IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received January 18, 2018, accepted February 16, 2018, date of publication February 28, 2018, date of current version March 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2810702

Representing Social Network Patient Data as Evidence-Based Knowledge to Support Decision Making in Disease Progression for Comorbidities

MADAN KRISHNAMURTHY[1], PAWEL MARCINEK[2], KHALID MAHMOOD MALIK[1], (Member, IEEE), AND MUHAMMAD AFZAL[3], (Member, IEEE)

[1]Department of Computer Science and Engineering, Oakland University, Rochester, MI 48309, USA
[2]Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309, USA
[3]Department of Software, Sejong University, Seoul 143-747, South Korea

Corresponding author: Madan Krishnamurthy (mkrishna@Oakland.edu)

ABSTRACT Social network patient data for comorbid studies is a sparsely explored avenue. This can provide unprecedented insight into disease conditions and their progression, hence facilitating improvement of healthcare and public health services. Structuring scattered social network data and mapping with standard disease ontologies to build reference-able knowledge base can be used in evidence-based decision support systems. In this paper, we attempt to address this direction of application where patient and time relationships are established between conditions to predict disease progression trends for comorbidities. Our prediction analysis is based on statistical modeling of the constructed knowledge base. It can be utilized towards driving personalized healthcare by applying life streams to patient journals that can provide a timed pattern of disease progression which can be used relatively and statistically for decision making and educational insight. We present and validate our approach using case-study of Brain Aneurysm with performance in terms of sensitivity and time-probability measures.

INDEX TERMS Clinical decision support systems, clustering, comorbidity, knowledge base, knowledge representation, ontology, social network, statistical analysis, and time progression.

I. INTRODUCTION

According to IBM, 90% of the data in the world was created within the last two years [23], [35]. These statistics clearly hint at the internet explosion with over 1.2 billion websites that continue to grow by the second [18]. The enormous ‘‘data exhaust’’ needs conversion into machine readable formats to be usable and provide coherent perspectives. Progression of web data from a ‘‘form’’ based approach to a smart-interactive approach has made the internet a forum for social media. Social media has been instrumental in changing the way we look at medical data. Websites such as *PatientsLikeMe* (PLM) [33] encourage people to have an open forum of conversations to compare similar health scenarios and be aware of options and concerns that can be pre-mediated. Effectiveness of social media in decision making towards disease diagnosis and treatment has been recently assessed in the domains of healthcare and public health. While social network data alone cannot be considered an accurate and reliable predictor for disease outcome studies, it can render useful insights by contributions from proven-research data, and provider notes. Study [20] supports this by demonstrating the concept of integrating social network data and mobile sensor data of patients to aid in clinical decision making. Likewise, healthcare providers may benefit from individual patient experiences that are relevant to the care they deliver [44]. Social network data can play a crucial role in areas such as public health where decisions are based on multiple influential data sources such as clinical data, environmental data, community-data and individual case-data which clinicians depend upon at the point of care [16]. Research works such as [12] have presented a high degree of agreement between social network patient data and data obtained from corresponding healthcare insurance claims to match disease diagnosis and treatment. Nakamura *et al.* [8] suggest a step ahead and state the need of both providers' and users' perspectives to improve the efficiency of treating complex diseases such as *Amyotrophic Lateral Sclerosis (ALS).* These studies are advancing towards maturing recognized social network patient data, also coined as patient-powered research data in [12] as important evidence in *Clinical Decision Support Systems* (CDSS) laterally with medical data.

Healthcare providers often engage in laborious surveys and prolonged data gathering from patients and clinical notes to recognize a specific behavior based on diagnostics characteristics [9], [15], [21]. While disease treatments continue to be more advanced and effective, cost and time associated with identifying disease progression and manifestations to other co-occurring conditions has been challenging [37], [46]. For example, authors in [40] mention that over 70 % of people with asthma have concomitant rhinitis; about 15 - 40% of rhinitis patients have clinically verified asthma. It is stated in [32] that knowledge about time progressions in comorbidities can aid physicians to take preventive actions to tackle future repercussions in patients. Studies like these can yield vital information that may be overlooked. Relationships between conditions help in proactively projecting trends of disease transformations to equip healthcare providers with vital information for treatment suggestions. Also, identifying valid relationships between progressions versus co-incidental conditions is challenging. Clustering related conditions or applying exclusion rules becomes important to avoid false-positive predictions. Incorporating social network data with computational knowledge concepts such as ontology and knowledge base can aid in suggesting futuristic disease progression trends.

Human Disease Ontology (DOID) furnishes a reliable, reusable and viable data source, mapping human disease descriptions to medical terms [6]. This integrates disease and medical vocabularies by semantic cross-mapping of terms to various standard medical vocabularies such as MeSH, ICD, SNOMEDCT, OMIM etc. [7]. Similarly, symptom and treatment ontologies such as SYMP, CSSO, CPT, APATREATMENT, etc. [7] can be considered optional extensions. CDSS rely on knowledge-based studies that empower medical professionals to improve clinical workflow scenarios and determine efficient diagnostic solutions [28], [31]. Integrating knowledge-backed solutions in CDSS equips healthcare providers with a highly granular approach in tackling complex, co-mingled and multi-dimensional healthcare problems. It can include practice-based and scientific literature-based evidences that are comprised of standardized, real-time, transparent and global data. Evidence-based medicine enabled by CDSS provides a wider acumen to enhance quality and efficiency of healthcare by optimizing the decision making process [17], [24].

In this paper, we build an extended knowledge base upon the existing disease ontologies by leveraging on the principles of knowledge and ontology engineering. The two knowledge base components – terminological knowledge, also known as TBox - refers to timeless taxonomies mapped to disease ontology and; ABox or assertion knowledge is highly specific, contingent and dependent upon a single set of circumstances subject to change [14], and referred to condition, patient and time instances in our case. We leverage on the triplet principles of *Resource Description Framework* (RDF) [39] to establish the subject-object-predicate relationship between conditions and patient. The main advantage of the knowledge base approach as we perceive, is that information extraction can be highly specific and parallelized. Mathematical evaluation on the available real-time data generates factual analysis that can be irrationally accepted. Additionally, including social network data as evidence alongside literature and practice-based data can provide an added dimension to pro-actively infer trends, treatments and preventive services based on patient records. Overall, we present a novel CDSS approach that uses a unique knowledge base built using social network data to assist in evidence-based decision making for comorbid studies. To the best of our knowledge, there is no concrete published literature to perform a baseline comparison with our current work. This further augments the significance and need for the intended research.

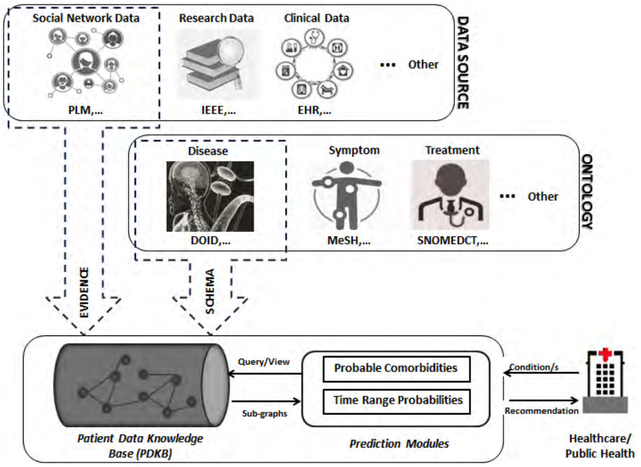


FIGURE 1. Proposed System Architecture.

We consider the wide-spread and easily accessible social network patient data as evidence in CDSS. Our work aims at predicting progression of a condition into other co-occurring conditions based on social network patient data adding time as a second variant. This is a prototype to represent the logic in decision-making and disease prediction that our method can render. The proposed system can be conceptually visualized as an enriched CDSS that recommends prediction related to comorbid conditions using statistical modeling on the patient knowledge base. Fig. 1 represents the system architecture of the conceptualized CDSS that may include multiple variables of evidence and ontology. Our study uses social network data as the evidence and DOID as the disease ontology. It is imperative to carefully integrate social network based evidence with clinical data (such as Electronic Health Record/EHR) and/or literature based evidences for a more holistic approach. However, our intent is to only represent the logic behind disease prediction and validation of appropriate evidence is beyond the scope of this paper.

Our system mainly comprises two components:

A. PATIENT DATA KNOWLEDGE BASE (PDKB)

Here, we extract patient information from PLM that includes a specific disease condition, its manifestation and progression to other conditions with time frame details all providing a specific patient record. Each of these conditions are mapped to standard DOID classes and further, two relationship types (patient and time) are applied to represent it as a customized knowledge base.

B. PROBABLE COMORBID PREDICTION AND TIME RANGE PROBABILITY PREDICTION MODULES

In the prediction modules, directed sub graphs [39] with statistical analysis are used to render useful information from the knowledge-based mappings to suggest disease progression trends to other co-occurring conditions and time predictions.

A sample case study for Brain Aneurysm based comorbid studies is provided to substantiate our methodology with evaluation and inferred results covered in the subsequent sections.

Overall, the paper is organized as follows. Section II covers related work pertaining to our research. Section III introduces the methodology comprised of PDKB build, defining object properties and instance associations along with prediction algorithms that form the basis of our evaluation considering sample data. Section IV addresses real data collection, data pre-processing, knowledge base representation and application of the prediction modules to define the objectives of the proposed methodology with inferred results. Section V presents an overview of our work with areas of enhancements laid out as Conclusion & Future work. Finally, Section VI lists out the references used in our current work.

II. RELATED WORK

There have been several works on evidence-based CDSS, semantic knowledge base management, onto-clustering based data analysis and medical comorbid studies that pertain to our focus of research [24], [26], [34]. In our proposed work we present a novel approach of constructing unique disease knowledge base that when used in conjunction with supported prediction algorithms integrated with CDSS can suggest disease progression trends based on input query.

