Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2022, 2022:53
https://doi.org/10.28919/cmbn/7504
ISSN: 2052-2541

DEFINING AND ANALYSIS OF MULTIMORBIDITY PATTERN OF DISEASES USING MARKOV RANDOM FIELD APPROACH: A COMPARATIVE ANALYSIS

FAOUZI MARZOUKI*, OMAR BOUATTANE

Laboratory of Electrical Engineering and Intelligent Systems, ENSET, Mohammedia,

Hassan II University of Casablanca, Morocco

Abstract. Aim: Multi-morbidity remains poorly understood due to the multifactorial complexity of this phenomenon and the lack of a standardized methodology for building and analysing Multimorbidity network. A comparative analysis of methods of modeling Multimorbidity network in literature may help to understand the pros and cons of these methods, then to facilitate a consensus about a standardized methodology. We propose to study two approaches for building Multimorbidity network focusing in their technical specificities.

Subject and Methods: We propose to model Multimorbidity using Ising Model, a Markov Random field based approach, and to compare its performance to the approach consisting in building a network of co-occurence using pairwise association strength estimated by Multimorbidity Coefficient. Besides, we illustrate how to use network science techniques to extract structural knowledge from Multimorbidity network.

Results: The results show that the Ising model is able to detect a similar structural patern as the approach of computing Multimorbidity coefficient for all paires of diseases. An evaluation of the stability and precision of the obtained comorbidity network has proved its reliability.

Conclusion: Defining methods and algorithms of detecting Multimorbidity network in formal language may help interdisciplinary cooperative research. Ising Model is a machine learning based on a probabilistic formalism

*Corresponding author

E-mail address: faouzi8marzouki@gmail.com

Received May 16, 2022

capable of detecting the same pattern as traditional approaches in Multimorbidity research literature. Understanding how diseases co-occur at the same time will help physicians to reason on multimorbidity burden as a complex system rather than reasoning on diseases as single and isolated entities.

Keywords: multimorbidity; comorbidity; centrality analysis; machine learning; graph theory.

2010 AMS Subject Classification: 92C42.

1. INTRODUCTION

With the increase of the average life expectancy, the aging phenomenon has led to a substantial increase in chronic diseases, therefore rising the prevalence of multimorbidity. Multimorbidity, i.e. two or more than two diseases in the same patient are diagnosed at the same time [37], is a significant health problem in modern medicine. It has been associated with poor prognosis, lower quality of life [29], increased health care costs, polypharmacy and the risk of premature death [11]. The management of multimorbidity is a complex process, and has become an emerging priority for public healthcare professionals. Unfortunately, multimorbidity remains not well understood due to its multifactorial complexity aspects, and to the health systems that are still designed in a single disease paradigm rather than multimorbidity. However, the transition from disease-centered care, to patient-centered care is ongoing [37]. Recently, increasing initiatives to exploit data-driven techniques and the increasing amount of electronic healthcare data [33] are taken to get more insight into this phenomenon.

There is a debate in the literature which has called for consistent and replicable methodology for the study of multimorbidity [18, 32, 45]. In a recent review, Jones et al. [19] highlighted the lack of a consensus in defining and measuring Multimoridity which results in no recommended standard method for calculating networks in multimorbidity. One way to facilitate a consensus about a standard methodology is to express the problem in a formal langage. In this work we try to adress this problem by proposing a formal definition of building a network of Multimorbidity, formal definitions helps in expliciting definitions and hypothesis about the problem and thus building common terminologies for researchers that may facilitate communicating their findings. Further, comparative analysis of different approaches in literature can reveal better understanding of pros and cons of each approach.

In particular, we propose to model comorbidity pattern detection using Ising model, a pairwise Markov random field-based approach (pMRF). It is a machine learning algorithm that estimates from binary data an undirected network of conditional dependences using regularization techniques. In addition, we compare the proposed approach with a baseline algorithm based on estimating Multimorbidity Coefficient (MC) between all pairs of diseases (will be denoted $MC-Algorithm$ in the paper). The studied methods will be applied on a case study of real medical data to detect comorbidity pattern of some selected valvular heart related diseases. After performing the two proposed approaches, we will compare the outcomes of their corresponding algorithms. Besides, we will analyze structural information characteristics revealed by the obtained networks. The results show that Ising and $MC-Algorithm$ outputted the same Comorbidity Disease Network. Further, we will illustrate how a mesoscopic analysis of the obtained multimorbidity network/pattern can be used to suggest an individual care strategy to manage patients who suffer from multimorbidity.

In section 2 we review literature related to the multimorbidity modeling. Section 3 presents an overview of the studied diseases in this work. Section 4 is devoted to data and methods used in this work. We define mathematically automatic detection of Multimorbidity pattern problem, and then develop its corresponding algorithm. We present and discuss the obtained results in section 5 before concluding in section 6.

2. Related Works

Recently, deep studies in medical literature were conducted to tackle multimorbidity burden, to explore its risk factors, its impact on quality of life in terms of mortality, costs and healthcare utility [46]. More technically, the methods and models differ either on the data, (cross sectional, temporal dimension, etc.), or whether the goal of the model is to explain, explore or to predict.

Earlier medical research relied on regression models, which were applied on single diseases and which ignore the hidden structure of the multimorbidity complexity. Recently, combinations of traditional data analysis and machine learning were proposed as multimorbidity research methods. In [49] 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 affect costs and inpatient utilization.

In [47] 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 population groups from others. In [50] 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 [48] a multilevel analysis of the influence of individual and area level factors on patterns of physical–mental multimorbidity and healthcare used in the general population. Applying this method allows detecting the isolated and combined influence of variables of each level on the outcome variables.

Other approaches in literature focused on probabilistic formulation and longitudinal data. In [51], Lappenschaar et al. summarize and classify some terminologies used in definitions of concepts of multimorbidity. They proposed a probabilistic framework to model these concepts using causal Bayesian network [54]. In [52], the authors proposed Bayesian network structure learning methods for modeling the interactions between risk factors explaining co-occurrences of malignant tumors in oncological area. This model was extended with a temporal dimension in [53]. Authors in [55] proposed a latent-based approach to model multimorbidity related events in temporal electronic health records. They introduced the notion of clusters of hidden states allowing the exploration of multiple dynamics that underlie events in data.

Network science is a relatively new approach to investigate Multimorbidity. To build Multimorbidity network, researchers estimate association strengths between diseases like Salton Cosine Index [20], odd ratio [1], Relative Risk [35], the standardized lift and confidence scores of the association rules as a probabilistic measuring of how conditionally the diseases are related [17]. Then the obtained network can be analyzed to reveal some structural caracteristics using for instance weighted degree, closeness and betweenness centrality [20], clustering coefficient, Page Rank and degree centrality [1], community detection algorithms [17].

In this work, we propose to study two methods to build Multimorbidity network. The first based on estimating Multimorbidity Coefficient (MC) as association strength of all paires of diseases. The second is based on estimating an Ising Model for the co-occurence of the diseases.

Ising model was used as a data analytic model to estimate dependencies between binary variables [14]. This method is becoming frequentely in psychology [38]. For example, it was

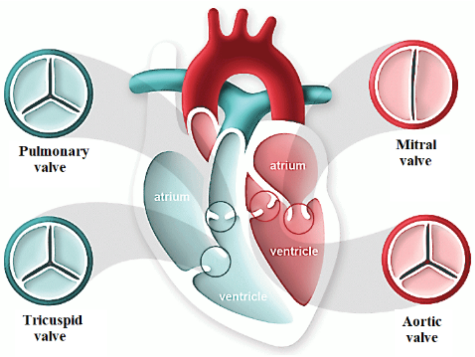
used to model theoretical assumptions of political beliefs, attitudes, and depression [6, 9]; and as an analytic tool in psychometric network [5, 22, 43]. We apply the two methods to build network co-occurence of valvular heart diseases, then we investigate if they output the same Multimorbidity network.

3. COMORBID VALVULAR HEART DISEASES OVERVIEW

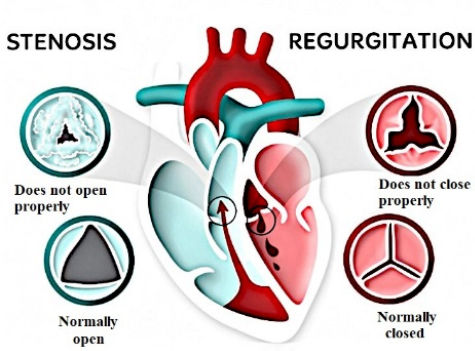
The human heart is viewed as a pump consisting of four chambers and four valves keeping enough blood flowing in a one-way direction: mitral, pulmonary, tricuspid and aortic as in Figure 1a. During a heartbeat, valves open to let blood flow from their chamber and close to stop the blood flowing backwards. Diseased or damaged valves impair the heart function [10]; this makes heart muscles become overworked and cannot pump properly. This may generate other problems like pulmonary hypertension, heart failure, stroke and others [16, 26, 28, 39, 41].

In Figure 1 there are two types of heart valve dysfunction: valvular stenosis (Figure 1a), and regurgitation or insufficiency (Figure 1b). Valvular stenosis refers to narrowing in the valve, which does not open enough and blood flow is therefore slowed. Insufficiency (leakage) occurs when the valve doesn't close properly and so the heart has to work harder to work properly.

