Harnessing Social Media for Drug-Drug Interactions Detection

Haodong Yang  
College of Information Science and Technology  
Drexel University  
Philadelphia, PA 19104, USA  
Haodong.Yang@drexel.edu

Christopher C. Yang  
College of Information Science and Technology  
Drexel University  
Philadelphia, PA 19104, USA  
Chris.Yang@drexel.edu

Abstract—Adverse drug reactions (ADRs) are causing a substantial amount of hospital admissions and deaths, which cannot be underestimated. Drug-drug interactions (DDIs) are an important patient safety problem and have been reported to cause a large portion of patient adverse events resulting in warning notices or the withdrawal of many drugs from the market. Currently, DDIs detection mainly depends on four kinds of data sources – clinical trial data, spontaneous reporting systems, electronic medical records, and chemical/pharmacologic databases, all of which have some limitations such as cohort biases, low reporting ratio, access issue, etc. In this study, we propose to detect DDIs signals from consumer contributed contents in online healthcare communities using association mining. We conduct an experiment with thirteen drugs and three DDI associations. Leverage, lift and interaction ratio are used in the experiment. DrugBank is used as gold standard to test the performance of the approach. The results show that our techniques are promising to detect signals of DDIs and the proposed measure, interaction ratio, performs better than leverage and lift.

Keywords—drug-drug interactions; social media; online health community; association mining

I. INTRODUCTION

Adverse drug reaction (ADR) is defined 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 [1]. It has been long recognized that ADRs has been presenting a serious health problem worldwide [2], and these harmful reactions would complicate patients’ medical conditions, increase hospital admission, and contribute to more morbidity, even death [3]. Drug-drug interactions (DDIs) are a significant drug safety problem for health consumers and may account for up to 30% of unexpected adverse drug reactions [4, 5]. DDIs are common and often caused by shared pathways of metabolism or interesting pathways of drug action [6]. DDIs 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 [6], and therefore has become an important research area.

Some efforts have been made in clinical trials to detect DDIs [7-10]. However, clinical trials primarily focus on ADRs detection of single drugs and do not typically investigate DDIs [11]. Also, sample size, cohort biases, time spans, and financial limit could be some of the crucial factors that obstacle the discovery of DDIs [12, 13]. Therefore, DDIs detection mainly depends on post-marketing surveillance. Currently in US, the Food and Drug Administration’s (FDA) Adverse Event Reporting System (FAERS) is used as the main source for post-marketing safety surveillance. Any adverse drug reactions of either single drugs or drug-drug combinations could be reported though FDA’s MedWatch site by both health professionals and consumers voluntarily. Many endeavors have been dedicated to detecting DDIs using reports retrieved from such kind of spontaneous reporting system and fruitful contributions have been displayed. However, it has also been recognized that the biggest limitation of spontaneous reports lies in its nature of passiveness which directly leads to low reporting ratio. It has been reported that only 1 to 10 percent of all reportable adverse effects were reported to MedWatch, and the majority of these reports came from drug companies [14]. 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 discovered to supplement the drug safety surveillance from spontaneous reporting systems.

Nowadays, the advance of Web 2.0 technology breeds the introduction and flourishing of online social media such as Facebook, Twitter, LinkedIn, and so on. Also, health-related online social networking sites such as MedHelp, PatientsLikeMe, DailyStrength, etc. are developing very fast and attracting millions of users to join these communities which provide a great platform for both health professionals and consumers to offer, seek, and discuss about any health-related information such as medical history, treatment experience, emotion change etc. Undoubtedly, it also includes sharing usage experience of drugs and discussions about adverse reactions cause by single drug as well as drug-drug

1 http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surv eillance/AdverseDrugEffects/default.htm  
2 https://www.accessdata.fda.gov/scripts/medwatch/medwatch-online.htm  
3 http://www.medhelp.org/  
4 http://www.patientslikeme.com/  
5 http://www.dailystrength.org/
combinations. In other words, there would be an ocean of valuable and timely healthcare information available on these online communities and if we make good use of it, we may detect novel and rare DDIs more effectively than only depending on spontaneous reporting systems to support post-marketing surveillance.

