Decision Support Systems 148 (2021) 113583

Contents lists available at ScienceDirect

Decision Support Systems

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

Clustering temporal disease networks to assist clinical decision support systems in visual analytics of comorbidity progression

Yajun Lu [a], Suhao Chen [b,c], Zhuqi Miao [c,*], Dursun Delen [c,d], Andrew Gin [c]

[a] Department of Management and Marketing, Jacksonville State University, Jacksonville, AL, USA
[b] School of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK, USA
[c] Center for Health Systems Innovation, Oklahoma State University, Stillwater, OK, USA
[d] Department of Management Science and Information Systems, Oklahoma State University, Tulsa, OK, USA

ARTICLE INFO

Keywords:
Clinical decision support
Visual analytics
Comorbidity progression
Temporal disease network
Clustering

ABSTRACT

Detection and characterization of comorbidity, the presence of more than one distinct disorder or illness concurrently occurring among a specific cohort of patients, is an invaluable decision aid and a prominent challenge in healthcare research and practice. The aim of this paper is to design a novel visual analytics system that can support efficient pattern detection and intuitive visualization of comorbidity progression modeled via temporal disease networks (TDNs). In the underlying system, we proposed two new clustering technologies—temporal clustering and disease clustering to detect the time of notable progression changes and simplify the visualization of TDNs. Through two case studies on Clostridioides Difficile and stroke, we demonstrate that the proposed system is able to provide evidence-based and visual insights regarding comorbidity progression effectively for clinical decision support.

1. Introduction

The widespread adoption of electronic health records (EHR) [1,2] and the increasing emphasis on the use of clinical decision support systems (CDSS) [3–5] have been two of the most remarkable outcomes of healthcare reform in the U.S. during the past decade. The adoption rate of basic EHR systems among U.S. hospitals has surged from 9.4% in 2008 to 83.8% in 2015 [6]. The ubiquitous adoption of EHR by health systems has generated an unprecedented amount of health data, which provide the longitudinal picture of patients' journeys, treatment pathways and care outcomes [7]. A CDSS refers to "any electronic system designed to aid directly in clinical decision making, in which characteristics of individual patients are used to generate patient-specific assessments or recommendations that are then presented to clinicians for consideration [8]". The abundance and comprehensiveness of EHR data, in conjunction with recent advances in CDSS, has offered researchers and practitioners an ideal platform to mine actionable insights to improve clinical decision making for better healthcare outcomes [4,9,10]. Specific applications include test ordering [11], therapy management [12], improving care delivery and access [13,14], detecting and predicting health conditions [15,16], and medication evaluation [17], among others.

Visual analytics (VA) can reduce the information overload on memory and cognition, and leverage the power of human perception [18,19]. Nowadays, it has become an integral component of CDSS [20–24]. For example, through VA, large volumes of data and complex ideas in healthcare settings can be presented with clarity, accuracy, and efficiency in visual diagrams [25,26]. Furthermore, VA dashboards allow real-time monitoring and tracking of healthcare information, such as hospital-specific antibiograms [27], adverse drug events [28], and departmental performance metrics [29]. It also has been reported that visualized data improved recall of important clinical information [30].

An emerging and important direction of VA in clinical decision making is to visualize and mine comorbidity progression patterns [31–35]. Comorbidity refers to one or more other health conditions coexisting with a particular index disease under investigation [36]. Comorbidity has been increasingly prevalent [37] and consistently challenging healthcare practice and research by leading to worse health outcomes, complicating diagnostics and treatments, and misleading medical statistics [36,38]. As a result, great efforts have been devoted to exploring effective methodology to handle comorbidity to improve clinical decision making during the past few decades [39–41]. Network modeling represents an intuitive and useful approach to investigate

* Corresponding author.
E-mail addresses: ylu@jsu.edu (Y. Lu), suhao.chen@okstate.edu (S. Chen), zhuqi.miao@okstate.edu (Z. Miao), dursun.delen@okstate.edu (D. Delen), andrew.gin@okstate.edu (A. Gin).

https://doi.org/10.1016/j.dss.2021.113583
Received 24 August 2020; Received in revised form 31 March 2021; Accepted 27 April 2021
Available online 6 May 2021

comorbid diseases and their progression patterns [42–44] mainly for the following advantages:

- *User-friendly presentation of disease associations.* By modeling comorbid diseases as nodes and their pairwise associations as edges, network models can present comorbidity visually. The nodes and edges can further carry attributes to express specific features of diseases and disease associations. Examples of such attributes include node size for disease prevalence and edge weight for association strength [45,46]. Furthermore, edges can be directed to represent the dynamic (e.g. causal or sequential) interactions among diseases [33,47].
- *Support for disease progression analysis.* By discretizing the entire time frame of the index disease into different time windows, modeling comorbidity within each window as a disease network (hereafter referred to as *temporal disease network*, TDN), then comparing the dynamics through the TDN sequence across different windows, researchers are able to show and analyze the progression of the index condition and comorbid diseases. This approach has been applied to chronic conditions, such as cancer and mental disorders, which often come with long period and multiple, comorbid diseases [49,50].
- *Capability to incorporate additional biomarkers.* In addition to diseases, other biomarkers, such as genes and symptoms, can also be modeled as nodes and incorporated into the disease network by establishing edges that are representative of associations between the diseases and the biomarkers. For example, “diseasome” networks incorporate genes and/or proteins as nodes, and link them with diseases [51,52], and psychiatric symptom networks include symptoms, drugs and even adverse effects of drugs in addition to diseases [42,53].

The objective of this research is to design a VA system that can efficiently detect and visualize comorbidity progressions using TDN models. The VA system incorporates two novel TDN clustering technologies—*temporal clustering* and *disease clustering*. The temporal clustering identifies notable changes during the comorbidity progression and aggregates windows to phases based on the time of the changes. On the other hand, the disease clustering captures higher-level structures of TDNs by clustering highly coexisting conditions and simplifies the TDN visualization based on the identified structures. The developed VA system can be integrated into CDSS to provide evidence-based, visual insights regarding the timeline and patterns of comorbidity progression to support the decision making in healthcare settings.

The remainder of this article is organized as follows. In Section 2, we provide a literature review of related work in the area of TDNs, and show the intellectual gaps that we are addressing in this research. The system design and proposed clustering technologies are presented in detail in Section 3, followed by two case studies on Clostridioides Difficile (C. Diff) and stroke in Section 4. Finally, Sections 5 and 6 include the discussion and conclusion of this study, respectively.

2. Related work and gaps

2.1. Network theory and TDN

A most fundamental network (aka graph) model, denoted by G in this article, is a mathematical structure composed of a set of nodes $V(G)$ that model the objects of interest and a set of undirected edges $E(G)$ representing the pairwise relationships among the objects [54]. The number of nodes and the number of edges included in a network are called the *order* and the *size* of the network, and are denoted by $|V(G)|$ and $|E(G)|$ respectively. Given a subset of nodes $S \subseteq V(G)$, we herein denote by $G[S]$ the subgraph induced by S, i.e. a subgraph obtained by dropping nodes in $V(G)\backslash S$ and their incident edges from G. For a node $v \in V(G)$, the neighbors of v, $N_G(v)$ refers to the set of nodes adjacent to v and its cardinality is called the *degree* of v, denoted by $\deg_G(v)$ herein. In this article, node v and its neighbors $N_G(v)$ together are referred to as the *closed neighborhood* of v, denoted by $N_G[v]$, and its induced subgraph is called the ego network of v, denoted by $ego_G(v)$. In other words, $N_G[v]{:=}N_G(v) \cup \{v\}$ and $ego_G(v){:=}G[N_G[v]]$.

In a basic undirected, unweighted network modeling comorbidity, nodes represent comorbid diseases, while edges manifest the coexistence relationships among diseases in a certain patient cohort. The coexistence relationship is usually evaluated using a statistical measure, such as relative risk [55], Pearson’s correlation [32], and Salton Cosine Index (SCI) [56], among others. Then, a threshold is used to eliminate trivial coexistences and retain the significant ones as edges. Comorbidity networks are often large, dense, and complicated. To facilitate the analysis and visualization of complex comorbidity networks, graph clustering methods have been often used to detect comorbidity patterns [56,57] and reduce network complexity [58]. A commonly used network clustering model is the clique, i.e. a complete graph, in which all nodes are pairwise interconnected [59,60]. For instance, in Fig. 1, the TDN at Window 1 is a clique of three nodes.

Given a sequence of TDNs across different time windows, progression analysis often involves comparing how much the TDNs are dissimilar from each other. There have been abundant approaches proposed in literature to measure the network dissimilarity [61,62]. A majority of these methods summarize the structural features of a network into a vector of statistics, then define the dissimilarity between a pair of networks as the distance (e.g. Euclidean or Manhattan distance) between the two vectors associated with the networks. In addition to basic structures in network theory, e.g. node degree and network diameter, the literature also used many advanced structural features, including cluster coefficient [63], graphlet [64], and graph kernel [65,66], to name a few.

2.2. Intellectual gaps

Through a thorough literature review, we found that in the area of TDN modeling and analysis, there has been limited work to:

- *Outline progression phases.* Most TDN-related studies [49,67,68] predefined a granularity parameter m, then discretized the entire time frame of the study cohort into m windows of even length or even sample size, without providing algorithms that can detect at which window(s) notable changes of TDNs had occurred. Another issue brought by the simple m-window discretization method is that when the granularity is high, many windows come with very similar TDNs, which increases the redundancy of visualization, especially at late stages of the time frame, when the number of comorbid diseases grows to a stable level.
- *Streamline the visualization.* Network clustering methods, such as the clique model, can be used to streamline the visualization of a single network as discussed earlier, but the extension to TDNs across multiple time windows is not straightforward. A confusion is that a clique in one window may be divided in another window, as shown in Fig. 1. For complex TDNs with large size and many windows, the

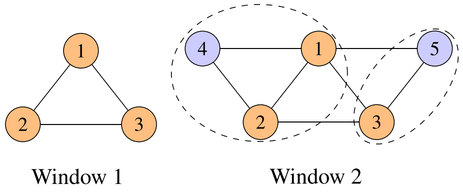


Fig. 1. Two TDNs in different windows. The one at Window 1 is a clique with three nodes. In Window 2, a clique enumeration algorithm may detect two cliques as circled, which is a split of the clique at the earlier window, causing analysts to lose track of the disease cluster implied by the clique.

confusion will be much deteriorated, leading to the loss of track of certain disease clusters.

The new technologies developed in this research are dedicated to filling in these two gaps and providing an effective VA tool to discover insights in comorbidity progression.

3. Methodology

The proposed VA system consists of four modules as shown in Fig. 2. Module 1 receives data from clinical data warehouses and prepares the data for subsequent analysis and visualization. Module 2 then builds TDNs with sufficient granularity using the preprocessed data, while Module 3 identifies highly similar TDNs and clusters corresponding windows into phases, followed by Module 4 that visualizes TDNs over the phases. The role of clinical domain experts is to guide the process of modules by determining the initialization parameters and examine the output for validity, while the system eventually provides visual insights regarding comorbidity progression back to the clinical experts to support evidence-based decision making. Since the technological contribution of our work mainly revolves around Modules 2, 3 and 4, the rest of this section will focus on elaborating the methodology we employed to design these modules.

