A Framework for Social Network-Based Dynamic Modeling and Prediction of Communicable Diseases

Samy Ghoniemy and Noha Gamal

***Abstract*—It was published lately in 2016 that there are approximately 3.7 million of deaths caused by communicable diseases annually. Unfortunately, currently there is no automated method for the detection and tracking of communicable diseases progression. In this paper, a framework is proposed, that is based on social network analysis, different biological sensors, and big data analytics as for predicting and analyzing communicable disease and to facilitate the process of managing, preventing and predicting risks of communicable disease progression. The proposed framework is largely based on graph theory and social network analysis algorithms to model and dynamically predict communicable disease risk for diagnosed and non-diagnosed patients. In this research, a global graph structure that maps a whole friendship network is proposed, and the suitable algorithms to identify and continuously monitor a certain communicable disease progression rate. This research can potentially be useful for forming a methodology for early intervention and prevention policies targeted at patients that can potentially divert them from the disease pathway. The interpretation and dynamic utilities offered by the framework and its predictive capability are considered a remarkable and promising broad model highlighting potential pathways linking social support, biological sensors and data sciences to physical health.**

***Index Terms*—Social network analysis, graph theory, communicable disease progression, healthcare, big data analytics.**

I. Introduction

Communicable diseases are the leading reasons of death worldwide. In spite of the popular belief that communicable diseases affect mostly high-income countries and older people, but the reality always gives quite a different picture. As it was lately published in [1], 80% of deaths from such diseases occur in low- and moderate-income countries, and these are mostly concentrated among the poor.

Almost half of the deaths occur at an age under 70 years. The premature mortality and reduced quality of life for patients not only results in an epidemiological burden, but also exerts a significant impact on national economies [2]. Therefore, communicable disease prevention and management has been a major concern for governments and related international organizations.

Although communicable diseases can be categorized in different ways, WHO uses three guiding principles for prioritization: i) diseases with a large-scale impact on mortality, morbidity and disability, such as (HIV), (TB), ii) diseases that can potentially cause epidemics, such as influenza and cholera, and iii) diseases that can be effectively controlled with available cost-effective interventions, such as diarrheal diseases.

This research focuses on using social network and big data analytics for predicting and analyzing to facilitate the process of managing, preventing and predicting risks of communicable disease progression [3]-[5]. Though, traditional methods of clinical diagnosis and regular monitoring of a large population are often resource-intensive in terms of available clinical necessities and economic capabilities [6].

One potential alternative comes from a data and information engineering perspectives on healthcare information systems, more specifically hospital admission data, which carries rich semantic information about the patients' overall health status and diagnosis information in the form of standardized codes [7].

This vast amount of systematically and social networks generated data can help us to understand the disease footprints left by infected patients and can then be utilized to evaluate and predict the health status of another population [8], [9]. At present, very limited work has been done in realizing these particular potentials of this healthcare data [10]-[12].

This research, therefore, asks the questions: how various types of healthcare data can generate knowledge that can help us to tracking the communicable disease pathways? What is the best techniques to organize and filter heterogenous healthcare data for most effective knowledge discovery? Is it possible to reasonably predict communicable disease risks by intelligently leveraging tracking of disease progression? What is the effect of utilizing graph theory and social networks analysis algorithms to correlate collected patients' health paths with already diagnosed patients? To what extent do behavioral and demographic risk factors contribute in the progression of communicable diseases? Finally, how can a generic and adaptive predictive framework to assess communicable disease be construct based on various health care data?

II. Research Focus and Proposed Framework

This research attempts to answer these questions by presenting a proposed framework. Which has two major goals: i) to analysis, track and dynamically model the progression rate of a particular communicable disease, and ii) to model and dynamically predict communicable disease risk for non-diagnosed patients. The proposed framework is largely based on graph theory and social network analysis algorithms [13], [14] for both parts.

Manuscript received March 20, 2018; revised August 23, 2018.

Samy Ghoniemy and Noha Gamal are with the British University, Egypt (e-mail: samy.ghoniemy@Bue.edu.eg, noha.gamal@acu.edu.eg).

