Artificial Intelligence In Medicine 90 (2018) 42–52

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Mining heterogeneous networks with topological features constructed from patient-contributed content for pharmacovigilance

Christopher C. Yang*, Haodong Yang

College of Computing and Informatics, Drexel University, United States

ARTICLE INFO

Keywords:
Drug safety
Adverse drug reaction
Drug-drug interactions
Heterogeneous networks
Social network mining

ABSTRACT

Drug safety, also called pharmacovigilance, represents a serious health problem all over the world. Adverse drug reactions (ADRs) and drug-drug interactions (DDIs) are two important issues in pharmacovigilance, and how to detect drug safety signals has drawn many researchers' attention and efforts. Currently, methods proposed for ADR and DDI detection are mainly based on traditional data sources such as spontaneous reporting data, electronic health records, pharmaceutical databases, and biomedical literature. However, these data sources are either limited by under-reporting ratio, privacy issues, high cost, or long publication cycle. In this study, we propose a framework for drug safety signal detection by harnessing online health community data, a timely, informative, and publicly available data source. Concretely, we used MedHelp as the data source to collect patient-contributed content based on which a weighted heterogeneous network was constructed. We extracted topological features from the network, quantified them with different weighting methods, and used supervised learning method for both ADR and DDI signal detection. In addition, after identifying DDI signals, we proposed a new metric, named Interaction Ratio, to identify associated ADRs due to suspected interactions. The experiment results showed that our proposed techniques outperforms baseline methods.

1. Introduction

Drug safety, also known as pharmacovigilance, is defined by the World Health Organization (WHO) as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible drug-related problems" [1]. One important issue related to drug safety is how to detect signal of adverse drug reactions (ADRs). It has been long recognized that ADRs represent a significant world-wide health problem. In 2000, ADR was defined comprehensively by Edwards and Aronson [2] as: "an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product". In the United States, ADRs are considered to be a leading cause of death all over the country. For example, it is showed that approximately 2 million patients are affected each year by ADRs [3] and approximately 5.3% of hospital admissions are associated with ADRs [4]. The associated cost is up to about 75 billion dollars annually [5]. Therefore, how to effectively and efficiently detect ADR signals is of paramount importance for drug manufacturers, government agencies, as well as health consumers.

Drug-drug interactions (DDIs), alterations of the effects of a drug due to the recent or simultaneous use of one or more other drugs, is another significant drug safety problem. As an important subset of ADRs, DDIs may account for up to 30% of unexpected adverse drug reactions [6]. Because of common therapeutic and clinical multiple drug co-administrations, DDIs are also common and often caused by shared pathways of metabolism or intersecting pathways of drug action [7]. In some extreme cases, adverse reactions caused by DDIs have led to death. For example, drug cerivastatin caused 31 cases of fatal rhabdomyolysis prior to June 2001, 12 of which involved the concomitant use of cerivastatine and gemfibrozil [8]. DDI detection is also of great clinical importance because most interactions could result in precaution of prescription, absolute contraindications of combination use, or even drug withdrawal from market [7], and therefore has been becoming an important research area in pharmacovigilance.

Currently, there are two major approaches to pharmacovigilance process: pre-marketing review and post-marketing surveillance. Before any pharmaceutical new drugs are approved by Food and Drug Administration (FDA) for marketing, the pre-marketing review process is required. This process focuses on identifying the risk associated with drugs and the risks must be established and clearly communicated to prescribers and consumers. However, pre-marketing clinical trials are often conducted in selective patient populations, with relatively small numbers of patients, and a short duration of follow-up. Hence, the pre-

* Corresponding author.

https://doi.org/10.1016/j.artmed.2018.07.002
Received 24 January 2017; Received in revised form 12 July 2018; Accepted 18 July 2018

marketing review process is too constrained in terms of sample size, cohort biases, time spans, and financial limit to possibly identify all potential adverse reactions that may occur when the drug is used in clinical practice [9]. Furthermore, clinical trials primarily focus on ADR detection of single drugs and do not typically investigate DDIs [10,11]. Therefore, drug safety surveillance, both ADR and DDI detection, depends heavily on post-marketing surveillance to detect latent adverse reactions.

In the recent years, some traditional data sources are often mined for drug safety signal detection, such as spontaneous reporting systems, electronic health records, pharmaceutical databases, and biomedical literature. However, these sources bare their own limitations that to some extent hinder effective and confident signal detection. For instance, spontaneous reporting systems have extremely high under-reporting ratio systems [9], electronic health records are not accessible to everyone due to privacy issue, pharmaceutical databases are more focused on chemical and molecular level so that not everyone has such domain knowledge, and formally-written literature has long publishing cycle. Therefore, it is urgent to find alternative data sources to supplement drug safety surveillance. Nowadays, the advancement of Internet breeds a lot of online health communities (OHCs) such as MedHelp, WebMD, PatientsLikeMe, DailyStrength, etc. A recent survey by Pew Internet & American Life Project showed that 72% of internet users said they went online for health information in 2012, 13% of which said they began their information seeking by visiting a site that specializes in health information, like WebMD [12]. We can imagine that countless health consumers and professionals go to those OHCs frequently to either seek or offer healthcare information, experience, advice, support, and so on. Frequent visits on OHCs would inevitably produce a huge volume of health-related contents that might be even more informative than some administrative databases. If we can take good advantage of these patient-contributed content, we may be able to reveal interesting and timely knowledge, insights and patterns that may not be extracted from other data sources.

In light of the popularity of social media in Web 2.0 and Health 2.0 era, it is beneficial to explore the potential of using OHC data for drug safety signal detection. Some of our previous studies have shown that OHC data can be used for pharmacovigilance. Concretely, in [13–17], we applied association rule mining techniques directly to patient-contributed content extracted from OHCs for drug safety detection. In this study, we propose a framework to detect both ADR and DDI signals by mining the structural information of weighted heterogeneous healthcare networks built from OHC data.

2. Literature review

In this section, we provide a thorough literature review for both ADR and DDI detection. Since both of them are very important issues in pharmacovigilance, abundant efforts have been dedicated to this area. In terms of data sources used by researchers, these two topics are quite similar, i.e. four traditional data sources are often used for both ADR and DDI detection, namely spontaneous reporting systems, electronic health records, pharmaceutical databases, and biomedical literature. Although traditional data sources have been widely utilized for drug safety signal detection and abundant promising results have been shown, each of them suffers from certain limitations so that timely and effective signal detection will be hampered. More introductions of the traditional sources can be found in a recent survey [18]. This paper explores the potential of an emerging data source – patient-contributed content, so we focused on reviewing recent studies that used this type of data. Also, we provided a literature review on heterogeneous network since our method is built within this framework.

2.1. Pharmacovigilance using patient-contributed content

To the best of our knowledge, there are an increasing number of studies dedicated to pharmacovigilance using patient-contributed content from such platform in the recent years. However, the number of such studies is still limited, and more efforts need to be made.

2.1.1. ADR detection

Segura-Bedmar et al. proposed to detect drugs and adverse events from Spanish posts collected from a health social media [19]. However, this study only extracted drugs and adverse events separately rather than identified drug-ADR associations. A group from University of Pennsylvania has released a tool – Medpie – that can be used to collect a corpus of medical message board posts, anonymize the corpus, and extract information on potential adverse drug effects discussed by users [20]. Using a diabetes online community data, Liu et al. developed a framework – the AZDrugMiner system – based on statistical learning to extract adverse drug reactions in patient discussions [3]. Using DailyStrength as the source of user comments, Leaman et al. extracted adverse reactions by matching the terms in user comments with a lexicon that combined concepts and terms from four resources [21]. Further, they developed a system to automatically extract mentions of ADRs from user reviews about drugs by mining a set of language patterns [22]. Some Natural Language Processing (NLP) techniques such as linguistic dependency relations were also used for ADR detection from health-related social media [23]. Sarker and Gonzalez utilized machine learning algorithm to classify ADR assertive text segments [24]. They harnessed NLP techniques to generate useful features such as topics, concepts, sentiments, and polarities. They also showed that integration of multiple corpora can significantly improve classification performance. Liu et al. also leverage NLP techniques to extract various lexical, syntactic, and semantic features, based on which several classifier ensembles were used to distinguish between ADRs and non-ADRs in social media texts [25]. Liu and Chen developed a framework with advanced NLP techniques for ADR extraction from social media data [26]. The framework consists of three components, namely medical entity extraction, adverse drug event extraction, and report source classification. However, information extraction using NLP would miss important information captured in paraphrase or formulated in colloquial language [27]. Recently, with the advancement of word embedding, Nikfarjam et al. proposed to use sequential labeling techniques to label ADRs [28]. Specifically, they utilized Condition Random Fields (CRFs) to extract ADR concepts, and the performance could be boosted significantly by adding word-embedding-based word cluster features.

