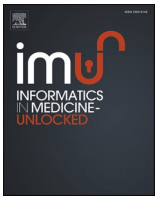
Informatics in Medicine Unlocked 39 (2023) 101243

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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



Summarizing multiple networks based on their underlying clustering structure to guide joint clustering of hospitals admissions

Nouf Albarakati, Avrum Gillespie, Zoran Obradovic *

Temple University, USA

ARTICLE INFO

Keywords:
Disease-based hospital clustering
Joint clustering
Graph matching
Network of networks

ABSTRACT

Healthcare services planning and regulation involve finding patterns in hospitals admission to detect their needs in a timely manner. Admission patterns for certain diseases are more precise than a general pattern including all diseases. Towards the objective of clustering hospitals based on their monthly admission behavior for different diseases, this study investigates the similarity among multiple disease-specific hospital networks to guide a joint clustering of hospitals. In this paper, the disease super network is generated from health records data using graph matching instead of relying on biomedical literature that is used in the previous work. The health records-based disease network is constructed using more than 7 million discharge records that are extracted from the California State Inpatient Database between 2009 and 2011. Comparison of the disease network results obtained using health records of different years shows consistency in clustering structure despite temporal changes in admission data. We show that the joint clustering guided by the health records-based similarity improves clustering group homogeneity measures as compared to the clustering guided by literature-based similarity (average homogeneity 53% vs 41%, respectively). The code used to conduct this work is available at https://github.com/Nouf-Barakati/JointCLusteringofHospitals.

1. Introduction

With the advancement of machine learning algorithms and the availability of Electronic Health Records (EHRs), healthcare organizations are investing in these products to extract insights and make informed decisions, utilize their resources effectively and efficiently, and enhance their services to meet the health needs of their population. For example, machine learning provides descriptive, predictive, and prescriptive analytics tools to understand and predict hospital admissions to help alleviate the current shortage of hospital beds which was exacerbated by the COVID-19 pandemic [1]. Understanding hospital admission behavior is critical to improving healthcare quality, reducing healthcare costs, and improving the overall health of the population [2]. By leveraging the insights gained from using machine learning tools on hospital admission data, healthcare providers and policymakers can make informed decisions that improve patient outcomes and enhance the effectiveness of the healthcare system [3].

Identifying groups of hospitals with similar monthly admission patterns would assist healthcare organizations in planning and regulating to improve their operational efficiencies and lower costs [4,5]. Clustering analysis has been applied to healthcare data to define groups of similar hospitals and detect different patterns in the hospital admissions [2,4,6]. Clustering is also utilized to identify factors associated with these patterns [6,7]. These admission patterns vary by conditions diagnosed at the admission time [8–10]. One of these factors is the principal diagnosis considered in the admission process [10,11]. For example, two hospitals could have similar numbers of monthly admission, yet have differences in monthly admission for cardiovascular disease or monthly admission for respiratory diseases. It is also evident from several studies that some hospitals have different admission variations for various diseases [10–12]. Therefore, it is imperative to cluster hospitals for different diseases.

To address this problem, we build a solution using layers of different methods to take advantage of each method's unique strengths and capabilities and build a more robust and flexible solution that suits this problem. First, because many medical and health-related phenomena involve interdependent entities, we build hospital similarity networks for different diseases included in the study [13]. Each hospital similarity network represents the similarity among hospitals' admission for a specific disease. In this representation, each similarity network is a

* Corresponding author. Data Analytics and Biomedical Informatics Center, 1925 N. 12th Street, SERC 386, Philadelphia, PA, 19122, USA.
E-mail addresses: nouf@temple.edu (N. Albarakati), Avrum.Gillespie@tuhs.temple.edu (A. Gillespie), zoran.obradovic@temple.edu (Z. Obradovic).

https://doi.org/10.1016/j.imu.2023.101243
Received 12 November 2022; Received in revised form 10 April 2023; Accepted 11 April 2023
Available online 14 April 2023

weighted graph where nodes represent hospitals admitting patients for a specific disease and edges represent the similarity among hospitals based on their monthly admission behavior for that specific disease. We use more than 7 million discharge records to extract monthly admission distributions for every disease in every hospital. These health records are obtained from the California State Inpatient Database between 2009 and 2011, where the records capture medical and sociodemographic information [14].

Then, instead of clustering individual networks separately, we jointly cluster all hospital networks considering the similarity among these networks to better reveal the underlying clustering structure. Clustering these hospital networks independently omits the global clustering structure shared among hospital networks [8]. Such an approach is motivated by an observation that joint clustering is often more efficient than independent clustering of individual networks [15]. This joint clustering takes advantage of the relationships and similarities between the networks, which can provide a more complete and accurate representation of the data. By clustering the data from multiple networks simultaneously, the joint clustering approach can capture complex patterns and structures that may not be apparent when analyzing each network individually. This can lead to more robust and accurate clustering results than independent clustering methods. In prior machine-learning studies, extensive experiments on synthetic and real-life data showed the effectiveness of joint clustering using various clustering performance measurements [16–18]. Joint clustering assumes that some networks share a common clustering structure where complementary information on this common structure is provided [16,17]. It identifies clusters that are consistent across similar networks while still preserving the network structure within each individual network.

Our previous work [8] utilized joint clustering to investigate the assumption that hospitals show similar behavior on their monthly admission distribution when similarity among disease symptoms is considered. The similarity among diseases was introduced from an external resource using a literature-based disease symptoms similarity network constructed using PubMed, an extensive medical bibliographic literature database [19]. Considering the similarity of the symptoms among diseases better revealed the underlying clustering structure for hospitals. Then, relying on the fact that health records provide opportunities to enhance and facilitate the clinical research [20,21], our follow-up study partially integrated the health record-based disease similarity network extracted from the external health records [9]. In this second study, a disease monthly-admission similarity was introduced to guide the joint clustering among hospital networks which improved the results [9]. However, the disease monthly-admission similarity didn't consider hospitals' different admission behavior among different diseases. Both prior studies [8,9] show consistent behavior among hospital networks when similarity among diseases is introduced from external resources in the clustering process.

These findings raised the question of what if we do not rely on external resources from PubMed or EHR to model the similarity among hospital networks. Would the clustering performance improve if disease similarity was calculated by analyzing the similarity among hospital networks and utilized to guide the joint clustering of hospital networks? Therefore, examining this question is the main contribution of this work. It extends our previous work to investigate the effect of disease similarity extracted from the similarity among disease-specific hospital networks on the joint clustering of these hospital networks. Graph matching is utilized to model the similarity among hospital networks and generate a health records-based disease network (HRDN). Graph matching finds the similarity between different hospital networks.

To carry out this study, data were represented as a Network of Networks (NoN) model. It is a multilayered network structure used to discover the hidden pattern in a heterogenous data [22]. This structure, illustrated in Fig. 1, is constructed of two layers of networks consisting of one super network and multiple sub-networks. The super network, disease network, at the top layer illustrates the similarity among different diseases, where each disease node represents a sub-network in the bottom layer. It mainly models the interconnectivity between the disease-specific hospital networks, which is used to guide joint clustering. Each subnetwork at the bottom layer is a disease-specific hospital network. Each disease-specific hospital network represents the similarity among hospital admission distributions for a specific disease. A method used to handle clustering heterogeneous multi-domain networks with an NoN-structure is the Network of Networks Clustering (NoNClus) [16]. This method permits multiple underlying clustering structures across different networks by clustering the super network and using it to guide clustering the multiple sub-networks [16]. That is, it groups multiple disease-specific hospital networks considering the grouping obtained for the disease network. This data model helped investigate the effect of different disease similarity networks by integrating a disease network with multiple disease-specific hospital networks.

In summary, the contributions of this paper are.