3.1. TDN construction

In order to quantify the coexistence relationship among comorbid diseases, we make use of SCI [56,69], which can be expressed as

$$SCI_{ij} = \frac{n_{ij}}{\sqrt{n_i n_j}} \tag{1}$$

where n_{ij} represents the number of hospital encounters with the onset of both diseases i and j, while n_i (or n_j) corresponds to the number of encounters with the onset of disease i (or j). When $SCI_{ij} = \gamma\%$, at least one of n_{ij}/n_i and n_{ij}/n_j is no less than $\gamma\%$. It implies that encounters with the onset of both diseases are at least γ percent of all encounters of one disease. *SCI* has been used as an alternative of Pearson's correlation coefficient (PCC) for disease network modeling because it avoids two potential issues of PCC: (i) sample size can have overly high impact on the PCC strength [69], and (ii) PCC may underestimate the coexistence of a pair of diseases, of which one is rare while the other is prevalent [70].

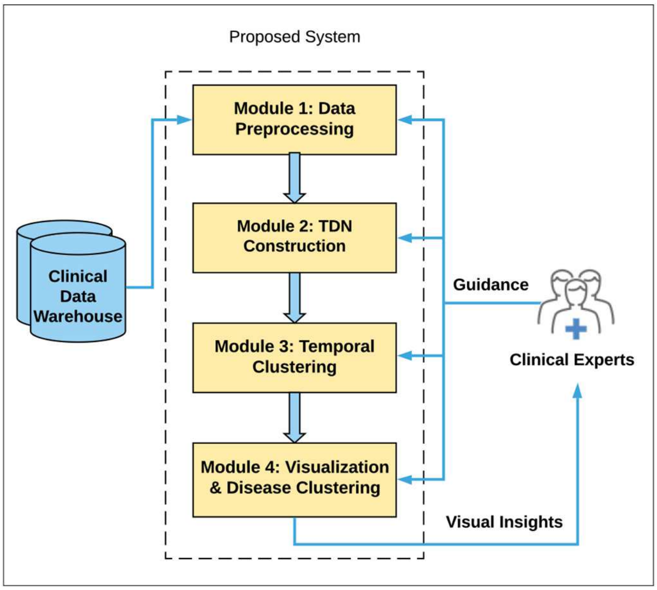

Proposed System
Module 1: Data Preprocessing
Module 2: TDN Construction
Module 3: Temporal Clustering
Module 4: Visualization & Disease Clustering
Clinical Data Warehouse
Guidance
Clinical Experts
Visual Insights

Fig. 2. The proposed TDN-based VA system for pattern detection and visualization of comorbidity progressions.

Given an *SCI* threshold θ determined under the clinical experts' guidance, we can then establish an edge between each pair of diseases (nodes) i and j such that $SCI_{ij} \geq \theta$. In addition to θ, the system also needs the advice from clinical experts to specify a value for the granularity parameter m to discretize the entire time frame into m windows that are as granular as possible. *Re*-organization of the windows will be accomplished by the Temporal Clustering Module of the system.

3.2. Temporal clustering

This module involves two techniques: (i) network dissimilarity measurement, and (ii) consecutive p-median clustering for time windows, as elaborated in the following.

3.2.1. Network dissimilarity measurement

In this research, we adapted and improved the *NetSimile* method proposed by Berlingerio et al. [63] to evaluate the network dissimilarity among different windows. The NetSimile method "quantifies" the structural features of a network G by calculating multiple statistical metrics (including median, mean, standard deviation, skewness, and kurtosis in this study) for a number of features associated with each node $v \in V(G)$. The specific features employed in this study include:

- The degree of v;
- Clustering coefficient of v, defined as $\frac{2}{deg_G(v)(deg_G(v)-1)}|E(G[N_G(v)])|$;
- The average degree of the neighbors of v;
- The average clustering coefficient of the neighbors of v;
- The size of the ego network of v;
- The number of edges connecting the nodes in $ego_G(v)$ and nodes not in $ego_G(v)$;
- The number of nodes that are not in $ego_G(v)$, but are neighbors of nodes in $ego_G(v)$.

The process results in a 35-entry vector of statistics that evaluates the structure of a network. The vector is herein referred to as the *signature vector*, and denoted by Z_G for a given network G.

In the classical NetSimile method, the dissimilarity between a pair of networks, G_i and G_j, was assessed using the Canberra distance of the corresponding signature vectors, defined as

$$\delta(G_i, G_j) = \frac{1}{35}\sum_{k=1}^{35}\frac{|Z_{G_i}[k] - Z_{G_j}[k]|}{|Z_{G_i}[k]| + |Z_{G_j}[k]|} \tag{2}$$

where $Z_{G_i}[k]$ (or $Z_{G_j}[k]$) indicates the kth entry of the vector Z_{G_i} (or Z_{G_j}). The similarity then can be expressed as $1-\delta(G_i, G_j)$. Nevertheless, the classical NetSimile method does not consider the disparity of node sets, thus can underestimate the dissimilarity when there are uncommon nodes between two networks. Considering the two TDNs, G_1 and G_2 shown in Fig. 3, $\delta(G_1, G_2) = 0$ indicating the "identical" edge structure between the two TDNs. However, the structure is based on different node sets (new diseases 4 and 5 are developed from G_1 to G_2, whereas diseases 1 and 2 become absent), thus are actually not the same. The

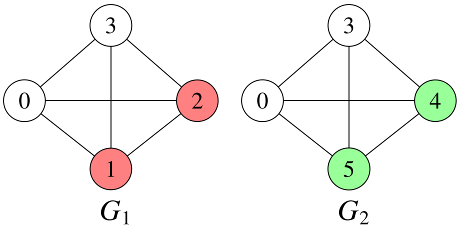


Fig. 3. Two TDNs that are cliques with different node sets.

classical NetSimile method fails to reflect such a disparity. This issue motivates us to introduce an overlapping factor to enhance the NetSimile method to handle the dissimilarity caused by the difference between node sets. The overlapping factor $\omega(G_i, G_j)$ is defined as follows,

$$\omega\left(G_i, G_j\right) = \frac{\left|E\left(G_i[D]\right)\right| + \left|E\left(G_j[D]\right)\right|}{\left|E\left(G_i\right)\right| + \left|E\left(G_j\right)\right|} \tag{3}$$

where $D = V(G_i) \cap V(G_j)$. Clearly, $0 \le \omega(G_i, G_j) \le 1$. By incorporating $\omega(G_i, G_j)$, the modified dissimilarity $d(G_i, G_j)$ is expressed as

$$d\left(G_i, G_j\right) = 1 - \omega\left(G_i, G_j\right) \times \left(1 - \delta\left(G_i[D], G_j[D]\right)\right) \tag{4}$$

The rationale behind the modified formula is straightforward: $1 - \delta(G_i[D], G_j[D])$ evaluates the similarity between the node-overlapping subgraphs of G_i and G_j. Because the rest parts are completely different, we scale down $1 - \delta(G_i[D], G_j[D])$ with the overlapping factor $\omega(G_i, G_j)$ to evaluate the overall similarity between the two entire networks. *Reconsidering* the two networks in Fig. 3, the dissimilarity between node-overlapping subgraphs $\delta(G_i[D], G_j[D]) = 0$ and $\omega(G_1, G_2) = 1/6$, therefore $d(G_1, G_2) = 5/6$ and the overall similarity between the two networks is 1/6, which is a better evaluation compared with that returned by the classical NetSimile method.

3.2.2. *Consecutive p-median clustering*

As we pointed out in Section 2.2, some consecutive windows can come with very similar TDNs, thus providing limited new information about comorbidity progression, and leading to visualization redundancy. In order to address the issue, we propose and solve a *consecutive p-median problem* (CPMP) defined as follows.

Problem: Consecutive *p*-median problem.

Input: A positive integer p, a collection of m objects $\mathscr{O} := \{O_1, O_2, \ldots, O_m\}$, and the distance between any two objects.

Output: From $\mathscr{O}$, find p objects with indices $\{j_1, j_2, \ldots, j_p\}$ as medians and assign the remaining $m - p$ objects to the medians such that.

(i) The total summation of distances from each O_i to its assigned median is minimized, and

(ii) When O_i is assigned to median O_{jq}, if $i \ge j_q$ then O_k for all k such that $j_q \le k < i$ must be assigned to O_{jq}, otherwise O_k for all k such that $i < k \le j_q$ must be assigned to O_{jq}.

The problem is an extension of the classical p-median problem that has been often used for clustering [71,72]. The extension is condition (ii) that impose the assignment of consecutive objects to medians. For example, if we would like to solve the consecutive 2-median problem on the TDNs shown in Fig. 4, we cannot assign the TDNs on Window 1 and Window 5 together even though they are identical. By applying CPMP to TDNs, we can group consecutive windows with highly similar TDNs into a same temporal cluster, which can be interpreted as a *phase* of comorbidity progression.

In this study, we developed a (linear) integer programming (IP) formulation (5)–(12) to model and solve the CPMP on a sequence of TDNs, $\mathscr{G} = \{G_1, G_2, \ldots, G_m\}$. In the formulation, the binary variable $x_{ij} = 1$ if and only if TDN G_i is assigned to median G_j, for any $i, j \in \{1, 2, \ldots, m\}$ such that $i \ne j$, otherwise $x_{ij} = 0$. When $x_{jj} = 1$, it indicates that G_j is designated as a median for any $j \in \{1, 2, \cdots, m\}$. The objective function (5) aims to minimize the total dissimilarity between TDNs and the medians to which the TDNs are assigned across all windows. Constraint (6) ensures that at most p TDNs are selected as medians. In Constraint (7), we force each TDN G_i to be assigned to exactly one median. While Constraint (8) guarantees that if G_i is assigned to G_j then G_j must be a median. Constraints (9) and (10) make sure that only consecutive TDNs can be grouped into a same cluster. In constraint (11), τ represents a threshold for *not* clustering. When the dissimilarity between two consecutive TDNs G_i and G_{i+1} is greater than or equal to τ, they will not be grouped into a same cluster. This constraint allows us to avoid clustering highly different TDNs.

$$min \sum_{i=1}^{m} \sum_{j=1}^{m} d\left(G_i, G_j\right) x_{ij} \tag{5}$$

$$\text{subject to :} \quad \sum_{j=1}^{m} x_{jj} \le p \tag{6}$$

$$\sum_{j=1}^{m} x_{ij} = 1 \quad \forall i \in \{1, 2, \cdots, m\} \tag{7}$$

$$x_{ij} \le x_{jj} \quad \forall i, j \in \{1, 2, \cdots, m\} | i \ne j \tag{8}$$

$$x_{ij} \le x_{kj} \quad \forall i \in \{1, 2, \cdots, m-2\}, j \in \{i+2, i+3, \cdots, m\}, k \in \{i+1, i+2, \cdots, j-1\} \tag{9}$$

$$x_{ij} \le x_{kj} \quad \forall i \in \{3, 4, \cdots, m\}, j \in \{1, 2, \cdots, i-2\}, k \in \{j+1, j+2, \cdots, i-1\} \tag{10}$$

$$x_{ij} + x_{i+1,j} \le 1 \quad \forall j \in \{1, 2, \cdots, m\}, i \in \{1, 2, \cdots, m-1\} | d(G_i G_{i+1}) \ge \tau \tag{11}$$

$$x_{ij} \in \{0, 1\} \quad \forall i, j \in \{1, 2, \cdots, m\} \tag{12}$$

3.2.3. *Selection of the value for p*

The parameter p determines how many clusters the initial time windows should be grouped into; in other words, how many phases the entire time frame is supposed to be broken down into. Usually, we are interested in a relatively small p to simplify the TDN sequence to allow us to capture the primary changes of comorbidity over time. Meanwhile, we need to avoid using a value that is too small, because an overly small p can result in very broad phases that combine windows little similar. The clinical advice from domain experts is essential to decide a proper value of p. Whereas, data analytic methods can also be used to support