In this work, we focus on the interaction between two drugs and propose to detect DDI signals through analyzing consumer contributed contents retrieved from online healthcare communities. If two drugs and an ADR are mentioned by health consumers in a designated field such as a thread in a forum frequently, an indication of an association could be inferred between the two drugs and the ADR. Therefore, we propose to use associations mining to detect DDIs from online healthcare communities.

II. LITERATURE REVIEW

Recently, there is a large number of research works that has been dedicated to the signal detection of DDIs. In terms of the data sources used, these studies could be grouped into four categories: clinical trial data, spontaneous reporting systems, electronic medical records, and chemical/pharmacological databases.

A. Clinical Trial Data

In order to determine whether elderly patients admitted to hospital with specific drug toxicities were likely to have been prescribed an interacting drug in the week prior to admission, Juurlink et al. [7] conducted three population-based, nested case-control studies on patients aged 66 years or older who were treated with glyburide and digoxin in Ontario. They found that many hospital admissions of elderly patients for drug toxicity occur after administration of a drug known to cause DDIs. In another study, Astrandet et al. [8] conducted a 15-month cohort study in which they collected and analyzed the prescriptions from all individuals to examine the frequency, distribution and determinants of potential drug interactions. Obrelli-Neto et al. [9] presented a cohort study which lasted a year on patients aged over 60 in Brazil to determine the incidence and characteristics of DDI-related ADRs among elderly outpatients as well as the factors associated with these reactions. Yukawa et al. [10] used a nonlinear mixed-effects modeling (NONMEM) to estimate the effects of drug-drug interaction on phenobarbitone clearance values, using 648 serum levels gathered during the routine clinical care of 349 pediatric and adult epileptic patients. They demonstrated that concomitant administration of phenobarbitone and valproic acid resulted in a 35.8% decrease of phenobarbitone clearance. The disadvantages of using this kind of data lies in the fact that clinical trials focus on establishing the safety and efficacy of single drugs, and do not typically investigate DDIs, and even when DDIs are suspected, sample sizes and cohort biases limit the ability to discovery rare adverse events [5]. Also, clinical trials are often quite time-consuming and costly.

B. Spontaneous Reporting Systems

Spontaneous reporting systems are a major repository for ADRs signal detection as well as DDIs signal detection. A variety of techniques have been proposed to identify DDI signals from such kind of data. Current post-marketing surveillance in United States primarily depends on FDA’s FAERS, and many approaches have been presented to detect DDI signals from it. For example, Tatonetti et al. [5] used FDA’s single-drug reports to build classification models for eight pre-defined adverse drug events, and then looked for pairs of drugs that match these single-drug profiles in order to predict potential interactions. Harpaz et al. [15] used a well-established data mining method – association rules mining and used lift as a measure to discover DDIs signals. Thakrar et al. [16] 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 other European and Asian countries. For example, Leone et al. [17] performed a search in an Italian spontaneous reporting database to verify which drug pairs are considered a potential DDI, using the online DRUGDEX system. The results showed that spontaneous reporting systems can be an important resource for detecting ADRs associated with the concomitant use of interacting drugs. Van Puijenbroek et al. [18, 19] used logistic regression to calculate ADR reporting odds ratios – the ratio of the exposure odds of reported cases of interest to the exposure odds of other ADRs – for detecting possible drug-drug interactions from spontaneous reports received by the Netherlands Pharmacovigilance Foundation LAREB. Qian et al. [20] developed a system to detect DDIs from reports submitted to Shanghai spontaneous reporting systems. They used logistic regression, the Ω shrinkage measure, an additive model and a multiplicative model for automatic detection of drug–drug interactions where two drugs were used concomitantly. Last but not the least, Norén et al. [21] used a statistical method and implemented and evaluated a shrinkage observed-to-expected ratio for exploratory analysis of suspected drug–drug interaction in a spontaneous reporting system – WHO database.

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 under-reporting 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 nature of the current reporting system [22].