Sim *et al.* [17] facilitate the use of CDSS using practice based and literature-based evidences recommended to enhance the quality of healthcare by developing cost-effective, yet maintainable decision-support systems. We suggest social network patient data as additional contributing evidence in CDSS to potentially avoid cognitive biasness that takes into consideration the aspects of clinical practice, research data and end-user aka patient input. Though, CDSS specific to comorbidity studies are uncommon, there have been a few works that recommend the need for such systems. Literature [2] provides a system architecture comprising of physical, data and application layer of a CDSS for comorbidity monitoring. The application layer discusses the methodology based on clinical practice guidelines (CPG) for multiple conditions co-existing in a patient. Similarly, Jafarpour and Abidi [3] present an ontology framework to merge digital CPGs related to comorbidities in CDSS to suggest therapy for patients exhibiting co-occurring conditions of Atrial Fibrillation and Chronic Heart Failure. They implement practitioner suggested merge representation ontology to define associated tasks in comorbid CPGs by creating a property called hasTask to merge CPGs with similar tasks. Though these analyses rely on neutralizing and merging CPGs across comorbid conditions, their work doesn't entail identifying relationships between conditions based on patient information which is a differentiator considered in our work.

Chen *et al.*[19] present an automated method to discover semantic relationships between two concepts in their knowledge base. The relationship defined between the two concepts constitutes a ranked approach where concepts are ordered in super-class and sub-class format with an "IS A" verb relationship. Similarly, in [29] Deshpande *et al.* relate their concepts using "fuzzy relationships" that still are verb representations. Also, literature [27] provides an ontology-based disease information system wherein classes (disease and symptom) are related using object property – has_symptom/is_a_symptom_of that are verbs. However, in our proposed work, we consider disease conditions as concepts/classes and patient-specific information which is a noun as the semantic relationship/object property. This is unique in the sense that no earlier attempts have been made to entail noun-based relationship mapping between instances. This approach is expected to provide a wider insight for knowledge representation and introduction to unforeseen trends.

Comorbid studies can aid healthcare providers in identifying the right disease inclinations and propose appropriate assistance. These studies mostly refer a survey-based approach or a clinical data analysis approach. Research work in [21] develops a study for improving clinical outcomes for patients with coexisting Diabetes Mellitus and End-stage Renal Disease. This involves observational studies including patient data and participant survey (patient, family and healthcare providers) and clinical data assessments. Study in [9] attempts to present the prevalence of comorbidity in patients identified with coronary heart diseases involving survey and lab tests over a sample of volunteers those were coherently mapped. Similarly, [15] pertains to the comorbid studies to relate diagnosis effectiveness for patients with *Autism Spectrum Disorders* (ASD) which involves detailed survey on 1366 children reported with ASD. These three works rely on exclusive participants' survey information along with available clinical research data to propose individual study objectives. This may often prove laborious, time consuming and expensive. Hence, in our current work, we leverage on the existing social media data that can provide similar survey information or more due to the presence of anonymity that aids patients in disclosing unbiased and true information without any geographic boundary. A similar attempt to study comorbidity in mental and behavioral health of patients using PLM data was performed in our earlier work [25] that emphasized the same utility of social network data analysis. We extend the same in our current research.

Ji *et al.* [45] present prediction of future comorbid disease progression by applying their collaborative prediction algorithm on patient social network data (PLM) and later inferring the validity of results hence obtained, by comparing with published medical literature. On the contrary, we consider a variable approach where instead of running our prediction algorithms against real dataset that could prove time/resource intensive for larger volumes of data; we construct a knowledge base out of the available social data to serve as a standard data reference that allows easy extraction of specific information by the use of sub-graphs. We deploy a test dataset to extensively evaluate our system performance instead of a direct factual comparison with medical literature. Also, we introduce the "time" aspect which according to us is vital in proactively generating clinical suggestions on when the disease progresses to another influenced by the time dimension.

We use hierarchical clustering analysis to provide the structuring required to classify and group nodes of input vectors using standard DOID ontology as the reference. By doing so, we provide an unrestricted framework for any disease type evaluation. Peleg *et al.*[26] follow an individualized approach of building their own ontology using medical literature and patient data specific to developmental disorders in children. They later combine this ontology with clustering methods (including domain expert inputs) to come up with golden standards pertaining to comorbid developmental disorders. This follows a more customized approach when compared to our open format and requires domain expert involvement for analysis. Also, this involves clustering of the entire ontology as opposed to the granular graph-clustering method we implement

III. METHODOLOGY

A. PATIENT DATA KNOWLEDGE BASE

Social network data of patients with comorbidity are captured from PLM, mapped to the standard disease ontology (DOID) [1] and represented as "knowledge" by leveraging on the open source ontology framework tool from Stanford Protégé [36]. In DOID, diseases are semantically classified based on medical vocabularies to comprehensively cross reference disease terminologies. The custom knowledge base, hence obtained is further subject to analysis using prediction algorithms to deduce vital disease progression patterns.

Fig. 2 represents the construction of *Patient Data Knowledge Base* (PDKB) where we explain an illustration of the knowledge base using sample patient data

The build of PDKB consists of five stages:

1) PATIENT DATA

Patients exhibiting multiple medical conditions are identified and data is extracted in a specified format with associated Patient ID, comorbid conditions and the diagnosed date for each medical condition. In the example representation patient PID1 is identified to have three medical conditions with diagnosis date – Cardiomyopathy (12/10/87); Otulipenia (07/07/88) and Angiodyspalsia (07/27/88). In the actual evaluation patient data is collected from PLM.

2) DISEASE ONTOLOGY

In the sample disease ontology created, we represent hierarchical alignment of diseases underneath three main classes, each disease expressed as a sub-class. For example, Otosclerosis is a sub-class under "nervous_system_disease". In the actual evaluation we use DOID to define the hierarchical classes.

3) OBJECT PROPERTIES

In this stage, the patient information obtained from Patient Data stage is denoted as an object property which indicates all the possible relationships between Conditions. PID# refers to "patient relationship" and PID#_TIME refers to "time relationship" in the object properties. In the example considered, patient with ID1, PID1 is the object property with three property associations – PID1 is the property association between conditions Cardiomyopathy, Otulipenia and Angiodyspalsia; PID1_20 property refers to progression from condition Otulipenia to Angiodyspalsia in 20 days and PID1_210 refers to progression from Cardiomyopathy to Otulipenia in 210 days for PID1.

4) INDIVIDUAL/INSTANCE CREATION

Here each medical condition pertains to a single instance. Each instance is mapped to the disease ontology classes using the "type" relationship. In the example considered, Cardiomyopathy is a type of Cardiovascular_system_disease as per the sample disease ontology considered. If there exists no mapping for individuals in the ontology, then the Instances exist independently as a "disease" class.

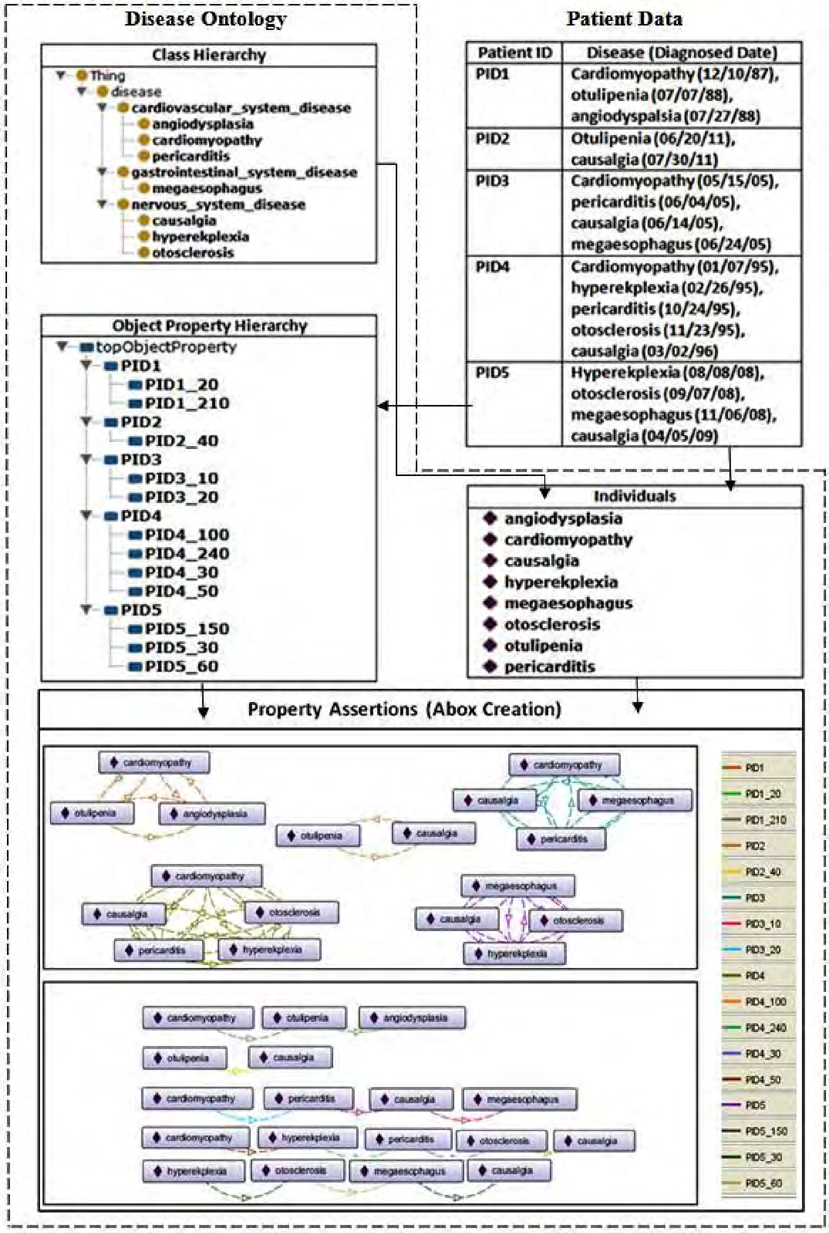
5) PROPERTY ASSERTION

The fifth stage involves linking properties to instances in order to create ABox component of PDKB.