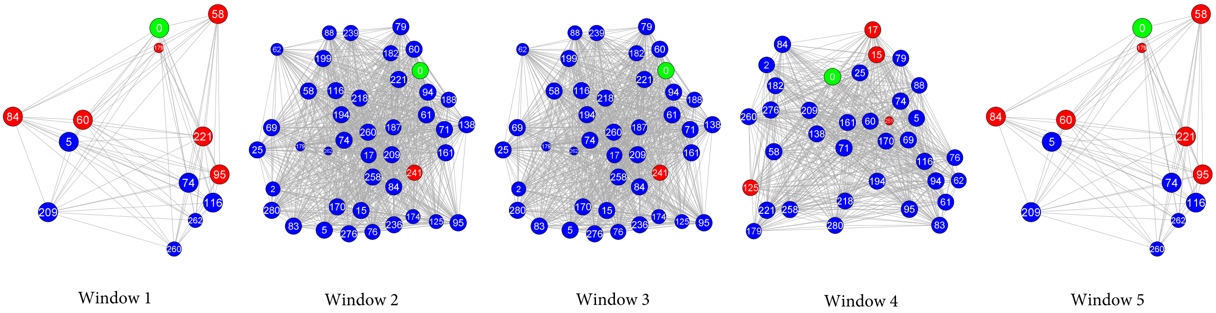


Fig. 4. A sequence of TDNs across five time windows, of which the ones on Window 1 and Window 5 are identical and the ones through Window 2 to Window 4 are highly similar.

the decision on this parameter.

The Silhouette Index (*SI*) has often been used in literature to determine the value of p for p-median models [72,73]. In this study, we adapted *SI* to find a proper value of p for our proposed CPMP. Let $\mathscr{C}(p) = \{C_1, C_2, \dots C_p\}$ be a clustering of TDNs $\mathscr{G} = \{G_1, G_2, \dots, G_m\}$; given a network $G \in \mathscr{G}$, let C_k represent the cluster that contains G and $\mathscr{G}_A$ denote the network(s) in $\mathscr{G}$ that are adjacent to G. Then, our adapted *SI* for G is defined as

$$SI_{\mathscr{G}}(\mathscr{C}(p), G) = \begin{cases} 0 & \text{if } |C_k| \geq \sigma|\mathscr{G}| \\ 0 & \text{if } |C_k| = 1 \text{ and } \exists \widehat{G} \in \mathscr{G}_A | d\left(G, \widehat{G}\right) < \tau \\ 1 & \text{if } |C_k| = 1 \text{ and } d\left(G, \widehat{G}\right) \geq \tau, \forall \widehat{G} \in \mathscr{G}_A \\ \dfrac{\Delta_{\mathscr{C} \setminus C_k}(G) - \Delta_{C_k}(G)}{max\left\{\Delta_{C_k}(G), \Delta_{\mathscr{C} \setminus C_k}(G)\right\}} & \text{if } 2 \leq |C_k| < \sigma|\mathscr{G}| \end{cases} \tag{13}$$

The adapted *SI* considers four scenarios: (i) When a cluster contains too many TDNs ($\sigma|\mathscr{G}|$ or more), or (ii) a cluster contains a single TDN, but this TDN does not differ much (dissimilarity is less than τ) from an adjacent TDN, then the *SI* is set to be 0 to discourage the scenarios. (iii) However, when a single TDN is too dissimilar (dissimilarity is τ or larger) from adjacent TDNs to be grouped into other clusters, we let *SI* be 1 to allow the TDN to form a cluster by itself. (iv) When a cluster is neither too large (less than $\sigma|\mathscr{G}|$) nor too small (size is at least 2), we compute an *SI* that measures how a TDN is similar to its assigned cluster compared with other clusters. $\Delta_{C_k}(G) = \frac{1}{|C_k|-1}\sum_{\widehat{G} \in C_k \setminus G} d\left(G, \widehat{G}\right)$ is the "internal distance" of G within its own cluster, defined as the average dissimilarity between G and the other networks in the cluster that G belongs to. While $\Delta_{\mathscr{C} \setminus C_k}(G) = min\left\{\frac{1}{|C_i|}\sum_{\widehat{G} \in C_i} d\left(G, \widehat{G}\right), \forall i \in \{1, 2, \cdots, p\} | i \neq k\right\}$ is the "external distance" of G, and is evaluated with the smallest average dissimilarities between G and the clusters to which G does not belong. The scenarios establish "soft" bounds of 2 (lower bound) and $\sigma|\mathscr{G}|$ (upper bound) for the cluster size. "Soft" means that though discouraged, the bounds still can be exceeded if necessary.

The overall clustering quality can be then evaluated using the average *SI* of all TDNs in the sequence $\mathscr{G}$, i.e.

$$SI_{\mathscr{G}}(\mathscr{C}(p)) = \frac{1}{|\mathscr{G}|}\sum_{G \in \mathscr{G}} SI_{\mathscr{G}}(\mathscr{C}(p), G) \tag{14}$$

The value of $SI_{\mathscr{G}}(\mathscr{C}(p))$ falls within the range of $[-1, 1]$. The higher is the $SI_{\mathscr{G}}(\mathscr{C}(p))$, the more likely are TDNs clustered properly such that TDNs are close within each cluster, but distant across different clusters. Note that, given a TDN sequence $\mathscr{G}$, $\mathscr{C}(p)$ is determined by the solution of IP formulation (5)–(12) with the input parameter p. Hence, $SI_{\mathscr{G}}(\mathscr{C}(p))$ is essentially a function of the number of clusters $p \in \{1, 2, \cdots, m\}$. The desired value for the parameter p, p^* should be the one that maximizes this function; in other words,

$$p^* = \underset{p \in \{1, 2, \dots, m\}}{argmax}\ SI_{\mathscr{G}}(\mathscr{C}(p)) \tag{15}$$

3.3. TDN visualization and disease clustering

A major challenge of TDN visualization is that complex networks may include too many nodes and edges to be displayed in an intuitive and orderly manner. In our TDN Visualization Module, we propose and solve a *minimum atomic clique partition problem* (MACPP) to address this challenge, as elaborated in the following.

Definition 1 (Atomic Clique). *Given a collection of networks, $\mathscr{G} = \{G_1, G_2, \cdots, G_m\}$, a subset $S \subseteq \cup_{i=1}^m V(G_i)$ is called an atomic clique if S is a clique in G_j, $\forall j \in M$, but $S \cap V(G_k) = \varnothing$, $\forall k \notin M$, where $M = \{i \in \{1, 2, \cdots, m\} | S \subseteq V(G_i)\}$.*

Definition 1 requires that in any network $G_i \in \mathscr{G}$, all nodes in an atomic clique S are either forming a clique or completely absent. For example, in Fig. 1, the atomic cliques across Window 1 and Window 2 are {1,2,3}, {4}, {5}. The clique {1,2,3} represents the initial comorbid diseases in Window 1, while {4} and {5} are newly developed diseases in Window 2. They are not interconnected directly, indicating that from diseases {1,2,3}, patients are very likely to develop either disease {4} or disease {5}, separately. Recall that in Section 2.2, we have shown that classical clique models could not necessarily capture this progression pattern. Instead, our proposed atomic clique model succeeds to address this challenge. Now, let us define MACPP that can decompose TDNs into a minimum set of atomic cliques.

Problem: Minimum atomic clique partition problem.

Input: A collection of networks, $\mathscr{G} = \{G_1, G_2, \cdots, G_m\}$.

Output: A collection of atomic cliques $\mathscr{K} = \{K_1, K_2, \dots, K_q\}$ such that.

- $K_i \cap K_j = \varnothing$, $\forall i, j \in \{1, 2, \dots, q\} | i \neq j$
- $\cup_{i=1}^q K_i = \cup_{j=1}^m V(G_j)$
- q is minimized.

The partition nature of the problem requires that the atomic cliques are mutually exclusive and in combination containing all nodes from the network collection. While the objective of minimizing the number of atomic cliques allows us to simplify the decomposition of the network collection as much as possible.

In this research, we developed an iterative algorithm—Algorithm 1—to find a feasible solution to MACPP. According to Definition 1, an atomic clique exists either in a single network or within an intersection of multiple networks. As a result, Algorithm 1 first finds a common node subset D across as many networks as possible through Lines 4–8. The initialization of D is performed at Line 4. Specifically, we assign the entire node set of the network G_k to D, where k is the smallest index of the networks remained in $\mathscr{G}$. The nested while loop from Line 9 to Line 16 then seeks an atomic clique partition on all D-induced subgraphs, $G_i[D]$, $\forall i \in M$. Once we narrow down to $G_i[D]$, we can iteratively detect and remove a maximum atomic clique across all $G_i[D]$ each time by leveraging an IP formulation until a partition is formed. After an atomic clique partition is found on $G_i[D]$, the algorithm excludes D and repeats previous steps until all $G_i \in \mathscr{G}$ are empty.

Algorithm 1: Atomic clique partition algorithm

```
   Input: A collection of networks $\mathcal{G} = \{G_1, G_2, \cdots, G_m\}$.
   Output: An atomic clique partition $\mathcal{K}$.
1  $\mathcal{K} \longleftarrow \emptyset$
2  while $\mathcal{G} \neq \emptyset$ do
3      $M \longleftarrow \emptyset$
4      $D \longleftarrow V(G_k)$, where $k = \min\{i \mid G_i \in \mathcal{G}\}$
5      for $G_i \in \mathcal{G}$ do
6          if $D \cap V(G_i) \neq \emptyset$ then
7              $D \longleftarrow D \cap V(G_i)$
8              $M \longleftarrow M \cup \{i\}$
9      while $D \neq \emptyset$ do
10         find a subset $K \subseteq D$ such that $K$ is a clique in $G_i[D], \forall i \in M$ and $|K|$ is maximized
             by solving formulation (16)–(18)
11         $\mathcal{K} \longleftarrow \mathcal{K} \cup K$
12         for $i \in M$ do
13             $G_i \longleftarrow G_i[V(G_i) \setminus K]$
14             if $V(G_i) = \emptyset$ then
15                 $\mathcal{G} \longleftarrow \mathcal{G} \setminus G_i$
16         $D \longleftarrow D \setminus K$
17 return $\mathcal{K}$
```

The IP formulation we used to find a maximum atomic clique across $G_i[D]$, $\forall\, i \in M$ is presented in (16)–(18). The binary variable $x_j = 1$ if and only if $j \in D$ is selected in the solution. Constraint (17) ensures that at most one of nodes $j, k \in D$ can be included in the solution if j and k are disconnected in any single network $G_i[D]$, so the solution will be guaranteed to be an atomic clique. While the objective function aims to maximize the cardinality of the atomic clique.

$$max \sum_{j \in D} x_j \tag{16}$$

$$x_j + x_k \leq 1 \quad \forall \{j,k\} \in Q = \{\{j,k\} \subseteq D \,|\, \exists i \in M \text{ such that } \{j,k\} \notin E(G_i)\,\} \tag{17}$$

$$x_j \in \{0,1\} \quad \forall j \in D \tag{18}$$

Algorithm 1 is essentially a *greedy* algorithm because the IP formulation tries to find a maximum atomic clique in each iteration of the nested loop through Lines 9–16. Clearly, the algorithm returns a feasible solution to MACPP because each K found in one iteration is isolated from that found in other iterations, and $\mathscr{K}$ exhausts all nodes in $\mathscr{G}$.