DOI: 10.7763/IJMO.2019.V9.679

For the first part, the concept of a "Communicable Disease Modeling CDM network" is proposed. CDM can effectively model the disease comorbidities and their transition patterns, thereby representing the disease progression. While generating the modelling network, it is proposed that not only look at the pattern of disease in the diagnosed patients, but also compare them with that of non-diagnosed patients to identify which comorbidities are more responsible for leading to the communicable disease pathway.

An individual communicable disease patient's health trajectory is not necessarily representing all comorbidities of that communicable disease. However, on a large population level, if individual patients' trajectories are merged by adding up edges between same disease pairs, an aggregated version of the health trajectory network can then be obtained. So, the CDM network should represent the overall health trajectory of patients having a particular communicable disease taking into consideration all available parameters and conditions.

For the second phase, the "modeling network" will be enhanced to be able to predict communicable disease risks in non-diagnosed patient networks. That is by building a predictive model that compares the CDM network with the medical history of a patient who is not yet diagnosed and calculates this patient's risk of developing the communicable disease in near future. For matching the networks, several graph theory and social network-based methods are utilized.

These methods look at multiple parameters, including the frequency of the comorbidities, clustering membership, and geographical, demographic and behavioral factors such as age, gender and smoking history. Individual risk scores against each of these parameters are then merged to generate single prediction score that represents the risk of communicable disease infection for non-diagnosed patients. Fig. 1, shows the functional diagram and the data flow from the inputs to resulting outputs of the proposed framework.