Valvular heart diseases can have various causal patterns: degenerative in origin, inflammatory or bacterial infection, like streptococcus pyogenes which can cause, in long term, rheumatic valvular heart diseases [10]. These functional disorders are accompanied by dilatation and cardiac fatigue: shortness of breath and risk of edema of the lower extremities, malaise, sometimes loss of consciousness, palpitations, congestive heart failure. Furthermore, multimorbid patients will suffer from increasing burden and lowering quality of life. Therefore, understanding the tendency of comorbidity between these diseases can help healthcare systems to anticipate valvular heart disease patients' needs and reduce unnecessary charges in managing multimorbid patients profiles.



(A) Normal valvular components



(B) Valvular stenosis and insufficiency (regurgitation)

FIGURE 1. Heart anatomy schematic, from [44] website

4. METHODOLOGY

4.1. Problem Setting.

4.1.1. *Problem Overview.* This section presents the mathematical formulation of the problem. We encourage the reader to read the appendices (Section 7) to take an overall view of the notations and abreviations used in this methodological section.

Let us start with the following assumptions: Let $|S|$ denotes the number of elements of a set S. Let $D = \{d_1, d_2, \ldots, d_{|D|}\}$ be a finite set containing $|D|$ number of diseases present in a medical dataset.

Let R be a k-ary relation over Cartesian product sets D^k. The diseases $d_1, d_2 \ldots d_k$ are related by the relation R if and only if they satisfy a predefined condition that depends on the context of the study. It can be for example the fact of being correlated, causally related, being in the same category or conditionally related. In this work R represent multimorbidity relation over k diseases. The relation R defines an undirected weighted hypergraph $G = (D, R)$ such that the vertices D are the nodes/diseases and hyperedges R represent the multimorbid diseases. The weights of the hypergraph are a measure of the strength of co-occurrence of these k multimorbid diseases in data. Each hypergraph G represents a Multimorbidity pattern. We define the Automatic Multimorbidity Pattern Detection problem (AMPD) as the conceptualization and the study of algorithms that automatically detect hidden Multimorbidity patterns given a medical data set.

We use in this work a probabilistic framework to model AMPD problem. Let $X = \{X_1, X_2, \ldots, X_{|X|}\}$ represent the set of $|X|$ patients. Each patient is characterized by a set of a tuple of diagnoses $X_i = (x_{1,i}, x_{2,i}, .., x_{|X_i|,i})$ where $x_{j,i} \in D$ is a random variable for the i^{th} patient of the j^{th} diagnosis. We suppose that X are random variables that are independent (in general it is supposed in the case in cross-sectional data like our data of application) and identically distributed (i.i.d) samples (the data are governed by the same underlying Multimorbidity Mechanism described by the joint probability distribution noted P^* over the variables X). The data X can be indexed by diseases $X = \{X_d | d \in D\}$ such that X_d is a binary random variable of the presence (or absence) of disease $d \in D$ over the patients (In the following the random variable X_d and the node/disease d will be used interchangeably).

Statistically, R is estimated in function of observations X. We consider the hypothesis space $H = \{R_\theta(X)|\theta$ the parameters of the model.$\}$. Given a family of models $R_\theta(X)$, our task is to learn some models $R^*_\theta(X) \in H$ that best fit the distribution P^* from which our data X were sampled. This is done by minimizing an expected loss function $L(X, R_\theta)$ which measures the loss that a model distribution R_θ makes on input observation X.

In this work, we focus on the special case of learning Comorbidity Disease Network (CDN) (thus learning an ordinary graph with ordinary edges). We define the binary relation between comorbid diseases using two approaches: Multimorbidity Coefficient based on strength computation approach and probabilistic dependency-based approaches. We will use the first approach as a baseline to evaluate the second proposed approach.

4.1.2. *Multimorbidity Coefficient-based approach.* Let $P(d_i)$ stands for the occurrence probability of the disease $d_i \in D$. The disease d_i is represented by the binary random variables for the presence (or absence) of the i^{th} disease. $P(d_1, d_2)$ stands for the occurrence probability of the diseases d_1 and d_2 at the same time.

We use Van Den Akker et al. definition of cluster comorbidity [42]: if d_1 has occured, then d_2 will be more likely to occur than what would be expected just by chance. We consider d_1 and d_2 are in positive comorbidity, i.e., they tend to appear together, if $P(d_1, d_2) > P(d_1)P(d_2)$. If $P(d_1)P(d_2) = P(d_1, d_2)$ we consider that the co-occurrence of the two diseases is what would be expected just by chance. 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 instance, myopia may be protective against diabetic retinopathy [24]).

To measure how strongly disorders are associated, a multimorbidity coefficient (MC) is calculated. MC is commonly used method for measuring pairwise association [2]. MC is defined as the division of observed rate of comorbidity (multimorbidity) by the rate which is expected under the null hypothesis of no association between the separate disorders. Using the Table 1 notations, the MC score for Disease 1 and Disease 2 is equal to:

$$MC = \frac{\frac{a}{N}}{\frac{a+c}{N} \cdot \frac{a+b}{N}} = \frac{aN}{(a+c) \cdot (a+b)} \tag{1}$$

TABLE 1. Cases of co-occurence of two diseases. For example, the number of cases in which disease 1 and disease 2 co-occur at the same time is a. The number of cases in which disease 1 is present and disease 2 is absent is 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

We implement this pairwise approach to learn the weighted undirected structure of the CDN (See MC-Algorithm in section 4.2). One of the advantages of this approach is its interpretability, since it mimics a clinical intuition: to decide if two diseases are comorbid, look for their co-occurrence frequency and decide, based on a threshold, if a multimorbidity pattern can be detected.

4.1.3. *Conditional dependence based approach.* Another possible definition of the binary relation R to model structure of CDN, rather than the association strength concept, is to fit inter-diseases probabilistic dependences/independences structures and search for the optimal structure given a Loss measure.

Let $\mathscr{P}(\mathscr{D})$ be the power set of diseases/binary random variables. Given the graph $G = (D,R)$, the set of random variables $(X_d)_{d\in D}$ form a Markov Random field in respect to G if any two subsets of variables are conditionally independent given a separating subset:

$$X_A \perp_P X_B | X_S \;\; A,B,S \in \mathscr{P}(\mathscr{D}) \tag{2}$$

Where every path from a node in A to a node in B passes through S. Let $P(X = x)$ be the probability of finding that the random variables X take on the particular value x. We suppose that the hidden distribution P^* underlying our multimorbid data can be factorized in the form:

$$P^*(X = x) = P^*(x_1, x_2, \ldots, x_{|D|}) = \frac{1}{Z} \prod_{c\in R} \Psi_c(x_c) \tag{3}$$

where R denotes the set of hyperedges of G, and each factor Ψ_c is a non-negative function over the variables in a hyperedges/multimorbid diseases c from R. The partition function

$$Z = \sum_{x_1, x_2, \ldots, x_{|D|}} \prod_{c \in R} \Psi_c(x_c) \tag{4}$$

is a normalizing constant that ensures that the distribution sums to one.

Markov Random Fields (MRFs) can compactly represent independence assumptions that Bayesian network (directed acyclic graph) cannot represent and visualize a probability distribution in undirected graph terminology: A Pairwise Markov Random Field (pMRF) is a network in which nodes represent variables, connected by undirected edges indicating conditional dependence structure. Two variables that are not connected (i.e. no edge between them) verify the Markov property: two nodes are conditionally independent given the set of all other nodes in the network. Further, pMRFs are well defined and have no equivalent models, unlike Bayesian network [25]. Therefore, they facilitate a clear interpretation of the edge-weight parameters as strength of unique associations between variables, which in turn may highlight potential causal relationships. Besides, Conditional independencies are also to be expected in many causal structures [34].

A pMRF can be parameterized as a product of strictly positive potential functions Ψ_i for all nodes i and j from D such that $i \neq j$ [31]:

$$P^*(X = x) = \frac{1}{Z} \prod_{i \in D} \Psi_i(x_i) \prod_{i,j \in D} \Psi_{i,j}(x_i, x_j) \tag{5}$$

Where: $\Psi_i(x_i)$ is the node potential function that can map a unique potential for every possible realization of X_i. $\Psi_{i,j}(x_i, x_j)$ the pairwise potential functions that can likewise map unique potentials to every possible pair of outcomes for X_i and X_j.

