Original article

Explaining provincial disease network via graph-theoretic analysis of disease occurrence in Iran

Arash Nazeri MD [a], Abdolreza Salahi-Moghadam PhD [b], Ali Reza Khoshdel MD, MPH, PhD [c,*], Mahtab Noorifard MD [c]

[a] Tehran University of Medical Sciences, Tehran, Iran
[b] Hormozgan University of Medical Sciences, Bandar-Abbas, Iran
[c] AJA University of Medical Sciences, Tehran, Iran

A B S T R A C T

Purpose: Social networks are increasingly being used to study the epidemiology of human diseases. These methods are usually limited to studying a single or a small group of diseases within a small community and may not fully capture effects of environmental factors that drift the epidemiology of diseases at the community level (rather than individual level). By introducing an ecological variant of the social network, we described provincial disease network (PDN) to study the similarities in regional occurrence of diseases.
Methods: In this network, nodes (provinces) are connected via edges together. Provinces that have similar pattern of disease prevalence and/or incidence tie stronger together. We sought to find modular organization of Iran's PDN and to identify factors (literacy, rural population percentage, smoking, geographical distance, and age distribution) that could predict the strength of interprovincial connections in the PDN.
Results: Provinces in Iran's PDN were segregated into five different modules. Geographic distance, differences in the literacy percentage and percentage of population in the 0–16 years age group showed significant inverse correlation with strength of connections in Iran's PDN.
Conclusions: Network-based approaches could provide important insights by identifying modular architecture within the PDN and underlying factors that drive similarity between disease patterns among regional entities.

© 2016 Elsevier Inc. All rights reserved.

Introduction

In the recent years, network analysis is increasingly being used to study epidemiology of diseases. Social networks have provided important insights to the dynamics of infectious disease transmission and spread of noninfectious diseases. These studies have shown that a variety of health-related phenomena, ranging from epidemics [1,2] to sexually transmitted diseases, smoking, obesity, drinking alcohol, and suicide, may spread along and within social networks [3–7].

All these studies have taken each individual as units of the complex networks and the interaction among these individuals as the links between them. However, there are many other important

factors (e.g., cultural factors, economic factors, and major health policies) that could affect burden of diseases at the societal levels. These factors may not be fully captured by individually mapping disease-based social networks. Furthermore, given that only a limited number of illnesses could co-occur in an individual, it is not possible to study multiple diseases simultaneously. One alternative to these approaches is to take societies (e.g., populations of cities, provinces, or even countries) as the units to build disease-based networks.

In this study, we provide a conceptual approach, by defining provincial disease network (PDN), which is defined as a set of nodes (provinces) and edges that link them together. In this network, provinces that have similar pattern of disease prevalence and/or incidence tie stronger together. Next, we try to elucidate community organization of Iran's PDN, which will allow us to categorize provinces based on the similarity between their disease patterns. Finally, we provide a framework by which we could find what factors determine network organization of the PDN.

Methods

Data sources

Disease prevalence and/or incidence data of the disease were used as a measure of occurrence of the disease in each province. The data set was extracted from resources of Iranian Ministry of Health and Medical Education. The final data set used in this study, included prevalence and/or incidence of 5 psychiatric illnesses (depression, anxiety disorders, psychotic disorders, addiction, and suicide), incidence of 13 types of malignancies (colorectal cancer, gastric cancer, esophageal cancer, brain tumors, lymphoma, leukemia and bladder cancer in both sexes, prostate cancer, lung cancer, and laryngeal cancer in men, breast, uterine and ovarian cancers in women), incidence of 23 infectious diseases (gonorrhea, syphilis, chlamydia, brucellosis, *Plasmodium vivax* malaria, *Plasmodium falciparum* malaria, hepatitis B, hepatitis C, typhoid fever, bordetellosis, tuberculosis, cutaneous leishmaniasis, echinococcosis, cutaneous anthrax, tetanus, diphtheria, measles, cholera, leprosy, polio, dysentery, meningococcal meningitis, and pneumococcal meningitis), and prevalence of three noncommunicable chronic diseases (obesity, diabetes mellitus, and hypertension), across 30 Iranian provinces (total number of disease entities = 45 composed of 5 main psychiatric, 23 infectious, 13 neoplastic, and 3 major noncommunicable diseases in the provinces). All the data used in this study correspond to years 2004–2007. We also considered available information of some potential risk factors for each province including the geographic longitude and latitude of the province center, demographic and lifestyle factors (alcohol, smoking, obesity), and education level for a diversity matrix.

