Drug-Drug Interactions Detection from Online Heterogeneous Healthcare Networks

Haodong Yang
College of Computing and Informatics
Drexel University
Philadelphia, PA 19104, USA
Haodong.Yang@drexel.edu

Christopher C. Yang
College of Computing and Informatics
Drexel University
Philadelphia, PA 19104, USA
Chris.Yang@drexel.edu

***Abstract*— Drug-drug interactions (DDIs) are a serious drug safety problem for health consumers and how to detect such interactions effectively and efficiently has been of great medical significance. Currently, methods proposed to detect DDIs are mainly based on data sources such as clinical trial data, spontaneous reporting systems, electronic medical records, and chemical/pharmacological databases. However, those data sources are limited either by cohort biases, low reporting ratio, or access issue. In this study, we propose to use online healthcare social media, an informative and publicly available data source, to detect DDI signals. We construct a heterogeneous healthcare network based on consumer contributed contents, develop heterogeneous topological features, and use logistic regression as prediction model for DDI detection. The experiment results show that the proposed heterogeneous topological features substantially outperform the homogenous ones in the training set but only slightly outperform the homogeneous ones in the testing set, and interesting heterogeneous paths with strong predictive power are discovered.**

Keywords—drug-drug interactions; healthcare social media; heterogeneous healthcare network; heterogeneous network mining; logistic regression

I. INTRODUCTION

In 2011, in a paper named "Can Computer Science Save Healthcare?", Howard Wactlar, Misha Pavel, and Will Barkis proposed a program of research and development along four technology thrusts to enable the improvement of American healthcare [1], two of which are listed as follows:

(1) a cyber-based empowering of patients and healthy individuals that enables them to play a substantial role in their own health and treatment; and

(2) utilizing diverse data to provide automated and augmented insight, discovery, and evidence-based health and wellness decision support.

The development of Web 2.0 and Health 2.0 technologies makes the first thrust promising. The advancement of Internet not only breeds the various online social media sites such as Facebook, Twitter, LinkedIn, and so on, but also leads to the flourishing of online healthcare social media sites such as MedHelp, PatientsLikeMe, DailyStrength, and so forth. A latest national survey, with regard to social life of health information, conducted in September 2012 by Pew Internet & American Life Project showed that with the year of 2012, 72% of adult internet users say they searched online for health information, and 26% adult internet users say they have read someone else's health experience about health or medical issues [2]. We can easily imagine that uncountable health consumers as well as health professionals go to those websites frequently to either seek or offer healthcare information. For example, since its introduction in 1994, MedHelp is the pioneer in online healthcare communities. Today, MedHelp empowers over 12 million people each month to take control over their health and find answers to their medical questions [3].

The development of Web 2.0 and Health 2.0 technologies also makes the second thrust promising. Frequent visits on healthcare social media would inevitably generate a huge collection of health-related contents that might be even more informative and timely than some administrative databases. Take adverse drug reactions (ADRs) as an example. It is highly possible that many patients choose social media over some available reporting systems such as Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) because of ignorance of these channels, embarrassment, perceptions of negative provider attitude, their extreme illness, etc [4]. Instead, they often resort to informal networks such as healthcare social media to discuss adverse reactions. If we can take good advantage of these consumer contributed contents, we may be able to discover interesting knowledge, insights and patterns that cannot be obtained from other data sources. Therefore, online health-related social media would serve as another data source to provide automated and augmented insight, discovery, and evidence-based health and wellness decision support.

Our previous research has shown that it is feasible to reveal knowledge from online healthcare social media. Drug safety surveillance is an important research area in pharmacovigilance. Our studies revealed that healthcare social media has become a reliable data source for effective ADRs detection [5, 6] and drug-drug interactions (DDIs) detection [7]. It has been long recognized that ADRs represent a significant health problem all over the world and they are considered to be a leading cause of death in the United States [8]. DDIs are a significant drug safety problem for health

consumers and may account for up to 30% of unexpected adverse drug reactions [9, 10], and they are common and often caused by shared pathways of metabolism or intersecting pathways of drug action [11]. DDI detection is of great clinical importance because most interactions could result in precaution of prescription, absolute contraindications of combination use, or even drug withdrawal from market [11], and therefore has become an important research area.

Some efforts have been made in clinical trials to detect DDIs [12-15]. However, clinical trials primarily focus on ADRs detection of single drugs and do not typically investigate DDIs [16]. Also, sample size, cohort biases, time spans, and financial limit could be some of the crucial factors that obstacle the discovery of DDIs [6, 17]. Therefore, post-marketing surveillance is a primary way of DDI detection. Current post-marketing drug surveillance in United States heavily depends on FAERS, but such spontaneous reporting data is to a great extent limited by its huge under-reporting ratio. It has been reported that only 1 to 10 percent of all reportable adverse effects were reported to FAERS, and the majority of these reports came from drug companies [18]. It means that many serious or rare adverse reactions may not be reported timely or even not reported at all, making detections difficult or even impossible. Therefore, alternative data source should be used to supplement the drug safety surveillance from spontaneous reporting systems. As we mentioned before, online healthcare social media provides a great platform for both health consumers and professional to discuss any health issues such as drugs, disease, ADRs etc., therefore it could be considered another data source for DDI detection.

In our previous work, we used content analysis for DDI detection. Concretely, we proposed to use association mining to identify DDIs from consumer contributed contents. In this paper, we focus on the interaction between two drugs and propose to detect DDI signals by analyzing the structure of a heterogeneous network that is constructed from online healthcare social media data.

II. Literature Review

In most of the current research on network science, social and information networks are usually assumed to be homogeneous, where nodes are objects of the same entity type and links are relationships from the same relation type. However, most real-word networks are heterogeneous, where nodes and relations are of different types [19]. For example, the network of Twitter consists of persons as well as tweets, photo, video, location, and so on, and the relationships could be following, followed, person-tweets, person-location, and so forth. Healthcare social media is also a heterogeneous network. Not only does it include drugs and ADRs discussed by consumers, but also contains other types of node such as diseases, treatments, communities and so on. Therefore, in addition to content analysis, heterogeneous network mining on healthcare social media provides another approach for discovering health-related knowledge such as DDI detection through link mining.

