A visual analytic approach for the identification of ICU patient subpopulations using ICD diagnostic codes

Daniel Alcaide [1], **Jan Aerts** [Corresp. 1, 2]

[1] Department of Electrical Engineering (ESAT) STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium

[2] UHasselt, I-BioStat, Data Science Institute, Hasselt, Belgium

Corresponding Author: Jan Aerts
Email address: jan.aerts@uhasselt.be

A large number of clinical concepts are categorized under standardized formats that ease the manipulation, understanding, analysis, and exchange of information. One of the most extended codifications is the International Classification of Diseases (ICD) used for characterizing diagnoses and clinical procedures. With formatted ICD concepts, a patient profile can be described through a set of standardized and sorted attributes according to the relevance or chronology of events. This structured data is fundamental to quantify the similarity between patients and detect relevant clinical characteristics. Data visualization tools allow the representation and comprehension of data patterns, usually of a high dimensional nature, where only a partial picture can be projected.

In this paper, we provide a visual analytics approach for the identification of homogeneous patient cohorts by combining custom distance metrics with a flexible dimensionality reduction technique. First we define a new metric to measure the similarity between diagnosis profiles through the concordance and relevance of events. Second we describe a variation of the STAD (Simplified Topological Abstraction of Data) dimensionality reduction technique to enhance the projection of signals preserving the global structure of data.

The MIMIC-III clinical database is used for implementing the analysis into an interactive dashboard, providing a highly expressive environment for the exploration and comparison of patients groups with at least one identical diagnostic ICD code. The combination of the distance metric and STAD not only allows the identification of patterns but also provides a new layer of information to establish additional relationships between patient cohorts. The method and tool presented here add a valuable new approach for exploring heterogeneous patient populations. In addition, the distance metric described can be applied in other domains that employ ordered lists of categorical data.

A visual analytic approach for the identification of ICU patient subpopulations using ICD diagnostic codes

Daniel Alcaide[1] **and Jan Aerts**[1,2]

[1]**Department of Electrical Engineering (ESAT) STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Belgium**
[2]**I-BioStat, Data Science Institute, UHasselt, Belgium**

Corresponding author:
Jan Aerts

Email address: jan.aerts@uhasselt.be

ABSTRACT

A large number of clinical concepts are categorized under standardized formats that ease the manipulation, understanding, analysis, and exchange of information. One of the most extended codifications is the International Classification of Diseases (ICD) used for characterizing diagnoses and clinical procedures. With formatted ICD concepts, a patient profile can be described through a set of standardized and sorted attributes according to the relevance or chronology of events. This structured data is fundamental to quantify the similarity between patients and detect relevant clinical characteristics. Data visualization tools allow the representation and comprehension of data patterns, usually of a high dimensional nature, where only a partial picture can be projected.

In this paper, we provide a visual analytics approach for the identification of homogeneous patient cohorts by combining custom distance metrics with a flexible dimensionality reduction technique. First we define a custom metric to measure the similarity between diagnosis profiles through the concordance and relevance of events. Second we describe a variation of the STAD (Simplified Topological Abstraction of Data) dimensionality reduction technique to enhance the projection of signals preserving the global structure of data.

An MIMIC-III clinical database is used to demonstrate the approach, presented as an interactive dashboard, providing a highly expressive environment for the exploration and comparison of patients groups with at least one identical diagnostic ICD code. The combination of the distance metric and STAD not only allows the identification of patterns but also provides a new layer of information to establish additional relationships between patient cohorts. The method and tool presented here add a potentially valuable new approach for exploring heterogeneous patient populations. In addition, the distance metric described can be applied in other domains that employ ordered lists of categorical data.

INTRODUCTION

Patient profiling and selection are a crucial step in the setup of clinical trials. The process involves analytical methods to handle the increasing amount of healthcare data but is still extremely labor-intensive (Sahoo et al., 2014). Nevertheless, the input from an expert in this selection is important.

To support the expert in the selection of suitable patients, visual analytics solutions can enable the exploration of a patient population, make recruitment consistent across studies, enhance selection accuracy, increase the number of selected participants, and significantly reduce the overall cost of the selection process (Fink et al., 2003; Damen et al., 2013). Visual analytics relies on interactive and integrated visualizations for exploratory data analysis in order to identify unexpected trends, outliers, or patterns. It can indicate relevant hypotheses that can be complemented with additional algorithms, and help define parameter spaces for these algorithms (Franken, 2009). A major challenge in creating visual solutions is to find effective tools which allow the projection of all data dimensions. One popular solution is to visualize the relationship between elements rather than raw data through similarity metrics which quantify the closeness between data objects (Liu et al., 2016). Similarity metrics are a fundamental part for most

of the case-based reasoning algorithms (Kolodner, 2014) such as the detection of consistent cohorts of patients within a patient population. One of the remaining open challenges in the analysis of patient similarity is to establish relevant and practical ways based on clinical concepts (Jia et al., 2019).

Many types of information about the patient profile such as diagnosis, procedures, and prescriptions are available under standardized categories contained in taxonomies or dictionaries, e.g., the International Classification of Diseases (ICD), Medical Dictionary for Regulatory Activities (MedDRA) and the Anatomical Therapeutic Chemical (ATC) Classification System. Each patient is for example linked to an ordered list of diagnoses, which are semantic concepts that are (in the case of MIMIC (Johnson et al., 2016)) ordered from most to least important (as per the MIMIC-III documentation "ICD diagnoses are ordered by priority - and the order does have an impact on the reimbursement for treatment"). These standardized formats provide a non-numerical data structure facilitating both understanding and management of the data. Several methods have been proposed to define similarity between lists of clinical concepts based on presence of absence of specific terms (Gottlieb et al. 2013; Zhang et al. 2014; Brown 2016; Girardi et al. 2016; Rivault et al. 2017; Jia et al. 2019). However, the diagnostic profile of a patient is not merely an independent list of semantic concepts but also includes an intrinsic order indicated by the position of the terms in the list reflecting the relevance vis-a-vis the actual patient status. To the best of our knowledge, no previous work has combined the categorical and ordinal nature of clinical events into a single distance function. This dualism can contribute to improving the detection of cohorts through diagnostic and procedural data. This can have a significant impact as diagnoses or procedures are part of recruitment criteria in most clinical trials (Boland et al., 2012).

In this paper, a novel approach for exploring clinical patient data is introduced. In particular, we focus on patient profiles represented by a set of diagnosis ICD codes sorted by relevance. The distance metric considers the sorted concepts as input, and the resulting pairwise values are projected into a dimensionality reduction graph.

The remaining part of this paper is organized as follows. In the section 'Background', we give an overview of related work in categorical events and graphical projections of patient similarity. The section 'Materials and Methods' describes the proposed distance metric and modifications applied on the base algorithms STAD for visualizing patient population. In 'Results', we demonstrate the effectiveness of the approach in a real-world dataset. The section 'Discussion' compares other methods and alternative metrics for similar data. Finally, the section 'Conclusion' presents conclusions and possible directions for future work.