2.1.2. DDI detection

Compared with ADR detection using patient-contributed content, much less efforts have been found for DDI detection using such data. White et al. demonstrated that Internet users are able to provide early clues about DDIs via their search logs [29,30]. In their study, they conducted a large-scale study of Web search log data gathered during 2010 and paid particular attention to the specific drug pairing of paroxetine and pravastatin, whose interaction was reported to cause hyperglycemia after the time period of the online logs used in the analysis. Then they used Reporting Ratio (RR) to assess the increased chance of a user searching for hyperglycemia-related terms given that they searched for both pravastatin and paroxetine. The experiment results showed that logs of the search activities of populations of computer users can contribute to drug safety surveillance.

Saker et al. conducted a thorough review on pharmacovigilance utilizing social media data. Out of the 15 studies that were published within the last two years, as many as 11 (73.3%) used annotated data that requires a lot of expert efforts [5]. Our previous endeavors do not rely on expert annotation. We proposed to mine associations between drugs and adverse reactions and to utilize measures such as support, confidence, leverage, lift, etc. to identify FDA alerted ADRs and DDI signals [13–17]. No matter which measure we use, one crucial factor is the number of forum threads that contain direct association between drugs and ADRs. For example, in ADR detection, we are counting the number of threads that contain both a drug and an ADR whereas in DDI detection, we are counting the number of threads that contain two drugs and an ADR.

2.2. Link prediction in heterogeneous networks

An online health community itself is also a social network. Besides discovering knowledge by directly mining patient-contributed content, the structure of the network could also provide valuable information. 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. 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. Given a dataset consisting of patient-contributed content, if we can extract from it different types of nodes such as drugs, ADRs, users, diseases, etc. and identify the relationships among them, we could view our data as a heterogeneous network.

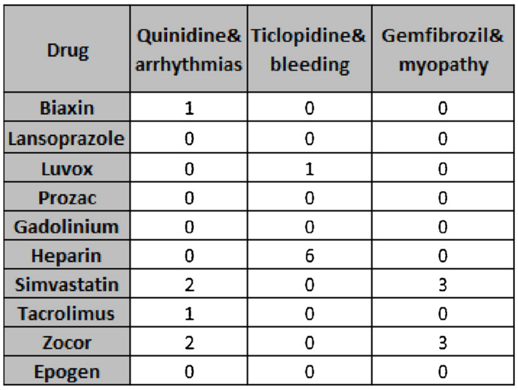
Given the problem of drug safety signal detection, we are predicting if there is an association between a drug and an ADR or between two different drugs. Such problem can be formulated as link prediction. Link prediction, dedicated to addressing the question of whether a link will be formed in the future, 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 [31]. As one of the early researchers who started working on link mining, Liben-Nowell and Kleinberg formalized link prediction problem [32]. In [32], 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.

However, most link predictions are formulated in homogenous network [32], and not until recent years are a few researchers dedicated to this problem in heterogeneous network. In [33], 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 [34]. 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.

To the best of our knowledge, there has been only a very limited number of research that uses techniques of heterogeneous network mining on OHC for knowledge discovery such as [35]. However, no studies have been found that provide a framework for drug safety signal detection by leveraging heterogeneous healthcare network. In this section, we introduce in detail the definition of heterogeneous healthcare network, the topological features extracted from such network, and the model for both ADR and DDI detection tasks in such network setting.

3. A framework for drug safety signal detection

In our previous studies, it is important to count the number of threads that contain drugs and ADRs. However, due to the openness and casualness of Internet, consumers could talk about anything, not necessary a specific drug or ADR, not to mention very rare ADRs, which makes it challenging to extract direct associations between drugs and ADRs. Fig. 1 shows part of our datasets in DDI detection experiment. Each cell represents the number of threads that contain both two drugs and the ADR. As we can see, a large number of cells are 0. It is probably because consumers are not aware of that the ADR is caused by drug-drug interactions, and they only mention one drug in the thread, thus making it difficult to extract direct associations between two drugs and an ADR. However, consumers may talk about two different drugs and the same ADR in separate threads, meaning the two drugs are linked by the same ADR. If we can extract the indirect relationships between those drugs, such relationships may help us identify ADR or DDI signals. Therefore, in this study, the idea leads us to consider our dataset as a network, and then identify drug safety signals through link mining in such network.



Drug	Quinidine& arrhythmias	Ticlopidine& bleeding	Gemfibrozil& myopathy
Biaxin	1	0	0
Lansoprazole	0	0	0
Luvox	0	1	0
Prozac	0	0	0
Gadolinium	0	0	0
Heparin	0	6	0
Simvastatin	2	0	3
Tacrolimus	1	0	0
Zocor	2	0	3
Epogen	0	0	0

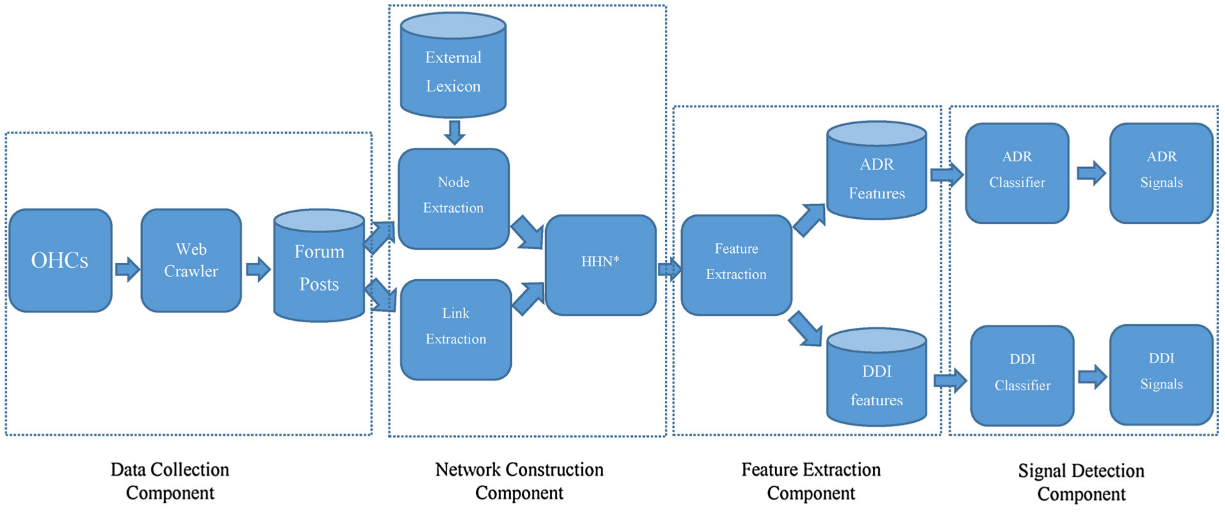
Fig. 1. Partial DDI Detection Dataset.

