Data science for modeling disease interactions: a baseline algorithm

Faouzi Marzouki [1,*] and *Omar* Bouattane[2]
[1,2] Laboratory of Signaux Systèmes Distribués et Intelligence Artificielle (SSDIA), Hassan II University of Casablanca, ENSET Mohammadia, Morocco

Abstract. Multimorbidity is one of the major problems in recent health care systems, the more conditions the patients suffer from, the worst psychological pressures are put upon these patients. We formulate Multimorbidity detection as a hypergraph learning problem. Then we propose an implementation of a multimorbidity pattern detection using Multimorbidity coefficient score. This pairwise based algorithm can be considered as a baseline to which other data-driven and machine learning techniques for multimorbidity pattern detection can be evaluated. We illustrate this algorithm by building a co-occurrence model for comorbid diseases over psycho-social profiles present in a real dataset. Based on the comorbidity network of diseases, we conducted mesoscopic analysis using centrality analysis of network disease/nodes and determined potential components of the network using community detection algorithms. The patterns detected in this work by the used algorithms reveal first, that the proposed algorithm can be used as a baseline to other approaches. Second, that aging does not influence the risk of developing Multimorbidity diseases just in quantity, but also in complexity.

1 Introduction

Coexisting diseases in the same patient during a given clinical stay is referred to as Multimorbidity [1]. In recent medicine, it is a challenging health problem and initiatives are ongoing to shift medical systems from disease focused care, to a patient focused care [2]. Recently, data-driven techniques have been explored due to increasing availability of data to get insight about Multimorbidity [3].

The relative risk [5] and odds ratios [6] and Multimorbidity coefficient (MC) [7,8] are all approaches to indicate the strength of a relationshi p between diseases. While the multimorbidity coefficient can assess magnitude of association strength of any number of diseases, Odds and Risk Ratio are ineffective at distinguishing clustering from coincidental comorbidity. Majority of researches represent the co-occurrence between diseases as a weighted graph such that nodes represent diseases and edges represent association strength. This representation is suitable more for comorbidity, but do not express higher order of co-occurrence suitable for multimorbidity.

We formulate Multimorbidity detection as a hypergraph learning problem. Then we propose an implementation of a multimorbidity pattern detection using Multimorbidity Coefficient Score. This pairwise based algorithm can be considered as a baseline to which other data-driven and machine learning techniques for multimorbidity pattern detection can be evaluated. We illustrate this algorithm by building a co-occurrence model for comorbid diseases over psycho-social profiles present in a real dataset. Erickson theory [4] divides human life in 8 life stages: Infancy, toddler, early childhood, middle childhood, adolescence, young adulthood, middle adulthood, older adulthood. The theory estimated the time interval of each stage, and each in between stage reflects psychological crisis ended up by a growth in maturation process. This crisis is a hall mark of the end of a stage in life and the beginning of the next.

In the following, we set multimorbidity detection problem in a formal framework (Section 2), we present the community detection algorithms (Section 3) and the centralities (Section 4) used to analyze the built comorbid disease network. We present preliminary results in Section 5, we present conclusion in Section 6.

2 METHODOLOGY

2.1 Problem setting

Let $D = \{d_1,d_2,\ldots,d_{|D|}\}$ be a finite set of $|D|$ distinct diseases. We consider $X=\{X^{(1)},X^{(2)},..,X^{(n)}\}$ as a set of n recorded diseases of n patients where $X^{(i)}=(x^{(1,i)}, x^{(2,i)},..,x^{(|Xi|,i)})$ and $x^{(j,i)}\in D$.

Let R(X) be a k-ary relation over Cartesian product D^k. We consider the hypothesis space $H = \{R_\Theta(X)\}$ such

* Corresponding author: faouzi8marzouki@gmail.com

that Θ encode the parameters related to the used model, for example threshold in pairwise methods, degrees of polynomial regression, number of axes in Principle component analysis [9], [10]. We define in this work the multimorbidity pattern detection as the research of $R_\Theta(X) \in H$ that best fit the data. The k-ary relation R(X) can be represented by a hypergraph G=(V,H) such that V=D and H is an element of the power set of D. In an ordinary graph, the edge connects exactly two nodes (which correspond to a comorbidity disease network). A hyper graph can join arbitrary number of nodes, thus it has more expressiveness of diseases association than ordinary graphs to represent Multimorbidity Diseases Network.

