures.

As for the standard Cosine similarity, the values of this weighted version range from −1 to 1.

Cosine similarity weighted by the Wu and Palmer measure

The Cosine similarity weighted by the Wu and Palmer measure is defined as follows for two patient vectors X and Y [9]:

$$\cos_{\theta,WP}(X,Y) = \frac{\sum_{i,j\in I} WP(ij)X_iY_j}{\sqrt{\sum_{i,j\in I} WP(i,j)X_iX_j}\sqrt{\sum_{i\in I} WP(ij)Y_iY_j}} \quad (3)$$

with $WP(i, j)$ being the Wu and Palmer measure between the labels of the two variables i and j of the set I.

The labels of the variables of the set I are organized into a classification tree consisting of successive levels (Fig. 1). The label composing the top level is the root and the labels composing the lowest level are the leaves. Each level in the classification is connected to the next level through edges representing the relationship between variable labels in the classification. A sequence of edges represents a path in the classification.

The Wu and Palmer measure is computed from the variable labels as follows:

$$WP(i,j) = \frac{2 \times depth(LCA(i,j))}{depth(i) + depth(j)} \quad (4)$$

with $depth(z) = E_z/E$ where E_z is the number of edges between the root and the variable label z in the classification tree and E is the total depth in the classification tree (i.e., the number of edges in the shortest path from the root to the leaves); $LCA(i, j)$ is the Lowest Common Ancestor of the labels of the variables i and j in the classification (i.e., the lowest label of the variable of set I that has both i and j as descendants). For example, the Wu and Palmer measure between the medicine labels B1 and B22 from the classification of the Fig. 1 is computed as follows:

$$WP(B1,B22) = \frac{2 \times depth(B)}{depth(B1) + depth(B22)} = \frac{2 \times (1/3)}{(2/3) + (3/3)} = 0.40.$$

As for the standard Cosine similarity, the values of this weighted version range from −1 to 1.

Cosine similarity weighted by the Lin measure

The Cosine similarity weighted by the Lin measure is defined as follows for two patient vectors X and Y:

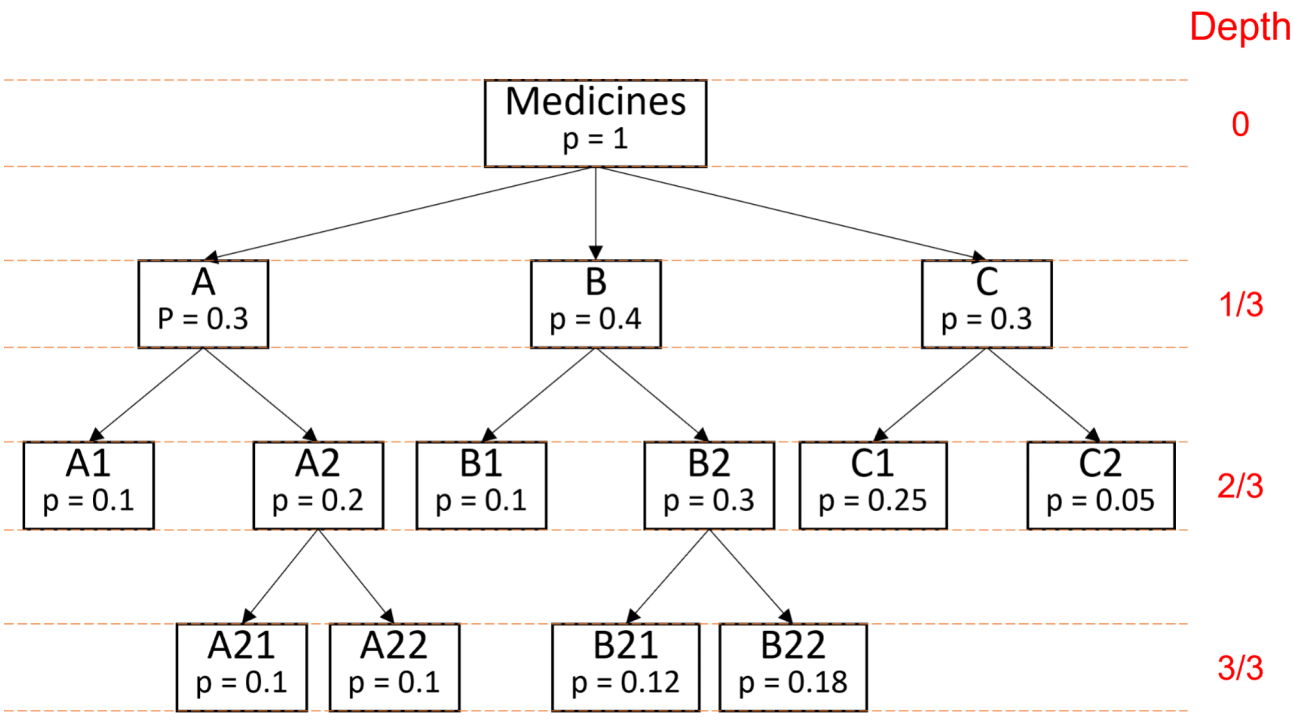

Fig. 1 Example of a classification tree of medicine data. The classification is composed of several medicine labels organized in successive levels interconnected by edges. The depth of a given variable label is the number of edges between the root (i.e., label "Medicines") and that given variable label, divided by the number of edges in the shortest path from the root to the leaves. *p* is the medicine label frequency

$$\cos_{\theta,Lin}(X,Y) = \frac{\sum_{i,j\in I} Lin(ij)X_iY_j}{\sqrt{\sum_{i,j\in I} Lin(i,j)X_iX_j}\sqrt{\sum_{i,j\in I} Lin(ij)Y_iY_j}} \quad (5)$$

With $Lin(i, j)$ the Lin measure between the labels of the two variables i and j of the set I.