4.2. Methods.

4.2.1. *Ising Model.* For the estimation of CDN, we used the Ising model to model the probability distribution for the comorbidity pattern of valvular heart disease and some related conditions. The Ising model can be used to estimate the pairwise Markov Random Field (pMRF)

for binary variables. Taking into account the equation 5, the potential functions are represented in Ising Model by log-linear model, such that: for all nodes $i \in D$: $ln\psi_i(x_i) = \alpha_i x_i$ and $ln\psi_{i,j}(x_i,x_j) = \omega_{i,j}x_i x_j$. This results in:

$$P^*(X = x) = \frac{1}{Z}\prod_{i\in D}\Psi_i(x_i)\prod_{i,j\in D}\Psi_{i,j}(x_i,x_j) \tag{6}$$

$$= \frac{1}{Z}e^{\left[\sum_{i\in D}\alpha_i x_i + \sum_{i,j\in D}\omega_{i,j}x_i x_j\right]} \tag{7}$$

Where Z is a normalizing constant and defined as:

$$Z = \sum_{x_1,x_2,\ldots,x_{|D|}} e^{\left[\sum_{i\in D}\alpha_i x_i + \sum_{i,j\in D}\omega_{i,j}x_i x_j\right]} \tag{8}$$

We can view the Ising model as a probability distribution that is governed by main effects α_i and pairwise interactions/edges $\omega_{i,j}$. The model can be represented by an undirected weighted graph $G = (D,R)$ such that edge weight $\omega_{i,j}$ are a real valued measures of dependence between nodes/diseases i and j:

$$\omega_{i,j} = \begin{cases} 0 & if(i,j) \notin R \\ a\ non\ zero\ real\ valued\ number & if(i,j) \in R \end{cases} \tag{9}$$

The higher (lower) $\omega_{i,j}$ becomes, the more nodes X_i and X_j prefer to be in the same (different) state (they tend to be both present or both absent in the same patient at the same time) and α_i can be viewed as a threshold parameter that denotes the tendency for node i to be in some state (the tendency to be present or absent in a patient). Ravikumar et al. used L1 regularized logistic regression [36] to estimate the structure of the Ising model (known as the least absolute shrinkage and selection operator [40]). The pseudo likelihood is used to approximate the full likelihood. For each node i, the expression which is maximized is [15]:

$$\max_{\alpha_i,\omega_i} Likelihood_i(\omega_i,\alpha_i,x) - \lambda Pen(\omega_i) \tag{10}$$

The pseudolikelihood PL approximates the likelihood with the product of univariate conditional likelihoods:

$$lnPL = \sum_{i=1}^{|D|} Likelihood_i(\omega_i,\alpha_i,x) \tag{11}$$

Where ω_i is the i^{th} row (or column due to symmetry) of the Ising structure network ω. The λ is the regularization tuning parameter and $Pen(\omega_i)$ denotes the penalty function, which is defined in terms of the LASSO as follows:

$$Pen(\omega_i) = \| \omega_i \|_1 = \sum_{j=1, j\neq i}^{|D|} |\omega_{i,j}| \tag{12}$$

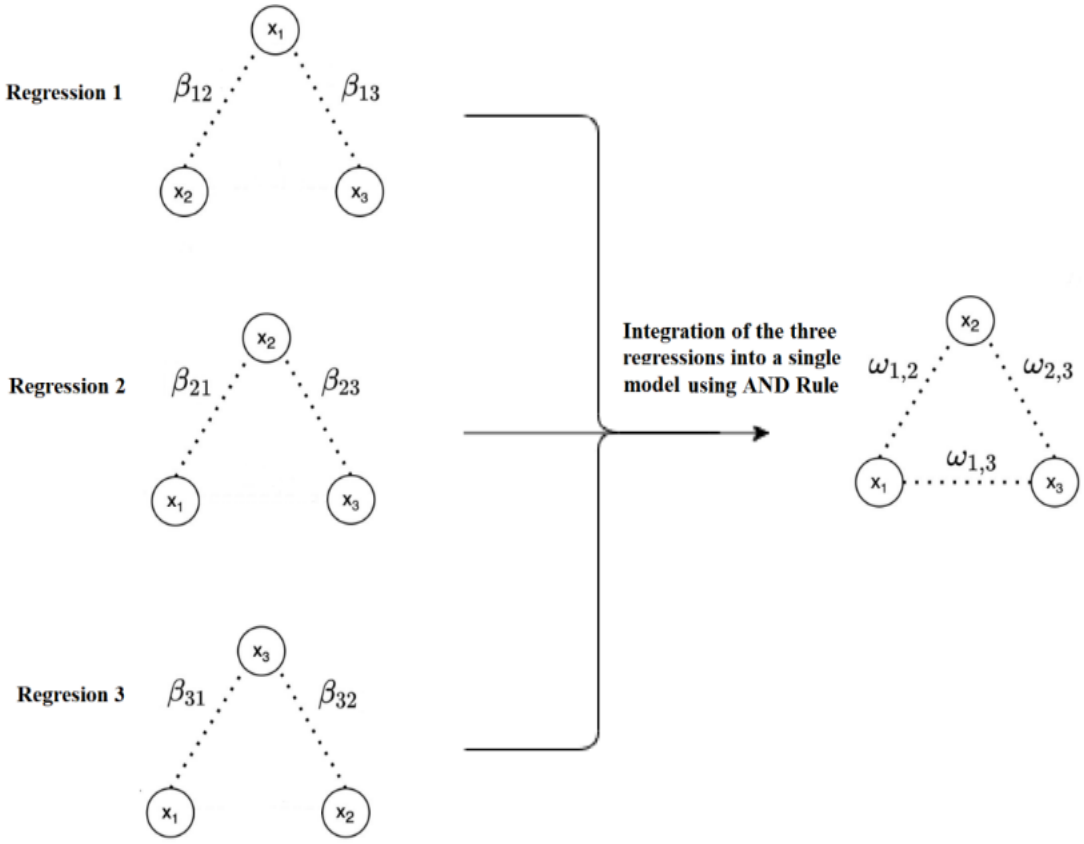


FIGURE 2. With a pseudo-likelihood estimation, we first estimate the neighborhood of each diseases node. This is performed through one logistic regression per node (here, we consider a simple three-node example). We combine the neighborhoods into a single network model whose weight matrix is ω, through the AND rule: an edge is present if both $\beta_{i,j}$ and $\beta_{j,i}$ are non-zero. This step is necessary because each node is both the dependent and independent variables; hence, we have two β estimates: $\beta_{i,j}$ and $\beta_{j,i}$.

$$\omega_{i,j} = \begin{cases} \frac{1}{2}\beta_{i,j}\beta_{j,i} & if \beta_{i,j} \neq 0 \; and \; \beta_{j,i} \neq 0 \\ 0 & otherwise \end{cases} \tag{13}$$

In Figure 2, this conditional probability is shown for each node of a simple three-node network. More importantly, if x_1, x_2 and x_3 are observed variables, this equation translates directly into a logistic regression. Hence, we can estimate β and α with one regression per node. For example, in Figure 2 network estimation is performed on three nodes by three regressions:

$$Regression\ 1: \; P(x_1|x_2,x_3) = \frac{x_1x_2\beta_{1,2}+x_1x_3\beta_{1,3}+\alpha_1x_1}{1+e^{x_1x_2\beta_{1,2}+x_1x_3\beta_{1,3}+\alpha_1x_1}} \tag{14}$$

$$Regression\ 2: \; P(x_2|x_1,x_3) = \frac{x_2x_1\beta_{2,1}+x_2x_3\beta_{2,3}+\alpha_2x_2}{1+e^{x_2x_1\beta_{2,1}+x_2x_3\beta_{2,3}+\alpha_2x_2}} \tag{15}$$

$$Regression\ 3: \; P(x_3|x_1,x_2) = \frac{x_3x_1\beta_{3,1}+x_3x_2\beta_{3,2}+\alpha_3x_3}{1+e^{x_3x_1\beta_{3,1}+x_3x_2\beta_{3,2}+\alpha_3x_3}} \tag{16}$$

A caveat is that we will have two estimates per edge, $\beta_{i,j}$and $\beta_{j,i}$, because each node serves both as the dependent and independent variables. However, they will converge as sample size goes to infinity. To solve this and complete the weight matrix ω, researchers in [43] used the and-rule: any edge is the average β if both $\beta_{i,j}$and $\beta_{j,i}$ are non-zero, otherwise $\omega_{i,j} = 0$. This multiple logistic regression approach inflates the rate of false positives, because of performing multiple testing. Authors in [43] used LASSO regularization (Least Absolute Shrinkage and Selection Operator) to overcome this problem. The LASSO procedure shrinks estimates towards zero, lowering the amount of detected edges and their strengths [40], and thus dropping out of the model the spurious edges letting just interpretable and important edges [43], limiting effect on over-fitting for smaller samples and leading to better out-of-sample generalizability [12]. The amount of shrinkage depends on a hyper-parameter γ which determines whether false positives or negatives are preferred. Lower γ favours more edges, while higher γ favours stronger shrinkage and hence, sparser networks. Therefore, a low γ increases the false-positive rate, while a high γ inflates the false-negative rate. This tuning parameter γ can be selected by minimizing the Extended Bayesian Information Criterion (EBIC) [7], such that

$$EBIC = -2Likelihood(X) + |\omega_i| ln(|X|) + 2\gamma|\omega_i| ln(|D| - 1) \tag{17}$$

in which $|\omega_i|$ is the number of nonzero parameters in ω_i and $|X|$ is the number of observations. The EBIC has been shown to be consistent for model selection (e.g. in psychometric domain [3, 43]), and to perform best with hyper parameter $\gamma = 0.25$ for the Ising model [3].

4.2.2. *Multimorbidity Coefficient based Algorithm (MC-Algorithm).* In this section we define and implement the pairwise association strength computation methodology for building Comorbidity Disease Network which will be considered as a benchmark to compare to. We will call this algorithm $MC - Algorithm$.

Let N_{diag} denote the number of diagnoses and N_{dis} the maximum number of diagnoses per patient in the dataset. Let $D = \{d_1, d_2, d_3, ..., d_{N_{dis}}\}$ the disease set present in the dataset.

$M_k \subset D$ such $k > 1$, is the subset of size k diseases from D. e.g M_2 is the subset of possible comorbidities. Let $f : I \subset \mathbb{N} \to \{D_1, D_2, ..., D_{N_{diag}}\} \subset \mathscr{P}(\mathscr{D})$ be an application that maps every patient $i \in I$ to his recorded diagnoses $f(i) = \{x^i_1, x^i_2, ..., x^i_{N_{diag}}\}$. The MC-Algorithm search for $\frac{N_{dis}!}{(N_{dis}-2)!2!} = \mathscr{O}(N^2_{dis})$ comorbid distinct combinations and estimates co-occurrence strength by attributing MC for each combination. If the MC is significantly higher than 1 then the algorithm considers that these two diagnoses are in comorbidity. If the MC is significantly less than 1 then we consider that these two diagnoses are in protective comorbidity. The bigger this number is, the stronger the association is considered. We will focus our analysis on positive comorbidity only.

