Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2021, 2021:91
https://doi.org/10.28919/cmbn/6839
ISSN: 2052-2541

STRUCTURAL KNOWLEDGE ANALYSIS AND MODELING OF MULTIMORBIDITY USING GRAPH THEORY BASED TECHNIQUES

FAOUZI MARZOUKI*, OMAR BOUATTANE

Laboratory SSDIA, ENSET Mohammedia, University Hassan II of Casablanca, Mohammedia, Morocco

Abstract. Multimorbidity is one of the major problems in the modern medical system. The more conditions the patient has, the greater the psychological pressure will be. We propose a formal definition of the general case of Multimorbidity Disease Network detection. Based on pairwise association method, we constructed an undirected weighted graph of co-occurrence for comorbidity based on the socio-psychological profile existing in a real data set. Based on the obtained network, we used the centrality analysis of the network nodes to conduct a mesoscopic-analysis, and used the community detection algorithm to determine potential components of the network. The main results show first, that algorithms used can be helpful for extracting models of multimorbidity. Second, that aging process not only affects the number of diseases, but can also influence Multimorbidity Burden and its complexity pattern.

Keywords: multimorbidity; comorbidity; centrality analysis; community detection; graph theory.

2010 AMS Subject Classification: 05C90.

1. Introduction

Comorbidity is defined as "any distinct additional clinical entity that has existed or that may occur during the clinical course of a patient who has the index disease under study" [1]. It is a major health problem in modern medicine: the more conditions are (i.e. Multimorbidity), the

*Corresponding author

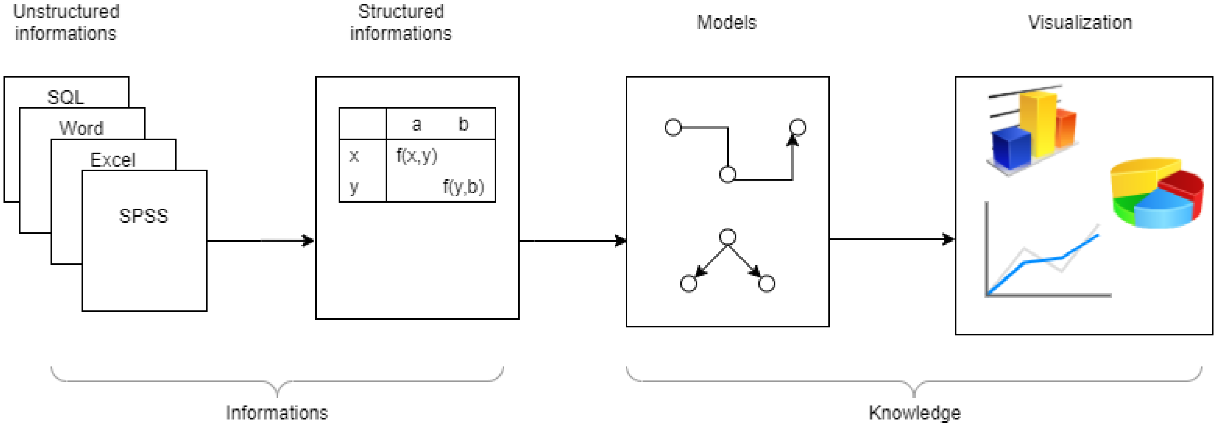
E-mail address: faouzi8marzouki@gmail.com

Received September 29, 2021

more burdens are put on patient and healthcare system. Moreover, healthcare systems are still designed in a single disease paradigm rather than Multimorbidity. However, the transition from a disease-centered care, to a patient-centered care is ongoing, but generally appears complicated [2].

Recently, initiatives are made to make use of the increasing amount of electronic health care data to get insight about this phenomena based on big data tools [3]. As general methodology, researchers in data science collect suitable data and tries to clean and transform them to suitable format, then apply models and algorithms to extract knowledge from primitive information gained from raw unstructured data. (See figure1). We try to contribute by this work in exploring the use of analytical tools in data science for multimorbidity modeling.

FIGURE 1. General methodology in data science followed in this work.



The increasing prevalence of multiple diseases poses problems for healthcare providers, care systems, and policies. It has impact as well for patients in well-being and quality of life [4, 5]. It seems to be obvious that understanding multimorbidity in light of psycho-social profiles provides more accurate understanding of psycho-medical needs of patients, thus to better manage cases with multimorbidity. Under this perspective, we investigate patterns of multimorbidity across psycho-social profiles of patients across life span time, according to Erickson theory [6] of psycho-social development. The mentioned theory classifies human life in 8 stages: Infancy (0-1.5 years), toddler (1.5-3 years) , early childhood (3-5 years), middle childhood (5-12 years), adolescence (13-18 years), young adulthood (18-40 years), middle adulthood (40-65 years), older adulthood ($>$ 65 years). Each stage reflects relatively stable psychological growth, in which individuals have general psychological properties and patterns. Edges between these stages reflect a transition period, a shift characterized often by a crisis considered as challenges in maturation process.

There are several ways to express association between diseases in multimorbidity research. The relative risk (RR), odds ratios (OR) [7, 8], and Multimorbidity coefficient (MC) are famous ones. Risk ratios is used for quantifying association among two disorders, which calculates the ratio of the risk of occurrence of a disease among exposed group of people to that among the unexposed [9]. Odds ratio is defined as the odds of being exposed to disorder D_2 if one has disease D_1, divided by the odds of being diagnosed with D_2 if one does not have D_1. The popularity of the odds ratio is due to the ease of calculation and consisting of a good estimate of the relative risk [10].

Odds and risk ratios estimate the overall strength of association between disorders but fail to separate clusters from coincidental comorbidity. Multimorbidity coefficients can express association among any number of diseases by dividing the actual rate of multimorbidity by expected numbers of cases. The MC favors pairs of low prevalence. To reduce this tendency a pseudo-count can be added to the numerator and denominator of the MC [11].

Our objectives are then twofold: first, to present the multimorbidity patterns detection problem in formal terminology. Second, to understand how multimorbidity develops across one's life stages. This work is conducted in two steps: first, constructing network model for co-occurencing diseases from the studied dataset. Second, analyzing network patterns using centrality analysis and detecting communities in the network.

In the remainder of this paper, Section 2 presents some related works. Section 3 is devoted to putting multimorbidity problem in formal framework. In Section 4 and Section 5, we present the used algorithms in community detection and the used centralities. We present results in Section 6. A general conclusion is in Section 7.

2. Related Works

Recently, increasing studies in medical literature were conducted to tackle multimorbidity burden, for exploring risk factors of multimorbidity, their impact on quality of life, mortality or costs and health care utility [12]. More technically, the methods and models differ either on the data, be it cross sectional or have a temporal dimension, or whether the goal of the model is to explain, explore or to predict.

Earlier medical research relied on regression models that are applied on single diseases, which ignore the big picture of multimorbidity complexity. Recently, combinations of traditional data analysis and machine learning were proposed as multimorbidity research methods. In [13] the authors used Classification/ regression trees and random forest applied to data of elderly adults to model and identify how specific combinations of chronic conditions, functional limitations, and geriatric syndromes affects costs and inpatient utilization. In [14] applied non-hierarchical cluster analysis based on k-means on Cross-sectional study using electronic health records of patients aged between 45 and 64 years to identify and separate certain population groups from others. In [15] added fuzziness upon k-means algorithm to estimate clusters of patients as well as membership matrix indicating the membership degree of a patient to a given cluster. In [16] a multilevel analysis of the influence of individual- and area-level factors on patterns of physical–mental multimorbidity and health-care use in the general population. Applying this method allows detecting the isolated and combined influence of variables of each level on the outcome variable.

In other approach, network science was a fertile domain to use in order to draw insights about comorbidity disease network. Hernández et al. [17] proposed an analysis of comorbidity patterns using network analysis and the use of association rules was performed to study disease associations in 6,101 Irish adults aged more than 50 years. They perform Louvain algorithm to detect clusters of diseases from the disease network. The standardized lift and confidence scores of the association rules was considered as probabilistic measures of how conditionally the diseases are related. In [18] logistic regression models, adjusted by age and sex, and odd ratio as strength association, were used to estimate the comorbidity network. They used some metrics to analyze the network such as clustering coefficient, Page Rank and degree centrality. In [19] used Salton Cosine Index as a comorbidity strength to build comorbidity network and weighted degree, closeness and betweenness centrality for a microscopic analysis of the comorbidity network.

Other approaches in literature focused in probabilistic formulation and longitudinal data. In the important work of [20], Lappenschaar et al. summarized and classified some terminologies used in definitions of concepts of multimorbidity, and proposed probabilistic framework to model these concepts using causal Bayesian network [21]. In [22], the authors proposed Bayesian network structure learning methods for modeling the interactions between risk factors explaining co-occurrences of malignant tumours in oncological area. This model was extended with a temporal dimension in [23]. Authors in [24] proposed a latent-based approach to model multimorbidity related event in temporal electronic health records and introduced the notion of clusters of hidden states allowing for an exploration of the multiple dynamics that underlie events in data.

