Predicting Comorbid Conditions and Trajectories using Social Health Records

Xiang Ji¹, Soon Ae Chun², James Geller¹,

¹ Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA
² Department of Information Systems and Informatics, City University of New York, Staten Island, NY, USA
xj25@njit.edu, soon.chun@csi.cuny.edu, james.geller@njit.edu,

Abstract—Many patients suffer from comorbidity conditions, for example, obese patients often develop type-2 diabetes and hypertension. In the US, 80% of Medicare spending is for managing patients with these multiple coexisting conditions. Predicting potential comorbidity conditions for an individual patient can promote preventive care and reduce costs. Predicting possible comorbidity progression paths can provide important insights into population health and aid with decisions in public health policies. Discovering the comorbidity relationships is complex and difficult, due to limited access to Electronic Health Records by privacy laws. In this paper, we present a collaborative comorbidity prediction method to predict likely comorbid conditions for individual patients, and a trajectory prediction graph model to reveal progression paths of comorbid conditions. Our prediction approaches utilize patient generated health reports on online social media, called Social Health Records (SHR). The experimental results based on one SHR source show that our method is able to predict future comorbid conditions for a patient with coverage values of 48% and 75% for a top-20 and a top-100 ranked list, respectively. For risk trajectory prediction, our approach is able to reveal each potential progression trajectory between any two conditions and infer the confidence of the future trajectory, given any observed condition. The predicted trajectories are validated with existing comorbidity relations from the medical literature.

Keywords—Disease Progression; Collaborative Prediction; Comorbidity Prediction; Mining Social Media; Trajectory Prediction

I. INTRODUCTION

Conditions of patients are diagnosed, and these diagnoses are recorded, e.g., in an electronic health record (EHR). Research has shown that some conditions are correlated with each other to a measureable degree (“comorbidities”) [1]. It is difficult to find the trajectory of conditions that may be correlated. Research has been carried out on predicting incidence [2-5] and progression trajectories [6, 7] from EHR data. However, EHR datasets are usually limited to one medical site or network and have limited coverage of population and time period. Moreover, because of HIPAA laws, EHR datasets are rarely accessible to non-affiliated researchers, thus the opportunities for research are often quite limited. To solve these problems, we tapped into self-posted medical histories on a well-known medical social media site, which may cover patients all around the world. Compared with EHRs, the data on social media have the advantages of open access and the lack of privacy issues. In patient social media, the patients voluntarily post their health status, with the purpose of letting others view and analyse the data and possibly provide advice. We define this kind of social media-based health data as Social Health Records (SHR). The SHRs contain self-reported data by patients, and are created to be shared with other patients on social media platforms, while EHRs are usually created by clinicians or health care providers, using specialized EHR systems, and shared among clinical staff for care and billing purposes.

To address the challenges of predicting a comorbid medical condition incidence and progression of medical conditions, we developed two prediction models for comorbidity relationships and condition trajectories based on SHRs, self-posted data available on patient-oriented social media sites. To the best of our knowledge, this is the first non-disease-specific work on modeling comorbidity based on patients’ social media postings. First, to predict the most probable conditions a patient will develop in the future, given the available medical history posted on his/her social media site, we utilized a collaborative filtering technique, which is widely used in applications such as TV programs [8], books [9], and online dating [10]. For this problem, patients are viewed as users, diagnosed conditions as items, and presence or absence of a condition as a rating with binary values. We calculated the similarity between a patient’s record and other patients’ records and derived the risk of a certain medical condition. The output is a ranked list of medical conditions for a patient.

Secondly, to infer medical condition progression trajectories given a certain observed medical condition, we propose an algorithm to build the trajectory model from the online patients’ diagnosis history. There are three steps to generate and make use of the trajectory model. The first step is called Edge Discovery, which identifies directional edges of comorbidities that co-occur for individual patients. The second step is called Linking, in which the generated edges are recursively linked to build the condition trajectory tree by recognizing the common node (condition) in two edges. In the last step, the trajectory model is used to infer the confidence value and support of potential progression trajectories given an observed condition. The predictive models of medical
condition incidence and progression trajectory based on patients’ social media data can provide insights for doctors and patients, to identify potential risks more quickly and to intervene and mitigate the risks at the earliest possible stage.

In Section II, related work is discussed. In Section III, we introduce a collaborative filtering-based method for predicting medical condition incidence. In Section IV, a trajectory model for inferring future medical condition progression trajectories is presented. In Section V, the evaluation study results are illustrated, interpreted, and discussed. Section VI contains conclusions and suggestions for future research.