Algorithm 1 MC-Algorithm

Require: *a patient − diagnosis function map* $f : I \subset \mathbb{N} \to \{D_1; D_2; ...; D_{N_{diag}}\}$ *a disease set* D,

Ensure: *a Comorbidity Disease Network* $G = (V, E)$

for $M_2 \in \mathscr{P}(\mathscr{D})$ **do**
  $W_{expected} \leftarrow \prod_{d \in M_2} Count(\{d\}, I)$
  $W_{observed} \leftarrow Count(M_2, I)$
  $MC \leftarrow \frac{W_{observed} * N_{diag}}{W_{expected}}$
  if H_0 : "$W_{expected} \geq W_{observed}$" is rejected at risk α **then**
    $E_{d_1,d_2} \leftarrow MC$, such that $\{d_1, d_2\} = M_2$
  end if
end for

procedure $Count(M_k, I)$
  $S_{occurence} \leftarrow \emptyset$
  for $X \in M_k$ **do**
    for $i \in I$ **do**
      $S_{occurence} \leftarrow S_{occurence} \bigcup (X \bigcap f(i))$
    end for
  end for
  return $|S_{occurence}|$
end procedure

The $count(S : I)$ procedure counts incrementally the number of occurrences of a disease $d \in S$ in diagnosis records indexed by $i \in I$. a sequential search will count the number of occurrences by iterating over the diagnosis records $f(I)$ resulting in $\mathscr{O}(|X| max\{f(i)\})$. This algorithm can be easily generalized from comorbidity to Multi-morbidity using hypergraph formalism. Interested reader can be referred to [30].

4.3. Data. The analysis was applied in a case study of real medical dataset [4], a hospital inpatients' diagnosis dataset. Each diagnosis of an admitted patient is encoded by ehe Tenth

Revision of the International Classification of Diseases (ICD 10). The data contain 78451patients (34639 males, and 43812 females). The maximum number of registered diagnosis per admission is 20. This study is applied to the analysis of the following diseases node: Non-rheumatic mitral and tricuspid and aortic (valve) insufficiency (coded respectively in ICD10 as I34.0 and I36.1 and I35.1). Non-rheumatic aortic (valve) stenosis (I35.0). Rheumatic tricuspid insufficiency (I07.1). Rheumatic disorders of both mitral and tricuspid valves (I08.1). Combined rheumatic disorders of mitral, aortic and tricuspid valves (I08.3). Other pulmonary hypertension (I27.2). Other ill-defined heart diseases (I51.89). these diseases have the potential to exhibit causal relationships [26, 27, 39, 41].

4.4. Evaluation methodology. After performing Ising model to estimate pairwise Markov random field representing joint probability distribution for the studied valvular heart diseases, we compared the CDN outputted by Ising Model and MC-Algorithm (ω^{Ising} and ω^{MC-Alg}).

If we conceptualize the outputted network as a distance matrices such that the distance in each pair of diseases corresponds to similarities in their level of co-occurrence, ranging from 0 (does not co-occur) to 1 (co-occur almost always), then we can perform Mantel test to measure the similarity of the two outputted networks [23].

The Mantel test is used for correlation between two proximity matrices ω^{Ising} and ω^{MC-Alg} and tests the null hypothesis H_0: "proximity among diseases in the matrix ω^{Ising} are not linearly related to the corresponding proximity in the matrix ω^{MC-Alg}" against the alternative hypothesis H_1: "proximity among diseases in matrix ω^{Ising} are linearly correlated to the corresponding proximity in the matrix ω^{MC-Alg}". The Mantel statistic can be normalized to range between -1 and +1:

(18)
$$r = \frac{1}{n-1} \sum_{i}^{|D|} \sum_{j}^{|D|} \frac{\omega_{i,j}^{Ising} - average(\omega^{Ising})}{std(\omega^{Ising})} \frac{\omega_{i,j}^{MC-Alg} - average(\omega^{MC-Alg})}{std(\omega^{MC-Alg})}$$

with $i \neq j$ and i and j are the row and column indices and n the number of distances in one of the matrix ω without accounting for the diagonal. $std()$ is the standard deviation. The statistical significance of the Mantel coefficient can be tested by performing a permutation test. It consists in simulating the realizations of the null hypothesis by repeated permutations of the lines and

columns in one of the matrices ω^{Ising} and ω^{MC-Alg} and recomputing the Mantel statistic. The result is a sampling distribution of the Mantel statistic under the null hypothesis. If there is no relationship between the matrices, the observed r value is near the center of the sampling distribution, while if a relationship is present, one would expect the observed value to be more extreme than most of the values obtained by permutation.

After that, we assessed how accurate networks are estimated, and how stable inferences from the network structure are. We estimate the accuracy of edge weights using bootstrapped Confidence intervals (CI) [13]. Then we estimated if edge-weights/centralities significantly differ. This can be done by checking if zero value is in the constructed bootstrapped CI of differences of edge-weights/centralities [8]. Once a network is computed, centrality indices can give information on the importance of each node. We will consider three main centrality indices. 1) Strength centrality is defined as the sum of the absolute weights of the edges incident to a node, a node has a high strength if it has strong connections with many others. 2) closeness centrality is the inverse of the sum of the distances of the a node from all the others. A node has a high closeness centrality if it is well connected to other nodes either by strong direct edges or by short indirect paths. The closeness of each node depends on the connectivity of all others to the network: Increasing the distance of a node from the rest of the network makes the closeness of all other nodes increase as well. 3) Betweenness centrality is the number of times a node occuring in the shortest path between two nodes, thus quantifying how much a node is important for other nodes to affect each other.

While assessing accuracy of edges can be measured by bootstrapped Confidence interval, centrality measure can result in biased results. Researchers in [13] suggested to assess stability of centralities by the coefficient of stability (CS). The correlation stability coefficient (CS coefficient) is based on case dropping bootstrap and is defined as the maximum proportion of cases that can be dropped such that the resulting centrality estimate correlates more than 0.70 with the original centrality estimate with 95% probability. Cutoff values of 0.25 and 0.50, respectively, have been suggested to indicate sufficient stability and good stability [13]. The stability analysis indicates if the order of centrality indices does not change after re-estimating the Comorbidity network using less data.

5. Results and Discussion

We performed MC-Algorithm on the valvular heart diseases data mentioned in section 4.3. Visualizations are performed using R Studio software. We see in Figure 3 the weighted graph of detected comorbidity disease network for male patients (older adulthood more than 65 years). The MC-Algorithm outputted a graph with 12 non-zero edges over 36 possible edges. The thickness of edges shows visual differences of calculated MC weights between each pair of diseases. If an edge is absent between two pairs, this means either their association is random or not significant (the null hypothesis was not rejected at risk 0.01) or both cases. In Figure 3 for example, Non rheumatic tricuspid insufficiency co-occurs with non-rheumatic mitral insufficiency 20.8 times more than what would be expected just by chance. This high score can indicate a potential causal relationship, which can be explained by functional abnormalities in heart functioning. The magnitude of weights show the strength of relationship between the studied valvular heart related diseases. Some diseases have strong connections (for example I27.2 – I36.1, I27.2-I08.1 and I34.0-I36.1), whereas others have weaker connections (e.g. I34.0-I35.0 and I27.2-I35.1). The nodes I27.2 and I34.0 have the highest degree centrality in the graph, which suggests their potential importance in the obtained graph.

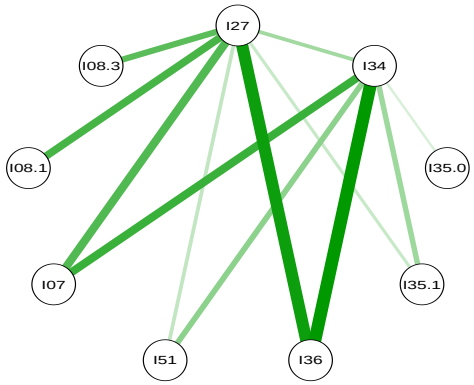


FIGURE 3. Comorbidity Disease Network detected by MC-Algorithm for males patients aged $\geq$ 65 years. See section 4.3 for more details about the abreviations of the codes. MC scores are presented by thickness of lines between nodes (See Table 2 for the values).

TABLE 2. MC scores computed by MC-algorithm (section 4.2.2) for the structure of co-occurrence of Fig. 3. Since the built graph is undirected, the weight matrix is symmetric (just the lower triangular is filled). The cases with "-" symbols mean that either the MC score is equal to one, or the null hypothesis is not rejected, or both.