In the aforementioned studies in medical field, the authors focused on the representability of concepts related to the multimorbidity problem, and were concerned more about the fidelity of the proposed model to the real world characteristics, and on the interpretability of the outcomes of proposed methods. They choose mathematical techniques as tools to draw conclusions when applying them to data, then interpreting the outcomes from medical point of view. The chosen model itself is a case study of a given analytical tool. In this work, our methodology is quite the opposite, we study the tools in a computational point of view and assess them on a case study of a real medical dataset.

In theoretical computer science, there are many ways to learn the network's structure, and during the last 20 years, many new advances in structure learning algorithms based on different principles have been proposed. The main question here is whether the performance of such learning algorithms still of significance and feasibility to the real world multimorbidity data. For example, where in some domains structure learning algorithms that favour sparsity of the network are more suitable, one cannot be sure whether this holds for Multimorbidity domain. Compared to other domains, medical data has a large range in frequency of occurrence of particular events. From a medical perspective, any significant association helping to understand multimorbid diseases network would be of great value. Besides, medical data often is characterized by having extreme prevalences. For example, 96 % of nearly 9800 diseases, in ICD-10 coding, have a prevalence of less than 1%. Further, co-occurrences of diseases of less frequent diseases are likely to be even rarer. Moreover, how these approaches perform in large size of real world datasets with very large number of variables is a great point to assess its feasibility. Therefore, it is necessary to know which of the learning algorithms are most suitable for the ultimate purpose of building computerized decision support systems. The feasibility of these models performance on very large and heterogenic real world data is questionable.

Our contributions, besides the different methodological concerns, are: we provide a general machine learning framework that describes the learning of structural knowledge of multimorbid data. This framework can express many of the methods proposed in medical literature in a general way. Then we give an implementation of a pairwise approach to our data and discuss some of its computational properties. Then we move to treat special case of the comorbidity network, and discuss some techniques of microscopic and mesoscopic analysis of the obtained comorbidity network. We choose centrality and community detection as techniques of analysis. For community detection, we compare four well known algorithms, representing different approaches in community detection literature. Finally, we applied some centralities and we try briefly and carefully to interpret them in light of our data of interest.

3. Data and Methods

3.1. Problem setting. To understand co-occurrence between multimorbid diseases (let us consider k multimorbid diseases), we computed associations strength of all combinations of k diseases in the data, per life stages for both genders across a given time period (2016).

In the following, we note $|S|$ the number of elements of a given set S. Let $D = \{d_1, d_2, \dots, d_{|D|}\}$ a finite set contining $|D|$ number of diseases present in medical dataset. Let $X = \{X_1, X_2, \dots, X_{|X|}\}$ the set of all observations of patients such that $X_i = (x_{1,i}, x_{2,i}, \dots, x_{|X_i|,i})$ with $x_{j,i} \in D$ a tuple of observed diagnosis for the patient i. Let $X^v \subset X$ the subset of patients with a certain profile v (a profile can be any patient's characteristics). $X^v = \{X_1^v, X_2^v, \dots, X_{|X_v|}^v\}$ is a finite subset of X indexed by v=(a, b) such that a represents in this work, gender:
$a \in \{female, male\}$ and b represents life stage: $b \in \{$Infancy, toddler, early childhood, middle childhood, adolescence, young adulthood, middle adulthood, older adulthood$\}$.

We assume that data X are independent and identically distributed (i.i.d), i.e samples from the datasets are generated by the same generative law, which has no memory of previously generated samples. To put it in another way, we base our analysis in this paper on the premise that every patient sample consists of distinct cases, each of which is caused by the same underlying Multimorbidity Mechanism.

Let R consists of a binary relation over Cartesian product sets $D \times D$. Two diseases d_1 and d_2 are related with the relation R if and only if they satisfy a predefined condition. This condition depends on the context of the study. It can be for example the fact of being correlated, or causally related, or conditionally related. This relation R is usually assessed by a metric to measure its strength. Statistically, R is estimated in function of observations X. We consider the hypothesis space $H = \{R_\theta(X) | R_\theta(X)$ depend on parameters $\theta = (\theta_1, \ldots, \theta_k)$ is a statistical model explaining the observed co-occurencing of diseases $\}$. Parameters θ encode the parameters related to the chosen model: it could be logistic coefficients or Bayesian networks probabilistic table, degree of polynomial regression, Hidden Markov models [24], latent class [25], principal component analysis [16], threshold in significant associations [18, 26]. Technically, statistical models R are grouped as family of equations and H is framed based on assumptions underlying the problem of interest. We use this binary relation over the disease set to define a Comorbidity Disease Network (CDN). Generally, we use a k-ary relation over $D \times D \times ... \times D = D^k$ to define a Multimorbidity Disease Network (k-morbidity disease network).

We define in this work the binary relation $R(d_1, d_2)$ as follows: two diseases d_1 and d_2 are related by the binary relation R if and only if $p(d_1, d_2) > p(d_1).p(d_2)$ is statistically true, such that d_1 and d_2 are assimilated by two binary random variables and $p(d_i)$ stands for the occurence probability of the disease $d_i \in D$ and $p(d_1, d_2)$ stands for the occurrence probability of d_1 and d_2 in the same time. This definition can be easily generalized in case of k-ary relation (this is done in algorithm 1 in section 3.3).

This definition coincides with Van Den Akker et al. definition [27] of cluster comorbidity: if d_1 has occured, then d_2 will be more likely to occur than what would be expected just by chance. If $p(d_1)p(d_2) \simeq p(d_1, d_2)$ we consider that the two conditions are randomely co-occurencing. d_1 and d_2 as positively comorbid diseases, if $p(d_1, d_2) > p(d_1)p(d_2)$, The final case $p(d_1, d_2) <$

$p(d_1)p(d_2)$ can be interpreted as d_1 and d_2 are in protective comorbidity (for example myopia may be protective against diabetic retinopathy [28]).

To measure the strength of this relation/association, multimorbidity coefficients (MC) is calculated. MC measures pairwise associations [29], i.e. how strongly disorders are linked. It is defined as the division of observed rate of co-morbidity (multimorbidity) by the rate which is expected, under the null hypothesis of no association between the separate disorders. See table 1 for illustration.

TABLE 1. Taking into account the notation of this table we have the Mutimorbidity coefficient score for Disease 1 and Disease 2 is:

$$\text{MC} = \frac{\frac{a}{N}}{\frac{a+c}{N} * \frac{a+b}{N}} = \frac{aN}{(a+c)*(a+b)}$$

		Disease2		
		Occurence	Absence	Total
Disease1	Occurence	a	b	a+b
	Absence	c	d	c+d
	Total	a+c	b+d	a+b+c+d = N

3.2. Data.

The analysis was applied in a case study of real medical dataset [30], a hospital inpatients' diagnosis dataset, consist originally of admissions in The National Health Service hospitals of Madrid, in Spain during 2016. Each diagnosis of an admitted patient is encoded by The International Classification of Diseases, Tenth Revision (ICD 10). The data contain 78451 patients (34639 males, and 43812 females). The maximum number of registered diagnosis per admission is 20.

3.3. Methods.

3.3.1. *Building Multimorbidity Disease Network (MDN) algorithm.*

Let $D = \{d_1, d_2, d_3, ..., d_n\}$ the set of disease. $M_k \subset D$ such that $k > 1$, is the subset of distinct k diseases from D. (e.g M_2 is the subset of all possible co-morbidities, M_3 is the subset of all possible Tri-morbidities, and so on). Let $f : I \subset N \to \{D_1, D_2, ..., D_{N_{diag}}\} \subset P(D)$an application that maps every patient $i \in I$ to is recorded diagnosis $f(i) = \{x_1^{<i>}, x_2^{<i>}, ..., x_{N_{diag}}^{<i>}\}$. (In our data of application $N_{diag} = 20$).

