

Social Health Records: Gaining Insights into Public Health Behaviors, Emotions, and Disease Trajectories

Soon Ae Chun, James Geller and Xiang Ji

Abstract Social media and personal health monitoring devices (e.g., Fitbit) provide abundant patient-generated health-related data. These open health data, generated via patient engagement and sharing, are referred to as *Social Health Records (SHR)* as opposed to the EHR (Electronic Health Records) that are created and entered by clinicians. SHRs are changing the healthcare paradigm from the authoritative provider-centric model to a collaborative and patient-oriented healthcare framework. This chapter proposes an *SHR Integration and Analytics Framework* to leverage Social Health Records for gaining insights into population-level and individual-level healthcare practices and behaviors, as well as emotions. The framework defines a pipeline for generating knowledge from the social health data sources to the end users, including the patients themselves, public health officials, and healthcare providers. The SHR integration and analytics framework build a coherent knowledge base, linking the Social Health Records that are “spilled” in distributed online social media, with other online health information sources, such as results from authoritative medical research. The semantic integration model of heterogeneous health data sources provides population-level health analytics and reasoning capabilities to gain intelligence on public healthcare issues and practices. The SHR is shown to be a valuable resource for epidemic surveillance systems with real-time monitoring. We focus on an approach to quantifying the SHR-based public emotions for measuring health concern levels and for tracking them, and propose SHR-based predictive models to infer individual-level and population-level comorbidity predictions and comorbidity progression trajectories.

S.A. Chun (✉)

City University of New York, New York, NY, USA
e-mail: Soon.Chun@csi.cuny.edu

J. Geller

New Jersey Institute of Technology, Newark, NJ, USA
e-mail: james.geller@njit.edu

X. Ji

The Bloomberg L.P., New York, NY, USA

Keywords Social Health Records (SHR) • Social media analytics • Social media content mining • Public health monitoring • Semantic integration • Social health knowledge base • Linked health data

1 Introduction

There is a large amount of health information available for any patient to address his/her health concerns. The freely available health datasets include open government health datasets, at the national, state or community level, such as OpenHealthdata.gov ranging from Medicare data to epidemiology; Web health resources curated by experts such as WebMD; and the personal health records shared by the patients on open or registered online social media services such as PatientsLikeMe. These are so-called *open health data*, which are readily accessible and downloadable. The patient-generated and shared data include the conditions, treatments, side effects, health histories, and personal physical, psychological, emotional and relationship experiences of individual patients. This data resembles the Electronic Health Record (EHR), which is defined as an electronic version of a patient's medical history that is collected and maintained by the provider (e.g., clinicians) over time.

The EHR system allows capturing the key administrative and clinical data relevant to that person's care, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports. Since the open online health records shared by patients or family care givers capture similar data about the patients, we call this *Social Health Record (SHR)* to distinguish from the closed EHR. Some of the key characteristics of EHR and SHR are shown in Table 1.

The SHR is capturing many instances of personal healthcare experiences, practices and other health-related behaviors, while the EHR is capturing the clinical data necessary to provide care. Even though a doctor prescribes a medicine X, the patient may consume a substitute medicine Y. The intention of sharing the Social Health Records is support-oriented with information and experience sharing, while the EHR is primarily care-oriented to address the conditions. The SHR expresses emotional and psychological attitudes, opinions and comments in ordinary language riddled with ambiguities, while the EHR may capture mostly the factual statements in expert language to avoid vagueness or ambiguities. Another difference is that the EHR is hard to share, protected by the HIPAA and HITECH regulations and locked into different EHR systems. This creates a silo effect and causes difficulty for interoperability and sharing EHRs. On the other hand, the SHR is an open system based on online services, so the data is easily shared over lightweight clients, e.g., a Web browser.

The EHR focuses on individual patients, and it is difficult to connect and aggregate EHRs of many patients unless one has all the access privileges. On the other hand, the SHRs are inherently crowdsourced data due to their base in social

Table 1 Characteristics of EHR and SHR

| EHR | SHR |
|---|--|
| Generated by clinicians or medical experts | Self-reported by patients, public, government |
| Clinical Data: Diagnoses, prescribed medications, allergies, problems, procedures, chart notes, clinical alert notes, lab results, and images | Experience and Behavior Data: <ul style="list-style-type: none"> • Health status reports <ul style="list-style-type: none"> – Experienced symptoms, side effects – Diagnosis reports • Healthcare practice data <ul style="list-style-type: none"> – Actual medications, treatments • Health-related behaviors/habits <ul style="list-style-type: none"> – Drinking, smoking, exercises, etc. – Nutritional data |
| Factual statements | Statements on emotional, psychological attitudes, comments, opinions |
| Uses medical expert language e.g., Myocardial infarction | Informal everyday language e.g., Heart attack |
| Comparatively unambiguous e.g., ICD9 code for a disease | Ambiguous or vague e.g., Diabetes (type 1 or 2?) Hepatitis (A or C?) |
| Closed data <ul style="list-style-type: none"> – Difficult to share patient data (e.g., due to HIPAA, HITECH regulations) – Often locked in siloed systems | Open data <ul style="list-style-type: none"> – Membership based sharing – Open sharing – Open system based on Web browsers |
| Care-oriented | Support-oriented |
| Individual records | Crowdsourced data |

media so they can reveal the aggregated information of the crowd. For instance, the forum entry SHR data in one community group (e.g., cancer patient groups) from many patients may easily reveal the major types of issues and popular treatment options for many patients. The SHRs can provide a unique opportunity to look into health care from the patients' perspectives to identify healthcare-related issues and improve the quality of care. The SHR data from the crowd can facilitate the ability to "connect the dots" among and across many patients and allow gaining public health intelligence and insights, such as detecting disease outbreaks and understanding population-related health trends. The crowdsourced SHRs can be a great asset for public health intelligence. Some examples of potential healthcare benefits of aggregating and mining SHRs include the following:

- determine which health topics are of greatest current concern
- identify a high-risk group of patients
- identify health trends both in the general public and at the individual level
- identify how patients view or feel about particular treatments and practices
- track adverse drug events
- identify the perceived quality of healthcare services, e.g., most desirable outcomes
- create education campaigns and interventions

- offer insights into the relationship between an individual's health and their everyday lifestyles
- reveal patients' attitudes toward health.

These datasets can help to assess and improve healthcare quality, as well as help to modify health-related policies. There are also patient-generated datasets, accessible through social media. Clinicians and healthcare providers may benefit from being aware of national health trends and individual healthcare experiences that are relevant to their current patients. The available open health datasets vary from structured to highly unstructured. Due to this variability, an information seeker has to spend time visiting many, possibly irrelevant, Websites and has to select information from each and integrate it into a coherent mental model.

In this chapter, we discuss an approach to integrating these openly available but widely dispersed health data sources, where health data is created and shared by patients voluntarily, and open knowledge and expertise shared by healthcare providers and professionals. The goal of developing the integrated data sources is to provide answers to information and knowledge needs of end users, to provide insights on public health through diverse analytics on social behaviors, and behavior models learned from the social data to predict trends. The insights are presented to convey an intuitive understanding of the public health trends and alerts for physicians, healthcare staff, health policy workers, and individual patients.

Our approach to integrating diverse open health data sources is through Linked Data principles and Semantic Web technologies. In Sect. 2, we present a brief summary of related works, and in Sects. 3 and 4, we provide the data collection and our approach on how to construct a linked data model which is then used as the basis for developing a set of analytics. In Sect. 5, we present the architecture for our social health data analytics platform and Sect. 6 describes the prototype system and analytics tools. The analytics tools include "Social InfoButtons," which provides awareness of both community and patient health issues, the population health concern trend analyses using machine learning of sentiment classification, and a comorbidity trajectory analysis using tree analysis that may shed light on the population health trends, but also may be useful in predicting a specific patient disease progression. The proposed social health analytics platform provides patients, public health officials, and healthcare specialists with a unified view of health-related information from both official scientific sources and social networks, and provides the capability of exploring the current data along multiple dimensions, such as time and geographical location.