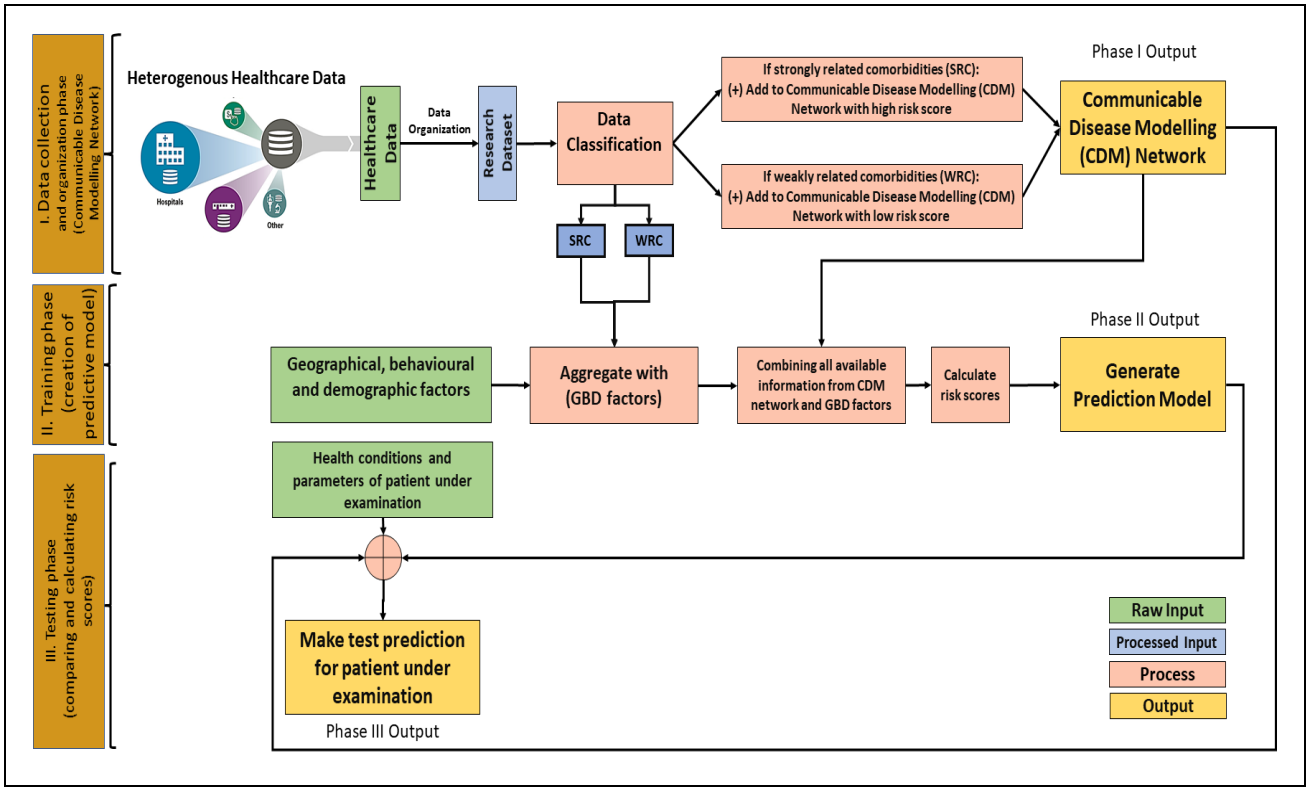

Fig. 1. Functional diagram for the proposed framework.

III. Functional Components of the Proposed Framework

The functional components of the framework presented in this section, are derived from the abovementioned concepts and how they fit together in our research. The input to the framework is presented by various types of healthcare data It is assumed that for each patient this should at least contain disease codes representing the health condition of the patient at a point of time (e.g., during hospital admission), along with vital demographic information like age, gender and smoking status. The output of the framework takes three forms: First, it creates a communicable disease modelling (CDM) network, which represents a combined health path for all communicable disease patients. Second, the framework is designed to combine healthcare classified data with geographical, behavioral and demographic risk factors using graph theories and SNA algorithms, that to calculate the risk scores of different parameters, such that; the framework will be able to generate prediction model given actual risk based on the CDM network. Third, the framework should provide an interface for predicting the probability of communicable disease development for a new test patient who is as yet non-diagnosed. The interpretation and utility of the modelling network and its predictive capability will depend on the dataset, context and interests of the stakeholders.

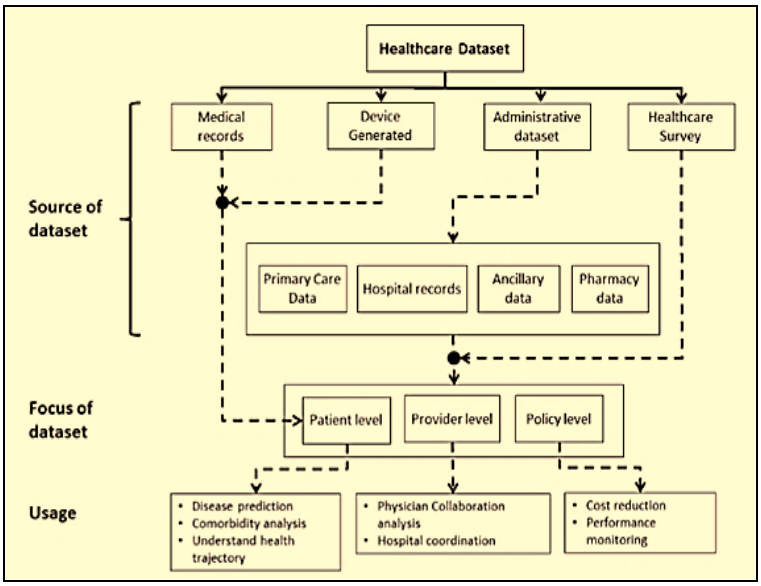

Fig. 2. Types of healthcare data and their prospective use by different stakeholders [22], [23].

In general, the predictive model compares the CDM network with the medical history of a patient who is not yet diagnosed with the particular communicable disease and calculates this patient's risk of developing the disease in near future. Along with these scores, the framework should also consider demographic and behavioral risk factors of the patients undertest, such as age, gender or smoking history. Since, there are multiple risk scores for each individual patient, it is proposed to develop a suitable linear prediction model from these scores to determine the overall risk of communicable disease.

In parallel with achieving the research goals mentioned earlier, the proposed framework aims to manipulate and alter data collected from healthcare centers in order to align them with our research methodology. Further to networked data, our research also considers several parameters based on demographic and behavior risk factors.

The goal of the parallel study, is to understand and quantify different parameters affect in disease progression and also, it may be discussed that how some parameters make better predictors than others. Our research finally aims to build a flexible and generic framework, that can be applied to various communicable diseases data and contexts.