a: PATIENT RELATIONSHIP

PID1 has conditions Cardiomyopathy, as well as Otulipenia; hence they can be represented by a bidirectional relationship, PID1 known as the "patient relationship". Since PID1 is known to exhibit three different conditions, there exist six different connections or "triplets" pertaining to condition → relationship → condition. This is shown in Fig. 3

The total number of triplets that can be derived from patient relationship can be obtained using the mesh-topology



Patient ID	Disease (Diagnosed Date)
PID1	Cardiomyopathy (12/10/87), otulipenia (07/07/88), angiodyspalsia (07/27/88)
PID2	Otulipenia (06/20/11), causalgia (07/30/11)
PID3	Cardiomyopathy (05/15/05), pericarditis (06/04/05), causalgia (06/14/05), megaesophagus (06/24/05)
PID4	Cardiomyopathy (01/07/95), hyperekplexia (02/26/95), pericarditis (10/24/95), otosclerosis (11/23/95), causalgia (03/02/96)
PID5	Hyperekplexia (08/08/08), otosclerosis (09/07/08), megaesophagus (11/06/08), causalgia (04/05/09)

FIGURE 2. Process of building PDKB.

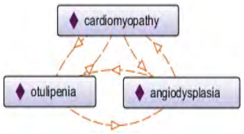


FIGURE 3. Example of Patient Relationship.

connection formula represented in Equation 1.

$$Cn(Cn - 1) \tag{1}$$

Where, Cn = Total number of conditions in a patient

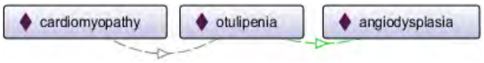


FIGURE 4. Example of Time Relationship.

b: TIME RELATIONSHIP

In the same lines, "time relationship" can be represented as the time taken to progress from one condition to another within a patient. In patient PID1, Cardiomyopathy to Otulipenia needed a time progression of 210 days and Otulipenia to Angiodyspalsia was 20 days. The time relationship is unidirectional and is represented by two triplets as per Fig. 4.

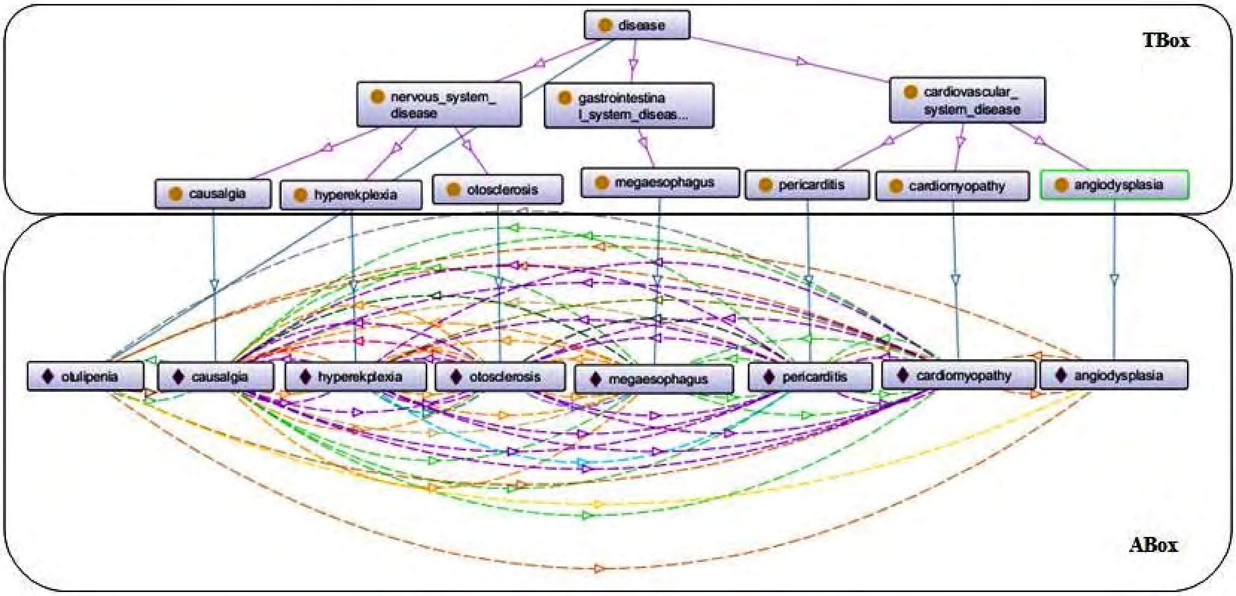


FIGURE 5. Illustration of sample PDKB.

TABLE 1. Number of triplets in sample knowledge base.

Patient ID	Number of conditions	Patient Relation–Triplets (from Eq.1)	Time Relation–Triplets (from Eq.2)
PID1	3	6	2
PID2	2	2	1
PID3	4	12	3
PID4	5	20	4
PID5	4	12	3
Total		52	13

The total number of triplets in time relation is as obtained by Equation 2.

$$Cn - 1 \tag{2}$$

Where, Cn = Total number of conditions in a patient

The values obtained for both patient and time relationships are captured in Table 1.

Hence, from the above description we have arrived at a total of 65 triplets for the sample data comprising of 5 patients and 8 conditions/instances. This is captured in the ABox section our custom knowledge base which is demonstrated as per Fig. 5.

The five-stage PDKB build illustrated for sample data, is followed to generate the actual custom knowledge base for patient data obtained from PLM. The TBox section of the knowledge base will comprise of DOID, both of which will be considered in our evaluation.

B. PREDICTION MODULES

Once the PDKB is built, it can be used to deduce important relationships based on graph theory principles. Sub-graphs can be derived from PDKB using visualization tools such as Protégé-OntoGraf for efficient knowledge representation. In our current work, we use Protégé-OWL version 4.3 with OntoGraf 1.0.1, a versatile plugin to interactively and automatically organize relationships as sub-graphs [11]. In each of the sub-graphs owing to the specific input criteria, "nodes" represent conditions and "edges" denote patient and/or time relations. The unidirectional relationships between nodes also known as "out-degree" and "in-degree" form a crucial part in influencing predictions. From the two modules we try to establish three different predictions based on input condition, visualized as a sub-graph projecting patient and time correlations. Each of the modules is discussed below:

1) PROBABLE COMORBIDITIES

In this section, we predict the probability of occurrence of a new condition in a patient with an already existing condition, hence exhibiting comorbidity. This is achieved by considering the patient-relationships between conditions in the Sub-graph for a given input condition. The percentage of occurrence of a new condition in a patient is obtained by the number of edges between new condition and the given input condition divided by the total number of edges depicted in the sub-graph.

In the sample sub-graph denoted in Fig. 6, input condition "Cardiomyopathy" is linked to all other nodes/conditions by edges representing patient relationship. There are a total of 8 nodes with 9 edges as per this example. The percentage of probability for a patient with Cardiomyopathy to acquire Causalgia is predicted to be 22.2% as per our algorithm illustrated in Algorithm 1. Table 2, represents the percentage of occurrence of a new condition in a patient with given input condition Cardiomyopathy from our sample knowledge base.

Further, we can extend the algorithm to predict DOID class comorbidities based on the conditions using Hierarchical

Algorithm 1 Chance of Occurrence (%) of Each Comorbid Condition

```
CI[i] // Input Conditions
For (i = 0, i = No. of Input Conditions CI[i], i++) {
    a = COUNT (No. of edges of CI[i])
    CC[j] //Comorbid Conditions of CI[i]
        For (j = 0, j = No. of Comorbid Conditions CC[j],j++) {
    b = COUNT (No. of edges between CI[i] and CC[j])
    P [i, j] = (b/a)*100 //Chance of Occurrence (%)
        }
}
```

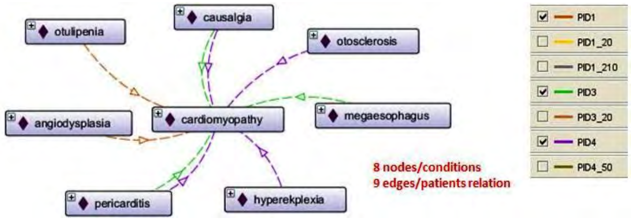


FIGURE 6. Sub-graph of input condition and its comorbidities.

TABLE 2. Comorbidities and their probable occurrence in %

Input Condition: Cardiomyopathy		
Comorbid Condition	Number of Patients	Chance of Occurrence (%)
Hyperekplexia	1	1/9*100=11.1
Pericarditis	2	2/9*100=22.2
Otosclerosis	1	1/9*100=11.1
Otulipenia	1	1/9*100=11.1
Angiodysplasia	1	1/9*100=11.1
Causalgia	2	2/9*100=22.2
Megaesophagus	1	1/9*100=11.1

Clustering Analysis (HCA) method. Hierarchical clustering is a data analysis method of grouping like entities into tiers of hierarchical relationships. HCA rely on the linking criteria between different elements to extract meaningful interpretations [38]. It can either follow a divisive approach where the representation starts top-down using cluster splitting or an agglomerative approach where the representation is bottom-up using cluster merging [10]. We use the agglomerative approach to merge data within a single cluster until every target element is fused under an umbrella class. In real sense, we map disease conditions to their parent classes based on ontology representations to infer hierarchical relationships between comorbid conditions for a given input condition. In our approach, the sub-graphs obtained for probable comorbid condition prediction is subject to HCA by relating disease conditions to their respective parent classes using the referenced disease ontology. This provides us with a clustered tree structure of the sub-graph.