2 Related Works

Integrating data from the Social Web is a challenging task that requires information extraction and data integration. Research works [1–3] extract health information from different sources including the web and social media, sensors, healthcare

claims and lab images, and physician notes that provide useful health information. There are still notable differences between professional experts and Web health users. Smith and Wicks [4] found that only 43% of the symptom terms (e.g., PatientsLikeMe) are present in the Unified Medical Language System Metathesaurus (UMLS). Their study reaffirmed the challenges of reconciling the differences between unfettered natural language descriptions and restricted terminologies as well as formalized knowledge sources.

For the data integration, the Semantic Web has been used as a framework for data integration, e.g., Linked Open Data (LOD) [5, 6], to create links between resources distributed in heterogeneous data sources. LOD principles require using URIs to identify resources, RDFs to represent information, and typically use of SPARQL to access the information. Many research works use the Semantic Web for data integration in various fields, e.g., geospatial data integration [7], folksonomies in a social tagging system with an ontology [8] and the fields of solar physics, space physics, and solar terrestrial physics [9]. In health informatics, a semantic integration model of different health data sources is used for annotating social health blogs [10]; a clinical trial knowledge repository is constructed integrating data from clinical trials and from side effect information [11]; and in [12], clinical trial data is integrated with drug data to support end users in finding an appropriate clinical trial for them to participate in.

Even though there are many sentiment analyses of Tweets in general area [13, 14] and in the health domain [15] using data mining and machine learning approach, most of works do not apply the results of the sentiment analysis to measure the degree of public concerns or anxiety toward disease, as an emotional health indicator as we propose.

Data mining and machine learning techniques are used to predict disease risks for individuals or to rank diseases by their risks. For instance, in [16, 17], a condition for one patient is predicted using similar patients, based on 13+ million elderly patients' hospital visit records. In Hassan and Syed [18], collaborative filtering (CF) is used for predicting cardiac death and recurrent myocardial infarction, based on demographics, comorbidity, lab test results, and outcomes from a real-world dataset containing 4557 patients' records. Other studies include the k-means algorithm to cluster patients and applied association rule analysis to predict disease for patients in each cluster [19], the patient risk prediction study in the context of active learning with relative similarities in [20], and the Chronic Disease Recommender System to suggest medical advice and diagnoses to patients [21].

Another set of research attempts to reveal and infer condition progression trajectories. The study in [22] investigated the temporal trajectory patterns of all diseases for the entire country of Denmark, using the Markov Cluster algorithm to identify the five largest clusters of disease trajectories, while a disease progression model is based on a Bipartite Bayesian Network [23] to identify a few comorbidities and infer the progression trajectory and comorbidity onset of individual patients. A number of disease progression models such as path models, oncogenetic

tree models, distance-based trees, directed acyclic graph model, etc. are reported in [24]. The progression model for comorbidities we conduct is intended to identify the population-level comorbidity trajectories using large datasets of social media source, using a lightweight tree-based model [25, 26].

3 Multiple Data Sources

The users of public health decision support may include community-based health providers, local, state, and federal government officials as well as the patients. Table 2 summarizes a few typical public health-related questions that these end users may pose to gain public health intelligence for their health decisions.

The content coverage of each individual data source is disparate and not sufficient to address the public health intelligence-related questions shown in Table 2. Table 3 shows the counts of datasets and the content types covered by each individual data source.

PatientsLikeMe, which is a medical, patient-centric, social network, provides patients’ personal and medical data and tracks the patients’ interactions with their associated conditions, treatments, and symptoms. MedHelp is a platform that hosts

Table 2 SHR-based public health intelligence types

| Category | Questions |
|---------------------|---|
| Statistics | <ul style="list-style-type: none">• What are the top conditions with the most patients?• How many patients are suffering from the condition X?• What are the most frequently cited symptoms of the condition X?• What is the percentage distribution of treatment options for the condition X?• Was the public health policy well received (e.g., positive responses) among population groups? |
| Demographics | <ul style="list-style-type: none">• Who are the patients suffering from this condition X?• What is the gender distribution of the patients with X? |
| Geospatial Analysis | <ul style="list-style-type: none">• How are patients with condition X distributed at the state/country level?• What is the average distance to travel for the patients to a treatment facility? |
| Correlation | <ul style="list-style-type: none">• Does gender play a role in choosing treatment options for a condition X?• Is there any difference in treatment options reported in social and official data sources?• Is there any comorbid relationship between two conditions or multiple conditions? |
| Trends | <ul style="list-style-type: none">• What are the changes in the number of cancer type X patients in a community over time by gender?• How did the popular treatments for a condition X change over time, and by location?• How is the disease X spreading in time and by location?• What are the citizens’ anxiety level changes over time?• Is our community health improving in terms of morbidity and mortality? |

Table 3 Content types in data sources

| Data Source | Patient | Condition | Treatment | Symptom | Review | Community | Post | Prevalence |
|----------------|---------|-----------|-----------|---------|--------|-----------|--------|------------|
| PatientsLikeMe | 17,407 | 1228 | 5608 | 2176 | n/a | n/a | n/a | n/a |
| MedHelp | n/a | n/a | n/a | n/a | n/a | 365 | 69,243 | n/a |
| WebMD | n/a | 647 | 180 | n/a | 86,715 | n/a | n/a | n/a |
| Mayo Clinic | n/a | 1116 | 2496 | 5426 | n/a | n/a | n/a | n/a |
| CDC | n/a | n/a | n/a | n/a | n/a | n/a | n/a | 52 |

discussion boards (e.g., forums) among patients and health professionals on various aspects of each specific condition or its category. WebMD is an online service providing information about drugs along with users' reviews of each drug, in addition to condition types and typical treatments. The CDC provides statistics of statewide prevalence of diseases. PubMed serves as a repository of comprehensive data on the medical and clinical scientific literature. In many cases, complete publications are accessible. Twitter is a real-time microblogging platform that can be used to monitor disease outbreaks [27] and disease sentiment trends [3], although it is not healthcare-specific. Among the information provided by Twitter, there are user posts, physical locations, and topics.

In summary, one data source alone may not answer the public health-related questions as shown in Table 2, because each source may cover some content but does not support cross-content queries often needed to gain public health intelligence. In other words, public health intelligence requires, as many applications do, integrated data from disparate data sources, to provide value for different communities and users concerned with questions about public health statistics, trends, correlations, and distributions.

To develop the social health knowledge graph, publicly available data was extracted from *PatientsLikeMe*¹, *PubMed*, *WebMD*, the CDC website, and the UMLS *Metathesaurus*. The PHP HTML DOM Parser [28] was utilized to extract relevant information from the above websites. The retrieved structured data was stored in a Jena triple store. To extract the relevant PubMed documents about conditions, we searched PubMed using 1228 condition names collected from the PatientsLikeMe list of conditions. For each condition, the available information such as PubMed URL, title, author, and conference/journal of the top 20 matched documents were collected and stored in the Jena triple store. WebMD resources were retrieved in a similar fashion. Furthermore, the CDC BRFSS prevalence data was also collected through scraping, since that data is published in tables, such as the prevalence data of Asthma in 2010 [29].