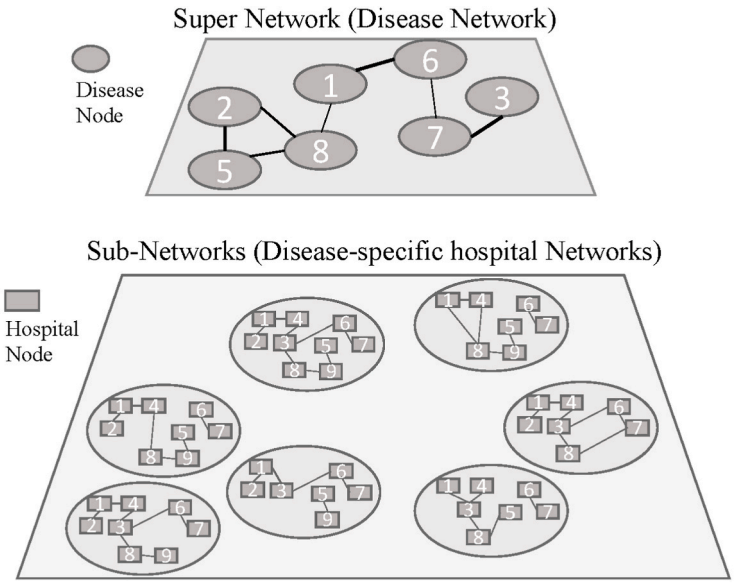


Fig. 1. Disease Network of Hospital Networks data model.

I. Summarizing multiple disease-specific hospital networks to generate a health records-based disease network using graph matching concept;
II. Comparing the literature-based disease network constructed from medical bibliographic literature and a health records-based disease network built from the health records;
III. Investigating the effect of the two disease networks on the joint clustering of multiple disease-specific hospital networks using the NoNClus method.

This paper is organized as follows: Section 2 provides an in-depth investigation of the related work, and Section 3 discusses the methodology and details the data model and algorithms used in carrying out this study. Section 4 illustrates the results, Section 5 interprets the results and discusses their significance, and Section 6 concludes this study.

2. Related work

Machine learning is increasingly being used in healthcare to improve the quality of care and increase operational efficiency. In particular, machine learning methods have been applied to cluster hospitals to better understand the relationships among healthcare facilities and identify improvement opportunities [4,5]. Clustering analysis has been applied to healthcare data to define groups of similar hospitals and detect different patterns in the hospital admissions [2,4,6]. Clustering is also utilized to identify factors associated with these different patterns [6,7]. Methods for clustering hospitals using multivariate data received attention in the 1970s and 1980s [5,6]. Another round of work studied clustering hospitals in the past 20 years by applying various clustering algorithms on patient-level or hospital-level data to identify subgroups of hospitals with similar characteristics [2,4,7,23,24]. Some of clustering algorithms used in these studies include k-means clustering, hierarchical clustering, and density-based clustering. k-means algorithm has been used to cluster hospital data [4,23], while multiple hierarchical algorithms [7,24] were used and compared in clustering hospitals.

While social network analysis and graph theory provides a powerful framework to study patterns, we are aware of only one published work that utilized social network analysis for clustering hospitals [23]. In that paper, hospitals are represented as nodes in a network, and the connections between nodes are used to capture the similarities between hospitals. However, the social network analysis of the hospital network was conducted after clustering hospitals using the k-means algorithm to understand mobility patterns in hospital clusters. Therefore, our current paper fills this gap of not utilizing the great potential in exploring relationships and patterns by using graph networks in clustering hospitals. In our approach hospitals are represented as nodes in a network and the connections between nodes are used to represent the similarities between hospitals.

However, most of the published work done in clustering hospitals was based on a single view of the hospitals' data [2,4,7,23,24]. A single view of hospital data represents a single hospital similarity network and only describes hospitals from a single aspect that does not accurately grasp the comprehensive information of hospitals. This limitation is addressed in this study. Conversely, the assumption we hold in conducting this work is that there are multi-views of hospital data with different similarities among hospitals for different diseases. To fulfill that assumption, joint clustering is used instead of using K-mean or DBSCAN algorithms. Using K-means and DBSCAN violate the assumption that some hospital networks share a common clustering structure and cannot cluster these hospital networks simultaneously. However, joint clustering is the method that allows clustering hospitals that belong to multiple networks simultaneously, considering the similarity among these networks.

Joint clustering algorithms are derived from spectral clustering and other graph-based methods [18,25–27]. Some methods consider joint clustering as multi-view clustering, where the constraint is to have the same number of nodes in all networks [25,26,28]. Other methods consider joint clustering as multi-domain clustering, where different networks have different sizes [16,17,29]. In this work, we consider multi-domain joint clustering algorithms to cluster multiple hospital networks assuming that different hospital networks have different numbers of hospitals. Moreover, not all joint clustering methods assumes that some networks share a common clustering structure. To address the two mentioned assumptions, we use a solution proposed by Ni et al. [16], a Network of Networks Clustering (NoNClus) method that handles clustering heterogeneous multi-domain networks with different numbers of nodes in each network. Further, the connectivity among these networks is modeled using the Network of Networks data structure to regularize the clustering structures in different networks. The NoNClus method permits multiple underlying clustering structures across different networks by clustering the super network and using it to guide clustering the multiple sub-networks. That is, it groups multiple sub-networks considering the grouping obtained for the super-network.

A Network of Networks (NoN) is a multilayered network structure used to discover the hidden pattern in heterogenous data [22]. This structure is constructed of a top-layer super network and multiple bottom-layer sub-networks. The super network in this structure models the interconnectivity between the sub-networks. Studies show that the top-layer super-network in the NoN structure represents critical information concerning shared clustering structure across all sub-networks [6,7,16]. Our previous work [8,9] utilized literature-based and health record-based super networks from external resources to guide the joint clustering of hospitals in the disease-specific hospital networks.

Besides the objective of clustering hospitals for different diseases using graph networks, the main contribution of this work is to investigate the joint clustering of hospitals guided by the health record-based super network extracted from the bottom-layer disease-specific hospital networks to calculate the similarities among these networks using the graph matching concept. Graph matching is the problem of finding a similarity between graphs. It has numerous applications in diverse fields, and therefore many algorithms and similarity measures were proposed to handle this problem [30]. Many studies were devoted to investigate exact and inexact graph matching by considering nodes, edges, and their attributes to analyze their similarities [31]. Theoretically, the graph-matching problem can be solved by comprehensively searching the entire solution space. However, this approach is practically unfeasible because the solution space expands exponentially as the size of input data increases. Therefore, prior studies use various approximation techniques to solve the problem [30,31]. In this study, the graph-matching problem is simplified by decomposing networks to a lower-rank approximation of a symmetric nonnegative matrix and then calculating the distance between these approximations because we are interested in examining the similarity of the underlying clustering structure of these networks.

3. Methodology

The main objective of this study is to jointly cluster hospitals based on their monthly admission behavior for different diseases considering the similarity among these diseases. To address this objective, we proposed a framework that is built using layers of different methods that leverages the strength of these methods. This section explains this proposed multi-layer framework. First, we built disease-specific hospital similarity networks. Then, we define the similarity among these different hospital networks using a health records-based disease network (HRDN) extracted from these hospital networks using a graph-matching algorithm. Also, we use a literature-based disease network (LDN) extracted from the human symptoms disease network to compare the results. Then, we leveraged the Network of Networks Data Model to jointly cluster these hospital networks using the NONClus method. Finally, we use the clustering homogeneity measure to measure the goodness of the clustering result.

3.1. Disease-specific hospital networks

The first layer in the proposed framework is using social network graphs to build hospital similarity networks. Data used to build these networks are extracted from the California State Inpatient Database (SID) as part of the Healthcare Cost and Utilization Project (HCUP) provided by the Agency for Healthcare Research and Quality (AHRQ) [14]. Over 7 million single-patient discharge records from the emergency department between 2008 and 2011 were used, including medical and sociodemographic information. There are 145 diseases included in this study. Hospitals' monthly admission distributions for each disease are aggregated for the principal diagnosis. The total number of hospitals used in this study was 152 out of 500 California hospitals, where hospitals with insufficient admission records for some diseases over the study period were excluded. A hospital was excluded from the study when it was represented in less than 50% of disease-specific hospital networks. This decision was made to ensure that there is enough data to measure the homogeneity among different networks.

Since hospitals have different admission distributions for different diseases,145 disease-specific hospital networks were built. Every disease-specific network has nodes representing hospitals with admission for the corresponding diseases considered in this study. Edges between these nodes represent similarities between hospitals' monthly admissions for the specific disease. Kullback-Leibler divergence was used to measure how one hospital's monthly admission distribution diverges from a second hospital's monthly admission distribution for every year separately.