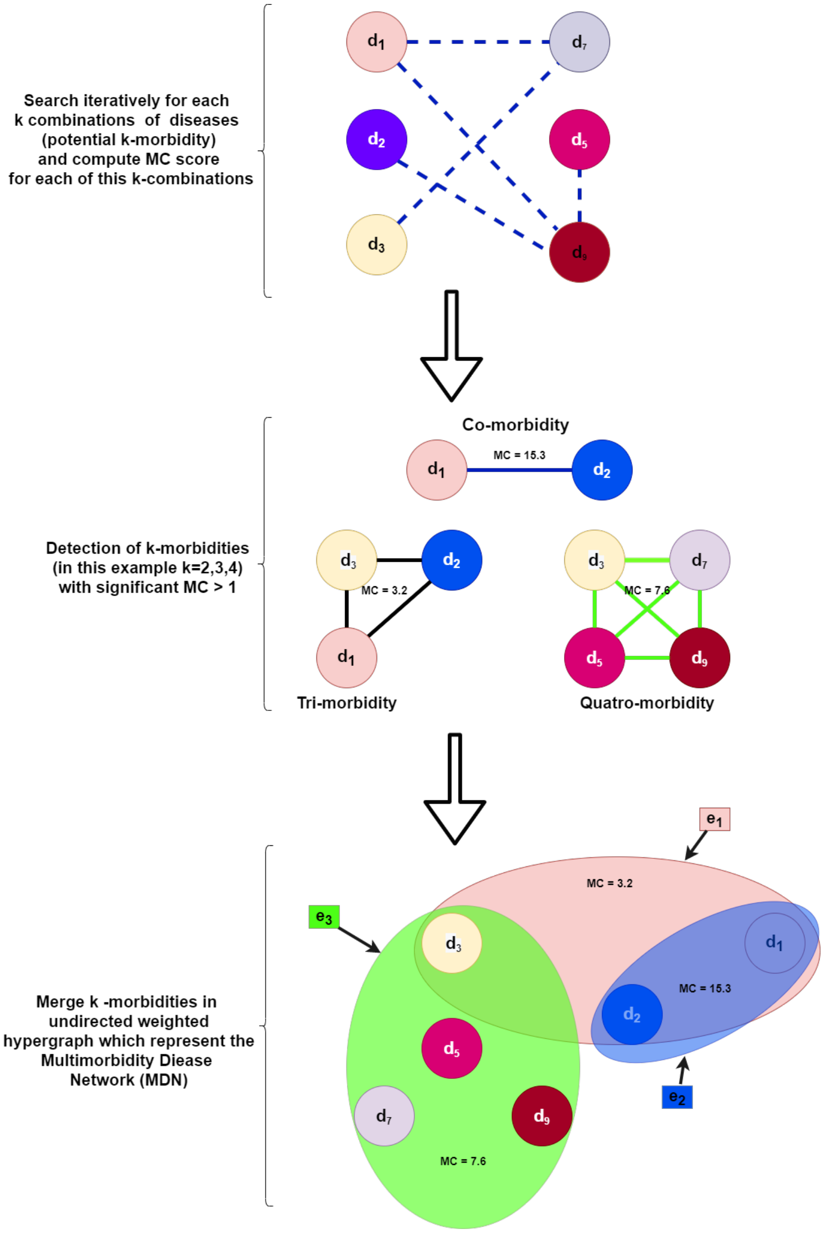
The algorithm search for every $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ combinations of distinct $n > k > 1$ diseases among n total diseases, and compute their MC score by mining their co-occurrence. If the MC is significantly higher than 1, then we consider that these diseases are in Multimorbidity. If the MC score is less than 1 then we consider that they are in protective Multimorbidity. The bigger this number is, the stronger the association is considered. We are more interested by positive comorbidity. For each single k-morbid disease (k is fix in this case), the algorithm has to verify $\frac{n!}{(n-k)!k!} = o(n^k)$ possible combinations.

The k-morbid disease sets are represented by an undirected weighted hypergraph. $H = (V, E)$ such that the vertices $V = D$ and hyperedges $e_{i,j} \in E$ represent the association strength MC. a hyper graph is a generalization of a graph in which an edge can join any number of nodes. In contrast, an edge in an ordinary graph connects exactly two nodes. We can easily notice in this proposed model that, while a Hyper graph corresponds to Multimorbidity Disease Network (MDN), an ordinary graph corresponds to a Comorbidity Disease Network (CDN). See Algorithm1. The algorithm repeats this instructions $|D|$ times which corresponds to the maximum number of multimorbidity (all diseases present in the dataset co-occur in the same time), which results in

$$\sum_{1<k<|D|} \binom{|D|}{k} = 2^{|D|} - |D| - 1$$

multimorbid diseases configurations to verify by algorithm 1. This exponential nature of this algorithm in respect to the number of investigated diseases makes it infeasible for large amount of epidemiological data. Two solutions can be adopted in this case. Either to reduce the search space to include just what is unknown associations in expert knowledge, or parallelization and distributed systems can be used to reduce the temporal complexity, especially with the fact that iterations over the k – morbidities can be computed separately.

FIGURE 2. Steps of Algorithm1: first the algorithm search for all combinations of k diseases and compute their MC. Second, it detects significant non random associations of the k diseases. Third, merge the detected Multimorbidities in one hypergraph $H = (V,E)$. In this example we have $V = \{d_1,d_2,d_3,d_4,d_5,d_7,d_9\}$ and the hyper edges $E = \{e_1,e_2,e_3\} = \{\{d_1,d_2,d_3\},\{d_1,d_2\},\{d_3,d_5,d_7,d_9\}\}$



Algorithm 1 Build Multimorbidity Disease Network

```
Require: $a\ patient\ to\ diagnosis\ function\ map\ f : I \subset \mathbb{N} \rightarrow \{D_1; D_2; ...; D_{N_{diag}}\}\ a\ disease\ set\ D,$
Ensure: $weighted\ undirected\ hypergraph\ H = (V, E)$
for each k from 2 to $|D|$ do
    for each $M_k \in \mathscr{P}(D)$ do
        $W_{expected} \leftarrow \prod_{d \in M_k} Count_{occ}(\{d\}, I)$
        $W_{observed} \leftarrow Count_{occ}(M_k, I)$
        if $H_0$ : "$W_{expected} \geq W_{observed}$" is rejected at risk $\alpha$ then
            $V^k_{M_k} \leftarrow M_k$
            $E^k_{M_k} \leftarrow MC$
        end if
    end for
end for

procedure MERGE($G_k = (V^k_{M_k}, E^k_{M_k})$)
    $V \leftarrow \cup_k V^k_{M_k}$
    $E \leftarrow \cup_k$ as.one.hyperedge($E^k_{M_k}$) // each clique $E^k_{M_k}$ from graph $G^k$ will be transformed into a one hyperedge in hypergraph $G$.
end procedure
procedure $Count_{occ}(S : a\ set, I : a\ subset\ of\ integers)$
    $S_{occurences} \leftarrow \emptyset$
    for each $X \in S$ do
        for each $i \in I$ do
            $S_{occurences} \leftarrow S_{occurences} \bigcup (X \bigcap f(i))$
        end for
    end for
    return $|S_{occurences}|$
end procedure
```

The $count_{occ}(S,I)$ procedure count the number of occurences of a disease $d \in S$ in diagnosis records indexed by $i \in I$. One way to implement this procedure is to implement it as a search algorithm, in which the sequential search will count the number of occurrences by iterating over the diagnosis records f(I) resulting in $\sum_{i \leq |X|} |f(i)| = O(|X|.max(|f(i)|))$. If $(D, \leq)$ is a totally ordered set such that $\leq$ represent totally order relation (such as lexicographic order of Diseases code, or some ranking score for diseases.), a binary search algorithm will take advantage of order relation to reduce search space logarithmically in sorted data; resulting in $\sum_{i \leq |X|} log|f(i)| = log(\prod_{i \leq |X|} |f(i)|) = O(|X|log(maxf(i)))$.

In the remainder of this paper, we will focus the analysis in the special case of Comorbidity Disease Network.

4. Community in Multimorbidity Disease Network

4.1. Overview.

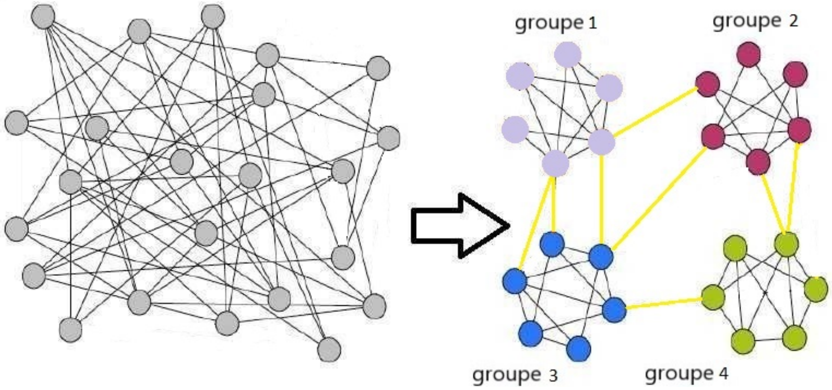
A community is a topological property of a network. It is defined as a group of nodes having strong connection and denser from the rest of the network. A community in the context of this paper is the set of diseases that interact with each other more than the others in the diseases network. Identifying community structures is a step towards the understanding of different structures and function of this network.

Each algorithm in this approach works on the multimorbidity network to detect subsets forming typical groups. So, this algorithm will assume that diseases present in dataset can be split into homogenous groups.