If d_i is represented by a binary random variable such that the probability $p(d_i)$ is defined as the probability of occurrence of the disease $d_i \in D$. We consider $d_1, d_2, .., d_n$ as positively comorbid diseases, if we have $p(d_1, d_2, .., d_n) > p(d_1).P(d_2) \ldots . p(d_n)$ i.e. the diseases occur more likely in the same time than what would be expected by chance only [11].

If $p(d_1, d_2, .., d_n) = p(d_1).p(d_2) \ldots . p(d_n)$ we consider that these conditions are randomly co-occurencing. The final case $p(d_1, d_2, .., d_n) < p(d_1).p(d_2) \ldots . p(d_n)$ refers to a protective comorbidity [12]. The magnitude of association between diseases is measured by Multimorbidity Coefficients (MC). It is the observed rate of occurrence of diseases divided by the rate that is expected under the assumption that there is no association between disorders. The algorithm calculates MC score for every combination of diseases and draws a hyper-edge for every hyper-edge with significantly MC score more than 1. See figure 1. Finally, Multiple testing corrections (e.g. Bonferroni adjustment) can be used to adjust p-values measured using multiple tests in order correct for false positives inflation.

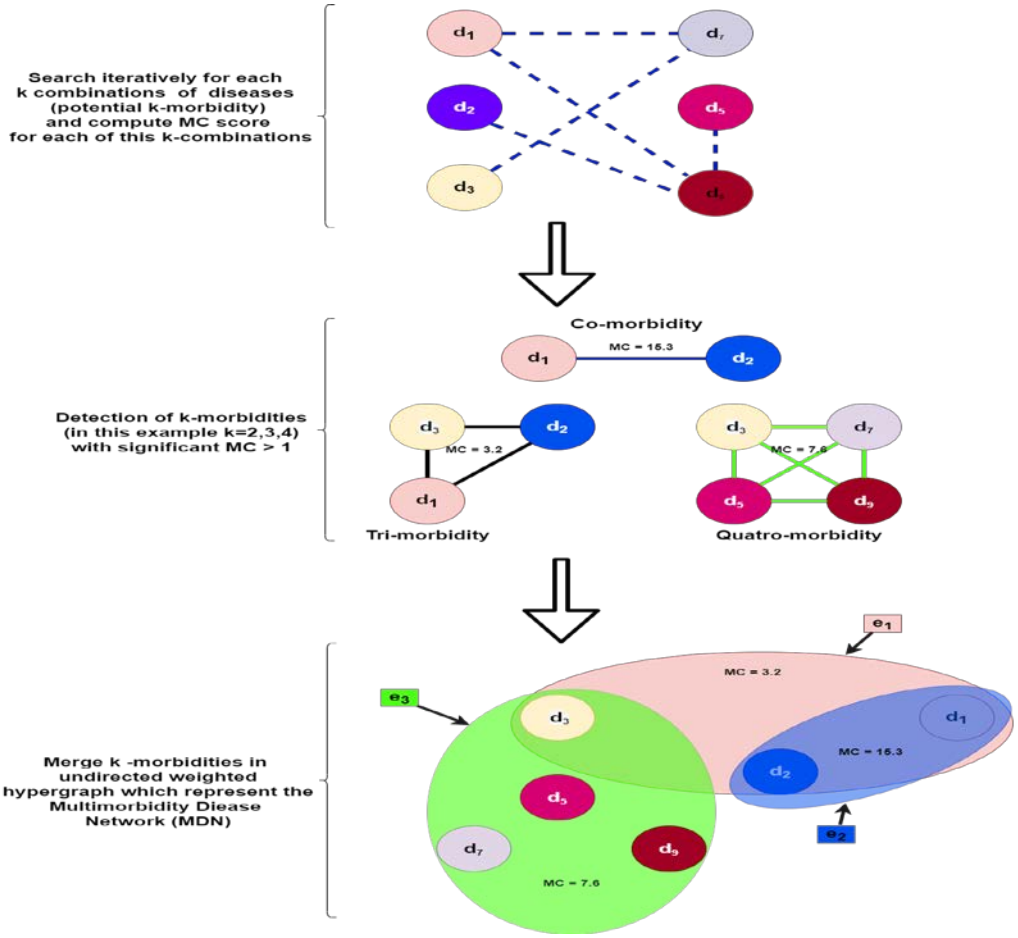

Fig. 1 Steps of the proposed MC-Algorithm: firstly, the algorithm computes MC score for all combinations of k distinct diseases. Secondly, the algorithm detects significant non-random associations of the current combination. Thirdly, it merges the Multimorbidities detected into one single hypergraph with vertices V and hyperedges E. In this illustration $V = \{d_1,d_2,d_3,d_4,d_5,d_7,d_9\}$ and $E = \{e_1,e_2,e_3\} = \{\{d_1,d_2,d_3\};\{d_1,d_2\};\{d_3,d_5,d_7,d_9\}\}$