I34.0	7.70							
I35.0	-	3.17						
I35.1	4.54	7.97	-					
I36.1	19.07	20.8	-	-				
I51.89	5.03	9.42	-	-	-			
I07.1	14.21	16.13	-	-	-	-		
I08.1	15.07	-	-	-	-	-	-	
I08.3	13.42	-	-	-	-	-	-	-
	I27.2	I34.0	I35.0	I35.1	I36.1	I51.89	I07.1	I08.1

To assess the performance of Ising Model to detect comorbidity in valvular hear diseases, we applied Ising Model to the same data as MC-Algorithm, and we compared the outputted weighted graphs. The Figure 4 bring together MC-Algorithm (Figure 4a), Ising Model (Figure 4b) and Polychoric correlation (Figure 4c) based graphs. To be able to compare in a fair way the outputs, we rescaled the weights $W_{i,j}$ of the graphs using Min-Max normalization to the same interval $[a,b]$:

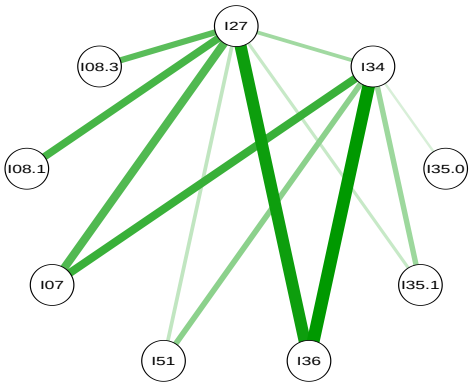
$$W_{i,j}^{transformed} = a - \frac{[W_{i,j}^{original} - min(W_{i,j}^{original})](b-a)}{max(W_{i,j}^{original}) - min(W_{i,j}^{original})} \tag{19}$$

Typically, these intervals reflect a scale between the absence (a=0) or the opposite direction of association (a=-1) to the full presence/positive strength (b=1). In Figure 4 the weights are rescaled in [a,b] = [0,1]. MC-Algorithm and Ising Model outputted the same structural comorbidity pattern, the same detected skeleton of the CDN and some differences between edge weights (12/36 non-zero edges and mean weight of 0.6702). Comparing weights of graph is a

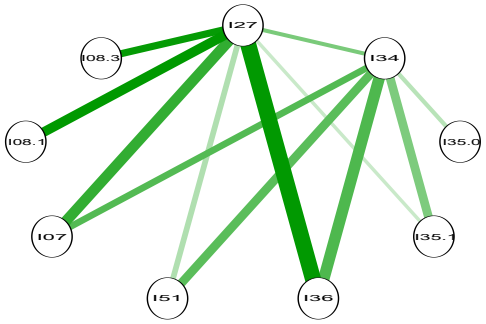
challenging task. The Figure 5a presents weight distribution for edges for MC-Algorithm and Ising Model both rescaled in [0,1]. A Two-sample Kolmogorov-Smirnov test resulted in D = 0.25, p-value = 0.8475, then we did not reject the null hypothesis that the samples are drawn from the same distribution. In Figure 5b we plotted the difference between MC-Algorithm and Ising Model adjacency matrices respectively. The difference fluctuates around Zero. If we consider the two adjacency matrices as a distance matrices between all pairs of diseases (diseases with strong weight edges have longer distance between them and vice versa), then a Mantel test between the two matrices can be used to test their similarity under the null hypothesis of there being no relation between the two matrices. We obtained significant p-value ($\leq$0.01) for Mantel score, thus we rejected the null hypothesis "there is no correlation between the weights of the two matrices ω^{Ising} and ω^{MC-Alg}".

As a conclusion, the outputted CDN structure by Ising and MC-Algorithm are significantly similar. However, without knowing the accuracy of the network structure and the stability of the centrality estimates, we cannot conclude whether the differences of centrality estimates are really interpretable or not.

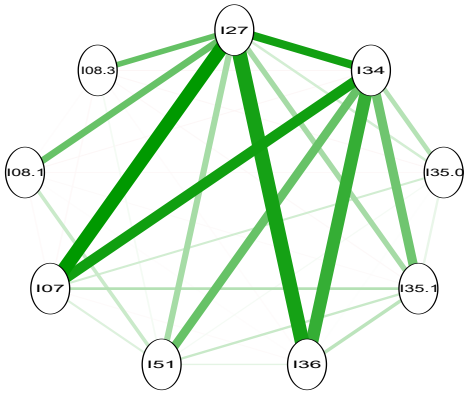
To assess the accuracy of the obtained CDN we used bootstrapped methods to construct confident intervals (CIs). Figure 6b shows the edge value estimated in the sample after performing 1000 bootstrapped networks. The lines surrounding the dots (i.e. means) indicate the width of the bootstrapped CIs. Many of edges are estimated as zero (e.g. I34.0-I08.1). Some edges are larger than zero, but their bootstrapped CIs contain zero value (e.g. I27.2-I51.89 and I34.0-I35.0). For a smaller number of edges, the estimates are larger than 0 and the CIs do not include zero (e.g. I27.2-I08.1 and I27.2-I34.0). Edges I27.2-I36.1 and I27.2-I08.1 are significantly stronger than almost all of other detected edges in the CDN. The absence of an edge does not present evidence that the edge is in fact exactly zero. Some edges were estimated to be zero but have bootstrapped mean more than zero (I51.89-I08.1, I35.1-I07.1, I35.1-I36.1, I35.1- I51.89, I35.0-I07.1, I27.2-I35.0). Due to shrinkage of LASSO, these small quantities were set to be zeros. Similarly, MC-Algorithm outputted these edges to be either not significant or MC score was equal to 1, i.e. they randomly co-occur. We can interpret these results by supposing that these edges being spurious edges and their occurrences can be due to chance or other con unmeasured



(A) MC-Algorithm



(B) Ising Model

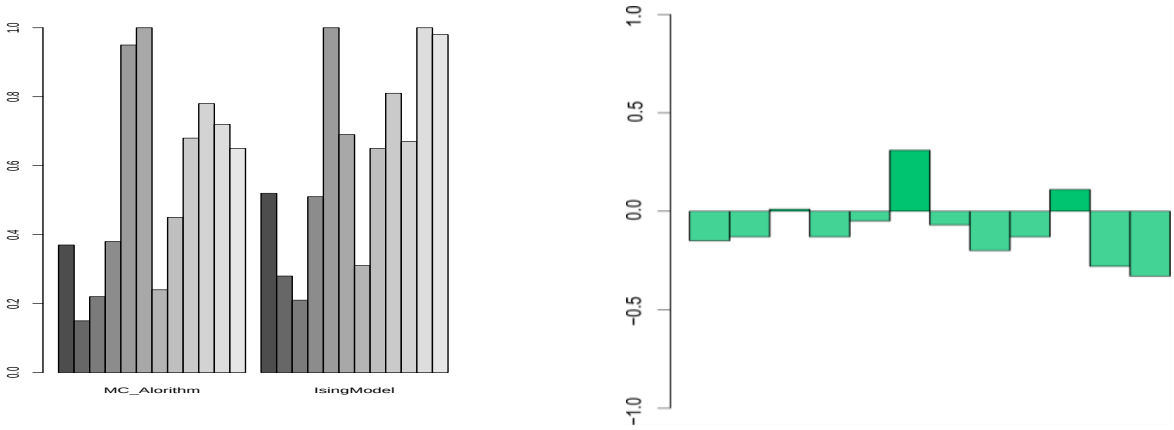


(C) polychoric correlation based graph.

FIGURE 4. Comparison of the outputted CDN for MC-Algorithm and Ising Model

confounding variable. This latter is supported by the fact that they do have connections in the Polychoric correlation network.

Other edges have low weights edges but their confidence interval contains zero, which suggests that they should be interpreted as weak edges (visually weak thickness in Figure 4). These edges were detected by all outputted networks. Besides, some edges have moderate weights and long bootstrapped Confidence Interval (e.g. I34.0-I36.1), which require more careful interpretations. Other edges have strong weights (I27.2-I08.1) and relatively narrow Bootstrapped CI (e.g. I27.2-I34.0). Many edges have overlapping Bootstrapped CI and different means, which suggests that these means may be not significantly different (e.g. I27.2-I08.1 and I27.2-I36.1).



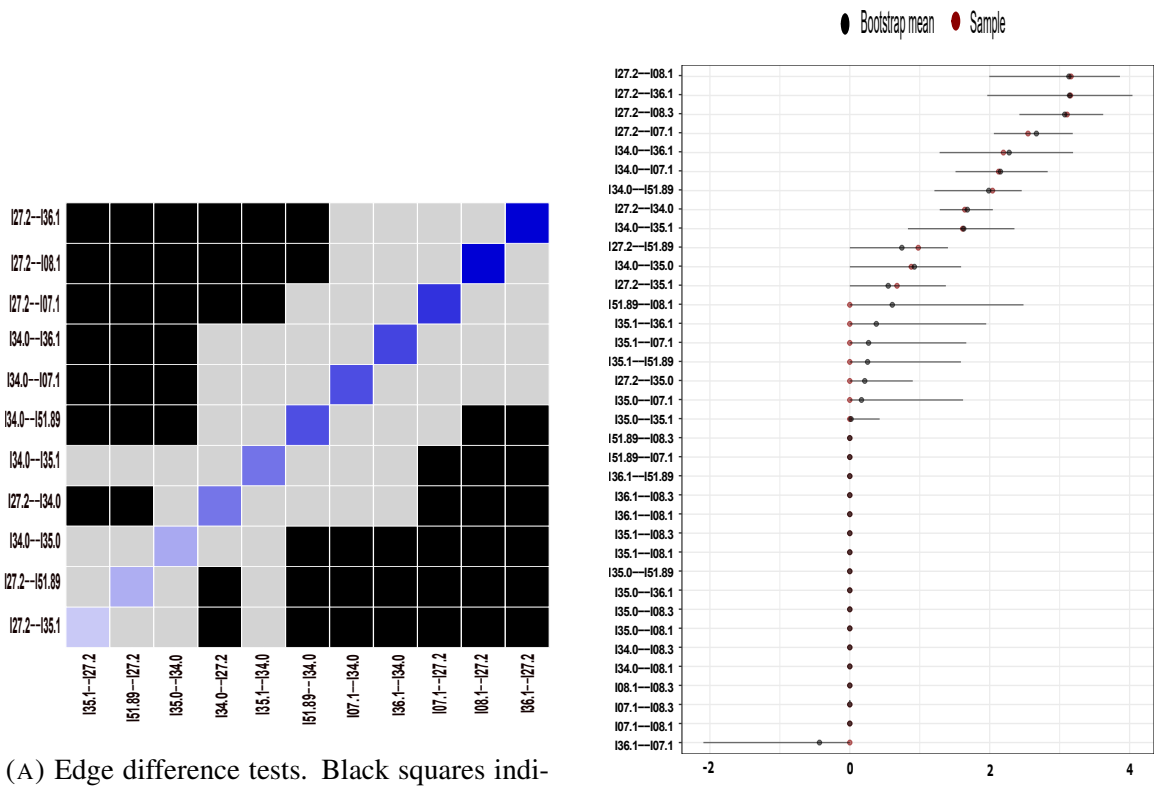
(A) Weight distribution for edges in for between MC-Algorithm and Ising Model

(B) Difference between MC-Algorithm and Ising Model adjacency matrices respectively.