For the input condition "Cardiomyopathy" considered in our example, we obtain the sub-graph as in Fig. 7. Each of the comorbid conditions associated with Cardiomyopathy has a relation to its parent class in our sample ontology, except for "Otulipenia" that is considered an independent linkage. It may be noticed that this is built from our sample ontology, where there are a total of 11 classes (1 super class + 3 parent class + 7 child-classes) with a maximum hierarchical depth of 2 levels. From sub-graph in Fig. 7 we can obtain the parent classes for each of the comorbid conditions using Algorithm 2. The classes are tabulated in Table 3.

Algorithm 2 Chance of Occurrence (%) of Each Comorbid Condition

```
CI[i] // Input Conditions
For (i = 0, i = No. of Input Conditions CI[i], i++) {
    CC[j] //Comorbid Conditions of CI[i]
    For (j = 0, j = No. of Comorbid Conditions CC[j], j++) {
    //Bioportal APIs
    BioportalSearchResult
[j] = get (http://data.bioontology. org+ "/search?q = " + CC[j] + "&ontology = "+ DOID));
        //Parse through collection to get Parent Class
        ParentClassofCC[j] = BioportalSearchResult [j].getAncestors ();
        }
    }
```

TABLE 3. Comorbidities and their ancestor classes.

Comorbid Conditions	Parent Class
Angiodyspalsia	Cardiovascular System Disease
Pericarditis	Cardiovascular System Disease
Causalgia	Nervous System Disease
Hyperekplexia	Nervous System Disease
Otosclerosis	Nervous System Disease
Megaesophagus	Gastrointestinal System Disease
Otulipenia	

TABLE 4. Cluster representation of comorbid instances.

Sample Ontology Classes			Instances	
Super Class	Parent Class	Child Class	No. of Comorbidities	Distribution in %
Disease	Cardiovascular System Disease	N/A	2	2/7*100=28.57
	Nervous System Disease	N/A	3	3/7*100=42.85
	Gastrointestinal System Disease	N/A	1	1/7*100=14.28
	N/A	N/A	1	1/7*100=14.28

When the suggested HCA is applied to the data above, we obtain the clustered tree with class associations represented in Table 4. Using this we may deduce that a patient diagnosed with Cardiomyopathy, has a higher probability of acquiring a nervous system disease (42.85%).

In our evaluation, we apply our HCA method on the subgraphs obtained for real patient data from PLM, mapping comorbid conditions to the ancestor classes defined in DOID ontology. Currently, there are about 12489 classes; maximum

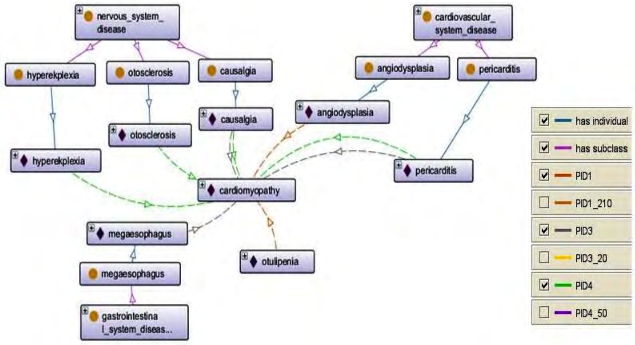


FIGURE 7. Sub-graph of input condition and its comorbidities with corresponding ontology ancestor classes.

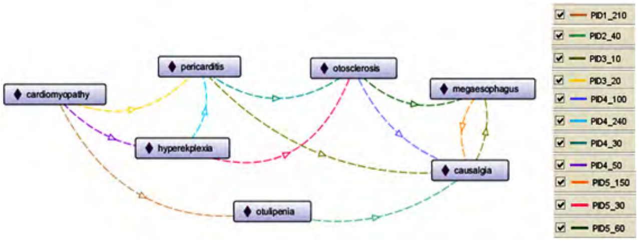


FIGURE 8. Sub-graph of time progression from Cardiomyopathy to Causalgia.

hierarchical depth of 12; average number of children in a single class is 4 as captured in the statistical ontology metrics for BioPortal DOID ontology [5], [6]. These metrics are the key considerations for our clustering analysis. We use patient condition as the search criteria in the BioPortal API (RESTful web service) to lookup for parent classes in DOID ontology [4]. The result obtained from the API call is in *Java Object Notation* (JSON) format that is further parsed to extract only the top three parent classes for each associated comorbid condition.

2) TIME RANGE PROBABILITIES

The second module aims at evaluating the time taken to progress from one identified condition to another predicted condition in a given patient. This uses unidirectional adjacency matrix representation of the sub-graph to depict path traversal from one condition to another. The sub-graphs obtained from our previous analysis are simple unidirectional graph segments with no self-loop. This allows us to represent the sub-graphs as symmetric square matrices with rows (i) and columns (j) defined by graph nodes having a 0 or 1 value based on whether the entries (conditions) are adjacent or not. In our initial observation we consider that two nodes/conditions can have only one path defined between them, hence the adjacency matrix obtained will have all 0s on the diagonal [42]. Also, if multiple patients have the same path traversal between conditions, average time between the progressions are considered. Fig. 8 represents the sub-graph obtained for disease progression from source node

$$
\begin{array}{l|ccccccc|}
 & \text{Cardiomyopathy} & \text{Hyperekplexia} & \text{Pericarditis} & \text{Otosclerosis} & \text{Otulipenia} & \text{Megaesophagus} & \text{Causalgia} \\
\text{Cardiomyopathy} & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\
\text{Hyperekplexia} & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
\text{Pericarditis} & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
\text{Otosclerosis} & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
\text{Otulipenia} & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\text{Megaesophagus} & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\text{Causalgia} & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
$$

FIGURE 9. Adjacency Matrix representation of sub-graph of time progression from Cardiomyopathy to Causalgia.

$$
A = \begin{pmatrix}
0 & 1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
$$

$$
\underset{\text{For } A^2,\ a_{1,7}=2}{\begin{pmatrix}
0 & 0 & 1 & 2 & 0 & 0 & 2 \\
0 & 0 & 0 & 1 & 0 & 1 & 2 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}}
\underset{\text{For } A^3,\ a_{1,7}=3}{\begin{pmatrix}
0 & 0 & 0 & 1 & 0 & 2 & 3 \\
0 & 0 & 0 & 0 & 0 & 1 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}}
\underset{\text{For } A^4,\ a_{1,7}=3}{\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 1 & 3 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}}
\underset{\text{For } A^5,\ a_{1,7}=1}{\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}}
\underset{\text{For } A^6,\ a_{1,7}=0}{\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}}
$$

FIGURE 10. Powers of matrix A to determine number of paths.

Cardiomyopathy to target node Causalgia in our example case scenario along with its adjacency matrix (7x7 diagonal matrix) notation as shown in Fig. 9.

The total number of paths between two conditions can be calculated using the power of matrices [30]. If A is the adjacency matrix obtained for a given sub-graph with elements of the matrix denoted by a_{ij} for source node s and target node t, the number of unique edges/paths with hops k can be represented as the i,j element of A^k. For example, the number of distinct paths between s and t nodes with 2 hops is A^2; similarly the number of paths with three hops is A^3. The total number of paths between source node and destination node can be evaluated by Equation 3.

$$
P = \sum_{k=1}^{n-1} A^k \tag{3}
$$

Where,

$P =$ total number of paths
$k =$ number of edges
$n =$ number of nodes
$A =$ Adjacency matrix

For the sample case scenario considered, the total number of paths traversed during progression from condition "Cardiomyopathy" to "Causalgia" can be derived as 9. The analysis is inferred using the following matrix evaluation in Fig. 10.

The total number of available paths for disease progression from source condition to target condition is determined using the *FindPath* function from Mathematica. Wolfram Mathematica provides technical computing tools using over 5000 well-defined functions, using powerful algorithms catering a myriad of areas such as mathematics, data science, image processing, neural networks, visualization etc. [41] The *FindPath* function from Mathematica used to

TABLE 5. Paths between source and target nodes, and time of progression for each path.

Path	Paths with Time Taken	Total Time
1	Cardiomyopathy $\xrightarrow{210}$ Otulipenia $\xrightarrow{40}$ Causalgia	250
2	Cardiomyopathy $\xrightarrow{20}$ Pericarditis $\xrightarrow{10}$ Causalgia	30
3	Cardiomyopathy $\xrightarrow{20}$ Pericarditis $\xrightarrow{30}$ Otosclerosis $\xrightarrow{100}$ Causalgia	150
4	Cardiomyopathy $\xrightarrow{20}$ Pericarditis $\xrightarrow{30}$ Otosclerosis $\xrightarrow{60}$ Megaesophagus $\xrightarrow{150}$ Causalgia	260
5	Cardiomyopathy $\xrightarrow{50}$ Hyperekplexia $\xrightarrow{240}$ Pericarditis $\xrightarrow{10}$ Causalgia	300
6	Cardiomyopathy $\xrightarrow{50}$ Hyperekplexia $\xrightarrow{240}$ Pericarditis $\xrightarrow{30}$ Otosclerosis $\xrightarrow{100}$ Causalgia	420
7	Cardiomyopathy $\xrightarrow{50}$ Hyperekplexia $\xrightarrow{240}$ Pericarditis $\xrightarrow{30}$ Otosclerosis $\xrightarrow{60}$ Megaesophagus $\xrightarrow{150}$ Causalgia	530
8	Cardiomyopathy $\xrightarrow{50}$ Hyperekplexia $\xrightarrow{30}$ Otosclerosis $\xrightarrow{100}$ Causalgia	180
9	Cardiomyopathy $\xrightarrow{50}$ Hyperekplexia $\xrightarrow{30}$ Otosclerosis $\xrightarrow{60}$ Megaesophagus $\xrightarrow{150}$ Causalgia	290

determine n number of paths is represented as below:

$$FindPath\ [g, s, t, Kspec, n]$$

$$\begin{gathered}\text{Where, } g = \text{graph}\\ s = \text{Source node}\\ t = \text{target node}\\ Kspec = \text{length of path which is number of edges it}\\ \text{contains}\\ n = \text{number of paths}\end{gathered}$$