The main concepts of graph theory and network measures

In summary, a graph (network) contains a set of nodes that are interconnected with edges. The edges may be binary (0 or 1) or weighted (e.g., ranging from 0–1), signed (negative values for edges are allowed) or unsigned (only positive values are allowed), and directed (with directionality) or undirected (without directionality). The degree of a node is defined as the number of direct connections with other elements of a network (i.e., neighbors). Nodes with high degrees in a network are called hubs. A module is defined as a set of densely connected nodes with relatively few numbers of connections with nodes of other modules. Strength of degrees can be calculated by a variety of methods (e.g., Pearson correlation and Spearman rank). The strength of connection between each pair of nodes within a network can be represented by adjacency matrix.

Graph theory has been used in several branches of biologic studies such as gene expression networks, metabolic networks, and neural networks. After the naming convention used for many of these networks (e.g., connectome, metabolome, and diseasome) the presented network in this study, which is based on epidemiologic data from major territorial units (i.e., provinces) could be named "Geomatome" as a new concept in Health Geomatics.

Provincial disease network

PDN constitutes of a set of provinces (nodes) that are connected by a set of edges to one another. In this network, strength of an edge is indexed by a "similarity measure" in "pattern of disease occurrence" between each pair of provinces. In this study, as the metric of association between provinces we used pairwise correlation coefficient (Spearman's rho) with each disease prevalence and/or incidence serving as an observation (Fig. 1). To make disease frequencies more comparable among illnesses, Z-score
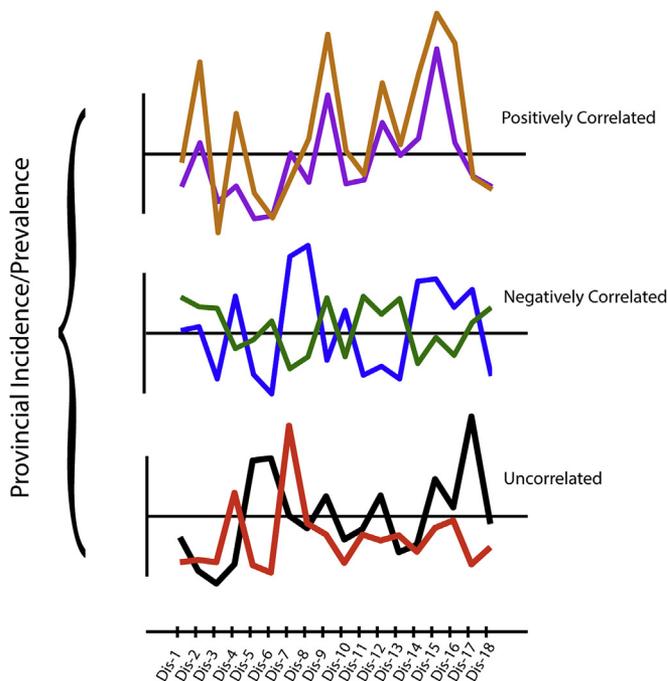

Fig. 1. A schematic diagram depicting three possible categories of pairwise interrelationships between a pair of provinces: (i) positively correlated, in which relative prevalence/incidence of the diseases move in the same direction; (ii) negatively correlated, in which relative prevalence/incidence of the diseases move in the opposite direction; and (iii) uncorrelated, in which relative prevalence/incidence of the disease are not associated between the two provinces.

transformation was applied to the raw prevalence and/or incidence values of each disease across the country.

For each pair of provinces i and j, we computed the Spearman's rank correlation coefficient, ρ_{ij} (i ≠ j). In other words, for each disease, the 30 provinces were ordered with regard to the prevalence and/or incidence of the disease. It was this schedule of 45 ranks (in disease occurrence) for the province i that was then correlated with the schedule of 45 ranks (in disease occurrence) with province j to form the connection between the two provinces. This resulted in an adjacency matrix of 30 × 30 Spearman's rank correlation coefficients representing the associations in Z-transformed provincial prevalence and/or incidence of diseases between each of the 435 unique pairs of provinces [i.e., (30 × 29)/2]; resulting in a signed weighted network was drawn. To compute node degree (i.e., the number of connections that link each node to the rest of the network), nonsignificant edges ($P_{ij} > .05$) were removed. Next, for each province, the numbers of negative and positive connections were summed separately to calculate negative and positive node degree, respectively. For visualization purposes, edges corresponding to nonsignificant pairwise correlation ($P_{ij} > .05$) were discarded.

Modularity and community structure of the PDN

As discussed previously, one of the applications of the graph theory to PDN is to group provinces into communities (modules) that have similar pattern of disease prevalence and/or incidence. In the graph theory, a module is defined as a set of nodes in a network that are more densely connected to one another than to the nodes from other modules [8]. In networks with no negative weights, modularity (Q) measures the fraction of the edges in the network that connect nodes within communities minus the expected value of the same quantity in a network with random connections but