

RESEARCH

Open Access



A network analysis to identify lung cancer comorbid diseases

Heru C. Rustamaji^{1,6}, Yustina S. Suharini^{1,7}, Angga A. Permana^{1,8}, Wisnu A. Kusuma^{1,4*}, Sri Nurdianti²,
Irmanida Batubara^{3,4} and Taufik Djatna⁵

*Correspondence:
ananta@apps.ipb.ac.id

¹ Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia
Full list of author information is available at the end of the article

Abstract

Cancer patients with comorbidities face various life problems, health costs, and quality of life. Therefore, determining comorbid diseases would significantly affect the treatment of cancer patients. Because cancer disease is very complex, we can represent the relationship between cancer and its comorbidities as a network. Furthermore, the network analysis can be employed to determine comorbidities as a community detection problem because the relationship between cancer and its comorbidities forms a community. This study investigates which community detection algorithms are more appropriate to determine the comorbid of cancer. Given different community findings, this study attempted to analyze the modularity generated by the algorithm to decide the significant comorbid diseases. We retrieved lung cancer comorbid data on the basis of text mining manuscripts in PubMed, searched through disease ontologies, and calculated disease similarity. We investigate 20 algorithms using five modularity metrics and 16 fitness function evaluations to determine the significant comorbid diseases. The results show the five best modularity algorithms, namely label propagation, spinglass, Chinese whispers, Louvain, RB Pots. These five algorithms found significant comorbidities: blood vessels, immune system, bone, pancreas, and metabolic disorders, atrial cardiac septal defect, atrial fibrillation respiratory system, interstitial lung, and diabetes mellitus. The fitness function justifies the results of the community algorithm, and the ones that have a significant effect are average internal degree, size, and edges inside. This study contributes to more comprehensive knowledge and management of diseases in the healthcare context.

Keywords: Network analysis, Comorbid, Lung cancer, Community algorithms, Modularity, Fitness function

Introduction

It is challenging to derive the collective behavior from knowing the system's components in complex systems, such as proteins and disease networks. We will never fully comprehend the complex systems unless we understand the networks that underpin them (Barabasi 2016). The network often conceptualizes system interactions, expressed as vertices (nodes) and edges (links) between pairs of nodes (Loe and Jensen 2015). The nodes represent the elements that make up the system, and the links describe their interactions. In a disease network, the nodes represent disease, and the links represent disease

similarities between the corresponding illnesses, constructed via Disease Ontology (DO). DO, an authoritative disease curation service, established curation to coordinate disease representation across biomedical resource (Schriml and Mitraka 2015). DO enables researchers to analyze the disease similarity through semantic similarity measures, expanding our understanding of the relationships between various diseases and classifying them (Li et al. 2011).

Network analysis reveals the network's core features, allowing complex relationships and network structure to be estimated (Hevey 2018). The network analysis also identifies groups of nodes strongly connected to the rest of the network. These interrelated groups characterize communities (Yang et al. 2016). A community is a local subgraph densely connected in a network. The community detection aims to expose the community structure attached to the network. Most of the techniques do not specify the number and size of communities (Barabasi 2016).

The community is essential in the medical field, with diseases as complex as cancer and several comorbidities. Comorbidity refers to the existence of a long-term health condition in the presence of a primary disease of interest. Having comorbidities may influence the patient's prognosis for primary diseases such as cancer (Fowler et al. 2020). According to the World Health Organization, cancer was the first or second cause of death before the age of 70 in 112 of 183 countries worldwide in 2019. Lung cancer remained the leading cause of cancer death, with an estimated 1.8 million deaths (18%) (Sung et al. 2021). Various studies, revealed that many cancer patients have comorbidities. Chronic obstructive pulmonary and cardiovascular disorders are the most common comorbidities among patient with lung cancer (Pavia et al. 2007). Other comorbid diseases are immune system (Jacob et al. 2020), bone diseases (Kuchuk et al. 2013), pancreatic disease (Bang et al. 2014), metabolic disease (Feng et al. 2020), atrial cardiac septal defect (Inafuku et al. 2016), interstitial lung disease (Margaritopoulos et al. 2017), familial atrial fibrillation (Bandyopadhyay et al. 2019), respiratory system disease (Leduc et al. 2017), diabetes mellitus (Hatlen et al. 2011), and hyperlipidemia (Huang et al. 2016). It is also possible for patient with lung cancer to have multiple overlapping conditions (Sigel et al. 2017). Comorbidities can affect the stage of cancer. Patients with comorbidities face a poorer quality of life and require higher healthcare costs, resulting in shorter patient survival (Sarfati et al. 2016). Thus, understanding the comorbidities that coexist with lung cancer is necessary for screening and disease management.

Disease comorbidity is a complex system because it involves various components in the body. Behind this complex system there is a network that defines the interactions between components. With the network representation, the structure of the relationship among diseases can be known and can be analyzed. Exploring the network structure is an efficient approach to identifying the complex disease networks by identifying the highly connected individual nodes and the specific node communities (Barabási et al. 2011; Mu et al. 2020). Chen and Xu (Chen et al. 2015) explored and analyzed the comorbidities pattern in colorectal cancer. There is also relationship between hepatocellular carcinoma and medical comorbidities based on community detection in a comorbid network (Mu et al. 2020). Comorbid networks were grouped using a community detection algorithm and evaluated using disease-gene associations. The disease comorbidity network shows a genetic link between colorectal cancer and metabolic disorders. Community detection

from a network determines the cancer subtypes using multi-omics data (Nguyen et al. 2020). Human diseases frequently arise from protein dysfunction and can be expressed in the community. Tripathi (Tripathi et al. 2019) comprehensively assessed many classical community detection algorithms for biological networks to recognize non-overlapping communities and proposed a heuristic algorithm to identify structurally small and well-defined communities. The network and tree approach is a tool for inference in decision support systems related to comorbidities because it involves uncertainty in diagnosis and treatment (Capobianco and Liò 2015).

There are many community algorithms, however, not every algorithm is suitable for performing community clustering in all fields. An efficient approach to measuring a community quality is known as modularity (Newman 2006). Some algorithms give optimum results in one area but are less than optimum if applied to other problem areas. This paper investigates which community detection algorithm is more suitable for the health sector, especially for comorbidities determination. The network community is evaluated based on the value of modularity. In network community, besides modularity, it is essential to evaluate various characteristics reflected in the fitness function of the set of communities.

This study consists of five stages (1) data preprocessing, (2) develop a network based on the calculation of similarity between diseases, (3) determine communities using various algorithms and measure modularity, (4) determine significant comorbidities based on communities, and (5) determine various fitness functions that correlate with cluster formation. The contribution of our findings is to provide an alternative use of networks in biomedical problems, especially in determining comorbid lung cancer. We hope that this study will aid in the better understanding and management of diseases in the clinical context.

Materials and methods

Materials

Data acquisition

We searched the list of diseases through text mining of manuscripts in PubMed via Pubtator Central (PTC) <https://www.ncbi.nlm.nih.gov/research/pubtator/>. PTC performs automatic annotations to provide six bioconcepts: disease, gene, species, mutation, chemical, and cell line (Wei et al. 2019). We focus on the disease since our goal is to obtain lung cancer comorbidities.

Data preprocessing

We cleaned the data and identified the comorbidities. The first cleaning was to remove the words death, mortality, lung cancer, considering that they are not comorbid diseases of lung cancer. Furthermore, for each disease found in the text mining stage, a disease ontology search was conducted through <https://disease-ontology.org/> to find the DOID. Disease Ontology (DO) is a framework for describing gene products from a disease perspective. It is critical for supporting functional genomics in disease contexts. Accurate disease descriptions can lead to the discovery of novel links between genes and disease, as well as new functions for previously unknown genes and alleles. DO is structured as a directed acyclic graph, which lays the groundwork for quantitative disease knowledge

computing. For instance, pneumonia is a disease with DOID:552, which also has synonyms with acute pneumonia (Table 1). The aim was to determine the disease term. We also investigated for names based on their synonyms. Finally, we eliminated diseases that cannot be traced through disease ontologies ended up with 395 lung cancer comorbid diseases identified using DOID.

Methods

Develop a network based on the calculation of similarity between diseases

We calculated the similarity on the basis of the DO generated from the previous stage. Afterward, we created a similarity matrix using the R program, mainly the doSim function in the DOSE library downloaded from Bioconductor (Yu et al. 2015). There are five calculation algorithms in the doSim function as developed by Wang (Wang et al. 2007), Jiang (Jiang and Conrath 1997) , Lin (Lin 1998), Resnik (Grabowski 1995) and Rel (Schlicker et al. 2006). Wang computes the semantic similarity of two DO terms based on their positions in the DO directed acyclic graph and their relationships to their ancestor terms. Four other methods based on information content are based on the frequencies of two DO terms and their closest common ancestor term in a corpus of DO annotations. The negative log likelihood of a DO term occurring in the DO corpus is used to calculate the information content of the term. The weight/value of this similarity shows that these two comorbid disease terms have semantic relationship, phenotype characteristics, relationships between genes and disease, and related medical vocabulary disease concepts. The result of this stage was the comorbidity similarity matrix. The matrix elements ranged from 0 to 1, with 0 indicating that the two comorbidities were not identical and 1 indicating that they were. Then, the matrix was analyzed using the applied threshold. Finally, we constructed a network based on the similarity of the disease matrices, followed by building network formation using the Cytoscape application (Shannon et al. 2003).



Determine communities using various algorithms and measure modularity

Various algorithms determine the network community using the cdlb library (Rossetti et al. 2019). Among the 42 crisp discovery community algorithms, we used 20 of them, namely fluid (Parés et al. 2018), belief community (Zhang and Moore 2014), constant Potts model (CPM) (Traag et al. 2011), Chinese Whispers (Biemann 2006), diffusion entropy reducer (DER) (Kozdoba and Mannor 2015), Eigenvector (Newman 2006), expectation-maximization (EM) (Newman and Leicht 2007), genetic algorithm (GA) (Pizzuti 2008), Girvan Newman (Girvan and Newman 2002), greedy modularity

Table 1 Disease ontology data on pneumonia

| Metadata | Data |
|------------|---------------------------------|
| ID | DOID:552 |
| Name | Pneumonia |
| Alternates | DOID:10509 DOID:11742 DOID:5871 |
| Synonym | Acute pneumonia [EXACT] |

It has one main DOID and three alternates DOID. All of them refer to same disease

(Clauset et al. 2004), Kcut (Ruan and Zhang 2007), label propagation (Raghavan et al. 2007), Leiden (Traag et al. 2019), Louvain (Blondel et al. 2008), Markov clustering (Enright et al. 2002), RBER Pots (Reichardt and Bornholdt 2006), RB Pots (Leicht and Newman 2008), significance (Traag et al. 2013), spinglass (Reichardt and Bornholdt 2006), surprise (Traag et al. 2015), and walktrap (Pons and Latapy 2006). We use these 21 algorithms to investigate various algorithm approaches, i.e., propagation based, statistical inference, modularity maximization, minimum cut method and Girvan Newman approach. A description of each algorithmsm, type approach and limitation is found in Additional file 1.