II. RELATED WORK

One thread of related research is to utilize data mining techniques to predict disease risks for individuals or to rank diseases by their risks. Davis et al. [2, 3] proposed CARE, which is the first well-known system for patient disease prediction using 13+ million elderly patients’ hospital visit records. They developed a method to predict the disease risk of one patient based on the disease risks of other similar patients. Hassan and Syed [5] summarized the reasons why collaborative filtering (CF) can be used to solve the problem of ranking patients along a continuum of risk for adverse outcomes. They incorporated demographics, comorbidity, lab test results, and outcomes into the feature space of their method. They concluded that collaborative filtering is the best method in predicting sudden cardiac death and recurrent myocardial infarction on a real-world dataset containing 4,557 patients’ records. Folino and Pizzuti [4] utilized the K-Means algorithm to cluster patients and applied association rule analysis to predict disease for patients in each cluster. Duan et al. [11] proposed to use correlations among nursing diagnoses, outcomes, and interventions to create a recommender system for constructing nursing care plans. Wiesner and Pfeifer [12] introduced a graph-based data structure of health-related concepts extracted from information in Wikipedia. Based on the health graph, they presented a recommendation procedure that makes use of a similarity measure to compute relevance with regard to users’ information needs. Qian et al. [13] investigated the patient risk prediction problem in the context of active learning with relative similarities. Hussein et al. [14] developed the Chronic Disease Recommender System to suggest medical advice and diagnoses to patients.

The above projects predict medical condition incidence but are not able to predict the medical condition progression trajectory. Another thread of research attempts to reveal and infer condition progression trajectories. Jensen et al. [6] investigated the temporal trajectory patterns of all diseases for the entire country of Denmark. They stratified the diagnoses by gender, age, and hospital encounter type, and identified 1,171 significant trajectories. Then they used the Markov Cluster algorithm to identify the five largest clusters of disease trajectories that centered on a small number of key diagnoses. Wang et al. [7] developed a disease progression model based on a Bipartite Bayesian Network; their model was able to identify a few comorbidities and infer the progression trajectory and comorbidity onset of individual patients. Hainke et al. [15] reviewed a number of disease progression models, which include path models, oncogenetic tree models, distance based trees, directed acyclic graph model, etc.

The existing research suffers from the following limitations. (1) Most of the above methods were developed on a single EHR dataset, which usually has limited coverage in terms of population and is hard to be integrated with other datasets due to different formats. Our method collected and preprocessed publicly available patients’ records, which can potentially cover patients around the world. (2) Since the graph-based progression trajectory construction process of Jensen et al. [6] is relatively difficult to explain and interpret, we propose a lightweight tree-based model inspired by oncogenetic tree models [15] to help reveal trajectory patterns in an intuitive and efficient manner. Different from oncogenetic trees, the actual trajectories and the patients who experience the trajectories were stored in the tree-based model. This was found to be an efficient method for calculating the confidence of future trajectories. This paper extends the preliminary study of comorbidity prediction models reported by Ji et al. [16]. It introduces the algorithm for a collaborative prediction model, and provides stratified comorbidity prediction models by gender.

III. COLLABORATIVE PREDICTION OF MEDICAL CONDITION

We present the collaborative prediction model to predict a ranked list of potential conditions for a patient, given a patient’s profile. In the first step, patients’ medical histories on their profiles were scraped. After data cleaning and filtering, the patients’ profiles were fed into the collaborative filtering model, training it to predict comorbidities. When the model is applied to a new patient’s record, the collaborative filtering computes the similarity between her/him and other similar patients and selects the neighborhood of patients who are most similar to the specific patient. Finally, the likelihood of each possible medical condition is calculated, and a ranked list based on the likelihood is built for this patient.

A. Collaborative Prediction Model

Patients and conditions are represented as a matrix, $M = I \times J$, where $I = \{\text{all patients}\}$ and $J = \{\text{all the possible conditions}\}$. $I = \{I_1, I_2, \ldots, I_n\}$ represents all the conditions of patient $i$ ordered by diagnosis date. Note that $J = \bigcup_{i \in I} \bar{J}(i)$. To predict medical conditions for a new patient $0$, we define $J_0 = \{C_1, C_2, \ldots, C_n\}$ as the existing conditions of patient $0$ and we define $H_0 = \{C_1, C_2, \ldots, C_k\}$ where $k \leq n$ to represent a head sequence of conditions that will be used as an input for the collaborative prediction model. In this case $H_0 = J_0$, since all of patient 0’s diagnosed conditions are used for the prediction model. We define $T_0 \subseteq J - H_0$ as a set of predicted conditions for patient $0$.