BACKGROUND

The exploration and analysis of patients through similarity measures has been presented in different areas of bioinformatics and biomedicine, and also data mining and information visualization. In this section, we review the related literature on these areas below, and we focus on the notion of similarity measures for categorical events and graphical representation of patient similarity.

Patient similarity and distance measures for categorical events

Different distance metrics exist for unordered lists of categorical data, including the overlap coefficient (Vijaymeena and Kavitha, 2016), the Jaccard index (Real and Vargas, 1996), and the simple matching coefficient (Šulc and Řezanková, 2014). These methods compute the number of matched attributes between two lists using different criteria. Although they treat each entry in the list as independent of the others, they have been used successfully to measure patient similarity to support clinical decision making and have demonstrated their effectiveness in exploratory and predictive analytics (Zhang et al. 2014; Lee et al. 2015). Similarly, different ways of computing distances between ordered lists are available (Van Dongen and Enright, 2012). The Spearman's rank coefficient (Corder and Foreman, 2014) is useful for both numerical and categorical data and has been used in clinical studies (Mukaka, 2012). However, correlation between ordered lists cannot be calculated when the lists are of different lengths (Pereira et al., 2009).

In the context of medical diagnoses, the ICD (International Classification of Diseases) codes have been widely used for describing patient similarity. However, these typically consider the hierarchical structure of the ICD codes. Gottlieb et al. (2013), for example, proposed a method combining the Jaccard score of two lists with the nearest common ancestor in the ICD hierarchy. The similarity measure for the ICD ontology was previously presented in Popescu and Khalilia (2011). Each term is assigned to a weight

based on its importance within the hierarchy, which was defined as $1 - 1/n$ where n corresponded to its level in the hierarchy.

In our work, however, we will not leverage the hierarchical structure of the ICD codes, but employ the ICD grouping as described by Healthcare Cost and Utilization Project (2019). Our approach takes the position of the term in the list of diagnoses into account, which is a proxy to their relevance for the patient status. The metric assigns a higher weight to terms located ~~in first positions than the last ones.~~

Alternative approaches such as those by Le and Ho (2005) and Ahmad and Dey (2007) consider the similarity between two attributes as the shared relationship with the other elements in the sample, i.e., two elements are similar if they appear with a common set of attributes. From a different perspective, the latent concept of these metrics is also present in the identification of comorbidity diseases (Moni et al. 2014; Ronzano et al. 2019) although these studies aim to find heterogeneous types of diseases rather than different profiles of patients. The main drawback of metrics based on co-occurrence is the assumption of an intrinsic dependency between attributes without considering their relevance. The work presented by Ienco et al. (2012) and Jia et al. (2015) use the notion of context which identifies the set of relevant categories to a defined attribute. The similarity measure in Jia et al. (2015) is determined by the correlation of their context attributes.

Graphical projections of patient similarity

Visually representing pairwise distance matrices remains a challenge. Most often, dimensionality reduction techniques are used to bring the number of dimensions down to two so that the data can be represented in a scatterplot (Nguyen et al. 2014; Girardi et al. 2016; Urpa and Anders 2019). Such scatterplots can not only indicate clusters and outliers, but are also very useful for assessing sample quality. In the case of patient data, each point in such plot represents a patient, and relative positions between them in the 2D plane correspond to the distance between them in the original higher dimensional space. Multidimensional scaling (MDS) is arguably one of the most commonly used dimensionality reduction methods (Mukherjee et al., 2018). It arranges points on two or three dimensions by minimizing the discrepancy between the original distance space and the distance in the two-dimensional space. Derived MDS methods have been presented, proposing modified versions of the minimization function but conserving the initial aim (Saeed et al., 2018). Besides MDS, recent methods have been proposed to highlight the local structure of the different patterns in high-dimensional data. For example, t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008) and uniform manifold approximation (UMAP) (McInnes et al., 2018) have been used in many publications on heterogeneous patient data (Abdelmoula et al. 2016; Simoni et al. 2018; Becht et al. 2019). Unlike MDS, t-SNE projects the conditional probability instead of the distances between points by centering a normalized Gaussian distribution for each point based on a predefined number of nearest neighbors. This approach generates robustness in the projection, which allows the preservation of local structure in the data. In a similar fashion, UMAP aims to detect the local clusters but at the same time generates a better intuition of the global structure of data.

In addition to scatterplot representations, alternative visual solutions are also possible, for example heatmaps (Baker and Porollo, 2018), treemaps (Zillner et al., 2008), and networks. The latter are often built using a combination of dimensionality reduction and topological methods (Li et al. 2015; Nielson et al. 2015; Dagliati et al. 2019). This approach has for example been used with success to visually validate the automated patient classification in analytical pipelines (Pai and Bader 2018; Pai et al. 2019). In general, the created network encodes the distance between two datapoints in high-dimensional space into an edge between them and the full dataset can therefore be represented as a fully connected graph. The STAD method (Alcaide and Aerts, 2020) reduces the number of edges allowing a more scalable visualization of distances. The original distance in high-dimensional space between two datapoints is correspondent to the path-length in the resulting graph between these datapoints. The main advantage of networks to display high-dimensional data is that users not only can perceive patterns by the location of points but also by the connection of elements, thereby increasing trust in the data signals.

MATERIAL AND METHODS

The International Classification of Diseases (ICD) is a diagnosis and procedure coding system used by hospitals to bill for care provided. They are further used by health researchers in the study of electronic medical records (EMR) due to the ease of eliciting clinical information regarding patient status. Although these administrative databases were not designed for research purposes, their efficiency compared to the

manual review of records and demonstrated reliability of information extracted have democratized the analysis of health data in this way (Humphries et al., 2000). Even though ICD codification is hierarchically organized, some concepts in the database may be under-reported (Campbell et al., 2011). To make analysis feasible, the ICD codes are in practice often grouped in higher categories to reduce noise and facilitate the comparison and analysis with automatic systems (Choi et al. 2016; Miotto et al. 2016; Baumel et al. 2018). In our approach, we adopt the ICD generalization introduced by the Clinical Classification Software (CSS) which groups diseases and procedures into clinically meaningful sections (Healthcare Cost and Utilization Project, 2019). Here we introduce a method to compare unequal sets of ordered lists of categories and explore the different cohorts of patients through visual representations of data. This approach employs a custom distance metric presented in section 'Diagnosis similarity and distances' within the visual analytics method as presented in section 'Spanning Trees as Abstraction of Data'.

Diagnosis similarity and distances

In the MIMIC dataset which was used for this work (Johnson et al., 2016), each patient's diagnosis is a list of ICD codes, as exemplified in Table 1. The average number of concepts per profile in the MIMIC III dataset is 13 with a standard deviation of 5. Diagnoses are sorted by relevance for the patient status. This order determines the reimbursement for treatment, and, from an analysis perspective, can help us to distinguish similar medical profiles even with different initial causes. The similarity metric presented in this work takes this duality into account and provides support for comparing profiles with an unequal length of elements.