The Lin measure analyzes the relationship between two variables i and j by considering the information content (IC) of the labels of the two variables and the information content of their lowest common ancestor [10]:

$$Lin(i,j) = \frac{2 \times IC(LCA(i,j))}{IC(i) + IC(j)} \quad (6)$$

with $IC(LCA(i, j))$ defined as the Resnik measure [19] and $IC(z) = -\log P(z)$ where $P(z)$ is the probability of occurrence of the variable z estimated by its frequency. For example, the Lin measure between the medicine labels B1 and B22 from the classification of the Fig. 1 is computed as follows:

$$Lin(B1,B22) = \frac{2 \times IC(B)}{IC(B1) + IC(B2)} = \frac{2 \times \log(0.4)}{\log(0.1) + \log(0.18)} = 0.46$$

As for the standard Cosine similarity, the values of this weighted version range from −1 to 1.

Identifying clusters of patients from patient networks

Various clustering methods can be used to identify clusters of patients from their similarity measures. Some examples include K-means, hierarchical clustering, or the Markov cluster algorithm applied to patient networks [20, 21]. In a previous work, we showed that building patient networks using Cosine similarity on medicine data and clustering the networks was a pertinent approach to identify patient clusters and trajectories [22]. Therefore, here, we build patient networks using Cosine similarity or its weighted versions on medicine data, and cluster the networks with the Markov Cluster algorithm to identify clusters of patients.

Constructing patient networks

A patient network is a graph $G = (V, E)$ with V patient nodes and E edges representing interactions between patient nodes. The network is constructed using a similarity matrix. Let $M = [m_{X,Y}]^n$ be the similarity matrix where n is the number of patients and $m_{X,Y}$ is the similarity between patient vectors X and Y. This similarity matrix is symmetrical, with $m_{X,Y} = m_{Y,X}$. We compute four similarity matrices, each corresponding to a specific similarity measure. We then apply a threshold t to the similarity matrices to construct the patient networks. Two patients are connected in the network (i.e., an edge between the patients is present) if their similarity is above

the threshold t. The connection between patients X and Y is weighted by the value $m_{X,Y}$ of the matrix.

To ensure comparable networks for the different similarity measures, we select a distinct threshold t for each measure. These thresholds are chosen to obtain approximately 5000 patient nodes in the largest connected component of each network (supplementary Table S1).

Clustering patient networks

We apply the Markov Cluster algorithm (MCL) [23] on the largest connected component of each patient network. The MCL algorithm uses random walks to simulate flows on the network. The flows allow to distinguish network areas where nodes are strongly connected, which correspond to the clusters. We use the version 0.0.6.dev0 of the "markov-clustering" Python package with the default parameters.

Cluster enrichment analysis

Let external variables be binary variables that are not used to compute the similarities between patients. The aim of the enrichment analysis is to assess if each external variable has a frequency higher than expected in a cluster. For each external variable and each cluster of patients, we compare patients inside and outside the cluster using Fisher's exact test [24]. This procedure involves performing a number of tests equal to the product of the number of clusters times the number of external variables. We adjust this multiple testing with the Benjamini-Hochberg procedure. We consider that a variable is enriched in a given cluster if its adjusted p-value is lower than 0.05.

Use-case: the Échantillon Généraliste Des Bénéficiaires

We use health data from the Échantillon Généraliste des Bénéficiaires (EGB), a French medico-administrative database. The EGB is a random sample of the French health insurance database [25]. It is representative of the French population and contains approximately 660,000 individuals followed over a period of 11 years.

We extract from the EGB data on medicine reimbursements between 2008 and 2018 (Fig. 2), including the date of reimbursement and the medicine classification in the Anatomical Therapeutic Chemical (ATC) class (see example Table 1). The ATC class is an international classification of medicines established by the World Health Organization (WHO) [26]. We exploit this classification in the patient similarity measures. We then select patients aged 60 during the study period, defined as those within the age range [60;61[during the years 2008–2018. All included patients were followed for one year after their 60th birthday. We include only patients who had received reimbursement for at least one medicine for two or more consecutive months. We therefore keep only patients with sustained reimbursements. We also extract chronic disease diagnoses declared by the patient to the French health insurance. These diagnoses are coded with the 10*th* revision of the international statistical classification of diseases and related health problems (i.e., using ICD-10 code). We thus exclude from our analysis the patients with no declared chronic diseases. Importantly, diabetes appears as the most frequent chronic disease observed within the population. We analyze female and male patient datasets separately. In each dataset, we calculate for each patient, the number of reimbursements they had for each medicine at age 60 (see example Table 2).