Recently, heterogeneous information network mining has been drawing increasing attention. Heterogeneous networks are more increasingly favored by researchers over homogeneous counterpart as they represent real-word networks in a more complete manner and carry much richer information, thus unveiling more interesting and otherwise hidden knowledge and patterns. Therefore, in this work, we propose to leverage heterogeneous network mining approach to detect drug safety signals. Fig. 2 shows our method schema that contains four primary components, namely data collection, network construction, feature extraction, and signal detection. Data collection component aims to collect data from online health communities. Depending on whether the websites provide API or not, we can develop different web crawlers for data collection. Extracted forum posts are stored in well-designed databases. Taking forum posts as input, network construction component aims to build a heterogeneous healthcare network that contains rich information. We first use external lexicons to extract different types of nodes from posts, such as drugs, ADRs, diseases, etc., and then link different nodes together based on their co-occurrence in an analysis unit. After the network is constructed, feature extraction component steps in to extract features for ADR and DDI detection respectively. At last, given extracted features, signal detection component performs binary classification to predict signals.

3.1. Heterogeneous healthcare network definition

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

An OHC can be modeled as a heterogeneous healthcare network in



*: HHN denotes Heterogeneous Healthcare Network

Fig. 2. ADR and DDI Detection Schema.

which there are a set of node types, such as *Drug*, *ADR*, *Disease*, *Treatment*, *Diagnostics*, *Users*, etc. and a set of relation types, such as *cause* or *is caused* between *Drug* and *ADR*, *treat* or *is treated* between *Treatment* and *Disease*, *use* or *is used* between *User* and *Drug*, *have* or *is had* between *User* and *Disease*, etc.

3.2. Heterogeneous healthcare network model

A network model $M_G = (\mathscr{T}, \mathscr{R})$ is a compressed representation for a heterogeneous network $G = (\mathscr{N}, \mathscr{L})$, which is a directional or non-directional graph consisting of node types $\mathscr{T}$, with links as relations from $\mathscr{R}$. Fig. 3 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*.

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 can 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. Detailed examples of bibliographic heterogeneous network mining can be found in [33,34]. However, not all heterogeneous networks contain explicit relations among different types of nodes, i.e. the semantic meaning of the relation could not be easily determined. Under such circumstances, the heterogeneous network could be represented as a non-directional network model and the relation between nodes can be the same kind of associations. For example, given a dataset of patient-contributed, it is not an easy task to accurately determine the explicit relations between nodes without using sophisticated natural language processing (NLP) techniques or thorough human annotation. However, it is still challenging to use NLP tools to analyze social media data [27] and thorough human annotation would be very time consuming. In our work, we propose to analyze a non-directional heterogeneous healthcare network (Fig. 4) that contains 4 types of nodes (namely *R*, *A*, *D*, and *U*) joined together by their co-occurrence in an analysis unit – a forum thread consisting of the original post and all following comments and replies. For R, A, and D, there will be a link between any two of them if co-occurrence is identified. For U, if there is a link between two users, there could be two scenarios: (1) one user is thread originator, another user is a commenter, and (2) both users are commenters of the same thread. We can always expand the network by adding more types of nodes and relations in the future.

3.3. Topological features

Topological features are also called structural features, which are extracted connectivity properties for pairs of objects in the networks [34]. Based on homogeneous network which only contains a specific

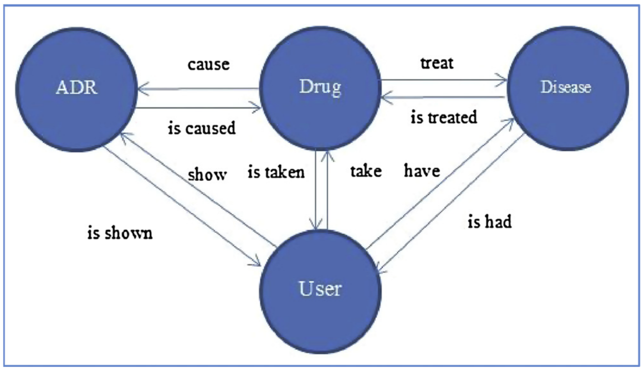


Fig. 3. Directional Network Model for Heterogeneous Healthcare Network.

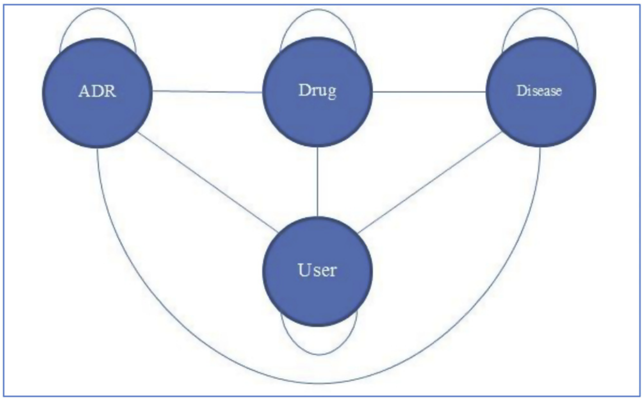


Fig. 4. Non-directional Network Model for Heterogeneous Healthcare Network.

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$_\beta$ [32] and propflow [36] or neighbor-based, such as common neighbors, Jaccard's coefficient, Adamic/Adar, preferential attachment, and SimRank$_\gamma$ [32]. However, in a heterogeneous network, as a neighbor of one node could belong to different types and a path could also flow through 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 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 our work, 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 $M_G = (\mathscr{T}, \mathscr{R})$, and is presented in the form of $T_s \xrightarrow{R_1} T_1 \xrightarrow{R_2} \ldots \xrightarrow{R_{L-1}} T_{L-1} \xrightarrow{R_L} T_d$. When the specific types of relations and directions cannot be determined between nodes, $T_1T_2{-}Path{-}L$ takes the form of $T_s{-}T_1{-}\ldots{-}T_{L-1}{-}T_d$ with links denoting associations between nodes.

3.3.1. Topological features for ADR detection

In [8], Vilar et al. proposed to detect drug-drug interaction signals through molecular structure similarity analysis. The basic assumption of their method is that if drug **a** interacts with drug **b** to cause a specific ADR, and drug **c** is structurally similar to **a**, then **c** is likely to interact with **b** to produce the same ADR. In this work, we focus on social media data. If one drug is identified to have an association with an ADR, it is possible that other drugs that are associated to this drug through other associations of diseases, users, and ADRs may cause the same ADR but not explicitly discussed in social media due to the limited medical knowledge of health consumers. In a heterogeneous healthcare network, we are going to predict the drug and ADR association through the path associations. If we know that drug **b** would cause the ADR and drug **a** and drug **b** are highly associated, it is possible that drug **a** would cause the same ADR. For example, if a user is taking drug **a** and drug **b** and she is experiencing an ADR that is a common side effect of drug **b**, there is a path of drug **a** – user – drug **b** – ADR on the heterogeneous healthcare network. It is possible that the ADR is also caused by drug **a** but there is not an explicit association in the heterogeneous network. Therefore, we construct a topological feature set by considering all 16 symmetric $RR{-}$Path with length less than 5 denoting associations between drugs and concatenating the targeted association, $R{-}A$, to its left side or right side. Since we are dealing with an non-directional heterogeneous network, we only consider concatenating $R{-}A$ to the right side of each $R{-}$Path. For example, by concatenating $R{-}D{-}U{-}D{-}R$ and $R{-}A$ together, we obtain a topological feature $-Path{-}5$: $R{-}D{-}U{-}D{-}R{-}A$. In total, including the targeted association $-A$, we have 17 topological features (Table 1).

3.3.2. Topological features for DDI detection

After we have the topological features for ADR detection, we can easily extract features for DDI detection. We just need to remove association $R{-}A$ from all features in Table 1, and the rest can be used to represent associations between two drugs. Therefore, in DDI detection problem, we extracted all the symmetric $R_sR_d{-}Path$ with length 1 to length 4, and there are 16 such paths in total given 4 different types of nodes R, A, D, and U, such as R-R, R-D-R, R-A-D-A-R, etc. (Table 2) The link existing between two nodes specifies the co-occurrence association between them.

Table 1
Topological Features for ADR Detection: $RA{-}Path{-}1$ to $RA{-}Path{-}5$ Denoting Associations between Drug and ADR.