We calculated and compared the modularity of the community outcomes formed by each algorithm. The higher the modularity, the better and optimal the community structure. There were five modularity algorithms: Newman Girvan (Newman and Girvan 2004), Erdos Renyi modularity (Erdos and Rényi 2011), link modularity (Nicosia et al. 2009), modularity density (Zhang et al. 2010), and Z modularity (Miyauchi and Kawase 2016). Then, we performed principal component analysis (PCA) to calculate the eigenvalues, which will be the weight of each modularity in calculating the overall modularity (Ramadhani et al. 2021). Nevertheless, PCA is usually used for dimensional reduction (Ahmadi et al. 2021). We sorted and selected the five best algorithms, each of which compared the results. At this stage, we prepared a clustering heatmap.

$$Q(S)_{NewmanGirvan} = \frac{1}{m} \sum_{c \in S} \left(m_s - \frac{(2m_s + l_s)^2}{4m} \right) \quad (1)$$

$$Q(S)_{ErdosRenyi} = \frac{1}{m} \sum_{c \in S} \left(m_s - \frac{(mn_s(n_s - 1))}{n(n - 1)} \right) \quad (2)$$

$$Q(S)_{LinkModularity} = \frac{1}{2m} \sum_{i,j \in V} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (3)$$

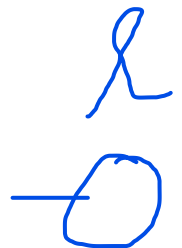
$$Q(S)_{ModularityDensity} = \sum_{c \in S} \frac{1}{n_c} \left(\sum_{i \in C} 2 * \lambda * k_{iC}^{int} - \sum_{i \in C} 2 * \lambda * k_{iC}^{out} \right) \quad (4)$$

$$Z(C)_{ZModularity} = \frac{\sum_{c \in C} \frac{m_c}{m} - \sum_{c \in C} (\frac{D_c}{2m})^2}{\sqrt{\sum_{c \in C} (\frac{D_c}{2m})^2 (1 - \sum_{c \in C} (\frac{D_c}{2m})^2)}} \quad (5)$$

where m is the number of graph edges, m_s is the number of community edges, l_s is the number of edges from nodes in S to nodes outside S , n_c is the number of nodes in C , K_{iC}^{int} is the degree of node i within C , K_{iC}^{out} are the degree of node i outside C , and λ is a parameter that allows for tuning of the measure resolution.

Determine significant comorbidities based on communities

The list of significant comorbid lung cancer in each community was calculated on the basis of centrality in each community formed, betweenness, degree, closeness, and



eigenvector centrality. The diseases found were relatively consistent in each community formed by algorithms such as label propagation, spinglass, Chinese whisper, Louvain, and RB POTS.

Determine various fitness functions that correlate with cluster formation

We compared the proximity level of the five algorithms by calculating fitness scores, such as average internal degree, internal edge density, edges inside, expansion (Radicchi et al. 2004), conductance (Shi and Malik 2000), cut ratio (Fortunato 2010), a fraction over median degree, triangle participation ratio, (Yang and Leskovec 2015), normalized cut (Shi and Malik 2000), max ODF, avg ODF, flake ODF (Flake et al. 2000), average embeddedness, average transitivity, scaled density, and size (Rossetti et al. 2019). Information regarding each fitness function can be found in Additional file 2. Additionally, we used PCA to determine the eigenvector (Gan and Djauhari 2012), representing the weight assigned to each fitness function when computing the overall fitness functions, and pick several fitness functions strongly related to the findings of the community algorithm.

Results

Data acquisition

We compiled the list of diseases by performing text mining on PubMed publications using Pubtator Central (PTC) <https://www.ncbi.nlm.nih.gov/research/pubtator/>. Text mining of manuscripts in PubMed search using the keywords “comorbid lung cancer” from PTC yielded 150 manuscripts (filter the full text) and 551 manuscripts (abstract). There is the name of lung cancer and other diseases that accompany each of these manuscripts; we take the comorbid disease from 551 manuscript. We found a list of 7183 disease names, with 1151 unique disease data. One of the manuscripts with PMID 34439135 obtained three disease names from the PTC automatic annotation results (Table 2).

Network based on the calculation of similarity between diseases

Following the 395 comorbid disease data with known DOID, the doSim function in the DOSE library calculated the matrix (Yu et al. 2015) to determine DO similarity. There are five calculation algorithms in the doSim: Wang, Jiang, Lin, Resnik, and Rel. According

Table 2 Automatic bioconcept annotation from Pubtator

| Metadata | Data |
|------------|--|
| PMID | 34439135 |
| Title | Current Treatment Strategies for Non-Small-Cell Lung Cancer with Comorbid Interstitial Pneumonia |
| Bioconcept | |
| Gene | – |
| Disease | Non-Small Cell Lung Cancer (4 occurrences), pneumonia (2), lung cancer (1) |
| Chemical | Carboplatin (1), paclitaxel (1) |
| Mutation | – |
| Species | Patients (4), honeycomb (1) |
| Cell line | – |

We are concerned about pneumonia disease, which is a comorbid lung cancer, not lung cancer

to the number of diseases, the calculation gives a symmetrical matrix size 395×395 . Each matrix element has a value range between 0 and 1, indicating a similarity level. The higher the value, the more each pair of diseases has a high similarity, and vice versa. At the time of the matrix calculation, 41 diseases did not have a matrix value. Hence, they are removed. A pair of identical illnesses should have a similarity value of 1. The results of calculations using Wang, Jiang, and Lin show that the diagonal of the matrix is worth 1. Table 3 presents an example of the data pieces for the Wang method.

Nevertheless, this is not the case with the Rel and Resnik methods. Neither method assigns a value of 1 to a pair of identical diseases. In the Rel method, diagonals contain values close to 1; e.g., in DOID 14667 (the value is 0.951) and 50117 (0.911). Even the Resnik method gives far from accurate results, for instance, on DOID 14667 (0.312), 50117 (0.250), 50127 (0.656), 50156 (0.886), and 9970 (0.799). Based on these considerations, the Rel and Resnik methods are removed for further calculations.

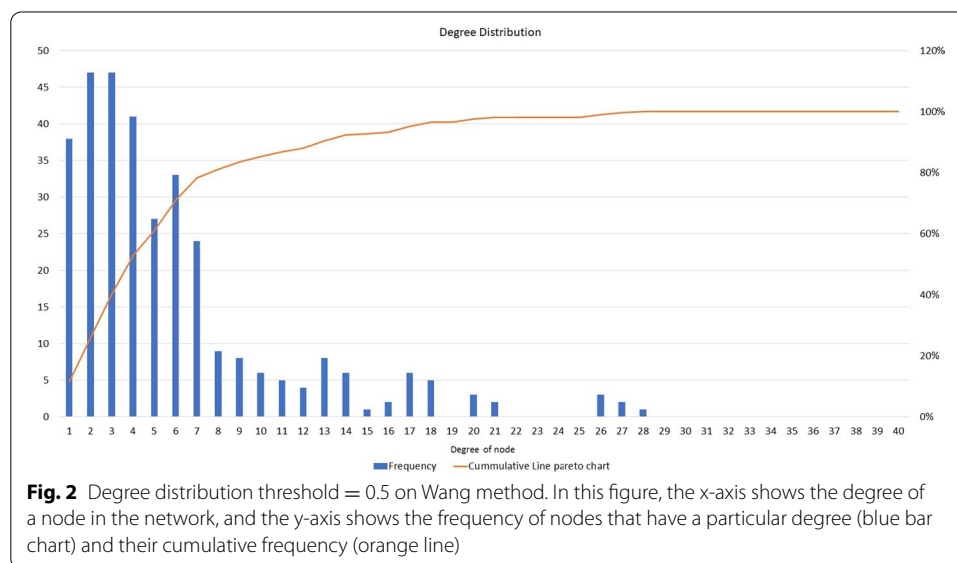
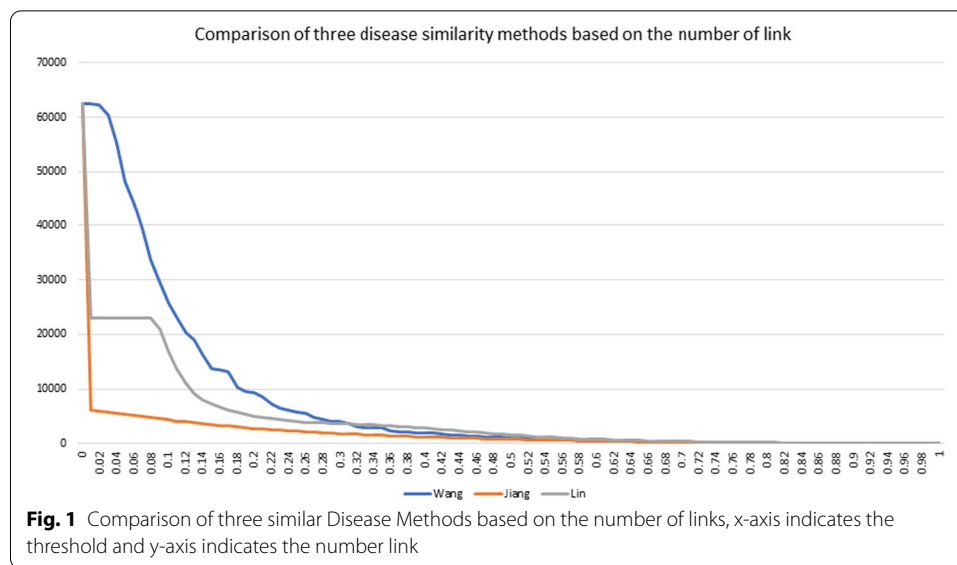
A graph/network was created on the basis of the the threshold value of the similarity matrix. The threshold value displays the connectivity of two nodes in a network from 0 to 1. If the matrix element value is above the threshold, the two nodes are connected, and vice versa. The threshold 0 indicates that all nodes connect to other nodes, with as many links as $n(n-1)/2$ where n represents the number of nodes. Threshold 1 causes all nodes to be disconnected and form a null graph. Figure 1 compares the number of links between calculations using the Jiang, Lin, and Wang methods. Among three approaches, the Wang's method seems more feasible because it forms a smooth and unbroken curve, where for a small threshold, there are many pairs of nodes connected to a link. Meanwhile, in the Jiang and Lin methods, the number of links suddenly drops drastically and breaks with a slight increase in the threshold.