3.2. HRDN: health records-based disease network using graph matching

The second layer of this framework defines the similarity among disease-specific hospital networks. To do so, we extract a disease network directly from the health records used to build disease-specific hospital networks using graph matching. Graph matching algorithms work by comparing the adjacency matrices or node adjacency lists of the two graphs/networks and finding the best match between nodes in the two graphs based on some criteria, such as minimizing the number of edges that need to be added or removed to make the two graphs isomorphic. A limitation of a graph matching-based analysis is potentially exponential complexity concerning the number of nodes in the graph. In our study, the similarity analysis of the underlying clustering structures is simplified by decomposing networks to a lower-rank approximation of a symmetric nonnegative matrix and then calculating the distance between these approximations. Then, we use the similarity matrix produced by the factorized graph matching method to examine its effect on clustering multi-domain disease-specific hospital networks. These steps of the second layer are summarized in Algorithm 1 and explained in more detail below.

Input: Multiple disease-specific hospital networks

Step 1: Acquire the lower-rank approximation of SNMF

For each disease-specific hospital networks, A_i, acquire low rank approximate H_i,
Do:

$$\min_{H\geq 0} \left\|A_i - H_i H_i^T\right\|_F^2$$

end

Step 2: Utilize the Euclidean distance to measure distance among networks based on clustering probabilities

For each disease-specific hospital networks, s,
Do:

For each disease-specific hospital networks, t,
Do:

$$d(s,t) = \min \sum_{i=1}^{k} \sum_{j=1}^{p} \sqrt{(h_{si} - h_{tj})^2}^{\,2},$$

s.t., k=number of clusters, p=permutations of clusters

end
end

Step 1: Symmetric Nonnegative matrix factorization (SNMF) is used for the graph clustering [25]. It provides a lower-rank approximation of a nonnegative matrix. It has been successfully used as a graph clustering method that takes an adjacency matrix as an input and produces clustering factors. It enforces nonnegativity on the clustering assignment matrix. We used SNMF to obtain a lower-rank approximation matrix for disease-specific hospital networks separately. For every disease-specific hospital network A_i, we acquire the lower-rank approximation H_i of SNMF using the objective function:

$$\min_{H\geq 0} \left\|A_i - H_i H_i^T\right\|_F^2$$

where A_i is the adjacency matrix for the i-th disease-specific hospital network, and H_i $(n \times k)$ is the lower-rank approximation (factor) of that network. Each row in H_i represents a hospital in the i-th disease-specific hospital network, and each column specifies the probability of a hospital belonging to a cluster.

Step 2: Factorized Graph Matching algorithm determines the degree of similarity between two graphs. Similarity score takes values in [0,1], where 0 means the two graphs are completely dissimilar, while 1 means they are identical. There are many studies to investigate exact and inexact graph matching by studying nodes, edges, and their attributes to analyze their similarities [31]. Our study is focused on the similarity of the underlying clustering structure among different networks where the similarity of the underlying clustering structure for two networks is measured by a set of nodes that belongs to the same cluster in both networks.

After obtaining a lower-rank approximation H_i in the first step, we measure the distance among every cluster of different networks to find the similarity in clustering structure among different networks. The lower-rank approximation can be easily interpreted in the context of clustering, where the largest entry h_{ij} indicates that node (hospital) i belongs to cluster j. So, each column in H represents different probabilities of different nodes belonging to that cluster.

Considering the way SNMF is calculated, it is necessary to calculate the distance between two networks for different permutations and find the minimum distance in these different permutations. For example, as an illustration assume that there are three clusters of hospitals in network 1 and network 2. The same set of hospitals may be in cluster 1 at network 1 and in cluster 1 in network 2, cluster 2 in network 2, or cluster 3 in network 2. Therefore, for 3 clusters (k = 3: (1,2,3)) in the first network, there are 6 permutations (p = 6: (1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2),(3,2,1)) of possible clusters in the second network to which measure the distance against.

Given the same clusters on different networks, s and t, and their corresponding vectors h_s and h_t in H, we say those two networks have identical underlying clustering structures if the minimum Euclidean distance between nodes probabilities of one network and permutation of probabilities of nodes in the other network equals zero. The distance can be written as:

$$d(\mathrm{s},\mathrm{t}) = \min \sum_{i=1}^{k} \sum_{j=1}^{p} \sqrt[2]{\left(h_{si} - h_{tj}\right)^2}$$

This distance is in the range of [0,∞). To convert this distance to a [0,1] similarity score, we normalized all distance scores using min-max normalization and converted the distance (d) to similarity (s) using the formula s = 1- d.

3.3. LDN: literature-based disease network

Considering the promising results obtained at our previous work [8,9] using the literature-based disease network (LDN), we compared the disease network obtained from the health records (HRDN) using our proposed summarizing method with one obtained from the medical bibliographic literature databases.

The literature-based disease network (LDN) is extracted from the human symptoms disease network. It is used as a weighted undirected disease network to model the interconnectivity among the bottom layer disease-specific hospital networks. The reason behind using a literature-based disease network was influenced by the fact that disease symptoms are one of the critical factors for admission decisions. It is used in this study for a comparison reason.

LDN is constructed by Zhou et al. [19] using the PubMed database and MeSH terminology. MeSH terminology indexed all articles in PubMed for over four thousand disease terms and over three hundred symptom terms. Then, the association between diseases and symptoms was identified, where a vector of related symptoms described every disease. The similarity between vectors of diseases was calculated as cosine ranging from 0 with no shared symptoms to 1, which means both diseases shared identical symptoms. Finally, it is important to mention that no patient records are utilized in generating LDN [19].

Clinical Classifications Software (CCS) code is used to categorize diseases because the disease-specific hospital networks are constructed using the California State Inpatient Database. Therefore, only 145 disease nodes are extracted from 1596 distinct diseases represented in the human symptoms network. The matching between the CCS codes and the MeSH terminology was done manually, and the average of similarities in some cases where the matching was not one-to-one was calculated [32].

3.4. NoNClus: Network of Networks Clustering

The third layer of the proposed framework is aimed to leverage the Network of Network data model and NoNClus clustering method to model the clustering structure in the disease-specific hospital networks. The clustering structure of the disease network is used to guide the clustering of different disease-specific hospital networks at the bottom layer. Fig. 2 shows the disease network of hospital networks data model where the NoNClus is used to cluster the disease network and all hospital networks into three clusters. The clustering of the disease network guides the clustering of disease-specific hospital networks.

The predefined number of main clusters or disease clusters (DC) for the top-layer network is k = 3. This simple number has been tested and given meaningful results. It implies that the underlying clustering structure among different disease-specific hospital networks is different. However, some networks may share the same underlying clustering structure if these networks belong to same group. For example, in Fig. 2, disease-specific hospital networks that represent diseases number 2 and 5, i.e., belong to the same top disease cluster, may share the same underlying clustering structure. Although the NoNClus method allows specifying different numbers of clusters among disease-specific hospital networks, our experiments were unified, and the number of hospital clusters was predefined as t = 3. This predefined cluster number has been chosen to keep this setting as simple as possible.

We built a NoN-data model using HRDN disease network as a top-layer network and 145 disease-specific hospital networks at the bottom layer. Another NoN-data model using LDN disease network and the same 145 disease-specific hospital networks at the bottom layers. Then we used NoNClus method for the joint clustering of these two NoN data models and compared the results.

The NoNClus [16] works in two stages. In the first stage, the NoNClus method clusters the top-layer disease network using a symmetric nonnegative matrix factorization by minimizing the objective function J:

$$\min \mathrm{J} = \left\| G - HH^T \right\|_F^2$$

where G is the adjacency matrix of the disease network. This network has g diseases/nodes, while H is the non-negative low-dimensional factor matrix of G with k clusters for g diseases. These factors define each disease node's probability to belong to one of the k main clusters.