Results

Our two use-case datasets are composed of 8,872 female and 9,765 male patients. For each dataset, we compute the similarity between patients, build networks, and identify clusters. We assess the performance of the different patient similarity measures with enrichment tests on the patient clusters using declared chronic diseases.

Similarity measures including variable label relationships have higher similarity values in the use-case populations

We first compare patient similarities computed from medicine reimbursements using four similarity measures, i.e., the standard Cosine similarity and its weighted versions (*Material and methods 2.1*).

In the dataset of female patients, the Cosine similarity weighted by the Wu and Palmer measure and the Cosine similarity weighted by the Lin measure identify more patient pairs with similarities with non-zero values ($n_0 = 3.89 \times 10^7$ for these two measures) as compared to the Cosine similarity and the Cosine similarity weighted by IDF ($n_0 = 3.02 \times 10^7$ for the two other measures) (Fig. 3). Additionally, the Cosine similarity weighted by the Wu and Palmer measure and the Cosine similarity weighted by the Lin measure show a higher variability. Thus, the weighted Cosine similarity measures that include variable label relationships information have higher similarity values. Similar results are observed in the dataset of male patients (Figure S1).

Similarity measures including variable label relationships improve patient cluster quality

A patient network is constructed for each of the four similarity measures, in both male and female datasets, leading to 8 different patient networks (*Material and methods 2.1*). The Fig. 4 shows the two networks constructed for the dataset of female patients using the Cosine similarity and the Cosine similarity weighted by the Wu and Palmer measure. The network constructed with the Cosine similarity (Fig. 4A) displays a highly connected structure. Conversely, the network constructed with the Cosine similarity weighted by the Wu and Palmer measure (Fig. 4B) reveals distinct subnetworks.