The goal of the collaborative prediction model is to predict the likelihood and the rank of each condition in $T_0$. Fig. 1 shows the process and components of the collaborative prediction model.
For each condition \( c \) in \( T_0 \), the neighbors \( N_c = \{ i | i \in I \land c \in J_1 \} \) are all other patients with condition \( c \). The probability of patient \( 0 \) having condition \( c \) in the future (represented as \( P_{0,c} \)) is calculated by the following equation:

\[
P_{0,c} = k \sum_{i \in N_c} w(0,i)
\]

where \( k \) is a normalizing factor, and is defined as reciprocal of the total number of patients in the neighborhood, formally \( k = 1/|N_c| \). \( w(0,i) \) is a measure of the similarity between patient \( 0 \) and patient \( i \), and is defined as the proportion of conditions of patient \( i \) to the conditions in head set of patient \( 0 \). Formally the similarity of patient \( i \) and patient \( j \) is defined in the following equation:

\[
w(i,j) = \frac{|\{ x | x \in H_0 \land x \in J_0 \}|}{|H_0|}
\]

where \( H_0 \) is the head set for patient \( 0 \). Then \( c \)'s support \( S_c \) is:

\[
S_c = \frac{1}{|I|} \sum_{i \in I} f(i)
\]

where \( f(i) \) is an indicator function. \( f(i) = 1 \) if \( c \in J_i \) and \( f(i) = 0 \) otherwise. The tuple \( <0, c, P_{0,c}, S_c> \) represents the fact that patient \( 0 \) has the probability of \( P_{0,c} \) of getting condition \( c \), and the condition \( c \)'s support is \( S_c \). After \( P_{0,c} \) and \( S_c \) have been computed, the list of potential conditions \( C* \) is defined as set of tuples \( C* = \{ <0, c, P_{0,c}, S_c> | c \in T_0 \} \) where \( c \) ranges over every condition in \( T_0 \). The likelihood of patient \( 0 \) developing condition \( c \) in the future is defined by the equation:

\[
L_{0,c} = P_{0,c} \times S_c
\]

To walk through the collaborative prediction algorithm, we use the dataset in Table I. The conditions of each patient are ordered by the patient’s diagnosis date for the condition. For example, patient \( P_2 \) was first diagnosed with \( C_1 \), then diagnosed with \( C_3 \), and then diagnosed with \( C_7 \) etc. The set of all patients is \( I = \{ P_1, P_2, P_3, P_4, P_5 \} \), and the set of all possible conditions is \( J = \bigcup_{i \in I} J_i = \{ C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8 \} \). When a new patient \( P_0 \) inputs diagnosed conditions \( C_1 \) and \( C_3 \), then \( H_0 = \{ C_1, C_3 \} \). Target \( T_0 \subseteq J - H_0 = \{ C_2, C_4, C_5, C_6, C_7, C_8 \} \).

<table>
<thead>
<tr>
<th>Patient</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_1 )</td>
<td>( C_1, C_2, C_4, C_7 )</td>
</tr>
<tr>
<td>( P_2 )</td>
<td>( C_1, C_3, C_5, C_2 )</td>
</tr>
<tr>
<td>( P_3 )</td>
<td>( C_2, C_3, C_4 )</td>
</tr>
<tr>
<td>( P_4 )</td>
<td>( C_2, C_6, C_7 )</td>
</tr>
<tr>
<td>( P_5 )</td>
<td>( C_2, C_5 )</td>
</tr>
</tbody>
</table>

Algorithm 1 Predicting Ranked List of Potential Medical Conditions of a Patient

**Input:** set of all possible conditions \( J \), set of existing conditions of patient \( 0: J_0 \)

**Output:** a ranked list of potential conditions \( C* \)

begin
initialize a set of target conditions \( T_0 \), which is calculated by the equation: \( T_0 = J - J_0 \);
for each condition \( c \) in \( T_0 \)
generate \( N_c \), which is the set of all other patients with condition \( c \);
for each patient \( i \) in \( N_c \)
calculate \( w(0,i) \), which is the similarity between patient \( 0 \) and patient \( i \);
end for
calculate probability \( (P_{0,c}, \text{support} \ (S_{0,c}), \text{and the likelihood} \ (L_{0,c})) \) of patient \( 0 \) having condition \( c \);
add tuple \( <0, c, L_{0,c}> \) to conditions set \( C* \);
end for
sort \( C* \) by likelihood \( L_{0,c} \);
return