In the second stage, the non-negative low-dimensional factor matrix of the disease network, H, is used as a regularization to guide the cluster of all disease-specific hospital networks and get a factor matrix of each disease-specific hospital network. The objective function for that second stage is:

$$\min_{\substack{U^{(i)} \geq 0 \ (i=1,\ldots,g) \\ V^{(j)} \geq 0 \ (j=1,\ldots,k)}} J = \sum_{i=1}^{g} \left\| A^i - U^i \left(U^i\right)' \right\|_F^2 + a \sum_{i=1}^{g} \sum_{j=1}^{k} h_{ij} \left\| U^i - V^j \right\|_F^2$$

where A^i represents the similarity in hospital admission for the ith disease, i.e., the adjacency matrix of every disease-specific hospital network, U^i is the factor matrix of ith disease-specific hospital networks and $V^{(j)}$ are introduced as k hidden clusters to represent the underlying structure of disease-specific hospital networks in the main cluster.

In this objective function, the first part deals with individually clustering the disease-specific hospital networks based on a similarity

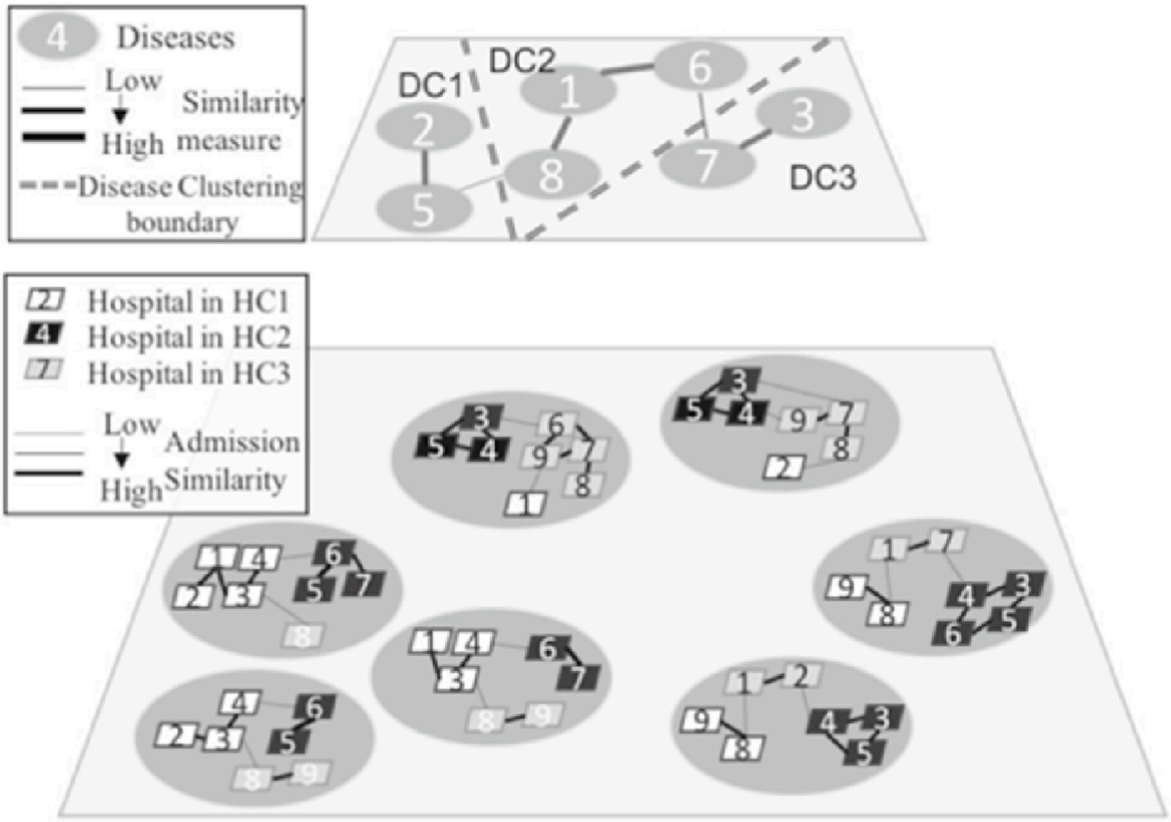


Fig. 2. Joint clustering of disease network of hospital networks.

matrix of hospital admission for each disease. The second part regularizes the factor matrix of each disease-specific hospital network, using the main clustering structure defined in the factor matrix of the main disease network, *H*, and the underlying clustering structure of domain-specific networks of the main cluster, *V* [16].

3.5. Clustering homogeneity measurement

Clustering is an unsupervised machine learning method that does not have a direct way to measure the goodness of the output. However, in this work, we evaluate the validity of clustering results using the clustering homogeneity measure. The NoNClus method performs clustering in two stages, clustering disease networks in the first stage and clustering hospitals in the second stage. Therefore, the group homogeneity measure is used to compare two disease networks clustered in the first stage and to compare hospital networks clustered in the second stage.

To compare the literature-based disease network (LDN) and health records-based disease network (HRDN), group homogeneity is defined as the percentage of the largest group of disease nodes belonging to the same cluster across different disease networks. Among different hospital networks, group homogeneity is the percentage of the largest hospital group that belongs to the same cluster across different networks.

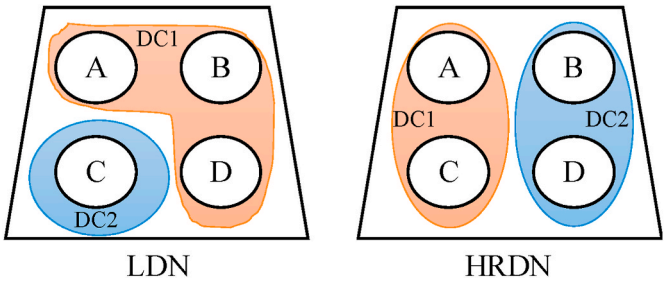


Fig. 3. Hypothetical example to explain clustering homogeneity visualizing two disease networks that have four disease nodes and two clusters.

The following is a hypothetical example of two disease networks with four disease nodes in each of them to explain the group homogeneity measure. Fig. 3 shows a visualization of this example where there are four disease nodes (A, B, C, and D) in both the literature-based disease network (LDN) and health records-based disease network (HRDN). In the LDN network, disease nodes A, B, and D belong to disease cluster DC1 while C belongs to disease cluster DC2. In the HRDN network, disease nodes A and C belong to disease cluster DC1, while B and D belong to disease cluster DC2. The group homogeneity value between LDN and HRDN networks is the maximum number of disease nodes that belong to the same disease clusters across both networks. In this example, there are three sets of disease nodes that belong to the same disease clusters across networks. In the first set, node C belongs to DC2 in the LDN network and to DC1 in the HRDN network. In the second set, node A belongs to DC1 in both networks, and in the third set, nodes B and D belong to DC1 in the LDN network and to DC2 in the HRDN network. To find the percentage, the maximum number of disease nodes that belongs to the same clusters across different networks is divided by the total number of disease nodes in both networks, which are nodes A, B, C, and D. That is, group homogeneity between these two networks is computed as the maximum of a different set of disease nodes grouped together in both networks divided by the total number of hospitals in both networks. For the previous example, group homogeneity between LDN and HRDN networks is max (1,1, 2)/4 = 2/4 or 50%.

4. Results and discussion

Towards the objective of clustering hospitals based on their monthly admission behavior for different diseases, we investigated the similarity among multiple disease-specific hospital networks to guide a joint

clustering of hospitals. We conducted two experiments using two different NoN data models. In the first experiment, the NoN data model has the HRDN disease network that was generated using the summarizing method we proposed as a super network to guide the joint clustering of the disease-specific hospital networks in the bottom layer. In the second experiment, the NoN data model has the LDN disease network as a super network to guide the joint clustering of the same set of disease-specific hospital networks in the bottom layer. As explained before, the joint clustering algorithm, NoNClus, clusters the disease super network (HRDN or LDN) to guide the clustering of disease-specific hospital sub-networks. Then, it uses disease clusters to guide the joint cluster of different disease-specific hospital networks. Therefore, this section presents the results of the clustering of two different disease networks used in these experiments and then the results of the joint clustering of hospital networks guided by HRDN and LDN disease networks.