Patient A (115057)			**Patient B (117154)**		
	ICD section	Label (ICD9)		ICD section	Label (ICD9)
1	996-999.	Infection and inflammatory reaction due to other vascular device, implant, and graft (99662)	1	430-438.	Unspecified intracranial hemorrhage (4329)
2	**990-995.**	**Sepsis (99591)**	2	430-438.	Cerebral artery occlusion, unspecified with cerebral infarction (43491)
3	**590-599.**	**Urinary tract infection, site not specified (5990)**	3	996-999.	Iatrogenic cerebrovascular infarction or hemorrhage (99702)
4	**401-405.**	**Unspecified essential hypertension (4019)**	4	**990-995.**	**Sepsis (99591)**
			5	**590-599.**	**Urinary tract infection, site not specified (5990)**
			6	**401-405.**	**Unspecified essential hypertension (4019)**

Table 1. Objective function in STAD and STAD-R. The correlation ρ is computed between the original distance matrix D_X and the distance matrix derived from the shortest path graph in D_U. The ratio R is calculated from the network at each iteration considering the edges included in the network. Note that distance $d_{network\,edge}$ are normalized values between zero and one.

The similarity between two patients (diagnosis profiles) A and B is based on which diagnoses (i.e. ICD9 codes) are present in both, as well as the position of these elements in the list. Consider a match M between two concepts c_A and c_B, which contributes to the similarity according to the following formula:

$$M_C(A,B) = ln\left(1 + \frac{1}{max(\,position(c_A),\,position(c_B)\,)}\right)$$

The position mentioned in the formula corresponds to the positional index in the list. As an example, the individual contribution of the concept "Sepsis" for patients A and B in Table 1 is $M_{Sepsis} = ln\left(1+\frac{1}{max(2,4))}\right) = ln\,1.25$. The total similarity between patients is the sum of individual contributions

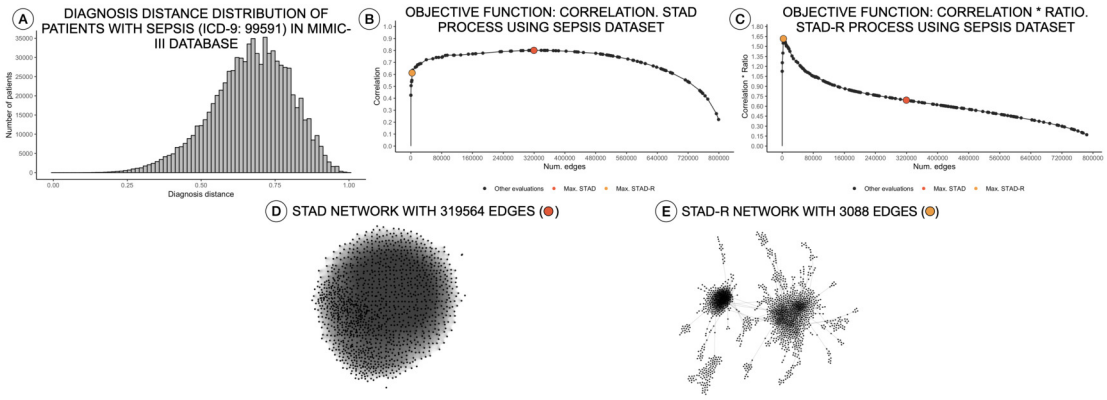

Figure 1. Distance distributions of a population of patients with sepsis, STAD, and STAD-R projections. The dataset is composed of a selection of 1,271 patients from MIMIC-III diagnosed with sepsis (ICD-9: 99591). Predefined conditions cause more homogeneous populations that mitigate the skewness of the diagnosis similarity distribution. (A) Distribution of diagnosis distance. (B) Correlation between original distance matrix and distance matrix based on STAD graph, given different numbers of edges. (C) Idem as (B) using STAD-R. (D) STAD network. (E) STAD-R network.

from the matched concepts $S(X,Y) = \sum_{n}^{i=1} M(X \bigcap Y)$. Applying this formula to the example in Table 1 gives: $S(PatientA, PatientB) = M_{Sepsis} + M_{Urinarytractinfection} + M_{Hypertension} = ln\,1.25 + ln\,1.20 + ln\,1.17 \simeq 0.56$

To perform the patient analysis in STAD (Section 'Simplified Topological Abstraction of Data'), the similarity measure S needs to be converted into a distance measure $D = 1 - S_{normalized}$ where $S_{normalized} = S/max(S)$.

Distance measures in categorical variables are built based on a binary statement of zero or one. Unlike other data types, categorical data generate a bimodal distribution, which can be considered as a normal when the element contains multiple dimensions (Schork and Zapala, 2012). The similarity in diagnosis metric not only depends on the matching of elements but also on their positions on the list. These two conditions tend to generate left-skewed distance distributions, as shown in (Figure 1A). In other words, most patients are very different from other patients.

Simplified Topological Abstraction of Data

Simplified Topological Abstraction of Data (STAD) (Alcaide and Aerts, 2020) is a dimensionality reduction method which projects the structure of a distance matrix D_X into a graph U. This method converts datapoints in multi-dimensional space into an unweighted graph in which nearby points in input space are mapped to neighboring vertices in graph space. This is achieved by maximizing the Pearson correlation between the original distance matrix and a distance matrix based on the shortest paths between any two nodes in the graph (which is the objective function to be optimized). STAD projections of multi-dimensional data allow the extraction of complex patterns. The input for a STAD transformation consists of a distance matrix of the original data, which in this case is based on the metric as defined in the previous section.

As mentioned above, high dissimilarity between datapoints (i.e. patients) results in a left-skewed distance distribution. Unfortunately, this skew poses a problem for STAD analysis. As mentioned above, the STAD method visualizes the distances between elements by means of the path length between nodes. Hence, to represent a big distance between two elements, STAD needs to use a set of intermediate connections that help to describe a long path. In case no intermediate nodes can be found, the algorithm forces a direct connection between the two nodes. As a result, in a left-skewed distribution, STAD tends to generate networks with an excessively high number of links, even when high correlation can be achieved as shown in Figure 1B and D. This means that the principle that nodes that are closely linked are also close in the original space (i.e. are similar) does not hold anymore (Koffka, 2013).