C. Electronic Medical Records

Several studies proposed to mine DDIs from Electronic Medical Records (EMR) data. For example, Chan et al. [23] 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. Their findings showed that ten out of the 17 antidepressant-oral anticancer drug pairs could potentially cause pharmacokinetic interactions, and the rest were pharmacodynamic interactions, with only three out of the 17 drug pairs were clinically documented to cause interacting events. Zwart-van Rijkom et al. [24] 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. They found that the most
frequently occurring potential clinical consequence of the DDIs was an increased risk of side-effects such as increased bleeding risk, hypotension, nephrotoxicity and electrolyte disturbances. EMR is often difficult to access because of privacy issues that it is usually available only to those research groups who have cooperation with hospitals, clinics or any other health organizations and communities [2].

D. Chemical/Pharmacological Databases

Some studies also focus on chemical/pharmacologic databases to detect signals of DDIs and DrugBank is a typical exemplar (Detailed introduction of DrugBank can be found in experiment section). Using DrugBank as a data source, Vilar et al. [4] presented a methodology applicable on a large scale that identifies novel DDIs based on molecular structural similarity to drugs involved in established DDIs. The underlying assumption is that if drug A and drug B interact to produce a specific biological effect, then drugs similar to drug A (or drug B) are likely to interact with drug B (or drug A) to produce the same effect. Segura-Bednar et al. [25-27] proposed two different methods – pattern matching and supervised machine learning (shallow linguistic kernel) – to automatically extract DDIs from biomedical texts retrieved from DrugBank. In order to evaluate their methods, they created the first annotated corpus, the DrugDDI corpus using documents in DrugBank, to study the phenomenon of interactions among drugs. 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.

As stated before, online health communities are providing a huge collection of valuable information for us to detect drug and drug interaction. However, this data still remain untapped by researchers in medical and healthcare domains. To the best of our knowledge, there are very few studies focusing on exploiting online health communities for DDI detection. Therefore, in this work, we propose to use associations mining to detect signals of DDIs from online healthcare communities. When health consumers discuss about ADRs of a concomitant use of two drugs, the co-occurrence of the drugs and their ADR in the posts or comments of a health-related social networking sites could be regarded as an association, and its interestingness and impressiveness [3] can be measured by investigating some measures such as support, confidence, leverage, lift etc.

III. METHODOLOGY

A. Data Preparation

Our previous studies have proved that online health communities, such as MedHelp, PatientsLikeMe and DailyStrength, are great resources that could provide abundant information for detecting signals of adverse drug reactions [2, 13]. More specifically, drug forums on those online social health websites are what we are looking for. Using these interacting platforms for sharing their health-related experience, health consumers discuss all possible aspects about drugs they took or are taking by starting new post followed by other people’s comments or by commenting on other people’s posts. Not only could they talk about some adverse reactions of a specific drug, it’s highly possible that they are also discussing some ADRs caused by drug and drug interactions, which therefore makes it promising to identify DDIs from health consumers’ discussions. In this study, online health community is used as the data source to collect forum discussions about drugs to detect DDIs and a topic thread, which consists of a post and all the following comments, is considers as an analysis unit.

As all the threads are composed of free texts, certain natural language processing techniques need to be utilized for data preprocessing. Punctuations and stop words are first removed. Next, n-grams (n=1, 2, ... m) are generated from all the words in each thread, where the maximum number m is determined empirically. Then the n-grams are matched with drug names and ADR lexicon respectively to identify DDIs.

B. Consumer Health Vocabulary

In order to detect ADRs caused by drug and drug interaction from forum thread, in this study, we use matching technique to find ADRs terms by referring to an external ADRs vocabulary. However, since Internet is characterized by casualness and openness, it’s been long recognized that laypersons (i.e. health consumers) and healthcare professionals think about and express health-related concepts quite differently [28]. In other words, consumers are using a vocabulary that is very different from what health professional are referencing. For example, in terms of arrhythmia which is often used by professionals and is a preferred name by Unified Medical Language System (UMLS), consumers usually prefer to express this ADR as abnormal heart rhythms, irregular heartbeat and so on. Therefore, standard medical lexicon used by professionals like UMLS is not applicable in this case, which means the vocabulary gap between the consumers and professionals would hinder effective ADRs detection from online healthcare communities, and we should look for another type of vocabulary which consists of consumer contributed expressions of ADRs.