4.1. Clustering of disease networks

A disease super network is used to guide the joint clustering of different disease-specific hospital sub-networks. Therefore, this subsection describes the clustering result of the two different disease networks, the Literature-based Disease Network (LDN) and the Health Record-based Disease Network (HRDN). In previous studies [8,9], the LDN disease network was introduced from an external resource to represent the similarity among disease-specific hospital networks. It is used in one experiment to compare its results with the second experiment's results. The HRDN disease similarity network is extracted from the disease-specific hospital networks using the graph-matching concept to characterize the effect of the learned similarity on the joint clustering of hospital networks. In the following subsections, we present, discuss, and compare the results of clustering HRDN with LDN done in the first step in the NoNClus method.

4.1.1. Clustering of literature-based disease network (LDN)

The literature-based disease network (LDN), used as a super network in one experiment, was extracted from the human symptoms network [19]. It is a super-network of 145 nodes where each node represents a disease. This weighted undirected disease network represents the similarity among diseases based on the similarity of the symptoms extracted from the medical bibliographic literature. Using NoNClus, 145 diseases are grouped into three main clusters in the first stage. One cluster has 34 diseases, the second cluster has 57 and the third cluster has 54 diseases. The top five diseases that have the highest probability to belong to each of the three clusters are listed in Table 1. The results show that the three clusters have a disease that shares similar symptoms. The first cluster comprises diseases that are related to pregnancy and childbirth complications. In the second cluster, the top diseases are related to respiratory and diseases of white blood cells. In the last cluster, top diseases vary more but have generally similar symptoms such as neurologic, nonspecific chest pain, and orthopedic conditions which may have occurred from a fall from the aforementioned neurologic conditions.

Table 1
Membership probability for the top five diseases in each of the three clusters of the LDN disease network.

CCS code	Disease name	LDN cluster	Probability
183	Hypertension complicating pregnancy, childbirth, and the puerperium	1	1.00
186	Diabetes or abnormal glucose tolerance complicating pregnancy; childbirth; or the puerperium	1	1.00
189	Previous C-section	1	0.95
190	Fetal distress and abnormal forces of labor	1	0.95
192	Umbilical cord complication	1	0.95
63	Diseases of white blood cells	2	1.00
125	Acute bronchitis	2	1.00
126	Other upper respiratory infections	2	1.00
133	Other lower respiratory disease	2	1.00
134	Other upper respiratory disease	2	1.00
79	Parkinson's disease	3	1.00
81	Other hereditary and degenerative nervous system conditions	3	1.00
82	Paralysis	3	1.00
102	Nonspecific chest pain	3	1.00
225	Joint disorders and dislocations; trauma related	3	1.00

4.1.2. Clustering of health record-based disease network (HRDN)

A health records-based disease network (HRDN), used as a super network in the second experiment, is obtained by summarizing multiple disease-based hospital networks into nodes according to the similarity among their underlying clustering structure. Four different HRDNs are generated for different years, from 2008 to 2011. As explained in Algorithm 1 in section 3.2, we proposed a method that calculates the similarity of the underlying clustering patterns among all disease-specific hospital sub-networks at the bottom layer to build a disease super-network. This disease network is used to guide clustering in different disease-specific hospital sub-networks using the NoNClus method. Table 2 lists the top five diseases that have the highest probability to belong to each of the three clusters in 2008, and to which clusters (with corresponding probabilities) these diseases belong in the following three years, 2009, 2010, and 2011. This table shows a significant consistency in the clustering results of the super network nodes among different years, as opposed to our previous work that generated a disease network from aggregated monthly admission for separate diseases over all hospitals and fused it into a single disease distribution network for every year. Cluster 1 is again pregnancy and its complication, while cluster 2 is now respiratory failure, esophageal disorders, secondary malignancies, and hip fractures; and Cluster 3 are infections, disorders of the breast including cancers, rheumatologic disorders, cancers of the kidney, and infections. Clustering results in the previous work were inconsistent over years. This observation shows that the summarizing method we proposed in this study generates a more robust disease network over years.

4.1.3. Comparison of LDN and HRDN-based clustering patterns

Clustering the two disease networks in stage 1 of the NoNClus method revealed some similarities and differences between LDN and HRDN. Table 3 shows the distribution of disease nodes in the three clusters for the two different disease networks obtained in the two experiments. Using NoNClus, 145 diseases in both the LDN and HRDN disease networks are grouped into three clusters. For the LDN disease network, one cluster has 34 diseases, the second cluster has 57 and the third cluster has 54 diseases. For the HRDN disease network, results show more consistency in the distribution of HRDN over the 4 years calculated. The numbers of nodes are between 19 and 25 in the first cluster over the four years, between 65 and 74 in the second cluster, and between 46 and 58 in the third cluster.

To investigate the clustering results among these networks, group homogeneity is calculated. Among LDN and HRDN networks, clustering homogeneity is defined as the percentage of the largest group of disease nodes belonging to the same cluster across both LDN and HRDN disease networks. Table 4 shows the group homogeneity measures among LDN and HRDN networks. The percentage of the different sets of disease nodes grouped together in both LDN disease network, and the four years of HRDN disease networks range between 39% and 45%. There was greater group homogeneity among the HRDN-based clustering, ranging between 78% and 86%, over the four years (2008, 2009, 2010, and 2011). This shows the consistency in the hospital admission pattern by disease over years and the differences in the clustering homogeneity among the clustering of LDN and clustering four years of HRDN.

However, Table 5 shows that the similarity between the top five

Table 2
Membership probability for the top five diseases in each of the three clusters of the HRDN disease network. The stability of the cluster assignment by the HRDN and the assignment probabilities are compared over four years (2008–2011).

CCS code	Disease name	2008 HRDN Clust	2008 Prob	2009 HRDN Clust	2009 Prob	2010 HRDN Clust	2010 Prob	2011 HRDN Clust	2011 Prob
185	Prolonged pregnancy	1	1.00	1	1.00	1	1.00	1	1.00
189	Previous C-section	1	1.00	1	1.00	1	0.99	1	1.00
190	Fetal distress and abnormal forces of labor	1	1.00	1	1.00	1	1.00	1	1.00
191	Polyhydramnios, other problems of amniotic cavity	1	1.00	1	1.00	1	0.89	1	1.00
192	Umbilical cord complication	1	1.00	1	1.00	1	1.00	1	1.00
42	Secondary malignancies	2	1.00	2	1.00	2	0.71	2	1.00
131	Respiratory failure; insufficiency; arrest	2	1.00	1	0.38	2	0.62	2	0.71
138	Esophageal disorders	2	1.00	2	0.93	2	0.88	2	0.70
143	Abdominal hernia	2	1.00	2	0.60	2	0.81	2	1.00
226	Fracture of neck of femur (hip)	2	1.00	2	0.77	2	0.92	2	0.87
8	Other infections; including parasitic	3	0.83	3	0.63	3	0.72	1	0.74
24	Cancer of breast	3	0.75	3	0.69	1	0.55	3	0.71
210	Systemic lupus erythematosus and connective tissue disorders	3	0.73	3	0.75	3	0.89	3	0.75
33	Cancer of kidney and renal pelvis	3	0.72	3	0.83	3	0.64	3	0.98
167	Nonmalignant breast conditions	3	0.70	1	1.00	3	0.64	3	0.59

Table 3
The number of disease nodes (percentage) in clusters found by the LDN and the HRDN.

	LDN	HRDN 2008	HRDN 2009	HRDN 2010	HRDN 2011
Cluster 1	34 (23.45%)	19 (13.10%)	22 (15.17%)	25 (17.24%)	21 (14.48%)
Cluster 2	57 (39.31%)	74 (51.03%)	65 (44.83%)	74 (51.03%)	69 (47.59%)
Cluster 3	54 (37.24%)	52 (35.86%)	58 (40.00%)	46 (31.72%)	55 (37.93%)

Table 4
Group Homogeneity of the LDN-based and the HRDN-based disease networks.