A. Link Prediction in Heterogeneous Information Network

Link prediction, dedicated to addressing the question of whether a link will be formed in the future, is an important subtask in link mining. It is defined as predicting the emergence of links in a network based on certain current or historical network information [20]. Given the problem of DDI detection, we are actually predicting if there will be link between two different drugs. Therefore, we can formulate the DDI detection as a problem of link prediction. In the recent years, link prediction has been a popular research theme due to the fast development of social networks, and a number of methods have been designed for link prediction.

As one of the early researchers who started working on link prediction, in 2003, Liben-Nowell and Kleinberg formalized the link prediction problem [21]. In [21], they used an unsupervised approach to predict the links based on a set of network topology features such as graph distance, common neighbors, Jaccard's coefficient, preferential attachment, etc. in co-authorship networks. In another study, Hasan et al. [22] used a supervised learning approach for co-authorship link prediction based on a set of easily computed features. Based on those features, Decision Tree, k-Nearest Neighbors, Multilayer Perceptron, Support Vector Machine, and RBF network were used as classification models for link prediction.

Most existing link prediction approaches are designed for homogeneous networks that only contain one type of objects such as authors in citation networks or social media users in friendship networks. However, as stated before, in our real world, most networks are heterogeneous where nodes and relationship are of different types, and there are only a limited number of studies that have leveraged heterogeneous network for link prediction.

In [19], Sun et al. studied the problem of co-authorship prediction in heterogeneous bibliographic network. Specifically, they first used a structure called *network schema* to summarize the heterogeneous network and proposed a new concept called *meta path* that can be extracted from *network schema*. Then they proposed 4 topological measures on those *meta paths*, which are path count, normalized path count, random walk, and symmetric random walk. At last, the authors viewed the link prediction as a binary classification problem and proposed to use logistic regression model as the supervised prediction model. Other than predicting **whether** a link will be built in the future, Sun et al. also conducted a study addressing the problem of **when** it will happen. In [23], they used meta path-based topological features and a generalized linear model (exponential distribution, Weibull distribution and geometric distribution) based supervised framework to predict the building time of author citation relationship.

Biomedical text mining is one of a few of health-related research fields that have utilized techniques of heterogeneous network mining for knowledge discovery. Different semantic types of biomedical concepts (e.g. drugs, genes, proteins, etc.) extracted from biomedical literatures and relationships between them (e.g. biochemical associations, regulatory relationship, etc.) form biomedical heterogeneous concept networks, from which knowledge could be revealed by using various heterogeneous network mining techniques. The Link prediction problem in biomedical concept networks could be referred to as hypotheses generation [24]. For example, Katukuri et al. [25] modeled a biomedical literature repository as a comprehensive network of different semantic types of biomedical concepts and formulated hypotheses generation as a process of link discovery on the concept network.

To the best of our knowledge, there have been only a very limited number of research that uses techniques of heterogeneous network mining on online healthcare social media for knowledge discovery. No studies have been found that use heterogeneous network information for DDI detection.

B. DDI Detection

Recently, there is a large number of research works that have been dedicated to the signal detection of DDIs. Similar with single-drug-single-ADR detection studies, in terms of the data sources used, DDI detection studies could be roughly grouped into five categories: (1) clinical trial data, (2) spontaneous reporting systems, (3) electronic medical records, (4) chemical/pharmacological databases, and (5) consumer contributed contents.

1) Clinical Trial Data

Some efforts have been made in clinical trials to detect DDI signals [12-15]. The disadvantages of using this kind of data lies in the fact that clinical trials primarily focus on establishing the safety and efficacy of single drugs, and do not typically investigate DDIs [10, 17]. Also, sample size, cohort biases, time spans, and financial limit could be some of the crucial factors that obstacle the discovery of DDIs [6, 10, 17].

2) Spontaneous Reporting Systems

Spontaneous reporting systems are a major repository for DDIs signal detection, and many endeavors have been dedicated to detecting DDIs using such data source. For example, Tatonetti et al. [10] used FDA's single-drug reports to build classification models for some pre-defined adverse drug events, and then performed logistic regression to look for pairs of drugs that match these single-drug profiles in order to predict potential interactions. Harpaz et al. [26] used a well-established data mining method – association rules mining – and used *support* and *lift* as the measures to discover DDIs signals. Thakrar et al. [27] investigated two models, a multiplicative and an additive model, to detect signals of DDIs from FDA's spontaneous reporting system. There are also spontaneous reporting centers in some European and Asian countries, and fruitful contributions have been made by using those reports [28-31]. Spontaneous reporting systems have been proved to be a useful source to detect DDIs. However, the nature of passiveness of these systems directly caused the extremely high underreporting ratio. It is especially difficult to detect new and emerging signals because a large number of interesting cases cannot be timely collected due to the underreporting ratio of the current reporting system [32].

3) Electronic Medical Record

Several studies proposed to mine DDIs from Electronic Medical Records (EMR). For example, Chan et al. [33] retrieved drug utilization reports from EMR to determine the patients who were prescribed with antidepressants and oral anticancer drugs between 2006 and 2009 at a cancer center. Zwart-van Rijkom et al. [34] used EMR of all patients hospitalized in the University Medical Centre Utrecht in 2006 who were prescribed at least one medication to calculated the percentage of patients experiencing at least one DDI and the percentage of prescriptions generating a DDI alert. The disadvantage of such data source is that EMR is often difficult to access because of privacy issues that it is usually available only to those research groups that have cooperation with hospitals, clinics or any other health organizations and communities [5].

4) Chemical/Pharmacological Databases

Some studies also focus on chemical/pharmacologic databases to detect signals of DDIs and DrugBank database is a typical exemplar. For example, Segura-Bedmar et al. [35-37] proposed two different methods – pattern matching and supervised machine learning (shallow linguistic kernel) – to automatically extract DDIs from biomedical texts retrieved from DrugBank. Vilar et al. [9] presented a methodology applicable on a large scale that identifies novel DDIs based on molecular structural similarity to drugs involved in established DDIs. Access issue is also one of the disadvantages of this kind of data because not all the chemical/pharmacological databases are free and public to everyone. Also, this kind of database more focuses on the chemical aspect such as drug structure than textual aspect.