FIGURE 5. Comparison of the outputted CDN for MC-Algorithm and Ising Model

In contrast, some Bootstrapped CI were not overlapped (e.g. I27.2-I08.1 and I34.0-I35.1) which suggests that differences in edge weights are meaningful. To investigate more the differences in edge weights, we conducted bootstrapped difference test. We set the following hypothesis null: $H_0^{i,j,k,l}$: "edge weights $e_{i,j}$ and $e_{k,l}$ are equal" for all nodes k,l,i,j, with significance threshold $\alpha = 0.05$. Figure 6a shows the results. Black squares indicate significance and gray ones indicates non significance. The blue squares in the diagonal represent the weights. For example, I34.0-I35.1 bootsrtrapped confidence interval overlaps with almost all other edges, and which makes its interpretation misleading. But for difference significance testing, it is significantly different from I27.2- I07.1 and I27.2–I08.1.

Besides investigating accuracy of weights of CDN, misoscopic analysis of components/nodes of the CDN can give insights about the importance of nodes contribution to form the structural information hidden in the connections pattern between diseases in CDN. The importance of individual nodes in the network can be assessed by investigating the node centrality. Figure 7a presents Centralities normalized as Z-score. For example, I07.1 and I36.1 have medium strength centrality and there is not significant difference between them. Nodes I27.2 and I34.0 were the most performing in closeness and betweenness centralities, this is because except I27.2 and I34.0, all nodes have one single connection. I35.0 have the smallest closeness centrality



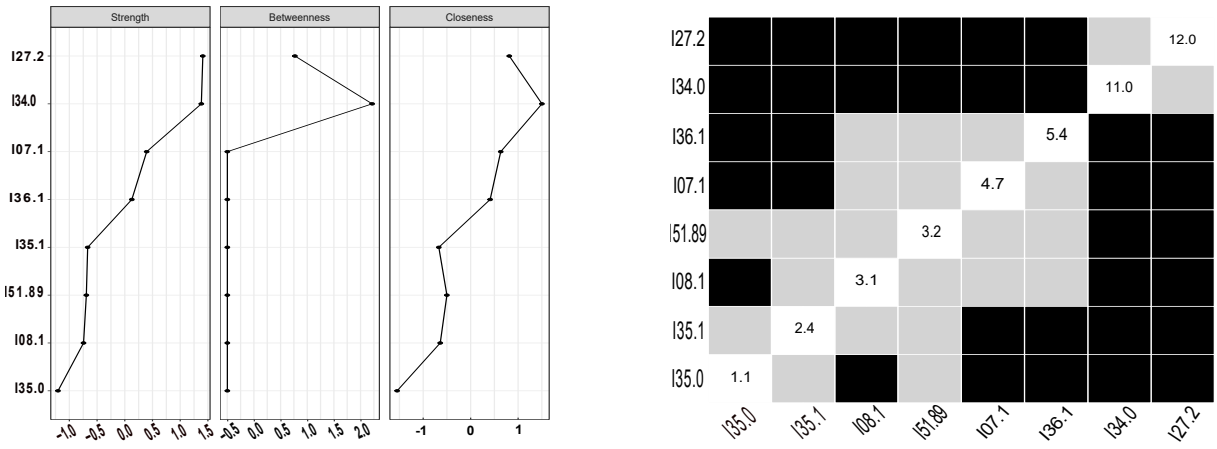
(A) Edge difference tests. Black squares indicate significance and gray ones indicates non significance. The blue squares in the diagonal represent the weights.

(B) Accuracy of the edge weight estimates (dots) by Ising Model and the 95% confidence intervals (lines) for the estimates.

FIGURE 6. Accuracy of edge weights and their differences.

because it has the weakest connection towards other nodes (MC=3.14) which makes her almost an isolated node. According to strength centrality, I27.2 and I34.0 were the most central nodes, followed by I36.1. Their centrality indicates that these nodes were those more likely to affect or to be affected directly by other nodes in the network. Nodes strength has gradually increasing scores which makes the interpretation of differences in these scores are not straightforward.

To investigate more the strength scores, we performed hypothesis testing -similar to edge weights difference test-. The null hypothesis in which we set $\alpha = 0.05$, $H_0^{i,j}$: " strength centrality scores of node i and j are equal" for all nodes with i and j. Figure 7b shows the results. While I27.2 and I34.0 strength centralities are not significantly different from each other, these two nodes are significantly different from and stronger than all other nodes in the CDN, which supports our earlier remark about their potential importance in the network. A straightforward



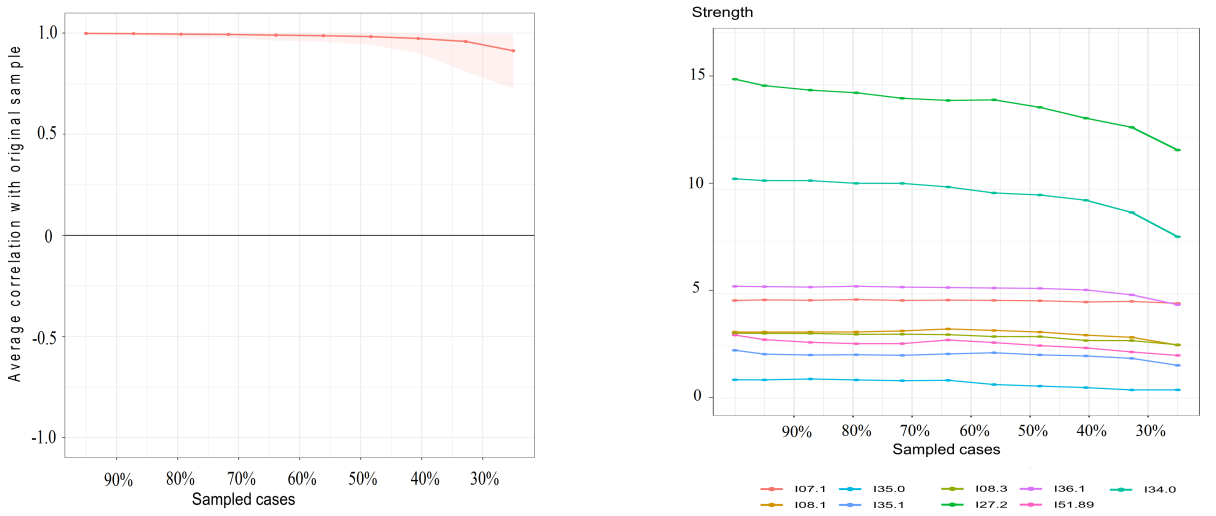
(A) Node centralities of CDN normalized in Z-score scale.

(B) Significance testing of strength nodes centralities of CDN.

FIGURE 7. Centrality measures of nodes of CDN and their difference significance of strength centrality.

conclusion is that other secondary pulmonary hypertension I27.2 and non-rheumatic mitral insufficiency I34.0 are key players in the detected CDN pattern and their activation has crucial influence in the other nodes in the network. Thus, they should be prioritized to be targeted in a therapeutical operation whenever a patient case exhibits these comorbid diseases. Targeting this highest centrality is a strategy to read off the comorbidity pattern and thus help to deconstruct these comorbid components and alleviates the multimorbidity burden put upon the patients.

While assessing accuracy of centrality measure can result in biased results, researchers in [13] suggest assessing the stability of centralities by the coefficient of stability CS. Figure 8a shows the results of the average correlation with the original sample for strength centrality of our data. Figure 8b shows the resulting plot of the strength indices for each edge in the network. The percentage of the sample included in the estimates decreases (the subset samples decrease from 95% of the original sample to 20% of the sample). We can see that the average correlation with original sample is excellent and close to 1, and start to decline by 45% of sampled cases in more than 0.8 such that bootstrapped confidence interval can reach 0.7. Although there is a drop in the correlation between the subsample estimate and the estimate from the original entire sample and even for the samples including only 50% of the individuals of the complete sample, the correlation with the centrality indices of the samples stays strong.



(A) (B)

FIGURE 8. Stability plot of the centrality estimates of the CDN. (a) Overall average correlation stability with the original sample in respect to cases dropped. The lines represent the variations of the mean correlations between the given sampled percentage and the complete (original) sample and the shades represent the area between the 2.5% and 97.5% percentile of the sampled. (b) Average correlation stability for each edge in the CDN.