Consider the first condition \( C_2 \) in \( T \), then the following other patients \( N_{c2} = \{ P_1, P_3 \} \) have this condition \( C_2 \). \( w(P_0, P_1) = 1 \) because all conditions are common between \( P_0 \) and \( P_1 \), and \( w(P_0, P_3) = 0 \). Then \( P_{0,c2} = (1+0)/2 = 0.5; S_{c2} = 2/5 = 0.4 \). The tuples for condition \( C_2 \) is therefore \( <0, C_2, 0.5, 0.4> \). Similarly, the tuples for other conditions in \( T \) are \( <0, C_2, 0.5, 0.4>, <0, C_3, 0.25, 0.4>, <0, C_6, 0.5, 0.2>, <0, C_7, 0.5, 0.8> \), and \( <0, C_8, 0.5, 0.4> \). The likelihoods of \( P_0 \) developing conditions are as follows: \( C_2: 0.5*0.4=0.2, C_3: 0.5*0.4=0.2, C_5: 0.25*0.4=0.1, C_6: 0.5*0.2=0.1, C_7: 0.5*0.8=0.4, C_8: 0.5*0.4=0.2 \). Therefore the ranked list of predicted conditions for \( P_0 \) is \( (C_5, C_2, C_6, C_8, \ldots) \).
C C2, C3). The order is not unique as, e.g., likelihoods are the same for C2 and C3, and we can only select top-K conditions.

IV. TRAJECTORY MODEL FOR PREDICTING PROGRESSION OF CONDITIONS

In many situations it is more desirable to predict a medical condition progression trajectory, instead of a single condition. The trajectories stemming from a medical condition can provide a potential set of paths the patient may traverse, as well as explain the likelihood of paths to a final condition for a patient who suffers from an initial condition. We propose a trajectory model to track the progression and infer the most probable future trajectories from a patient’s observed history. A trajectory from a condition c is modeled as a tree T(c) = (N, E) where N = {C1, C2, ... , Cn} is a set of nodes that represent the conditions and E = {e1, e2, ... , en} is a set of edges where each edge e = (Cj, Ck) represents a progression from condition Cj to condition Ck.

There are three steps to generate and make use of the trajectory tree. The first step is to discover edges of conditions from patients’ diagnoses histories as made public in their profiles. The second step is to generate the trajectory model, based on the edges created in the first step. In the last step, the trajectory model is used to infer the confidence value and support of potential progression trajectories given a patient’s diagnosis history. More in detail:

1) Edge Discovery: This step helps identify directional edges of comorbidities, which co-occur for individual patients. A directed edge e is defined as: e = {(Cj, Ck) | A patient was diagnosed with conditions Cj and Ck and Cj preceded Ck in temporal order}. In order to calculate the confidence value and support of a trajectory, the patients with these edges are defined as: I(ej) = {Patients who have the edge ej}.

2) Linking: The generated edges are recursively linked to build the condition trajectory tree T by recognizing the common node (condition) in two edges. For example, given edges ej=(Cj, Ck), ej=(Cj, Cl), and ej=(Ck, Cl) a tree is built with an edge trajectory ej→ek→el resulting in a condition trajectory (Cj→Ck→Cl). Note that we use edges for implementation purpose. For interpretation, the edge trajectory is converted into a condition trajectory.

Algorithm 1 is used for building the edge trajectory tree. The current edge ce = (Cj, Ck) and the new edge ne = (Ck, Ch) of conditions are linkable if Cj = Ck, and ne will not create a cycle in the current path. The trajectory model can be used to infer the confidence value of a medical condition trajectory given a certain observed condition. Given an edge trajectory ti = [e1, e2, e3, ..., en], then Utj (the set of patients who have trajectory ti) is the intersection of the sets of the patients who have the same chain of linkable edges. Formally:

\[ U_{ti} = \bigcap \{ I(e_i) \text{ where } e_i \text{ is an edge in trajectory } t_i \} \]  \hspace{1cm} (4)

3) Inference: We are defining the support of trajectory t (slightly differently from the standard definition) as \(|U_{t1}|\). The confidence value C of the edge trajectory (e1→e2→e3→...→en) given an observed condition c is calculated as a conditional probability, where ej is the starting edge and ej = (null, c).

\[ C(e_1 \rightarrow e_2 \rightarrow e_3, ..., \rightarrow e_n \mid c) = \left| U_{t1} \right| / |I(e_1)| \]  \hspace{1cm} (5)

The comorbidity index (CI) of trajectory (e1→e2→e3,...→en) is defined as follows:

\[ CI(e_1 \rightarrow e_2 \rightarrow e_3, ..., \rightarrow e_n) = \left| U_{t1} \right| / \sum_{\text{conditions}} \ PS(c) \]  \hspace{1cm} (6)

where c is any condition in trajectory t and PS(c) is set of patients with c. The comorbidity index is used to achieve higher scores for trajectories with less-frequent conditions.