5) Consumer Contributed Contents

As stated before, consumer contributed contents provide a great asset for us to mine knowledge and patterns. In the research area of DDI detection, such data has also been used to achieve the goal although the number of such works is very limited. In 2013, White et al. [38] demonstrated that Internet users are able to provide early clues about adverse drug reactions via their search logs that are generated by consumers. This strong evidence proved that consumer contributed content could be used for DDI detection. Also, one of our previous studies proposed to detect DDIs signals from consumer contributed contents in online health communities (e.g. MedHelp) using associations mining [7]. We conducted an experiment with 13 drugs and 3 DDI associations. *Leverage*, *lift* and *interaction ratio* were used in the experiment. DrugBank was used as gold standard to test the performance of the approach. The results showed that our techniques are promising to detect signals of DDIs and the proposed measure, *interaction ratio*, performs better than *leverage* and *lift*.

Given the gap we found that very few studies use techniques of heterogeneous network mining on online healthcare social media and also very few works use such data source for DDI detection, in this paper, we combine the techniques and the data source, and propose to identify DDI signals by mining the structure of a heterogeneous network that is extracted from online healthcare social media data.

III. METHODOLOGY

In this section, we introduce in detail the definition of heterogeneous healthcare network, the topological features extracted from such network, and the model for DDI detection task in such network setting.

A. Heterogeneous Healthcare Network

A heterogeneous network is defined as a graph $G = (\mathcal{N}, \mathcal{L})$ consisting of nodes joined by links, where $N = \{n_1, n_2, \dots, n_g\}$, $L = \{l_1, l_2, \dots, l_L\}$ and l_i can be directional or non-directional. In the graph G, each node $n_i \in \mathcal{N}$ belongs to one particular type from $\mathcal{T}$, each link $l_i \in \mathcal{L}$ belongs to one particular relation from $\mathcal{R}$, and the number of the types of nodes $|\mathcal{T}| > 1$ or the number of types of relations $|\mathcal{R}| > 1$.

A healthcare social media can be modeled as a heterogeneous healthcare network in which there are a set of node types, such as *Drug*, *Disease*, *ADR*, *Treatment*, *Diagnostics*, *Users*, etc. and a set of relation types, such as *treat* or *is treated* between *Treatment* and *Disease*, *cause* or *is caused* between *Drug* and *ADR*, *use* or *is used* between *User* and *Drug*, *have* or *had* between *User* and *Disease*, etc.

B. Healthcare Network Model

A network model $M_G = (\mathcal{T}, \mathcal{R})$ is a compressed representation for a heterogeneous network $G = (\mathcal{N}, \mathcal{L})$, which is a directional or non-directional graph consisting of node types $\mathcal{T}$, with links as relations from $\mathcal{R}$.

Fig. 1 succinctly presents a directional network model of a heterogeneous healthcare network. As we can see, the network includes four types of nodes, namely *Drug*, *ADR*, *Disease*, and *User*. For abbreviation, we use a capital letter to represent each node type, i.e. *R* for *Drug*, *A* for *ADR*, *D* for *Disease*, and *U* for *User*. The relations in this network contain *cause* or *is caused* between *R* and *A*, *treat* or *is treated* between *R* and *D*, *show* or *is shown* between *U* and *A*, *have* or *is had* between *U* and *D*, and *take* or *is taken* between *U* and *R*.

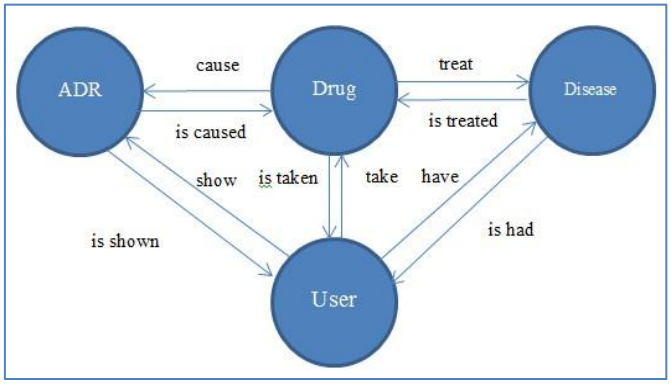

Fig. 1 Directional Network Model for Heterogeneous Healthcare Network

A directional network model can be extracted from a heterogeneous network only when the relation between a pair of different types of node can be determined. For example, a bibliographic network is a typical example of heterogeneous network and could be represented by a directional network model. The relations among different types of node, such as *paper*, *author*, *venue*, and *topic*, can be explicitly and easily determined. For example, nodes *author* and *paper* are linked together by the relation of *write* or *is written*, *paper* and *venue* by *publish* or *is published*, *paper* and *topic* by *mention* or *is mentioned*, etc. (Detailed example of bibliographic heterogeneous network mining can be found in [19, 23]) However, not all heterogeneous networks contain explicit relations among different types of nodes, which mean the semantic meaning of the relation between two nodes could not be easily determined. Under such circumstances, the heterogeneous network could be represented as a non-directional network model and the relation between two different types of node can be the association between them. For example, given a dataset of consumer contributed contents collected from an online healthcare forum, which contain different types of nodes such as *R*, *A*, *D*, and *U*, it is not an easy task to accurately determine the explicit relations between two nodes without using sophisticated natural language processing techniques or thorough human annotation. However, as we know, it is still challenging to use natural language processing techniques to analyze social media data and thorough human annotation would be very time consuming. In such case, the heterogeneous network can be summarized as a non-directional network model and the co-occurrence of two nodes that can be easily determined in an analysis unit (e.g. a thread or a message) could be regarded as their relation. Fig. 2 shows a heterogeneous healthcare network that is represented as a non-directional network. To summarize, in this work we aim at analyzing a non-directional heterogeneous healthcare network which contains four types of nodes (namely R, A, D, and U) joined together by their co-occurrence in an analysis unit.