Community detection is a network analysis method often confused with graph partitioning and graph clustering. Graph partitioning and graph clustering is the task of dividing the vertices of the graph in a given number of mutually exclusive subsets of a given size such that the number of linking edges is minimal. Whereas Community detection is attributing to every node a given class (community), and overlapping is allowed, and no class number or sizes are predefined.

Many algorithms and approaches have been developed to detect community in networks, each has its pros and cons, and performance depends as well on the topological properties of each network. Thus, a comparative analysis is often required for selecting the most suited method. Community detection is often formulated as a combinatorial optimization problem, in which an optimization method tries to optimize a criterion (cut, conductance, modularity, etc) and the exact solution is NP hard, thus the focus on greedy, approximation algorithms and heuristics in literature.

FIGURE 3. Community detection problem. In the left a set of disease nodes and their associative edges. Community detection consists of revealing hidden diseases groups. Yellow edges must tend to a min score. These groups are interpreted as multimorbid diseases categories.



4.2. Girvan-Newman algorithm.

Introduced by Girvan and Newman [31]. A traditional divisive algorithm uses edge betweeness as a metric to identify the boundaries of communities. These methods try to identify network elements playing the role of linking communities. The Newman-Girvan algorithm detects communities by iteratively removing those important linking edges from the original network. Those edges that are most likely between communities are determined based on betweeness measures, which is the number of shortest paths between all nodes pairs that run along the edge, the edges connecting communities are then expected to have high edge betweenness. For the removal of each edge, the calculation of edge betweenness is $O(EN)$, therefore, this algorithm's time complexity is $O(E^2N)$, and $O(N^3)$ in a sparse network.

4.3. Label Propagation Algorithm.

This algorithm was introduced by Raghavan et al [32]. This algorithm starts with initializing a distinct community labels for each node in the network. Then, listing the nodes in the network in a random order. Afterwards, through the random sequence, each node takes the label of the majority of its neighbors. The above step will stop once each node has the same label as the majority of its neighbors. The computational complexity is O(E).

The advantage of this algorithm is that it is quite fast because it doesn't collect prior information about the network. Different community structures are reachable from the same initial condition, this is a limitation of this algorithm. The algorithm uses the network structure to guide its progress and does not optimize any chosen measure of community quality. The problem however is that subgraphs in the network that is bi-partite or nearly bi-partite in structure lead to oscillations of labels.

4.4. Louvain Algorithm. One of the popular algorithms with multilevel hierarchical strategy. It was introduced by Blondel et al [33]. It is a greedy algorithm attempting to optimize the modularity score of a partition of a network. The optimization is executed in two steps. First, the method looks for small communities by optimizing local modularity. Secondly, it aggregates nodes belonging to the same community and builds a new network whose nodes are the aggregated communities. These steps are repeated until a maximum of modularity is attained resulting in a hierarchy of communities is produced. The computational complexity of the multilevel algorithm is $O(NlogN)$.

4.5. Walktrap Algorithm. Proposed by Pon Latapy [34], a hierarchical clustering algorithm based on predefined random walk based similarity measure, which is a distance between nodes and sets of nodes (communities). The basic idea is that short distance random walks tend to stay in the same community. Starting from a totally non-clustered partition, the distances between all adjacent nodes are computed. Then, two adjacent communities are chosen, they are merged and the distances between communities are updated. If this step is repeated $N+1$ times, then the computational complexity of this algorithm is $O(EN^2)$. And $O(N^2log(N))$ For sparse networks.

4.6. Comparing community quality. One of the most popular measure of community strength is modularity. The basic idea is to compare the fraction of edges within the cluster to the expected fraction in a random graph with identical degree distribution. Its value can be either positive or negative. Positive value of modularity implies the presence of a community structure shape. The modularity Q is used to compare quality of communities detected but also as an objective function to optimize. It is defined as follows:

$$
\begin{aligned}
Q &= \sum_u (e_{uu} - a_u^2) \\
&= \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})\delta(C_i, C_j)
\end{aligned}
$$

where m is the number of edges, A_{ij} is the adjacency matrix, k_i is the degree of node i, C_i is community/group i, δ is the Kronecker delta such that $\delta(x,y) = 1$ if $x = y$ and 0 otherwise; e_{uu} is the observed fraction of edges within the group/community u, a_u^2 is the expected fraction for the same group, e_{ij} is the fraction of edges that associates groups i and j. $Q = 0$ implies that this group is what would be expected by a random attribution of edges. $Q \simeq 1$ implies that the group is very well shaped as a community inside the graph.

5. Centrality Metrics

5.1. Centrality measures. Centrality indices characterize an important vertex in a graph by defining a real-valued function $C : V \to R$ on the vertices of a graph, where the score $C(v)$ provide a ranking that identifies a scale of node importance. In general, the importance is defined in relation to the information flow across the network, or to the cohesiveness of the network. In this work, we calculated the following centralities: edge betweennes, closeness, degree, eigen vector.

5.2. Betweenness. Betweenness centrality quantifies how many times a node play a role of bridge in the shortest path between two other nodes. It was introduced originally as a measure to quantify the control of a human on the communication flow between other humans in a social network by Linton Freeman [35]. Formally, if σ_{st} is the total number of shortest paths that run from node s to node t, and $\sigma_{st}(v)$ is the number of those paths that pass through v, the betweenness can be defined as $C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$.

5.3. Closeness. Closeness is defined as the average length of the shortest path between a node and all other nodes in the graph. i.e. the closer a node is to all other nodes ,the more central is. if $d(v_1, v_2)$ is the distance between vertices $v_1 and v_2$ then closeness of a vertex v can be defined formally [36] as: $C(x) = \frac{1}{\sum_y d(y,x)}$ for all vertices y in V. Computing betweenness and closeness of vertices involve computing the shortest paths between all pairs of vertices, which requires $O(V^3)$ using the Floyd–Warshall algorithm [37], or $O(VE)$ on unweighted graphs with Brande's algorithm [38].

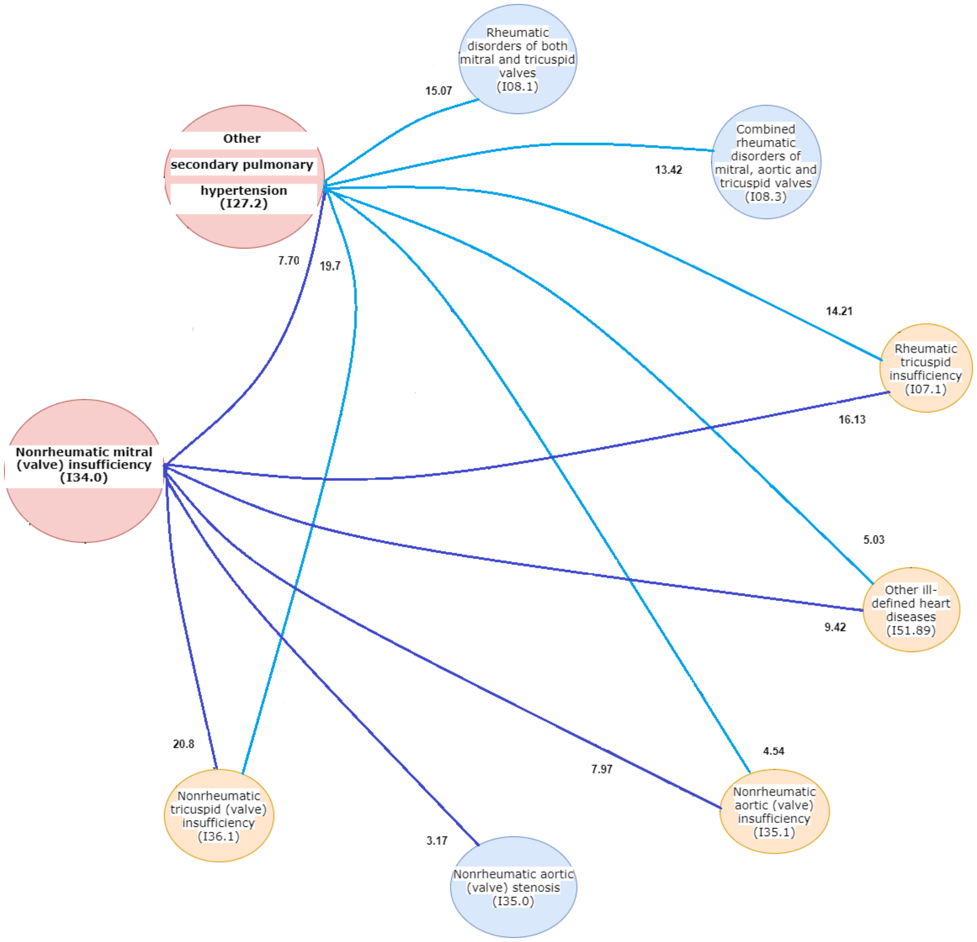
5.4. Eigen vector. Its basic idea is that connections to influential nodes will lend a node more influence than connections to less influential nodes. If we denote the centrality of vertex i by x_i, then we can model influential effect by making xi proportional to the average of the centralities of i's network neighbors: If $A = (a_{i,j})$ is the adjacency matrix, then $x_v = \frac{1}{\lambda} \sum_j a_{v,j}\, x_j$, or equivalently in matrix notation $Ax = \lambda x$, with $x = (x_1, x_2, ..)$ the vector of centralities. Since the adjacency matrix is non-negative, by the Perron-Frobenius theorem there is a unique largest real positive eigenvalue λ. In this way, the eigenvector centrality accords for each vertex a centrality that depends on both the quality and the number of its connections [39].