Measuring disease similarity is based on functional associations between genes, and it is a disease data source for the building of biomedical databases. In the terminology graph, the similarity of the two diseases is represented by a link that connects the nodes of the two diseases. An edge connects two nodes because they have disease similarities calculated from the disease ontology. The more significant similarity between two diseases means that the more closely related they are, the more common information they have (Su et al. 2019). On the other hand, the smaller the similarity value, the less similarity between the two diseases. Moreover, a threshold value determines the link. The lower the threshold will result in a denser network. However, a

Table 3 Snippet of the matrix calculated using the Wang method

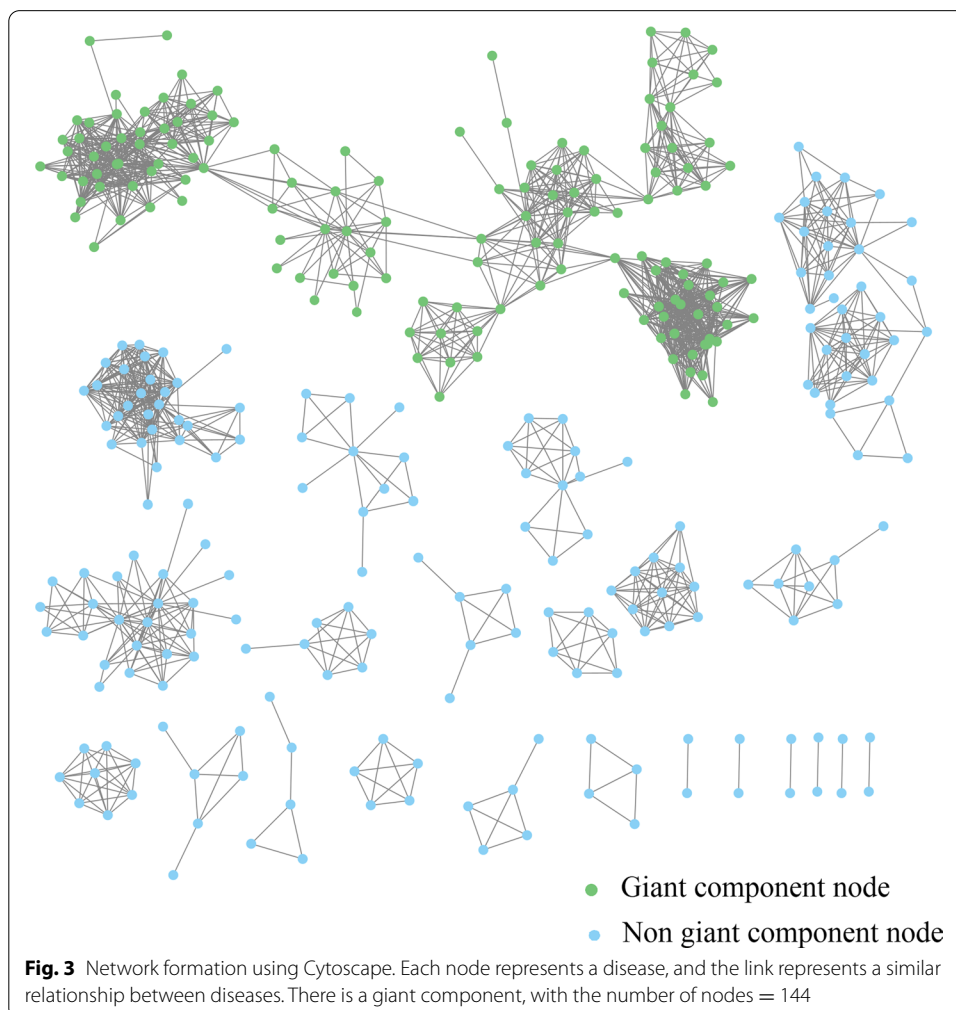
| DOID | 14667 | 50073 | 50117 | 50127 | 50152 | 50156 | ... | 9970 |
|-------|-------|-------|-------|-------|-------|-------|-----|-------|
| 14667 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0.562 |
| 50073 | 0 | 1 | 0.401 | 0 | 0 | 0 | ... | 0 |
| 50117 | 0 | 0.401 | 1 | 0 | 0 | 0 | ... | 0 |
| 50127 | 0 | 0 | 0 | 1 | 0.448 | 0.46 | ... | 0 |
| 50152 | 0 | 0 | 0 | 0.448 | 1 | 0.628 | ... | 0 |
| 50156 | 0 | 0 | 0 | 0.46 | 0.628 | 1 | ... | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 9970 | 0.562 | 0 | 0 | 0 | 0 | 0 | ... | 1 |

Rows and columns show the disease represented by DOID. The value in the matrix cell is the level of similarity. The higher the more similar



low threshold allows two diseases with low similarity to form a link (false positive). On the other hand, a high threshold will form fewer nodes clusters and give lower modularity. With these considerations, the moderate threshold used is 0.5 (Additional file 3). The following is a distribution of degrees with a threshold of 0.5, used in constructing the community networks (Fig. 2). Threshold minimizing the proportions of false-positive and false-negative (Bettembourg et al. 2015). A similarity threshold set to 0.5 filters low diseases similarities that do not well represent a link on a network (Zhao and Wang 2018). This threshold will change the network structure significantly. For example, if the threshold is 0 then the network will be a complete graph, whereas if the threshold is 1 then it will be a null graph.

A network was developed based on matrix similarity between diseases calculated using the Wang method. In the network, there are 338 nodes and 1609 edges; the average number of neighbors is 13,639, and the clustering coefficient is 0.796. The number of connected components, The subgraph in which every pair of nodes has a path connecting them, is 23, each of which contains 144, 35, 29, 25, 11, 11, 11, 8, 8, 7, 6, 6, 6, 5, 5, 5, 4, 2, 2, 2, 2, 2, and 2 nodes (Fig. 3). In this study, we selected the highest connected component, containing 144 nodes. The graph's largest connected components have a distinct community structure, as opposed to the second or third. This is accomplished by grouping nodes belonging to the largest components into nonoverlapping cohesive subgroups. Most identified groups are strong because each node collaborates with nodes from their group more frequently than with nodes from other groups (Savić et al. 2015). The highest modularity is the first most significant component, and the smaller the number of nodes, the lower the modularity value (Additional file 4). The disease group obtained in the first most significant component is heterogeneous, and the second and third largest components have clustered like a group of cancers other than lung and psychological disorders.



Network community and modularity

Communities were determined using the *cdlib* library (Rossetti et al. 2019). Algorithms include fluid (Parés et al. 2018), belief community (Zhang and Moore 2014), constant Potts model (CPM) (Traag et al. 2011), Chinese whispers (Biemann 2020), diffusion entropy reducer (DER) (Kozdoba and Mannor 2015), eigenvector (Newman 2006), expectation-maximization (EM) (Newman and Leicht 2007), genetic algorithm (GA) (Pizzuti 2008), Girvan Newman (Girvan and Newman 2002), greedy modularity (Clauset et al. 2004), Kcut (Ruan and Zhang 2007), label propagation (Raghavan et al. 2007), Leiden (Traag et al. 2019), Louvain (Blondel et al. 2008), Markov clustering (Enright et al. 2002), RBER Pots (Reichardt and Bornholdt 2006), RB Pots (Leicht and Newman 2008), significance (Traag et al. 2013), spinglass (Reichardt and Bornholdt 2006), surprise (Traag et al. 2015), and walktrap (Pons and Latapy 2006). Each algorithm generates a different community. For each community, we measured modularity using several modularity algorithms. A higher modularity measurement for a particular network indicates a better community structure. There are five modularity algorithms used, namely Newman Girvan (Newman and Girvan 2004), Erdos Renyi Modularity (Erdos and Rényi 2011), Link Modularity (Nicosia et al. 2009), Modularity density (Zhang et al. 2010), and Z Modularity (Miyachi and Kawase 2016) (Table 4). The modularity calculation formula is expressed in Eqs. (1)–(5).

The calculations and sorting based on each modularity formula produced a sequence of 20 different community algorithms and overall modularity calculated by PCA. The calculation gives an eigenvalue of 4.322 and eigenvector for each analysis of modularity Newman Girvan, Erdos Renyi, link modularity, modularity density, and Z modularity

Table 4 Modularity measurement

| Algorithm | Newman Girvan | Erdos Renyi | Link | Density | Z |
|-------------------|---------------|-------------|-------|---------|---------|
| Fluid | 0.548 | 0.510 | 0.121 | 24.106 | 1.207 |
| Belief | 0.701 | 0.760 | 0.140 | 59.499 | 1.580 |
| CPM | − 0.010 | 0.000 | 0.000 | − 1964 | − 0.098 |
| Chinese whispers | 0.711 | 0.773 | 0.135 | 70.804 | 1.715 |
| DER | 0.417 | 0.491 | 0.143 | 26.175 | 0.844 |
| Eigenvector | 0.707 | 0.779 | 0.140 | 65.678 | 1.601 |
| EM | 0.549 | 0.591 | 0.136 | 34.864 | 1.126 |
| ga | 0.649 | 0.723 | 0.124 | 54.424 | 1.587 |
| Girvan Newman | 0.703 | 0.762 | 0.142 | 59.117 | 1.571 |
| Greedy modularity | 0.678 | 0.711 | 0.132 | 61.066 | 1.593 |
| Kcut | 0.000 | 0.007 | 0.144 | 6.587 | 0.000 |
| Label propagation | 0.710 | 0.781 | 0.134 | 73.272 | 1.716 |
| Leiden | 0.707 | 0.757 | 0.135 | 66.314 | 1.683 |
| Louvain | 0.711 | 0.773 | 0.135 | 70.804 | 1.715 |
| Markov Clustering | 0.707 | 0.786 | 0.140 | 68.667 | 1.603 |
| RBER Pots | 0.702 | 0.790 | 0.139 | 63.449 | 1.597 |
| RB Pots | 0.711 | 0.773 | 0.135 | 70.804 | 1.715 |
| Significance | 0.622 | 0.719 | 0.118 | 5.528 | 1.567 |
| Spinglass | 0.710 | 0.781 | 0.134 | 73.272 | 1.716 |
| Surprise | 0.697 | 0.780 | 0.132 | 59.834 | 1.691 |
| Walktrap | 0.707 | 0.786 | 0.140 | 68.667 | 1.603 |

0.4681, -0.4669 , -0.4025 , -0.42912 , and -0.46538 , respectively. We select the best five out of the 20 community algorithms from these: Label Propagation, Spinglass, Chinese whisper, Louvain, and RB Pots. The five algorithms form different communities. Each algorithm clusters seven or nine groups. In Fig. 4, the left side expresses the complete nodes in each community, and the right side illustrates the relationship between each cluster. The best five algorithms is found in Additional file 5.