In our example, each of the nodes of the sub graph in Figure 8 is converted to numbered vertices for assessment: 1− > *Cardiomyopathy*; 2− > *Hyperekplexia*; 3− > *Pericarditis*; 4− > *Otosclerosis*; 5− > *Otulipenia*; 6− > *Megaesophagus and* 7− > *Causalgia*. *FindPath* returns a list of all paths from s to t. For our sample data:

$$FindPath\ [g, 1, 7, Infinity, All]$$

$$\begin{gathered}\text{Where, } g = \text{Graph }[\{1 \to 2, 1 \to 3, 1 \to 5, 2 \to 3, 2 \to 4,\\ 3 \to 4, 3 \to 7, 4 \to 6, 4 \to 7, 5 \to 7, 6 \to 7\}]\\ s = 1(\text{Cardiomyopathy})\\ t = 7(\text{Causalgia})\\ Kspec = \text{Infinity}\\ \text{n} = \text{All}\end{gathered}$$

Hence, *FindPath* returns a list of all nine paths from source condition 1 to target condition 7 as below: *Output* = {{1,5,7}, {1,3,7}, {1,3,4,7}, {1,2,4,7}, {1,2,3,7}, {1,3,4,6,7}, {1,2,4,6,7}, {1,2,3,4,6,7}}. Once the paths are determined, we now overlay the time factor for each of the paths.

Table 5 represents each of the paths traversed from our example source condition Cardiomyopathy to target Causalgia with time associations between the different nodes/conditions. The total time per path is the summation of individual times between two nodes. The next section of the prediction mechanism entails calculation of the probable progression time from source to target conditions.

Time range probability calculations can either be continuous or discrete based upon the input datasets. For large available datasets, time variations can be represented using *Gaussian Normal Distribution* and standardized Z scores can be applied to obtain the *Probability Density Function* [22]. *Z-scores* are expressed as the ratio of mean score to standard deviation obtained by the normal distribution, hence defining exact location of a value within the distribution. If a disease progression depicts a large number of total paths, the bell curve for normal distribution inherits a symmetric representation. Alternatively, for smaller discrete datasets, time probability calculations assume asymmetric notation and can be evaluated using five-point summary based statistical analysis [13]. This principle is used to evaluate what suits best for the available data paths to obtain optimum time probability.

IV. EVALUATION

A. DATASET COLLECTION AND PRE-PROCESSING

Social network data obtained from PLM for patients identified with Brain Aneurysm is the source dataset for our analysis as a sample case study. As of April 2017 (during time of data extraction), there were 472,470 registered patients on PLM. Of these patients, 271 patients were registered with Brain Aneurysm, furnishing vital information about their individual conditions, treatment and diagnosis. A total of 47.97% of Brain Aneurysm patients indicated positive diagnosis of other co-existing conditions, providing a number at 130. About 36.92% of these 130 (48 patients) provided date of diagnosis and remaining 63.08% (82 patients) did not include the time factor. This available data is segregated to generate a small test dataset (15%) to be applied against the rest (85%). Hence, data for 19 patients (8 with and 11 without date of diagnosis) constitutes the test dataset and rest 111 records is the data that is subject to analysis. For our study, we assume that the social data log is accurate and no variation in terms of the specificities as disclosed by the patient. Also, terms "disease" and "condition" are interchangeably used.

From data extracted for the 111 patients, we define two object properties – patient relation and time relation. As stated earlier, patient relation refers to the unique patient ID (PID#) and time relation pertains to the time of disease progression from one condition to another in each of the 40 patients who have published diagnosed dates. Hence, we arrive at 111 PIDs or patient relations and 150 time relations as the object properties. Also, there were 290 unique conditions identified in the 111 patients and of the 290 conditions only 211 conditions could be mapped in the DOID framework. The remaining 79 conditions were categorized as superclass "disease" under DOID. Table 6 represents the test dataset.

The real-time data obtained is pre-processed to generate "triplets" that constitute the PDKB build. The number of triplets for patient relation and time relation are obtained

TABLE 6. Test dataset from PLM.

Patient	Conditions with/without diagnosed date
1	Brain Aneurysm (02/15/2008), Migraine (07/01/2002)
2	Brain Aneurysm (06/23/1994), Epilepsy (08/15/1994)
3	Brain Aneurysm (04/01/2002), Diabetes Type 2 (12/30/2001), Hypercholesterolemia (07/01/2005)
4	Brain Aneurysm (07/01/2013), Multiple Sclerosis (07/01/1998), Osteoarthritis (07/01/2010)
5	Brain Aneurysm (02/15/2012), Diabetes Type 2 (12/12/2009), Epilepsy (03/01/2012)
6	Brain Aneurysm (07/16/2009), Hemorrhagic Stroke (07/16/2009), Rheumatoid Arthritis (12/15/2009), Sjogren's Syndrome (02/15/2010)
7	Brain Aneurysm (03/26/2007), Epilepsy (05/01/2009), Major Depressive Disorder (03/14/2012), Spinal Stenosis (03/26/2009)
8	Brain Aneurysm (06/10/2007), Fibromyalgia (08/15/2007), Hypertension (08/15/2008), Hypercholesterolemia (05/15/2008), Irritable Bowel Syndrome (06/15/2012)
9	Brain Aneurysm, Epilepsy
10	Brain Aneurysm, Hemorrhagic Stroke
11	Brain Aneurysm, Hypertension
12	Brain Aneurysm, Multiple Sclerosis
13	Brain Aneurysm, Stroke
14	Brain Aneurysm, Subarachnoid Hemorrhage
15	Brain Aneurysm, Migraine, Subarachnoid Hemorrhage
16	Brain Aneurysm, Diabetes Type 2, Insomnia
17	Brain Aneurysm, Stroke, Systemic Lupus Erythematosus
18	Brain Aneurysm, Epilepsy, Hydrocephalus
19	Brain Aneurysm, Diabetes Type 2, Fibromyalgia, Osteoarthritis

TABLE 7. Number of triplets in PDKB having patient relation.

No. of Conditions in a patient	No. of Triplets based on No. of conditions (from Eq.1)	No. of Patients with No. of Conditions specified in column 1	Number of Triplets in PDKB
2	2	19	2*19=38
3	6	15	6*15=90
4	12	8	12*8=96
5	20	9	20*9=180
6	30	9	30*9=270
7	42	9	42*9=378
8	56	7	56*7=392
9	72	11	72*11=792
10	90	7	90*7=630
11	110	2	110*2=220
13	156	2	156*2=312
14	182	1	182*1=182
16	240	1	240*1=240
17	272	2	272*1=272
18	306	1	306*1=306
20	380	3	380*3=1140
21	420	1	420*1=420
25	600	2	600*2=1200
38	1406	1	1406*1=1406
39	1482	1	1482*1=1482
Total		111	10,046

using Equation 1 and 2 respectively. Tables 7 and 8 represent the values generated for the current Brain Aneurysm case study.

TABLE 8. Number of triplets in PDKB having time relation.

No. of Conditions in a patient	No. of Triplets based on No. of conditions (from Eq.2)	No. of Patients with No. of Conditions specified in column 1	Number of Triplets in PDKB
2	1	6	1*6=6
3	2	7	2*7=14
4	3	6	3*6=18
5	4	5	4*5=20
6	5	3	5*3=15
7	6	2	6*2=12
8	7	1	7*1=7
9	8	6	8*6=48
10	9	1	9*1=9
11	10	1	10*1=10
17	16	1	16*1=16
18	17	1	17*1=17
Total		40	192

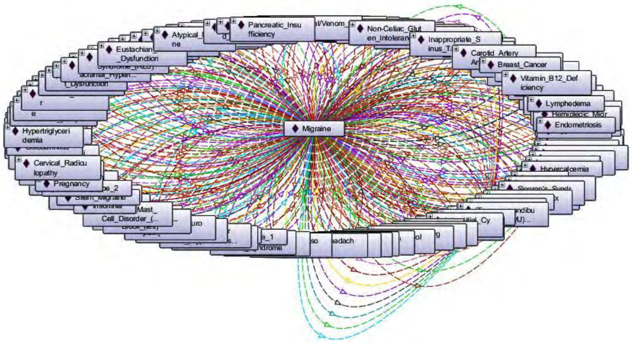


FIGURE 11. Sub-graph of comorbidities for Migraine from PDKB.

B. PROBABLE COMORBIDITIES

In this section, we address the prediction of probable comorbidities for a given input condition from the test dataset. Considering Migraine as the input condition, the sub-graph mapping all its comorbidities is represented in Fig. 11. There were a total of 15 patients (edges) identified with Migraine (source node) also reporting co-existence of 125 other conditions (target nodes) that are represented in the sub-graph.

Table 9 captures the node-edge relationship between Migraine and 125 identified conditions with percentage of co-occurrence in the 15 patients.

For our evaluation we consider a cutoff of 10% chance of occurrence which yields 36 conditions in the case of Migraine. The remaining 89 conditions with <10% probability are delisted. Similarly, we considered all conditions enlisted in the test dataset to obtain the >10% chance of occurrence for comorbidity prediction. From the test dataset (Table 6), for each of the input condition in a patient, all expected output condition/s are considered. Samples for four patients are shown in Table 10

Every condition observed as an ''Expected Output Condition'' for the 19 patients, total number of expected matches is identified in Table 11. Hence, for the 20 distinct conditions in test dataset, a total of 114 matches are expected

Confusion matrix is used to obtain the ''true'' match between the expected output condition versus the

TABLE 9. Comorbidities of migraine and their probable occurrence in %.