In this study, in order to handle above problem, we choose Consumer Health Vocabulary (CHV) Wiki6 to build our ADRs lexicon. Consumer Health Vocabulary (CHV) is a computerized collection of health expressions derived from actual consumer utterances (authored by consumers), linked to professional concepts, and reviewed and validated by professionals and consumers [28]. It’s 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” [28]. The first-generation CHV was developed by Zeng et al [29]. CHV reflects the difference between consumers and professionals in expressing health concepts and helps to bridge the vocabulary gap. In this work, we used CHV to build up our ADRs lexicon. We search for some specific ADRs in CHV Wiki to gather together all the expressions that are most used by consumers, and then add them into our ADRs lexicon. For example, we find 19 different expressions for ADR arrhythmias in CHV such as abnormal cardiac rhythm, arrhythmia, abnormal heartbeat, etc.

http://consumerhealthvocab.chpc.utah.edu/CHVwiki/
abnormal heart rhythms, heart rhythm disorder, irregular heartbeat and so forth.

C. Associations Mining for DDI

Associations mining technique is widely used in the field of data mining to discover interesting patterns [30]. When health consumers discuss interaction of two drugs, the co-occurrence of the pair of drugs and their consequential adverse reaction in the posts or comments of an online health community forum could be regarded as an association, and its interestingness and impressiveness can be measured by investigating some measures. Then the signal strength of the ADRs could be reflected by the value of those measures. Therefore, in this study, we adopt this method and measures of support, confidence, leverage, and lift to detect signals of DDIs.

1) Association Mining

Association mining can be defined mathematically as follow. Let \( I = \{ I_1, I_2, ..., I_m \} \) 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 is an implication of the form \( A \Rightarrow B \), where \( A \subseteq I \), \( B \subseteq I \), and \( A \cap B = \emptyset \), where both \( A \) and \( B \) are a set of items, which is referred to as an itemset [30]. An itemset that contains \( k \) items is a \( k \)-itemset. For example, the set \{simvastatin\} is a 1-itemset, the set \{simvastatin, myopathy\} is a 2-itemset, and the set \{simvastatin, gemfibrozil, myopathy\} is a 3-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset [30].

In this work, we apply the associations mining to DDI signals detection and constrain the number of interacting drugs as two. In this case, 1) \( I = \{ I_1, I_2, ..., I_m \} \) denotes a set of n-gram terms, \( n=1, 2, 3, ... \). Each \( I_i \) is a word or multi-word term extracted from all the threads we collected, and it could be a drug, an ADR, or any other n-gram term; 2) \( X \) is the dataset including all threads of all drugs; 3) each transaction \( T \) represents a thread, and the items in each \( T \) are n-gram terms; and 4) we are interested in mining the association \((D_1 \cup D_2) \Rightarrow R \) (or \((Drug_1 \cup Drug_2) \Rightarrow ADR \)) in which \( Drug_1 \) and \( Drug_2 \) represent two drugs \( Drug_1 \) and \( Drug_2 \) and \( R \) stands for an ADR. Note that symbol “\( \cup \)” here means that a transaction contains both \( Drug_1 \) and \( Drug_2 \).

2) Measures

There are two basic measures used in associations mining: support and confidence. In this study, support is defined as the percentage of transactions in \( X \) that contain \( D_1 \cup D_2 \cup R \):

\[
\text{support}((D_1 \cup D_2) \Rightarrow R) = \frac{\text{count}(D_1 \cup D_2 \cup R)}{\text{total count}}
\]

where \( \text{count}(D_1 \cup D_2 \cup R) \) is the number of threads that contain \( D_1, D_2 \) and \( R \), and \( \text{total count} \) is the total number of threads in the whole dataset. A low value of \( \text{support}((D_1 \cup D_2) \Rightarrow R) \) suggests that association \((D_1 \cup D_2) \Rightarrow R \) may occur simply by chance and it does not indicate a strong signal of ADR. In addition, other measures should be taken into consideration. In this study, confidence is defined as the percentage of transactions in \( X \) containing \( D_1 \cup D_2 \) that also contain \( R \):

\[
\text{confidence}((D_1 \cup D_2) \Rightarrow R) = \frac{\text{support}((D_1 \cup D_2) \Rightarrow R)}{\text{support}(D_1 \cup D_2)}
\]