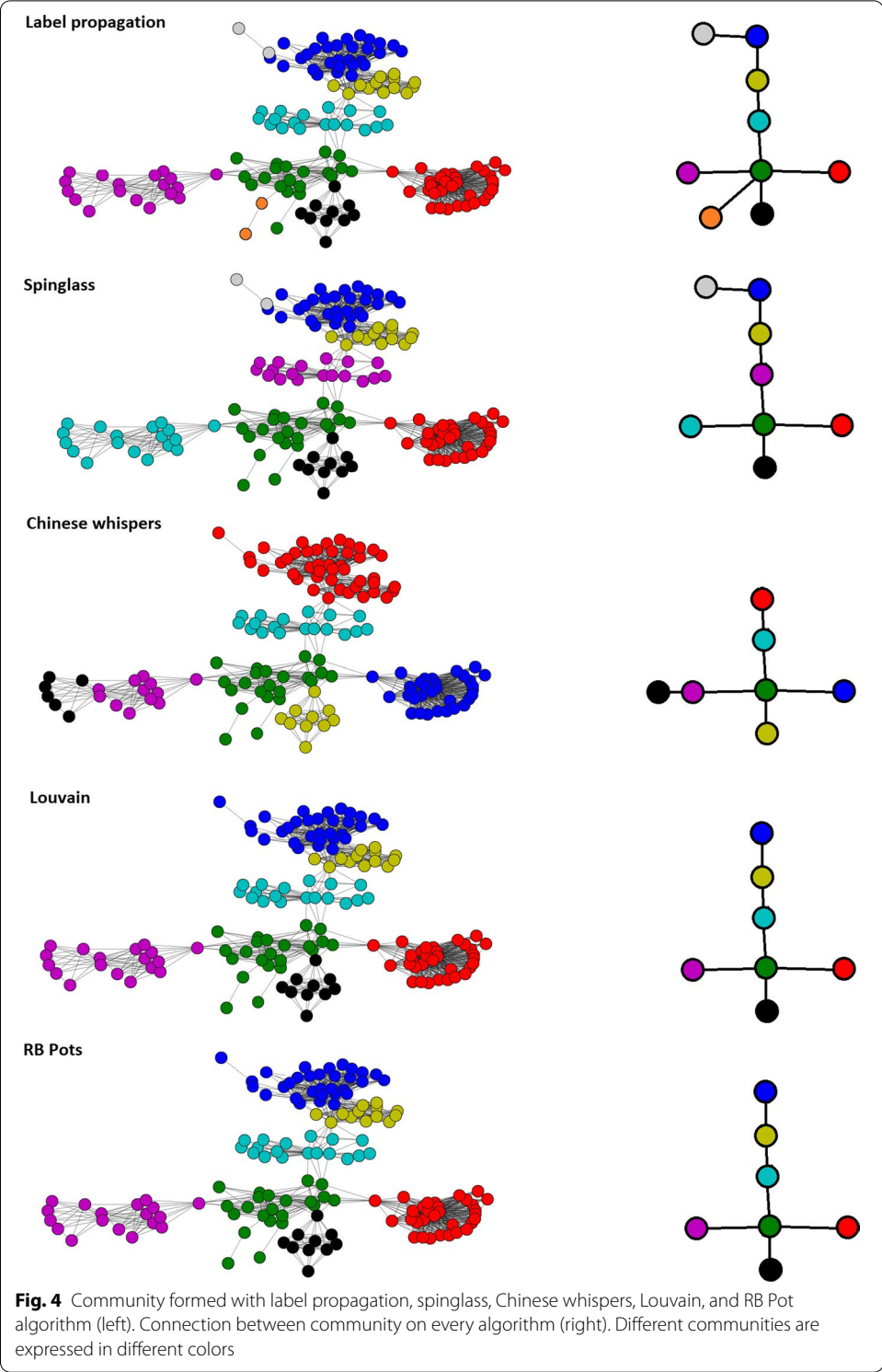
The label propagation algorithm divides the community into nine clusters associated with respiratory system disease, vascular disease, immune system disease, bone disease, metabolism disease, atrial heart septal defect, pancreatic disease, familial atrial fibrillation, and persistent generalized lymphadenopathy. The spinglass divides the community into eight clusters associated with interstitial lung disease, vascular disease, immune system disease, metabolism disease, bone disease, atrial heart septal defect, pancreatic disease, and familial atrial fibrillation. The Louvain and RB Pots algorithms also produced seven similar clusters associated with interstitial lung disease, vascular disease, immune system disease, bone disease, metabolism disease, atrial heart septal defect, and pancreatic disease. The essential diseases are diabetes mellitus, vascular disease, respiratory system disease, immune system disease, bone disease, diabetes mellitus, pancreatic disease, and familial hyperlipidemia. Cluster 2 relates to vascular becomes central in lung cancer comorbid diseases in every community algorithm applied.

List of comorbid diseases in each community

Our intention is to find diseases that is considered significant among existing diseases that has the greatest centrality in each community. The list of significant comorbid lung cancer in each community is calculated on the basis of centrality in each community formed based on betweenness, degree, closeness, and eigenvector centrality. The diseases found were relatively consistent in each community formed from community algorithms such as label propagation, spinglass, Chinese whisper, Louvain, and RB POTS. Conversely, vascular, immune system, disease, and pancreatic disease are commonly encountered based on differences in community algorithms and centrality (Table 5). These five algorithms can find community patterns that are relatively similar in finding significant comorbid diseases. Diseases that occur in a community, have similarities with each other. For example, in communities with respiratory/interstitial system diseases primary lung disease associated respiratory disease, emphysema and pneumonia (Fig. 5). Disease groups formed in each cluster can be further investigated for their relationship, especially on the similarity of symptoms, anatomy, cells, genes, phenotypes, and potential for treatment. It is important in relation to precision medicine for cancer comorbid patients, especially to improve diagnosis and safe therapy.

Determine various fitness functions

We calculated the fitness score, consisting of average embeddedness, average internal degree, average transitivity, conductance, cut ratio, edges inside, expansion, a fraction over median degree, internal edge density, normalized cut, max ODE, avg ODE, Flake ODE, scaled density, size, and triangle participation ratio (Table 6). The higher the fitness score, the better the results. Several fitness functions chosen are those with a significant relationship to the community. The calculation using PCA gives eigenvalue of 5.333



with each eigenvector of 0.447169; 0.447227; 0.447224; 0.447224; and 0.447224. Based on these results, the most significant fitness sequences that correlate with community formation are average internal degree, size and edges inside.

Table 5 List of significant diseases in each community, ordered according to nodes obtained from various centralities

| No | Betweenness centrality | Degree centrality | Closeness | Eigenvector |
|-------------|------------------------------|------------------------------|------------------------------|------------------------------|
| Label | Respiratory syst | Respiratory syst | Respiratory syst | Respiratory syst |
| Propagation | Vascular disease | Vascular disease | Vascular disease | Vascular disease |
| N = 9 | Severe combined immunodef. | Immune system | Immune system | Immune system |
| | Bone disease | Bone disease | Bone disease | Bone disease |
| | Disease of metabolism | Disease of metabolism | Disease of metabolism | Disease of metabolism |
| | Atrial heart septal defect | Atrial heart septal defect | Atrial heart septal defect | Atrial heart septal defect |
| | Pancreas disease | Pancreas disease | Pancreas disease | Pancreas disease |
| | Familial atrial fibrillation | Familial atrial fibrillation | Familial atrial fibrillation | Familial atrial fibrillation |
| | p gnrl. lymphadenopathy | p gnrl. lymphadenopathy | p gnrl. lymphadenopathy | p gnrl. lymphadenopathy |
| Spinglass | Interstitial lung disease | Interstitial lung disease | Interstitial lung disease | Interstitial lung disease |
| N = 8 | Vascular disease | Vascular disease | Vascular disease | Vascular disease |
| | Immune system | Immune system | Immune system | Immune system |
| | Disease of metabolism | Disease of metabolism | Disease of metabolism | Disease of metabolism |
| | Bone disease | Bone disease | Bone disease | Bone disease |
| | Atrial heart septal defect | Atrial heart septal defect | Atrial heart septal defect | Atrial heart septal defect |
| | Pancreas disease | Pancreas disease | Pancreas disease | Pancreas disease |
| | Familial atrial fibrillation | Familial atrial fibrillation | Familial atrial fibrillation | Familial atrial fibrillation |
| Chinese | Cardiovascular system | Vascular disease | Vascular disease | Vascular disease |
| Whisper | Respiratory system | Respiratory system | Respiratory system | Respiratory system |
| N = 7 | Immune system disease | Immune system disease | Immune system disease | Immune system disease |
| | Bone disease | Bone disease | Bone disease | Bone disease |
| | Diabetes mellitus | Diabetes mellitus | Diabetes mellitus | Diabetes mellitus |
| | Pancreas disease | Pancreas disease | Pancreas disease | Pancreas disease |
| | Familial hyperlipidemia | Familial hyperlipidemia | Familial hyperlipidemia | Familial hyperlipidemia |
| Louvain | Interstitial lung disease | Interstitial lung disease | Interstitial lung disease | Interstitial lung disease |
| N = 7 | Cardiovascular system | Vascular disease | Vascular disease | Vascular disease |
| | Immune system | Immune system | Immune system | Immune system |
| | Bone disease | Bone disease | Bone disease | Bone disease |
| | Disease of metabolism | Disease of metabolism | Disease of metabolism | Disease of metabolism |
| | Atrial heart septal defect | Atrial heart septal defect | Atrial heart septal defect | Atrial heart septal defect |
| | Pancreas disease | Pancreas disease | Pancreas disease | Pancreas disease |
| RB POTS | Interstitial lung disease | Interstitial lung disease | Interstitial lung disease | Interstitial lung disease |
| N = 7 | Cardiovascular system | Vascular disease | Vascular disease | Vascular disease |
| | Immune system | Immune system | Immune system | Immune system |
| | Bone disease | Bone disease | Bone disease | Bone disease |
| | Disease of metabolism | Disease of metabolism | Disease of metabolism | Disease of metabolism |
| | Atrial heart septal defect | Atrial heart septal defect | Atrial heart septal defect | Atrial heart septal defect |
| | Pancreas disease | Pancreas disease | Pancreas disease | Pancreas disease |