4. Case studies

To assess the effectiveness of our proposed system, we applied it to two case studies on analyzing and visualizing the comorbidity progressions during hospitalizations for C. Diff and stroke patients, respectively. In the case studies, our system was implemented using Python 3.7, and the IP formulations involved were solved using a state-of-the-art optimization solver—Gurobi 8.1.1 [74].

4.1. Data cohorts and data preparation

We integrated Cerner Health Facts® EHR data warehouse as the data source into our system. Health Facts® contains clinical data extracted directly from the U.S. hospitals that operate on Cerner EHR systems. Cerner Corporation collects and integrates the data through its established operations in compliance with the Health Insurance Portability and Accountability Act (HIPAA) laws. Because the data has been completely de-identified according to HIPAA regulations, the Institutional Review Boards (IRB) at Oklahoma State University exempted the study from review.

C. Diff is a bacterial infection that are mostly hospital-acquired among senior patients [75], while stroke is one of the leading chronic

Table 1
The statistics of encounters and TDNs in each window (TDNs are ego networks).

Window	C. Diff - Senior Female Cohort				Stroke - Senior Female Cohort			
	Enct #	Diag #	|V|	|E|	Enct #	Diag #	|V|	|E|
1	158,408	1,088,614	6	15	13,070	57,703	105	602
2	173,611	1,400,515	7	21	11,641	55,473	110	727
3	200,907	1,625,593	9	34	11,721	55,131	108	723
4	196,913	1738,845	14	90	11,201	54,259	113	776
5	242,541	2,056,285	12	65	10,251	50,313	113	821
6	173,887	1,612,831	26	321	8697	42,569	114	843
7	164,309	1,533,572	28	372	7392	36,497	117	889
8	128,370	1,246,318	32	482	6348	31,215	119	936
9	116,790	1147,284	34	547	5259	26,660	120	1072
10	95,243	962,605	39	692	4680	23,286	122	1101
11	86,439	883,730	35	584	3850	19,480	120	1147
12	72,727	750,208	42	811	3507	17,752	123	1250
13	66,558	703,248	44	931	2868	14,422	118	1308
14	55,194	581,826	50	1195	2660	13,407	120	1371
15	47,954	519,614	54	1401	2061	10,627	121	1551
16	39,010	423,053	54	1360	1849	9288	117	1516
17	33,844	380,795	61	1738	1501	7780	124	1660
18	30,021	332,668	62	1754	1502	7637	125	1655
19	25,702	295,288	62	1806	1157	6028	131	1831
20	23,662	265,881	61	1739	1155	6013	133	1864
21	20,171	233,745	68	2155	924	4816	117	1682
22	18,604	213,280	66	2039	936	4894	125	1859
23	15,760	187,325	70	2275	738	3957	133	1893
24	15,514	177,554	74	2484	839	4361	122	1771
25	13,080	155,944	75	2595	586	3076	123	1707
26	13,832	160,786	67	2139	869	4542	121	1770
Total	2,229,051	20,677,407	–	–	117,262	571,186	–	–

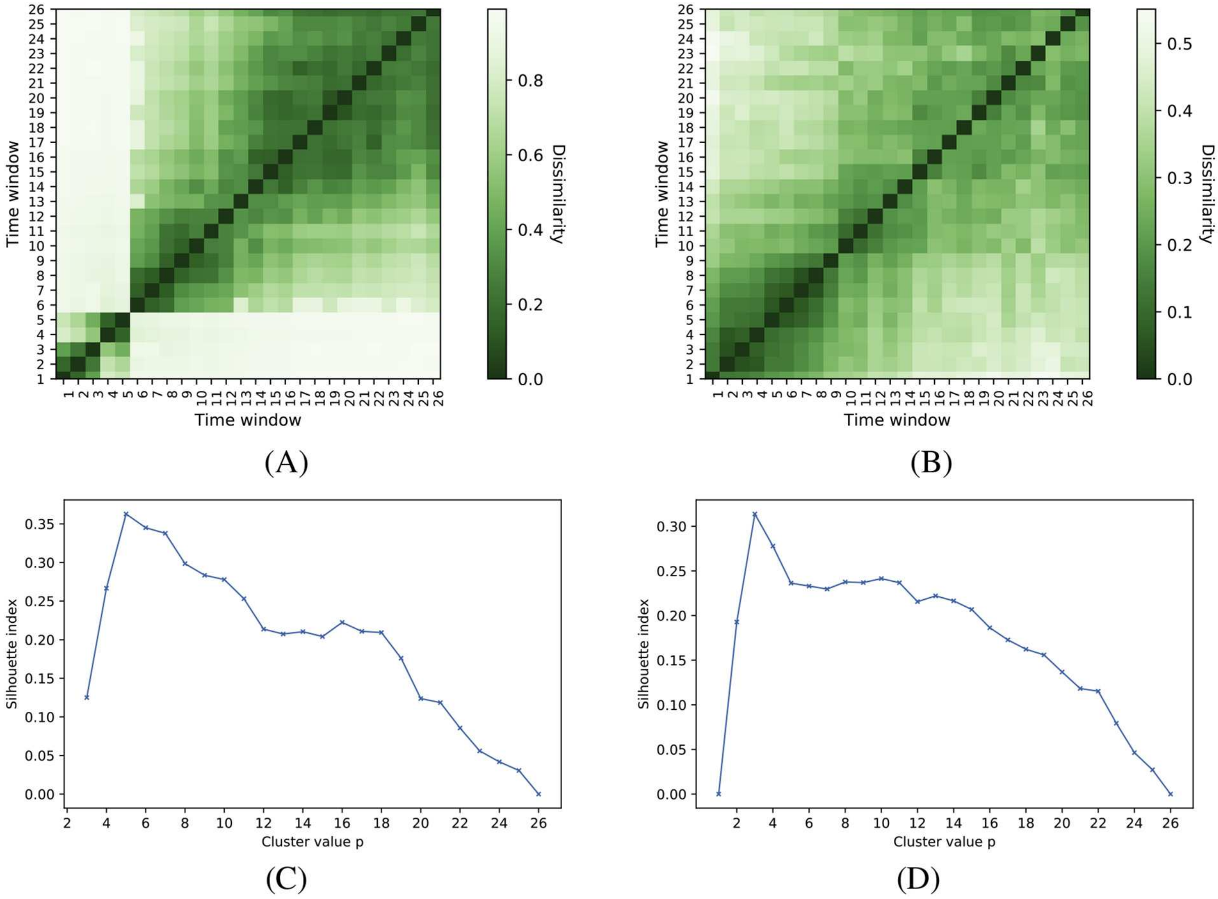


Fig. 5. The heat maps of dissimilarities among TDNs and the SI charts for different values of p. (A) and (C) are diagrams for C. Diff; (B) and (D) are diagrams for stroke.

conditions for death/disability in the U.S. [76]. Our C. Diff and stroke study cohorts were extracted from Health Facts® using International Classification of Diseases 9th/10th Revision (ICD-9/10) codes (the ICD-9/10 codes are listed in the Supplementary Material). The cohorts included hospitalized encounters of female patients aged 65 or older with the onset of C. Diff/stroke between November 1999 and August 2017. Patient age, length of stay (LOS), and all diagnoses associated with the encounters were exported as well.

Our data preprocessing mainly dealt with outlying LOS, erroneous diagnoses, and diagnosis combination. In order to exclude extreme outliers in LOS, we restricted analysis to the encounters of LOS within the range of 24 h to 14 days, which is a common range for inpatient hospital stays. We noticed that the data included some infeasible diagnoses, such as birth/labor-related diagnoses and male conditions. Encounters with such erroneous diagnoses were excluded from the study cohorts. Furthermore, the ICD-9/10 codes used in Health Facts® can be overly specific to express disease states in the usual sense. We used the Clinical Classifications Software (CCS) [77] to aggregate ICD-9/10 codes into relatively high-level disease states. For example, CCS combines malignant neoplasms at different locations of esophagus together as the "cancer of esophagus". Our data extraction and preprocessing eventually resulted in two large datasets containing hundreds of thousands or millions of encounters and diagnosis records as shown in Table 1 (under the "Enct #" and "Diag #" columns).

4.2. TDN construction

In Health Facts®, diagnoses were recorded in encounters, but lacking specific timestamps about at what time during the encounter a condition was diagnosed. In other words, given time points $t_1 < t_2 < \ldots < t_m$ within an encounter, we cannot tell what diagnoses occurred exactly during a time interval $[t_i, t_{i+1}]$. Therefore, we defined the windows based on LOS as Warner et al. did in their studies on hospital-acquired complications [31,78]. In particular, Window i includes all encounters with LOS $\in [l + (i-1)\epsilon, l + i\epsilon)$, where l is the smallest LOS included for analysis (24 h in our case studies in light of the data preparation). The rational is that when a large sample is included in a window, the statistical results based on the sample can be considered as the expected values of the attributes of a general population in the window. Then, the changes newly happened to Window $i + 1$ from Window i can be well representative of the events occurring within the interval $[l + i\epsilon, l + (i + 1)\epsilon)$ for the population. In our case studies, we specified $\epsilon = 12$ hours, which resulted in 26 windows in total, i.e. $m = 26$.

Then, we built networks over the 26 windows with *SCI* threshold $\theta = 0.05$. Since our interest was concentrated on the progression of C. Diff/stroke and its strongly coexisting diseases, we only considered the ego networks of C. Diff/stroke as the TDNs for analysis and visualization henceforth. The TDNs constructed based on our C. Diff and stroke cohorts are visualized in Figs. A1 and A2 respectively in the Appendix, while the data (including the edge lists and the mapping between nodes and diseases) of the TDNs are provided in the Supplementary Material. The orders and sizes of the TDNs are listed in Table 1.