4 Social Health Knowledge Graph

In order to query any individual social health-related record or to gain public health intelligence, we developed a social health knowledge graph, which serves as integrated knowledge base consisting of health records, extracted from multiple user-generated health contents on their social media data sources and data and expertise (knowledge) from other open health data sources. We use a lightweight ontology that contains the health record-related concepts and relationships, which serves as a semantic schema for integration. Figure 1 shows a snippet of the ontology.

The *Condition*, *Treatment*, and *Patient* classes are the central concepts in the semantic model. The *Condition* class has the “*isDiagnosedTo*” relationship to the *Patient* class and has the “*exhibit*” relationship to the *Symptom* class, and the

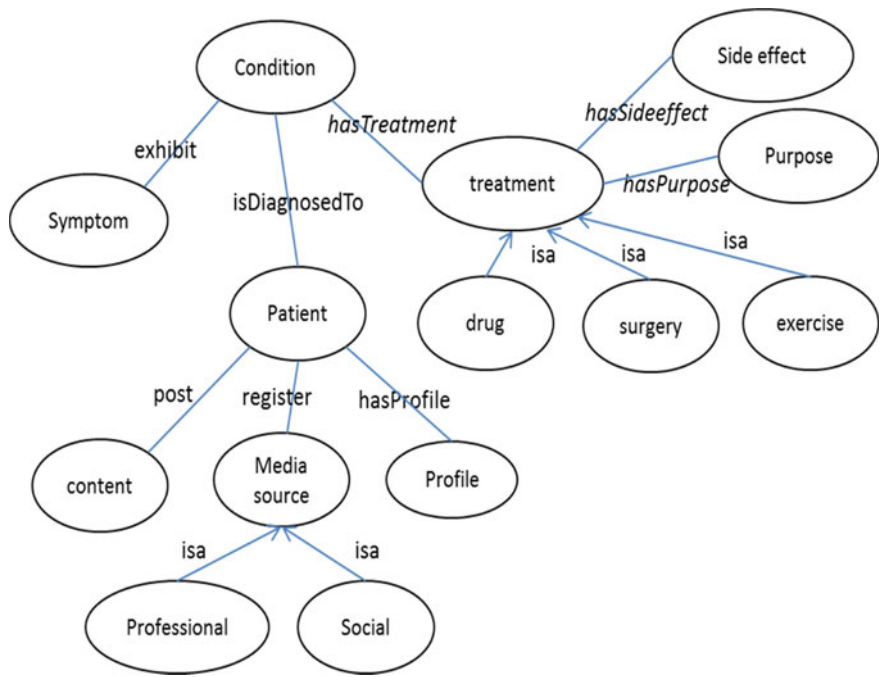


Fig. 1 Social Health Record ontology for semantic integration

“*hasTreatment*” relationship to the *Treatment* class. The *Treatment* class has the “*hasSideEffect*” relationship to the *SideEffect* class and the “*hasPurpose*” relationship to the *Purpose* class. The *Treatment* class also has subclasses indicating different categories of treatments, including *Procedure*, *Exercise*, *Drug*, *Surgery*, etc. The *Drug* class and *Therapy* class also have their own subclasses.

The concepts in the ontology are used to recognize and extract the entities, and the relationships defined in the ontology between concepts help relate the recognized entities.

4.1 Social Health Record Model

Each health record is modeled as a Linked Data assertion represented as a triple $\langle \text{subject}, \text{predicate}, \text{object} \rangle$, denoting the atomic knowledge unit which states that the “subject” entity is related to the “object” entity by the “predicate” relationship. The subject or object represents a class in ontologies, and a predicate is a property of a class or between classes which states the relationship in existence between two entities. To instantiate the health record model, we extracted the health-related concepts with their URIs and represented them as triples.

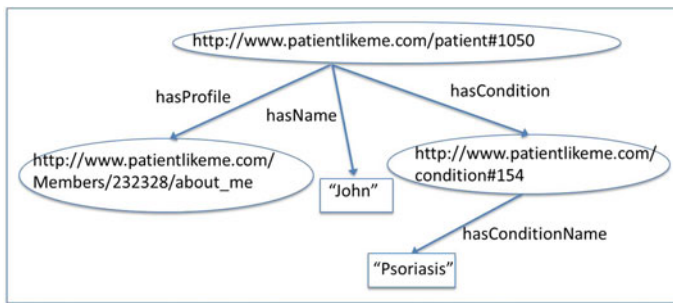


Fig. 2 Example of linked triples

For example, the URI₁ at <http://www.patientlikeme.com/patient#1050> describes the patient named “John” and his profile is described at the URI₂, http://www.patientlikeme.com/members/232328/about_me, and he has the “*Psoriasis*” condition described in URI₃ <http://www.patientlikeme.com/condition#154>. This information is represented as triples $\langle \text{URI}_1, \text{hasName}, \text{John} \rangle$, $\langle \text{URI}_1, \text{hasProfile}, \text{URI}_2 \rangle$, and $\langle \text{URI}_1, \text{hasCondition}, \text{URI}_3 \rangle$. A group of such triples can be used to describe the patient. The triples corresponding to statements about the patient “John” are shown in Fig. 2.

In order to integrate disparate data sources, entity resolution is used to recognize the terms from different resources that actually represent the same concept. For instance, consider a term for a condition extracted from PatientsLikeMe and another condition retrieved from the CDC website [30]. The PatientsLikeMe condition is referred to “Human immunodeficiency virus”, while the CDC refers to it as “HIV”. A knowledgeable human can identify these two terms as both referring to the same concept, but for computers, it is harder to capture the underlying identity, especially when two names do not have any literal similarity. For example, “ALS” and “Lou Gehrig’s Disease” are two different names but they refer to the same concept.

In general, the problem described above is called the entity consolidation/resolution or entity disambiguation problem. Rao et al. [31] reviewed the common approaches to entity disambiguation. For entity consolidation in linked open data, Hogan et al. [32] developed a method to use explicit owl:sameAs relations to perform consolidation. In the domain of medical informatics, Hassanzadeh et al. [33] reported on the LinkedCT project, which utilized exact match, string match, and semantic match to discover links between clinical trial entities, such as trials, conditions, interventions, primary outcomes, etc.

As in previous work by Chun and MacKellar [34], the UMLS [35], which contains the Metathesaurus of medical concepts, is used in this research to provide a common vocabulary and semantics for multiple terms that refer to the same concept. Ji et al. [36] developed a term matching algorithm by using the UMLS to recognize identical concepts. CUIs, which are concept unique identifiers for medical concepts in the UMLS, are used by the algorithm to identify the same concept with different terms. To discover the “sameAs” links, we apply two rules: (i) If two

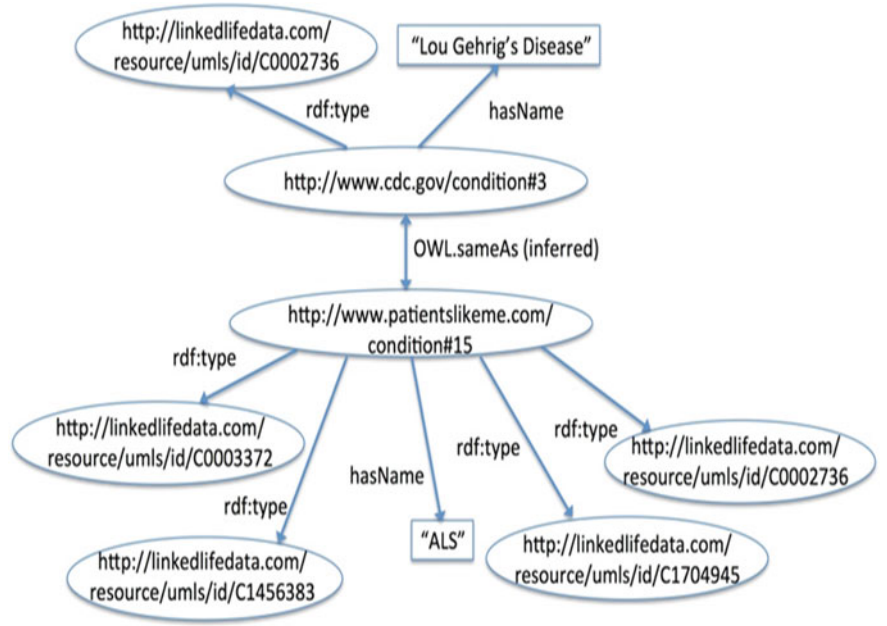


Fig. 3 Entity linking across data sources

conditions in two datasets are of the same name (after necessary stemming and preprocessing), they are regarded as the same concept, and a linkage between the two conditions is discovered; and (ii) When the same CUI is associated with two different condition terms from different datasets, the sameAs link is inferred between the two terms because each concept in the UMLS is uniquely identified by a CUI.