Source Condition	Comorbid condition/s	No. of edges/ patients	Chance of occurrence (%)
Migraine (15 patients)	Brain Aneurysm (1 condition)	13	(13/15)*100=86.66
	Fibromyalgia (1 condition)	8	(8/15)*100=53.33
	Osteoarthritis (1 condition)	5	(5/15)*100=33.33
	Major Depressive Disorder & Peripheral Neuropathy (2 conditions)	4	(4/15)*100=26.66
	Asthma, Degenerative Disc Disease, Dysautonomia, Ehlers Danlos Syndrome, GERD, Generalized Anxiety Disorder, Hypertension, IBS, Kidney stone, Rheumatoid Arthritis (RA),Seasonal Allergy & Systemic Lupus Erythematosus (12 conditions)	3	(3/15)*100=20
	Autonomic Neuropathy, Cataract, Dry Eye Syndrome, Epilepsy, Food allergy, Hypermobility Syndrome, Hypothyroidism, Insomnia, Myalgic Encephalomyelitis, Chronic Fatigue Syndrome, Orthostatic Hypotension, Osteopenia, Panic Disorder, Pregnancy, Restless Legs Syndrome (RLS),Sacroiliac Joint Dysfunction, Sjogren's Syndrome, Spinal Stenosis, Stroke & Temporomandibular joint (TMJ) Syndrome (19 conditions)	2	(2/15)*100=13.33
	Other (89 conditions)	1	(1/15)*100=6.66

predicted output condition for the >10% probabilities as shown in Table 12. This provides a visual interpretation of the performance of our prediction method [43]. Further sensitivity measure is calculated to obtain the true positive rate of our evaluation using Equation 4.

$$Sensitivity\% = \left(\frac{No.\ of\ predicted\ `true'\ matches}{No.\ of\ expected\ `true'\ matches} \right) \times 100 \tag{4}$$

TABLE 10. Expected output conditions for a sample of four patients.

Patient	Input Condition	Expected Output Condition
1	Brain Aneurysm	Migraine
	Migraine	Brain Aneurysm
3	Brain Aneurysm	Diabetes Type 2, Hypercholesterolemia
	Diabetes Type 2	Brain Aneurysm, Hypercholesterolemia
	Hypercholesterolemia	Brain Aneurysm, Diabetes Type 2
6	Brain Aneurysm	Hemorrhagic Stroke, Rheumatoid Arthritis, Sjogren's Syndrome
	Hemorrhagic Stroke	Brain Aneurysm, Rheumatoid Arthritis, Sjogren's Syndrome
	Rheumatoid Arthritis	Brain Aneurysm, Hemorrhagic Stroke, Sjogren's Syndrome
	Sjogren's Syndrome	Brain Aneurysm, Hemorrhagic Stroke, Rheumatoid Arthritis
8	Brain Aneurysm	Fibromyalgia, Hypertension, Hypercholesterolemia, Irritable Bowel Syndrome
	Fibromyalgia	Brain Aneurysm, Hypertension, Hypercholesterolemia, Irritable Bowel Syndrome
	Hypertension	Brain Aneurysm, Fibromyalgia, Hypercholesterolemia, Irritable Bowel Syndrome
	Hypercholesterolemia	Brain Aneurysm, Fibromyalgia, Hypertension, Irritable Bowel Syndrome
	Irritable Bowel Syndrome	Brain Aneurysm, Fibromyalgia, Hypertension, Hypercholesterolemia

TABLE 11. Conditions and expected matches.

Condition	No. of Expected matches
Brain Aneurysm	35
Migraine	3
Epilepsy	9
Diabetes Type 2	9
Hypercholesterolemia	6
Multiple Sclerosis	3
Osteoarthritis	5
Hemorrhagic Stroke	4
Rheumatoid Arthritis	3
Sjogren's Syndrome	3
Major Depressive Disorder	3
Spinal Stenosis	3
Fibromyalgia	7
Hypertension	5
Stroke	3
Insomnia	2
Hydrocephalus	2
Subarachnoid Hemorrhage	3
Systemic Lupus Erythematosus	2
Irritable Bowel Syndrome	4
	Total: 114

TABLE 12. Confusion matrix of probable comorbidities.

Confusion Matrix		**Predicted**	
Total expected match = 114		TRUE	FALSE
Expected	TRUE	73	41
	FALSE	0	0

Overall, for the considered PDKB construction, the sensitivity percentage in predicting probable comorbid conditions for a scenario with >10% co-occurrence was obtained to be 64%.

TABLE 13. Cluster representation of comorbid instances for Migraine.

DOID Classes			Instances	
Super Class	Parent Class	Child Class	No. of Comorbidities	Distribution in %
Disease	Disease by infectious agent	Viral infectious disease	1	(1/125)*100 =0.8
	Disease of anatomical entity	Cardiovascular system disease	12	(12/125)*100 =9.6
		Endocrine system disease	3	(3/125)*100 =2.4
		Gastrointestinal system disease	3	(3/125)*100 =2.4
		Hematopoietic system disease	2	(2/125)*100 =1.6
		Immune system disease	15	(15/125)*100 =12
		Integumentary system disease	11	(11/125)*100 =8.8
		Musculoskeletal system disease	4	(4/125)*100 =3.2
		Nervous system disease	11	(11/125)*100 =8.8
		Reproductive system disease	1	(1/125)*100 =0.8
		Respiratory system disease	5	(5/125)*100 =4
		Urinary system disease	4	(4/125)*100 =3.2
	Disease of cellular proliferation	Benign neoplasm	1	(1/125)*100 =0.8
		Cancer	1	(1/125)*100 =0.8
	Disease of mental health	Cognitive disorder	8	(8/125)*100 =6.4
		Developmental disorder	1	(1/125)*100 =0.8
		Sleep disorder	2	(2/125)*100 =1.6
	Disease of metabolism	Acquired metabolic disease	4	(4/125)*100 =3.2
		Inherited metabolic disorder	1	(1/125)*100 =0.8
	Syndrome		5	(5/125)*100 =4
			30	(30/125)*100 =24

For HCA evaluation we group all the identified comorbidities associated with the given input condition to their respective DOID classes. As stated in methodology, DOID can have several parent classes and child classes represented hierarchically. However, we consider only the top 3 DOID classes for our evaluation (named as super class, parent class and child class respectively). Considering the same example for analysis, we already know that 15 patients and 125 comorbidities are associated with Migraine. Each of these 125 comorbid conditions is mapped to the three DOID classes with individual distribution percentages as in Table 13.

In this evaluation, we consider only those entries with a threshold of 5% and above of instance distribution. Also, conditions that only have an association with the super class "Disease" and no subsequent parent and child classes are excluded from our analysis. From the test dataset (Table 6), for each input condition, the expected DOID class for every associated comorbid condition is obtained. Sample associations for patients 1, 3, 6 and 8 are represented in Table 14.

TABLE 14. Expected DOID classes for a sample of four patients.

Patient	Input Condition	Expected DOID Class
1	Brain Aneurysm	immune system disease (Migraine)
	Migraine	cardiovascular system disease (Brain Aneurysm)
3	Brain Aneurysm	acquired metabolic disease (Diabetes Type 2), Disease (Hypercholesterolemia)
	Diabetes Type 2	cardiovascular system disease (Brain Aneurysm), Disease (Hypercholesterolemia)
	Hypercholesterolemia	cardiovascular system disease (Brain Aneurysm), acquired metabolic disease (Diabetes Type 2)
6	Brain Aneurysm	cardiovascular system disease (Hemorrhagic Stroke), integumentary system disease (Rheumatoid Arthritis), immune system disease (Sjogren's Syndrome)
	Hemorrhagic Stroke	cardiovascular system disease (Brain Aneurysm), integumentary system disease (Rheumatoid Arthritis), immune system disease (Sjogren's Syndrome)
	Rheumatoid Arthritis	cardiovascular system disease (Brain Aneurysm, Hemorrhagic Stroke), immune system disease (Sjogren's Syndrome)
	Sjogren's Syndrome	cardiovascular system disease (Brain Aneurysm, Hemorrhagic
		Stroke), integumentary system disease (Rheumatoid Arthritis)
8	Brain Aneurysm	musculoskeletal system disease (Fibromyalgia), cardiovascular system disease (Hypertension), Disease (Hypercholesterolemia), Syndrome (Irritable Bowel Syndrome)
	Fibromyalgia	cardiovascular system disease (Brain Aneurysm, Hypertension), Disease (Hypercholesterolemia), Syndrome (Irritable Bowel Syndrome)
	Hypertension	cardiovascular system disease (Brain Aneurysm), musculoskeletal system disease (Fibromyalgia), Disease (Hypercholesterolemia), Syndrome (Irritable Bowel Syndrome)
	Hypercholesterolemia	cardiovascular system disease (Brain Aneurysm, Hypertension), musculoskeletal system disease (Fibromyalgia), Syndrome (Irritable Bowel Syndrome)
	Irritable Bowel Syndrome	cardiovascular system disease (Brain Aneurysm, Hypertension), musculoskeletal system disease (Fibromyalgia), Disease (Hypercholesterolemia)

Every class (condition) captured in "Expected DOID Class" for 19 patients, the total number of expected matches is identified and presented in Table 15. For the 20 conditions in test dataset, a total of 106 matches are expected. However, 10 out of the 106 matches have a single DOID class mapping to "Disease", hence not considered and leaves us with a total of 96 expected DOID class matches.