4.3. Temporal clustering

The dissimilarity between each pair of the TDNs of the C. Diff cohort is calculated and plotted as a heat map shown in Fig. 5 (A). From the heat map, we may roughly observe that (i) there exist a few dark blocks, which correspond to clusters of windows that may imply progression phases; and (ii) the phases tend to include more windows over time, indicating that comorbidity evolves more rapidly at earlier phases compared with later phases. We now present the CPMP results on this TDN sequence to demonstrate CPMP's effectiveness to capture the observations algorithmically. In order to solve the CPMP on this TDN sequence, we firstly used the *SI* method described in Section 3.2.3 to determine a proper p^* for the TDNs. During the calculation of *SI*, we let both the parameters τ and σ be 0.5, meaning we do not intent to cluster a window with its adjacent window(s) if the dissimilarity is no less than 0.5, and we discourage a cluster that includes half or more of all windows since it might be overly broad. The result in Fig. 5 (C) shows that $p^* = 5$, indicating that the entire window sequence should be clustered into five phases. Given $p = p^* = 5$, the CPMP solution is: Phase 1 includes Windows 1–3, Phase 2 contains Windows 4–5, Phase 3 consists of Windows 6–11, Phase 4 is comprised of Windows 12–20, and Phase 5 includes Windows 21–26. The corresponding days of the phases are shown in Table 2. The results are aligned with the observations we can inspect from Fig. 5 (A), demonstrating that the proposed consecutive p-median model is capable to identify the progression patterns algorithmically.

Table 2
Phases and corresponding windows and days.

Cohorts	Time Unit	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
C. Diff	Window	1–3	4–5	6–11	12–20	21–26
	Day	2–3	3–4	4–7	7–11	12–14
Stroke	Window	1–8	9–15	16–26	–	–
	Day	2–5	6–9	9–14	–	–

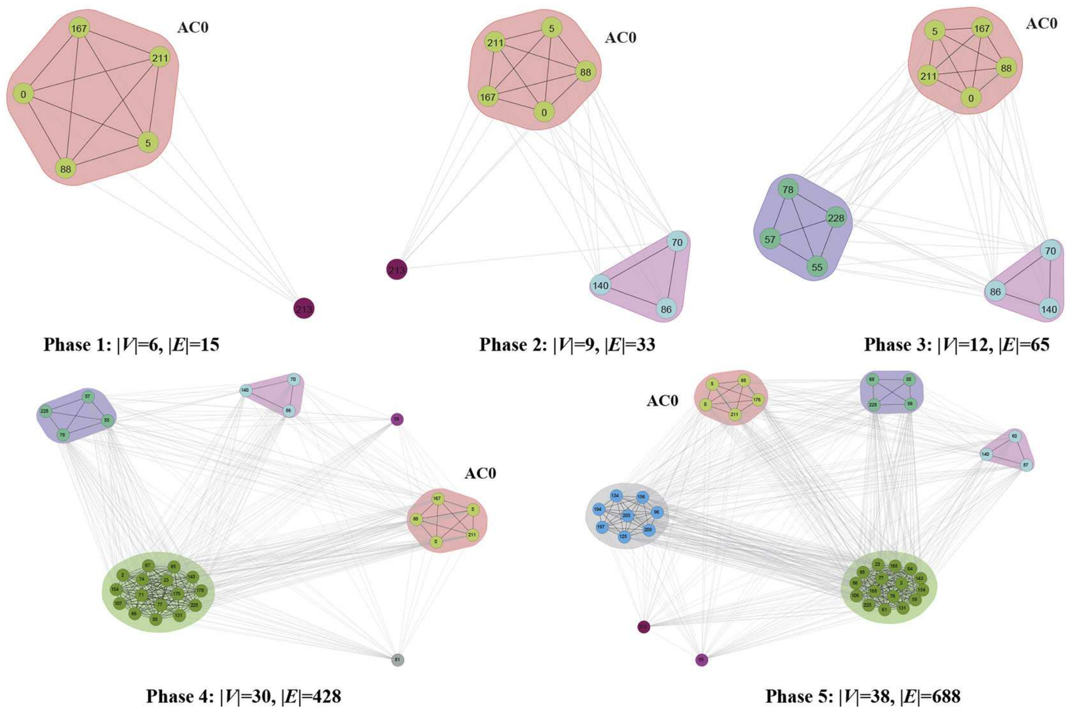
The stroke results are presented in Fig. 5 (B) and Fig. 5 (D). Fig. 5 (D) shows that $p^* = 3$, implying that hospitalized stroke patients may experience three phases: Phase 1 includes Windows 1–8, Phase 2 contains Windows 9–15, and Phase 3 consists of Windows 16–26, as shown in Table 2. Similar to the C. Diff results, the phases outlined by the proposed consecutive p-median model are also in line with what we can observe from Fig. 5 (B).

4.4. Visualization of TDNs in phases

By visualizing TDNs on the identified phases, we can reduce the complexity of the entire TDN sequence over time. However, the complexity inside a single TDN remains because some TDNs can include many nodes and edges. For example, the C. Diff TDN at Phase 5 includes 688 edges incident to 38 nodes. Visualizing such dense networks in a user-friendly format will significantly facilitate subsequent inspection and analysis. To that end, we firstly found an atomic clique partition using Algorithm 1. Then, for the TDN at every phase, we plotted each atomic clique together in a compact, shaded space. In addition, to keep consistency, each atomic clique was rendered in the same color across all phases.

The C. Diff comorbidity progression is visualized in Fig. 6, from which we can observe that acute renal failure (node 5), fluid and electrolyte disorder (node 88), other gastrointestinal disorders (node 167), and septicemia (node 211) along with C. Diff (node 0) form an atomic clique that occurs persistently across all phases (marked as AC0 in Fig. 6). It implies that these diseases are highly coexisting with C. Diff throughout the entire time frame. Many clinical studies [79,80] have reported similar findings that these diseases are highly associated with C. Diff, thus validating our VA results. Another interesting progression pattern we can inspect from Fig. 6 is that instead of occurring independently, the comorbid diseases appeared at later phases tend to form atomic cliques as well. In other words, the onset of one of these diseases may indicate one or more other conditions in the same atomic clique. For example, urinary tract infections (UTI, node 228) appears in Phases 3–5, which echoes a previous study finding that UTI is associated with prolonged hospitalization of C. Diff patients [31]. Furthermore, our approach discovers that UTI occurs in an atomic clique that also includes cardiac dysrhythmias (node 55), chronic kidney disease (node 57), and disorders of lipid metabolism (node 78). It suggests that doctors should pay attention to not only UTI but also these UTI-associated diseases to prevent prolonged hospitalization.

The stroke comorbidity progression is visualized in Fig. 7, which shows that a few diseases start to be highly coexisting with stroke after



Phase 1: |V|=6, |E|=15

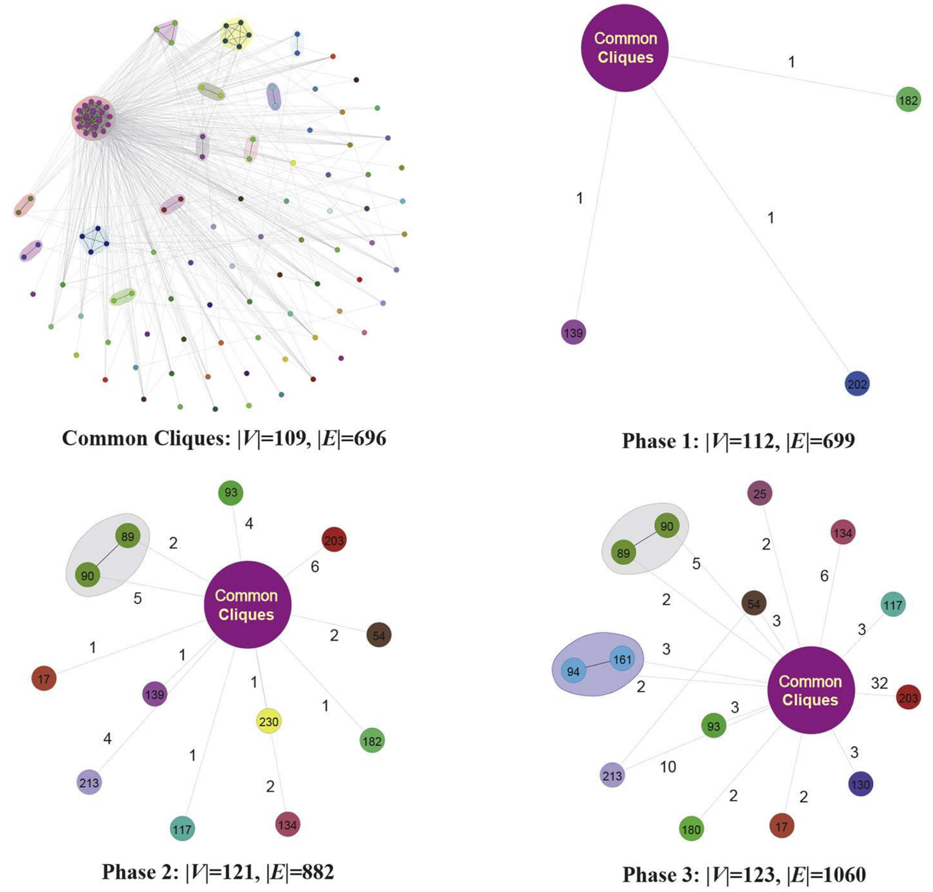
Phase 2: |V|=9, |E|=33

Phase 3: |V|=12, |E|=65

Phase 4: |V|=30, |E|=428

Phase 5: |V|=38, |E|=688

Fig. 6. TDNs constructed on the phases of the C.Diff cohort.



Common Cliques: |V|=109, |E|=696

Phase 1: |V|=112, |E|=699

Phase 2: |V|=121, |E|=882

Phase 3: |V|=123, |E|=1060

Fig. 7. TDNs constructed on the phases of the stroke cohort. There exists a set of *common cliques* throughout all the phases. The common cliques are visualized in detail at the upper left part of the figure and simplified as a large node in the TDNs across the phases. The edge weight in the TDNs indicates how many nodes inside the set of common cliques are connected to a node outside the common cliques.

Phase 1. It implies that these disease states are highly associated with prolonged hospitalization more than one week of stroke patients. This association of some of the diseases, such as mental health disorders (node 130) and shock (node 213), are also supported by other clinical studies [81,82]. Furthermore, other two risk factors for prolonged hospitalizations—fracture of lower limb (node 89) and fracture of hip (node 90)—occur in the same atomic clique. It indicates that these two conditions are very likely to occur together, which may be resulted from post-stroke fall [83].

5. Discussion

Our proposed VA system for comorbidity progression has significant implications in both the technology advance and healthcare application, as discussed in the following.

Technical Contributions: The highlight of this research from technical perspective is that we look into the temporal and disease clustering of TDNs for the *first* time. In this effort, two new problems and associated algorithms, rooted from methodology for the single network, were extended to network sequences (i.e. TDNs) to address the challenges in implementing the temporal and disease clustering of TDNs:

- *The consecutive p-median problem* was extended from the classical *p*-median problem, by requiring each cluster to only include consecutive objects (TDNs in our case) to model temporal clustering. An IP formulation was developed to solve the problem, and the classical Silhouette Index was modified to determine a suitable value for the parameter *p*.
- *The minimum atomic clique partition problem* was extended from the minimum clique partition problem for a single network to clustering diseases across a sequence of TDNs. A greedy heuristic algorithm was developed to find a feasible solution for the problem.