The proposed framework in the research can potentially be useful for forming an early intervention and prevention policies targeted at those patients that can potentially divert them from the disease pathway and reduce healthcare costs from both provider and consumer perspective.

IV. Data Collection

Affordable and quality healthcare system is one of the most important development paths in any country [15], [16]. Broad goals of a healthcare system are to improve citizens' health using responsive, financially fair, and efficient ways. Modern healthcare systems evolved and revolutionized since the evolving of computing and networking technologies [17].

Large amounts of data are generated by the modern healthcare systems. Healthcare data are broadly categorized into four groups based on the way they are collected; survey data, device-generated data, medical records, and administrative data. Following, each group of healthcare data is explained in brief [18], [19].

Fig. 2 shows the four types of healthcare data and their prospective use by different stakeholders. In our research, all mentioned types of healthcare data will be used to provide the required knowledge discovery [20]-[23].

A. Survey Data

Many countries maintain active registry of databases through health conditions related surveys. For example, the statistical survey conducts in Egypt every 15 years that collect information about each family member, such as, age, chronic disease condition, gender, professional, etc. all these pieces of information are analyzed to build a database for the population that can be so beneficial in detecting healthcare conditions.

B. Device-Generated Data

Data generated by electronic devices in an automated way. For example, data generated from modern medical devices located in healthcare centers that can generate periodic scan of patient vital signs. Such data can also be generated by personal electronic devices such as, smart phones, smart watches, etc. such devices can generate and share real-time health related information.

C. Medical Records

A medical record is generated for each patient individually in healthcare centers including medical conditions, diagnosis, and treatment. Each patient has a unique number identifying his medical information (Medical Record Number, MRN) in order to link patient's information in different consultations.

The advancement of information technology helped the development of electronic healthcare concept (e-health). As a consequence, medical records are broadly unified, centrally stored and worldwide shared. Consistent accessibility to up-to-date patient's medical records is provided. So, the healthcare service providers and emergency departments can have access to patient's complete medical history.

D. Administrative Data

Administrative data is collected from billing, auditing, quality assurance and other administrative purposes. The data contain service-level information such as fees, charges and summary of provided services.

V. Conclusion and Future Work

The integration of graph theory and SNA-based methodologies in the field of disease progression and prediction is the primary contribution of our research. This network-based approach has not been widely used in this field, especially using different types of data mentioned earlier in section 4.

Consequently, this research is likely to contribute to the scientific community by showing the potential use of correlated data in analyzing communicable disease progressions and predicting risk through the proposed framework.

A three phases frame work has been presented, that is designed to: create a communicable disease modelling (CDM) network, combine healthcare data with geographical, behavioral and demographic risk factors using graph theory and SNA algorithms, that to calculate the risk scores of different parameters, and finally it can provide an interface for predicting the probability of communicable disease development for a new test patient who is as yet non-diagnosed.

A full model of this framework is being implemented and the resulting outcome will be compared with similar systems in order to measure its efficiency and all the results will be published in our next publication.

References

[1] World health organization. [Online]. Available: http://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_part2.pdf

[2] Cancer is a leading cause of death worldwide, accounting for an estimated 9.6 million deaths in 2018. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/

[3] J. G. Anderson, "Evaluation in health informatics: Social network analysis," *Computers in Biology and Medicine*, vol. 32, no. 3. pp. 179-193, 2016.

[4] S. Uddin, A. Khan, and M. Piraveenan, "Administrative claim data to learn about effective healthcare collaboration and coordination through social network," in *Proc. Hawaii international Conference on System Sciences (HICSS-48)*, Hawaii, United States, 2015.

[5] T. A. Snijders, "Testing for change in a digraph at two time points," *Social Networks*, vol. 12, no, 4. pp. 359-373,158.

[6] T. A. Snijders, G. G. V. D. Bunt, and C. E. Steglich, "Introduction to stochastic actorbased models for network dynamics," *Social Networks*, vol. 32, no. 1, pp. 44-60.

[7] A. Khan, S. Uddin, and U. Srinivasan, "Understanding chronic disease comorbidities from baseline networks: Knowledge discovery utilising administrative healthcare data," in *Proc. the Australasian Computer Science Week Multiconference*, 2017.