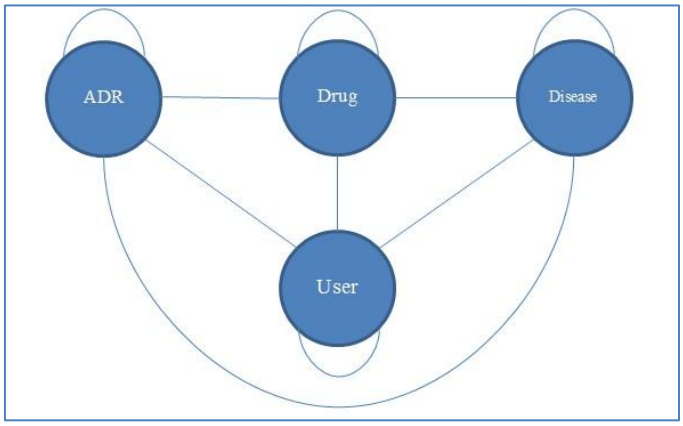

Fig. 2 Non-directional Network Model for Heterogeneous Healthcare Network

C. Topological Features in Heterogeneous Healthcare Network

Topological features are also called structural features, which are extracted connectivity properties for pairs of objects in the networks [23]. Based on homogeneous network which only contains a specific type of nodes, there are a number of well-known and frequently used topological features. Most of the features are either path-based, such as graph distance, Katz_β [21] and propflow [39] or neighbor-based, such as common neighbors, Jaccard's coefficient, Adamic/Adar, preferential attachment, and SimRank_γ [21]. However, in a heterogeneous network, as a neighbor of one node could belong to different types and a path could also flow though different types of nodes, the commonly used features in homogeneous networks may no longer be applicable in such situation. For instance, in a heterogeneous healthcare network, two different drugs could be related by the path $R - D - U - D - R$ because of the co-occurrence of each two adjacent nodes in analysis units, and the **possible** (we cannot know for sure until we read the contents posted by the user) semantic meaning of such path could be explained as "a user has two different diseases which are treated by two different drugs respectively." However, such information cannot be inferred from a homogeneous healthcare network that only consists of drugs. Therefore, some novel features that can reflect the characteristics of a heterogeneous network should be designed.

In this study, we define $T_sT_d - Path - L$ as a topological feature of a heterogeneous network. A $T_sT_d - Path - L$ is an abstract path defined between two types of nodes T_s and T_d

with length L. It is extracted from the network model $M_G = (\mathcal{T}, \mathcal{R})$, and is presented in the form of $T_s \xrightarrow{R_1} T_1 \xrightarrow{R_2} \dots \xrightarrow{R_{L-1}} T_{L-1} \xrightarrow{R_L} T_d$. When the specific types of relations and directions cannot be determined between nodes, $T_1T_2 - Path - L$ takes the form of $T_s - T_1 - \cdots - T_{L-1} - T_d$ with links denoting associations between nodes. Table 1 lists all the symmetric $R_sR_d - Path$ with length 1 to length 4, and there are 16 such paths in total given four different types of nodes *R*, *A*, *D*, and *U*. The link existing between two nodes specifies the co-occurrence association between them.

Table 1 Symmetric $R_sR_d - Path - 1$ to $R_sR_d - Path - 4$ in a Heterogeneous Healthcare Network

Path	Length
R – R	1
R – A – R	2
R – D – R	2
R – U – R	2
R – A – A – R	3
R – D – D – R	3
R – U – U – R	3
R – A – A – A – R	4
R – A – D – A – R	4
R – A – U – A – R	4
R – D – A – D – R	4
R – D – D – D – R	4
R – D – U – D – R	4
R – U – A – U – R	4
R – U – D – U – R	4
R – U – U – U – R	4

There are several ways of quantifying the topological features in a heterogeneous network. Sun et al. [23] proposed to use such measure as path count, normalized path count, random walk, and symmetric random walk to quantify the features given two types of nodes and the path between them. In this work, without loss of generality, we use the count of path between two nodes as the measure. Fig. 3 illustrates an example of $R_sR_d - Path - 4$ between two different nodes: path $R - D - U - D - R$, and we can easily calculate the count the path instances between R_1 and R_2, which is 4.

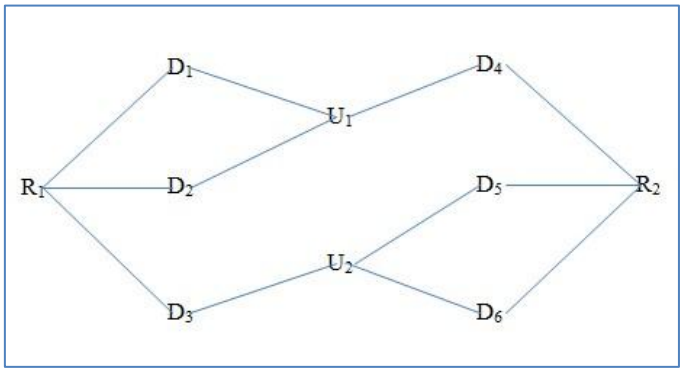

Fig. 3 $R - D - U - D - R$: an Example of $R_sR_d - Path - 4$ between Two Different Drugs

D. DDI detection Model

In this study, we view DDI detection as a binary classification problem. Concretely, given a pair of drug nodes, we use a classification model to label them as either "1" (interaction) or "0" (no interaction) based on their quantified topological features extracted from the heterogeneous healthcare network.

Logistic regression is utilized as our binary classification model which is able to output the probability of two drugs being interacting with each other. In logistic regression, we are trying to minimize the following cost function:

$$J(\theta) = -\frac{1}{m}\left(\sum_{i=1}^{m}(y^{(i)} log h_\theta(x^{(i)}) + (1 - y^{(i)})\log(1 - h_\theta(x^{(i)}))\right) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

where

- m is the number of training example (drug pairs);
- n is the number of topological features;
- $x^{(i)} \in \mathbb{R}^{n+1}$: a $n+1$ dimensional vector including a constant 1 and n topological features;
- $\theta \in \mathbb{R}^{n+1}$: a $n+1$ dimensional vector of parameters associated with constant 1 and each of the n topological features;
- $h_\theta(x^{(i)}) = g(\theta^T x^{(i)})$ where $g(z) = \frac{1}{1+e^{-z}}$ is called sigmoid function or logistic function and θ^T is the transpose of θ;
- For each training example (a pair of drug nodes $< R_1^{(i)}, R_2^{(i)} >$), $y^{(i)} = 1$ if they have interaction, and $y^{(i)} = 0$ otherwise;
- $\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$: regularization term for the purpose of preventing overfitting problem where λ is the regularization parameter.