	LDN	HRDN 2008	HRDN 2009	HRDN 2010	HRDN 2011
LDN	1.00	0.41	0.39	0.45	0.41
HRDN 2008	0.41	1.00	0.85	0.83	0.86
HRDN 2009	0.39	0.85	1.00	0.78	0.82
HRDN 2010	0.45	0.83	0.78	1.00	0.81
HRDN 2011	0.41	0.86	0.82	0.81	1.00

diseases in each of the three clusters of the literature-based disease network falls into the same cluster in the HRDN for 2008.

4.2. Clustering of disease-specific hospital networks

The LDN and HRDN were used in two different experiments to guide the joint clustering of different disease-specific hospital sub-networks. In the first experiment, the NoN data model is used with LDN as a super network at the top layer and disease-specific hospital networks at the bottom layer. In the second experiment, another NoN data model is used with HRDN as a super network at the top layer and disease-specific hospital networks at the bottom layer. Then, we jointly clustered hospitals using NoNClust method. The clustering results of disease-specific hospital networks in these two experiments and the comparison between them are listed in the following subsections.

4.2.1. Hospital joint clustering guided by the LDN

Clustering hospitals vary among multiple disease-specific hospital networks. Previous studies [6,7] showed that the underlying clustering structure is similar for hospital networks that are represented by disease nodes belonging to the same disease cluster at the disease super network in NoN data model. Hospital clustering analysis using a LDN disease network to guide the clustering in the NoNClus algorithm is shown in Table 6. The results are for the top five diseases that have a high probability to belong to each of the three clusters at the LDN disease network layer.

Row-wise, the table is organized into three main sections. The first five rows present data for the top five disease-specific hospital networks that represent the top five disease nodes belonging to the first LDN cluster. The next five rows are the results of the top five disease-specific hospital networks that represent the top five disease nodes belonging to the second LDN cluster. The third set of five rows of results is for the top five disease-specific hospital networks that represent the top five disease nodes belonging to the third main cluster of the literature-based disease network.

Table 5
Comparison of clustering results of the LDN and the HRDN. The numbers show the probability of the disease belonging to clusters (C1, C2, C3).

CCS code	Disease name	LDN	HRDN 2008
185	Prolonged pregnancy	0.95 (C1)	1.00 (C1)
189	Previous C-section	0.95 (C1)	1.00 (C1)
190	Fetal distress and abnormal forces of labor	0.95 (C1)	1.00 (C1)
191	Polyhydramnios and other problems of amniotic cavity	0.95 (C1)	1.00 (C1)
192	Umbilical cord complication	0.95 (C1)	1.00 (C1)
126	Other upper respiratory infections	1.00 (C2)	0.66 (C2)
133	Other lower respiratory disease	1.00 (C2)	0.81 (C2)
139	Gastroduodenal ulcer (except hemorrhage)	1.00 (C2)	0.73 (C2)
140	Gastritis and duodenitis	1.00 (C2)	0.92 (C2)
154	Noninfectious gastroenteritis	1.00 (C2)	0.57 (C2)
81	Other hereditary and degenerative nervous system conditions	1.00 (C3)	0.56 (C3)
90	Inflammation: infection of eye (except that caused by tuberculosis or sexually transmitted disease)	0.97 (C3)	0.63 (C3)
167	Nonmalignant breast conditions	0.94 (C3)	0.70 (C3)
204	Other non-traumatic joint disorders	0.96 (C3)	0.53 (C3)
225	Joint disorders and dislocations; trauma-related	1.00 (C3)	1.60 (C3)

Table 6
Hospital clustering analysis using the LDN-based disease network. The number of hospitals (the percentage of the hospitals).

CCS code	Disease-specific hospital network	LDN Clust.	Prob. LDN Clust.	No. Hosp. in network (% out of 152)	No. of Hosp. in Clus.1 (%)	No. of Hosp. in Clus.2 (%)	No. of Hosp. in Clus.3 (%)
183	Hypertension complicating pregnancy, childbirth, and the puerperium	1	1.00	97 (0.64)	32 (0.33)	35 (0.36)	30 (0.31)
186	Diabetes or abnormal glucose tolerance complicating pregnancy; childbirth; or the puerperium	1	1.00	65 (0.43)	18 (0.28)	25 (0.38)	22 (0.34)
189	Previous C-section	1	0.95	55 (0.36)	16 (0.29)	24 (0.44)	15 (0.27)
190	Fetal distress and abnormal forces of labor	1	0.95	52 (0.34)	14 (0.27)	22 (0.42)	16 (0.31)
192	Umbilical cord complication	1	0.95	47 (0.31)	15 (0.32)	18 (0.38)	14 (0.30)
63	Diseases of white blood cells	2	1.00	141 (0.93)	52 (0.37)	41 (0.29)	48 (0.34)
125	Acute bronchitis	2	1.00	149 (0.98)	52 (0.35)	44 (0.30)	53 (0.36)
126	Other upper respiratory infections	2	1.00	144 (0.95)	63 (0.44)	35 (0.24)	46 (0.32)
133	Other lower respiratory disease	2	1.00	151 (0.99)	61 (0.40)	43 (0.28)	47 (0.31)
134	Other upper respiratory disease	2	1.00	139 (0.91)	50 (0.36)	41 (0.29)	48 (0.35)
79	Parkinson's disease	3	1.00	108 (0.71)	33 (0.31)	31 (0.29)	44 (0.41)
81	Other hereditary and degenerative nervous system conditions	3	1.00	135 (0.89)	44 (0.33)	41 (0.30)	50 (0.37)
82	Paralysis	3	1.00	83 (0.55)	22 (0.27)	27 (0.33)	34 (0.41)
102	Nonspecific chest pain	3	1.00	152 (1.00)	53 (0.35)	50 (0.33)	49 (0.32)
225	Joint disorders and dislocations; trauma related	3	1.00	123 (0.81)	39 (0.32)	39 (0.32)	45 (0.37)

Column-wise, this table lists the information about each disease-specific hospital network and then details about hospitals in the network. For each disease-specific hospital network, CCS code and the name of the disease for that network are listed first. then, the LDN cluster to which this disease belongs in the LDN disease network at the top layer of NoN data model and the probability to belong to that cluster are listed. After explaining the network, the rest of the columns present the results of hospitals clustering inside the network. First, the number of hospitals in the network is listed with a percentage of the hospitals in the network out of 152 total hospitals used in the study. Then the number of hospitals in each cluster is listed with the percentage of these hospitals out of the number of hospitals in the network.

Table 6 shows similar clustering patterns among the disease-specific hospital networks that belong to the same LDN disease cluster. The distribution of hospitals in the three clusters is similar among the set of disease-specific hospital networks that belongs to the first cluster in the LDN disease network. Also, this consistency in the distribution among different hospital clusters is extended to the next two sets of disease-specific hospital networks that belong to the second and third LDN clusters.

As shown in Table 6, the first set of disease-specific hospital networks that represent the top five diseases belonging to LDN cluster 1 has a low number of hospitals that range between 29% and 38% of hospitals out of 152 hospitals in each of the five networks. The distribution of these hospitals in each cluster is consistent among the first set of networks. On average, 30% of hospitals are in the first cluster, 40% of hospitals in the second cluster, and 30% of hospitals are grouped in the third cluster. For example, for the diagnosis of previous C-section (189), there are 55 hospitals that treated patients with that disease, 16 (29%) hospitals (Cluster 1) have different admission patterns for this diagnosis compared to the 24 (44%) hospitals in cluster 2, and 15 (27%) hospitals in cluster 3. The differences in admission patterns in these hospital clusters can be used to inform the staffing of nurses and other maternity resources. Considering the low representation of these hospitals in this set, it is important to mention that the total number of hospitals used in this study is 152 hospitals out of 500 California hospitals as there some hospitals had no admission records for some diseases included in this study over the study period. Also, hospitals were eliminated from the study if they are not represented in more than 50% of disease-specific hospital networks reflecting that not all hospitals have maternity services.

The second set of networks, that represent the top five diseases that belong to LDN cluster 2, has the highest number of hospitals in each network; The percentage of hospitals in these networks are between 93% and 99% of 152 hospitals. Also, the distribution of these hospitals in each cluster is considerably consistent. The average percentage of hospitals grouped in the first hospitals cluster in every network is 38%. In average, 29% of hospitals are grouped in the second cluster, and 34%

Table 7
Hospital clustering analysis using the HRDN. The number of hospitals (the percentage of the hospitals).