To better illustrate the above method, let us consider the example dataset presented in Table I. After we applied the pair generation process on this dataset, 20 edges were generated (sorted by number of patients): Eex = {(C1, C3), (C1, C7), (C2, C7), (C2, C3), (C2, C6), (C3, C6), (C6, Ca), (C6, Cb), (C7, Cb), (C7, Ca), (C7, Cy), (Ca, Cy), (Cb, Cy), (Ca, Cb)}.

Fig. 2. The example of condition trajectories starting from condition C2.

Algorithm 2  Build Condition Trajectory Tree

<table>
<thead>
<tr>
<th>Input:</th>
<th>set of edges E, current depth cd, maximum depth md, current edge ce, path pa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output:</td>
<td>trajectory model</td>
</tr>
<tr>
<td>begin</td>
<td>/* limit trajectories to a certain length*/</td>
</tr>
<tr>
<td>if</td>
<td>current depth cd is equal to maximum depth md</td>
</tr>
<tr>
<td>return</td>
<td></td>
</tr>
<tr>
<td>end if</td>
<td>for each edge ne in edge set E</td>
</tr>
<tr>
<td>/* if two edges can be linked, recursively build tree */</td>
<td></td>
</tr>
<tr>
<td>if</td>
<td>(ne is linkable with current edge ce)</td>
</tr>
<tr>
<td>add ne as a child of current edge ce</td>
<td></td>
</tr>
<tr>
<td>/* path is used for tracking patients of trajectory*/</td>
<td></td>
</tr>
<tr>
<td>append ne to the tail of path pa</td>
<td></td>
</tr>
<tr>
<td>call Algorithm 1 with input E, cd+1, md, ne, pa</td>
<td></td>
</tr>
<tr>
<td>remove ne from the path pa</td>
<td></td>
</tr>
<tr>
<td>end if</td>
<td></td>
</tr>
<tr>
<td>end for</td>
<td></td>
</tr>
<tr>
<td>return</td>
<td></td>
</tr>
</tbody>
</table>

In Fig. 2, the number in parentheses indicates the number of patients having the trajectory from the root to the current node. For example, C2(1) indicates that there is one patient (P1) having trajectory (C1→C2). The confidence of trajectory
(C_2 \rightarrow C_3 \rightarrow C_4) is 2/2 = 1, and the confidence of trajectory (C_2 \rightarrow C_3 \rightarrow C_4 \rightarrow C_5) is 1/2 = 0.5.

V. EVALUATION STUDY

A. Dataset Preprocessing

The collected dataset is from the patients’ self-posted data in PatientsLikeMe website. It contains 17,418 patients’ information, including id, username, gender, age and location and 35,606 diagnoses for these patients. Each diagnosis contains 6 attributes: PatientId, HasCondition, ConditionId, IsPrimaryCondition, FirstSymptomDate, and DiagnosisDate, for example, “ID: 8, HasCondition: Stroke, ConditionId: 48, IsPrimaryCondition: 0, FirstSymptomDate: May 1998, DiagnosisDate: Sep 1998”.

As in all social networks, the data collected is not clean, especially the Date field. Some of the FirstSymptomDate are specified by the user as “?” or blank, and some of DiagnosisDate are marked as “?” or “Undiagnosed”. A gap-based method is developed to preprocess the diagnosis dataset. Date format, such as “?” or “Undiagnosed”, or blank is marked as improper dates. In diagnosis dataset, 17,146 (48%) diagnoses have improper symptom dates, 17,443 (49%) diagnoses have improper diagnosis dates, and 14,775 (41%) diagnoses have both improper symptom date and improper diagnosis date. A gap-based method is developed to fill the missing dates and works as follows. Suppose D_i is all the diagnoses of patient i, and D_i contains four subset D_{i1}, D_{i2}, D_{i3}, D_{i4} mathematically speaking, D_i = U_{j=1}^{4} D_{ij}. For each D_{ij}, D_{ij} is the diagnoses that both symptom date and diagnosis date are proper. D_{i2} is the diagnoses that symptom date is improper and diagnosis date is proper. D_{i3} is the diagnoses that symptom date is proper and diagnosis date is improper. D_{i4} is the diagnoses that both symptom date and diagnosis date are improper. First, D_{i4} is deleted from diagnosis dataset. Then to fill the missing date values of D_{i2} and D_{i4}, the local average gap LAP_i is computed. LAP_i is calculated by the following formula:

\[
LAP_i = \frac{\sum_{g \in G} g}{|G|} \quad (7)
\]