2.2 Data

Our work was applied to a hospital inpatients diagnosis dataset, published by the national health service authorities in Madrid. The data contain 78 451 patients (34639 males, and 43812 females). The diseases are encoded in ICD10 (12763 unique ICD10 code).

2.4 Mesoscopic analysis of the Comorbidity Disease Network

We applied the special case of MC-algorithm on real medical data to extract Comorbidity Disease Network, and then we analyzed the obtained Network using centrality and community analysis. A community is formed by the subset of diseases that interact more than the other subsets in the network. Community detection can reveal important topological properties such as group of nodes which have strong interconnections from the rest of the network, and may potentially represent a latent common cause. In this work, community detection consists in revealing potential hidden multimorbid groups diseases. We compared in this work the following four well-known community detection algorithms: algorithm of Girvan and Newman [13], Label Propagation [14], Louvain Algorithm [15], Walktrap Algorithm [16]. We compared these algorithms based on modularity measure Q ($0 \leq Q \leq 1$) which is defined as the number of edges within the community divided by the expected number in a random graph with the same distribution of degrees. $Q = 0$ indicates the current group of nodes is what would be expected if edges were assigned randomly. $Q = 1$ indicates a perfect community of nodes.

3 Preliminary results and discussion

We applied MC-algorithm to extract Comorbidity Disease Graph (CDG) for males (represented by an ordinary graph). Figures 2 and 3 show the CDG for males across life stages (according to Erickson theory). A detected edge in the CDG indicates that there is significant score MC between the two diseases (we set the significance level < 0.01). The figure shows increasing amount of diseases/nodes and density of edges/associations as age increases (while in older males graph there are 447 diseases and 1468 significant association, early infancy stage has 7 diseases and 4 significant associations).

We found that females have average of communities more than males with good scores of Modularity score (around 0.78-0.90), which indicate a modular topology of CDN. The Increasing numbers of communities reveals additional layers of multimorbidity burden. For instance, the node I08.1 (Rheumatic disorders of mitral and tricuspid in the same time) co-occurs 15.07 times with the node I27.2 (other secondary pulmonary diseases), more than what would be expected by chance only.

The MC-Algorithm can be considered as a baseline to which other data driven and machine learning can be compared to. For example, in figure 5, we applied Bayesian network learning structure and Markov random field network estimation and to data of males aged more than 65 years. They reveal comparable structural patterns of CDN as MC-Algorithm. Except the I27 – I35.0 edge dependence detected by Bayesian network, the rest of the skeleton is the same outputted skeleton (the same independence map).

These three algorithms were applied to the following valvular heart related diseases for males more than 65 years: 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 mitral and tricuspid valves (I08.1). Combined rheumatic disorders of tricuspid, mitral and aortic valves (I08.3). Other pulmonary hypertension (I27.2). Other ill-defined heart diseases (I51.89).

4 Conclusion and perspectives

In this work, we represented Multimorbidity pattern Detection as a hypergraph learning problem. We proposed an implementation of a multimorbidity pattern detection using Multimorbidity Coefficient Score. We applied this algorithm to real medical data in order to build a co-occurrence model for comorbid diseases over psycho-social profiles according to Erickson theory. Based on the outputted network of multimorbidity of diseases, we analyzed the Comorbidity Disease Network using centrality and community detection algorithms. The analysis reveals that aging does not influence the risk of developing Multimorbidity diseases just in quantity, but also in complexity.

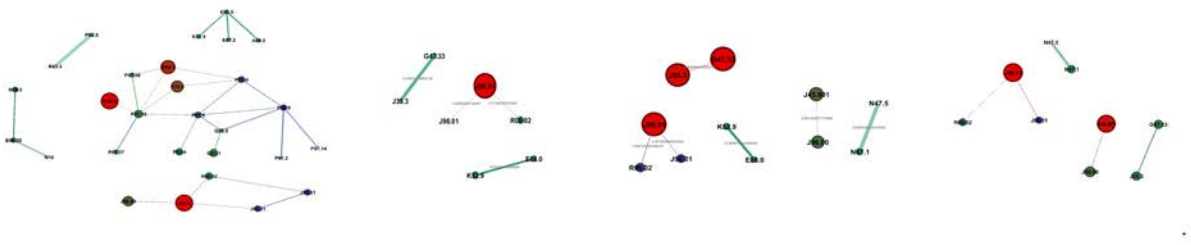

Fig. 2. From infancy stage (left) to adolescence (right).