CCS code	Disease-specific hospital network	HRDN Clust.	Prob. HRDN Clust.	No. Hosp. in network (% out of 152)	No. of Hosp. in Clus.1 (%)	No. of Hosp. in Clus.2 (%)	No. of Hosp. in Clus.3 (%)
185	Prolonged pregnancy	1	1.00	44 (0.29)	16 (0.36)	13 (0.30)	15 (0.34)
189	Previous C-section	1	1.00	55 (0.36)	25 (0.45)	14 (0.25)	16 (0.29)
190	Fetal distress and abnormal forces of labor	1	1.00	52 (0.34)	21 (0.40)	16 (0.31)	15 (0.29)
191	Polyhydramnios and other problems of amniotic cavity	1	1.00	57 (0.38)	23 (0.40)	17 (0.30)	17 (0.30)
192	Umbilical cord complication	1	1.00	47 (0.31)	20 (0.43)	16 (0.34)	11 (0.23)
42	Secondary malignancies	2	1.00	147 (0.97)	48 (0.33)	50 (0.34)	49 (0.33)
131	Respiratory failure; insufficiency; arrest (adult)	2	1.00	152 (1.00)	48 (0.32)	50 (0.33)	54 (0.36)
138	Esophageal disorders	2	1.00	149 (0.98)	47 (0.32)	44 (0.30)	58 (0.39)
143	Abdominal hernia	2	1.00	151 (0.99)	48 (0.32)	49 (0.32)	54 (0.36)
226	Fracture of neck of femur (hip)	2	1.00	147 (0.97)	48 (0.33)	42 (0.29)	57 (0.39)
8	Other infections; including parasitic	3	0.830	101 (0.66)	30 (0.30)	38 (0.38)	33 (0.33)
24	Cancer of breast	3	0.748	91 (0.60)	26 (0.29)	37 (0.41)	28 (0.31)
210	Systemic lupus erythematosus and connective tissue disorders	3	0.732	109 (0.72)	31 (0.28)	41 (0.38)	37 (0.34)
33	Cancer of kidney and renal pelvis	3	0.724	101 (0.66)	29 (0.29)	37 (0.37)	35 (0.35)
167	Nonmalignant breast conditions	3	0.704	103 (0.68)	27 (0.26)	38 (0.37)	38 (0.37)

of hospitals are grouped in the third cluster.

The third set of disease-specific hospital networks, that represent the top five diseases that belong to LDN cluster 3, has a hospital number that ranges between 55% and 100% of 152 hospitals. Also, the distribution of these hospitals in each cluster is considerably consistent. The average percentage of hospitals grouped in every network is 31% in the first cluster, 31% in the second cluster, and 38% in the third cluster.

4.2.2. Hospital joint clustering guided by the HRDN

The result of hospital clustering analysis using the HRDN to guide the clustering process is shown in Table 7. This hospital clustering is regularized by clusters of the summarized disease network calculated from the similarity of the clustering structures among different disease-specific hospital sub-networks at the bottom layer using the proposed summarizing algorithm to build the HRDN disease super-network. This table lists the results of joint clustering disease-specific hospital networks for the top five diseases that have a high probability to belong to each of the three HRDN clusters at the disease network layer.

Row-wise, the table is organized into three main sections. The first five rows present data for the top five disease-specific hospital networks that represent the top five disease nodes belonging to the first HRDN cluster of the HRDN network. The next five rows are the results of the top five disease-specific hospital networks that represent the top five disease nodes belonging to the second HRDN cluster. The third set of five rows of results is for the top five disease-specific hospital networks that represent the top five disease nodes belonging to the third HRDN cluster.

Column-wise, this table also lists the information about each disease-specific hospital network and then details about hospitals in the network. For each disease-specific hospital network, the CCS code and the name of the disease for that network are listed first. Then, the HRDN cluster to which this disease belongs in the disease network at the top layer of NoN data model and the probability to belong to that cluster are listed. After explaining the network, the rest of the columns present the results of hospitals inside the network. First, the number of hospitals in the network is listed with a percentage of the hospitals in the network out of 152 total hospitals used in the experiments. Then, the rest of the columns show the number of hospitals in each cluster with the percentage of these hospitals out of the number of hospitals in the network.

As shown in Table 7, the first set of disease-specific hospital networks that represent the top five diseases belonging to the first HRDN cluster has a low number of hospitals that ranges between 29% and 38% of hospitals out of 152 hospitals in each of the five networks. The distribution of these hospitals in each cluster is consistent among the first set of networks. On average, 40% of hospitals are grouped in the first cluster, and 30% in the second and third clusters, respectively.

The second set of networks, which represents the top five diseases that belong to HRDN cluster 2, has the highest number of hospitals in each network; The percentages of hospitals in these networks are between 97% and 100% of 152 hospitals. Also, the distribution of these hospitals in each cluster is considerably consistent. The average percentage of hospitals grouped in every network is 32% in the first cluster, 32% in the second cluster, and 35% in the third cluster.

The third set of disease-specific hospital networks, which represents the top five diseases that belong to HRDN cluster 3, has a hospital number between 60% and 72% of 152 hospitals. Also, the distribution of these hospitals in each hospital cluster is considerably consistent. The average percentage of hospitals grouped in every network is 28% in the first cluster, 38% in the second cluster, and 34% in the third cluster.

4.2.3. Comparing hospital clustering homogeneity when using LDN versus using HRDN

As mentioned in section 3.5, we compared the clustering results of the two NoN data models by evaluating the sense of belongingness using the clustering homogeneity measure. Among different hospital networks, group homogeneity is the percentage of the largest hospital group that belongs to the same cluster across different networks.

Table 8 shows three sets of group homogeneity measures among five disease-specific hospital networks that represent the top five diseases belonging to each cluster of the LDN disease network. In other words, the left table in Table 8 shows the group homogeneity measures for disease-specific hospital networks that belong to the top five diseases belonging to the first LDN cluster, the second table for the diseases belonging to the second LDN cluster, and the third table for the third LDN cluster. In the first table, group homogeneity measures among these networks range between 38% and 47%, meaning that between 38% and 47% of hospitals in these networks belong to the same cluster over these networks. The second set of disease-specific hospitals has group homogeneity measures between 36% and 42% while the third set has group homogeneity measures that range between 35% and 47%. For example, in Table 8 hospital admissions for diagnosis 189 Previous C-sections occurs with a diagnosis 192 umbilical cord complication 38% of the time.

Comparing the two models, Table 9 shows three sets of group homogeneity measures among five disease-specific hospital networks that represent the top five diseases belonging to each cluster of the HRDN disease network. Group homogametic measures in this table show a slight improvement in using the HRDN overusing a LDN disease network. In the first left table inside Table 9, the group homogeneity measures for disease-specific hospital networks that belong to the top five diseases belonging to the first HRDN cluster range between 43% and 61% which means that between 43% and 61% of hospitals in these networks belong to the same cluster over these networks. The second set of disease-specific hospitals that belongs to the top five diseases belonging to the second HRDN cluster has group homogeneity measures between 38% and 58% while the third set has group homogeneity measures that range between 41% and 59%. Some notable association from cluster 1 of the HRDN is that 185 Prolonged labor is associated with the diagnosis of 190 fetal distress suggesting potential fetal complications of prolonged labor. In cluster 2, diagnosis 131 Respiratory failure occurs 48% of the time with an abdominal hernia which could indicate a potential surgical complication after hernia repair. For cluster 3, the diagnosis of 8 other infections occurs with 210 Systemic Lupus Erythematous (SLE) 59% of the time, signaling that SLE patients may frequently suffer from infections related to immunosuppressive therapies.

Studying group homogeneity of all disease-specific hospital networks, group homogeneity measures ranging between 35% and 47% with average measures close to 41% for the clustering guided using a LDN disease network. Group homogeneity measures ranged between 39% and 61% with average measures close to 53% for the clustering guided using the HRDN. These numbers provide evidence that the HDRN model slightly outperformed the LDN model. Therefore, a joint cluster of

Table 8
Homogeneity of three clusters found using the LDN.