Then we use the following gradient descent to find the optimal $\hat{\theta}$ that minimizes the cost function:

$$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)}, j = 0;$$

$$\theta_j := \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)} + \frac{\lambda}{m}\theta_j\right], j = 1,2,3,\dots,n;$$

where α is the learning rate. At last we applied the learned $\hat{\theta}$ to predict the drug pairs in our test set using the hypothesis $h_\theta(x^{test}) = \frac{1}{1+e^{-\theta^T x^{test}}}$, and $y^{test} = 1$ if $h_\theta(x^{test}) \geq 0.5$ and $y^{test} = 0$ otherwise.

IV. EXPERIMENT

A. Dataset Collection

In this study, MedHelp, an online healthcare social media site, is used as the source of health consumers' posts and comments. : "Every day, members come to MedHelp to receive the support they need from other patients like them, to search information on drugs and health topics, to document their medical history, and to share their knowledge with others in need. [3]" We focus on the drug section, which is one of the most important and popular components in MedHelp, and use the search engine[1] provided by the website to search for all the

[1] http://www.medhelp.org/search

threads about a certain drug. To effectively detect DDI signals, the drugs should bear active discussion in MedHelp. Therefore, we targeted 20 drugs that have more than 500 threads for each of them, and collected all the original posts and following comments of those drugs. The 20 drugs include Adenosine, Biaxin, Cialis, Concerta, Elidel, Epogen, Gadolinium, Geodon, Heparin, Lansoprazole, Lantus, Lunest, Luvox, Prozac, Risperdal, Simvastatin, Tacrolimus, Vyvanse, Zocor, and Zyprexa. The names of those drugs come from FDA's website, which includes an index of drugs that have been the subject of a Drug Safety Communication, Healthcare Professional Information sheet, Early Communication About an Ongoing Safety Review, or other important information[2]. Those drug names could be generic name or brand name. Fig. 4 shows an example of thread about Biaxin. In total, there are 16,344 threads which range from the year of 1995 to 2012.

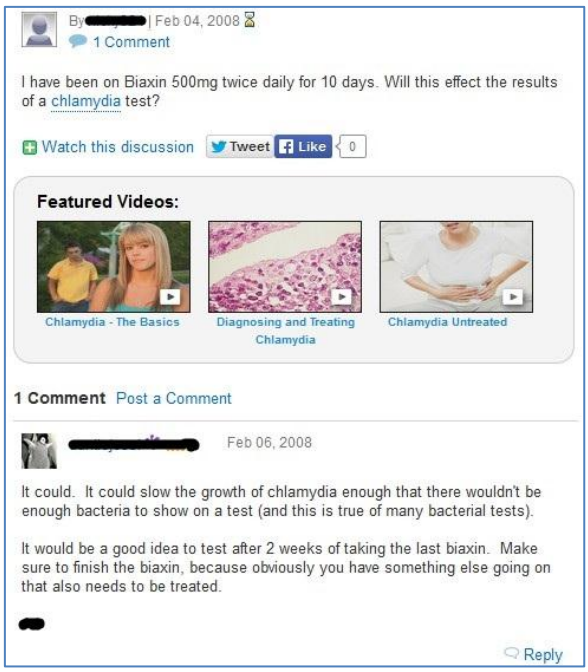

Fig. 4 A Thread of Biaxin

B. Network Construction

To construct the heterogeneous healthcare network, we need to extract different types of nodes and their relations from all of the threads.

1) Nodes

In this work, we focus on four different types of nodes, namely *R*, *A*, *D*, and *U*. External lexicons are used to extract those nodes. For *R*, besides the 20 drug names collected, we also add three other drugs (i.e. Quinidine, Ticlopidine, and Gemfibrozil) that could interact with some of the 20 drugs into our drug list. For more information about the three drugs and the interactions, please refer to [7]. For *A*, we focus on 5 ADRs (i.e. Depression, Diarrhea, Heart Disease, Kidney Disease, and Suicidal), and use Consumer Health Vocabulary (CHV) Wiki[3] to build our ADR lexicon. CHV is defined as "a collection of forms used in health-oriented communication for a particular task or need (e.g., information retrieval) by a substantial percentage of consumers from a specific discourse group and the relationship of the forms to professional concepts" [40]. It reflects the difference between consumers and professionals in expressing health concepts and helps to bridge this vocabulary gap. Therefore, using CHV is able to help us capture more consumers' expressions and better extract ADR terms. For *D*, we used a list of disease collected from MedHelp, which includes 47 diseases, such as HIV, IBS, Bipolar Disorder, Stomach Ulcers, OCD, Stroke, and so on. For *U*, we extract the user names, including thread originator and all following commenters from each thread. The dataset is de-identified before conducting the experiment.

2) Links

As discussed earlier, given a dataset of consumer contributed contents collected from an online healthcare forum, it is a challenging task to accurately determine the explicit relations between two nodes without using sophisticated natural language processing techniques or thorough human annotation. In this study, we treat our heterogeneous healthcare network as non-directional, and two nodes are linked together if they co-occur in the same thread, including both original post and following comments. It means that all the different types of nodes extracted from a thread will form a complete graph, i.e. each node has a link to all other nodes in the same thread. Each thread forms a sub-graph and then all these sub-graphs are combined together to generate our ultimate heterogeneous healthcare network.

C. Gold Standard

In terms of co-authorship prediction in a heterogeneous bibliographic network, the gold standard can be straightforwardly and conveniently set up, because the relations between nodes are explicit in terms of semantic meaning: if two different authors are linked to the same paper, they are co-authors; otherwise they are not. However, in our healthcare network, the links between nodes are based on co-occurrence and their semantic meanings are implicit, so even if two different nodes are linked together, it does not mean that they will interact with each other. Therefore, an external database DrugBank is used to set up the gold standard. The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information [41-43].We search for all the 23 drugs to see if one drug is reported to interact with any other drugs using the *Interax Interaction Search*[4] engine provided by the DrugBank. If two drugs are reported to have interaction, we label the pair of drugs nodes as "1", and otherwise as "0". For example, Biaxin is reported by interact with Quinidine to cause Arrhythmias, Simvastatin is reported to interact with Gemfibrozil to cause Myopathy, etc.