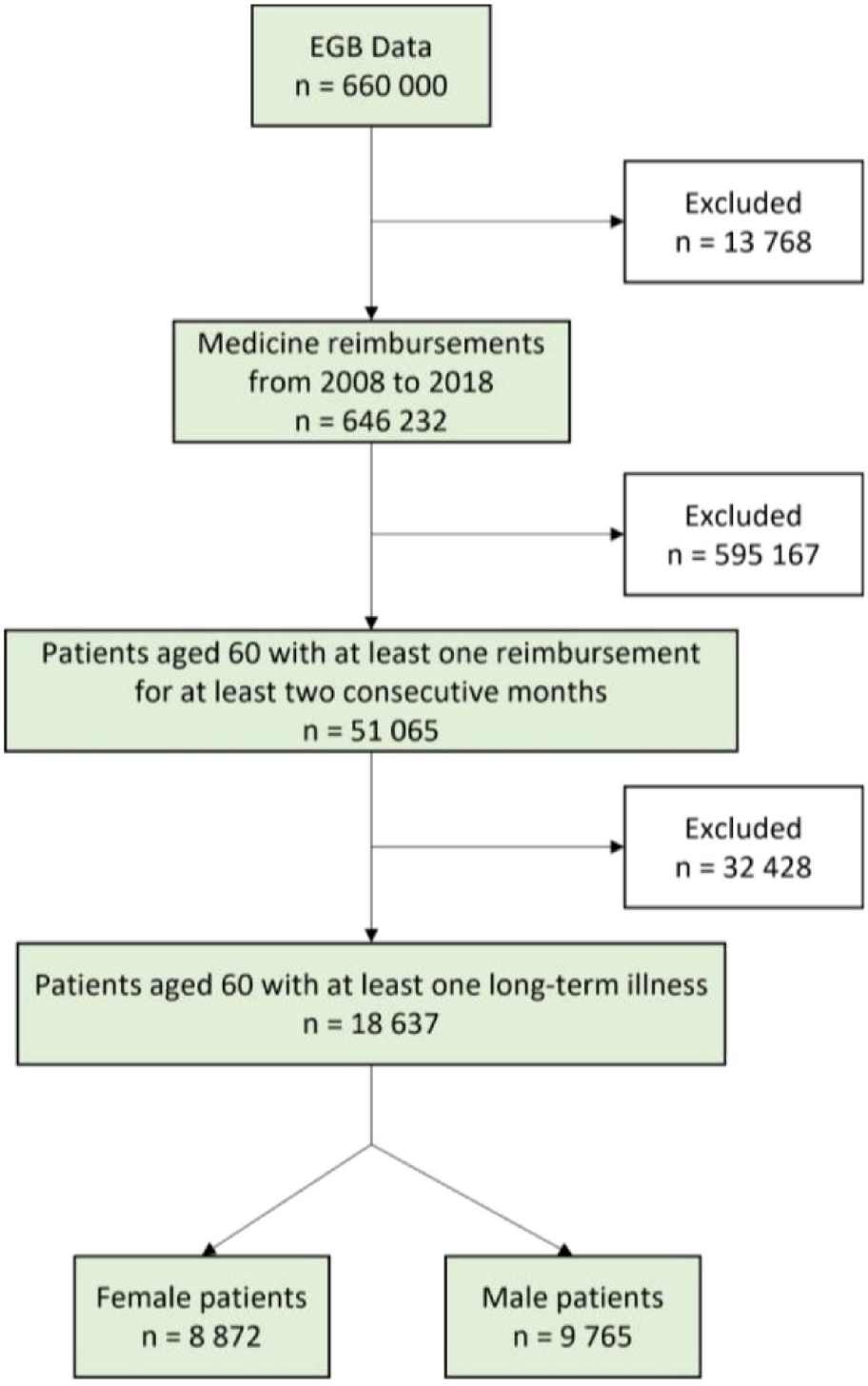

Fig. 2 Flowchart of the medicine data extraction process from the Échantillon Généraliste des Bénéficiaires (EGB). Exclusions were made to include only (i) patients aged 60 (hence creating a manageable subset of data given the large size of the EGB database), (ii) patients who had received reimbursement for at least one medicine for two or more consecutive months (hence ensuring sustained reimbursements) and (iii) having at least one long-term illness (hence focusing on chronic conditions)

Table 1 Example of medicine reimbursements contained in the Échantillon Généraliste Des Bénéficiaires (EGB). ATC: Anatomical Therapeutic Chemical

Patient ID	Reimbursement date	ATC class	Medicine name
P_1	01/04/2008	M01AE01	Ibuprofen
P_1	01/12/2015	B01AC06	Aspirin
P_2	01/02/2010	N02AX02	Tramadol
P_3	01/05/2016	B01AC04	Clopidogrel

Table 2 Example of total number of reimbursements that three patients aged 60 years received for three different medicines

Patient ID	Tramadol	Aspirin	Ibuprofen
P_1	0	10	5
P_2	1	8	4
P_3	2	6	3

In the network of female patients built with the Cosine similarity, we identify 12 clusters composed of at least 50 patients. We carry out an enrichment analysis to identify, in each cluster, potential enrichments in chronic diseases (Fig. 5A). The enrichment analysis reveals several significant enrichments. For instance, we observe significant enrichments of patients with thyroid and breast cancers in cluster 1, cerebrovascular diseases in cluster 5 and depressive episodes in cluster 2. Similarly, in the dataset of male patients, we identify 11 clusters (Fig. 6A). The enrichment analysis reveals significant enrichments of patients with cerebrovascular diseases in cluster 5, atherosclerosis in cluster 6, prostate cancer in cluster 8 and thyroid cancer in cluster 9. Of note, we identify several clusters with the same enriched chronic diseases. For instance, female clusters 2, 3, 9 and 11, and male clusters 2, 4, 7, 10 and 11 are all enriched in type 2 diabetes patients. Female clusters 6 and 7 are enriched in breast cancer patients, female clusters 10 and 12 in autoimmune disorder patients and male clusters 1 and 3 in coronary diseases patients. Overall, the use of Cosine similarity allows to identify clusters of similar patients. However, several clusters are redundant regarding their chronic disease enrichments. Similar results are obtained with the Cosine similarity weighted by IDF (Figs. 5B and 6B).

Patient networks constructed with the Cosine similarity weighted by the Wu and Palmer measure and the Lin measure result in a higher number of clusters significantly