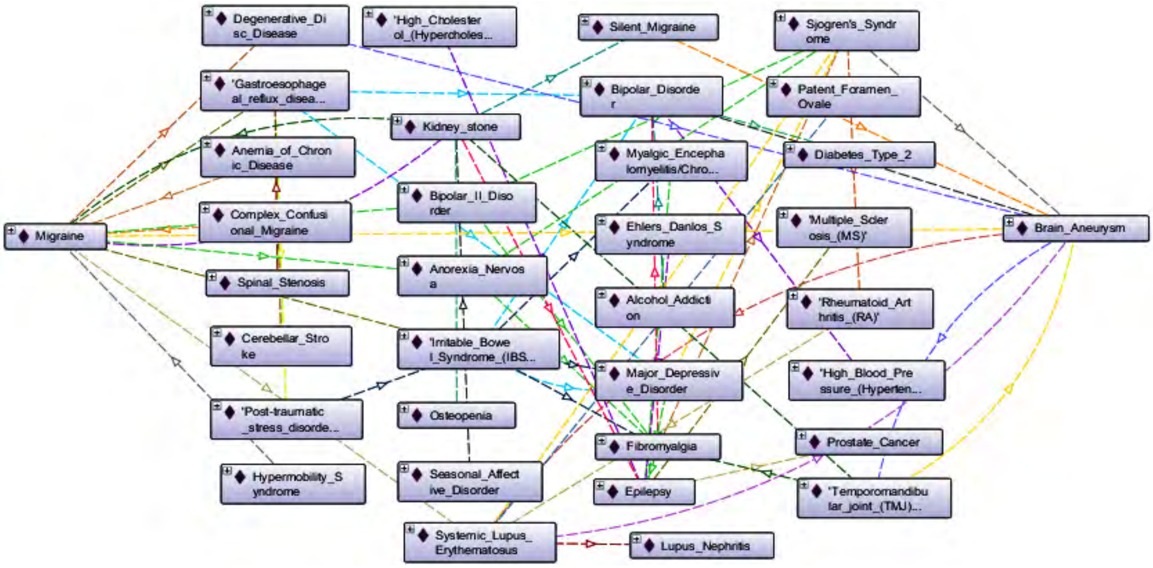

Degenerative_Disc_Disease
'High_Cholesterol_(Hypercholes...
Silent_Migraine
Sjogren's_Syndrome
'Gastroesophageal_reflux_disea...
Bipolar_Disorder
Patent_Foramen_Ovale
Kidney_stone
Anemia_of_Chronic_Disease
Myalgic_Encephalomyelitis/Chro...
Diabetes_Type_2
Bipolar_II_Disorder
Complex_Confusional_Migraine
Ehlers_Danlos_Syndrome
'Multiple_Sclerosis_(MS)'
Brain_Aneurysm
Migraine
Anorexia_Nervosa
Spinal_Stenosis
Alcohol_Addiction
'Rheumatoid_Arthritis_(RA)'
Cerebellar_Stroke
'Irritable_Bowel_Syndrome_(IBS...
Major_Depressive_Disorder
'High_Blood_Pressure_(Hyperten...
'Post-traumatic_stress_disorde...
Osteopenia
Fibromyalgia
Prostate_Cancer
Hypermobility_Syndrome
Seasonal_Affective_Disorder
Epilepsy
'Temporomandibular_joint_(TMJ)...
Systemic_Lupus_Erythematosus
Lupus_Nephritis

FIGURE 12. Sub-graph of time progression from Migraine to Brain Aneurysm.

TABLE 15. DOID classes and expected matches.

DOID Classes	No. of Expected True
Immune system disease	21
Cardiovascular system disease	41
Acquired metabolic disease	9
Integumentary system disease	11
Cognitive disorder	3
Musculoskeletal system disease	7
Syndrome	4
	Total: 96

TABLE 16. Confusion matrix of probable DOID classes.

Confusion Matrix		**Predicted**	
Total expected match = 96		TRUE	FALSE
Expected	TRUE	76	20
	FALSE	0	0

The confusion matrix for our HCA is represented in Table 16.

Hence, for the case study, we arrived at a sensitivity percentage of 79% in predicting DOID class associations for the comorbid conditions for a scenario with >5% distribution.

C. TIME RANGE PROBABILITIES

This section covers the third objective of our research to evaluate the time progression from source condition to target condition. Here, we study the time progression from Migraine (source) to Brain Aneurysm (target) using sub-graphs, matrix representation, *FindPath* function and normal distribution discussed in Methodology section. There are 52 identified nodes (conditions), connected by 120 unique edges between Migraine and Brain Aneurysm as per sub-graph depicted in Fig. 12.

TABLE 17. Details of path from migraine to brain aneurysm.

Parameter	Value
Source Node s	Migraine
Target Node t	Brain Aneurysm
Total Nodes between s & t	52
Total Edges between s & t (g)	120
Max. length of path *Kspec*	7
Length (paths between s & t)	No. of paths
1	1
2	2
3	13
4	32
5	70
6	119
7	230
TOTAL	467

FindPath function is applied to obtain all paths from Migraine to Brain Aneurysm. Plugging in the corresponding values for expression 1, graph value g substitutes to 120 edges and we consider *Kspec* value to 7 instead of Infinity to provide an easily decipherable snapshot of our analysis. Table 17 provides an overview of our current consideration which yields a total of 467 time-weighted paths of disease progression from Migraine to Brain Aneurysm. Each path has a time progression value in days and the end to end progression time is the summation of all times associated with individual edges comprising the path.

Gaussian normal distribution is applied to all the 467 time associations for each path to obtain the bell curve plot representing the time progression for entire population range. The probability density function for normal distribution is represented in Equation 5 [22].

$$PDF = \frac{1}{\sqrt{2\pi\sigma^2}}(e \cdot \frac{-\mu\,(x-)^2}{2\sigma^2}) \tag{5}$$

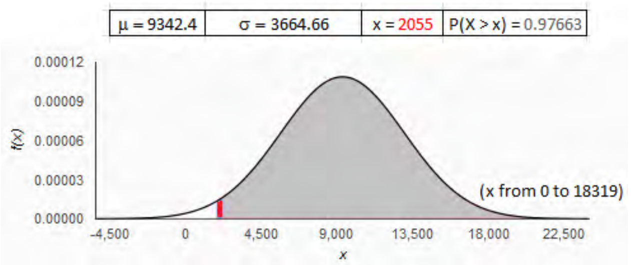


FIGURE 13. Bell curve representing time progression of 467 Paths between Migraine and Brain Aneurysm.

$$\begin{aligned} \text{Where, PDF} &= \text{Probability Density Function} \\ \sigma &= \text{Standard Deviation} \\ \mu &= \text{Mean} \end{aligned}$$

For our case, σ has a value of 3664.66 and μ has a corresponding value of 9342.40. Substituting in the Equation 5 we obtain the plot as in Fig. 13. Based on test data, we have a single patient who progresses from Migraine to Brain Aneurysm via a single path taking 2055 days (Migraine diagnosed on 07/01/2002 and Brain Aneurysm on 02/15/2008). This is the observation index value (x) for that patient that is positioned under the bell curve of PDF. The z-score analysis for this case is represented in the same plot and discussed subsequently.

In methodology, we defined the significance of z-scores in predicting the probability of disease progression from source to target when the PDF can be represented as a normal distribution. The z-score can be calculated using Equation 6 [22].

$$Z\ score = (x - \mu)/\sigma \tag{6}$$

In the case of the single patient whose progression from Migraine to Brain Aneurysm took 2055 days, the Z-score obtained is −1.9886. This implies that there is a 97.7% probability of a person with Migraine to acquire Brain Aneurysm within 2055 days based on population data represented in the bell curve.

V. CONCLUSION AND FUTURE WORK

In this paper, we discussed the importance of representing patient social network data as a significant contributor for evidence-based healthcare studies in CDSS. Knowledge and ontology engineering fundamentals with statistical modeling were used to predict condition and time of disease progression based on comorbidity behaviors for a sample case study considering Brain Aneurysm. The outcomes obtained from our evaluation convey the impact of patient and time relationships in predicting disease progression trends. Our current work is a suggested research approach implying the importance of knowledge engineering of patient data to infer futuristic patterns of comorbid progression. This can be further extended to other specific healthcare studies as part of future work. We recommend a few areas of enrichment that can be applied for our current work.

In our work, we consider only the DOID ontology framework for mapping disease conditions; this can be enhanced to include other standard ontologies such as SNOMEDCT, NIH MeSH etc. to encompass diverse range of identified conditions that can provide holistic associations. Also, we have considered only the diagnostic aspect of conditions; the same approach can be extended to map disease symptoms and treatments by referring to related ontologies. The greater the number of reference ontologies and higher the variance of expected outcomes, more comprehensive and spectral can be the area of application.

The dataset used for our evaluation was obtained from PLM for only one disease condition of Brain Aneurysm to demonstrate the methodology as a sample case study. PLM in its entirety can be considered to provide disease progression trends for over 2,700 conditions listed in its database. Likewise, the system can be realized to include practice based and literature-based evidence with social network data to obtain an enriched inference.

Overall, our current work aims at providing a foundation for evidence-based CDSS in predicting disease progression trends. This can be improved using automation tools, enhanced by the application of customized inference engines and extended to other methodologies such as graph similarity to provide a more granular approach.

REFERENCES

[1] *Disease Ontology*. Accessed: 2017. [Online]. Available: http://disease-ontology.org/

[2] A. Fan, D. Lin, and Y. Tang, "Clinical decision support systems for comorbidity: Architecture, algorithms, and applications," *Int. J. Telemed. Appl.*, vol. 2017, Mar. 2017, Art. no. 1562919.

[3] B. Jafarpour and S. S. R. Abidi, "Merging disease-specific clinical guidelines to handle comorbidities in a clinical decision support setting," in *Artificial Intelligence in Medicine* (Lecture Notes in Computer Science), vol. 7885. Berlin, Germany: Springer, 2013, pp. 28–32.

[4] *Bioontology*. Accessed: 2017. [Online]. Available: http://data.bioontology.org/documentation

[5] *Bioontology Wiki*. Accessed: 2017. [Online]. Available: https://www.bioontology.org/wiki/index.php/Ontology_Metrics#Ontology_Metrics_in_BioPortal

[6] *BioPortal*. Accessed: 2017. [Online]. Available: https://bioportal.bioontology.org/ontologies/DOID

[7] *Bioportal Ontologies*. Accessed: 2017. [Online]. Available: https://bioportal.bioontology.org/ontologies

[8] C. Nakamura, M. Bromberg, S. Bhargava, P. Wicks, and Q. Zeng-Treitler, "Mining online social network data for biomedical research: A comparison of clinicians' and patients' perceptions about amyotrophic lateral sclerosis treatments," *J. Med. Internet Res.*, vol. 14, no. 3, p. e90, 2012.