D. Experiment Setting

In order to control the size of the constructed heterogeneous healthcare network, we only retain the nodes with frequency larger than 20 and links with frequency larger than 15. Frequency here means the number of threads in which the node

2 http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm111085.htm#B

3 http://consumerhealthvocab.chpc.utah.edu/CHVwiki/

4 http://www.drugbank.ca/interax/drug_lookup

or link appears. After filtering, there are 394 nodes and 3255 links in our final network. For each pair of nodes, we use all the 16 symmetric $R_s R_d - Path - L$ (Table 1) as their features and count the number of path instances for each feature. In the experiment, we found that our dataset is unbalanced. Compared with all the possible drug pairs, those who can interact with each other only account for a small portion (only 17 such pairs), which means the positive drug pairs are far fewer than the negative ones, and the ratio of those two in our experiment is approximately 1:12. Therefore, based on the original unbalanced dataset, we build a new dataset that contains an equal sized set of positive pairs and negative pairs by undersampling. The whole process of dataset preparation is summarized as follows:

a. Given the positive drug pairs, randomly select an equal sized set of negative pairs to form a new dataset;

b. From the new dataset, randomly select 70% of positive and negative pairs respectively to form the training set, and the remaining 30% of drug pairs to form the test set;

c. Train a logistic regression model using the training set and apply the learned parameters on the test set.

To evaluate the classification accuracy, we use two measures – accuracy and F1 score – to assess the proposed techniques. They are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

and

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where TP is true positive, FP is false positive, and FN is false negative.

We also use sensitivity and specificity that are often used in medical area to evaluate the proposed techniques.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

In our experiment, we repeat above process of a to c 50 times to obtain the average value of accuracy and F1 score, sensitivity and specificity on both training set and test set.

E. Results and Discussion

1) Interesting Features

In logistic regression, the probability of an example being classified as "1" is modeled as follows:

$$P\left(y^{(i)} = 1 \middle| x^{(i)}; \theta\right) = \frac{1}{1 + e^{-\theta^T x^{(i)}}}$$

when we compute the odds of an example being classified as "1", we can derive the following equation:

$$\frac{P(y^{(i)} = 1)}{P(y^{(i)} = 0)} = \frac{P\left(y^{(i)} = 1 \middle| x^{(i)}; \theta\right)}{1 - P(y^{(i)} = 1 | x^{(i)}; \theta)} = e^{\theta^T x^{(i)}}$$
$$= e^{\theta_0} e^{\theta_1 x_1^{(i)}} \dots e^{\theta_n x_n^{(i)}}$$

where e^{θ_j} is called **odds ratio**, which specifies how much the odds changes multiplicatively with a one-unit change in the feature $x_j^{(i)}$, and θ_j itself is **log-odds ratio**. Therefore, the parameter vector θ is log-odds ratios. Positive values indicate a positive correlation between the probability of being classified as "1" and the corresponding features; whereas negative values indicate a negative correlation. The larger the absolute value of the θ_j is, the more predictive power the corresponding topological feature has. A θ_j with value close to zero means the corresponding feature is relatively neutral in terms of prediction power.

After looking into the optimal parameter vector $\widehat{\boldsymbol{\theta}}$, we found several interesting features that have strong predictive power in terms of DDI detection. Table 2 and Table 3 present the top topological features with positive and negative $\boldsymbol{\theta}$ values respectively. As we can see, path R – A – R is strongly positively correlated with the probability of a drug pair being classified as "interaction". The possible semantic meaning of this path could be explained as two different drugs causing the same ADRs. It means that the more same ADRs two different drugs are associated (co-occur) with, the more likely that those two drugs could interact with each other. This pattern also makes sense physiologically: when two drugs operate on the same pathway or are cleared by the same pathway, they are likely to cause the same adverse reactions - administering both at the same time is therefore likely to cause an adverse effect (and therefore an interaction) since the same pathways are involved. However, when two drugs are associated with different ADRs (feature R – A – A – R), it is less likely that there is an interaction between them. Also, when two drugs are associated with a single disease, either linked directly (R – D – R) or through users (R – U – D – U – R), it is less likely that those two drugs have interaction. A possible reason is that if two drugs are used to treat the same disease, they might share similar internal molecular structures so that they are less likely to interact with each other. When associated with two (R – D – D – R) or three (R – D – D – D – R) different diseases and the more such paths exist, the two drugs could be quite different and the more likely that they would interact. However, when more users are involved in such paths (R – D – U – D – R), the less likely the interaction will happen, because if a user who is having two diseases are prescribed a couple of drugs as separate treatments, the drugs are not supposed to be able to interact. We also notice that in homogeneous network, path R – R has the highest $\boldsymbol{\theta}$ value, which mean it is the most powerful predictor among all the 4 features. However, in the heterogeneous healthcare network, the $\boldsymbol{\theta}$ value of path R – R is 0.74 whose predictive power is weaker than many other paths. In Table 2 and Table 3, we listed the ratio of $\boldsymbol{\theta}$ values between those high-predictive-power paths and R – R. As we can see, for instance, the predictive power of paths R – A – R and R – D – R is as about 7.31 and 7.14 times as that of path R – R respectively.

Table 2 Top Topological Features with Positive θ Value

Features	θ Value	θ/θ^{R-R}
R – A – R	5.41	7.31
R – D – D – R	1.96	2.65
R – D – D – D – R	5.63	7.61

Table 3 Top Topological Features with Negative θ Value

Features	θ Value	θ/θ^{R-R}
R – D – R	-5.28	7.14
R – A – A – R	-5.34	7.22
R – D – U – D – R	-3.55	4.80
R – U – D – U – R	-3.25	4.39

2) Accuracy, F1 Score, Sensitivity, and Specificity

We compare the classification results using heterogeneous topological features with that using homogeneous ones. For the heterogeneous features, we count the number of path instances for all the 16 symmetric $R_sR_d - Path - L$; whereas for the homogeneous features, we construct a homogeneous network that only contain one type of node – drug, and count the number of path instances between each drug pairs to quantify the homogeneous topological features with length no longer than 4, namely R – R, R – R – R, R – R – R – R, and R – R – R – R – R. Table 4 and Table 5 present the comparison results between heterogeneous topological features and homogeneous topological features in terms of classification accuracy, F1 score, sensitivity and specificity on training set and test set respectively. As we can see, in the training set, heterogeneous features substantially outperform homogeneous one in terms of all the four measures; in the test set, except sensitivity, heterogeneous features slightly outperform homogeneous ones in terms of other three measures.