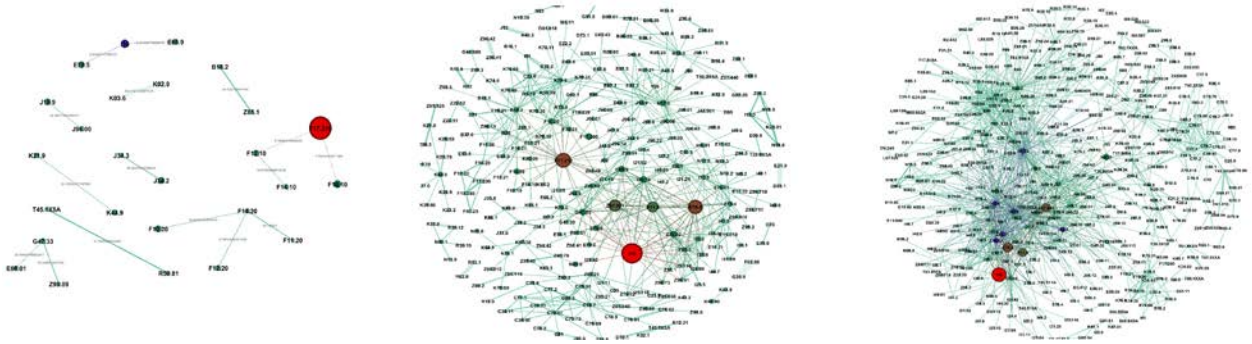

Fig. 3. From young adulthood (left) stage to older adulthood stage (right).

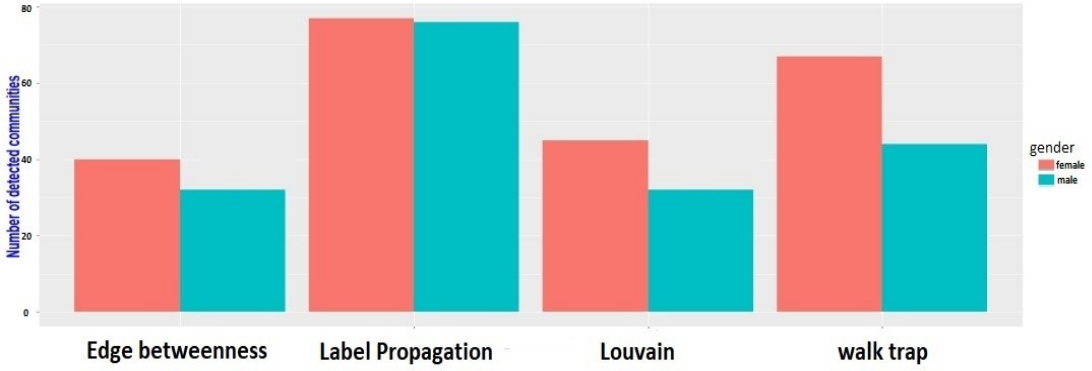

Fig. 4. Communities number detected by the algorithms. Data of application were for older adulthood (more than 65 years).