5.5. Degree. It is the simplest centrality measure. It is the count of how many edges a node has. The degree can be interpreted as the ability of a node for catching flowing through the network (such as a virus, or some information propagation). Calculating degree centrality for all the nodes takes $\Theta(V^2)$ in a dense adjacency.

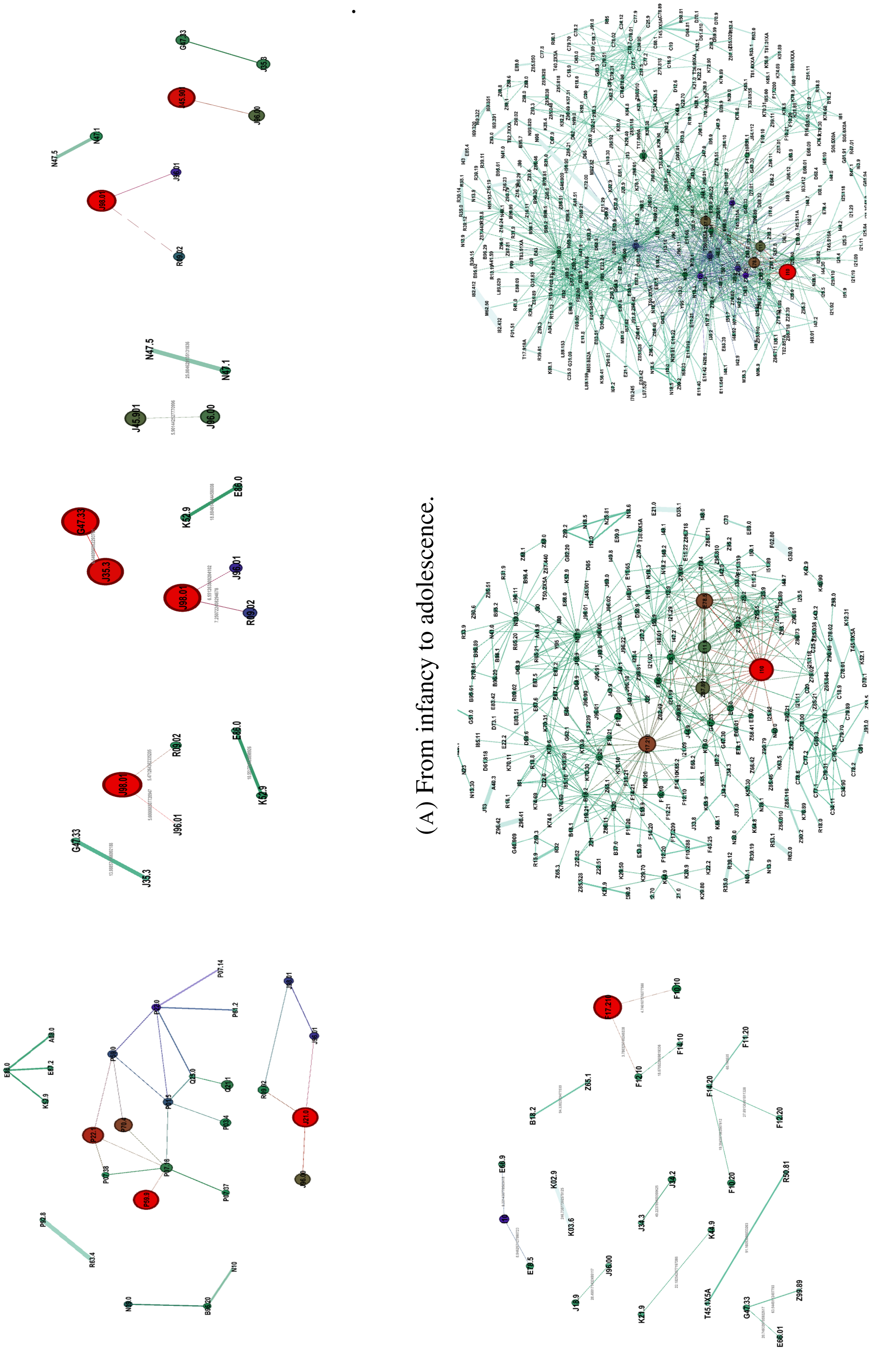
6. Results and Discussion

We applied algorithm 1 in the special case of Comorbidity Disease Network. (i.e $M_k = M_2$) to our data. Figure 4 presents example of 9 diseases and their co-occurrence strength (MC) for older males ($>$ 65 years). The absence of an edge between two diseases means, according to our available data, that either the association is not significant or this association is what would be expected just by chance. The presence of edge is considered as significant non random association comorbidity between the two diseases linked by this edge. For example, Rheumatic disorders of both mitral and tricuspid (valves) co-occur 15.07 times more than what would be expected just by chance with other secondary pulmonary diseases.

FIGURE 4. Example of co-occurrence of some ICD 10 codded diseases for males in older adulthood (> 65 years). For example, non rheumatic aortic stenosis co-occurs with non-rheumatic mitral insufficiency other secondary pulmonary diseases 3.17 times more than what would be expected just by chance.


Rheumatic disorders of both mitral and tricuspid valves (I08.1)
15.07
Other secondary pulmonary hypertension (I27.2)
Combined rheumatic disorders of mitral, aortic and tricuspid valves (I08.3)
13.42
7.70
19.7
14.21
Rheumatic tricuspid insufficiency (I07.1)
16.13
Nonrheumatic mitral (valve) insufficiency (I34.0)
5.03
Other ill-defined heart diseases (I51.89)
9.42
20.8
4.54
7.97
Nonrheumatic aortic (valve) insufficiency (I35.1)
Nonrheumatic tricuspid (valve) insufficiency (I36.1)
3.17
Nonrheumatic aortic (valve) stenosis (I35.0)

If we zoom out the whole detected comorbid diseases we can see visually the increasing quantity and complexity of diseases across psycho-social profiles. Figure 5 show visually the obtained zoom out Comorbidity Disease Network for males across life cycles. Each edge in the graph represents a significant association between two diseases ($p - value < 0.01$).

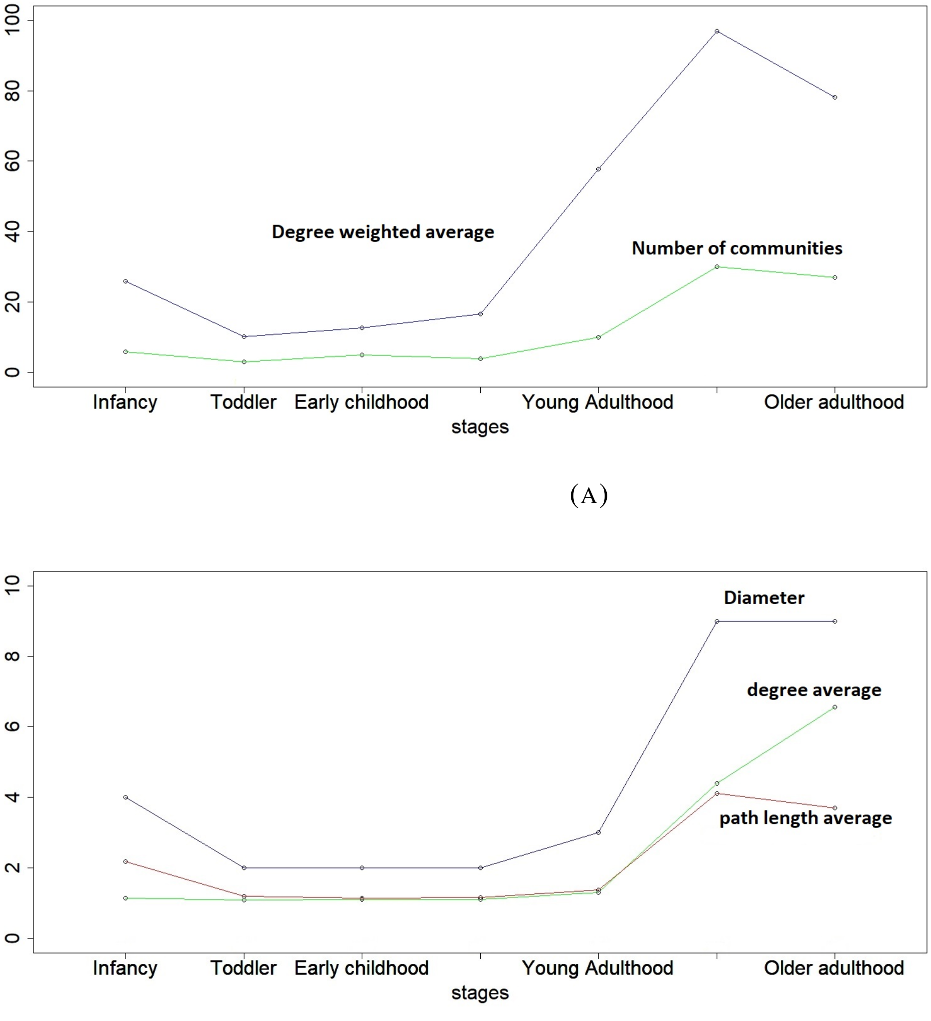


(A) From infancy to adolescence.