Therefore, we propose a modification of the STAD algorithm, named STAD-R (where the R stands for "Ratio"), which avoids the problem on datasets of dissimilar items through the use of a modified objective function. To reduce the number of links between dissimilar datapoints we alter the STAD method to

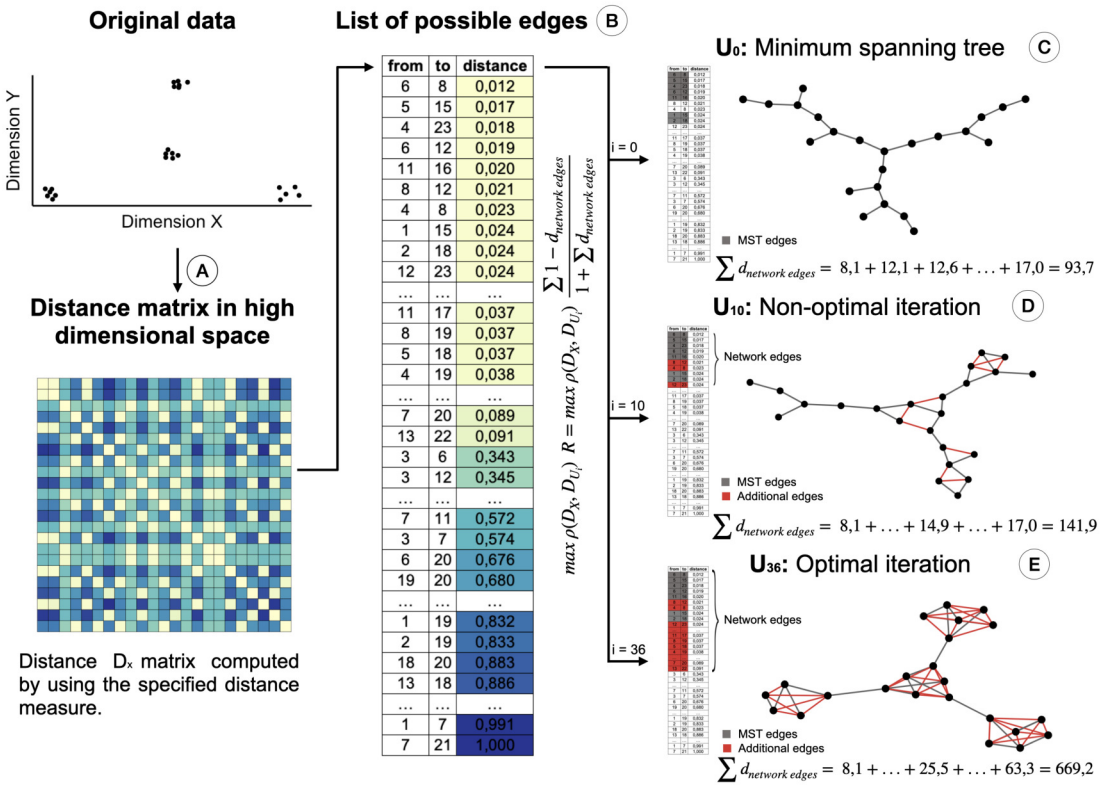

Figure 2. Creation of the STAD-R network for different iterations. (A) Distance matrix D_X: Pairwise distances between all elements in a point cloud are calculated using a defined distance metric. (B) Distance list: Transformation of the matrix into a edges list. Edges are sorted by their distance. Smaller distances are first candidates to become part of the network U. (C) The Minimum spanning tree connects all nodes with minimum distance. It guarantees that a path exists between all nodes and becomes the initial iteration in the evaluation of the optimal STAD network (D) The addition of edges over the MST may improve the correlation between the two distance matrices. Edges are added in sequential order following the list in B. (E) The optimal network is found at the iteration with the maximum combination of correlation between D_X and D_U and the ratio R.

incorporate the ratio $R = \frac{\sum 1 - d_{network\,edge}}{\sum 1 + d_{network\,edge}}$, in which the sum of $d_{network\,edge}$ refers to the sum of distances of edges included in the network (see Figure 2). Note that edges represent the distance between two elements of the dataset and constitute a cell in the pairwise distance matrix.

This ratio R is added to the objective function of the algorithm, which maximizes the correlation ρ between the distance matrices D_X (of the input dataset) and D_U (based on shortest path distances in the graph). When including the ratio R, the objective function in STAD-R is not only a maximization problem based on the Pearson correlation but also a maximization of ratio R. Table 2 shows the difference between STAD and STAD-R.

The ratio R is the sum of those distances of datapoints in D_X that are directly connected in network U. Figure 2 ~~provides an intuition of~~ the creation of a STAD-R network during different iterations.

The result of STAD-R over STAD is presented in Figure 1E. The network has ~~a~~ considerable~~, lower number of~~ links (Figure 1C), and patterns in the data are much more apparent.

The STAD-R algorithm generates networks with considerably lower number of links compared to the correlation-based version. The ratio R restricts the inclusion of dissimilarities and therefore, the number of edges in the network. This new constraint also alters the number of edges in networks generated from other distributions types, e.g., right-skewed or normal. Nevertheless, the general ”shape” of the resulting network remains the same. An example is presented in Figure 3A, showing a right-skewed distance

STAD	STAD-R
$max\,\rho(D_X, D_U)$	$max\,\rho(D_X, D_U)R = max\,\rho\frac{\sum 1-d_{network\,edges}}{\sum 1+d_{network\,edges}}$

Table 2. Objective function in STAD and STAD-R. The correlation ρ is computed between the original distance matrix D_X and the distance matrix derived from the shortest path graph in D_U. The ratio R is calculated from the network at each iteration considering the edges included in the network. Note that distance $d_{network\,edge}$ are normalized values between zero and one.

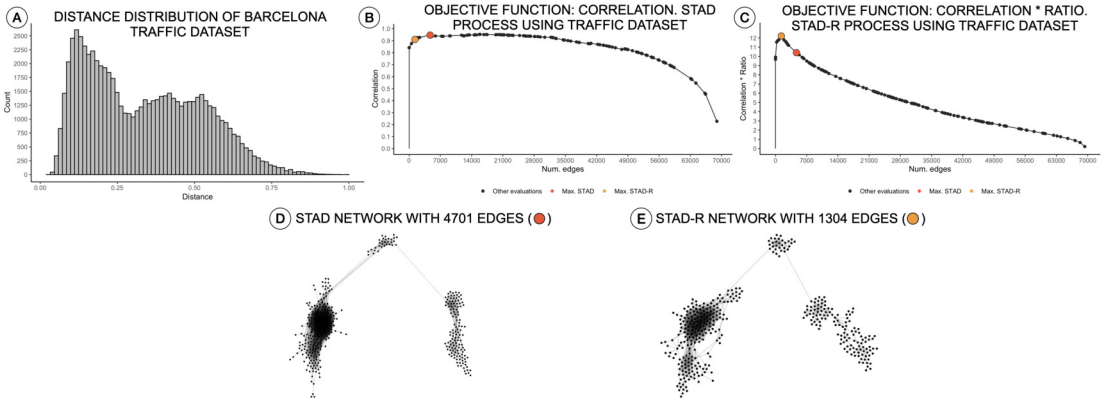

Figure 3. Distance distributions of traffic activity, STAD, and STAD-R projections. The dataset contains the traffic activity in the city of Barcelona from October 2017 until November 2018. The dataset was presented and analyzed in (Alcaide and Aerts, 2020). (A) Distribution of diagnosis distance. (B) Correlation between original distance matrix and distance matrix based on STAD graph, given different numbers of edges. (C) Idem as (B) using STAD-R. (D) STAD network. (E) STAD-R network.

distribution, leading to networks with different numbers of edges for STAD and STAD-R, respectively. However, the structure is still preserved in both networks (Figure 3D and E).