Confidence determines the extent to which the appearance of \( D_1 \cup D_2 \Rightarrow R \) implies the appearance of \( R \). Based on these two measures, an association could be identified if both of its support and confidence values exceed a pre-determined threshold.

Another two commonly used measures are leverage and lift which are defined as follows respectively:

\[
\text{leverage}((D_1 \cup D_2) \Rightarrow R) = \frac{\text{support}((D_1 \cup D_2) \Rightarrow R)}{\text{support}(D_1 \cup D_2) \times \text{support}(R)}
\]

\[
\text{lift}((D_1 \cup D_2) \Rightarrow R) = \frac{\text{count}(D_1 \cup D_2 \cup R)}{\text{count}(D_1 \cup D_2) \times \text{count}(R)}
\]

Both leverage and lift consider the correlation between itemsets \( D_1 \cup D_2 \) and \( R \). Leverage indicates the proportion of transactions in \( X \) that contain \( D_1 \cup D_2 \cup R \) by excluding probability that if \( D_1 \cup D_2 \) and \( R \) are independent with each other whereas lift considers the ratio of those two. For example, note that lift can also be written as:

\[
\text{lift}((D_1 \cup D_2) \Rightarrow R) = \frac{\text{support}((D_1 \cup D_2) \Rightarrow R)}{\text{count}(D_1 \cup D_2) \times \text{count}(R)}
\]

Large values indicate that the occurrence of the \((D_1 \cup D_2) \Rightarrow R \) association has unlikely occurred by chance. Roughly, \( \text{lift}((D_1 \cup D_2) \Rightarrow R) = 1 \) indicates that the two drugs and ADR are statistically independent with each other, \( \text{lift}((D_1 \cup D_2) \Rightarrow R) > 1 \) that the drugs and ADR are positively correlated, and \( \text{lift}((D_1 \cup D_2) \Rightarrow R) < 1 \) that they are negatively correlated. For both leverage and lift, the higher the values are, the stronger the DDIs signals are. There two measures are used in this study.

To sum up, using above four measures, we are interested in the occurrence of three kinds of itemsets: the first one is a 1-itemset containing one ADR such as \{myopathy\}, the second one is a 2-itemset including two certain drugs such as \{simvastatin, gemfibrozil\}, and the third one is a 3-itemset containing two drugs and an ADR such as \{simvastatin, gemfibrozil, myopathy\}. Given the total number of threads in the whole dataset, we can obtain the value of above four measures by calculating \( \text{count}(R) \) (number of threads containing the ADR), \( \text{count}(D_1 \cup D_2) \) (number of threads containing both two drug names), and \( \text{count}(D_1 \cup D_2 \cup R) \) (number of threads containing both two drug names and the ADR).
Although our previous research has demonstrated that the four measures, especially leverage and lift, could effectively detect ADRs reported by FDA [2, 13], we were dealing with a single drug and its adverse reaction. Also, there are some limitations about above four measures. For example, support and confidence work well when the frequency of ADRs is high on a particular drug. However, health consumers discuss about every aspect of drugs in the threads, such as drug dosage and drug prescription. It is very likely that threads that are related to ADR are only a small number of the total number of threads, especially for those rare ADRs. The value of leverage could also be very low and even negative values would be obtained which makes it very difficult to interpret the results [2, 13].