(B) From young adulthood to older adulthood.

FIGURE 5. CDN for Males patients. The red color indicates the diseases with the highest frequent diseases.

FIGURE 7. Diameter degree and weighted, degree average, path length average and number of communities evolution over stages of life, detected for males in the studied dataset.



(A)

(B)

In figure 7 represent quantitatively some descriptions of this obtained CDN. Visually, we observe increasing amount of conditions as well as the density interactions in function of age: for instance, older males CDN has 447 diseases and 1468 significant associations detected, while in early infancy resulted in a CDN with 7 diseases and 4 significant associations. We can see from the indicators of figure 7, three phases: in infancy, slightly high indicators, then from toddler, to young adulthood a stable indicators overs CDN, then middle and older adulthood know a sudden increasing in indicators values. The vulnerability of infancy and older adulthood profiles can explain these three phases. Besides, degree distribution of diseases nodes increases from average of 1.5 in childhood to 6 in older adulthood, which reveal increasing potential pathways and complicated possible disease scenarios as age increases. The diameter increases from 2 to 9 (i.e the maximum number of edges between two nodes/diseases in CDN is of 9 significant associations).This confirms, unfortunately the increasing possible pathways, but fortunately reveals (since the diameters is 9 which is very low to the possible highest diameter) that diseases act like “islands” of diseases, which can be manageable if these islands are detected and understood from data.

In table 2 we present some descriptive statistics and diseases with high centrality measures of our data according life stages. The hardest and maybe the misleading part of in handling centrality measure is its interpretation. Since the importance of a node in a graph (which is reflected partially by centrality and other statistics) is based upon assumptions about what is important to consider in a specific domain of interest. Since our work is essentially technical, we will be careful in interpretations and we let physicians to consider centralities in light of their specialty. However, in this work we adopted the following basic interpretations to justify the chosen centralities: diseases with high degree centrality are diseases that have high probability of appearing in multiple comorbidities. Closeness centrality reflects the degree of contagion of diseases over its comorbid diseases. Diseases with high edge betweenness are diseases which should be targeted in therapeutic interventions, since they act like bridges connecting other diseases which increase the multimorbidity burden of patients. Finally, diseases and conditions with high Eigen vector centrality are conditions related to influential diseases, which may indicates a diseases related by direct causations. Notice that nodes in the graph may have heterogenic centrality profiles, and nodes with the highest score in most centralities tend to play a role of a “trouble makers” in the CDN. Examples of diseases with highest scores in multiple centralities in the same time are primary hypertension (I10), heart failure (I50.9), acute kidney failure (N17.9). It is known in medical literature that these diseases generate other diseases because of damaging impact of the instability of vital organs at the core of the circulatory and genitourinary systems.

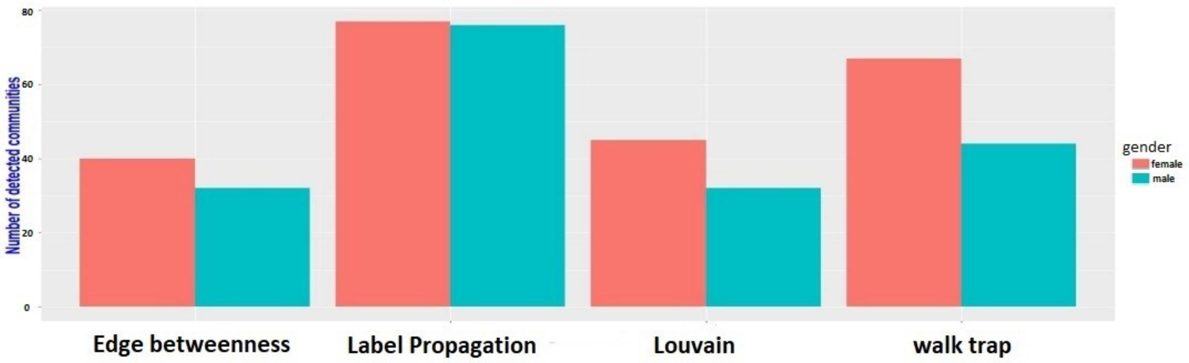
TABLE 2. Some centrality statistics about graph diseases. Example of high scores ICD10 coded diseases are given.

			# of detected	Example of	Example of diseases with high centrality				
			communities	comorbid diseases	betweeness	closeness	degree	eigenvector	eccentricity
life cycle	Infancy	f	7	(E86.0, E87.2, K52.9) (J98.01, R09.02),(J21.0, J96.01)	Q25.0-Q21.1-P07.18	RO9.02-J21.0	Q21.1-Q25.0	Q25.0-P22	P61.2-P22.0
		m	5	(J21.0-R09.02-J96.01) (P59.9-P70.4-P59-P22-Q25)	P01.5-P59.0-Q25.0-J21	E86.0-B96.20-R63.4- P92.8-J21.0	D07.18-P22.0- E86.0-Q25.0	P09.18-P59.0-P22.0	P22.1-P70.4-Q21.1 J98
	Toddeler	f	3	(J35.3-G47.33)-(R09.02-J98.01) (B96.20-N39.0)	-	-	-	-	-
		m	3	(J98.01-J96.01-R09.02)-(J35.3-G47.33) (K52.9-E86.0)	-	-	-	-	-
	Early childhood	f	2	(J35.3-G47)-(J98.01-J96.01-R09.02)	J98.01	J98.01-J35.3	J98.01	J98.01	R09.02-J96
		m	5	(N47.5-N47.1)-(J35.3-G47)- (J45.9-J96.00)	-	-	J98.01	J98.01	-
	Middle childhood	f	1	J35.3-G47.33	-	-	-	-	-
		m	4	(J98.01-J96.01-R09.02)-(N47.5-N47.1) (G47.33-J35)	J98.01	G47.3-J35-N47.5	J98.01	J98.01	R09.02
	Adolescence	f	1	Z37.0	Z37.0-O48.0	G47.3	-	-	-
		m	1	K02.9-K03.6 (p-value<0.05)	-	-	-	-	-
	Young adulthood	f	31.25	(Z37.0-O48.0-Z3A.07-E03.9)- (Z37-030-060) (J45-Z88-O99) -(Z33.02-O35)	Z37.0-090.81-060.14 F17-D64.9	E66.01-O99.354-Z85-K83 C73-E60.3-E11	Z37.0-O48.0- Z3A.40-O77.0	Z37.0-OO9.523 -O48.0-O77.0	D50.9-O99.3 -Z88.8-O60.14
		m	10	(F17.210-F10.10-F12.10-F14.10)- (T45.1-R50.81) I10-E66.9-E78)-(G47.33-E66.01-Z99.8)	F14.20	F14.20-G47.33-I10-J34.3	F14.20-F17.210	F14.20-F17.20	F14-F12-F17
	Middle adulthood	f	39,75	(I10-E66.9-E09-M19-E78-I25-E66.3-J45) (F17-E78.5-J44.1-O99.324)	Z37.0-E03.7-I10- F17-Z92-N17.9	M20.11-M77.42-Z88-N18 I12.0-D35.1-K82.4	Z37.0-I10-OO9.523- E78-J22	Z37.0-009.523 -048-099-Z67.40	030.43-C56.1- K31.89-I85.10
		m	41.25	(I10-E78.5-Z87.891-K76.0-N40.0)- (F17.210-K76.0-J44)	Z87.89-Z92.21-N17.9- K76.6-F10.20	K44.9-I12.0-N40.1-E21.0 G30.9	F17.210-N17.9-Z87.89 B18.2-E78.5-I10	E78.5-Z87.891-I10 E11.9-I25.2	F10.288-F12.20 Z22.5-D78
	Older adulthood	f	60	(I10-E78.0-E66.9-E11.9-E78.5-Z86.73) (N39.0-E11-N17.9-R32-K59-G30-L89.15)	Z90.710-I50.9-Z79.82 N17.9-C78.7-K44.9	M20.12-M48.06-E89.0-I68.0 L89.621-S06.6	I50.9-N17.9-N39-I27.2 Z59.3-R32-E46	I50.9-N17.9-Z59.3- N39-I12.9-I27-D50.9	B18.2-K74.60-I85.1 K31.89
		m	44.75	(I10-I25.2-E78.0-Z79.5)- (Z87.891-G47.33-N40-R91.1)	Z87.891-N39.0-I50.9 C78.7-F10.21	Z88.1-I69.351-R62.0-I43-R55	I50.9-N17.9-N39.0- I12.9-Z87.891-I27.2	I50.9-N17.9-I12.9 -Z87.891-N18.9	K26.4-K57.31-N99.0

Further, we applied community detection approach to detect potential components of MDN. Since algorithms in literature have different assumptions, pros and cons, we conducted comparative analysis of four well known algorithms. Figure 9 present number of communities detected for each algorithm.