Application in Healthcare: Supported by the temporal clustering module, our proposed system can automatically detect the comorbidity progression phases. Because the disease states and coexistence relationships are highly similar within each phase while remarkably distinct across different phases, the end of a phase can indicate a beginning time point of significant progression changes. Furthermore, through our visualization module, we are able to show the comorbidity coexistence relationships and progression patterns visually and concisely. It can help doctors understand when and what diseases are most likely to be comorbid with the index disease, and plan prevention and treatments in advance. For example, in our stroke case study, the VA results in Fig. 7 show that fractures are associated with prolonged hospitalization more than one week. Furthermore, the fractures often include both lower limb and hip fractures. By being aware of this fact, hospitals and doctors can prepare proper care resources to prevent/handle both types of fractures during patients' hospitalizations. In addition, the TDNs can be used to compare different subgroups of patients, such as matched case-control cohorts based on a certain treatment [84] to evaluate the treatment's efficacy or different gender groups [69] to reveal progression disparities between genders.

Limitations: This research mainly has two limitations. First, in literature there are many approaches for the network dissimilarity measurement. The choices of the method may influence the temporal clustering results. However, a systematic review and comparison of all the methods on our problem is beyond the scope of this study. Second, our temporal and disease clustering approaches only work on undirected, unweighted networks. TDNs can be more sophisticated by carrying node attributes (like disease frequency), edge weight (like *SCI* value), and edge direction (like presence order). Performing temporal and disease clustering on such complex TDNs requires corresponding dissimilarity measurement methods and graphical cluster models. Nevertheless, many of the approaches are either still absent or requiring much effort for suitable adaptions. As a result, we leave these challenges for future work.

6. Conclusion

Comorbidity is a prominent challenge in healthcare practice and research. In this work, we modeled comorbidity progression as a sequence of TDNs, and designed a VA system, which integrates novel temporal and disease clustering technologies to mine and visualize progression patterns from the TDN sequence. Two case studies of applying the system to C. Diff and stroke demonstrate the effectiveness of the system. Based on the discussion in Section 5, we summarize two directions for our future work—*healthcare application* and *technical improvement*. From the healthcare application perspective, we plan to apply the proposed system to more diseases to mine useful insights for healthcare practice. We will also incorporate more biomarkers besides comorbidity during the applications to reveal more progression patterns. In order to improve the proposed technologies, we plan to extend our temporal and disease clustering approaches to more sophisticated TDNs that can carry node attributes and edge weights. In this study, we proposed a heuristic algorithm for MACPP, which does not necessarily find a minimized solution. Hence, we are interested in developing exact algorithms, such as IP formulations, which are able to provide optimal solutions for MACPP in our future work.

Declaration of Competing Interest

None.

Acknowledgments

The authors gratefully acknowledge Cerner Corporation for sharing Health Facts® EHR database to support this research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Cerner Corporation. We adapted certain Python code shared by Alexander Veremyev to plot Figs. 6 and 7; his help with the code is highly appreciated.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dss.2021.113583.

Appendix A. TDNs Established on the Initial 26 Time Windows

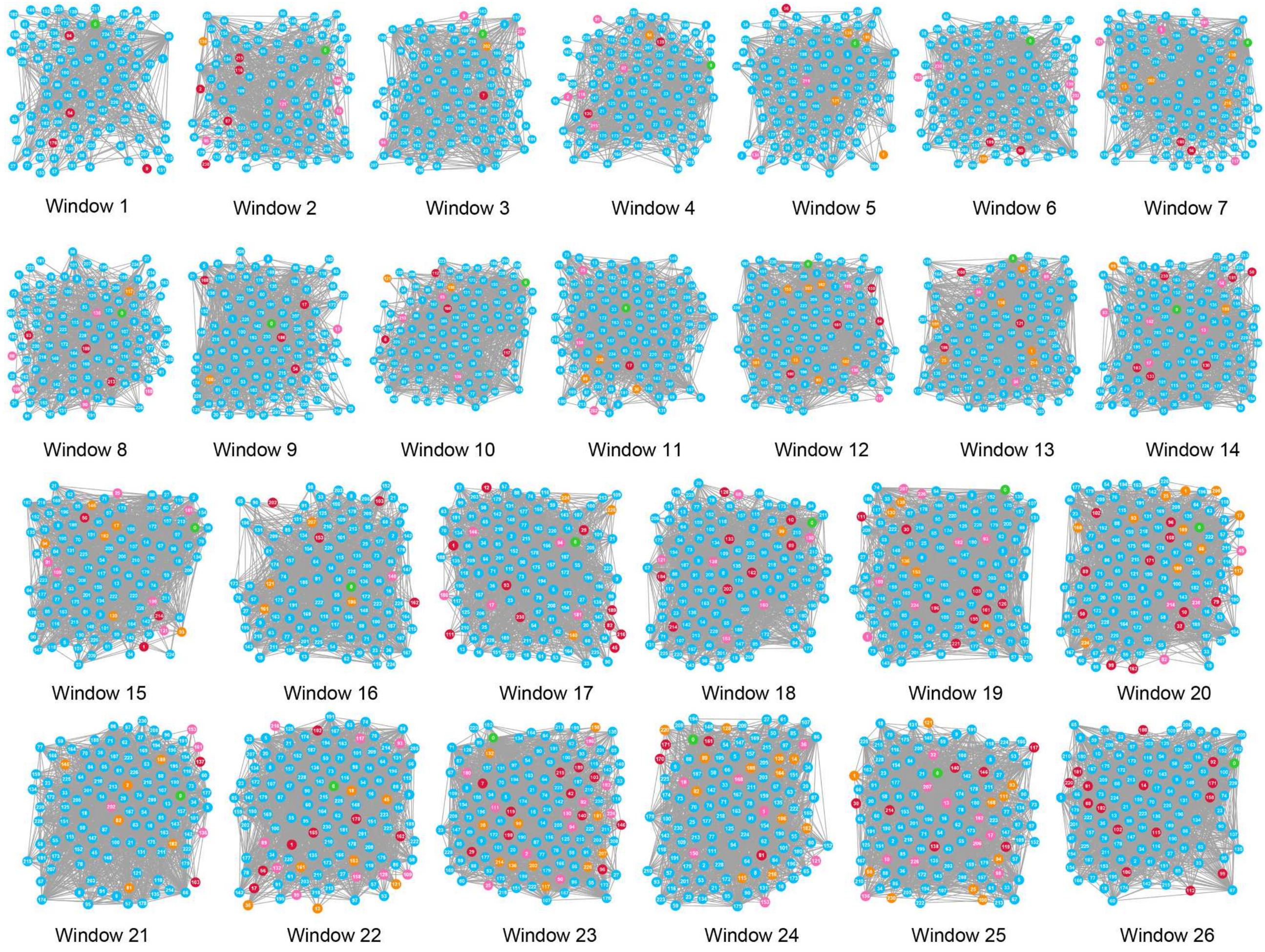


Fig. A1. TDNs (ego networks) constructed on the 26 windows for the C. Diff cohort (senior female patients). The node color is used to indicate the existence pattern of a node in adjacent windows: C.Diff node is in green color through all windows. Given a window, a blue node indicates that the node also appears in both adjacent windows or the unique adjacent window. A red node means that the node does not appear in any adjacent window(s). Pink means that the node also appears in the next window but not in the previous window, while orange indicates that the node also occurs in the previous window but not in the next window. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

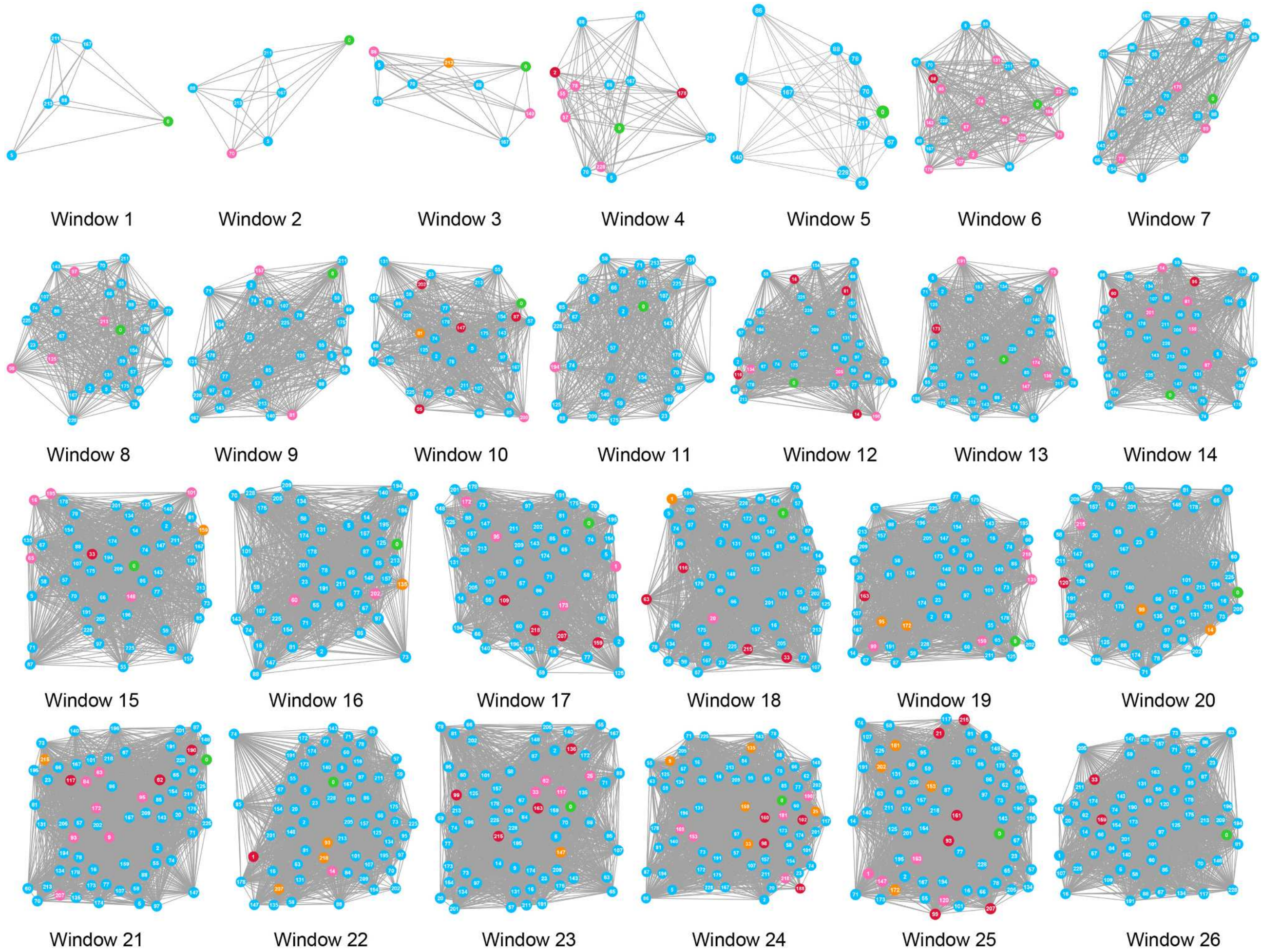


Fig. A2. TDNs (ego networks) constructed on the 26 windows for the stroke cohort (senior female patients).