Path	Length
R – A	1
R – R – A	2
R – A – R – A	3
R – D – R – A	3
R – U – R – A	3
R – A – A – R – A	4
R – D – D – R – A	4
R – U – U – R – A	4
R – A – A – A – R – A	5
R – A – D – A – R – A	5
R – A – U – A – R – A	5
R – D – A – D – R – A	5
R – D – D – D – R – A	5
R – D – U – D – R – A	5
R – U – A – U – R – A	5
R – U – D – U – R – A	5
R – U – U – U – R – A	5

Table 2
Topological Features for DDI Detection: $R_sR_d{-}Path{-}1$ to $R_sR_d{-}Path{-}4$ Denoting Associations between Drug and Drug.

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

3.4. Weighted heterogeneous healthcare network and feature quantification

There are two possible types of heterogeneous networks: non-weighted and weighted. A non-weighted network means the links do not carry weight information whereas a weighted one is a network in which the links between any pairs of nodes have weights assigned to them. In most real-world networks, the strength of associations between different pairs of nodes is not entirely the same when links exist between them. For example, given a heterogeneous healthcare network in Fig. 5, the number next to the link denotes the link frequency. If we don't consider the weight, $PC(R_1, A_3)$ under path $R{-}A{-}R{-}A$ is 2, and $PC(R_1, A_3)$ under path $R{-}D{-}R{-}A$ is also 2. However, to some extent, path $R{-}D{-}R{-}A$ is more interesting in this case because the nodes under the path co-occurred more frequently. Therefore, drug safety detection based on a weighted heterogeneous network could achieve better performance by considering the paths with different strength of associations. In this study, we propose to use three different metrics to weight the network: link frequency (LF), link leverage (LV), and link lift (LT).

Let l_{ab} be a link between nodes T_a and T_b and considering a thread of an OHC forum as an analysis unit, LF is the number of threads in which nodes T_a and T_b co-occur. Leverage and lift are often used in association rule mining, one of the most important and well researched techniques of data mining. Association rule mining was first introduced by Agrawal

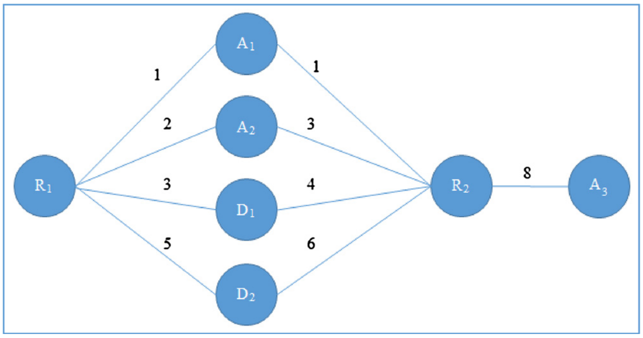


Fig. 5. An Example of A Heterogeneous Healthcare Network.

et al. when they were trying to identify significant purchasing pattern from a large database of consumer transactions [37]. This technique aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories and is widely used in various areas such as telecommunication networks, market and risk management, inventory control, etc [38]. Mathematically, let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of items. Let X, the task–relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \emptyset$, where both A and B are a set of items, which is referred to as an itemset. Leverage and lift are often used to measure the interestingness and impressiveness of an association. In this study, we use these two to measure the association strength between a pair of nodes as represented by a link in a heterogeneous network. Give a link l_{ab} between nodes T_a and T_b, LV and LT are defined respectively as:

$$LV(l_{ab}) = support(l_{ab}) – support(T_a) \times support(T_b)$$

$$LT(l_{ab}) = \frac{support(l_{ab})}{support(T_a) \times support(T_b)}$$

where

$$support(l_{ab}) = \frac{LF(l_{ab})}{Z}$$

$$support(T_a) = \frac{NF(T_a)}{Z}$$

$$support(T_b) = \frac{NF(T_b)}{Z}$$

where $NF(T_a)$ and $NF(T_b)$ denote node frequency (number of threads that contain the node) of node T_a and T_b respectively, and Z is the total number of threads in the dataset. For both LV and LT, the higher the value is, the more importance the link will be.

There are several ways of quantifying the topological features in a heterogeneous network. In [34], Sun et al. proposed to use such measures as path count, normalized path count, random walk, and symmetric random walk to quantify the features. There measures could also be applied into our non-directional heterogeneous network with some modifications. In this work, after adding weight to the network, we propose to use Weighted Path Count (*WPC*) to quantify the extracted topological features. Given a $T_s T_d–Path–L$, the *WPC* is defined as:

$$WPC(T_s, T_d) = \frac{1}{L} \sum_{P} \sum_{i=1}^{L} w(n_i, n_{i+1})$$

where P denotes a specific path, L is the length of P, n_i and n_{i+1} are two directly connected nodes following P, and $w(n_i, n_{i+1})$ is weight of the corresponding link connecting node n_i and n_{i+1}. Take the network in Fig. 5 as an example. If we use link frequency as the weight, (1) under path $R–A–R–A$, $WPC(R_1, A_3) = \frac{1}{3}(1 + 1 + 8 + 2 + 3 + 8) = \frac{23}{3}$, and (2) under path $R–D–R–A$, $WPC(R_1, A_3) = \frac{1}{3}(3 + 4 + 8 + 5 + 6 + 8) = \frac{34}{3}$. In this way, we can tell that for drug R_1 and ADR A_2 path $R–D–R–A$ has stronger association than $R–A–R–A$.

3.5. Drug safety signal detection model

We model drug safety signal detection as a binary classification problem.

3.5.1. ADR detection

Given a drug-ADR pair, we use a classification model to label them as either “1″ (drug causes the ADR) or “0″ (drug does not cause the ADR) based on their quantified topological features extracted from the heterogeneous healthcare network.

3.5.2. DDI detection

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. Various classification models could be used such as Logistical Regression (LR) [39], Naïve Bayes (NB) [39], Support Vector Machine (SVM) [40], etc.

3.6. Associated adverse reaction for DDI detection

3.6.1. Association rule mining metrics

Above techniques for DDI detection could only predict if two drugs are interacting through mining the weighted heterogeneous healthcare network. However, after such signals are detected, we are also interested in identifying what consequent adverse reaction would be caused due to interaction, which would lead to better further investigation. Therefore, we also propose to apply association mining to associated ADR detection. Three metrics are often used to capture the association strength, namely confidence, leverage, and lift. The metrics here are different from those in section 3.4 that only measure the strength of two nodes. Instead, they are trying calculate the association strength of two drugs causing an ADR, and are defined as follows.

$$confidence((R_1 \cup R_2) \Rightarrow A) = \frac{support((R_1 \cup R_2) \Rightarrow A)}{support(R_1 \cup R_2)}$$

$$\begin{aligned} leverage((R_1 \cup R_2) \Rightarrow A) &= support((R_1 \cup R_2) \Rightarrow A) \\ &\quad - support((R_1 \cup R_2) \times support(A) \end{aligned}$$

$$lift((R_1 \cup R_2) \Rightarrow A) = \frac{support((R_1 \cup R_2) \Rightarrow A)}{support((R_1 \cup R_2) \times support(A)}$$

where

$$support((R_1 \cup R_2) \Rightarrow A) = \frac{count(R_1 \cup R_2 \cup A)}{total\ count}$$

where $count(R_1 \cup R_2 \cup A)$ is the number of threads that contain R_1, R_2 and A, and *total count* is the total number of threads in the whole dataset.

Confidence determines the extent to which the appearance of $R_1 \cup R_2$ implies the appearance of. Both leverage and lift consider the correlation between $R_1 \cup R_2$ and A. Leverage indicates the proportion of threads that contain $R_1 \cup R_2 \cup A$ by excluding probability that if $R_1 \cup R_2$ and A are independent with each other whereas lift considers the ratio of those two. For example, note that lift can also be written as:

$$\begin{aligned} lift((R_1 \cup R_2) \Rightarrow A) &= \frac{support((R_1 \cup R_2) \Rightarrow A)}{support((R_1 \cup R_2) \times support(A)} \\ &= \frac{P(R_1, R_2, A)}{P(R_1, R_2) \times P(A)} = \frac{P(A|R_1, R_2)}{P(A)} \end{aligned}$$

Large values indicate that the occurrence of the $(R_1 \cup R_2) \Rightarrow A$ association has unlikely occurred by chance. Roughly, $lift((R_1 \cup R_2) \Rightarrow A) = 1$ indicates that the two drugs and ADR are statistically independent with each other, $lift((R_1 \cup R_2) \Rightarrow A) > 1$ that the drugs and ADR are positively correlated, and $lift((R_1 \cup R_2) \Rightarrow A) < 1$ that they are negatively correlated. For both leverage and lift, the higher the values are, the stronger the DDIs signals are.