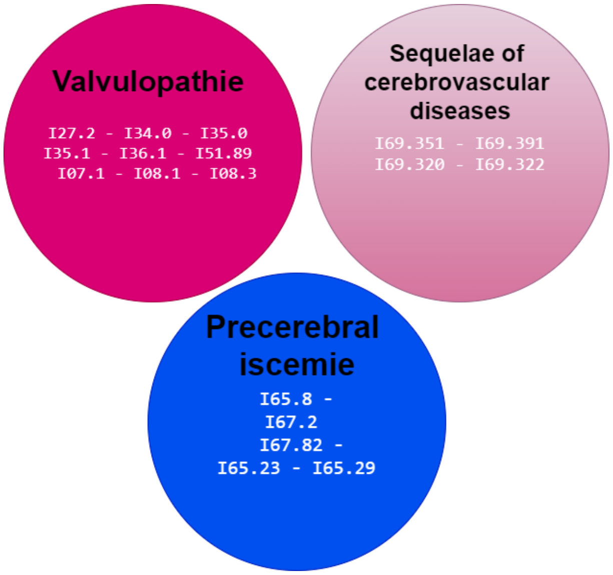
We observe from figure 9 that females have in general more average of communities than males, and modularity values had a good scores around 0.78-0.90, which can be explained by a modular topology aspect of CDN, i.e groups of diseases are easily distinguishable (and acts like islands) which causes the algorithm performances are comparable. We think that increasing numbers of clusters reveals increasing types of multimorbidity diseases and additional layers of multimorbidity burden as age increases.

FIGURE 9. Comparison of algorithms of community detection for older adult patients (>65 years).



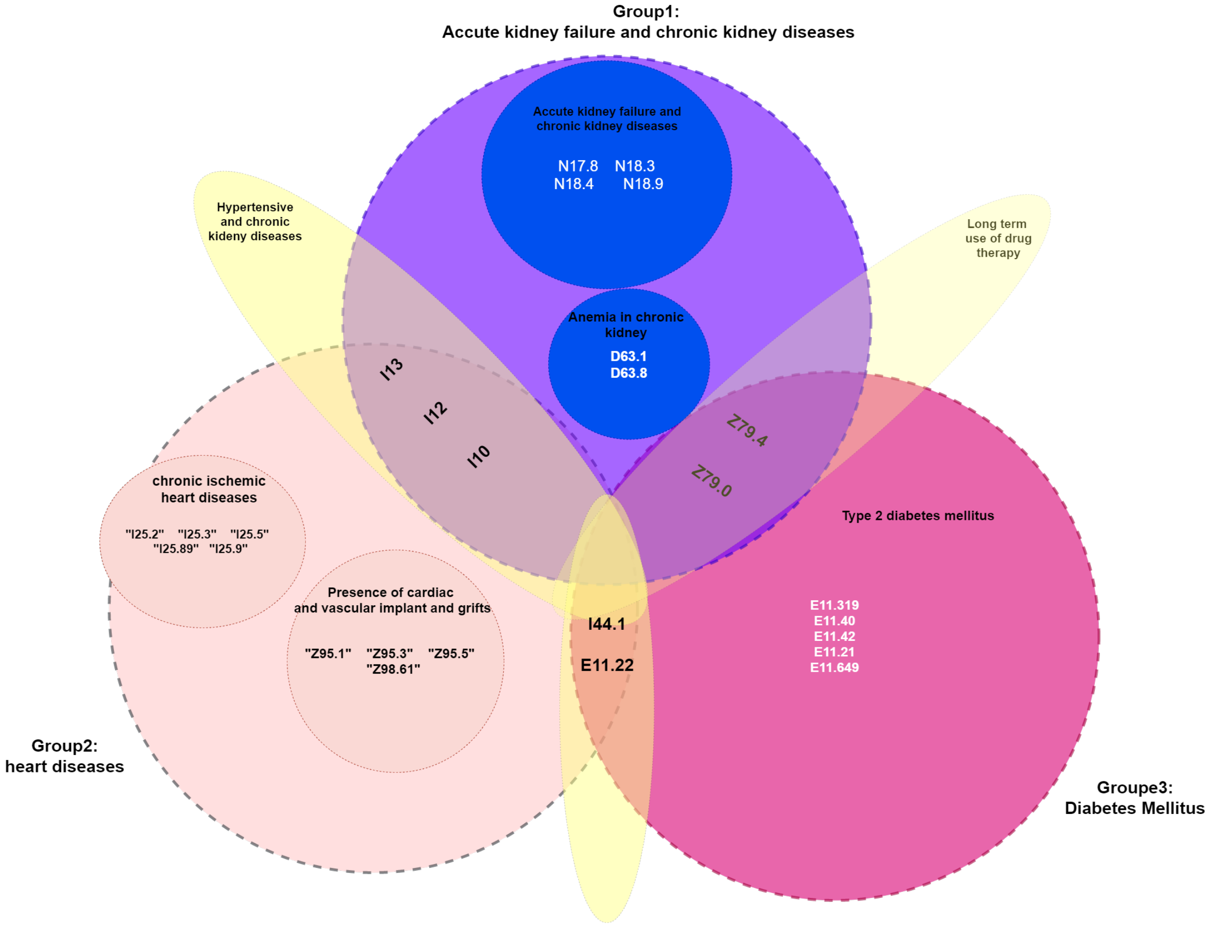
On the other hand, best performances for label propagation are when the label is initialized in the community's gravity center: initializing this algorithm in nodes with high closeness centrality can yields the best results, because label propagation is proportional to the node's capacity to propagate information flow across the network. Edge betweeness target edges that connect communities and could be of medical importance. Determining the communities boundaries can be helpful in the treatment of edges diseases and, thus, the prevention of potential complications. For example, hypertension can act as edge diseases between cardiovascular and cerebral diseases, then targeting hypertension for a patient with cardiovascular diseases can help him avoid complications in cerebral related problems and putting another burden on him.

FIGURE 10. Some communities detection and interpretations which we propose to them.



In term of interpretability, we found some communities we think easily interpretable (see figure 10), whereas others consist of heterogeneous entities, which we think that reflect the complex interactions between diseases. Figure 11 present an example of interactions between three examples of communities detected by Louvain algorithm. Interactions are presented as overlapping circles. Each circle presents an interpretation of detected community and diseases in edges (colored in yellow ellipses) can be thought of as bridging diseases. For example, type 2 Diabetes Mellitus and Cardiovascular Diseases could be linked because of genetic and epigenetic factors linking [40]. Hypertensive kidney and heart diseases [41] have hypertensive perturbation as common, and long term use of insulin can be one reason that relates kidney and diabetes [42]. See figure 11.

FIGURE 11. Example of three overlapped communities. For each group we proposed an interpretation. Yellow ellipses are an interpretation and a proposed explanation of the reason behind intersection parts detected of these three communities. Here some code traductions. Z79.4 = Long term (current) use of insulin. Z79.0 = Long term (current) use of anticoagulants and antithrombotics/antiplatelets , I13 = Hypertensive heart and chronic kidney disease, I12 = Hypertensive chronic kidney disease, I10 = Essential (primary) hypertension.



As a main conclusion, aging process do not increases additional condition average burden, but also can have effect on complexity alongside with quantity. This is important theoretically, since aging process will be considered as an example of a hidden Multimorbidity Mechanism that generates Multimorbid datasets, which is very important to suppose before any analysis of this type of data.