Where G = \{d, diagnosisdate - d, symptomdate | d \in D_{ij}\}, which is the set of time gaps between corresponding pairs of proper diagnosis date and proper symptom date for patient i. Finally LAP_i is used as the offset to fill the missing symptom dates in D_{i2} and the missing diagnosis dates in D_{i3}. After the data cleaning is done, the diagnosis dataset ends up with 20,816 diagnoses, in which both FirstSymptomDate and DiagnosisDate are proper dates.

B. Collaborative Prediction of Medical Condition Incidence

To evaluate the collaborative prediction approach, we used a leave-one-patient-out validation strategy similar to [3]. Recall that head of patient i is H_i and the head size |H_i| is a parameter in our experiments. Only the patients that have |H_i| + 1 conditions are used for validation. Among them, each time one active patient i is taken out and the other patients are used for training. Then the first |H_i| conditions of patient i are fed into the trained model and the remaining |J_i| - |H_i| conditions of patient i are considered as future conditions and used for evaluation. The top-K conditions in the predicted ranked list are considered. We used coverage and rank to evaluate the prediction performance for each patient. The coverage is the proportion of correct future conditions in the top-K ranked list to the total number of correct future conditions. The rank is the average rank of all correct future conditions in the ranked list for this patient. The process is repeated for each patient, and averages of coverages and ranks are computed.

The results are in Table II, where K is the size of the predicted ranked list and the head size is 2. The collaborative prediction model achieves a coverage value of 48% and 75% for top-20 and top-100 ranked lists, respectively.

**TABLE II. CONDITION INCIDENCE PREDICTION RESULTS**

<table>
<thead>
<tr>
<th>Top-K</th>
<th>Average Coverage</th>
<th>Average Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>48%</td>
<td>7.25</td>
</tr>
<tr>
<td>100</td>
<td>75%</td>
<td>21.59</td>
</tr>
<tr>
<td>All</td>
<td>100%</td>
<td>123.29</td>
</tr>
</tbody>
</table>

**TABLE III. EXAMPLE OF PREDICTIONS FOR INDIVIDUAL PATIENTS**

<table>
<thead>
<tr>
<th>Id</th>
<th>Diagnosed Conditions</th>
<th>Top 2 Predicted Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>296</td>
<td>Migraine, Fibromyalgia</td>
<td>Chronic Fatigue Syndrome, Generalized Anxiety Disorder</td>
</tr>
<tr>
<td>42</td>
<td>Eating Disorder, Phobic disorder</td>
<td>Social Anxiety Disorder, Post-Traumatic Stress Disorder</td>
</tr>
<tr>
<td>50</td>
<td>HIV, Seborrheic Dermatitis</td>
<td>Bipolar Disorder, Lactose Intolerance</td>
</tr>
</tbody>
</table>

**TABLE IV. COMORBIDITIES FROM MEDICAL LITERATURE**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Comorbidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major Depressive Disorder (MDD)</td>
<td>Dysthymia, Panic Disorder, Agoraphobia, Social Anxiety, Obsessive–Compulsive Disorder, Generalized Anxiety Disorder, and Post-Traumatic Stress Disorder, Alcohol Dependence, Psychotic Disorder, Antisocial personality, Eating Disorders, Borderline Personality Disorder</td>
</tr>
<tr>
<td>Irritable Bowel Syndrome (IBS)</td>
<td>Major Depression, Anxiety, Somatiform Disorders, Fibromyalgia, Chronic Fatigue Syndrome, Gastroesophageal Reflux Disease, Restless Legs Syndrome</td>
</tr>
<tr>
<td>Eating Disorder (ED)</td>
<td>Obessive–Compulsive Disorder, Bipolar Disorder, Substance Abuse, Diabetes, Bone Disease, Cardiac Complications, Gastrointestinal Distress</td>
</tr>
<tr>
<td>Obesity</td>
<td>Type 2 Diabetes Mellitus, Hypertension, Dyslipidemia, Cardiovascular Disease, Stroke, Sleep Apnea, Gallbladder Disease, Hyperuricemia And Gout, Osteoarthritis, IBS, Sleep Apnea Disorder</td>
</tr>
<tr>
<td>Bipolar I</td>
<td>Substance Abuse, Generalized Anxiety Disorder, Simple Phobia, Social Phobia, Obsessive–Compulsive Disorder, PTSD, Panic Disorder</td>
</tr>
<tr>
<td>Migraine</td>
<td>Stroke, Sub-Clinical Vascular Brain Lesions, Coronary Heart Disease, Hypertension, Psychiatric Diseases, RLS, Obesity, Epilepsy, Asthma, Irritable Bowel Disease, Chronic Fatigue Syndrome, Fibromyalgia</td>
</tr>
</tbody>
</table>