	183	186	189	190	192		63	125	126	133	134		79	81	82	102	225
183	1.00	0.42	0.44	0.39	0.47	63	1.00	0.42	0.39	0.39	0.36	79	1.00	0.39	0.47	0.36	0.39
186	0.42	1.00	0.43	0.41	0.44	125	0.42	1.00	0.37	0.41	0.36	81	0.39	1.00	0.44	0.36	0.39
189	0.44	0.43	1.00	0.45	0.38	126	0.39	0.37	1.00	0.42	0.42	82	0.47	0.44	1.00	0.42	0.46
190	0.39	0.41	0.45	1.00	0.41	133	0.39	0.41	0.42	1.00	0.39	102	0.36	0.36	0.42	1.00	0.35
192	0.47	0.44	0.38	0.41	1.00	134	0.36	0.36	0.42	0.39	1.00	225	0.39	0.39	0.46	0.35	1.00

Table 9
Homogeneity of three clusters found using the HRDN.

	185	189	190	191	192		42	131	138	143	226		8	24	210	33	167
185	1.00	0.50	0.61	0.45	0.43	42	1.00	0.39	0.58	0.39	0.56	8	1.00	0.42	0.59	0.43	0.42
189	0.50	1.00	0.45	0.44	0.59	131	0.39	1.00	0.44	0.48	0.41	24	0.42	1.00	0.41	0.41	0.47
190	0.61	0.45	1.00	0.58	0.44	138	0.58	0.44	1.00	0.47	0.38	210	0.59	0.41	1.00	0.46	0.43
191	0.45	0.44	0.58	1.00	0.45	143	0.39	0.48	0.47	1.00	0.47	33	0.43	0.41	0.46	1.00	0.58
192	0.43	0.59	0.44	0.45	1.00	226	0.56	0.41	0.38	0.47	1.00	167	0.42	0.47	0.43	0.58	1.00

hospitals guided by the HRDN can replace a joint cluster of hospitals guided by a LDN disease network. For example, in comparison to Table 8, in Table 9 hospital admissions for diagnosis 189 Previous C-sections occurs with diagnosis 192 umbilical cord complication 59% of the time.

5. Conclusion

The objective of this study was to cluster hospitals based on their monthly admission behavior for different diseases to better reveal hidden patterns and assist healthcare organizations in planning and regulation. We proposed a method to summarize multiple disease-specific hospital networks and generate a health records-based disease network (HRDN) that is used to guide a joint clustering of hospital networks. We compared the joint clustering guided by HRDN and the literature-based disease network (LDN) constructed from medical bibliographic literature. The experimental results show the enhancement in clustering homogeneity when the disease network was extracted from health records by summarizing the underlying clustering structure of different hospital networks. This is significant because it better revealed the hidden underlying hospital clustering structure for specific diseases without the need to get external data. We found unique hospital clusters with different admission for the same disease as well as unique disease clusters among diagnoses which could inform hospital policies and procedures. Further the proposed approach could provide potential solutions to other similar problems. In future research, we need to automate the proposed method to summarize multiple hospital networks using different similarity measures and with the ability to evaluate the optimal number of clusters. This will open the road to exploring more dimensions of hospital and disease clustering to improve healthcare.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Jay Patel for his editing help, insightful suggestions, feedback sessions, and moral support.

References

[1] Karan A, Wadhera RK. Healthcare system stress due to Covid-19: evading an evolving crisis. J Hosp Med 2021;16(2):127.
[2] Belciug S, Gorunescu F. Improving hospital bed occupancy and resource utilization through queuing modeling and evolutionary computation. J Biomed Inf 2015;53:261–9.
[3] Cunningham P, Sammut J. Inadequate acute hospital beds and the limits of primary care and prevention. Emerg Med Australasia (EMA) 2012;24(5):566–72.
[4] Delamater PL, Shortridge AM, Messina JP. Regional health care planning: a methodology to cluster facilities using community utilization patterns. BMC Health Serv Res 2013;13(1):1–16.
[5] Thomas JW, Griffith JR, Durance P. Defining hospital clusters and associated service communities in metropolitan areas. Soc Econ Plann Sci 1981;15(2):45–51.
[6] Phillip PJ, Mullner R, Andes S. Toward a better understanding of hospital occupancy rates. Health Care Financ Rev 1984;5(4):53.
[7] Shay PD. More than just hospitals: an examination of cluster components and configurations. Virginia Commonwealth University; 2014.
[8] Albarakati N, Obradovic Z. Disease-based clustering of hospital admission: disease network of hospital networks approach. In: 2017 IEEE 30th international symposium on computer-based medical systems (CBMS). IEEE; 2017.
[9] Albarakati N, Obradovic Z. Multi-domain and multi-view networks model for clustering hospital admissions from the emergency department. International Journal of Data Science and Analytics 2019;8(4):385–403.
[10] Bourgeois FT, et al. Variation in emergency department admission rates in US children's hospitals. Pediatrics 2014;134(3):539–45.
[11] Sabbatini AK, Nallamothu BK, Kocher KE. Reducing variation in hospital admissions from the emergency department for low-mortality conditions may produce savings. Health Aff 2014;33(9):1655–63.
[12] McMahon Jr LF, Wolfe RA, Tedeschi PJ. Variation in hospital admissions among small areas: a comparison of Maine and Michigan. Med Care 1989:623–31.
[13] Hu H, et al. Review of social networks of professionals in healthcare settings—where are we and what else is needed? Glob Health 2021;17:1–17.
[14] NIS HNIS. Healthcare cost and utilization project (HCUP). 2011.
[15] Sugano Y, Matsushita Y, Sato Y. Graph-based joint clustering of fixations and visual entities. Trans Appl Percept 2013;10(2):1–16.
[16] Ni J, et al. Flexible and robust multi-network clustering. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining; 2015.
[17] Ni J, et al. Self-grouping multi-network clustering. In: 2016 IEEE 16th international conference on data mining (ICDM). IEEE; 2016.
[18] Zhang G-Y, et al. Joint representation learning for multi-view subspace clustering. Expert Syst Appl 2021;166:113913.
[19] Zhou X, et al. Human symptoms–disease network. Nat Commun 2014;5(1):1–10.
[20] Cowie MR, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017;106(1):1–9.
[21] Bartlett VL, et al. Feasibility of using real-world data to replicate clinical trial evidence. JAMA Netw Open 2019;2(10). e1912869-e1912869.
[22] Kivelä M, et al. Multilayer networks. Journal of complex networks 2014;2(3):203–71.
[23] Chrusciel J, et al. Making sense of the French public hospital system: a network-based approach to hospital clustering using unsupervised learning methods. BMC Health Serv Res 2021;21:1–12.
[24] Carr BG, et al. Quality through coopetition: an empiric approach to measure population outcomes for emergency care–sensitive conditions. Ann Emerg Med 2018;72(3):237–45.
[25] Gligorijević V, Panagakis Y, Zafeiriou S. Non-negative matrix factorizations for multiplex network analysis. IEEE Trans Pattern Anal Mach Intell 2018;41(4):928–40.
[26] Li Y, et al. Large-scale multi-view spectral clustering via bipartite graph. In: Proceedings of the AAAI conference on artificial intelligence; 2015.
[27] Wang R, Yan J, Yang X. Graduated assignment for joint multi-graph matching and clustering with application to unsupervised graph matching network learning. Adv Neural Inf Process Syst 2020;33:19908–19.
[28] Hung MH, et al. Prediction of masked hypertension and masked uncontrolled hypertension using machine learning. Front Cardiovasc Med 2021;8:778306.
[29] Ni J, et al. Inside the atoms: ranking on a network of networks. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining; 2014.
[30] Yan J, et al. A short survey of recent advances in graph matching. In: Proceedings of the 2016 ACM on international conference on multimedia retrieval; 2016.
[31] Grohe M, Rattan G, Woeginger GJ. Graph similarity and approximate isomorphism. arXiv preprint arXiv:1802.08509 2018.
[32] Glass J, et al. Extending the modelling capacity of Gaussian conditional random fields while learning faster. In: Thirtieth AAAI conference on artificial intelligence; 2016.