Figure 3 shows that “ALS” from PatientsLikeMe and “Lou Gehrig’s Disease” in the UMLS are identified as the same entity, after the sameAs link has been inferred.

5 Architecture of Social Health Analytics Platform

To enable end users like health officials or epidemiologists to draw public health intelligence to better understand the population’s health status or to get data-driven insights into the social health behaviors, the *social health analytics platform* is proposed. Figure 4 shows the major components consisting of data extraction, linking, and discovering additional links through inference to construct an integrated connected knowledge graph, and the analytics component where the machine learning component builds the models to automate the data processing to not only summarize, but also to predict sentiments, and diseases that may be correlated with

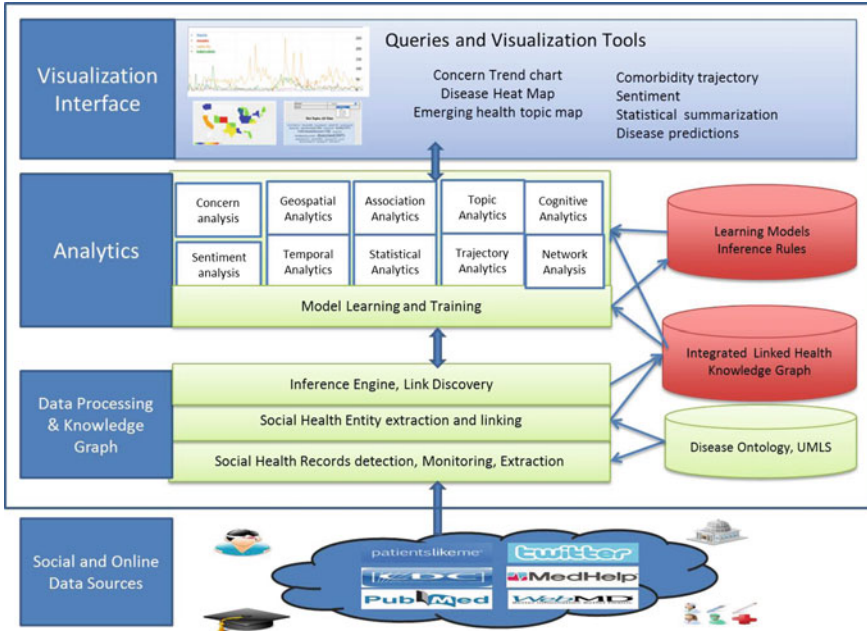


Fig. 4 Social health analytics platform architecture

other diseases. This system is intended to provide answers to various health-related questions as shown in Table 3.

We have implemented a prototype system. The data processing layer at the bottom layer of the architecture is responsible for monitoring the social data sources and extracting and ingesting the data into a staging database. The layer is composed of multiple connectors, one for each type of data source, through APIs or specialized extraction connectors to retrieve data from heterogeneous sources. Among others, we have a Web crawler that uses the PHP HTML DOM Parser to scrape Websites and to retrieve relevant information. Additional connectors can be developed as needed. Data sources currently accessed in our extraction routine include the social network site PatientsLikeMe and Twitter (through APIs), the health forum MedHelp, the government-maintained CDC site, the Mayo Clinic Website, the PubMed Website, and the patient resource portal WebMD. The incoming data, where applicable, goes through the geo-coding processor, where text-based location information is resolved to latitude and longitude coordinates (geo-coding) and, vice versa, coordinates are resolved to names of places (reverse geo-coding) by using third-party services. Geo-coding is required to enable geospatial analytics and to chart data on maps.

Data is then stored in RDF format in the Jena triple store [31] with 612,017 triples representing entities from different sources above mentioned and their relationships. From here, data is linked and augmented via the inference engine

component. The latter makes use of supplemental information specified in the UMLS, inference rules repositories, as well as of an entity resolution and a reasoning service. The inference engine is the place where data linkage is performed and additional facts are derived, thus enabling cross-dataset exploration and reasoning about data. Both the inference engine and the triple repository can be accessed via the analytics layer, which is why the analytics are deployed. At the higher level, users interact with the system via visualizations or the system interface, which invokes analytics operations according to the user's input.

6 Social Public Health Analytics

In this section, we provide a few analytical applications using the social health knowledge graph and SHR to illustrate how Social Health Records are used to provide public health intelligence.

6.1 *Social InfoButtons*

The integrated triple store of social health data can support the basic queries using SPARQL [37]. In addition, to provide answers to the basic queries about public health, the knowledge graph is exploited for knowledge navigation to answer various complex health questions listed in Table 3. It can be used to answer questions such as “What are the top diseases reported by other patients?” or “How many male patients with Asthma are in the state of New Jersey?”

Using these basic capabilities for question answering, we built a Social InfoButtons similar to InfoButtons [38] to provide social health information delivered in a context-aware fashion, e.g., in the clinical patient care context, in the government policy evaluation context, and in the personal information look-up context. Cimino et al. [39, 40] developed InfoButtons to complement the existing Electronic Health Records (EHR) systems and meet the clinicians' information needs in the context of patient care. Cimino et al. [41] described different information needs, their contexts, their resources, and the corresponding applicable methods. In Social InfoButtons, we implemented similar functionalities to provide context-aware information, but the information of Social InfoButtons covers patients' social health information at an aggregated level. This aggregated information includes the percentage of treatments or symptoms for a given disease self-reported by the patients, which can help clinicians to understand the context-specific disease and care patterns or trends from other similar patients at the point of care.

The Social InfoButtons system displays the current disease trends as a list of most common diseases, based on the statistics of the accessed social network sites, as shown in Fig. 5. It also provides disease-specific trends among patients, such as favorite drugs, symptoms, demographics, and geographical distribution of the

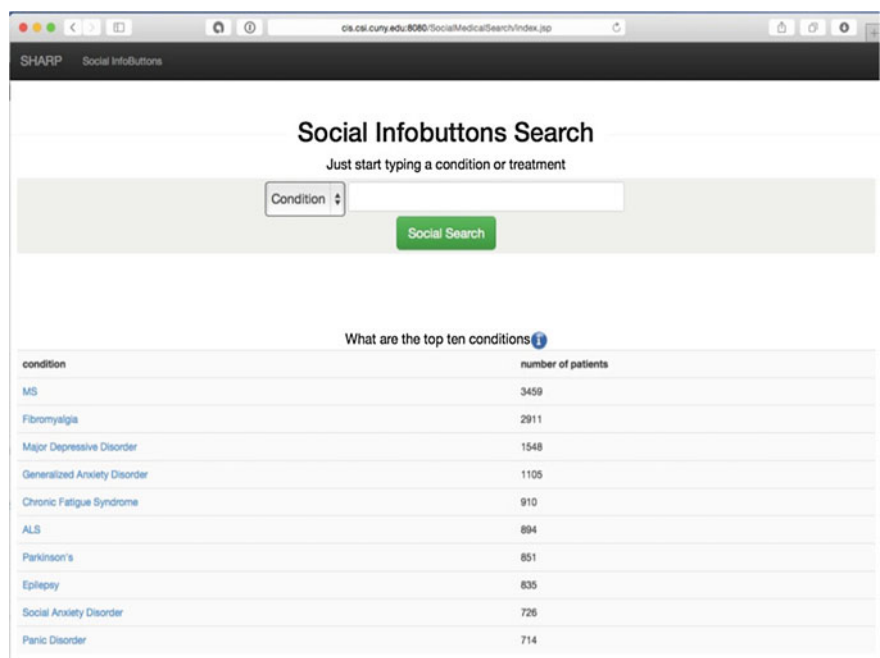


Fig. 5 Social InfoButtons search to provide social health behaviors

patients. The visualization of these social data is juxtaposed with open government data statistics or cutting edge research information from PubMed and WebMD, to allow comparative views.