RESULTS

We applied this approach to the MIMIC-III database (Johnson et al., 2016), which is a publicly available dataset developed by the MIT Lab for Computation Physiology, containing anonymized health data from intensive care unit admissions between 2008 and 2014. The MIMIC-III dataset includes the diagnosis profiles of 58,925 patients. Their diagnoses are described using the ICD-9 codification and sorted according to their relevance to the patient. To reduce the number of distinct terms in the list of diagnoses, ICD codes were first grouped as described in the ICD guidelines Healthcare Cost and Utilization Project (2019). The proof-of-principle interface as well as the underlying code can be found on http://vda-lab.be/mimic.html.

The interface is composed of two main parts: an overview node-link network visualization including all patients (Figure 4A), and a more detailed view of selected profile groups (Figure 4B). Networks for each ICD code are precomputed: for each ICD-9 code the relevant patient subpopulations were extracted from the data, diagnosis distances and the resulting graph were computed using STAD-R. When the user selects an ICD-9 code from the interface (in this case code 2910; alcohol withdrawal delirium), the corresponding precomputed network is displayed.

The output of Louvain community detection (De Meo et al., 2011) is added as post-hoc annotation to facilitate the selection and exploration of the most evident patterns. The Louvain algorithm defines clusters by measuring the density of links inside the group compared to the links between them, which is close to the user interpretation of networks. However, the interpretation of a STAD-R network is not limited to discrete clusters. It aims to represent all relationships between points, including other types of patterns, such as trends or loops. The user can subsequently select either a cluster in this visualisation or individual patients, which will then trigger the display of a barchart which gives more information for that particular cluster (Figure 4B). This stacked barchart shows how different ICD codes are spread across the

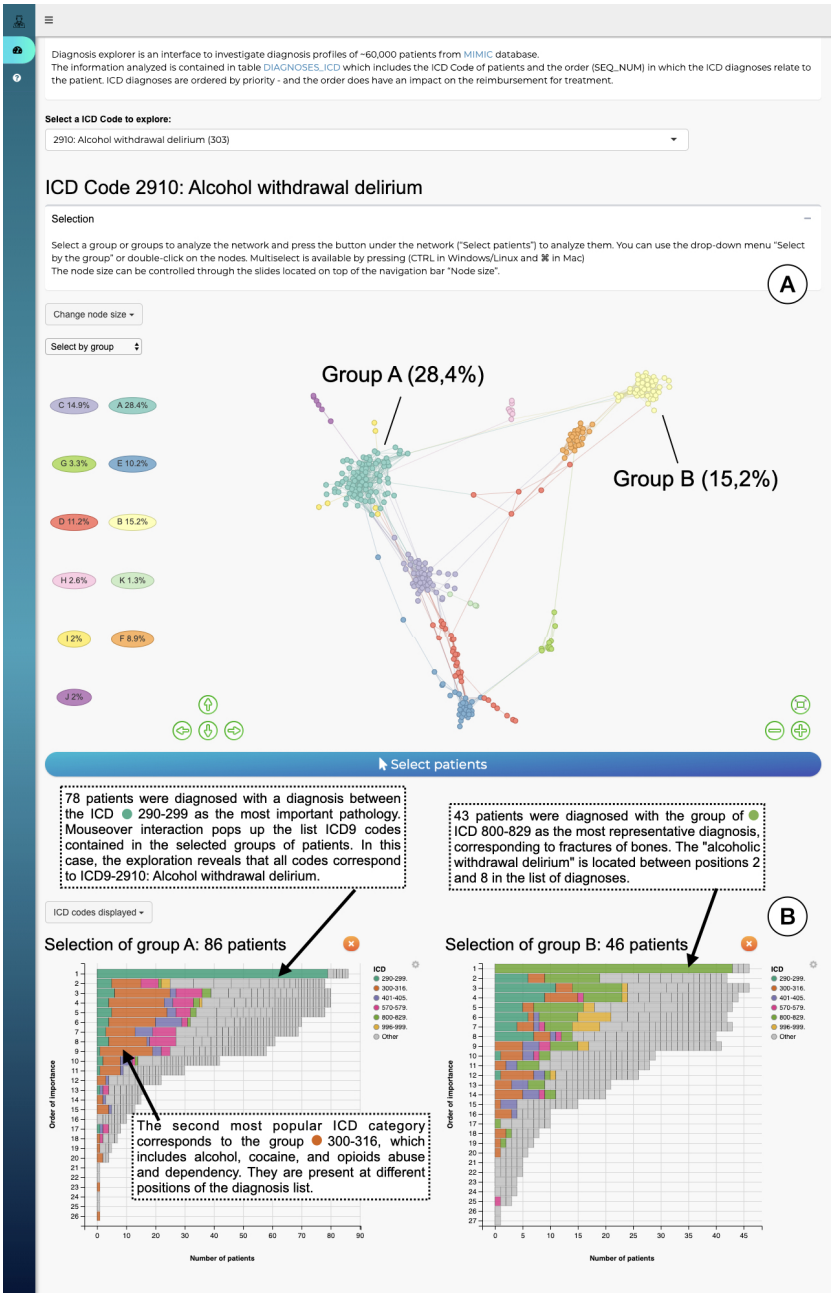

Figure 4. The interface to explore the diagnosis profiles in the MIMIC-III database. (A) Network visualization of those patients who have alcohol withdrawal delirium as one of their diagnoses. The network is visualized using a force-directed layout. Node colors are assigned automatically following Louvain community detection. (B) Bar-charts to compare the diagnosis profiles of selected groups in the network. Color corresponds to ICD category. In this example Group A contains patients with alcohol withdrawal delirium as the primary diagnosis; in contrast, Group B lists closed fractures as the most relevant diagnosis, and alcohol withdrawal delirium is only in the 2nd to 8th position.

different positions in the list of diagnoses: how many patients have code 2910 at the first position in the diagnosis list, how many at the second position, etc; the same goes for the other ICD codes. Total bar lengths decrease as the position in the list decreases due to the fact that different patients have different lengths of diagnosis lists.