[9] C. M. Boyd *et al.*, "Informing clinical practice guideline development and implementation: Prevalence of coexisting conditions among adults with coronary heart disease," *J. Amer. Geriartrics Soc.*, vol. 59, no. 5, pp. 797–805, 2011.

[10] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Hierarchical clustering," in *Cluster Analysis*, 5th ed. Chichester, U.K.: Wiley, 2011, ch. 4.

[11] S. Falconer. (Apr. 12, 2010). *ProtégéWiki*. Accessed: 2017. [Online]. Available: https://protegewiki.stanford.edu/wiki/OntoGraf

[12] G. S. Eichler *et al.*, "Exploring concordance of patient-reported information on patientslikeme and medical claims data at the patient level," *J. Med. Internet Res.*, vol. 18, no. 5, p. e110, 2016.

[13] A. Waigandt, *Statistics Primer*. San Jose, CA, USA: San Jose State Univ., 2016.

[14] G. De Giacomo and M. Lenzerini. "TBox and ABox reasoning in expressive description logics," AAAI, Menlo Park, CA, USA, Tech. Rep. WS-96-05, 1996.

[15] H. A. Close, L.-C. Lee, C. N. Kaufmann, and A. W. Zimmerman, "Co-occurring conditions and change in diagnosis in autism spectrum disorder," Amer. Acad. Pediatrics, Itasca, IL, USA, Tech. Rep., 2012, doi: 10.1542/peds.2011-1717.

[16] HLN Consulting, "Future information capabilities for public health. Clinical decision support," Joint Public Health Inform. Task Force, LLC, Palm Desert, CA, USA, Tech. Rep., 2013.

[17] I. Sim *et al.*, "Clinical decision support systems for the practice of evidence-based medicine," *J. Amer. Med. Inf. Assoc.*, vol. 8, no. 6, pp. 527–534, 2001.

[18] *InternetLiveStats*. Accessed: 2017. [Online]. Available: http://www.internetlivestats.com/

[19] J. Chen, J. Liu, and W. Yu, "Discovering semantic relationships for knowledgebase," in *Proc. IEEE ICPCA*, Oct. 2008, pp. 242–247.

[20] K. Denecke, "Integrating social media and mobile sensor data for clinical decision support: concept and requirements," in *Studies in Health Technology and Informatics*, vol. 225. Amsterdam, The Netherlands: IOS Press, 2016, pp. 562–566.

[21] K. Griva, N. Mooppil, E. Khoo, V. Y. W. Lee, A. W. C. Kang, and S. P. Newman, "Improving outcomes in patients with coexisting multimorbid conditions—The development and evaluation of the combined diabetes and renal control trial (C-DIRECT): Study protocol," *BMJ Open*, vol. 5, no. 2, p. e007253, 2015.

[22] *Laerd Statistics*. Accessed: 2017. [Online]. Available: https://statistics.laerd.com/statistical-guides/standard-score-2.php

[23] J. Loechner, "90% of today's data created in two years," MediaPost Commun., New York, NY, USA, Tech. Rep., Dec. 2016.

[24] M. Afzal, M. Hussain, R. B. Haynes, and S. Lee, "Context-aware grading of quality evidences for evidence-based decision-making," *Health Inform. J.*, p. 1460458217719560, Aug. 2017, doi: 10.1177/1460458217719560.

[25] M. Krishnamurthy, K. Mahmood, and P. Marcinek, "A hybrid statistical and semantic model for identification of mental health and behavioral disorders using social network analysis," in *Proc. ASONAM*, San Francisco, CA, USA, Aug. 2016, pp. 1019–1026.

[26] M. Peleg, N. Asbeh, T. Kuflik, and M. Schertz, "Onto-clust—A methodology for combining clustering analysis and ontological methods for identifying groups of comorbidities for developmental disorders," *J. Biomed. Inform.*, vol. 42, no. 1, pp. 165–175, 2009.

[27] M. Thirugnanam, M. Ramaiah, V. Pattabiraman, and R. Sivakumar, "Ontology based disease information system," in *Proc. Int. Conf. Modeling Optim. Comput.*, 2012, pp. 1–7.

[28] N. Mani and B. Lithgow, "Medical and healthcare applicationsof intelligent information systems: An overview," *Proc. Inaugral Conf. Victorian Chapter IEEE Eng. Med. Biol. Soc.*, Jan. 1999.

[29] O. Deshpande *et al.*, "Building, maintaining, and using knowledge bases: a report from the trenches," in *Proc. SIGMOD Int. Conf. Manage. Data*, New York, NY, USA, 2013, pp. 1209–1220.

[30] P. V. O'Neil, *Advanced Engineering Mathematics*, 7th ed. Pacific Grove, CA, USA: Brooks/Cole, 1995.

[31] *Openclinical*. Accessed: 2017. [Online]. Available: http://www.openclinical.org/dss.html

[32] P. Kalgotra, R. Sharda, B. Molaka, and S. Kathuri, "Time-based comorbidity in patients diagnosed with tobacco use disorder," in *Guide to Big Data Applications*. Cham, Switzerland: Springer, 2017, pp. 401–413.

[33] *PatientsLikeMe*. Accessed: 2017. [Online]. Available: https://www.patientslikeme.com/

[34] S. Van de Velde *et al.*, "Tailoring implementation strategies for evidence-based recommendations using computerised clinical decision support systems: Protocol for the development of the GUIDES tools," *Implement. Sci.*, vol. 11, p. 29, Mar. 2016.

[35] D. Selinger, "Big data: Getting ready for the 2013 big bang," Forbes, Jersey City, NJ, USA, Tech. Rep., Jan. 2013.

[36] Stanford University. *Protege*. Accessed: 2017. [Online]. Available: http://protege.stanford.edu/

[37] S. K. Gruschkus, J. M. Darragh, M. A. Kolodziej, R. A. Beveridge, M. Forsyth, and C. Reyes, "Impact of disease progression on healthcare cost and resource utilization among follicular NHL patients treated within the US oncology network," *Blood*, vol. 114, p. 4543, Oct. 2015.

[38] D. M. Blei, *Hierarchical Clustering*. Princeton, NJ, USA: Princeton Univ. Press, Feb. 28, 2008.

[39] W3C Semantic Web. (Fed. 25, 2014). *Resource Description Framework (RDF)*. Accessed: 2017. [Online]. Available: https://www.w3.org/RDF/

[40] WHO. (2001). *Chronic Respiratory Diseases*. [Online]. Available: http://www.who.int/gard/publications/chronic_respiratory_diseases.pdf

[41] *Wolfram*. Accessed: 2017. [Online]. Available: http://reference.wolfram.com/language/ref/FindPath.html

[42] *Wolfram Mathworld*. Accessed: 2017. [Online]. Available: http://mathworld.wolfram.com/AdjacencyMatrix.html

[43] *WordPress*. Accessed: 2017. [Online]. Available: https://classeval.wordpress.com/introduction/basic-evaluation-measures/

[44] X. Ji, S. A. Chun, P. Cappellari, and J. Geller, "Linking and using social media data for enhancing public health analytics," *J. Inf. Sci.*, vol. 43, no. 2, pp. 221–245, 2016.

[45] X. Ji, S. A. Chun, and J. Geller, "Predicting comorbid conditions and trajectories using social health records," *IEEE Trans. Nanobiosci.*, vol. 15, no. 4, pp. 371–379, Jun. 2016.

[46] Z. Zhou *et al.*, "Healthcare resource use, costs, and disease progression associated with diabetic nephropathy in adults with type 2 diabetes: A retrospective observational study," *Diabetes Therapy*, vol. 8, no. 3, pp. 555–571, 2017.

MADAN KRISHNAMURTHY received the master's degree in information technology from Towson University, Towson, MD, USA. He is currently pursuing the Ph.D. degree in computer science and informatics with Oakland University, Rochester, MI, USA. His research interests include health informatics and social network analysis in medical domain using data science and Semantic Web Technologies, such as knowledge engineering, ontology engineering, and linked data.

PAWEL MARCINEK received the master's degree in experimental physics from Jagiellonian University, Krakow, Poland. He is currently pursuing the Ph.D. degree in applied mathematics with Oakland University, Rochester, MI, USA. His research interests include numerical methods for partial differential equations, applications of PDE in physics and in mechanical engineering, cluster analysis, risk analysis, and statistical predictive modeling.

KHALID MAHMOOD MALIK is currently an Assistant Professor with the School of Engineering and Computer Science, Oakland University, Rochester, MI, USA. His research interests include integrated area of health informatics, cognitive mobile computing, autonomous decentralized systems (ADS), Internet of Things, and Semantic web. He lead various academic and industrial projects in area of ontology based information extraction for health informatics, Semantic based information security and data loss prevention, ADS based architecture and its applications, big data analytics using linked data, and Semantic based web filtering.

MUHAMMAD AFZAL received the B.S degree in computer science from the Kohat University of Science and Technology, Pakistan, the M.S. degree from the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Pakistan, in 2009, and the Ph.D. degree from Kyung Hee University, South Korea, in 2017 with Excellent Thesis Award. He is currently an Assistant Professor with Sejong University, Seoul, South Korea. He has been serving as a proctor of HL7 International for conducting HL7 Standard Certification Exams in Pakistan since 2010. He has co-authored over 60 publications in different reputed journals/conferences. His current research interests include evidence-adaptive decision support systems, knowledge acquisition and management, precision medicine, applications of machine learning, text processing, and information extraction.