These results have better coverage (7% and 15% increase) when compared with the results reported by Davis et al. [3].

The dataset was collected in the year 2012 and the data was processed right after.
who used EHR data. Our results show that the collaborative prediction model is able to make good predictions based on patients’ social media data. Table III shows examples of predictions.

C. Generating Comorbid Disease Trajectories

To the best of our knowledge, there is no study to directly compare our results with. Thus we show the medical condition progression trajectories generated by our tree-based model and compare the trajectory results with existing comorbidities reported in literature. We selected six conditions, namely “Major Depressive Disorder” [17, 18], “Migraine” [19], “Irritable Bowel Syndrome (IBS)” [20-22], “Eating Disorder” [23], “Obesity” [24-26] and “Bipolar I” [27, 28]. These conditions and their comorbidities are listed in Table IV. We chose each of these six conditions as the “starting condition” (i.e., a root) and generated the trajectory tree model (Algorithm 1) by setting the trajectory’s minimum support to 5. The trajectories (Table V) are first ranked in terms of their length. Within the same length, the top-2 trajectories in terms of the comorbidity index are shown.

The predicted trajectories cover most of the comorbidities in the literature. More importantly, different from previous research [6, 7], which predicts incidence or visualizes temporal trajectory patterns, our tree-based model predicts the confidence of the future trajectory and reveals every possible path between any two medical conditions. For example, for Eating Disorder (ED) and Panic Disorder (PD), there are at least two paths progressing from ED to PD (ED→Tobacco Addiction→Drug Addiction→PD and ED→Obsessive-Compulsive Disorder→PD). It means that Eating Disorder is possible to progress to Panic Disorder either via Tobacco and Drug Addiction or via Obsessive-Compulsive Disorder. This functionality can help doctors and patients better understand the potential trajectories between any medical conditions.

D. Comorbidity Progression Trajectory Analysis

To illustrate how the trajectories can be used to help doctors reveal the progression paths of medical conditions, we performed a case study on the progression trajectory starting with “Major Depressive Disorder” (MDD). The numbers in parenthesis on each node indicate the numbers of patients following the trajectory from the root to the current node, e.g., there are 17 patients with the trajectory (MDD->IBS). In Fig. 3, the most frequent length-2 trajectories are (MDD->GAD) (165 patients) and (MDD->Fibromyalgia) (127 patients). The most frequent length-3 trajectory is (MDD->GAD->PD); (PD=Panic Disorder). The confidence value of (MDD->GAD->Panic Disorder) given the observed condition MDD is 37/680 = 5.4%. The other length-3 trajectories between MDD and PD are (MDD->Dysthymia-

>PD) (3.4%), MDD->PTSD->PD) (3.2%), and MDD->Social Anxiety Disorder->PD) (2.5%).

One possible explanation of this result is that Bouchard et al. [29] found that in young adults with low levels of lead exposure, higher blood lead levels were associated with increased risks of MDD and Panic Disorder, which confirmed the comorbidity of MDD and PD.