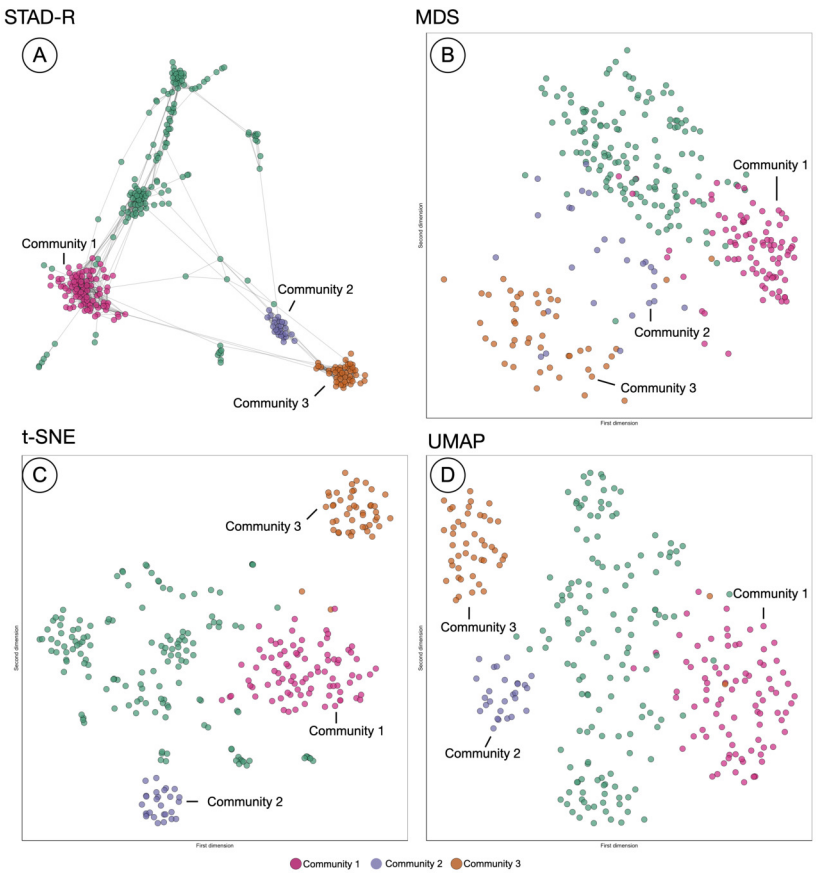

Figure 5. Comparison of STAD-R, MDS, t-SNE and UMAP using the population of patients with patients with alcohol withdrawal delirium (ICD-9 291.0). The three communities were determined by the Louvain algorithm. Community 1 are patients diagnosed with alcohol withdrawal delirium in the first positions of the list. Community 2 were patients with intracranial injuries as concussions. Community 3 are patients with fractures of bones as the primary diagnosis.

DISCUSSION

The definition of a custom similarity metric together with a flexible dimensionality reduction technique constitute the key elements of our approach. In this section, we evaluate the benefits of STAD to detect patterns in diagnostic data compared to other popular methods and further discuss the application of the presented distance metric in a different but similar context.

Comparing STAD to other dimensionality reduction methods

The projection of distances in STAD-R aims to enhance the representation of similarities using networks. Similar groups of patients tend to be inter-connected, which are perceived as a homogeneous cohort. The outputs of three popular algorithms (MDS, t-SNE, and UMAP) are compared with STAD-R in Figure 5. The population used in this example is the collection of MIMIC-III patients with alcohol withdrawal delirium (ICD-9 291.0), which was also used for Figure 4. The MDS projection endeavors to approximate all distances in data by defining the two most informative dimensions. Dimensionality methods such as t-SNE and UMAP favor the detection of local structures over the global, although UMAP also retains part of the general relations. Conversely, the abstract graph produced by STAD-R must still be embedded to be visualized, and the selection of the layout may produce slightly different results. Unlike scatterplots, node-link representations provide a more flexible platform for exploring data, especially when node positions can be readjusted according to the analyst and data needs (Henry et al., 2007).

In the four plots of Figure 5, the same points were highlighted to ease the comparison between them. These groups correspond to three communities identified by the Louvain method in the interface. For

instance, community 1 and 3 correspond to the patients analyzed in section 'Results'. Community 1 were patients diagnosed with alcohol withdrawal delirium as the primary diagnosis (Group A in Figure 4); community 3 are patients with fractures of bones as the primary diagnosis (Group B in Figure 4); community 2 are patients with intracranial injuries such as concussions. Despite the simple comparison presented, further analysis between these groups confirmed qualitative differences between profiles and a closer similarity between communities 2 and 3 than 1. The initial causes of communities 2 and 3 are associated with injuries while the primary diagnosis of patients in community 1 is the delirium itself.

In Figure 5, we can see that communities that are defined in the network (Figure 5A) are relatively well preserved in t-SNE (Figure 5C) but less so in MDS (Figure 5B). However, t-SNE does not take the global structure into account which is apparent from the fact that communities 2 and 3 are very far apart in t-SNE but actually are quite similar (STAD-R and MDS). UMAP (Figure 5D) improves on the t-SNE output and results in a view similar to MDS.

Although the interpretation of these visualizations is difficult to assess, quality metrics may help quantify the previous intuitions. Table 3 presents the quantitative measures for global distance and local distance preservation of projections in Figure 5. Global distance preservation was measured using the Spearman rank correlation (ρ_{Sp}). It compares the distances for every pair of points between the original data space and the two-dimensional projection (Zar, 2005). Local distance preservations were measured by the proportion of neighbors identified in the projection. This metric quantifies how many of the neighbors in the original space are neighbors in the projection (Espadoto et al., 2019). We evaluated this metric using a neighborhood of fourteen neighbors, ~~which~~ is the average cluster size in the MIMIC-III dataset using Louvain community detection ($14-nn$).

The richness of the node-link diagram representation of STAD-R cannot be captured using node position in the 2D plane alone. Therefore, STAD-R is analyzed from two perspective. First, the abstract graph ~~as~~ generated by STAD-R (STAD-R graph) and, second, the two-dimensional projection after graph drawing (STAD-R layout). The abstract graph only considers the connections between nodes to determine the distances between them, whereas the graph drawing results only consider the node placement in the 2D plane.

Based on the values from Table 3, the STAD-R obtained equivalent results to other dimensionality reduction methods in the preservation of the global and local structures. The abstract graph (STAD-R graph) is a not directly visible object as other methods. However, the node placement (STAD-R layout) is able to capture the information from the graph obtaining measures, in this example, close to UMAP, both global and local distances. ~~Besides,~~ node-link diagrams provide ~~the in~~tangible information ~~of~~ links ~~that~~ enhance the interpretation of relationships and allow thorough exploration through interactions such as drag nodes to other positions.

Global/local focus	Measure	MDS	t-SNE	UMAP	STAD-R graph	STAD-R layout
Global	ρ_{Sp}	0.54	0.41	0.47	0.52	0.47
Local	$14-nn$	0.34	0.60	0.53	0.62	0.52

Table 3. Distance preservation measures of projections in Figure 5. The table describes the Spearman's rank correlation (ρ_{Sp}) and the proportion of the first fourteen nearest neighbors preserved ($14-nn$). The selection of fourteen neighbors corresponds to the average cluster size in the MIMIC-III dataset using Louvain community detection. Column "STAD-R graph" represents the abstract graph and column "STAD-R layout" represents the node placement generated by a ForceAtlas2 layout (Jacomy et al., 2014) which is the layout implemented in the interface.