3.6.2. Interaction ratio

Although our previous research has demonstrated that the three measures, especially leverage and lift could effectively detect ADRs reported by FDA [13,14,16,17], we were dealing with a single drug and its adverse reaction. Also, there are some limitations about them. For example, in DDI detection, confidence could be very low because there may be very few consumers mentioning both drugs and associated ADR as they may not be aware of fact that the ADR is caused by drug-drug interaction. Also, leverage could be even negative that makes it very difficult to interpret the results. Therefore, in order to effectively identify associated ADR, we propose a new metric that is called Interaction Ratio and defined as:

$$IR_c((R_1 \cup R_2) \Rightarrow A) = \frac{confidence((R_1 \cup R_2) \Rightarrow A)}{confidence(R_1 \Rightarrow A) \times confidence(R_2 \Rightarrow A)}$$

where IR_c means Interaction Ratio, subscript c denotes confidence on which this formula is based, R_1 is one of the drugs in our collected dataset, R_2 is a drug which could interact with D_1 to generated ADR R, $confidence(R_1 \Rightarrow A)$ is the confidence value that A is caused by R_1, and $confidence(R_2 \Rightarrow A)$ is the confidence value that A is caused by R_2. The rationale behind this metric is that if an ADR is caused by the interaction of R_1 and R_2 rather than only by R_1 or R_2 alone, the value of $confidence((R_1 \cup R_2) \Rightarrow A)$ should be higher than that of $confidence(R_1 \Rightarrow A)$ or $onfidence(R_2 \Rightarrow A)$, and the division would boost the value of $IR_c((R_1 \cup R_2) \Rightarrow A)$.

4. Experiment

4.1. Data collection

In this study, MedHelp.org, a pioneer in online health communities, is used as the source of health-contributed contents. We focus on the drug section, which is one of the most important and popular components in MedHelp. To effectively detect drug safety signals, the drugs should bear active discussion. 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 threads. 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[1], 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. In total, there are 16,344 threads.

4.2. Network construction

To construct the heterogeneous healthcare network, we need to extract different types of nodes and their relations. In this work, we focus on four types of nodes, namely R, A, D, and U, and external lexicons are used to extract them. 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 to enrich our dataset for DDI detection. For A, we focus on 10 ADRs, namely Blurred Vision, Cancer, Depression, Diarrhea, Heart Disease, Hypertension, Kidney Disease, Skin Discoloration, Stroke, Suicide. Some of the drugs collected were alerted by FDA to cause some adverse reactions. For example, Lansoprazole and Heparin are both alerted to cause Diarrhea; Luvox and Prozac are both alerted to cause suicidal thoughts. Therefore, a part of our ADR list comes from FDA's official alert, whereas the rest is based on drug labeling revisions. The drug labeling revisions provide new ADRs added on the labels of drugs after the drugs are released. The labeling revision information could be found on FDA's website "Drugs@FDA". Then we use Consumer Health Vocabulary (CHV) Wiki to build our ADR lexicon. More introduction of CHV can be found in [41]. CHV reflects the difference between consumers and professionals in expressing health concepts and helps to bridge this vocabulary gap. Therefore, high quality CHV is able to help with capturing more consumers' expressions and better extracting ADR terms. Some studies are dedicated to expanding CHV by using social media data [42–44]. For D, we search for diseases that are treated by each of the 20 drugs in SIDER database to construct our disease lexicon. SIDER contains information such as adverse drug reactions and diseases on marketed medicines, and the information is extracted from public documents and package inserts [45]. At last, there are 205 diseases in total, such as Bipolar Disorder, Hyperactivity Disorder, Hypercholesterolaemia, and so on. For U, we extract all user names from each thread. The dataset is de-identified before conducting the experiment. For links, we treat our network as non-directional, and two nodes are linked together if they co-occur in the same thread.

In order to exclude the nodes and links that appear in the heterogeneous healthcare network rarely, we only retain the nodes and links with frequency larger than 15. After filtering, there are 511 nodes and 4378 links in our final network with density being 0.034. Then we weigh the network using leverage and lift respectively, and quantify the extracted features for both ADR and DDI detection with weighted path count.

4.3. Gold standard

4.3.1. ADR detection

As mentioned earlier, current post-marketing surveillance in United States primarily depends on FDA's FAERS system, and alerts will be released on FDA's website[2] if the ADR is confirmed after investigation. Out of the 20 drugs collected, 8 of them were alerted by FDA to cause some adverse reactions. Some of the drugs share the same alerted ADRs. For example, Lansoprazole and Heparin are both alerted to cause Diarrhea; Luvox and Prozac are both alerted to cause suicidal thoughts. We used 6 ADRs alerted by FDA to construct a part of our gold standard for evaluating the proposed techniques. Another part of the gold standard is based on drug labeling revisions. The drug labeling revisions provide new ADRs added on the labels of drugs after the drugs are released. The labeling revision information could be found on FDA's website "Drugs@FDA[3]". In this study, we used 4 other ADRs as another part of our gold standard, and we have 10 ADRs in total, namely Blurred Vision, Cancer, Depression, Diarrhea, Heart Disease, Hypertension, Kidney Disease, Skin Discoloration, Stroke, and Suicide.

4.3.2. DDI detection

In the constructed 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 would interact with each other. Therefore, an external database DrugBank is used to set up the gold standard. DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information [46].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]

[1] http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm111085.htm

[2] http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm111085.htm

[3] http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm

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

engine. If two drugs are reported to have interaction, we label the pair of drug nodes as "1″, and otherwise "0″. For example, Biaxin is reported by interact with Quinidine to cause Arrhythmias, Simvastatin is reported to interact with Gemfibrozil to cause Myopathy, etc. At last, 20 positive (drug pairs labeled as "1″) are identified. This information is also used as gold standard for associated adverse reaction detection.

4.4. Evaluation

To evaluate the effectiveness of the proposed techniques, we set up baseline for both ADR and DDI detection.

4.4.1. ADR detection

In our previous study [17], when using association rule mining to detect ADR signals, leverage has the best performance. Therefore, we first compare supervised techniques with unsupervised counterpart, namely association rule mining, and we use F1 score to compare their performance. Second, we set up a comparison between weighted and non-weighted heterogeneous healthcare networks: specifically, we construct a non-weighted heterogeneous network, extract all the 17 $RA-Path-L$ and used path count to quantify them, and then perform classification on it. We use both F1 score and area under the ROC curve (AUC) to evaluate the performance.

4.4.2. DDI detection

Here we have two steps for DDI detection. We first detect if two drugs interact, and then identify the associated adverse reactions due to interaction. Therefore, we develop baselines for each step.

For the first step, we set up two baselines for comparison:

(1) Comparison between heterogeneous and homogeneous networks: We compare the performances between heterogeneous and homogeneous networks. Specifically, we constructed an unweighted homogeneous network that only contains one type of node – drug. We counted the number of path instances for each drug pair 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. Then we perform classification on the dataset.

(2) Comparison between weighted and unweighted heterogeneous networks: We also compare the performances between weighted and unweighted heterogeneous networks. Specifically, we constructed an unweighted heterogeneous network, extracted all the 16 symmetric $\mathrm{R_sR_d}$−Path−L and used path count to quantify them. Then we perform classification on the dataset.

We use sensitivity, specificity, and AUC to evaluate the proposed methods.