The social information might be useful for clinicians as it provides them with a new perspective on the current condition/symptoms they are encountering. It is also helpful for patients because when patients are faced with a health concern, they usually want to know how similar patients are coping with the same concern, and how quickly they are recovering.

When a doctor is caring for a veteran who suffers from PTSD (Post-traumatic Stress Disorder), he can practice evidence-based medicine and explore the social trends and experiences of other patients like his patients. As shown in Fig. 6, the Social InfoButtons system can provide answers to typical questions that might be asked by the clinician, represented as information button icons.

This way, the nontechnical person who is not familiar with the SPARQL query language can query the knowledge graph to quickly get the desired answers. For example, the InfoButton icon next to “where is the individual patient?” can provide the doctor with a map (shown in Fig. 7) to indicate the location of all the patients who posted that s/he is suffering from PTSD.

The doctor can look at the total number of patients by region and is able to zoom in on the map for each patient level to view their profile information such as username, social network profile page, gender, age, and location. The treatments

Information of the patients with the condition: PTSD

How many patients? 73

who are these patients?

how are the patients distributed in state and country level?

where is the individual patient?

what is the patients' gender distribution?

Treatments of the condition: PTSD

Individual Therapy (Psychotherapy) [pubmed](#) [webmed](#)
 Evaluated By **95** patients
Side Effects:

Sertraline (Prescription Drug) [pubmed](#) [webmed](#)
 Evaluated By **25** patients
Side Effects:

(1) Weight gain 21%;
 (2) Dry mouth (xerostomia) 20%;
 (3) Loss of sex drive (libido) 18%;
 (4) Fatigue 17%;
 (5) Insomnia 11%;
 (6) Emotional withdrawal 11%;

Citalopram (Prescription Drug) [pubmed](#) [webmed](#)
 Evaluated By **22** patients
Side Effects:

(1) Fatigue 25%;
 (2) Sex drive (libido) decreased 17%;
 (3) Brain fog 15%;
 (4) Anxious mood 14%;
 (5) Weight gain 13%;
 (6) Dizziness 12%;

Fig. 6 PTSD-related Social InfoButtons

used by other similar patients for PTSD and how they reacted to them are also displayed to the doctor to make better informed decisions. For example, if his patient is from a particular location, the doctor can find out common characteristics of all patients in the close-by region, such as any common profile information, notable common symptoms, and treatments reported by other patients. The doctor can make a better recommendation on a treatment regimen that seems more acceptable and more effective to the particular group of patients in the region.

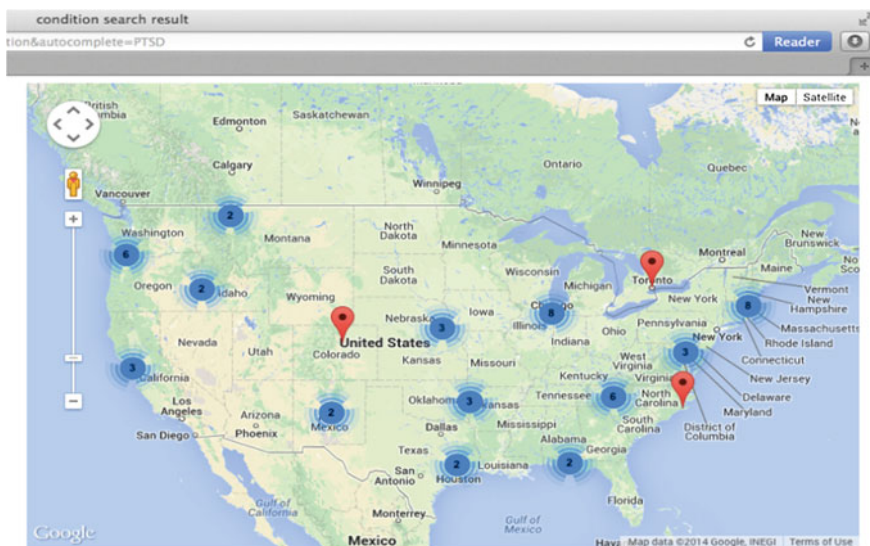


Fig. 7 Map of all patients who reported PTSD in their social health media

For PTSD, the most popular treatment, as shown in Fig. 6, is Individual Therapy, which is a form of Psychotherapy, evaluated by 95 patients, and has no side effect reports. The second most popular treatment, a prescription drug called Sertraline, has side effects such as weight gain (21%) and dry mouth (20%). The doctor can utilize the Social InfoButtons system to retrieve the symptoms and their severity levels. For PTSD, 297 patients reported severe flashbacks, 448 moderate flashbacks, 410 mild flashbacks, and 489 did not report flashbacks. He can compare his patient's symptoms with the other patients and learn that it is most likely his patient's flashback symptoms that may be mild. In summary, the Social InfoButtons system can help doctors to make decisions using knowledge of social trends and experiences of similar patients, using population-level intelligence as a benchmark, and compare it with diagnoses and treatment options for his patient.

A government agency can follow trends and understand whether discrepancies exist between official statistical data and social data. Social InfoButtons can allow officials to identify discrepancies, which may serve as a starting point for further investigations. For instance, there is no universally accepted treatment for Fibromyalgia, a common chronic pain condition. The government official or a researcher can query and browse query results and trigger queries that display analytics of contrasting data from official and social sources for Fibromyalgia. The analytic provides the list of treatments for the condition, ordered by popularity (defined as the number of treatment occurrences in the social space). Starting from this analytic, the knowledge worker can perform a comparison against authoritative sources. For the specific case, the user would discover that a treatment with

Table 4 Discrepancy on treatment types in Social Health Records and authoritative source

| Treatment in SI | # of Patients in SI | Appears in Authority |
|---------------------------|---------------------|----------------------|
| Duloxetine | 1058 | Yes |
| Pregabalin | 955 | Yes |
| Milnacipran | 357 | Yes |
| Gabapentin | 346 | Yes |
| Tramadol | 201 | Yes |
| Cyclobenzaprine | 188 | No |
| Amitriptyline | 141 | Yes |
| Hydrocodone–Acetaminophen | 128 | Yes |
| Naltrexone | 55 | No |
| Massage Therapy | 52 | No |
| Meloxicam | 50 | No |
| Venlafaxine | 46 | No |
| Carisoprodol | 43 | No |

Cyclobenzaprine is reported in social media data but not in official documents, as shown in Table 4.

Similarly, the agency may want to explore the distribution of the population reporting Asthma and how it compares with official data. An interactive map, supplemented with a heat map analysis, allows her to pinpoint the gender distribution by geographical area, and access contrast data via the given charts. Figure 8a, b shows the gender distribution for Asthma in the states of Ohio and Pennsylvania, respectively. From these two figures, it is interesting to note the following: first, there is a substantial difference between data from the official and the social sources; and, second, this difference is consistent across the states, i.e., Ohio and Pennsylvania.