Similarity measures for ICD procedures

The diagnosis similarity described in section 'Diagnosis similarity and distances' is designed for assessing distance between diagnosis profiles, but the principles presented here can be generalized to other terminologies. For example, the procedures which patients receive during a hospital stay are also recorded and also follow an ICD codification: they also contain a list of categories similar to diagnosis. Unlike ICD diagnoses list, which encode priority, the order of procedure code lists indicate the sequence in which encode procedures were performed. Thus the weight distribution in the similarity that was used for the diagnosis metric must be adapted to the nature of the procedure data. ~~Therefore, we~~ can alter the formula

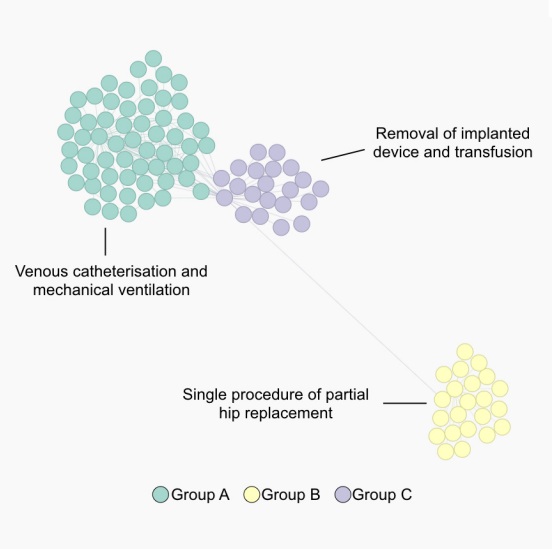

Figure 6. The population of patients who received a partial hip replacement (ICD 9: 81.52). The network was computed using STAD-R, and distances were estimated using an adapted version of diagnosis similarity for procedures. Color is based on Louvain community detection.

to include the relative distance between positions of matched elements instead of the top position in the diagnosis case. Formally, the similarity between two procedure concepts can be described as follows:

$$M_C(A,B) = ln\left(1 + \frac{1}{|position(C_A) - position(C_B)| + 1}\right)$$

As with diagnosis similarity, the metric is estimated as the sum of individual contributions of matched concepts, $S(X,Y) = \sum_{i=1}^{n} M(X \cap Y)$.

Figure 6 shows a STAD network generated using this adapted similarity for procedures. This example illustrates the population of patients with partial hip replacement (ICD 9: 81.52) in the MIMIC-III population. We can identify three clusters which describe three types of patients: group A are patients with the largest list of activities and are often characterized by venous catheterization and mechanical ventilation; patients in group B are mainly patients with a single procedure of partial hip replacement; patients in group C are characterized by the removal of an implanted device and a blood transfusion (data not shown).

CONCLUSIONS

In this paper, we introduced a custom distance metric for lists of diagnoses and procedures, as well as an extension to STAD for dissimilar datapoints. The diagnosis similarity measure can be applied to any ordered list of categories in a manner that is not possible with the measures available in the literature so far. The metric is designed to identify differences between patients through standardized concepts (diagnosis and procedures) where the weights of matching concepts are adapted to highlight the most relevant terms. As mentioned in Boriah et al. (2008), selecting a similarity measure must be based on an understanding of how it handles different data characteristics. The projection of data using STAD-R allows both for the detection of local structures and the representation of the global data structure. While no dimensionality reduction output from a high-dimensional dataset can completely project all relationships in the data, the connection of nodes in the graph allows a granular selection and exploration of cohorts. Furthermore, the embedding of the network into an interactive dashboard provides a level of convenience that supports interpretation of the analysis results of the network.

Moreover, as discussed previously, STAD-R can reveal equivalent data signals at multiple levels to other dimensionality reduction methods. Quantitative and qualitative (user) evaluation of the method can be further extended with other datasets to assess both the information captured by the graph and the benefits of node-links diagrams to represent the similarity between datapoints. Following this direction,

we plan to further explore STAD-R in collaboration with domain experts in diverse case studies. ~~Besides,~~ we plan to build a more robust interface that allows the computation and exploration of STAD-R networks ~~from~~ a ~~friendlier~~ environment.

ACKNOWLEDGEMENTS

This project is financed through the IWT SBO ACCUMULATE Grant nr 150056 and the Flemish Government "Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen" programme.

REFERENCES

Abdelmoula, W. M., Balluff, B., Englert, S., Dijkstra, J., Reinders, M. J., Walch, A., McDonnell, L. A., and Lelieveldt, B. P. (2016). Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences*, 113(43):12244–12249.

Ahmad, A. and Dey, L. (2007). A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recognition Letters*, 28(1):110–118.

Alcaide, D. and Aerts, J. (2020). Spanning trees as approximation of data structures. *IEEE Transactions on Visualization and Computer Graphics*.

Baker, F. and Porollo, A. (2018). Coeviz: A web-based integrative platform for interactive visualization of large similarity and distance matrices. *Data*, 3(1):4.

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., and Elhadad, N. (2018). Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38.

Boland, M. R., Tu, S. W., Carini, S., Sim, I., and Weng, C. (2012). Elixr-time: a temporal knowledge representation for clinical research eligibility criteria. *AMIA summits on translational science proceedings*, 2012:71.

Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM international conference on data mining*, pages 243–254. SIAM.

Brown, S.-A. (2016). Patient similarity: emerging concepts in systems and precision medicine. *Frontiers in physiology*, 7:561.

Campbell, P. G., Malone, J., Yadla, S., Chitale, R., Nasser, R., Maltenfort, M. G., Vaccaro, A., and Ratliff, J. K. (2011). Comparison of icd-9–based, retrospective, and prospective assessments of perioperative complications: assessment of accuracy in reporting. *Journal of Neurosurgery: Spine*, 14(1):16–22.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. (2016). Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318.

Corder, G. W. and Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons.

Dagliati, A., Geifman, N., Peek, N., Holmes, J. H., Sacchi, L., Sajjadi, S. E., and Tucker, A. (2019). Inferring temporal phenotypes with topological data analysis and pseudo time-series. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 399–409. Springer.

Damen, D., Luyckx, K., Hellebaut, G., and Van den Bulcke, T. (2013). Pastel: A semantic platform for assisted clinical trial patient recruitment. In *2013 IEEE International Conference on Healthcare Informatics*, pages 269–276. IEEE.

De Meo, P., Ferrara, E., Fiumara, G., and Provetti, A. (2011). Generalized louvain method for community detection in large networks. In *2011 11th International Conference on Intelligent Systems Design and Applications*, pages 88–93. IEEE.

Espadoto, M., Martins, R. M., Kerren, A., Hirata, N. S., and Telea, A. C. (2019). Towards a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–20.

Fink, E., Hall, L. O., Goldgof, D. B., Goswami, B. D., Boonstra, M., and Krischer, J. P. (2003). Experiments on the automated selection of patients for clinical trials. In *SMC'03 Conference Proceedings.*

2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483), volume 5, pages 4541–4545. IEEE.

Franken, N. (2009). Visual exploration of algorithm parameter space. In *2009 IEEE Congress on Evolutionary Computation*, pages 389–398. IEEE.