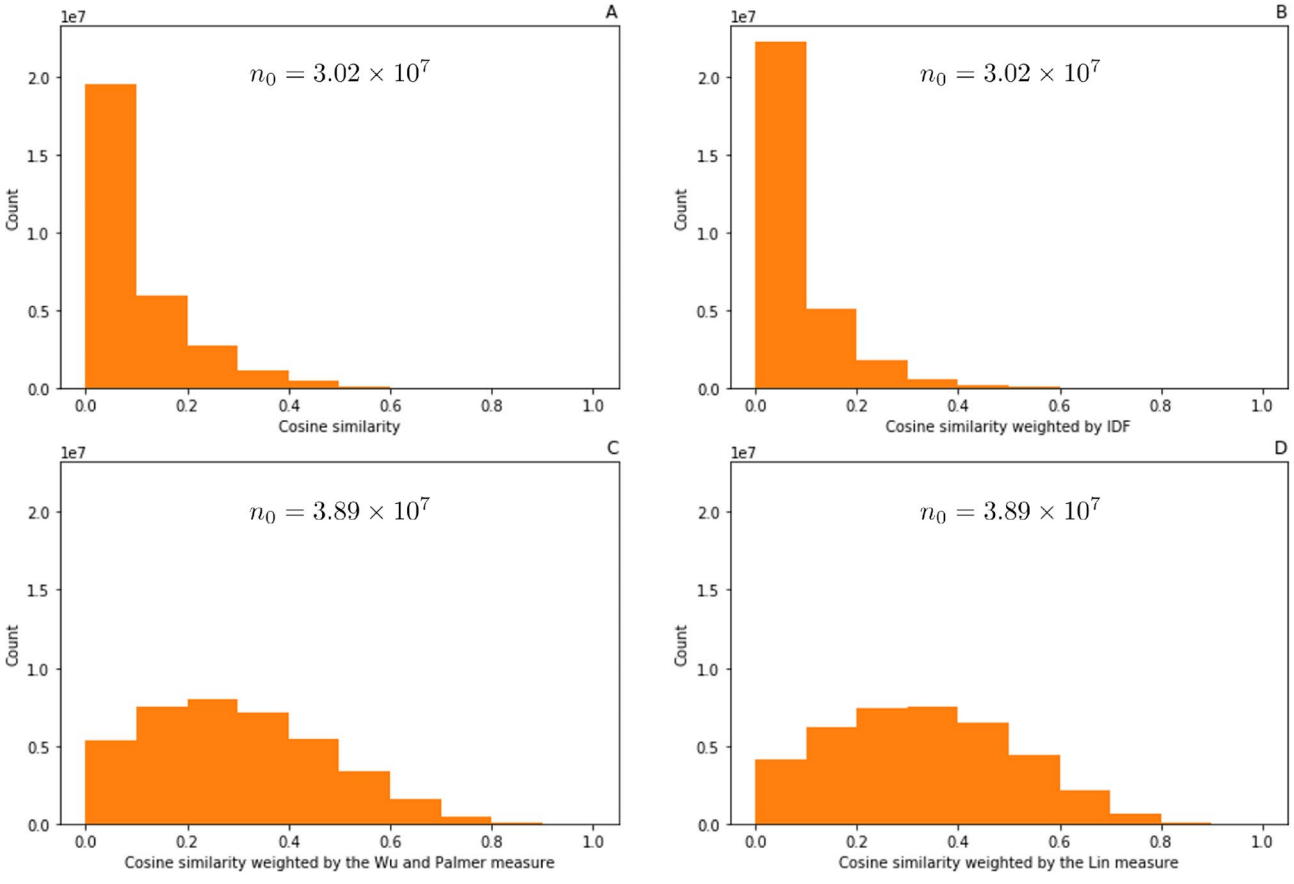

Fig. 3 Similarity distributions in the female patient dataset. **A**: Distribution of the Cosine similarity, **B**: Distribution of the Cosine similarity weighted by the Inverse Document Frequency (IDF), **C**: Distribution of the Cosine similarity weighted by the Wu and Palmer measure, **D**: Distribution of the Cosine similarity weighted by the Lin measure. n_0: Total number of patient pairwise similarities with non-zero values