Table 4 Performance on Training Set

	Accuracy	F1 Score	Sensitivity	Specificity
Heterogeneous Features	**0.91**	**0.91**	**0.94**	**0.88**
Homogeneous Features	0.77	0.79	0.87	0.67

Table 5 Performance on Test Set

	Accuracy	F1 Score	Sensitivity	Specificity
Heterogeneous Features	**0.79**	**0.81**	0.88	**0.69**
Homogeneous Features	0.78	0.80	**0.89**	0.67

As shown in the experiment results, using homogeneous features, we can achieve similar performance on both training set and test set because path R – R is the most powerful predictor and dominate the classification results. However, when switching to heterogeneous features, we achieved better performance on training set than on test set. It means although heterogeneous features contain more information than homogeneous ones, when we applied the parameters trained by logistic regression to another dataset (e.g. test set), the performance dropped and was slightly better than that of homogeneous features. Therefore, logistic regression may not be the most suitable algorithm that is capable of exploring the power of heterogeneous features, and in the future, we need to explore some other models for the prediction.

V. CONCLUSION

The development of Health 2.0 technologies leads the booming of online healthcare social media such as MedHelp, PatientsLikeMe and so on. Such platforms are not only empowering patients and individuals to play a substantial role in their own health and treatment, but also generating informative data that can be used providing automated insights, discovery, and health and wellness decision support. Drug-drug interactions are a serious drug safety problem for health consumers and how to detect such interactions effectively and efficiently has been of great medical significance and become an important research area. Currently, methods proposed to detect DDIs are mainly based on such data sources as clinical trial data, spontaneous reporting systems, electronic medical records, and chemical/pharmacological databases. However, those data sources are limited either by cohort biases, low reporting ratio, or by access issue. In this study, we proposed to harness online healthcare social media for DDI detection. We used MedHelp as our source to collect consumer contributed contents based on which a heterogeneous network was constructed. Then we extracted topological features from the network and used logistic regression as prediction model for DDI signal detection. The experiment results showed that the proposed heterogeneous topological features substantially outperform the homogeneous ones in the training set but only slightly outperformed the homogeneous ones in the testing set. More importantly, according to the parameters of prediction model, we discovered several paths that have strong predictive power, and these powerful predictors can be used in our future studies. This work can be extended in several directions in the future: (1) larger size datasets should be collected to assess the proposed techniques; (2) more types of nodes can be added into the heterogeneous network such as symptoms, diagnostics, treatments, etc.; (3) various co-occurrence levels can be considered, such as post, thread title, or sentence; and (4) extracted powerful predictors can be used in other

classification modes such as support vector machine, naïve Bayes and so on instead of logistic regression, as logistic regression model has limitations such as sensitivity to outliers and may not be the best one to explore the power of heterogeneous features.

REFERENCES

[1] H. Wactlar, M. Pavel, and W. Barkis, "Can computer science save healthcare?," *IEEE Intelligent Systems*, vol. 26, p. 79, 2011.

[2] S. Fox. (2014, 3/6/2014). *The social life of health information*. Available: http://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/

[3] MedHelp. (1994, 1/29/2014). *Medical Information, forums and communities: About Us*. Available: http://www.medhelp.org/aboutus.htm

[4] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard, and J. H. Holmes, "Identifying potential adverse effects using the web: A new approach to medical hypothesis generation," *Journal of Biomedical Informatics*, vol. 44, p. 989, 2011.

[5] C. C. Yang, L. Jiang, H. Yang, and X. Tang, "Detecting Signals of Adverse Drug Reactions from Health Consumer Contributed Content in Social Media," *Proceedings of ACM SIGKDD Workshop on Health Informatics, Beijing, August 12, 2012.*, 2012.

[6] C. C. Yang, H. Yang, L. Jiang, and M. Zhang, "Social Media Mining for Drug Safety Signal Detection," *Proceedings of ACM CIKM International Workshop on Smart Health and Wellbeing, Maui, Hawaii, October 29, 2012.*, 2012.

[7] H. Yang and C. C. Yang, "Harnessing Social Media for Drug-Drug Interactions Detection," *Proceedings of IEEE International Conference on Healthcare Informatics, Philadelphia, PA, September 8 - 11, 2013*, 2013.

[8] J. Lazarou, B. H. Pomeranz, and P. N. Corey, "Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies," *JAMA : the journal of the American Medical Association*, vol. 279, pp. 1200-1205, 1998.

[9] S. Vilar, R. Harpaz, E. Uriarte, L. Santana, R. Rabadan, and C. Friedman, "Drug-drug interaction through molecular structure similarity analysis," *Journal of the American Medical Informatics Association*, vol. 19, p. 1066, 2012.

[10] N. P. Tatonetti, G. H. Fernald, and R. B. Altman, "A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports," *Journal of the American Medical Informatics Association : JAMIA*, vol. 19, pp. 79-85, 2012.

[11] N. P. Tatonetti, D. M. Roden, R. B. Altman, J. C. Denny, S. N. Murphy, G. H. Fernald, G. Krishnan, V. Castro, P. Yue, P. S. Tsau, and I. Kohane, "Detecting Drug Interactions From Adverse-Event Reports: Interaction Between Paroxetine and Pravastatin Increases Blood Glucose Levels," *Clinical Pharmacology & Therapeutics*, vol. 90, pp. 133-142, 2011.

[12] D. N. Juurlink, M. Mamdani, A. Kopp, A. Laupacis, and D. A. Redelmeier, "Drug-Drug Interactions Among Elderly Patients Hospitalized for Drug Toxicity," *JAMA: The Journal of the American Medical Association*, vol. 289, pp. 1652-1658, 2003.

[13] B. Astrand, E. Astrand, K. Antonov, G. Petersson, i. Naturvetenskapliga, K. Högskolan i, and i. Humanvetenskapliga, "Detection of potential drug interactions - a model for a national pharmacy register," *European journal of clinical pharmacology*, vol. 62, pp. 749-756, 2006.