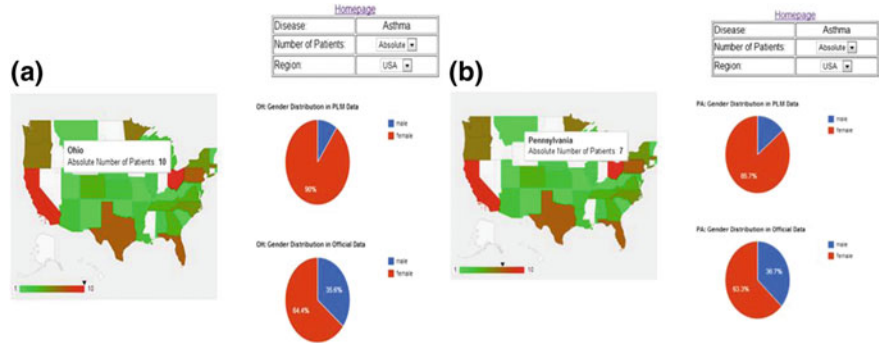


Fig. 8 Asthma distribution heat map and gender breakdown in **a** Ohio and **b** Pennsylvania

In addition, a patient wants to know more about his condition and he is interested in researching the scientific literature, joining social networks, exploring blogs or forums, etc. This can be a challenging task for a nonexpert. The plethora of information channels to consider that pose different levels of terminology issues, and his limited expertise can be prohibitive. Social InfoButtons can help this kind of individual to explore the knowledge, to gain an understanding of crowd level common behaviors or health practices through analytics provided by the system.

6.2 *Sentiment Analytics to Monitor Public Health Concerns*

We have also developed a sentiment analytics component named ESMOS (Epidemic Sentiment Monitoring System) to monitor the timeline and topic distribution of population-level public health concern [42, 43]. Using Twitter datasets, we developed a sentiment classification model, with unlabeled tweets and the subjective language as well as none-linguistic clues such as emoticons, to distinguish the personal from nonpersonal tweets (e.g., news tweets), and to distinguish positive from negative sentiments among personal tweets. The sentiment analysis results are used to calculate the population-level public concern toward a disease.

The ESMOS displays (1) a concern timeline chart to track the public concern trends on the timeline; (2) a tag cloud for discovering the popular topics within a certain time period with a capability to drill down to the individual tweets; and (3) a public health concern map to show the geographic distribution of particular disease concentrations with different granularities (e.g., state, county, or individual location level).

Figure 9 shows the different visual tools. The public health specialists can utilize the concern timeline chart, as shown in Fig. 9a, to monitor (e.g., identify concern peaks) and compare public concern timeline trends for various diseases. Then the specialists might be interested in what topics people are discussing on social media during the “unusual situations” discovered with the help of the concern timeline chart. To answer this question, they can use the word cloud analytics, as shown in Fig. 9b, to browse the top topics within a certain time period for different diseases and individual tweets. The public health concern heat map in Fig. 9c shows the state-level public concern levels.

This illustrates that Social Health Records, such as tweets, which may be considered as weak signals on their own, can be a source of population-level intelligence to understand the public health issues and attitudes toward a particular disease when analyzed in the large collective datasets. Here, again, each tweet analysis makes use of disease-related knowledge bases (e.g., disease ontology) and subjective language as background knowledge in classifying the tweets in building the classification models.

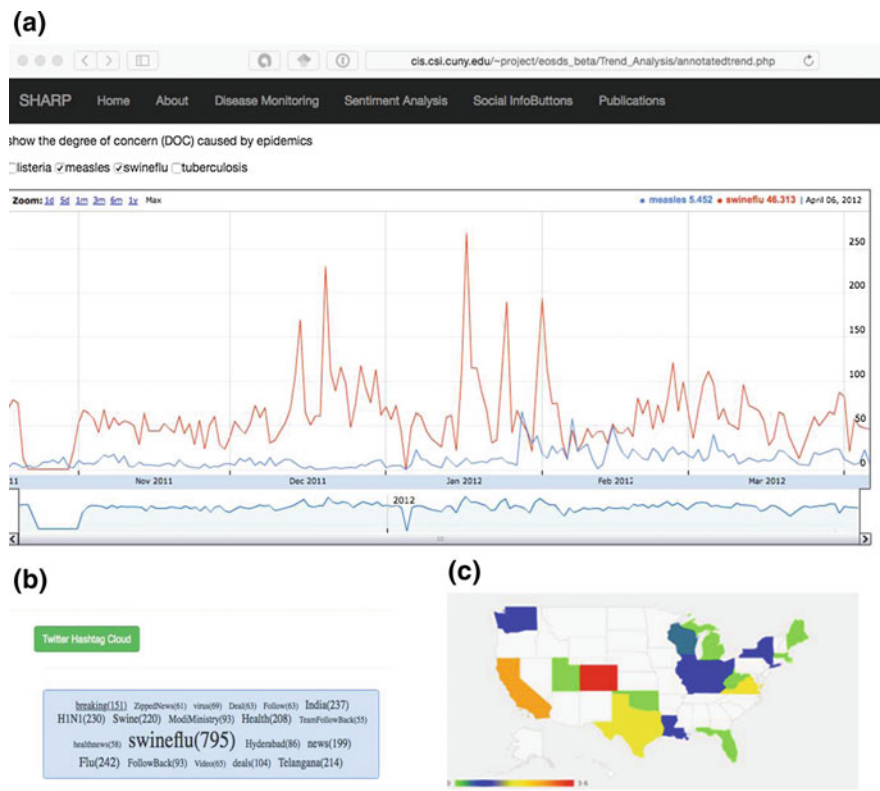


Fig. 9 a Public health concern trend line, b Topic trending, c Public health concern map

6.3 Comorbidity Prediction and Trajectory Analysis

Managing multiple coexisting conditions of one patient raises important public health issues, especially when conditions are associated with high costs. In the US, 80% of Medicare spending is expended for managing comorbidity of patients. For instance, obese patients often develop type-2 diabetes and hypertension. Thus, predicting potential comorbidity conditions for an individual patient or a group of patients with similar profiles can promote preventive care and reduce costs. In addition, predicting possible comorbidity progression paths using large datasets from the Social Health Records 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. With the SHRs, great opportunities are provided to study this kind of population-level predictive model building.

In building a prediction model for identifying a potential comorbid condition, or discovering all possible trajectory paths, we take two approaches [25, 26]: 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, i.e., the 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 comorbidity risk trajectory prediction, our approach uses a graph construction approach to build a connected graph from one condition to another condition, using edge discovery and linking discovered edges to reveal each potential progression trajectory between any two conditions and infer the confidence value of the future trajectory, given any observed condition. The predicted trajectories are validated with existing comorbidity relations from the medical literature.

The dataset from the patients’ self-posted data on the PatientsLikeMe website in 2012 included 17,418 patients’ information, including id, username, gender, age, and location, and 35,606 diagnosed conditions for these patients. Each diagnosis contains six attributes: PatientId, HasCondition, ConditionId, IsPrimaryCondition, FirstSymptomDate, and DiagnosisDate, for example, “ID: 8, HasCondition: Stroke, ConditionId: 48, IsPrimaryCondition: 0, FirstSymptomDate: May 1998, DiagnosisDate: Sep 1998”.

Using the variant algorithm of collaborative filtering approach, the top 2 predicted conditions are identified (see Table 5).