However, it is not known exactly which paths MDD patients go through to get PD. The trajectory model reveals possible intermediate nodes between these two conditions. Thus, MDD can progress to PD via GAD, Dysthymia, PTSD, or Social Anxiety Disorder. The likelihood going through GAD is higher than for other paths.
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2564299, IEEE Transactions on NanoBioscience

E. Discussion on Gender-specific Medical Incidence and Trajectories

In this section, we discuss stratifying the medical incidence and trajectories based on patients' gender to investigate the possible gender-specific disease incidence and progression trajectories. To compare the difference of medical incidence in terms of gender, we categorized 174 medical conditions into 18 categories. Fig. 5 shows the distributions of male and female patients. We found that most of the categories have similar ratios as the overall ratio (female/male=8023/3932=2.04) except for five categories: “muscle, bone, joint,” “digestive, intestinal,” “lungs, respiratory,” “women’s health” and “men’s health.” These five categories have gender ratios of 11.05, 5.62, 3.27, 27.09 and 0 respectively. For “muscle, bone, joint,” the reason is that most of the patients in this category have the condition “Fibromyalgia”; 90% of Fibromyalgia patients are female, which aligns well with our gender ratio. Most patients in “digestive, intestinal” have IBS (Irritable bowel syndrome) and the ratio of 5.62 is slightly higher than the ratio of 2 reported by Mayo Clinic. The significant gender difference for “women’s health” and “men’s health” is notable, and we found out that 13 male patients suffer from Postpartum Depression (sic!), a clinical depression after child birth and a “women’s health” condition.

For the progression trajectories, we run trajectory model separately on 8,023 female patients and on 3,932 male patients by specifying each condition as the root. The preliminary results show that the trajectory trees show significant
difference in terms of the size and progression courses across gender for most of the conditions. The male and female trajectory tree starting from Obsessive-Compulsory Disorder (OCD) is shown in Fig. 4. Male and female patients show the many identical trajectories (e.g., OCD->Generalized Anxiety Disorder (GAD)->Major Depressive Disorder (MDD), OCD->GAD->Social Anxiety Disorder (SAD), OCD->MDD->Panic Disorder, and OCD->MDD->PTSD). One exception is that the male patients show the trajectory of OCD->Dysthymia->SAD, which is not found in female patients. To validate this observation, we searched the related medical articles. Assuncao et al. found out that a third of OCD patients has social phobia (SAD), which was significantly associated with male gender, dysthymia, and generalized anxiety disorder (GAD). This study could possibly explain that why only male patients show the trajectory of OCD->Dysthymia->SAD. More experiments need to be carried out to systematically compare the gender-specific condition trajectories in the future.

Fig. 5. The number of (a) male and (b) female patients in each medical condition category.
F. Limitations

One limitation for this research is a data quality issue. For our dataset, before the April 2011, a given patient could only have one condition, or pick from a small cluster of conditions on the social network website. For instance, the patient suffering from a mood disorder can only select a list of other mood disorders they suffer as the same time, so they often co-occur. As a result, the site’s population is largely skewed towards people with conditions such as mood disorders and fibromyalgia. To mitigate this issue, we plan to select a subset of online patients who joined the site after April 2011, whose posted diagnosis could be a better representative of their comorbidities.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a collaborative comorbidity prediction and a comorbidity trajectory graph model to predict risks of medical condition incidence and trajectories using patients’ social media data. Different from other research, we only used publicly available patient-reported medical data, which we call SHR (Social Health Records) in contrast with EHR or PHR. A prediction model based on collaborative filtering (CF) is presented to predict a ranked list of future condition incidences. In addition, a trajectory prediction model and algorithm are presented for disease progression trajectories from a starting condition. The experimental results show that the collaborative prediction model for a condition incidence predicts future conditions with coverage of 48% (top-20) and 75% (top-100). The trajectory model reveals each possible progression trajectory for any two conditions. The top-ranked trajectories automatically discovered comorbidities that were validated by the medical literature. We also discussed the difference of trajectory results across patients’ gender.

Future work includes 1) Improve the tree-based trajectory model. Currently, the trajectories that have highly “popular” conditions (e.g., Fibromyalgia) tend to be ranked high in terms of confidence, because more conditions are paired with “popular” conditions. It is also possible to filter trajectories with low confidence levels. 2) Systematically evaluate the trajectory prediction model. The current evaluation is based on the reported comorbidities. More experiments for evaluating the quality of predicted trajectories will be designed and performed. Also, the trajectory results in different stratifications of patients (e.g., by gender or by age) will be compared to investigate whether there are any significant gender or age-specific differences in trajectories. 3) The progression trajectory can be regarded as a generic version of a collaborative prediction model, as it not only predicts the potential incidence but also the intermediate conditions. We plan to combine the two models into one single model and to utilize the time sequence feature in the future.

ACKNOWLEDGEMENT

This research has been partially funded by the PSC-CUNY Research Foundation under the awards 42-64266 and 43-65232, and partially conducted through NJIT sponsored by the Leir Charitable Foundations.

REFERENCES