Girardi, D., Wartner, S., Halmerbauer, G., Ehrenmüller, M., Kosorus, H., and Dreiseitl, S. (2016). Using concept hierarchies to improve calculation of patient similarity. *Journal of biomedical informatics*, 63:66–73.

Gottlieb, A., Stein, G. Y., Ruppin, E., Altman, R. B., and Sharan, R. (2013). A method for inferring medical diagnoses from patient similarities. *BMC medicine*, 11(1):194.

Healthcare Cost and Utilization Project (2019). Clinical classifications software (icd-9-cm) summary and download. summary and downloading information. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp. Accessed: September 11, 2020.

Henry, N., Fekete, J.-D., and McGuffin, M. J. (2007). Nodetrix: a hybrid visualization of social networks. *IEEE transactions on visualization and computer graphics*, 13(6):1302–1309.

Humphries, K. H., Rankin, J. M., Carere, R. G., Buller, C. E., Kiely, F. M., and Spinelli, J. J. (2000). Co-morbidity data in outcomes research are clinical data derived from administrative databases a reliable alternative to chart review? *Journal of clinical epidemiology*, 53(4):343–349.

Ienco, D., Pensa, R. G., and Meo, R. (2012). From context to distance: Learning dissimilarity for categorical data clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9(6):e98679.

Jia, H., Cheung, Y.-m., and Liu, J. (2015). A new distance metric for unsupervised learning of categorical data. *IEEE transactions on neural networks and learning systems*, 27(5):1065–1079.

Jia, Z., Lu, X., Duan, H., and Li, H. (2019). Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC medical informatics and decision making*, 19(1):91.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Koffka, K. (2013). *Principles of Gestalt psychology*. Routledge.

Kolodner, J. (2014). *Case-based reasoning*. Morgan Kaufmann.

Le, S. Q. and Ho, T. B. (2005). An association-based dissimilarity measure for categorical data. *Pattern Recognition Letters*, 26(16):2549–2557.

Lee, J., Maslove, D. M., and Dubin, J. A. (2015). Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PloS one*, 10(5):e0127428.

Li, L., Cheng, W.-Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., and Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, 7(311):311ra174–311ra174.

Liu, S., Maljovec, D., Wang, B., Bremer, P.-T., and Pascucci, V. (2016). Visualizing high-dimensional data: Advances in the past decade. *IEEE transactions on visualization and computer graphics*, 23(3):1249–1268.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094.

Moni, M. A., Xu, H., and Lio, P. (2014). Cytocom: a cytoscape app to visualize, query and analyse disease comorbidity networks. *Bioinformatics*, 31(6):969–971.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71.

Mukherjee, S., Sinha, B. K., and Chattopadhyay, A. K. (2018). Multidimensional Scaling. In *Statistical Methods in Social Science Research*, pages 113–122. Springer.

Nguyen, Q. V., Nelmes, G., Huang, M. L., Simoff, S., and Catchpoole, D. (2014). Interactive visualization for patient-to-patient comparison. *Genomics & informatics*, 12(1):21.

Nielson, J. L., Paquette, J., Liu, A. W., Guandique, C. F., Tovar, C. A., Inoue, T., Irvine, K.-A., Gensel,

J. C., Kloke, J., Petrossian, T. C., Lum, P. Y., Carlsson, G. E., Manley, G. T., Young, W., Beattie, M. S., Bresnahan, J. C., and Ferguson, A. R. (2015). Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature communications*, 6:8581.

Pai, S. and Bader, G. D. (2018). Patient similarity networks for precision medicine. *Journal of molecular biology*, 430(18):2924–2938.

Pai, S., Hui, S., Isserlin, R., Shah, M. A., Kaka, H., and Bader, G. D. (2019). netdx: Interpretable patient classification using integrated patient similarity networks. *Molecular systems biology*, 15(3).

Pereira, V., Waxman, D., and Eyre-Walker, A. (2009). A problem with the correlation coefficient as a measure of gene expression divergence. *Genetics*, 183(4):1597–1600.

Popescu, M. and Khalilia, M. (2011). Improving disease prediction using icd-9 ontological features. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, pages 1805–1809. IEEE.

Real, R. and Vargas, J. M. (1996). The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385.

Rivault, Y., Le Meur, N., and Dameron, O. (2017). A similarity measure based on care trajectories as sequences of sets. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 278–282. Springer.

Ronzano, F., Gutiérrez-Sacristán, A., and Furlong, L. I. (2019). Comorbidity4j: a tool for interactive analysis of disease comorbidities over large patient datasets. *Bioinformatics*.

Saeed, N., Nam, H., Haq, M. I. U., and Muhammad Saqib, D. B. (2018). A survey on multidimensional scaling. *ACM Computing Surveys (CSUR)*, 51(3):47.

Sahoo, S. S., Tao, S., Parchman, A., Luo, Z., Cui, L., Mergler, P., Lanese, R., Barnholtz-Sloan, J. S., Meropol, N. J., and Zhang, G.-Q. (2014). Trial prospector: matching patients with cancer research studies using an automated and scalable approach. *Cancer informatics*, 13:CIN–S19454.

Schork, N. J. and Zapala, M. A. (2012). Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. *Frontiers in genetics*, 3:190.

Simoni, Y., Becht, E., Fehlings, M., Loh, C. Y., Koo, S.-L., Teng, K. W. W., Yeong, J. P. S., Nahar, R., Zhang, T., Kared, H., Duan, K., Ang, N., Poidinger, M., Lee, Y. Y., Larbi, A., Khng, A. J., Tan, E., Fu, C., Mathew, R., Teo, M., Lim, W. T., Toh, C. K., Ong, B.-H., Koh, T., Hillmer, A. M., Takano, A., Lim, T. K. H., Tan, E. H., Zhai, W., Tan, D. S. W., Tan, I. B., and Newell, E. W. (2018). Bystander cd8+ t cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature*, 557(7706):575.

Šulc, Z. and Řezanková, h. (2014). Evaluation of recent similarity measures for categorical data. In *Proceedings of the 17th International Conference Applications of Mathematics and Statistics in Economics. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wroclaw*, pages 249–258.

Urpa, L. M. and Anders, S. (2019). Focused multidimensional scaling: interactive visualization for exploration of high-dimensional data. *BMC bioinformatics*, 20(1):221.

Van Dongen, S. and Enright, A. J. (2012). Metric distances derived from cosine similarity and pearson and spearman correlations. *arXiv preprint arXiv:1208.3145*.

Vijaymeena, M. and Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28.

Zar, J. H. (2005). Spearman rank correlation. *Encyclopedia of Biostatistics*, 7.

Zhang, P., Wang, F., Hu, J., and Sorrentino, R. (2014). Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Summits on Translational Science Proceedings*, 2014:132.

Zillner, S., Hauer, T., Rogulin, D., Tsymbal, A., Huber, M., and Solomonides, T. (2008). Semantic visualization of patient information. In *2008 21st IEEE International Symposium on Computer-Based Medical Systems*, pages 296–301. IEEE.