This result, based on the social data, has a good match with the official findings in medical literature as shown in Table 6. For instance, people with Fibromyalgia are predicted to have comorbidities of Chronic Fatigue Syndrome and Generalized Anxiety Disorder.

However, the medical literature results or the collaborative prediction model do not show the possible trajectory to show the progress from one condition to another, other than stating that these conditions likely co-occur.

The trajectory analyses using the SHRs have shown more promising transitional steps of the comorbidity direction. The following visual analysis using our approach in Fig. 10 shows the trajectory of the potential comorbidity progression for public health insights, using collective intelligence garnered from a large set of SHR records from many people.

Figure 10 shows the progression trajectory starting with “Major Depressive Disorder” (MDD). The numbers in parentheses on each node indicate the numbers

Table 5 Example of predicted comorbidity conditions associated with diagnosed conditions

| Id | Diagnosed Conditions | Top 2 Predicted Conditions |
|-----|----------------------------------|---|
| 296 | Migraine, Fibromyalgia | Chronic Fatigue Syndrome, Generalized Anxiety Disorder |
| 42 | Eating Disorder, Phobic disorder | Social Anxiety Disorder, PTSD |
| 50 | HIV, Seborrheic Dermatitis | Bipolar Disorder, Lactose Intolerance |

Table 6 Comorbidities from medical literature

| Condition Category | Comorbidity |
|---------------------------------|--|
| Major Depressive Disorder (MDD) | 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 |
| Irritable Bowel Syndrome (IBS) | Major Depression, Anxiety , Somatoform Disorders, Fibromyalgia , Chronic Fatigue Syndrome , Gastroesophageal Reflux Disease, Restless Legs Syndrome |

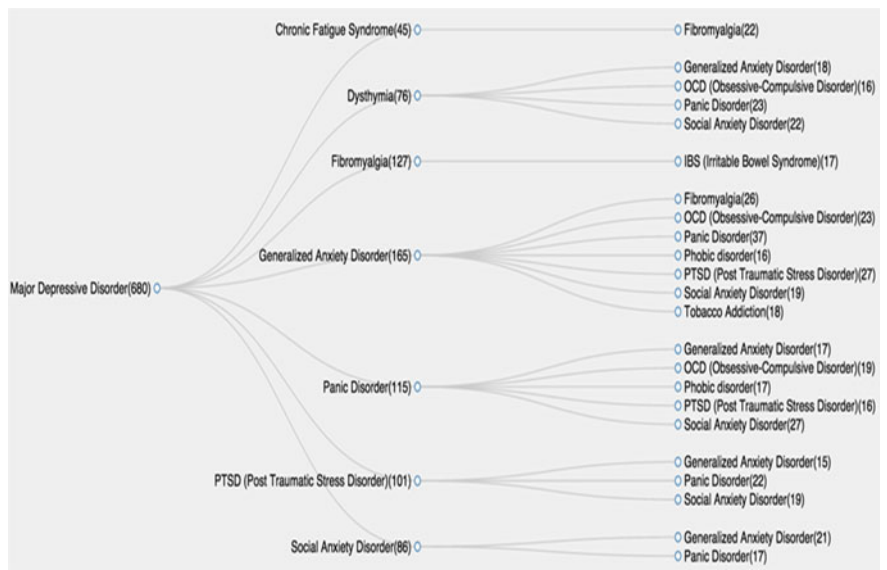


Fig. 10 The comorbidity progression trajectory model starting from “Major Depressive Disorder”

of patients following the trajectory from the root to the current node, e.g., there are 17 patients with the trajectory (MDD→Fibromyalgia→IBS). In Fig. 3, the most frequent length-2 trajectories are (MDD→GAD) (165 patients) followed by (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%), followed by (MDD→PTSD→PD) (3.2%) and (MDD→Social Anxiety Disorder→PD) (2.5%).

This comorbidity progression trajectory analysis results from our study as shown in Table 7 contrasting with those in Table 6, where the comorbid conditions are just listed without showing the progression. For instance, the trajectory MDD→Generalized Anxiety Disorder (GAD)*→Obsessive–Compulsive Disorder

Table 7 Trajectory analysis results for comorbidity prediction (comorbidity index in percentage/confidence value in percentage/support)

| Condition | Comorbidity |
|---------------------------------|---|
| Major Depressive Disorder (MDD) | MDD→Post-traumatic Stress Disorder (PTSD)*→Panic Disorder*→Social Anxiety Disorder* (0.25/1.3/9) MDD→PD*→SAD*→Phobic Disorder (0.23/1.1/8) MDD→Generalized Anxiety Disorder (GAD)*→Obsessive–Compulsive Disorder (OCD)* (0.7/3/23) MDD→PD*→OCD* (0.7/2/19) MDD→Bipolar II (1.7/4/21) MDD→Borderline Personality Disorder* (1.2/3/21) |
| Irritable Bowel Syndrome (IBS) | IBS→Gastroesophageal Reflux Disease (GERD)*→Restless Legs Syndrome (RLS)* (0.9/3/6) IBS→Fibromyalgia*→Chronic Fatigue Syndrome (CFS)* (0.3/9/17) IBS→RLS* (6/12/23) IBS→Osteoarthritis (3/10/18) |

*The comorbidity exists in medical literature

(OCD)* (0.7/3/23) shows that it is likely that patients will develop OCD through GAD from MDD.

The new insights from the SHRs can be used for anticipatory prevention measures with appropriate treatments.

7 Conclusions

In this chapter, we have shown that public health intelligence can be gathered from the Social Health Records shared by individuals on online social media, combined with the authoritative data shared by medical experts. We have presented a Social Health Analytics Platform for enabling the use of semantics in the analysis of Social Health Records to gain population-level health intelligence. The proposed Social Health Records Analytics Platform enables flexible collection of data from a variety of sources. Collected data is reconciled in a unified data model focusing on medical conditions and treatments and linked to create a knowledge base that enables cross-dataset exploration and analysis. Furthermore, the knowledge base can be extended by defining inference rules and using automatic reasoning.

We have illustrated the Social Health Analytics cases for population health, including the Social InfoButtons application to provide on-demand social health intelligence according to the information needs in different situations, sentiment analysis of Social Health Records to measure the population level of concern for health issues, and visual and trending analytics to provide situation awareness of disease evolution. We further discussed the comorbid progression analytics to predict the likely conditions to develop and the likely paths from one condition to another through time. The content of each individual Social Health Record

(SHR) may not provide many insights, but we showed that collectively the SHRs can bring great value for population health intelligence and understanding.

Many challenges still exist with using SHRs, because the data governance issues such as who owns the SHRs and who decides to share them need to be further addressed. So far, the use the SHRs has been relatively free of any governmental regulations and is subject to the data policies of online social media providers. However, there is concern that the existing privacy issues may prevent the effective utilization of SHRs for analytics in the future.

Acknowledgements The research work reported in this paper was partially funded by PSC-CUNY Research Foundation under the award numbers #64266 and #65232. The main research was carried out as part of the dissertation work by X. Ji at NJIT. The dataset was collected in the year 2012 when it is freely available. The data was processed right after.