For the second step, we compare our proposed techniques with Reporting Ratio (RR) that were used in [29] to detect DDI signal and associated ADR from Web search log data. They paid particular attention to the specific drug pairing of paroxetine and pravastatin, whose interaction was reported to cause hyperglycemia. Fig. 6 shows the Venn diagram of different user groups in the analysis conducted in [29]. *RR* is defined as observed/expected, i.e. $(a/b)/(c/d)$. Observed is defined as the fraction of users who searched for both pravastatin and paroxetine (b) who also queried for hyperglycemia symptoms (a), and expected is defined as the fraction of users who searched for pravastatin (d_1) who also searched for hyperglycemia symptoms (c_1), or the fraction of users who searched for paroxetine (d_2) who also searched for hyperglycemia symptoms (c_2) [29]. In this work, we replace *Pravastatin searchers* with R_1, *Paroxetine searchers* with R_2, and *All hyperglycemia searchers* with A. Therefore, $RR(R_1)$ is defined as $(a/b)/(c_1/d_1)$, where a denotes the number of threads that contain R_1, R_2 and A, b is the number of threads that contain both R_1 and R_2, c_1 is the number of threads that contain both R_1 and A, and d_1 is the number of threads that contain R_1. $RR(R_2)$ is defined in a similar way. Sensitivity, specificity, and AUC are also used to compare the performance of different metrics.

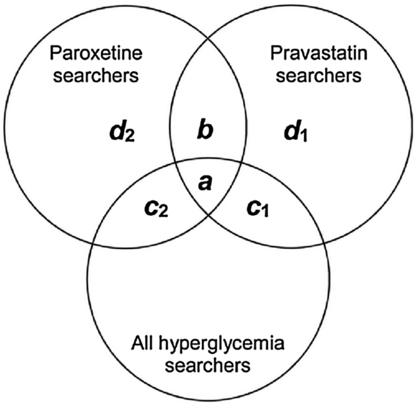


Fig. 6. Venn Diagram Showing the Different User Groups [29].

4.5. Results and discussion

4.5.1. ADR detection

During the experiment, we found that our dataset is highly imbalanced, and the ratio of positive (labeled as 1) and negative (labeled as 0) drug-ADR pairs is approximately 1:14. Therefore, we use undersampling to build a training dataset with an equal sized set of positive and negative pairs. Then we perform 5-fold cross validation using multiple classifiers, i.e. LR, NB, and SVM. The undersampling process is repeated 1000 times and the final performance is averaged.

When applying association rule mining with leverage as metric to our dataset, we can achieve 0.20 in F1 score. As shown in Table 3, no matter what weighting schema and classifier we use, the proposed supervised leaning techniques outperformed unsupervised one, especially when we use leverage and lift to weight the heterogeneous healthcare network. It is because supervised learning technique is able to recognize patterns of true positive and true negative signals by analyzing the proposed features. It is more powerful in predicting unseen new examples.

Figs. 7 and 8 show the performance comparisons of different network weighting schemas using different classifiers in terms of F1 score and AUC score respectively. We can observe that, in all scenarios, heterogeneous network weighted by leverage (Hete_LV) and heterogeneous network weighted by lift (Hete_LT) outperform non-weighted network and network weighted by link frequency except AUC score using Naïve Bayes as classifier.

We also conducted ANOVA analysis to see, for each classifier, if there is any significant difference between different network weighting schemas in terms of F1 score and AUC score. Both Welch procedure ($p =$ .000) and Brown-Forsythe procedure ($p =$ .000) showed that a statistically significant difference exists in terms of both F1 score and AUC no matter which classifier is used. Furthermore, for both F1 and AUC scores under all three classifiers, Games-Howell post-hoc tests demonstrated that (1) Hete_LV is statistically significant higher than all other three network weighting schemas ($p =$ .000 for all comparisons) except its AUC score is significantly lower than Hete_Non-weighted ($p =$ .000) and Hete_LT ($p =$ .000) when Naïve Bayes is used; (2) Hete_LT is significantly higher than both Hete-Non-weighted and Hete_LF ($p =$ .000 for all comparisons) except its AUC score is significantly lower than Hete_Non-weighted ($p =$ .000) when NB is used. Last but not the least, ANOVA analysis under network weighting schema Hete_LV showed that (1) for F1 score, NB is significantly higher than both LR ($p =$ .000) and SVM ($p =$ .000), and SVM is significantly higher than LR ($p =$ .000); (2) for AUC score, SVM is significantly higher than both LR ($p =$ .000) and NB ($p =$ .000), and LR is significantly higher than NB ($p =$ .000).

The results demonstrate that (1) the performance of different classifiers varies under different evaluation scenarios, and (2) leverage- and

Table 3
Comparison between Supervised Learning and Unsupervised Learning Using F1 Score.

	Hete_Non-Weighted	Hete_LF	Hete_LV	Hete_LT
LR	0.35	0.41	0.95	0.97
NB	0.75	0.43	0.98	1.00
SVM	0.45	0.41	0.97	0.92

Hete_Non-Weighted: heterogeneous network with no weight.
Hete_LF: heterogeneous network weighted with link frequency.
Hete_LV: heterogeneous network weighted with leverage.
Hete_LT: heterogeneous network weighted with lift.

lift-weighted heterogeneous healthcare networks are generally more effective in ADR detection than non-weighted and frequency-weighted heterogeneous network. Link frequency is proportional to its support value ($support(l_{ab}) = \frac{LF(l_{ab})}{Z}$), which is not enough to represent the information carried by links. For example, considering a link R – A, its support value could be very small, but it does not mean that this link is trivial, because this ADR could be one of the rare ADRs that would be caused by the drug. One limitation of support lies in the fact that it would work well when the ADR of the drug appear frequently in the dataset. However, health consumers discuss diverse aspects of drugs in online forum, such as drug dosage, drug prescription, concomitant use of different drugs, and so forth. It is very likely that threads that are related to the specific ADR are only a small portion of the total threads, especially for those rare ADRs. Leverage and lift could be used to address this problem because they incorporate the support of the ADR in the dataset. Both leverage and lift measure the strength of a link not only by looking at its support but also the correlation between the two nodes. Leverage measures the difference between the proportion of threads containing both nodes above those expected if the two nodes were independent of each other whereas lift calculate the ratio of these two. Therefore, both leverage- and lift-weighted heterogeneous networks perform better than frequency-weighted one. It is also worth noting that in ADR detection, we only utilized undersampling for handling imbalanced dataset. However, since undersampling only randomly select a small portion of data from the majority class, it would potentially remove very important data examples, thus hurting the training model. Therefore, in DDI detection, we also applied other methods for handling imbalanced dataset.

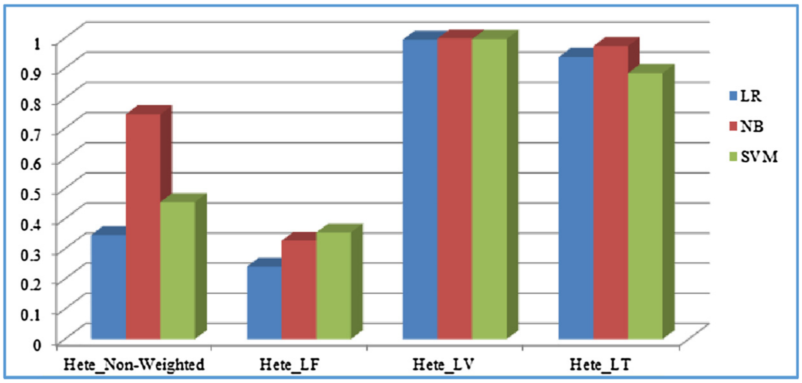


Fig. 7. F1 Score Comparison.