[14] P. R. Obreli-Neto, W. P. Gaeti, R. K. N. Cuman, A. Nobili, A. de Oliveira Baldoni, C. M. Guidoni, D. P. de Lyra Júnior, D. Pilger, J. Duzanski, M. Tettamanti, and J. M. Cruciol-Souza, "Adverse drug reactions caused by drug-drug interactions in elderly outpatients: a prospective cohort study," *European journal of clinical pharmacology*, vol. 68, p. 1667, 2012.

[15] E. Yukawa, H. To, S. Ohdo, S. Higuchi, and T. Aoyama, "Detection of a drug–drug interaction on population-based phenobarbitone clearance using nonlinear mixed-effects modeling," *European Journal of Clinical Pharmacology*, vol. 54, pp. 69-74, 1998.

[16] E. L. Olvey, S. Clauschee, and D. C. Malone, "Comparison of Critical Drug-Drug Interaction Listings: The Department of Veterans Affairs Medical System and Standard Reference Compendia," *Clinical Pharmacology & Therapeutics*, vol. 87, pp. 48-51, 2010.

[17] P. G. M. van der Heijden, E. P. van Puijenbroek, S. van Buuren, and J. W. van der Hofstede, "On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of under-reporting on odds ratios," *Statistics in Medicine*, vol. 21, pp. 2027-2044, 2002.

[18] A. J. J. Wood, "Thrombotic Thrombocytopenic Purpura and Clopidogrel — A Need for New Approaches to Drug Safety," *The New England Journal of Medicine*, vol. 342, pp. 1824-1826, 2000.

[19] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, 2011, pp. 121-128.

[20] Y. Sun and J. Han, "Mining heterogeneous information networks: principles and methodologies," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 3, pp. 1-159, 2012.

[21] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 1019-1031, 2007.

[22] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in

SDM'06: Workshop on Link Analysis, Counter-terrorism and Security, 2006.

[23] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, "When will it happen?: relationship prediction in heterogeneous information networks," in *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, pp. 663-672.

[24] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, pp. 57-71, 2005.

[25] J. R. Katukuri, Y. Xie, V. V. Raghavan, and A. Gupta, "Hypotheses generation as supervised link discovery with automated class labeling on large-scale biomedical concept networks," *BMC genomics*, vol. 13, p. S5, 2012.

[26] R. Harpaz, H. S. Chase, and C. Friedman, "Mining multi-item drug adverse effect associations in spontaneous reporting systems," *BMC Bioinformatics*, vol. 11 Suppl 9, pp. S7-S7, 2010.

[27] B. T. Thakrar, S. B. Grundschober, and L. Doessegger, "Detecting signals of drug-drug interactions in a spontaneous reports database," *British journal of clinical pharmacology*, vol. 64, pp. 489-495, 2007.

[28] R. Leone, L. Magro, U. Moretti, P. Cutroneo, M. Moschini, D. Motola, M. Tuccori, and A. Conforti, "Identifying adverse drug reactions associated with drug-drug interactions: data mining of a spontaneous reporting database in Italy," *Drug safety : an international journal of medical toxicology and drug experience*, vol. 33, pp. 667-675, 2010.

[29] E. P. van Puijenbroek, A. C. G. Egberts, R. H. B. Meyboom, and H. G. M. Leufkens, "Signalling possible drug–drug interactions in a spontaneous reporting system: delay of withdrawal bleeding during concomitant use of oral contraceptives and itraconazole," *Br J Clin Pharmacol*, vol. 47, pp. 689-693, 1999.

[30] E. P. van Puijenbroek, A. C. G. Egberts, E. R. Heerdink, and H. G. M. Leufkens, "Detecting drug-drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs," *European Journal of Clinical Pharmacology*, vol. 56, pp. 733-738, 2000.

[31] Y. Qian, J. He, X. Ye, W. Du, J. Ren, Y. Sun, H. Wang, B. Luo, Q. Gao, and M. Wu, "A computerized system for detecting signals due to drug-drug interactions in spontaneous reporting systems," *British journal of clinical pharmacology*, vol. 69, p. 67, 2010.

[32] Y. Ji, Y. Hao, P. Dews, A. Mansour, J. Tran, R. E. Miller, and R. M. Massanari, "A Potential Causal Association Mining Algorithm for Screening Adverse Drug Reactions in Postmarketing Surveillance," *IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE*, vol. 15, pp. 428-437, 2011.

[33] A. Chan, K. Y.-L. Yap, D. Koh, X. H. Low, and Y. T. Cheung, "Electronic database to detect drug-drug interactions between antidepressants and oral anticancer drugs from a cancer center in Singapore: implications to clinicians," *Pharmacoepidemiology and Drug Safety*, vol. 20, pp. 939-947, 2011.

[34] J. E. F. Zwart-van Rijkom, E. V. Uijtendaal, M. J. ten Berg, W. W. van Solinge, and A. C. G. Egberts, "Frequency and nature of drug-drug interactions in a Dutch university hospital," *British journal of clinical pharmacology*, vol. 68, pp. 187-193, 2009.

[35] I. Segura-Bedmar, P. Martínez, and C. de Pablo-Sánchez, "Extracting drug-drug interactions from biomedical texts," *BMC Bioinformatics*, vol. 11, pp. P9-P9, 2010.

[36] I. Segura-Bedmar, P. Martínez, and C. de Pablo-Sánchez, "Combining syntactic information and domain-specific lexical patterns to extract drug-drug interactions from biomedical texts," *In: Proceedings of the ACM fourth international workshop on data and text mining in biomedical informatics (DTMBIO'10)*, pp. 49-56, 2010.

[37] I. Segura-Bedmar, P. Martínez, and C. de Pablo-Sánchez, "Using a shallow linguistic kernel for drug-drug interaction extraction," *Journal of Biomedical Informatics*, vol. 44, pp. 789-804, 2011.

[38] R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz, "Web-scale pharmacovigilance: listening to signals from the crowd," *Journal of the American Medical Informatics Association*, vol. 20, p. 404, 2013.

[39] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 243-252.

[40] Q. T. Zeng and T. Tse, "Exploring and developing consumer health vocabularies," *Journal of the American Medical Informatics Association : JAMIA*, vol. 13, pp. 24-29, 2006.

[41] C. Knox, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, D. S. Wishart, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, and C. Mak, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic acids research*, vol. 39, p. D1035, 2011.

[42] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, pp. D901-D906, 2008.

[43] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, vol. 34, pp. D668-D672, 2006.