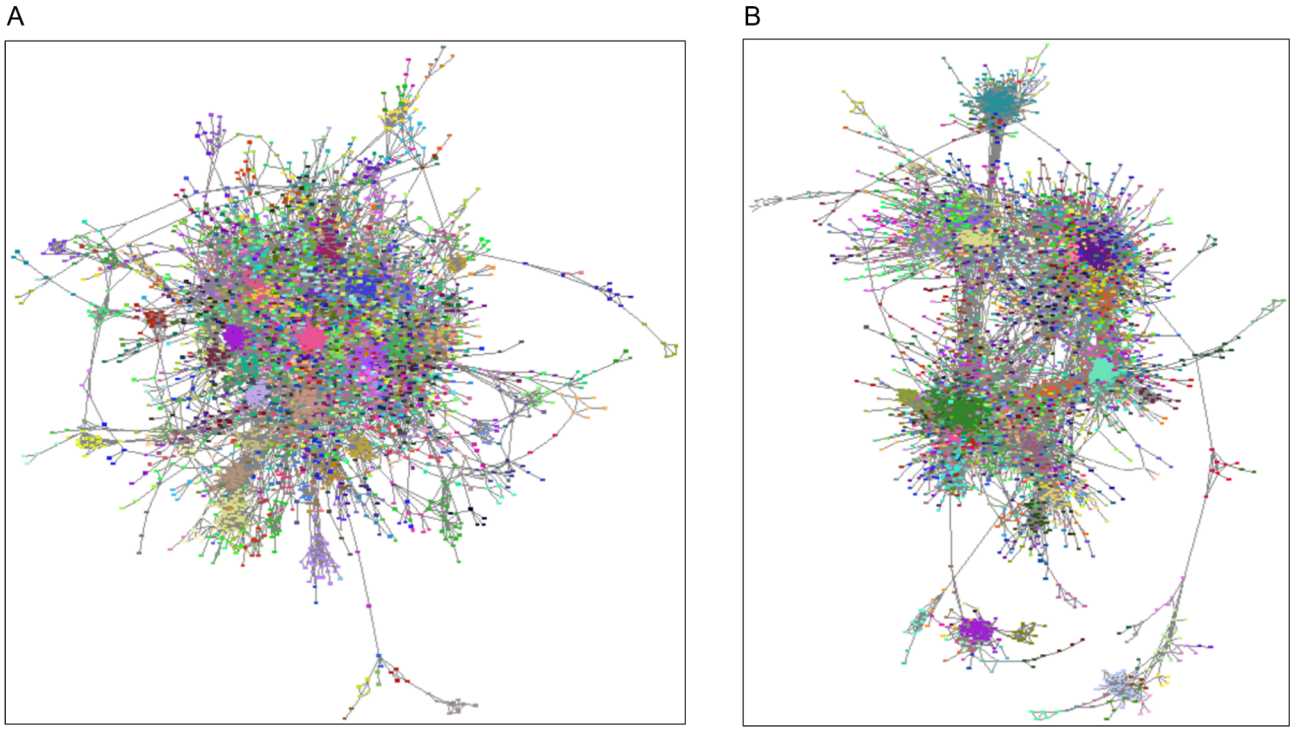

Fig. 4 Patient networks built from Cosine similarity and Cosine similarity weighted by the Wu and Palmer measure. Networks are built from Cosine similarity (**A**) and Cosine similarity weighted by the Wu and Palmer measure (**B**), on the female patient dataset. Nodes represent patients aged 60 and edges represent the interactions between those patients. The length of edges is inversely proportional to the Cosine similarity or the Cosine similarity weighted by the Wu and Palmer measure. Node colors represent the clusters identified with the Markov Clustering algorithm. For the sake of visualization, we only represent the largest connected component of each network

enriched in chronic diseases and less redundant clusters (Figs. 5C and D and 6C and D). Using the Cosine similarity weighted by the Wu and Palmer measure (Fig. 5C for the female dataset and Fig. 6C for the male dataset), the enrichment analysis reveals clusters significantly enriched in respiratory disease patients (female cluster 4 and male cluster 7), psychiatric disorders patients with psychotic side (female cluster 1 and male cluster 6), coronary disease patients (female cluster 5 and male cluster 1) and patients with type 2 diabetes associated with its comorbidities (female and male clusters 2). In the dataset of female patients, we find clusters significantly enriched in thyroid and breast cancer patients (clusters 3 and 6). In the dataset of male patients, we find clusters significantly enriched in atherosclerosis patients (cluster 3), depressive disorders patients (cluster 4) and heart failure patients (cluster 9). Similar results are obtained using the Cosine similarity weighted by the Lin measure (Figs. 5D and 6D). Notably, all clusters are significantly enriched in diabetes patients, in both datasets. This is explained by the fact that diabetes is the most frequent chronic disease in our population (see overall columns in Figs. 5 and 6).

Discussion

In this paper, we adapted two novel weighted similarity measures for health quantitative variables. These similarity measures were weighted by considering the variable labels relationships. They performed better in identifying clusters of patients suffering from different diseases as compared to unweighted similarity measures that did not consider label relationships. Overall, our analysis highlighted the interest of considering variable label relationships when calculating patient similarities to improve patient stratification.

In recent years, there has been a growing interest in computing patient similarities using Electronic Health Records (EHR). However, most papers focused on computing similarities using variables extracted from medical texts through natural language processing methods. These papers also focused on the development of methods to automatically learn variable label relationships from data [27]. However, using variable label relationships from existing medical classifications has been poorly addressed, despite the ready availability of this expert information. In this study, we underline the value of integrating such an expert knowledge into patient similarity measures to increase clustering performance. This is particularly relevant to analyze records obtained from administrative claim databases. Indeed, these databases

Fig. 5 Chronic disease enrichments in patient clusters obtained from the female patient dataset. Clusters are identified in networks built from Cosine similarity (**A**), Cosine similarity weighted by the Inverse Document Frequency (**B**), Cosine similarity weighted by the Wu and Palmer measure (**C**), and Cosine similarity weighted by the Lin measure (**D**), on the female patient dataset. The numbered columns represent the clusters composed of at least 50 patients, ranked from the largest to the smallest. The last column, named overall, represents all the patients found in the network's largest connected component. n: number of patients identified in each cluster or in the network largest connected component. The rows correspond to the chronic diseases. Box colors represent the proportion of patients with a given chronic disease. Stars represent significant enrichments (*p*-value lower than 0.05 after Benjamini-Hochberg correction). For the sake of visualization, we only represent chronic diseases that are significant in at least one cluster

gather medical variables with labels that are always organized into classifications such as SNOMED-CT or ICD-10.

While our study only considered variable label relationships organized according to a classification, other types of variable label organization exist. For instance, the Human Phenotype Ontology (HPO) is a directed acyclic graph (DAG). Previous studies have already proposed variable label relationship measures for this type of label organization. For example, Köhler et al. developed a variable label relationship measure that exploits the structure of HPO to improve clinical diagnostics [28]. Xue, Peng, and Shang derived another measure that exploits both the DAG structure and the phenotype term definition of HPO in order to improve the prediction of disease-related phenotypes [29]. However, these measures were originally designed for binary variables and would need to be adapted for quantitative variables commonly found in medical health records as well as in many biological and omics datasets. A recent work demonstrated the interest of considering prior knowledge representation in the context of omics data [13]. Of note, embedding approaches can also be used to encode structured background knowledge, e.g. opa2vec [30] or owl2vec [31]. However, we did not consider these approaches here as ATC classification is more simple than an ontology or a knowledge graph, so the differences between weights computed from distances in the direct or embedded spaces might be too subtle to be visible with our evaluation procedures.