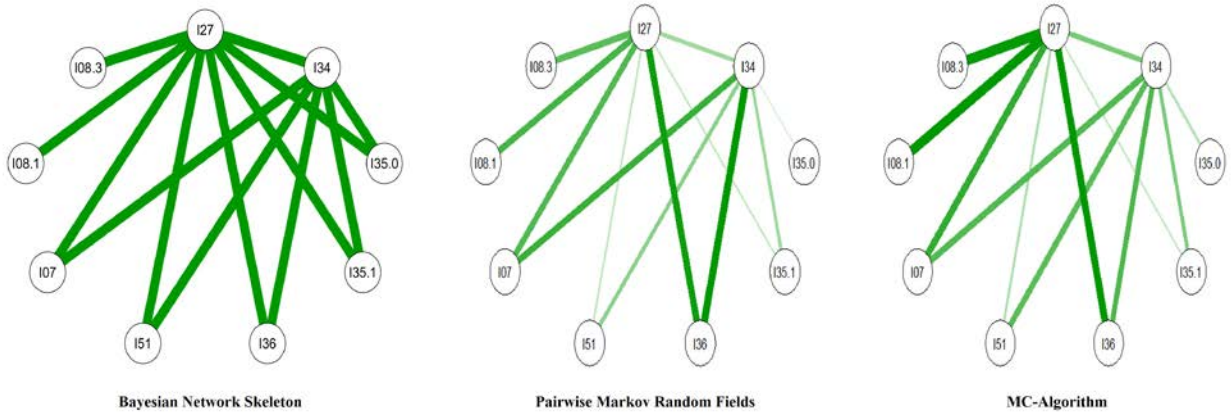

Fig. 5 Some Machine learning algorithm performance compared with MC-Algorithm. Each present edge indicates a significant detected comorbidity. An absent edge reflects non-association detection between two diseases/nodes. In the right, Bayesian network skeleton learnt from data using Hill Climbing algorithm. In the center a Markov Random field approaches using Ising Model. In the left the outputted CDN of the proposed MC-Algorithm for older adulthood (> 65 years). Note that in the Bayesian network all the edges are put in the same score which reflects in the same thickness of edge weights (we are interested in its skeleton).

References

1. A. R. Feinstein, "The pre-therapeutic classification of co-morbidity in chronic disease," J. Chronic Dis., vol. **23**, no. 7, pp. 455–468
2. M. Rijken et al., How to improve care for people with multimorbidity in Europe? European Observatory on Health Systems and Policies, [2017]. Available on: http://www.ncbi.nlm.nih.gov/books/NBK464548. Last accessed 12 aout 2021
3. R. Pastorino et al., " Benefits and challenges of Big Data in healthcare: an overview of the European initiatives ", Eur. J. Public Health, vol. **29**, no Supplement_3, p. 23-27
4. G. A. Orenstein et L. Lewis, "Eriksons Stages of Psychosocial Development ", in StatPearls, Treasure Island (FL): StatPearls Publishing, [2021].

* Corresponding author: faouzi8marzouki@gmail.com

5. R. Bonita, R. Beaglehole, T. Kjellström, et W. H. Organization, Basic epidemiology. World Health Organization, [2006].
6. H. A. Droogleever Fortuyn et al., " Severe fatigue in narcolepsy with cataplexy ", J. Sleep Res., vol. **21**, no 2, p. 163-169
7. F. S. Roque et al., "Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts ", PLoS Comput. Biol., vol.**7**, no 8, p. e1002141, [aug 2011]
8. A.-L. Barabási, N. Gulbahce, et J. Loscalzo, " Network medicine: a network-based approach to human disease ", Nat. Rev. Genet., vol. **12**, no 1, p. 56-68
9. A. Aguado, F. Moratalla-Navarro, F. López-Simarro, et V. Moreno, " MorbiNet: multi-morbidity networks in adult general population. Analysis of type 2 diabetes mellitus comorbidity , Sci. Rep., vol. **10**, no 1, p. 2416
10. C. Madlock-Brown et R. B. Reynolds, " Identifying obesity-related multimorbidity combinations in the United States ", Clin. Obes., vol. **9**, no 6, p. e12336
11. M. van den Akker, F. Buntinx, J. F. Metsemakers, S. Roos, et J. A. Knottnerus, " Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases ", J. Clin. Epidemiol., vol. **51**, no 5, p. 367-375
12. L. S. Lim, E. Lamoureux, S. M. Saw, W. T. Tay, P. Mitchell, et T. Y. Wong, " Are myopic eyes less likely to have diabetic retinopathy? ", Ophthalmology, vol. **117**, no 3, p. 524-530
13. M. Girvan et M. E. J. Newman, " Community structure in social and biological net-works ", Proc. Natl. Acad. Sci., vol. **99**, no 12, p. 7821-7826
14. U. N. Raghavan, R. Albert, et S. Kumara, " Near linear time algorithm to detect community structures in large-scale networks ", Phys. Rev. E, vol. **76**, no 3, p. 036106
15. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, et E. Lefebvre, " Fast unfolding of communities in large networks ", J. Stat. Mech. Theory Exp., vol. 2008, no **10**, p. P10008
16. P. Pons et M. Latapy, " Computing Communities in Large Networks Using Random Walks ", in Computer and Information Sciences - ISCIS 2005, Berlin, Heidelberg, [2005], p. 284-293.