9

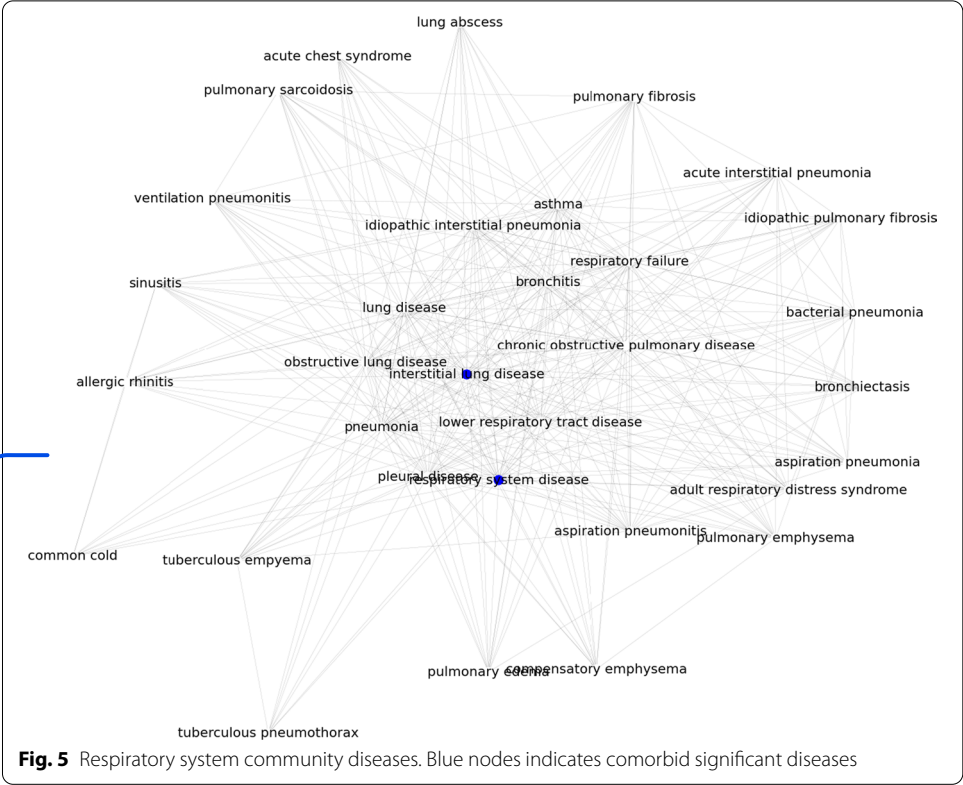


Table 6 Fitness score obtained from various fitness functions using each algorithm

| Fitness score | Label propagation | Spinglass | Chinese whispers | Louvain | RB Pots |
|------------------------------|-------------------|-----------|------------------|---------|---------|
| Average embeddedness | 0.888 | 0.721 | 0.929 | 0.929 | 0.929 |
| Average internal degree | 9.125 | 8.113 | 11.198 | 11.198 | 11.198 |
| Average transitivity | 0.667 | 0.601 | 0.838 | 0.838 | 0.838 |
| Conductance | 0.148 | 0.3 | 0.094 | 0.094 | 0.094 |
| Cut ratio | 0.008 | 0.008 | 0.009 | 0.009 | 0.009 |
| Edges inside | 101.333 | 91.1 | 130.571 | 130.571 | 130.571 |
| Expansion | 0.984 | 1.04 | 1.083 | 1.083 | 1.083 |
| Fraction over median dgr | 0.27 | 0.243 | 0.364 | 0.364 | 0.364 |
| Internal edge density | 0.709 | 0.538 | 0.603 | 0.603 | 0.603 |
| Normalized cut | 0.157 | 0.308 | 0.105 | 0.105 | 0.105 |
| Max ODF | 6.333 | 5.8 | 7.857 | 7.857 | 7.857 |
| AVG ODF | 0.984 | 1.04 | 1.083 | 1.083 | 1.083 |
| Flake ODF | 0.009 | 0.208 | 0.011 | 0.011 | 0.011 |
| Scaled density | 7.437 | 5.645 | 6.322 | 6.322 | 6.322 |
| Size | 16 | 14.4 | 20.571 | 20.571 | 20.571 |
| Triangle participation ratio | 0.772 | 0.695 | 0.971 | 0.971 | 0.971 |

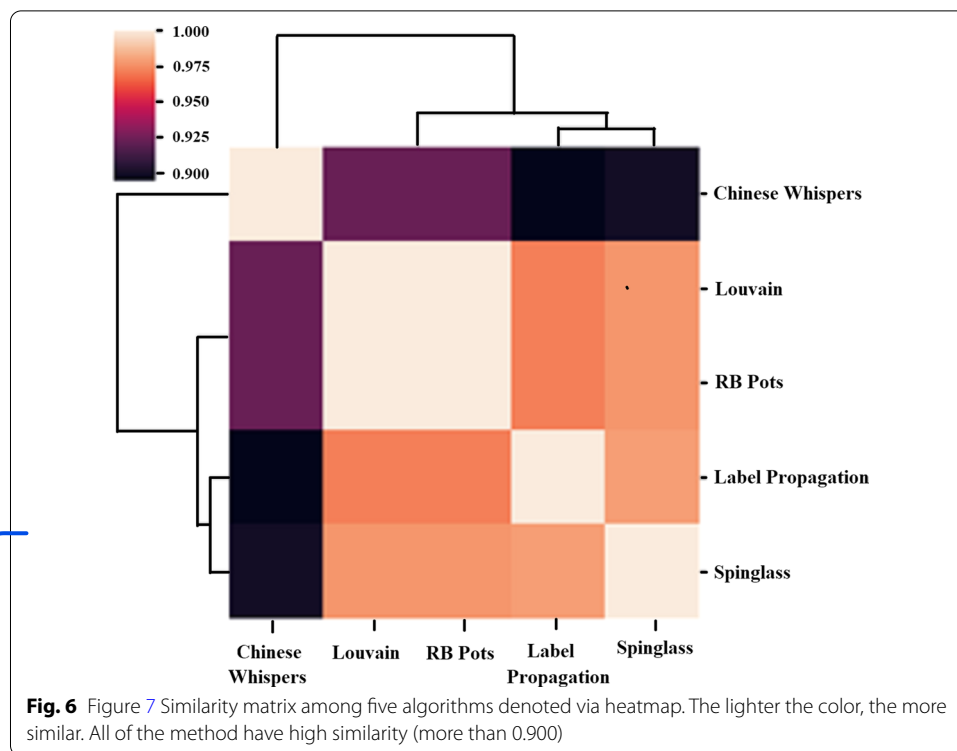
The higher the fitness score, the better the results

Discussion

Community algorithms comparison

Community algorithms heatmap reveals the closeness between algorithms (Fig. 6). In

6



this case, the label propagation algorithm has similarities with the spinglass, and the RB Post algorithm is similar to the results of the Louvain algorithm. The most different is the Chinese whisper algorithm.

Intersection result of comorbid disease among algorithms

A summary is visualized by Venn diagram using InteractiVenn (Heberle et al. 2015) based on the list of comorbidities, using the ensemble vote majority method. The numbers in this Venn diagram are the number of comorbid diseases produced by each community algorithm. The results are shown in Fig. 7, and the disease details are presented in Table 7. Every row shows the diseases identified by a particular algorithm in this table. For example, in the first row, all of the community algorithms found that vascular disease, immune system disease, bone disease, pancreas disease is significant comorbid lung cancer.

According to the existing references, these algorithms have succeeded in detecting various significant comorbid. Based on the DOID hierarchical structure, the results can be seen based on the DOID structure that have group/upper-level organization of diseases as in Additional file 6. The major comorbidity in patients with lung cancer is cardiovascular, approximately 23% (Pavia et al. 2007). The immune system dysregulation associated with autoimmune diseases increases the risk of cancer. Standardized incidence, standardized mortality, and hazard ratios indicated an increased risk of lung cancer (Hemminki et al. 2012). Bone metastases diseases are common in patient with lung cancer and have shorter overall survival (Kuchuk et al. 2013). Pancreatic metastases are found in advanced small cell lung cancer. In autopsy studies, pancreatic metastasis occurs between 1.6 and 10.6%. The primary tumor is usually in the left lung, and 15% of