In this study, we used the Cosine similarity because we have previously shown that this measure was more effective than others to deal with our specific use-cases [22]. However, depending on the data, other similarity measures could be used, and weighted, to compute similarities between patients. We also explored the interest of incorporating variable frequencies in the computation of patient similarities. Indeed, we weighted the Cosine similarity by the Inverse Document Frequency (IDF) to take into account the frequency of the usage of the medicines. Our hypothesis was that it would better capture similarity information as two patients taking the same uncommon medicine would be considered more similar than two patients taking the same common medicine. However, we observed that exploiting the medicine frequency did not enhance the performance of the similarity measures. This may be attributed to the fact that a single medicine can be used to treat multiple

Fig. 6 Chronic diseases enrichments in patient clusters obtained from the male patient dataset. Clusters are identified in networks built from Cosine similarity (**A**), Cosine similarity weighted by the Inverse Document Frequency (**B**), Cosine similarity weighted by the Wu and Palmer measure (**C**), and Cosine similarity weighted by the Lin measure (**D**), on the male patient dataset. The numbered columns represent the clusters composed of at least 50 patients, ranked from the largest to the smallest. The last column, named overall, represents all the patients found in the network's largest connected component. n: number of patients identified in each cluster or in the network largest connected component. The rows correspond to the chronic diseases. Box colors represent the proportion of patients with a given chronic disease. Stars represent significant enrichments (*p*-value lower than 0.05 after Benjamini-Hochberg correction). For the sake of visualization, we only represent chronic diseases that are significant in at least one cluster

conditions, making it challenging to associate a medicine with a specific pathology. However, considering the frequency of medicines may be more effective in the context of rare diseases [32]. In such cases, these diseases are typically treated with orphan medicines that have specific indications.

Of note, we only used three different weights in patient similarity computation. Therefore, our conclusion that considering medicine label relationships is more relevant than considering medicine frequency may not universally extend to all measures. However, the central message of our paper, demonstrating the effectiveness of these adapted label relationship measures in improving patient stratification, remains robust.

In this paper, we assessed the performance of the different similarity measures to cluster patients using external binary variables (i.e., chronic diseases). We employed these external variables in cluster enrichment analyses. This novel approach deviates from the typical reliance on internal criteria such as silhouette score, which do not offer clinically interpretable insights. Using these enrichment analyses, we were able to interpret the clusters clinically and to compare the different similarity measures from an expert point of view. Although previous works have already used enrichment analyses to study enrichments of HPO terms in the literature [33], to the best of our knowledge, it was never used on patient medical data. Finally, this study might be useful for data-based Clinical Decision Support Systems (CDSS) to support diagnosis inference [34].

A limitation of this study is the lack of data on other potentially contributing factors, such as social deprivation or additional demographic variables, which may influence clustering outcomes. These factors could provide deeper insights into patient stratification but were not available in the dataset.

As a conclusion, we recommend considering variable label relationships when computing patient similarities to improve stratification of patients regarding their health status.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-025-02459-8.

Supplementary Material 1

Acknowledgements
The authors would like to acknowledge Pierre Sabatier for extracting and formatting the data. The authors would also like to thank Morgane Terezol and Ozan Ozisik for their comments after proofreading the article. And finally, we would like to thank all the members of MMG and Heka teams for their feedback.

Author contributions
JL contributed to the writing-original draft preparation, the visualization, the conceptualization and the methodology of the study. AL, AB and AJ contributed to the supervision, the conceptualization, the methodology, the writing-reviewing and the editing of the study.

Funding
This work was supported by the Inserm cross-cutting program Genomic variability 2018 GOLD.
The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability
The datasets analyzed during the current study are not publicly available due to the access policy of the Échantillon Généraliste des Bénéficiaires database.

Declarations

Ethics approval and consent to participate
All methods were carried out in accordance with the RECORD (REporting of studies Conducted using Observational Routinely-collected Data) statement. This study has been declared to INSERM (Institut National de la Santé et de la Recherche Médicale, https://www.inserm.fr/). The data from this study are extracted from the EGB (Échantillon Généraliste des Bénéficiaires), a permanent 1/97 representative sample of the National Health Data System (Système National de Données de Santé, SNDS). The information provided to individuals in EGB on the possible re-use of their data and the procedures for exercising their rights comply with the legislative and regulatory provisions applicable to the processing of personal data in the SNDS. According to French regulations, informed consent is not required for secondary data reuse, but patient information is mandatory. Therefore, individuals in SNDS database are informed of the reuse of their data for research and can opposed to this reuse as defined by Articles 92 to 95 of Decree No. 2005 – 1309 of 20 October 2005 (https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000037300884/). EGB data, which is part of the SNDS, can be reused for research projects from authorized persons once the research project is declared to their institution (INSERM). Institutional review board approval was not required for this study because the research was based on the secondary use of the EGB (Échantillon Généraliste des Bénéficiaires) medico-administrative database, which is a random sample of 1/97 of the French National medico-administrative database.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Sorbonne Université, Université Paris Cité, INSERM, Centre de Recherche des Cordeliers, Paris F-75006, France
[2]HeKA, Inria Paris, Paris F-75015, France
[3]Aix Marseille Univ, INSERM, MMG, Marseille UMR1251, France
[4]Université Paris Cité, INSERM, NeuroDiderot, Paris UMR1141, 75019, France
[5]CNRS, Marseille, France
[6]Barcelona Supercomputing Center, Barcelona, Spain
[7]Université Paris Cité, Sorbonne Université, INSERM, Centre de Recherche des Cordeliers, F-75006 Paris, France
[8]French National Rare Disease Registry (BNDMR), Greater Paris University Hospitals (AP-HP), Paris, France