References

[1] I.R. Bardhan, M.F. Thouin, Health information technology and its impact on the quality and cost of healthcare delivery, Decis. Support. Syst. 55 (2013) 438–449.

[2] T.R. Huerta, M.A. Thompson, E.W. Ford, W.F. Ford, Electronic health record implementation and hospitals' total factor productivity, Decis. Support. Syst. 55 (2013) 450–458.

[3] T.J. Bright, A. Wong, R. Dhurjati, E. Bristow, L. Bastian, R.R. Coeytaux, G. Samsa, V. Hasselblad, J.W. Williams, M.D. Musty, L. Wing, A. Kendrick, G. Sanders, D. Lobach, Effect of clinical decision-support systems: a systematic review, Ann. Intern. Med. 157 (2012) 29–43.

[4] A. Gupta, R. Sharda, Improving the science of healthcare delivery and informatics using modeling approaches, Decis. Support. Syst. 2 (2013) 423–427.

[5] R.W. Grout, E.R. Cheng, A.E. Carroll, N.S. Bauer, S.M. Downs, A six-year repeated evaluation of computerized clinical decision support system user acceptability, Int. J. Med. Inform. 112 (2018) 74–81.

[6] J. Henry, Y. Pylypchuk, T. Searcy, V. Patel, Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015, ONC Data Brief 35 (2016) 1–9.

[7] T.T. Moores, Towards an integrated model of it acceptance in healthcare, Decis. Support. Syst. 53 (2012) 507–516.

[8] K. Kawamoto, C.A. Houlihan, E.A. Balas, D.F. Lobach, Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success, BMJ 330 (2005) 765.

[9] M.P. Johnson, K. Zheng, R. Padman, Modeling the longitudinality of user acceptance of technology with an evidence-adaptive clinical decision support system, Decis. Support. Syst. 57 (2014) 444–453.

[10] R.G. Fichman, R. Kohli, R. Krishnan, Editorial overview—the role of information systems in healthcare: current research and future trends, Inf. Syst. Res. 22 (2011) 419–428.

[11] Z.Y. Zhuang, C.L. Wilkin, A. Ceglowski, A framework for an intelligent decision support system: a case in pathology test ordering, Decis. Support. Syst. 55 (2013) 476–487.

[12] B. Yet, K. Bastani, H. Raharjo, S. Lifvergren, W. Marsh, B. Bergman, Decision support system for warfarin therapy management using bayesian networks, Decis. Support. Syst. 55 (2013) 488–498.

[13] J. Barjis, G. Kolfschoten, J. Maritz, A sustainable and affordable support system for rural healthcare delivery, Decis. Support. Syst. 56 (2013) 223–233.

[14] Y. Li, A. Vo, M. Randhawa, G. Fick, Designing utilization-based spatial healthcare accessibility decision support systems: a case of a regional health plan, Decis. Support. Syst. 99 (2017) 51–63.

[15] S. Piri, D. Delen, T. Liu, H.M. Zolbanin, A data analytics approach to building a clinical decision support system for diabetic retinopathy: developing and deploying a model ensemble, Decis. Support. Syst. 101 (2017) 12–27.

[16] K. Topuz, F.D. Zengul, A. Dag, A. Almehmi, M.B. Yildirim, Predicting graft survival among kidney transplant recipients: a bayesian decision support model, Decis. Support. Syst. 106 (2018) 97–109.

[17] G. Van Valkenhoef, T. Tervonen, T. Zwinkels, B. De Brock, H. Hillege, Addis: a decision support system for evidence-based medicine, Decis. Support. Syst. 55 (2013) 459–475.

[18] J.J. Caban, D. Gotz, Visual analytics in healthcare – opportunities and research challenges, J. Am. Med. Inform. Assoc. 22 (2015) 260–262.

[19] A.F. Simpao, L.M. Ahumada, J.A. Gálvez, M.A. Rehman, A review of analytics and clinical informatics in health care, J. Med. Syst. 38 (2014) 45.

[20] K.K. Mane, C. Bizon, C. Schmitt, P. Owen, B. Burchett, R. Pietrobon, K. Gersing, Visualdecisionlinc: a visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry, J. Biomed. Inform. 45 (2012) 101–106.

[21] A.F. Simpao, L. Ahumada, M. Rehman, Big data and visual analytics in anaesthesia and health care, Br. J. Anaesth. 115 (2015) 350–356.

[22] A.F. Simpao, L.M. Ahumada, B.R. Desai, C.P. Bonafide, J.A. Gálvez, M.A. Rehman, A.F. Jawad, K.L. Palma, E.D. Shelov, Optimization of drug–drug interaction alert rules in a pediatric hospital's electronic health record system using a visual analytics dashboard, J. Am. Med. Inform. Assoc. 22 (2015) 361–369.

[23] A. Rind, T.D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, B. Shneiderman, Interactive information visualization to explore and query

electronic health records, Found. Trend. Human-Computer Inter. 5 (2013) 207–298.

[24] O. Nelson, B. Sturgis, K. Gilbert, E. Henry, K. Clegg, J.M. Tan, J.O. Wasey, A. F. Simpao, J.A. Gálvez, A visual analytics dashboard to summarize serial anesthesia records in pediatric radiation treatment, Appl. Clin. Inform. 10 (2019) 563.

[25] M. Nadj, A. Maedche, C. Schieder, The effect of interactive analytical dashboard features on situation awareness and task performance, Decis. Support. Syst. 135 (2020) 113322.

[26] B. Kamsu-Foguem, G. Tchuenté-Foguem, L. Allart, Y. Zennir, C. Vilhelm, H. Mehdaoui, D. Zitouni, H. Hubert, M. Lemdani, P. Ravaux, User-centered visual analysis using a hybrid reasoning architecture for intensive care units, Decis. Support. Syst. 54 (2012) 496–509.

[27] A.F. Simpao, L.M. Ahumada, B.L. Martinez, A.M. Cardenas, T.A. Metjian, K. V. Sullivan, J.A. Gálvez, B.R. Desai, M.A. Rehman, J.S. Gerber, Design and implementation of a visual analytics electronic antibiogram within an electronic health record system at a tertiary pediatric hospital, Appl. Clin. Inform. 9 (2018) 37.

[28] A. Sorbello, A. Ripple, J. Tonning, M. Munoz, R. Hasan, T. Ly, H. Francis, O. Bodenreider, Harnessing scientific literature reports for pharmacovigilance: prototype software analytical tool development and usability testing, Appl. Clin. Inform. 8 (2017) 291.

[29] M. Karami, R. Safdari, From information management to information visualization: development of radiology dashboards, Appl. Clin. Inform. 7 (2016) 308.

[30] D.W. Tscholl, L. Handschin, P. Neubauer, M. Weiss, B. Seifert, D.R. Spahn, C. B. Noethiger, Using an animated patient avatar to improve perception of vital sign information by anaesthesia professionals, Br. J. Anaesth. 121 (2018) 662–671.

[31] J.L. Warner, A. Zollanvari, Q. Ding, P. Zhang, G.M. Snyder, G. Alterovitz, Temporal phenome analysis of a large electronic health record cohort enables identification of hospital-acquired complications, J. Am. Med. Inform. Assoc. 22 (2013) e281–e287.

[32] C.A. Hidalgo, N. Blumm, A.-L. Barabási, N.A. Christakis, A dynamic network approach for the study of human phenotypes, PLoS Comput. Biol. 5 (2009) 1–11.

[33] T. Wang, R.G. Qiu, M. Yu, R. Zhang, Directed disease networks to facilitate multiple-disease risk assessment modeling, Decis. Support. Syst. 129 (2020) 113171.

[34] M.E. Hossain, S. Uddin, A. Khan, M.A. Moni, A framework to understand the progression of cardiovascular disease for type 2 diabetes mellitus patients using a network approach, Int. J. Environ. Res. Public Health 17 (2020) 596.

[35] M. Krishnamurthy, P. Marcinek, K.M. Malik, M. Afzal, Representing social network patient data as evidence-based knowledge to support decision making in disease progression for comorbidities, IEEE Access 6 (2018) 12951–12965.

[36] A.R. Feinstein, The pre-therapeutic classification of co-morbidity in chronic disease, J. Chronic Dis. 23 (1970) 455–468.

[37] M.J. Divo, C.H. Martinez, D.M. Mannino, Ageing and the epidemiology of multimorbidity, Eur. Respir. J. 44 (2014) 1055–1068.

[38] R. Gijsen, N. Hoeymans, F.G. Schellevis, D. Ruwaard, W.A. Satariano, G.A.M. van den Bos, Causes and consequences of comorbidity: a review, J. Clin. Epidemiol. 54 (2001) 661–674.

[39] V. De Groot, H. Beckerman, G.J. Lankhorst, L.M. Bouter, How to measure comorbidity: a critical review of available methods, J. Clin. Epidemiol. 56 (2003) 221–229.

[40] E. Capobianco, P. Lio, Comorbidity: a multidimensional approach, Trends Mol. Med. 19 (2013) 515–521.

[41] H.M. Zolbanin, D. Delen, A.H. Zadeh, Predicting overall survivability in comorbidity of cancers: a data mining approach, Decis. Support. Syst. 74 (2015) 150–161.

[42] A. Cramer, L.J. Waldorp, H.L.J. van der Maas, D. Borsboom, Comorbidity: a network perspective, Behav. Brain Sci. 33 (2010) 137–193.

[43] A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nat. Rev. Genet. 12 (2011) 56–68.

[44] J.C. Brunson, R.C. Laubenbacher, Applications of network analysis to routinely collected health care data: a systematic review, J. Am. Med. Inform. Assoc. 25 (2018) 210–221.

[45] M.J. Divo, C. Casanova, J.M. Marin, V.M. Pinto-Plata, J.P. de Torres, J.J. Zulueta, C. Cabrera, J. Zagaceta, P. Sanchez-Salcedo, J. Berto, R.B. Davila, A.B. Alcaide, C. Cote, B.R. Celli, Copd comorbidities network, Eur. Respir. J. 46 (2015) 640–650.

[46] J.L. Warner, J.C. Denny, D.A. Kreda, G. Alterovitz, Seeing the forest through the trees:uncovering phenomic complexity through interactive network visualization, J. Am. Med. Inform. Assoc. 22 (2015) 324–329.

[47] A.B. Jensen, P.L. Moseley, T.I. Oprea, S.G. Ellesøe, R. Eriksson, H. Schmock, P. B. Jensen, L.J. Jensen, S. Brunak, Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients, Nat. Commun. 5 (2014).

[49] L. Chen, N. Blumm, N. Christakis, A. Barabási, T. Deisboeck, Cancer metastasis networks and the prediction of progression patterns, Br. J. Cancer 101 (2009) 749–758.

[50] A. Chmiel, P. Klimek, S. Thurner, Spreading of diseases through comorbidity networks across life and gender, New J. Phys. 16 (2014) 115013.