References

1. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. *Health Inf. Sci. Syst.* **2**, 1–10 (2014)
2. Househ, M., Borycki, E., Kushniruk, A.: Empowering patients through social media: the benefits and challenges. *Health Inf. J.* **20**, 50–58 (2014)
3. Ji X, Chun, S.A., Geller, J.: Monitoring public health concerns using Twitter sentiment classifications. In: *Proceedings of IEEE International Conference on Healthcare Informatics*, pp. 335–344. Philadelphia, PA (2013)
4. Smith, C.A., Wicks, P.J.: PatientsLikeMe: consumer health vocabulary as a folksonomy. In: *Proceedings of American Medical Informatics Association Annual Symposium*, pp. 682–686. Washington D.C. (2008)
5. Bizer, C.: Evolving the web into a global data space. In: Fernandes, A.A., Gray, A.G., Belhajjame, K. (eds.) *Proceedings of 28th British National Conference on Databases*, p. 1. Springer, Manchester, UK (2011)
6. Bizer, C., Heath, T., Berners-Lee, T.: Linked data—the story so far. *Int. J. Semant. Web Inf. Syst.* **5**, 1–22 (2009)
7. Harth, A., Gil, Y.: Geospatial data integration with linked data and provenance tracking. In: *W3C/OGC Linking Geospatial Data Workshop*, pp. 1–5 (2014)
8. Specia, L., Motta, E.: Integrating folksonomies with the semantic web. In: *Proceedings of the 4th European Conference on The Semantic Web: Research and Applications*, pp. 624–639. Springer, Innsbruck, Austria (2007)
9. Fox, P., McGuinness, D.L., Cinquini, L., et al.: Ontology-supported scientific data frameworks: the virtual solar-terrestrial observatory experience. *Comput. Geosci.* **35**, 724–738 (2009)
10. Chun, S.A., MacKellar, B.: Social health data integration using semantic Web. In: *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 392–397 (2012)
11. MacKellar, B., Schweikert, C., Chun, S.A.: Patient-centered clinical trials decision support using linked open data. *Int. J. Softw. Sci. Comput. Intell.* **6**, 31–48 (2014)
12. Tofferi, J.K., Jackson, J.L., O'Malley, P.G.: Treatment of fibromyalgia with cyclobenzaprine: a meta-analysis. *Arthritis Rheum.* **51**, 9–13 (2004)
13. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends Inf. Retrieval* **2**(1–2), 1–135 (2008)

14. Zhuang, L., Jing, F., Zhu, X.-Y.: Movie review mining and summarization. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 43–50. Arlington, VAS (2006)
15. Chew, C., Eysenbach, G.: Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS ONE* **5**(11), e14118 (2010)
16. Chawla, N.V., Davis, D.A.: Bringing big data to personalized healthcare: a patient-centered framework. *J. Gen. Intern. Med.* **28**, 660–665 (2013)
17. Davis, D.A., Chawla, N.V., Christakis, N.A., Barabasi, A.L.: Time to CARE: a collaborative engine for practical disease prediction. *Data Min. Knowl. Disc.* **20**, 388–415 (2010)
18. S. Hassan and Z. Syed, “From netflix to heart attacks: collaborative filtering in medical datasets,” in *Proceedings of the 1st ACM International Health Informatics Symposium*, Arlington, Virginia, USA, 2010, pp. 128–134
19. Folino, F., Pizzuti, C.: A comorbidity-based recommendation engine for disease prediction. In: Proceedings of the IEEE 23rd International Symposium on Computer-Based Medical Systems, pp. 6–12. Bentley, Australia (2010)
20. Qian, B., Wang, X., Cao, N., Li, H., Jiang, Y.-G.: A relative similarity based method for interactive patient risk prediction. *Data Min. Knowl. Disc.* **29**, 1070–1093 (2015)
21. Hussein, A.S., Omar, W.M., Li, X., Hatem, M.A.: Smart collaboration framework for managing chronic disease using recommender system. *Health Syst.* **3**, 12–17 (2014)
22. Jensen, A.B., Moseley, P.L., Oprea, T.I., Ellesøe, S.G., Eriksson, R., Schmock, H., et al.: Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat. Commun.* **5** (2014)
23. Wang, X., Sontag, D., Wang, F.: Unsupervised learning of disease progression models. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 85–94. New York, NY (2014)
24. Hainke, K., Rahnenführer, J., Fried, R.: Disease progression models: a review and comparison. Dortmund University, Technical Report (2011)
25. Ji, X., Chun, S.A., Geller, J., Oria, V.: Collaborative and trajectory prediction models of medical conditions by mining patients’ social data. In: Proceedings of 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 695–700. Washington D.C. (2015)
26. Ji, X., Chun, S., Geller, J.: Predicting comorbid conditions and trajectories using social health records. *IEEE Trans. Nanobiosci.* **15**(4):371–379 (2016)
27. Ji, X., Chun, S.A., Geller, J.: Epidemic outbreak and spread detection system based on twitter data. In: Proceedings of the First International Conference on Health Information Science, pp. 152–163. Beijing, China (2012)
28. PHP Simple HTML DOM Parser. <http://simplehtmldom.sourceforge.net>. Accessed 14 Apr 2014
29. CDC Prevalence Data of Asthma in 2010. <http://www.cdc.gov/asthma/brfss/2010/brfssdata.htm>. Accessed 14 Apr 2014
30. Behavioral Risk Factor Surveillance System. <http://www.cdc.gov/brfss/>. Accessed 14 Apr 2014
31. Rao, D., McNamee, P., Dredze, M.: Entity linking: finding extracted entities in a knowledge base. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds.) *Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing*, pp. 93–115. Springer, Berlin (2013)
32. Hogan, A., Zimmermann, A., Umbrich, J., Polleres, A., Decker, S.: Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *Web Semant.* **10**, 76–110 (2012). doi:[10.1016/j.websem.2011.11.002](https://doi.org/10.1016/j.websem.2011.11.002)
33. Hassanzadeh, O., Kementsietsidis, A., Lim, L., Miller, R.J., Wang, M.: LinkedCT: a linked data space for clinical trials. *CoRR* **abs/0908.0567** (2009)
34. Chun, S.A., MacKellar, B.: Social health data integration using semantic Web. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, pp. 392–397. Trento, Italy (2012)

35. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**(Database issue), D267–270 (2004). doi:[10.1093/nar/gkh061](https://doi.org/10.1093/nar/gkh061)
36. Ji, X., Chun, S.A., Geller, J.: Social InfoButtons: integrating open health data with social data using semantic technology. In: *Proceedings of the Fifth Workshop on Semantic Web Information Management*, New York (2013)
37. SPARQL Query Language for RDF. <http://www.w3.org/TR/rdf-sparql-query/>. Accessed 14 Apr 2014
38. Collins, S.A., Currie, L.M., Bakken, S., Cimino, J.J.: Information needs, Infobutton Manager use, and satisfaction by clinician type: a case study. (1067–5027 (Print)) (2009)
39. Cimino, J.J., Elhanan, G., Zeng, Q.: Supporting infobuttons with terminological knowledge. In: *Proceedings of AMIA Annual Fall Symposium*, pp. 528–532. AMIA, Bethesda, MD (1997)
40. Cimino, J.J.: Use, usability, usefulness, and impact of an infobutton manager. In: *Proceedings of American Medical Informatics Association Annual Symposium*, pp. 151–155. AMIA, Bethesda, MD (2006)
41. Cimino, J.J., Li, J., Allen, M., Currie, L.M., Graham, M., Janetzki, V., Lee, N.J., Bakken, S., Patel, V.L.: Practical considerations for exploiting the World Wide Web to create infobuttons. *Medinfo* **11**, 277–281 (2004)
42. Ji, X., Chun, S.A., Wei, Z., Geller, J.: Twitter sentiment classification for measuring public health concerns. *Soc. Netw. Anal. Min.* **5**, 1–25 (2015)
43. Ji, X., Chun, S., Geller, J.: Knowledge-based tweet classification for disease sentiment monitoring. In: *Pedrycz, W., Chen S.-M. (eds.) Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence*, pp. 425–454. Springer (2016)

Public Health Intelligence and the Internet

Shaban-Nejad, A.; Brownstein, J.S.; Buckeridge, D.L.
(Eds.)

2017, XII, 148 p. 36 illus., Hardcover

ISBN: 978-3-319-68602-8