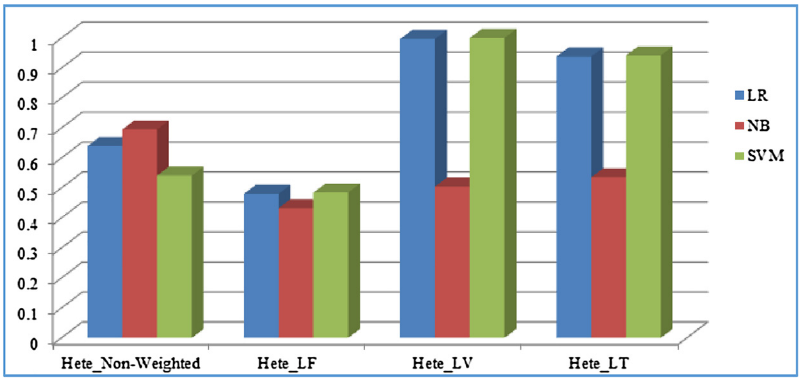


Fig. 8. AUC Score Comparison.

4.5.2. DDI detection

4.5.2.1. Interaction detection. Compared with adverse drug reactions, drug-drug interactions are more scarce [47], so we also expect our dataset to be highly imbalanced. Indeed, we found that the ratio of positive (drug pairs labeled as 1) and negative (drug pairs labeled as 0) examples is approximately 1:12 in our dataset. Besides undersampling, we also propose to use oversampling, MetaCost [48], and AdaBoost [49,50] techniques to learn from the imbalanced data.

Table 4 demonstrates the sensitivity and specificity of different methods in different network setting. Homo_NW and Hete_NW denote non-weighted homogeneous network and non-weighted heterogeneous network respectively, whereas Hete_LV and Hete_LT denote heterogeneous network weighted by leverage and lift respectively. As we can see, except AdaBoost that classified all examples as "no interaction", for the other three methods, Heterogeneous network generally performed better than the homogeneous counterpart. In particular, compared horizontally, leverage-weighted heterogeneous network has the best

Table 4
Sensitivity and Specificity of Different Methods in Different Network Settings.

	Sensitivity				Specificity			
	Homo_NW	Hete-NW	Hete_LV	Hete_LT	Homo_NW	Hete-NW	Hete_LV	Hete_LT
Undersampling	0.63	0.77	***0.83***	0.80	0.73	0.68	***0.74***	0.70
Oversampling	0.69	0.76	***0.86***	0.83	0.72	0.73	***0.78***	0.73
MetaCost	0.71	0.74	***0.95***	0.85	0.61	0.65	***0.66***	0.63
AdaBoost	0	0.23	0.2	0.30	1	0.97	0.97	0.96

Table 5
AUC Score of different Methods in Different Network Settings.

	Homo_NW	**Hete-NW**	**Hete_LV**	**Hete_LT**
Undersampling	0.73	0.78	***0.81***	0.78
Oversampling	0.76	0.85	***0.87***	0.85
MetaCost	0.73	0.81	***0.84***	0.80
AdaBoost	0.76	0.78	***0.87***	0.83

Table 6
Sensitivity, Specificity, and AUC of Different Metrics in Associated ADR Detection.

	Sensitivity	**Specificity**	**AUC**
$\boldsymbol{RR(R_1)}$	0.43	0.84	0.63
$\boldsymbol{RR(R_2)}$	0.43	0.82	0.63
Confidence	0.14	***0.96***	0.55
Leverage	0.48	0.82	0.65
Lift	0.48	0.82	0.65
$\boldsymbol{IR_c}$	***0.52***	0.83	***0.68***

performance for undersampling, oversampling, and MetaCost in terms of both sensitivity and specificity. Lift-weighted heterogeneous network also performed better than non-weighted heterogeneous network in sensitivity and comparable in specificity. Compared vertically, MetaCost under leverage-weighted heterogeneous network outperformed other methods in sensitivity whereas oversampling under the same network setting has the highest specificity. Table 5 illustrates the AUC scores of different methods in various network settings, which, again, shows that both non-weighted and weighted heterogeneous networks outperformed homogeneous counterpart, and leverage-weighted heterogeneous healthcare network has the highest AUC in all methods.

The results suggest that (1) heterogeneous network is more effective in DDI signal detection and (2) leverage- and lift-weighted heterogeneous network perform better than non-weighted one and leveraged-weighted network has the best performance. It suggests that heterogeneous healthcare networks carry richer information because it incorporates various types of nodes that could better represent the real-world network. Also, weighted links are more informative than binary representation. We also found that although AdaBoost method could achieve reasonable specificity and AUC score, it is not able to recall any true positive signals. Therefore, only AdaBoost may not achieve satisfactory performance and more techniques should be incorporated. For example, cost-sensitive methods and adaptive boosting could be integrated for improving the performance.

4.5.2.2. Associated ADR detection. Here we are more interested in recalling as many true positive examples as possible. Therefore, we apply MetaCost to our original leverage-weighted heterogeneous healthcare network because this combination has the highest cross validated sensitivity and fair specificity according to Table 4 and we are more interested in recalling as more true signals as possible. Our dataset consists of 20 positive and 233 negative drug pairs. At last, 18 true positive pairs out of 20 and 64 false positive pairs are detected by the algorithm. We believe that as we expand our dataset in the future, the proposed techniques are able to save a fair amount of computational resource without sacrificing too much sensitivity.

We set up a threshold for each metric by their characteristics and domain knowledge. Concretely, we assign 1 to both $RR(R_1)$ and $RR(R_2)$ because RR greater than 1 means a higher odds of interaction [51]; 0.3 to triad confidence based on expert opinion; 0 to leverage because leverage greater than 0 denotes a positive correlation between two drugs and ADR [17]; 1 to lift because lift greater than 1 also denotes a positive correlation between two drugs and ADR [17]; and 20 to IR (confidence) based on expert opinion. Table 6 shows the performance of different metrics in detecting associated ADRs due to drug-drug interaction. As we can see, association mining metrics except confidence have better performance than baseline methods. In particular, our proposed technique, IR_c, achieved the highest AUC and sensitivity and comparable specificity. It is also worth noting that although IR_c is based on confidence, it significantly outperformed confidence in AUC and sensitivity and has a slightly lower specificity

5. Conclusion

The development of Web 2.0 and Health 2.0 technologies leads the booming of OHCs such as MedHelp, WebMD and so on. Such platforms are not only empowering individuals to play a substantial role in their own health, but also generating informative patient-contributed content that can be utilized to mine timely and useful knowledge, thus providing automated insights and discovery. Since pharmacovigilance, namely ADRs and DDIs, represents a serious health problem all over the world, how to detect drug safety signals has drawn many researchers' attention and efforts. Currently, the methods proposed to detect ADR and DDI signals are mainly based on traditional data sources such as spontaneous reporting data, electronic health records, pharmaceutical databases, and biomedical literature. However, these data sources are either limited by under-reporting ratio, privacy issues, high cost, or long publication cycle. In this study, we propose a framework for drug safety signal detection by harnessing online health community data, a timely, informative, and publicly available data source. We used MedHelp as data source to collect patient-contributed content based on which a weighted heterogeneous network was constructed. Then we extracted topological features from the network, quantified them with different weighting methods, and used supervised learning method for both ADR and DDI signal detection. In addition, after identifying DDI signals, we proposed a new metric, named Interaction Ratio, to identify associated ADRs due to suspected interactions. The experiment results show that our proposed techniques outperform the baseline methods. Specifically, in ADR detection, supervised techniques outperformed unsupervised counterpart with association rule mining, and leverage- and lift-weighted networks could achieve better results than non-weighted network within supervised methods. In DDI detection, besides undersampling, we also utilized oversampling, cost-sensitive, and ensemble methods to deal with imbalanced dataset issue. The experiment results showed that, again, leverage-weighted network has better performance than both homogeneous network and non-weighted heterogeneous network. The advantage of heterogeneous-network-based approach is that it captures both direct and indirect relations among different types of nodes. Furthermore, compared with homogeneous and unweight network, weighted heterogeneous healthcare networks carry much rich information that could better represent the real-world networks and help us better identify drug safety signals. There are also some limitations in this study. First of all, the construction of network depends on extracted nodes, among which extraction of ADRs, drugs, and diseases highly relies on the quality of pre-compiled vocabularies. If some rare or undiscovered ADR or diseases are not included in the vocabulary, then we are not able to make the prediction. Second, social media data abounds with misspellings, omissions, newly invented words, etc. that could also affect the quality of extracted entities. Therefore, a very critical digression research would be focused on how to extract verbatim of adverse reactions, drugs, and diseases from patient-contributed contents. In the future, this work could be extended in several directions: (1) increase the scale of dataset and include additional types of nodes into the heterogeneous healthcare network such as diagnostics, symptoms, treatments, and (2) consider asymmetric topological features and other quantification methods in mining heterogeneous networks.