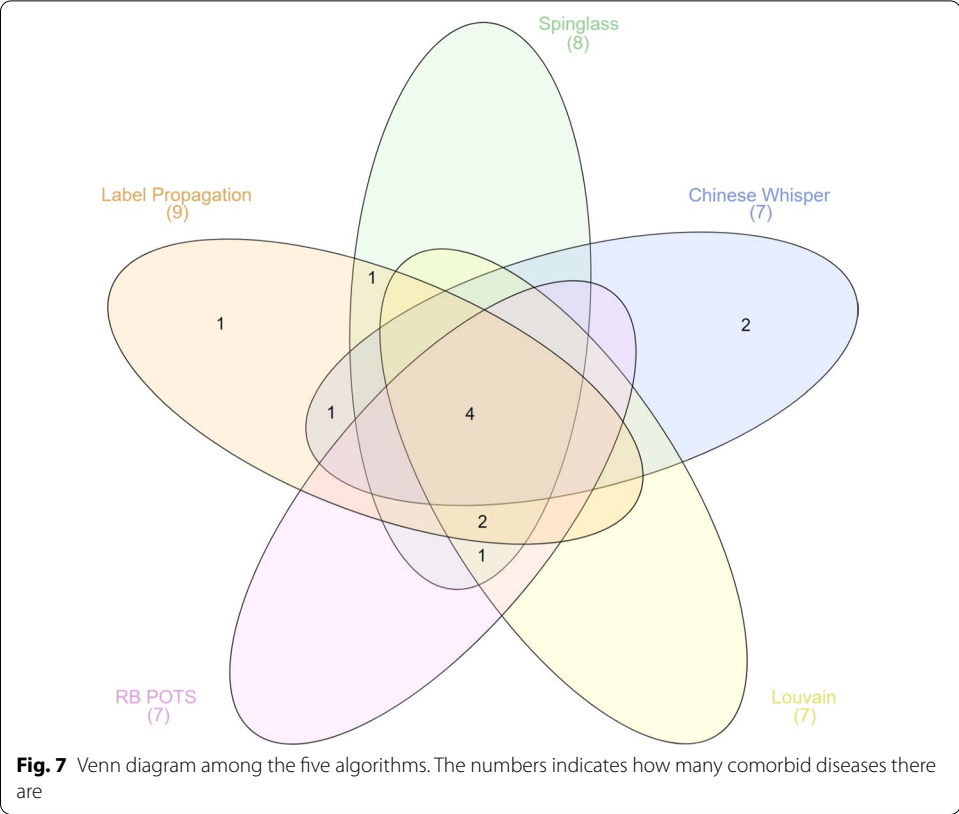


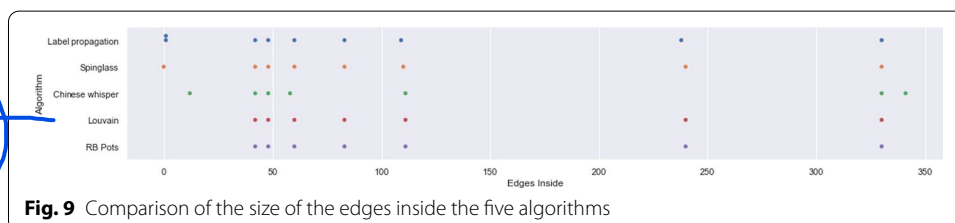
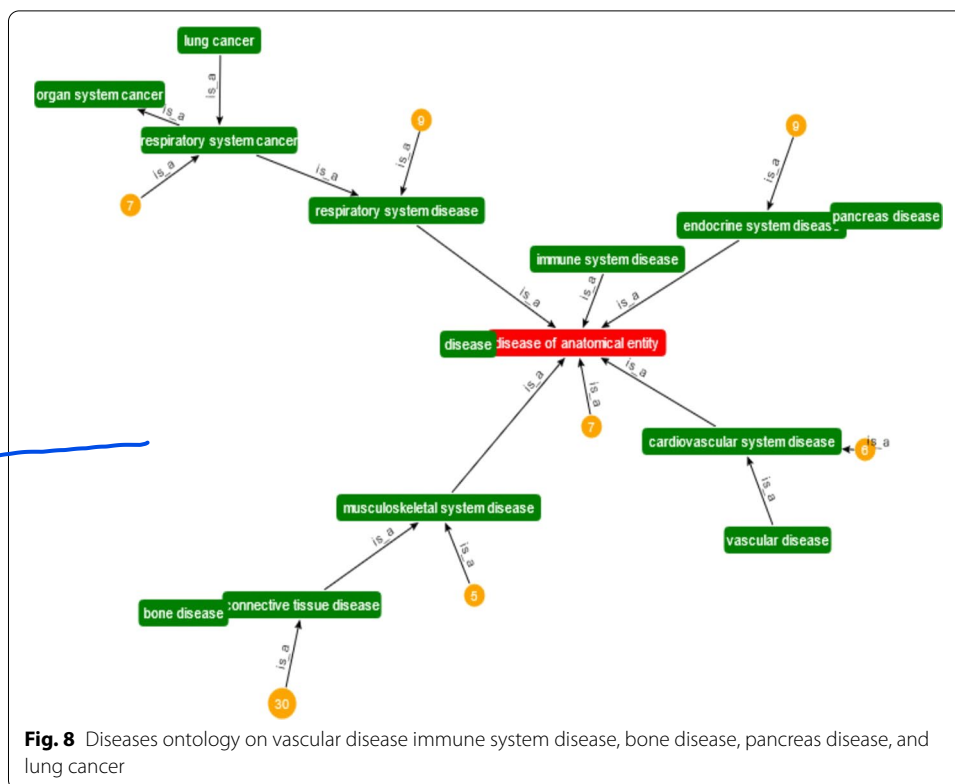
Table 7 Algorithm and significant disease

| Algorithm | Disease |
|---|---|
| Label Propagation and Spinglass and Chinese whisper and Louvain and RB POTS | Vascular disease, immune system disease, bone disease, pancreas disease |
| Label Propagation and Spinglass and Louvain and RB POTS | Disease of metabolism, atrial heart septal defect |
| Spinglass and Louvain and RB POTS | Interstitial lung disease |
| Label Propagation and Spinglass | Familial atrial fibrillation |
| Label Propagation and Chinese whisper | Respiratory system disease |
| Chinese whisper | Diabetes mellitus, familial hyperlipidemia |
| Label Propagation | Persistent generalized lymphadenopathy |

There are various lists of the same disease found in different algorithms

patients have pancreatic metastasis (Gonlugur et al. 2014). There is a lipid metabolism disorder in lung cancer (Merino Salvador et al. 2017). When patient with lung cancer have comorbid interstitial lung disease, the average survival at diagnosis is worse than without comorbidities (Margaritopoulos et al. 2017). Among 159,615 patients diagnosed with lung cancer in 2016, 10,050 (6.29%) patients had a concurrent diagnosis of atrial fibrillation (Bandyopadhyay et al. 2019). These patients frequently have tobacco-related illnesses (e.g., respiratory diseases) due to the much higher incidence of lung cancer in smokers and ex-smokers. (Leduc et al. 2017).

Conversely, patients with diabetes mellitus who have lung cancer have a higher survival rate than those without (Hatlen et al. 2011). Additionally, comorbid



hyperlipidemia is associated with a significant reduction in mortality in patients with lung cancer (Lazzarini et al. 2016). Specifically, COPD is a disease often found in comorbid cancer in the respiratory system disease group. Nevertheless, the community system limitations used that do not involve the prevalence of comorbid disease occurrence and its severity can be attached to the weight of nodes in the network. Structure of Directed Acyclic Graph in vascular, immune system, bone, pancreas disease, and Lung cancer can be described in terms of its relationship to the disease ontology as shown in the Fig. 8. All these diseases are a group of disease of anatomical entities.

Finally, the fitness functions in average internal degree, size, and edges inside correlate with the grouping between community algorithms by justifying the results. By a swarm plot, the Louvain and RB Pots algorithms have similar results in comparing the size of the edges inside, while label propagation and spinglass have identical results (Fig. 9). The size comparison of the five algorithms shows that the Chinese

whisper algorithm has significant differences from the results of the other algorithms (Fig. 10). Nevertheless, the Chinese whisper algorithm is close to the Louvain and RB Pots algorithms (Fig. 11).

Network Analysis has been used by Folino et al. (2010), to predict the risk of comorbid diseases suffered by patients and use association rules. reveal the comorbid network of occurrence of comorbidity. Ljubic et al. (2020) also conducted a network analysis to obtain genes that are associated with colorectal cancer and its comorbidities. Chmiel et al. (2014) conducted a research on Spreading of diseases through comorbidity networks across life and gender. However, the three studies used network analysis, but did not use community so that they could not reveal groups of diseases that have closeness which is the hallmark of this study. Table 8 provides a comparison among the research.

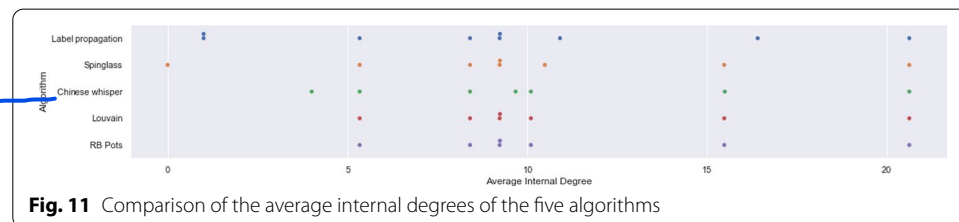


Table 8 Comparison of comorbidity network studies

| Author | Data | Disease | Network constructed | Analysis |
|---------------------------------|--|-------------------|---|--|
| Folino et al. (2010) | Medical records of 1462 patients | Common | Occurrence in patient and relative risk | Association rule |
| Ljubic et al. (2020) | SID California inpatient database (ICD-9) | Colorectal cancer | ϕ -correlation and Relative Risk (RR) | Centrality measurement |
| Chmiel et al. (2014) | Database of the Main Association of Austrian Social Security Institutions and Text Mining Pubmed | Common | Statistical multiplex network | Evolution disease network |
| Renteria-ramos et al. (2018) | Three independent administrative databases Risaralda province (2011–2016) | Common | k-Communities | Intensity analysis and motif coherence |
| Moratalla-Navarro et al. (2020) | 285,342 patients in Catalonia, Spain, (period: 2006–2017) | Hypo thyroidism | Comorbidity networks using logistic regression models | Multivariate logistic regression with LASSO |
| Our research | Pubtator text mining | Lung Cancer | Disease Similarity | Community Network and Centrality measurement |

The contribution of this study is that we can grouping diseases in communities and investigate community algorithms. While the limitations of this study are that it has not considered the background factors of the patient, the frequency of occurrence and severity of comorbidities as well as gene/protein interactions of each cancer comorbid disease.

Conclusions and suggestions

In this study, a disease network has been developed on the basis of the similarity of disease ontologies. In determining DO similarity, Wang algorithm performs better than Jiang, Lin, Resnik, and Rel algorithms, with a degree distribution threshold of 0.5. Modularity relevant in grouping comorbid lung cancer is label propagation, spinglass, Chinese whisper, Louvain, RB Pots, marked from five modularity measurements: Newman & Girvan, Erdos & Renyi, link modularity, modularity density, and Z Modularity. The calculation of the fitness score related to the modularity algorithm is the average internal degree, size, and edges inside. As determined, the significant comorbidities are vascular disease, immune system disease, bone disease, pancreatic disease (based on four algorithms); disease of metabolism and atrial heart septal defect (3); familial atrial fibrillation respiratory system disease and interstitial lung disease (2); diabetes mellitus, familial hyperlipidemia, and persistent generalized lymphadenopathy (1).

The investigation should be continued by searching disease-related genes that have been determined, represented in a multilayer community network, and looking for overlapping communities based on these genes and the diseases affecting them. For example, it is possible to look for candidate drugs, especially herbal drugs, for various significant comorbid diseases in lung cancer from the corresponding genes. This series of work will support the Sustainable Development Goals, especially for good health and well-being.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1007/s41109-022-00466-y>.

Additional file 1. Description of 21 algorithm used in the study.

Additional file 2. Fitness Function and explanation.

Additional file 3. The experiment for threshold was carried out using the Label Propagation, Chinese Whispers and RB Pots algorithms.

Additional file 4. Modularity on graph component.

Additional file 5. Explanation of five Community Detection Algorithm selected (Label propagation, Louvain, Spinglass, Chinese Whispers and RB Pots).

Additional file 6. Group diseases in Disease Ontology.

Acknowledgements

Thank you to the Tropical Biopharmaca Research Center (TropBRC) and Directorate of Scientific Publications and Strategic Information IPB University for providing supports and facilities in this research. Thanks also to Suminar Setiati Achmadi who has provided input on this manuscript from the aspect of language editing or grammar editing.

Author contributions

Conceptualization, WAK, HCR, YSS, AAP; methodology, HCR, and WAK; software, HCR; resources, WAK; data curation, HCR, YSS and AAP; writing-original draft preparation, HCR; writing-review and editing, HCR, YSS, AAP, WAK, SNI, INB, TD; visualization, HCR; supervision, WAK, SNI, INB, TD; project administration, WAK, INB. All authors have read and agreed to the published version of the manuscript. We attest that all authors contributed significantly to the creation of this manuscript. All named authors had substantial contributions to the conception of the work, the acquisition, analysis, and interpretation of data for the work. All authors wrote, read and approved the final manuscript. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare that they have no competing interests.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Author details

¹Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia.