[51] A.-L. Barabási, Network medicine — from obesity to the "diseasome", N. Engl. J. Med. 357 (2007) 404–407.

[52] Y. Nam, D.-g. Lee, S. Bang, J.H. Kim, J.-H. Kim, H. Shin, The translational network for metabolic disease – from protein interaction to disease co-occurrence, BMC Bioinform. 20 (2019).

[53] B. Davazdahemami, D. Delen, A chronological pharmacovigilance network analytics approach for predicting adverse drug events, J. Am. Med. Inform. Assoc. 25 (2018) 1311–1321.

[54] D.B. West, Introduction to Graph Theory, Volume 2, Prentice hall Upper Saddle River, NJ, 1996.

[55] E. Jeong, K. Ko, S. Oh, H.W. Han, Network-based analysis of diagnosis progression patterns using claims data, Sci. Rep. 7 (2017) 1–12.

[56] Y. Chen, X. Zhang, G. Zhang, R. Xu, Comparative analysis of a novel disease phenotype network based on clinical manifestations, J. Biomed. Inform. 53 (2015) 113–120.

[56] M. Guo, Y. Yu, T. Wen, X. Zhang, B. Liu, J. Zhang, R. Zhang, Y. Zhang, X. Zhou, Analysis of disease comorbidity patterns in a large-scale China population, BMC Med. Genet. 12 (2019) 177.

[57] Z. Shu, W. Liu, H. Wu, M. Xiao, D. Wu, T. Cao, M. Ren, J. Tao, C. Zhang, T. He, X. Li, R. Zhang, X. Zhou, Symptom-based network classification identifies distinct clinical subgroups of liver diseases with common molecular pathways, Comput. Methods Prog. Biomed. 174 (2019) 41–50.

[58] I. Schäfer, H. Kaduszkiewicz, H. Wagner, G. Schön, M. Scherer, H.V.D. Bussche, Reducing complexity: a visualisation of multimorbidity by combining disease clusters and triads, BMC Public Health 14 (2014).

[59] E. Sokolova, A.M. Oerlemans, N.N. Rommelse, P. Groot, C.A. Hartman, J. C. Glennon, T. Claassen, T. Heskes, J.K. Buitelaar, A causal and mediation analysis of the comorbidity between attention deficit hyperactivity disorder (adhd) and autism spectrum disorder (asd), J. Autism Dev. Disord. 47 (2017) 1595–1604.

[60] M. Peleg, N. Asbeh, T. Kuflik, M. Schertz, Onto-clust—a methodology for combining clustering analysis and ontological methods for identifying groups of comorbidities for developmental disorders, J. Biomed. Inform. 42 (2009) 165–175.

[61] M. Tantardini, F. Ieva, L. Tajoli, C. Piccardi, Comparing methods for comparing networks, Sci. Rep. 9 (2019).

[62] P. Wills, F.G. Meyer, Metrics for graph comparison: a practitioner's guide, PLoS One 15 (2020) 1–54.

[63] M. Berlingerio, D. Koutra, T. Eliassi-Rad, C. Faloutsos, Network similarity via multiple social theories, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013, pp. 1439–1440.

[64] N. Pržulj, D.G. Corneil, I. Jurisica, Modeling interactome: scale-free or geometric? Bioinformatics 20 (2004) 3508–3515.

[65] S.V.N. Vishwanathan, N.N. Schraudolph, R. Kondor, K.M. Borgwardt, Graph kernels, J. Machine Learning Res. 11 (2010) 1201–1242.

[66] S. Ghosh, N. Das, T. Gonçalves, P. Quaresma, M. Kundu, The journey of graph kernels through two decades, Computer Sci. Rev. 27 (2018) 88–111.

[67] M.M. Martel, C.A. Levinson, J.K. Langer, J.T. Nigg, A network analysis of developmental change in adhd symptom structure from preschool to adulthood, Clin. Psychol. Sci. 4 (2016) 988–1001.

[68] E. McElroy, P. Fearon, J. Belsky, P. Fonagy, P. Patalay, Networks of depression and anxiety symptoms across development, J. Am. Acad. Child Adolesc. Psychiatry 57 (2018) 964–973.

[69] P. Kalgotra, R. Sharda, J.M. Croff, Examining health disparities by gender: a multimorbidity network analysis of electronic medical record, Int. J. Med. Inform. 108 (2017) 22–28.

[70] B. Fotouhi, N. Momeni, M.A. Riolo, D.L. Buckeridge, Statistical methods for constructing disease comorbidity networks from longitudinal inpatient data, Appl. Network Sci. 3 (2018) 46.

[71] T.D. Klastorin, The *p*-median problem for cluster analysis: a comparative test using the mixture model approach, Manag. Sci. 31 (1985) 84–95.

[72] H. Köhn, D. Steinley, M.J. Brusco, The *p*-median model as a tool for clustering psychological data, Psychol. Methods 15 (2010) 87.

[73] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[74] Gurobi Optimization L.L.C., Gurobi Optimizer Reference Manual, URL: http://www.gurobi.com, 2020.

[75] Centers for Disease Control and Prevention, Antibiotic Resistance Threats in the United States, 2019.

[76] W.G. Members, D. Mozaffarian, E.J. Benjamin, A.S. Go, D.K. Arnett, M.J. Blaha, M. Cushman, S.R. Das, S. de Ferranti, J. Després, et al., Heart disease and stroke statistics-2016 update: a report from the american heart association, Circulation 133 (2016) e38–e360.

[77] Agency for Healthcare Research and Quality, Rockville, MD, HCUP Tools and Software. Healthcare Cost and Utilization Project (HCUP), URL: https://www.hcup-us.ahrq.gov/tools_software.jsp, 2020.

[78] J.L. Warner, P. Zhang, J. Liu, G. Alterovitz, Classification of hospital acquired complications using temporal clinical information from a large electronic health record, J. Biomed. Inform. 59 (2016) 209–217.

[79] M.P. Bauer, M.P. Hensgens, M.A. Miller, D.N. Gerding, M.H. Wilcox, A.P. Dale, W. N. Fawley, E.J. Kuijper, S.L. Gorbach, Renal failure and leukocytosis are predictors of a complicated course of *clostridium difficile* infection if measured on day of diagnosis, Clin. Infect. Dis. 55 (2012) S149–S153.

[80] R. Doshi, J. Desai, Y. Shah, D. Decter, S. Doshi, Incidence, features, in-hospital outcomes and predictors of in-hospital mortality associated with toxic megacolon hospitalizations in the United States, Intern. Emerg. Med. 13 (2018) 881–887.

[81] N. Siddiqui, M. Dwyer, J. Stankovich, G. Peterson, D. Greenfield, L. Si, L. Kinsman, Hospital length of stay variation and comorbidity of mental illness: a retrospective study of five common chronic medical conditions, BMC Health Serv. Res. 18 (2018) 1–10.

[82] P.K. Myint, S. Sheng, Y. Xian, R.A. Matsouaka, M.J. Reeves, J.L. Saver, D.L. Bhatt, G.C. Fonarow, L.H. Schwamm, E.E. Smith, Shock index predicts patient-related clinical outcomes in stroke, J. Am. Heart Assoc. 7 (2018), e007581.

[83] A.A. Schmid, C.K. Wells, J. Concato, M.I. Dallas, A.C. Lo, S.E. Nadeau, L. S. Williams, A.J. Peixoto, M. Gorman, J.L. Boice, et al., Prevalence, predictors, and outcomes of poststroke falls in acute hospital setting, J. Rehabil. Res. Dev. 47 (2010) 553–562.

[84] M. Kim, S. Banerjee, Y. Zhao, F. Wang, Y. Zhang, Y. Zhu, J. DeFerio, L. Evans, S. M. Park, J. Pathak, Association networks in a matched case-control design – co-occurrence patterns of preexisting chronic medical conditions in patients with major depression versus their matched controls, J. Biomed. Inform. 87 (2018) 88–95.

Dr. Yajun Lu is an Assistant Professor in the Department of Management and Marketing at Jacksonville State University (JSU). His primary research interests are in Network Optimization, Graph-based Data Mining, and Data Analytics of Complex Networks with applications in Social Network Analysis and Healthcare. He received his Ph.D. in Industrial Engineering and Management from Oklahoma State University in 2019 and M.S. degree in Industrial Engineering from Huazhong University of Science and Technology in 2011. Prior to joining JSU, he was a Visiting Assistant Professor in the Department of Analytics and Operations Management at Bucknell University from 2019 to 2021. Besides his academic positions, he had worked as an Industrial Engineer in Huawei Technologies Co., Ltd., China from March 2011 to July 2014.

Suhao Chen is currently a Ph.D. student in the School of Industrial Engineering and Management at Oklahoma State University. His research interest is clinical data analytics. He received a master's degree in Management in 2010 and a bachelor's degree in Information Management and Systems in 2007, from Shanghai Jiao Tong University and Nanjing University, respectively.

Dr. Zhuqi Miao is currently the Health Data Science Program Manager of the Institute for Predictive Medicine of the Center for Health Systems Innovation, Oklahoma State University. His research interests include clinical data analytics, predictive medicine, network analysis and optimization. Dr. Miao received his master's and doctoral degrees in Industrial Engineering and Management from Oklahoma State University in 2012 and 2016, and his master's degree in Automation and bachelor's degree in management science from Xiamen University, China in 2007 and 2010. His research has been published in many prestigious journals in management science and medicine, including INFORMS Journal on Computing, Networks, Annals of Operations Research, Journal of the American Academy of Orthopaedic Surgeons, Journal of Clinical Medicine, and PloS ONE, among others.

Dr. Dursun Delen is the holder of William S. Spears Endowed Chair in Business Administration, Patterson Family Endowed Chair in Business Analytics, Director of Research for the Center for Health Systems Innovation, and Regents Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University. Prior to his academic tenure at OSU, he worked for a privately-owned research and consultancy company as a research scientist for five years, during which he led a number of advanced analytics research projects funded by federal agencies including DoD and NASA. Dr. Delen has authored more than 160 peer reviewed articles and 10 books/textbooks in the broad area of Business Analytics and Data Science. He is often invited to companies for consultancy engagements and national and international conferences for keynote addresses. He regularly chairs tracks and minitracks at various business analytics and information systems conferences. Currently, he is the editor-in-chief for the Journal of Business Analytics and the AI in Business (in Frontiers in Artificial Intelligence), senior editor for the Journal of Decision Support Systems, Decision Sciences, and Journal of Business Research, associate editor for Decision Analytics and International Journal of RF Technologies, and is on the editorial boards of several other academic journals. He has been the recipient of several research and teaching awards including the prestigious Fulbright scholar, regents' distinguished teacher and researcher, president's outstanding research and teaching, and Big Data mentor awards.

Dr. Andrew Gin is the Medical Director of the Institute for Predictive Medicine of the Center for Health Systems Innovations, Oklahoma State University. He received his MD degree from the University of Oklahoma in 1976, completing his residency in neurology in 1980. He received a bachelor's degree in chemistry in 1972 and a master's degree in Predictive Analytics in 2017, both from Northwestern University. Since 1980, he has been practicing clinical neurology. His primary data science interests include mathematical modeling and artificial and human intelligence.