To further investigate the stability of the strength estimates, we can plot how much each strength estimate fluctuates in the different samples. As can be seen in Figure 8b, the ordering of the strength estimates is sufficiently stable: Only nodes that have the same strength estimates in the complete sample change their position in the different samples. I27.2 and I34.0 are clearly stable in comparison with the remaining nodes. The order starts to be instable by sampling 20% of the original sample. For example, I36.1 and I07.1 tend to the same strength once dropping 20% of the original sample. As a result, it is thus possible to interpret the estimated strength estimates of the CDN. Overall, the pattern suggests an excellent stability of the centrality indices for strength centrality.

Ising Model present some advantages for reducing spurious edges, and the regularization technique estimates a statistical model while including a penalty for model complexity has been shown to converge to the generating network structure under the assumption that the network

is sparse [36] and simulation studies reported that the LASSO has a low likelihood of false positives [21].Thus, The LASSO yields a more parsimonious graph (fewer connections between nodes) that reflects the most important empirical relationships hidden in the data. However, the nature of the relationship represented as an edge needs to be further investigated and interpreted (the edge could represent a direct causal pathway between nodes, or it could reflect the common effect of a latent variable not included in the network model).

As a conclusion, regularization technique reduces over-fitting and removes false positive edges, to exclude spurious relationships and to make networks more parsimonious, robust, and interpretable. It is however, important to consider that the detected morbidity can be affected by the number and the nature of the selected diseases. For example a strong relationship detected between two diseases can be caused by a third disease acting as a latent variable not included in the analyzed network. Thus the replication of this study by other researchers on more diseases is important and crucial for investigating such limits.

6. Conclusions and Perspectives

In this work, we proposed pairwise Markov Random field approach to detect the comorbidity pattern of some valvular heart diseases. Using Ising Model suited for binary data; the obtained results suggest that this model can detect potential causal links among diseases and provide Comorbidity Disease Network comparable to widespread and traditional approach in multimorbidity research. We applied these methods on a case study of a real dataset. An assessment of the stability and accuracy of the obtained Network suggests that the obtained network is reliable and the degree of confidence with which edge weight and centrality rankings can be interpreted is meaningful.

The proposed method presents advantages of lowering the risk for over-fitting due to regularization technique which is controlled by a hyper parameter included in the model, and more investigation is needed for the exploration of performance of this model in Multimorbidity field. Besides, we conducted a mesoscopic analysis of the Comorbidity disease Network to better understand important nodes in the emergence of the detected pattern. The results suggest that Secondary Pulmonary hypertension and Non rheumatic Mitral (valve) insufficiency played crucial role in the skeleton of the Network. Thus, prioritizing these diseases in patients with

valvular heart Multimorbidity may alleviates the burden of Multimorbidity resulted from the effect of a strong significant co-occurrence between diseases, and hopefully protect the patients from complicated states.

Further investigation of Ising Model to other data and other diseases is required. The replication of this study will constitute accumulative evidence to best understand to what extend network analysis and machine learning techniques offers the potential for insight, into structural relations among core Multimorbidity processes and build integral research framework, as mining tool, which help physicians to accumulate empirical evidences necessary to understand the complex phenomenon of Multimorbidity.

7. APPENDICES

7.1. Abbreviations. We collect in Table 3 some abbreviations used in this paper.

TABLE 3. Abbreviations.

Abbreviation	Meaning
AMPD	Automatic Multimorbidity Pattern Detection.
CDN	Comorbidity Disease Network.
ICD 10	The International Classication of Diseases, Tenth Revision.
MC	Multimorbidity Coefficient.
EBIC	Extended Bayesian Information Criterion.
CI	Confidence Interval.
CIs	Confidence Intervals.
CS	Correlation Stability coefficient.
pMRF	Pairwise Markov Random Field.

7.2. Notations used in Methodology sections (4.1).

7.2.1. *Notations of Problem Overview section (4.1.1).* This section introduce the general mathematical formulation of Automatic Multimorbidity Pattern Detection probem. See notations in Table 4.

TABLE 4. Notations of Problem Overview section (4.1.1).

Notation	Meaning
$\|S\|$	The number of elements of a set S.
$D = \{d_1, d_2, \ldots, d_{\|D\|}\}$	The set of the studied diseases.
R	k-ary relation over Cartesian product sets D^k.
d_i	The disease number i.
$G = (D, R)$	Hypergraph such that the vertices D are the nodes and hyperedges are represented by R.
$X = \{X_1, X_2, \ldots, X_{\|X\|}\}$	The set of observed data of patients.
$X_i = (x_{1,i}, x_{2,i}, .., x_{\|X_i\|,i})$	Tuple of diagnoses of the patient i.
$x_{j,i} \in D$	The diagnosis j for the i^{th} patient.
P^*	The joint probability distribution over X.
$X = \{X_d \| d \in D\}$	The data X indexed by diseases $d \in D$.
X_d	Binary random variable of the presence of disease $d \in D$ over the patients of the dataset.
$R_\theta(X)$	Family of models estimated from the data X.
H	The hypothesis space defined by a parametrization of the relation $R_\theta(X)$.
$R^*_\theta(X) \in H$	The model that best fit the distribution P^*.
$L(X, R_\theta)$	The expected loss function which measures the loss that a model distribution R_θ makes on input observation X.

7.2.2. *Notations of Multimorbidity Coefficient-based approach section (4.1.2).* This section introduce the mathematical definition of the Relation R (which defines a Multimorbidity pattern) using Multimorbidity Coefficient score. See notations in Table 5.

7.2.3. *Notations of Conditional dependence- based approach section (4.1.3).* This section introduce the mathematical definition of the Relation R using probabilistic framework. See notations in Table 6.

TABLE 5. Notations of Multimorbidity Coefficient-based approach section (4.1.2).

Notation	Meaning
d_i	Binary random variable for the presence or (absence) of the disease number i.
$P(d_i)$	The probability of observing d_i.
$P(d_i, d_j)$	The probability of observing d_i and d_j at the same time.

TABLE 6. Notations of Conditional dependence-based approach section (4.1.3).

Notation	Meaning
$\mathscr{P}(\mathscr{D})$	The power set of diseases/binary random variables.
$P(X = x)$	The probability that the random variables X take on the particular value x.
$c \in R$	An element of hyperedges R (i.e. multimorbid diseases).
Ψ_c	Non-negative function over the variables in a hyperedges c.
Z	Normalizing constant that ensures that the distribution sums to one.
$\Psi_i(x_i)$	The node potential function for a realization of x_i.
$\Psi_{i,j}(x_i, x_j)$	The pairwise potential functions for a pair of realizations of x_i and x_j.

7.2.4. *Notations of Ising model section (4.2.1).* This section proposes an algorithmic implementation of the Relation R using Ising Model which implements the Pairewise Markov Random Field. See notations in Table 7.

7.2.5. *Notations of Multimorbidity Coefficient based Algorithm (MC-Algorithm) section (4.2.2).* This section proposes an algorithmic implementation of the Relation R using Multimorbidity Coefficient score. See notations in Table 8.

TABLE 7. Notations of Ising model section (4.2.1).

Notation	Meaning
α_i	Estimate parameter of the logistic regression for the variable i.
$\beta_{i,j}$	Estimate parameter of the logistic regression for the dependent variable i and the independent variables j.
β	The set of estimte parameters $\beta_{i,j}$.
$\omega_{i,j}$	Edge weight of the Ising structure network that links nodes i and j.
ω	The weight matrix of the Ising structure network.
ω_i	The i^{th} row (or column due to symmetry) of the Ising structure network ω.
λ	The regularization tuning parameter.
$Pen(\omega_i)$	The penalty function which is defined in terms of the LASSO.
γ	The parameter of shrinkage.

TABLE 8. Notations of Multimorbidity Coefficient based Algorithm (MC-Algorithm) section (4.2.2).

Notation	Meaning
N_{diag}	The number of diagnoses in the dataset.
N_{dis}	The number of diseases presented in the dataset.
$D = \{d_1, d_2, d_3, ..., d_{N_{dis}}\}$	The diseases in the dataset.
$M_k \subset D$ such $k > 1$	The subset of size k diseases from D.
$\mathscr{P}(\mathscr{D})$	The power set of D.
$f : I \to \{D_1, D_2, ..., D_{N_{diag}}\}$	The application that maps every patient i to his recorded diagnosis $\{x^i_1, x^i_2, ..., x^i_{N_{diag}}\}$.
$\mathscr{O}(n^2)$	The upper bound computational complexity.
$W_{expected}$	Current expected Weight of d_1 and d_2.
$W_{observed}$	Current observed Weight of d_1 and d_2.
E_{d_1,d_2}	The weighted edge in the CDN/ the graph G(V, E) that maps two nodes/diseases d_1 and d_2.

CONFLICT OF INTERESTS

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

REFERENCES

[1] 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. https://doi.org/10.1038/s41598-020-59336-1.

[2] A.L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nat. Rev. Genet. 12 (2010), 56–68. https://doi.org/10.1038/nrg2918.

[3] R.F. Barber, M. Drton, High-dimensional Ising model selection with Bayesian information criteria, Electron. J. Statist. 9 (2015), 567-607. https://doi.org/10.1214/15-ejs1012.

[4] J. Bonis, https://github.com/drbonis/CMBD_MAD_2016.

[5] L. Boschloo, C.D. van Borkulo, D. Borsboom, R.A. Schoevers, A prospective study on how symptoms in a network predict the onset of depression, Psychother Psychosom. 85 (2016), 183–184. https://doi.org/10.1159/000442001.