Acknowledgements

This work was supported in part by the National Science Foundation under the Grant IIS-1741306, IIS-1650531, and DIBBs-1443019. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] WHO. (2002, 6/18/2014). The Importance of Pharmacovigilance. Available: http://apps.who.int/medicinedocs/pdf/s4893e/s4893e.pdf.

[2] Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet 2000;356(9237):1255–9.

[3] X. Liu and H. Chen, "AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums," in Smart Health, ed: Springer, 2013, pp. 134-150.

[4] Kongkaew C, Noyce PR, Ashcroft DM. Hospital admissions associated with adverse drug reactions: a systematic review of prospective observational studies. Ann Pharmacother 2008;vol. 42(7):1017–25.

[5] Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, Upadhaya T, Gonzalez G. Utilizing social media data for pharmacovigilance: a review. J Biomed Inform 2015;54:202–12.

[6] Tatonetti NP, Fernald GH, Altman RB. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. J Am Med Inform Assoc 2012;19(1):79–85.

[7] Tatonetti NP, Roden DM, Altman RB, Denny JC, Murphy SN, Fernald GH, Krishnan G, Castro V, Yue P, Tsau PS, Kohane I. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clin Pharmacol Ther 2011;90(1):133–42.

[8] Vilar S, Harpaz R, Uriarte E, Santana L, Rabadan R, Friedman C. Drug-drug interaction through molecular structure similarity analysis. J Am Med Inform Assoc 2012;19(6):1066.

[9] van der Heijden PGM, van Puijenbroek EP, van Buuren S, van der Hofstede JW. On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of under-reporting on odds ratios. Stat Med 2002;21(14):2027–44.

[10] Olvey EL, Clauschee S, Malone DC. Comparison of critical drug-Drug interaction listings: the department of veterans affairs medical system and standard reference compendia. Clin Pharmacol Ther 2010;87(1):48–51.

[11] van Puijenbroek EP, Egberts ACG, Heerdink ER, Leufkens HGM. Detecting drug-drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs. Eur J Clin Pharmacol 2000;56(9):733–8.

[12] S. Fox and M. Duggan. (2013, 5/20/2013). Health Online 2013. Available: http://www.pewinternet.org/2013/01/15/health-online-2013/.

[13] Yang CC, Jiang L, Yang H, Tang X. Detecting signals of adverse drug reactions from health consumer contributed content in social media. Presented at the ACM SIGKDD Workshop on Health Informatics. 2012.

[14] Yang CC, Yang H, Jiang L, Zhang M. Social media mining for drug safety Signal detection. Proceedings of the ACM CIKM International Workshop on Smart Health and Wellbeing. 2012. p. 33–40.

[15] Yang H, Yang CC. Harnessing social media for drug-drug interactions detection. Proceedings of the IEEE International Conference on Healthcare Informatics. 2013. p. 22–9.

[16] C. C. Yang, H. Yang, and L. Jiang, Postmarketing Drug Safety Surveillance Using Publicly Available Health-Consumer-Contributed Content in Social Media ACM Transactions on Management Information Systems (TMIS), 5 (1), 2:1-2:21, 2014.

[17] Yang H, Yang CC. Using health consumer contributed data to detect adverse drug reactions by association mining with temporal analysis. ACM Tran Intell Syst Technol (TIST) 2015;6(4). 55, 27 pages.

[18] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, and C. Paris, "Text and data mining techniques in adverse drug reaction detection," ACM Computing Surveys To appear. View in Article, 2015.

[19] Segura-Bedmar I, Revert R, Martínez P. Proceedings of the Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL. 2014. p. 106–15.

[20] Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, Leonard CE, Holmes JH. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. J Biomed Inform 2011;44(6):989.

[21] Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010;2010:117–25.

[22] Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of adverse drug reactions from user comments. AMIA Annu Symp Proc 2011;2011:1019–26.

[23] Yates A, Goharian N, Frieder O. Proceedings of the 2013 ACM SIGIR Workshop on Health Search and Discovery. 2013.

[24] Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 2015;53:196–207.

[25] Liu J, Zhao S, Zhang X. An ensemble method for extracting adverse drug events from social media. Artif Intell Med 2016;70:62–76.

[26] Liu X, Chen H. A research framework for pharmacovigilance in health social media: identification and evaluation of patient adverse drug event reports. J Biomed Inform 2015;58:268–79.

[27] Denecke K. Extracting medical concepts from medical social media with clinical NLP tools: a qualitative study. Presented at the The Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing. 2014.

[28] Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015;22(3):671–81.

[29] White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc 2013;20(3):404–8.

[30] White RW, Harpaz R, Shah NH, DuMouchel W, Horvitz E. Toward enhanced pharmacovigilance using patient-generated data on the internet. Clin Pharmacol Ther 2014;96(2):239–46.

[31] Sun Y, Han J. Mining heterogeneous information networks: principles and methodologies. Synth Lect Data Min Knowl Discov 2012;3(2):1–159.

[32] LibenNowell D, Kleinberg J. The link-prediction problem for social networks,". J Am Soc Inf Sci Technol 2007;58(7):1019–31.

[33] Sun Y, Barber R, Gupta M, Aggarwal CC, Han J. Co-author relationship prediction in heterogeneous bibliographic networks. Proceedings of the Advances in Social Networks Analysis and Mining (ASONAM). 2011. p. 121–8.

[34] Sun Y, Han J, Aggarwal CC, Chawla NV. Proceedings of the Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. 2012. p. 663–72.

[35] Kumar A, Zhao K. Making sense of a helathcare forum - smart keyword and user navigation graphs. Thirty Fourth Internationsl Conference on Information Systems. 2013.

[36] Lichtenwalter RN, Lussier JT, Chawla NV. Proceedings of the Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010. p. 243–52.

[37] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD Record. 1993. p. 207–16.

[38] Kotsiantis S, Kanellopoulos D. Association rules mining: a recent overview. GESTS International Transactions on Computer Science and Engineering 2006;32(1):71–82.

[39] Ng AY, Jordan MI. On discriminative vs. Generative classifiers: a comparison of logistic regression and naive bayes. Proceedings of the Advances in Neural Information Processing Systems. 2002. p. 841–8.

[40] Boser BE, Guyon IM, Vapnik VN. Proceedings of the Proceedings of the Fifth Annual Workshop on Computational Learning Theory. 1992. p. 144–52.

[41] Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13(1):24–9.

[42] Jiang L, Yang C. Using Co-occurrence analysis to expand consumer health vocabularies from social media data. Proceedings of the IEEE International Conference on Healthcare Informatics. 2013.

[43] Jiang L, Yang CC, Li J. Discovering consumer health expressions from consumer-contributed content. Proceedings of the Social Computing, Behavioral-Cultural Modeling and Prediction. 2013. p. 164–74.

[44] Jiang L, Yang CC. Expanding consumer health vocabularies by learning consumer health expressions from online health social media. Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction 2015. 2015. p. 314–20.

[45] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010;6(1).

[46] Knox C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C. DrugBank 3.0: a comprehensive resource for' omics' research on drugs. Nucleic Acids Res 2011;39. (Database issue), p. D1035.

[47] Zwart-van Rijkom JEF, Uijtendaal EV, ten Berg MJ, van Solinge WW, Egberts ACG. Frequency and nature of drug-drug interactions in a Dutch university hospital. Br J Clin Pharmacol 2009;68(2):187–93.

[48] Domingos P. Proceedings of the Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1999. p. 155–64.

[49] Freund Y, Schapire RE. Experiments with a new boosting algorithm. Proceedings of the ICML. 1996. p. 148–56.

[50] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55(1):119–39.

[51] Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry 2010;19(3):227–9.