7. Conclusion

We based our work by the conviction that psycho-social characteristics have to be taking into account when managing patients with multimorbidity, in order to valorize further human capital. We tried to investigate how diseases co-occur based on life stages corresponding to a psycho-social point of view. We classified these profiles based on Erickson theory of life stages. First, we gave a formal definition of detecting Multimorbidity Disease Network. We focused on pairwise associations, which output a Comorbidity Disease Network (CDN). Second, we presented how some centralities measures evolve over life stages the CDN, and we compared four algorithms that detect components/ communities in CDN. The main results indicate that processes of aging do not increase Multimorbidity in term of diseases count, but also in complexity of Multimorbidity and we showed qualitative and quantitative visualizations of such phenomenon. This work needs to be verified for generalizability for other empirical data, besides further investigation of the suited centralities for multimorbidity context.

Conflict of Interests

The author(s) declare that there is no conflict of interests.

References

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

[2] M. Rijken et al., How to improve care for people with multimorbidity in Europe? Copenhagen (Denmark): European Observatory on Health Systems and Policies, 2017. [Online]. available on: http://www.ncbi.nlm.nih.gov/books/NBK464548/

[3] R. Pastorino, C. De Vito, G. Migliara, K. Glocker, I. Binenbaum, W. Ricciardi, S. Boccia, Benefits and challenges of Big Data in healthcare: an overview of the European initiatives, Eur. J. Public Health. 29 (2019), 23–27.

[4] J. Suls, P.A. Green, C.M. Boyd, Multimorbidity: Implications and directions for health psychology and behavioral medicine, Health Psychol. 38 (2019), 772–782.

[5] T.T. Makovski, S. Schmitz, M.P. Zeegers, S. Stranges, M. van den Akker, Multimorbidity and quality of life: Systematic literature review and meta-analysis, Ageing Res. Rev. 53 (2019), 100903.

[6] G.A. Orenstein, L. Lewis, Eriksons Stages of Psychosocial Development, in StatPearls, Treasure Island (FL): StatPearls Publishing, 2021. Accessed: 13 jul, 2021. [Online]. Available on: http://www.ncbi.nlm.nih.gov/books/NBK556096/

[7] H.A. Droogleever Fortuyn, R. Fronczek, M. Smitshoek, S. Overeem, M. Lappenschaar, J. Kalkman, W. Renier, J. Buitelaar, G.J. Lammers, G. Bleijenberg, Severe fatigue in narcolepsy with cataplexy: Severe fatigue in narcolepsy with cataplexy, J. Sleep Res. 21 (2012), 163–169.

[8] C. Salisbury, L. Johnson, S. Purdy, J.M. Valderas, A.A. Montgomery, Epidemiology and impact of multimorbidity in primary care: a retrospective cohort study, Br. J. Gen. Pract. 61 (2011), e12–e21.

[9] R. Bonita, R. Beaglehole, T. Kjellström, et W. H. Organization, Basic epidemiology. World Health Organization, 2006. [Online]. Available On: https://apps.who.int/iris/handle/10665/43541.

[10] L. Batstra, E.H. Bos, J. Neeleman, Quantifying psychiatric comorbidity: Lessions from chronic disease epidemiology, Soc. Psychiatry Psychiatr. Epidemiol. 37 (2002), 105–111.

[11] F.S. Roque, P.B. Jensen, H. Schmock, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts, PLoS Comput Biol. 7 (2011), e1002141.

[12] D.L. Vetrano, K. Palmer, A. Marengoni, et al. Joint action advantage wp4 group, frailty and multimorbidity: a systematic review and meta-analysis, J. Gerontol.: Ser. A. 74 (2019), 659–666.

[13] N.K. Schiltz, D.F. Warner, J. Sun, et al. Identifying specific combinations of multimorbidity that contribute to health care resource utilization: an analytic approach, Med. Care. 55 (2017), 276–284.

[14] C. Violán, A. Roso-Llorach, Q. Foguet-Boreu, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis, BMC Fam. Pract. 19 (2018), 108.

[15] A. Marengoni, A. Roso-Llorach, D.L. Vetrano, et al. Patterns of multimorbidity in a population-based cohort of older people: sociodemographic, lifestyle, clinical, and functional differences, 75 (2020), 798–805.

[16] Y.-P. Wang, B.P. Nunes, B.M. Coêlho, et al. Multilevel analysis of the patterns of physical-mental multimorbidity in general population of são paulo metropolitan area, Brazil, Sci. Rep. 9 (2019), 2390.

[17] B. Hernández, R.B. Reilly, R.A. Kenny, Investigation of multimorbidity and prevalent disease combinations in older Irish adults using network analysis and association rules, Sci. Rep. 9 (2019), 14567.

[18] A. Aguado, F. Moratalla-Navarro, F. López-Simarro, V. Moreno, MorbiNet: multimorbidity networks in adult general population. Analysis of type 2 diabetes mellitus comorbidity, Sci. Rep. 10 (2020), 2416.

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

[20] M. Lappenschaar, A. Hommersom, P.J.F. Lucas, Probabilistic causal models of multimorbidity concepts, AMIA. Annu. Symp. Proc. 2012 (2012), 475-484.

[21] J. Pearl, Éd., The morgan kaufmann series in representation and reasoning, in Probabilistic Reasoning in Intelligent Systems, San Francisco (CA): Morgan Kaufmann, 1988, p. i. https://doi.org/10.1016/B978-0-08-051489-5.50001-1.

[22] M. Lappenschaar, A. Hommersom, J. Lagro, P.J.F. Lucas, Understanding the co-occurrence of diseases using structure learning, in: N. Peek, R. Marín Morales, M. Peleg (Eds.), Artificial Intelligence in Medicine, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013: pp. 135–144.

[23] M. Lappenschaar, A. Hommersom, P.J.F. Lucas, J. Lagro, S. Visscher, J.C. Korevaar, F.G. Schellevis, Multilevel temporal Bayesian networks can model longitudinal change in multimorbidity, J. Clinic. Epidemiol. 66 (2013), 1405–1416.

[24] M.L.P. Bueno, A. Hommersom, P.J.F. Lucas, M. Lobo, P.P. Rodrigues, Modeling the dynamics of multiple disease occurrence by latent states, in: D. Ciucci, G. Pasi, B. Vantaggi (Eds.), Scalable Uncertainty Management, Springer International Publishing, Cham, 2018: pp. 93–107.

[25] Z. Khorrami, M. Rezapour, K. Etemad, et al. The patterns of non-communicable disease multimorbidity in iran: a multilevel analysis, Sci. Rep. 10 (2020), 3034.

[26] C. Madlock-Brown, R.B. Reynolds, Identifying obesity-related multimorbidity combinations in the United States, Clin. Obes. 9 (2019), e12336.

[27] M. van den Akker, F. Buntinx, J.F.M. Metsemakers, S. Roos, J.A. Knottnerus, Multimorbidity in General Practice: Prevalence, Incidence, and Determinants of Co-Occurring Chronic and Recurrent Diseases, J. Clinic. Epidemiol. 51 (1998), 367–375.

[28] L.S. Lim, E. Lamoureux, S.M. Saw, W.T. Tay, P. Mitchell, T.Y. Wong, Are myopic eyes less likely to have diabetic retinopathy? Ophthalmology. 117 (2010), 524–530.

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

[30] J. Bonis, drbonis/CMBD_MAD_2016. 2019. [Online]. Available on: https://github.com/drbonis/CMBD_MAD_2016.

[31] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Nat. Acad. Sci. 99 (2002), 7821–7826.

[32] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E. 76 (2007), 036106.

[33] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. 2008 (2008), P10008.

[34] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: pInar Yolum, T. Güngör, F. Gürgen, C. Özturan (Eds.), Computer and Information Sciences - ISCIS 2005, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005: pp. 284–293.

[35] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry, 40 (1977), 35-41.

[36] G. Sabidussi, The centrality index of a graph, Psychometrika, 31 (1966), 581-603.

[37] R.W. Floyd, Algorithm 97: Shortest path, Commun. ACM, 5 (1962), 345.

[38] U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. 25 (2001), 163-177.

[39] M.E.J. Newman, Mathematics of Networks, in: Palgrave Macmillan (Ed.), The New Palgrave Dictionary of Economics, Palgrave Macmillan UK, London, 2008: pp. 1–8.

[40] S. De Rosa, B. Arcidiacono, E. Chiefari, A. Brunetti, C. Indolfi, D.P. Foti, Type 2 diabetes mellitus and cardiovascular disease: genetic and epigenetic links, Front. Endocrinol. 9 (2018), 2.

[41] S.M. Hamrahian, B. Falkner, Hypertension in Chronic Kidney Disease, in: Md.S. Islam (Ed.), Hypertension: From Basic Research to Clinical Practice, Springer International Publishing, Cham, 2016: pp. 307–325.

[42] A.H. Rubenstein, M.E. Mako, D.L. Horwitz, Insulin and the Kidney, Nephron. 15 (1975), 306–326.