²Department of Mathematics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia. ³Department of Chemistry, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia. ⁴Tropical Biopharmaceutical Research Center, IPB University, Bogor, Indonesia. ⁵Department of Agroindustrial Technology, Faculty of Agricultural Engineering and Technology, IPB University, Bogor, Indonesia. ⁶Department of Informatics, Faculty of Industrial Technology, UPN "Veteran" Yogyakarta, Yogyakarta, Indonesia. ⁷Department of Informatics Engineering, Institut Teknologi Indonesia, Tangerang Selatan, Indonesia. ⁸Department of Informatics, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia.

Received: 9 November 2021 Accepted: 22 April 2022

Published online: 18 May 2022

References

- Ahmadi M, Sharifi A, Jafarian Fard M, Soleimani N (2021) Detection of brain lesion location in MRI images using convolutional neural network and robust PCA. *Int J Neurosci* 1–12. ISSN 15635279. <https://doi.org/10.1080/00207454.2021.1883602>
- Bandyopadhyay D, Ball S, Hajra A, Chakraborty S, Dey AK, Ghosh RK, Palazzo AM (2019) Impact of atrial fibrillation in patients with lung cancer: insights from National Inpatient Sample. *UC Heart Vasc* 22:216–217. ISSN 23529067. <https://doi.org/10.1016/j.ijcha.2019.02.012>
- Bang UC, Benfield T, Hyldstrup L, Bendtsen F, Beck Jensen JE (2014) Mortality, cancer, and comorbidities associated with chronic pancreatitis: a Danish nationwide matched-cohort study. *Gastroenterology* 146(4):989–994.e1. ISSN 15280012. <https://doi.org/10.1053/j.gastro.2013.12.033>
- Barabási A-L (2016) *Network science*. Cambridge University Press, Cambridge
- Barabási AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. ISSN 14710056. <https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=78650373804&origin=inward>
- Buttembourg C, Diot C, Dameron O (2015) Optimal threshold determination for interpreting semantic similarity and particularity: application to the comparison of gene sets and metabolic pathways using GO and ChEBI. *PLoS ONE* 10(7). ISSN 19326203. <https://doi.org/10.1371/journal.pone.0133579>
- Biemann C (2006) Chinese Whispers—an efficient graph clustering algorithm and its application to natural language processing problems. In: *Proceedings of TextGraphs: the first workshop on graph based methods for natural language processing*. Association for Computational Linguistics, New York City, pp 73–80. <https://aclanthology.org/W06-3812>
- Biemann C (2020) Chinese whispers—an efficient graph clustering algorithm and its application to natural language processing problems. In: *Proceedings of TextGraphs: the 1st workshop on graph-based methods for natural language processing*, (June), pp 73–80
- Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008(10):1–12. ISSN 17425468. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Capobianco E, Liò P (2015) Comorbidity networks: beyond disease correlations. *J Complex Netw* 3(3):319–332, 01. ISSN 2051-1310. <https://doi.org/10.1093/comnet/cnu048>
- Chen Y, Li L, Xu R (2015) Disease comorbidity network guides the detection of molecular evidence for the link between colorectal cancer and obesity. In: *AMIA joint summits on translational science proceedings. AMIA joint summits on translational science*, vol 2015, pp 201–206. ISSN 2153-4063. <http://www.ncbi.nlm.nih.gov/pubmed/26306270%0A>. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4525229>
- Chmiel A, Klimek P, Thurner S (2014) Spreading of diseases through comorbidity networks across life and gender. *New J Phys* 16. ISSN 13672630. <https://doi.org/10.1088/1367-2630/16/11/115013>
- Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Top* 70(6):6. ISSN 1063651X. <https://doi.org/10.1103/PhysRevE.70.066111>
- Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7):1575–1584. ISSN 03051048. <https://doi.org/10.1093/nar/30.7.1575>
- Erdos P, Rényi A (2011) On the evolution of random graphs. *Struct Dyn Netw* 9781400841:38–82. <https://doi.org/10.1515/9781400841356.38>
- Feng J, Mu XM, Ma LL, Wang W (2020) Comorbidity patterns of older lung cancer patients in Northeast China: an association rules analysis based on electronic medical records. *Int J Environ Res Public Health* 17(23):1–14. ISSN 16604601. <https://doi.org/10.3390/ijerph17239119>
- Flake GW, Lawrence S, Giles CL (2000) Efficient identification of web communities. In: *Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '00*. Association for Computing Machinery, New York, pp 150–160. ISBN 1581132336. <https://doi.org/10.1145/347090.347121>
- Folino F, Pizzuti C, Ventura M (2010) A comorbidity network approach to predict disease risk. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, volume 6266 LNCS. Springer, Berlin, pp 102–109. ISBN 3642150195. https://doi.org/10.1007/978-3-642-15020-3_10

- Fortunato S (2010) Community detection in graphs. *Phys Rep* 486(3):75–174. ISSN 0370-1573. <https://doi.org/10.1016/j.physrep.2009.11.002>. <https://www.sciencedirect.com/science/article/pii/S0370157309002841>
- Fowler H, Belot A, Ellis L, Maringe C, Luque-Fernandez MA, Njagi EN, Navani N, Sarfati D, Rachet B (2020) Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. *BMC Cancer* 20(1):1–15. ISSN 14712407. <https://doi.org/10.1186/s12885-019-6472-9>
- Gan SL, Djauhari MA (2012) An overall centrality measure: the case of U.S. stock market. *Int J Basic Appl Sci* 12(06):99–104
- Girvan M, Newman ME (2002) Community structure in social and biological networks. *Proc Natl Acad Sci U S A* 99(12):7821–7826. ISSN 00278424. <https://doi.org/10.1073/pnas.122653799>
- Gonlugur U, Mirici A, Karaayvaz M (2014) Pancreatic involvement in small cell lung cancer. *Radiol Oncol* 48(1):11–19. ISSN 1318-2099. <https://doi.org/10.2478/raon-2013-0022>. <https://pubmed.ncbi.nlm.nih.gov/24587774>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3908842/>
- Grabowski HG (1995) Using information content to evaluate semtic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on artificial intelligence, vol 7, No. 1, pp 541–559
- Hatlen P, Grønberg BH, Langhammer A, Carlsen SM, Amundsen T (2011) Prolonged survival in patients with lung cancer with diabetes mellitus. *J Thorac Oncol* 6(11):1810–1817. ISSN 15561380. <https://doi.org/10.1097/JTO.0b013e31822a75be>
- Heberle H, Meirelles VG, da Silva FR, Telles GP, Minghim R (2015) InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinform* 16(1):1–7. ISSN 14712105. <https://doi.org/10.1186/s12859-015-0611-3>
- Hemminki K, Liu X, Ji J, Sundquist J, Sundquist K (2012) Effect of autoimmune diseases on risk and survival in histology-specific lung cancer. *Eur Respir J* 40(6):1489–1495. ISSN 1399-3003 (Electronic). <https://doi.org/10.1183/09031936.00222911>
- Hevey D (2018) Network analysis: a brief overview and tutorial. *Health Psychol Behav Med* 6(1):301–328. ISSN null. <https://doi.org/10.1080/21642850.2018.1521283>
- Huang W-Y, Li C-H, Lin C-L, Liang J-A (2016) Long-term statin use in patients with lung cancer and dyslipidemia reduces the risk of death. *Oncotarget* 7(27):42208
- Inafuku K, Morohoshi T, Adachi H, Koumori K, Masuda M (2016) Thoracoscopic lobectomy for lung cancer in a patient with a partial anomalous pulmonary venous connection: a case report. *J Cardiothorac Surg* 11(1):77–79. ISSN 17498090. <https://doi.org/10.1186/s13019-016-0527-7>
- Jacob S, Rahbari K, Tegtmeyer K, Zhao J, Tran S, Helenowski I, Zhang H, Walunas T, Varga J, Dematte J, Villafior V (2020) Lung cancer survival in patients with autoimmune disease. *JAMA Netw Open* 3(12):e2029917. ISSN 2574-3805. <https://doi.org/10.1001/jamanetworkopen.2020.29917>
- Jiang JJ, Conrath, DW (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th research on computational linguistics international conference, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), pp 19–33. <https://aclanthology.org/O97-1002>
- Kozdoba M, Mannor S (2015) Community detection via measure space embedding. In: Advances in neural information processing systems, 2015-January, pp 2890–2898. ISSN 10495258
- Kuchuk M, Addison CL, Clemons M, Kuchuk I, Wheatley-Price P (2013) Incidence and consequences of bone metastases in lung cancer patients. *J Bone Oncol* 2(1):22–29. ISSN 2212-1374. <https://doi.org/10.1016/j.jbo.2012.12.004>. <https://www.sciencedirect.com/science/article/pii/S2212137413000031>
- Lazzarini E, Carter P, De Boer M, Balbi C, Altieri P, Pfeffer U, Gambini E, Varesio L, Bosco M, Coviello D, Pompilio G, Brunelli C, Cancedda R, Ameri P, Bollini S, McGowan J, Uppal H, Chandran S, Sarma J, Potluri R, Octavia Y, De Kleijnen M, Van Thiel B, Ridwan Y, Te Lintel Hekkert M, Van Der Pluijm I, Essers J, Hoeijmakers J, Duncker D (2016) Mechanisms of cancer-related cardiomyopathy. *Cardiovasc Res* 111(suppl 1):S14–S15. ISSN 0008-6363. <https://doi.org/10.1093/cvr/cvw130>
- Leduc C, Antoni D, Charloux A, Falcoz PE, Quoix E (2017) Comorbidities in the management of patients with lung cancer. *Eur Respir J* 49(3). ISSN 13993003. <https://doi.org/10.1183/13993003.01721-2016>
- Leicht EA, Newman MEJ (2008) Community structure in directed networks. *Phys Rev Lett* 100(11):118703. ISSN 0031-9007 (Print). <https://doi.org/10.1103/PhysRevLett.100.118703>
- Li J, Gong B, Chen X, Liu T, Wu C, Zhang F, Li C, Li X, Rao S, Li X (2011) DOSim: an R package for similarity between diseases based on Disease Ontology. *BMC Bioinform* 12(1):1–10. ISSN 14712105. <https://doi.org/10.1186/1471-2105-12-266>
- Lin D (1998) An information-theoretic definition of similarity. In: ICML, pp 296–304
- Ljubic B, Pavlovski M, Alshehri J, Roychoudhury S, Bajic V, Van Neste C, Obradovic Z (2020) Comorbidity network analysis and genetics of colorectal cancer. *Inform Med Unlocked* 21:100492. ISSN 23529148. <https://doi.org/10.1016/j.imu.2020.100492>
- Loe CW, Jensen HJ (2015) Comparison of communities detection algorithms for multiplex. *Phys A Stat Mech Appl* 431(June 2014):29–45. ISSN 03784371. <https://doi.org/10.1016/j.physa.2015.02.089>
- Margaritopoulos GA, Antoniou KM, Wells AU (2017) Comorbidities in interstitial lung diseases. *Eur Respir Rev* 26(143):1–15. ISSN 16000617. <https://doi.org/10.1183/16000617.0027-2016>
- Merino Salvador M, Gómez de Cedrón M, Moreno Rubio J, Falagán Martínez S, Sánchez Martínez R, Casado E, Ramírez de Molina A, Sereno M (2017) Lipid metabolism and lung cancer. *Crit Rev Oncol Hematol* 112:31–40. ISSN 1040-8428. <https://doi.org/10.1016/j.critrevonc.2017.02.001>. <https://www.sciencedirect.com/science/article/pii/S1040842817300513>
- Miyauchi A, Kawase Y (2016) Z-score-based modularity for community detection in networks. *PLoS ONE* 11(1):1–17. ISSN 19326203. <https://doi.org/10.1371/journal.pone.0147805>
- Moratalla-Navarro F, Moreno V, López-Simarro F, Aguado A (2020) MorbiNet study hypothyroidism comorbidity networks in the adult general population. *J Clin Endocrinol Metab* 106(3):e1179–e1190, 12. ISSN 0021-972X. <https://doi.org/10.1210/clinem/dgaa927>
- Mu XM, Wang W, Jiang YY, Feng J (2020) Patterns of comorbidity in hepatocellular carcinoma: a network perspective. *Int J Environ Res Public Health* 17(9). ISSN 16604601. <https://doi.org/10.3390/ijerph17093108>
- Newman ME (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E Stat Non-linear Soft Matter Phys* 74(3). ISSN 15502376. <https://doi.org/10.1103/PhysRevE.74.036104>

- Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlinear Soft Matter Phys* 69(2 Pt 2):26113. ISSN 1539-3755 (Print). <https://doi.org/10.1103/PhysRevE.69.026113>
- Newman ME, Leicht EA (2007) Mixture models and exploratory analysis in networks. *Proc Natl Acad Sci U S A* 104(23):9564–9569. ISSN 00278424. <https://doi.org/10.1073/pnas.0610537104>
- Nguyen H, Tran B, Tran D, Nguyen QH, Le DH, Nguyen T (2020) Disease subtyping using community detection from consensus networks. In: *Proceedings—2020 12th international conference on knowledge and systems engineering, KSE 2020*, (June 2021), pp 318–323. <https://doi.org/10.1109/KSE50997.2020.9287843>
- Nicosia V, Mangioni G, Carchiolo V, Malgeri M (2009) Extending the definition of modularity to directed graphs with overlapping communities. *J Stat Mech Theory Exp* 2009(3). ISSN 17425468. <https://doi.org/10.1088/1742-5468/2009/03/P03024>
- Parés F, Gasulla DG, Vilalta A, Moreno J, Ayguadé E, Labarta J, Cortés U, Suzumura T (2018) Fluid communities: a competitive, scalable and diverse community detection algorithm BT—complex networks and their applications VI. Springer International Publishing, Cham, pp 229–240. ISBN 978-3-319-72150-7
- Pavia R, Spinelli F, Monaco M, Mondello B, Monaco F, Gaeta R (2007) Lung cancer and cardiovascular diseases: occurrence, comorbidity and surgical timing. *J Cardiovasc Surg* 48(2):227–231. ISSN 0021-9509 (Print)
- Pizzuti C (2008) GA-Net: a genetic algorithm for community detection in social networks BT—parallel problem solving from nature—PPSN X. Springer, Berlin, pp 1081–1090. ISBN 978-3-540-87700-4
- Pons P, Latapy M (2006) Computing communities in large networks using random walks. *J Graph Algorithms Appl* 10(2):191–218. ISSN 15261719. <https://doi.org/10.7155/jgaa.00124>
- Radicchi F, Castellano C, Cecconi F, Loreto V, Paris D (2004) Defining and identifying communities in networks. *Proc Natl Acad Sci U S A* 101(9):2658–2663. ISSN 00278424. <https://doi.org/10.1073/pnas.0400054101>
- Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. *Phys Rev E Stat Nonlinear Soft Matter Phys* 76(3):1–12. ISSN 15393755. <https://doi.org/10.1103/PhysRevE.76.036106>
- Ramadhani HF, Annisa, Kusuma WA (2021) Identification of significant proteins in coronavirus disease 2019 protein–protein interaction using principal component analysis and ClusterONE. *Bioinform Biomed Res J* 3(2):25–34. <https://doi.org/10.11594/bbrj.03.02.04>
- Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. *Phys Rev E Stat Nonlinear Soft Matter Phys* 74(1):1–16. ISSN 15393755. <https://doi.org/10.1103/PhysRevE.74.016110>
- Renteria-Ramos R, Hurtado RG, Urdinola P (2018) Epidemiology, public health and complex networks. (November). <https://doi.org/10.22490/25904779.3053>
- Rossetti G, Milli L, Cazabet R (2019) CDLIB: a python library to extract, compare and evaluate communities from complex networks. *Appl Netw Sci* 4(1). ISSN 23648228. <https://doi.org/10.1007/s41109-019-0165-9>
- Ruan J, Zhang W (2007) An efficient spectral algorithm for network community discovery and its applications to biological and social networks. In: *Proceedings—IEEE international conference on data mining, ICDM*, pp 643–648. ISSN 15504786. <https://doi.org/10.1109/ICDM.2007.72>
- Sarfati D, Koczwara B, Jackson C (2016) The impact of comorbidity on cancer and its treatment. *CA Cancer J Clin* 66(4):337–350. ISSN 0007-9235. <https://doi.org/10.3322/caac.21342>. <https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21342>
- Savić M, Ivanović M, Radovanović M, Ognjanović Z, Pejović A, Jakšić Krüger T (2015) Exploratory analysis of communities in co-authorship networks: a case study. *Adv Intell Syst Comput* 311:55–64. ISSN 21945357. https://doi.org/10.1007/978-3-319-09879-1_6
- Schlicker A, Domingues FS, Rahnenführer J, Lengauer T (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinform* 7:302. ISSN 1471-2105 (Electronic). <https://doi.org/10.1186/1471-2105-7-302>
- Schriml LM, Mittra E (2015) The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. *Mamm Genome* 26(9–10):584–589. ISSN 10889051. <https://doi.org/10.1007/s00335-015-9576-9>
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software Environment for integrated models of biomolecular interaction networks. *Genome Res* 13(11):2498–2504. ISSN 10889051. <https://doi.org/10.1101/gr.1239303>. <https://pubmed.ncbi.nlm.nih.gov/14597658/>
- Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 22(8):888–905. ISSN 01628828. <https://doi.org/10.1109/34.868688>
- Sigel K, Wisnivesky J, Crothers K, Gordon K, Brown ST, Rimland D, Rodríguez-Barradas MC, Gibert C, Goetz MB, Bedimo R, Park LS, Dubrow R (2017) Immunological and infectious risk factors for lung cancer in US veterans with HIV: a longitudinal cohort study. *Lancet. HIV* 4(2):e67–e73. ISSN 2352-3018 (Electronic). [https://doi.org/10.1016/S2352-3018\(16\)30215-6](https://doi.org/10.1016/S2352-3018(16)30215-6)
- Su S, Zhang L, Liu J (2019) An effective method to measure disease similarity using gene and phenotype associations. *Front Genet* 10. ISSN 1664-8021. <https://doi.org/10.3389/fgene.2019.00466>. <https://www.frontiersin.org/article/10.3389/fgene.2019.00466>
- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 71(3):209–249. ISSN 0007-9235. <https://doi.org/10.3322/caac.21660>
- Traag VA, Van Dooren P, Nesterov Y (2011) Narrow scope for resolution-limit-free community detection. *Phys Rev E Stat Nonlinear Soft Matter Phys* 84(1):1–9. ISSN 15393755. <https://doi.org/10.1103/PhysRevE.84.016114>
- Traag VA, Krings G, Van Dooren P (2013) Significant scales in community structure. *Sci Rep* 3:1–10. ISSN 20452322. <https://doi.org/10.1038/srep02930>
- Traag VA, Aldecoa R, Delvenne JC (2015) Detecting communities using asymptotical surprise. *Phys Rev E Stat Nonlinear Soft Matter Phys* 92(2). ISSN 15502376. <https://doi.org/10.1103/PhysRevE.92.022816>
- Traag VA, Waltman L, van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 9(1). ISSN 20452322. <https://doi.org/10.1038/s41598-019-41695-z>

- Tripathi B, Parthasarathy S, Sinha H, Raman K, Ravindran B (2019) Adapting community detection algorithms for disease module identification in heterogeneous biological networks. *Front Genet* 10(MAR). ISSN 16648021. <https://doi.org/10.3389/fgene.2019.00164>. https://api.elsevier.com/content/abstract/scopus_id/85066635283
- Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23(10):1274–1281. ISSN 13674803. <https://doi.org/10.1093/bioinformatics/btm087>. <http://www.godatabase.org>
- Wei CH, Allot A, Leaman R, Lu Z (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* 47(W1):W587–W593. ISSN 13624962. <https://doi.org/10.1093/nar/gkz389>. <https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/>
- Yang J, Leskovec J (2015) Defining and evaluating network communities based on ground-truth. *Knowl Inf Syst* 42(1):181–213. ISSN 0219-3116. <https://doi.org/10.1007/s10115-013-0693-z>
- Yang Z, Algesheimer R, Tessone CJ (2016) A comparative analysis of community detection algorithms on artificial networks. *Sci Rep* 6. ISSN 20452322. <https://doi.org/10.1038/srep30750>. https://api.elsevier.com/content/abstract/scopus_id/84982671578
- Yu G, Wang LG, Yan GR, He QY (2015) DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 31(4):608–609. ISSN 14602059. <https://doi.org/10.1093/bioinformatics/btu684>
- Zhang P, Moore C (2014) Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proc Natl Acad Sci U S A* 111(51):18144–18149. ISSN 10916490. <https://doi.org/10.1073/pnas.1409770111>
- Zhang S, Ning XM, Ding C, Zhang XS (2010) Determining modular organization of protein interaction networks by maximizing modularity density. *BMC Syst Biol* 4(SUPPL. 2). ISSN 17520509. <https://doi.org/10.1186/1752-0509-4-10>
- Zhao C, Wang Z (2018) GOGO: an improved algorithm to measure the semantic similarity between gene ontology terms. *Sci Rep* 8(1):1–10. ISSN 20452322. <https://doi.org/10.1038/s41598-018-33219-y>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