Received: 16 June 2023 / Accepted: 3 January 2025
Published online: 15 March 2025

References

1. Garcelon N, Neuraz A, Salomon R, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet J Rare Dis. 2018;13(1):1–11.
2. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med. 2015;7(311):311ra174.
3. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018May;6(5):361–9.
4. Hong Y, Zeng ML. International classification of diseases (ICD). KO Knowl Organ. 2023;49(7):496–528.
5. Donnelly K, et al. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform. 2006;121:279.
6. Irani J, Pise N, Phatak M. Clustering techniques and the similarity measures used in clustering: a survey. Int J Comput Appl. 2016;134(7):9–14.
7. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. J Doc. 1972;28(1):11–21.
8. Conroy B, Xu-Wilson M, Rahman A. Patient similarity using population statistics and multiple kernel learning. In: Machine Learning for Healthcare Conference. PMLR; 2017.pp. 191–203.
9. Wu Z, Palmer M. Verb semantics and lexical selection. arXiv preprint cmp-lg/9406033. 1994.
10. Lin D et al. An information-theoretic definition of similarity. In: Icml. 1998.pp. 296–304.
11. Ni J, Liu J, Zhang C et al. Fine-grained patient similarity measuring using deep metric learning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017.pp. 1189–98.
12. Girardi D, Wartner S, Halmerbauer G, et al. Using concept hierarchies to improve calculation of patient similarity. J Biomed Inform. 2016;63:66–73.
13. Kańduła MM, Aldoshin AD, Singh S et al. ViLoN-a multi-layer network approach to data integration demonstrated for patient stratification. Nucleic Acids Res. 2023Jan11;51(1):e6.
14. Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001;98(Sep11):10869–74.
15. Alonso I et, Contreras D. Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: an UMLS approach. Expert Syst Appl. 2016;44.p.386–99.
16. Mcinnes BT, Pedersen T, et Pakhomov SV. UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. AMIA annual symposium proceedings. American Medical Informatics Association.2009;p.431.
17. Pedersen T, Pakhomov S, Patwardhan S et al. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform. 2007;vol.40.3.p.288–99.
18. Singhal A, et al. Modern information retrieval: a brief overview. IEEE Data Eng Bull. 2001;24(4):35–43.
19. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007. 1995.
20. Xu R, Wunschll D. Survey of Clustering algorithms. IEEE Trans Neural Netw. 2005;16(3):645–78.
21. Schaeffer SE. Graph clustering. Comput Sci Rev. 2007Aug;1(1):27–64.
22. Lambert J, Leutenegger AL, Jannot AS, Baudot A. Tracking clusters of patients over time enables extracting information from medico-administrative databases. J Biomed Inform. 2023;139:104309.
23. vanDongen S. A cluster algorithm for graphs. Inform Syst [INS]. 2000;(R 0010).
24. Fisher RA. On the interpretation of χ 2 from contingency tables, and the calculation of P. J Roy Stat Soc. 1922;85(1):87–94.
25. Tuppin P, De Roquefeuil L, Weill A, et al. French national health insurance information system and the permanent beneficiaries sample. Revue D'epidemiologie et de sante Publique. 2010;58(4):286–90.
26. Skrbo A, Begović B, Skrbo S. Classification of drugs using the ATC system (anatomic, therapeutic, Chemical classification) and the latest changes. Med Arh. 2004;58(1Suppl 2):138–41.

27. Choi E, Bahadori MT, Searles E et al. Multi-layer representation learning for medical concepts. In: proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.pp. 1495–504.
28. Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85(4):457–64.
29. Xue H, Peng J, Shang X. Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO. BMC Syst Biol. 2019;13(2):1–12.
30. Smaili FZ, Gao X, et Hoehndorf R. OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. Bioinformatics. 2019;vol.35.no 12.pp. 2133–2140.
31. Chen J, Hu P, Jimenez-Ruiz E et al. Owl2vec*: Embedding of owl ontologies. Machine Learning. 2021;vol.110.no 7.pp. 1813–1845.
32. Chen X, Garcelon N, Neuraz A, et al. Phenotypic similarity for rare disease: ciliopathy diagnoses and subtyping. J Biomed Inform. 2019;100:103308.
33. Deng Y, Gao L, Wang B, et al. HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. PLoS ONE. 2015;10(2):e0115692.
34. Gräßer F, Tesch F, Schmitt J et al. A pharmaceutical therapy recommender system enabling shared decision-making. User Model User-Adapt Interact. 2022;pp. 1–44.

Publisher's note