[8] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *Proc. the International Conference on Computer Systems and Applications*, 2008.

[9] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," in *Proc. 1994 ACM Conference on Computer supported Cooperative Work*, 1994, ACM, pp. 175-186.

[10] S. Uddin, A. Khan, and L. A. Baur, "A framework to explore the knowledge structure of multidisciplinary research fields," *PLOS One*, vol. 10, no. 4.

[11] M. Baglioni, S. Pieroni, F. Geraci, F. Mariani, S. Molinaro, M. Pellegrini, and E. Lastres, "A new framework for distilling higher quality information from health data via social network analysis," in Proc. *2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW)*, 2013, pp. 48-55.

[12] J. J. Baker, "Medicare payment system for hospital inpatients: diagnosis-related groups," *Journal of health care finance*, vol. 28, no. 3, pp. 1-13.

[13] B. K. Armstrong, J. A. Gillespie, S. R. Leeder, G. L. Rubin, and L. M. Russell, "Challenges in health and health care for Australia," *Medical Journal of Australia*, vol. 187, no. 9, p. 485.

[14] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proc. the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998, pp. 43-52.

[15] J. Goecks A. Nekrutenko, and J. Taylor, "Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biol*, vol. 11, no. 8, 2010.

[16] R. S. Ledley and L. B. Lusted, "Reasoning foundations of medical diagnosis symbolic logic, probability, and value theory aid our understanding of how physicians reason," *Science*, vol. 130, pp. 9-21.

[17] H. Lee, Z. Tu, M. Deng, F. Sun, and T. Chen, "Diffusion kernel-based logistic regression models for protein function prediction," *Omics: A Journal of Integrative Biology*, vol. 10, no. 1, pp. 40-55.

[18] K. E. Nelson and C. Williams, *Infectious Disease Epidemiology*, Jones & Bartlett Publishers, 2013.

[19] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2.

[20] F. R. Pitts, "A graph theoretic approach to historical geography," *The Professional Geographer*, vol. 17, no. 5, pp. 15-20.

[21] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J. C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali, "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data," *Medical Care*.

[22] S. G. Rizzo, D. Montesi, A. Fabbri, and G. Marchesini, "ICD code retrieval: Novel approach for assisted disease classification," in *Proc. International Conference on Data Integration in the Life Sciences*, 2015.

[23] S. Uddin, "Social networks enabled coordination performance model for patient hospitalization," *Faculty of Engineering and IT*, The University of Sydney, Sydney, 2011.

Samy Ghoniemy was born on March 27, 1967, in Giza, Egypt. He received his bachelor, and master of engineering degrees in electrical engineering from Military Technical College (MTC), in 1990 and 1996 respectively. He also attended a three-terms condensed program studying optoelectronic courses at MTC in 1991. From 1991 to 2000 he was a junior lecturer in the Department of Optoelectronic, MTC, Cairo. In Oct. 2000, he was one of the recipients of a governmental scholarship award from Egypt, Ministry of defense to do his PhD in Carleton University, Canada. He is presently working toward the PhD degree in electrical engineering at the Department of Systems and Computer Engineering, Carleton University, Ottawa,Canada. His primary research interest is radio(mm-wave) over fiber-optic communication systems, with emphasis on laser transmitter, fiber optics, and optical modems in base stations, as well as the study of design and production of such systems. Mr. Ghoniemy is a student member of IEEE and its communications and laser and electro-optics societies. He is a member of SPIE "The international society for optical engineering". He is a member of OSA " Optical Society of America".

Noha Gamal was born on December 29, 1979, in Alex, Egypt. She received her bachelor degree in electrical engineering from Faculty of Engineering, Alexandria University in 2001. She received her master degree in communication and information technology from Nile University, Cairo, Egypt in 2015. She also attended a one year condensed program studying Information Technology at ITI in 2002. From 2002 to 2014 she worked as a communication engineer till she was information and communication manager in many local and international companies. Since 2014 till now she is working as a lecturer at Ahram Canadian University, Giza, Egypt. Currently, she is working on her earning her PhD from Faculty of Information and Computer Science, Ain Shams University, Cairo, Egypt.