[6] M.J. Brandt, W.W.A. Sleegers, Evaluating belief system networks as a theory of political belief system dynamics, Pers. Soc. Psychol. Rev. 25 (2021), 159–185. https://doi.org/10.1177/1088868321993751.

[7] J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, Biometrika. 95 (2008), 759–771. https://doi.org/10.1093/biomet/asn034.

[8] M.R. Chernick, Bootstrap methods: A guide for practitioners and researchers, John Wiley and Sons, (2011).

[9] J. Dalege, D. Borsboom, F. van Harreveld, et al. The attitudinal entropy (AE) framework: Clarifications, extensions, and future directions, Psychol. Inquiry. 29 (2018), 218–228. https://doi.org/10.1080/1047840x.2018.1542235.

[10] C. Dass, A. Kanmanthareddy, Rheumatic heart disease, StatPearls Publishing, Treasure Island (FL) (2021). https://www.ncbi.nlm.nih.gov/books/NBK538286.

[11] J. Doyle, E. Murphy, S. Smith, et al. Addressing medication management for older people with multimorbidities: a multi-stakeholder approach, in: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, ACM, Barcelona Spain, 2017: pp. 78–87. https://doi.org/10.1145/3154862.3154883.

[12] S. Epskamp, J. Kruis, M. Marsman, Estimating psychopathological networks: Be careful what you wish for, PLoS ONE. 12 (2017), e0179891. https://doi.org/10.1371/journal.pone.0179891.

[13] S. Epskamp, D. Borsboom, E.I. Fried, Estimating psychological networks and their accuracy: A tutorial paper, Behav. Res. 50 (2017), 195–212. https://doi.org/10.3758/s13428-017-0862-1.

[14] A. Finnemann, D. Borsboom, S. Epskamp, H.L.J. van der Maas, The theoretical and statistical ising model: A practical guide in R, Psych. 3 (2021), 594–618. https://doi.org/10.3390/psych3040039.

[15] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010), 1–22.

[16] A.S. Helms, D.S. Bach, Heart valve disease, Primary Care: Clinics in Office Practice. 40 (2013), 91–108. https://doi.org/10.1016/j.pop.2012.11.005.

[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. https://doi.org/10.1038/s41598-019-51135-7.

[18] M.C. Johnston, M. Crilly, C. Black, et al. Defining and measuring multimorbidity: a systematic review of systematic reviews, Eur. J. Public Health. 29 (2018), 182–189. https://doi.org/10.1093/eurpub/cky098.

[19] I. Jones, F. Cocker, M. Jose, M. Charleston, A.L. Neil, Methods of analysing patterns of multimorbidity using network analysis: a scoping review, J. Public Health (Berl.) (2022). https://doi.org/10.1007/s10389-021-01685-w.

[20] 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. https://doi.org/10.1016/j.ijmedinf.2017.09.014.

[21] N. Krämer, J. Schäfer, A.-L. Boulesteix, Regularized estimation of large-scale gene association networks using graphical Gaussian models, BMC Bioinformatics. 10 (2009), 384. https://doi.org/10.1186/1471-2105-10-384.

[22] J. Kruis, G. Maris, Three representations of the Ising model, Sci. Rep. 6 (2016), 34175. https://doi.org/10.1038/srep34175.

[23] P. Legendre, M.J. Fortin, Spatial pattern and ecological analysis, Vegetatio. 80 (1989), 107–138. https://doi.org/10.1007/bf00048036.

[24] L.S. Lim, E. Lamoureux, S.M. Saw, et al. Are myopic eyes less likely to have diabetic retinopathy? Ophthalmology. 117 (2010), 524–530. https://doi.org/10.1016/j.ophtha.2009.07.044.

[25] R.C. MacCallum, D.T. Wegener, B.N. Uchino, L.R. Fabrigar, The problem of equivalent models in applications of covariance structure analysis, Psychol. Bull. 114 (1993), 185–199. https://doi.org/10.1037/0033-2909.114.1.185.

[26] M.T. Maeder, L. Weber, M. Buser, et al. Pulmonary hypertension in aortic and mitral valve disease, Front. Cardiovasc. Med. 5 (2018), 40. https://doi.org/10.3389/fcvm.2018.00040.

[27] M.T. Maeder, L. Weber, H. Rickli, Pulmonary hypertension in aortic valve stenosis, Trends Cardiovasc. Med. 32 (2022), 73–81. https://doi.org/10.1016/j.tcm.2020.12.005.

[28] K. Maganti, V.H. Rigolin, M.E. Sarano, R.O. Bonow, Valvular Heart Disease: Diagnosis and Management, Mayo Clinic Proceedings. 85 (2010), 483–500. https://doi.org/10.4065/mcp.2009.0706.

[29] 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. https://doi.org/10.1016/j.arr.2019.04.005.

[30] F. Marzouki, O. Bouattane, Structural knowledge analysis and modeling of multimorbidity using graph theory based techniques, Commun. Math. Biol. Neurosci. 2021 (2021), Article ID 91. https://doi.org/10.28919/cmbn/6839.

[31] K.P. Murphy, Machine learning: a probabilistic perspective, MIT Press, Cambridge, MA (2012).

[32] K. Nicholson, T.T. Makovski, L.E. Griffith, et al. Multimorbidity and comorbidity revisited: refining the concepts for international health research, J. Clinic. Epidemiol. 105 (2019), 142–146. https://doi.org/10.1016/j.jclinepi.2018.09.008.

[33] R. Pastorino, C. De Vito, G. Migliara, et al. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives, Eur. J. Public Health. 29 (2019), 23–27. https://doi.org/10.1093/eurpub/ckz168.

[34] J. Pearl, Causality: Models, reasoning, and inference, Cambridge University Press, New York, NY, (2000).

[35] P. Planell-Morell, M. Bajekal, S. Denaxas, et al. Trajectories of disease accumulation using electronic health records, Stud. Health Technol. Inform. 270 (2020), 469-473. https://doi.org/10.3233/shti200204.

[36] P. Ravikumar, M.J. Wainwright, J.D. Lafferty, High-dimensional Ising model selection using ℓ_1-regularized logistic regression, Ann. Statist. 38 (2010), 1287-1319. https://doi.org/10.1214/09-aos691.

[37] M. Rijken, V. Struckmann, I. van der Heide, et al. How to improve care for people with multimor-bidity in Europe? European Observatory on Health Systems and Policies, Co-penhagen (Denmark) (2017). https://www.ncbi.nlm.nih.gov/books/NBK464548.

[38] D.J. Robinaugh, R.H.A. Hoekstra, E.R. Toner, D. Borsboom, The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research, Psychol. Med. 50 (2019), 353–366. https://doi.org/10.1017/s0033291719003404.

[39] S. Rosenkranz, J.S.R. Gibbs, R. Wachter, et al. Left ventricular heart failure and pulmonary hypertension, Eur. Heart J. 37 (2015), 942–954. https://doi.org/10.1093/eurheartj/ehv512.

[40] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc.: Ser. B (Methodol.) 58 (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[41] T. Tichelbäcker, D. Dumitrescu, F. Gerhardt, et al. Pulmonary hypertension and valvular heart disease, Herz. 44 (2019), 491–501. https://doi.org/10.1007/s00059-019-4823-6.

[42] M. van den Akker, F. Buntinx, J.F.M. Metsemakers, et al. Multimorbidity in general practice: Prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases, J. Clinic. Epidemiol. 51 (1998), 367–375. https://doi.org/10.1016/s0895-4356(97)00306-5.

[43] C.D. van Borkulo, D. Borsboom, S. Epskamp, T.F. Blanken, L. Boschloo, R.A. Schoevers, L.J. Waldorp, A new method for constructing networks from binary data, Sci. Rep. 4 (2014), 5918. https://doi.org/10.1038/srep05918.

[44] Heart Foundation, Heart value disease, https://www.heartfoundation.org.nz/your-heart/heart-conditions/heart-valve-disease, accessed 2021/10/08.

[45] C.J.M. Whitty, F.M. Watt, Map clusters of diseases to tackle multimorbidity, Nature. 579 (2020), 494–496. https://doi.org/10.1038/d41586-020-00837-4.

[46] D.L. Vetrano, K. Palmer, A. Marengoni, et al. Frailty and multimorbidity: A systematic review and meta-analysis, J. Gerontol.: Ser. A. 74 (2018), 659–666. https://doi.org/10.1093/gerona/gly110.

[47] C. Violán, A. Roso-Llorach, Q. Foguet-Boreu, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis, BMC Fam. Pract. 19 (2018), 108. https://doi.org/10.1186/s12875-018-0790-x.

[48] 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. https://doi.org/10.1038/s41598-019-39326-8.

[49] N.K. Schiltz, D.F. Warner, J. Sun, et al. Identifying specific combinations of multimorbidity that contribute to health care resource utilization, Med. Care. 55 (2017), 276–284. https://doi.org/10.1097/mlr.0000000000000660.

[50] 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, J. Gerontol.: Ser. A. 75 (2020), 798–805. https://doi.org/10.1093/gerona/glz137.

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

[52] 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. https://doi.org/10.1007/978-3-642-38326-7_21.

[53] M. Lappenschaar, A. Hommersom, P.J.F. Lucas, et al. Multilevel temporal Bayesian networks can model longitudinal change in multimorbidity, J. Clinic. Epidemiol. 66 (2013), 1405–1416. https://doi.org/10.1016/j.jclinepi.2013.06.018.

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

[55] M.L.P. Bueno, A. Hommersom, P.J.F. Lucas, et al. 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. https://doi.org/10.1007/978-3-030-00461-3_7.