Since this work is dedicated to identifying DDIs signals (i.e. association (D₁ ∪ D₂) ⇒ R), these four measures may not be sufficient to detect the signals. For example, the value of support could be even less because there may be very few consumers mentioning both drugs and their ADR of interaction as they may not be aware of fact that the ADR is caused by drug and drug interaction. Also, negative leverage values could not be avoided. Therefore, in order to identify DDIs signals, a variation of confidence is used in this work, which is called interaction ratio and defined as:

\[ R_c((D_1 ∪ D_2) ⇒ R) = \frac{\text{confidence}(D_1 ⇒ R) \times \text{confidence}(D_2 ⇒ R)}{\text{confidence}(D_1 ⇒ R) + \text{confidence}(D_2 ⇒ R)} \]

where \( R_c \) means interaction ratio, subscript c denotes confidence on which this formula is based, \( D_1 \) is one of the drugs in our collected dataset, \( D_2 \) is a drug which could interact with \( D_1 \) to generated ADR \( R \), confidence \( (D_1 ⇒ R) \) is the confidence value of association \( D_1 ⇒ R \) that \( R \) is only cause by \( D_1 \), and confidence \( (D_2 ⇒ R) \) is the confidence value of association \( D_2 ⇒ R \) that \( R \) is only cause by \( D_2 \). The rationale behind this measure is that if an ADR is caused by the interaction of \( D_1 \) and \( D_2 \) rather than only by \( D_1 \) or \( D_2 \) alone, then the value of confidence \( (D_1 ∪ D_2) ⇒ R \) should be higher than that of confidence \( (D_1 ⇒ R) \) or confidence \( (D_2 ⇒ R) \), which makes the values of \( R_c((D_1 ∪ D_2) ⇒ R) \) high.

Leverage ratio is not considered due to the problem of negative values. The ratio of support \( (R_s) \) and lift \( (R_l) \) could be derived as the product of \( R_c((D_1 ∪ D_2) ⇒ R) \) and a constant and therefore be proportional to \( R_c \). For example,

\[ R_s((D_1 ∪ D_2) ⇒ R) \]

In this study, health consumers’ posts and comments are collected from MedHelp, one of the largest online healthcare communities worldwide. Since its introduction in 1994, MedHelp is the pioneer in online health communities. Today, MedHelp empowers over 12 million people each month to take control over their health and find answers to their medical questions [31]. The search engine provided by MedHelp is used to search for all the threads about a certain drug. Fig. 1 demonstrates the returned results for drug Adenosine, from which we are able to get the number of total “Posts” (considered as threads in this study) and detailed content of each thread (including original post and all the following comments) after we click on a thread title.

Table 1 Dataset

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Number of Threads</th>
<th>Drug Name</th>
<th>Number of Threads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biaxin</td>
<td>686</td>
<td>Lansoprazole</td>
<td>592</td>
</tr>
<tr>
<td>Luvox</td>
<td>570</td>
<td>Prozac</td>
<td>718</td>
</tr>
<tr>
<td>Gadolinium</td>
<td>706</td>
<td>Heparin</td>
<td>1463</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>796</td>
<td>Tacrolimus</td>
<td>583</td>
</tr>
<tr>
<td>Zocor</td>
<td>919</td>
<td>Epogen</td>
<td>819</td>
</tr>
</tbody>
</table>

IV. EXPERIMENT

A. Dataset

To effectively detect DDIs with our proposed techniques, the drug should have active discussion in MedHelp. Therefore, we target ten drugs with more than five hundred threads for each of them, and collected all the original posts and comments of these drugs systematically. Table 1 shows the dataset for this study.

![Search Results for Adenosine](http://www.medhelp.org/search)
B. DrugBank Dataset

In this work, DrugBank database is used as a 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 [32-34]. We search for all the drugs that could interact with each of the ten drugs in our database using the Interax Interaction Search engine on the DrugBank Database. At last, three common adverse reactions and three corresponding drugs are selected. Table 2 illustrates the detail of dataset we are using, including $D_1$, $D_2$ that could interact with $D_1$, and corresponding adverse reactions. Drugs Quinidine, Ticlopidine and Gemfibrozil have 187, 97 and 117 threads in MedHelp respectively. As we can see, in this study we are trying to detect five associations that have been highlighted in Table 2.

<table>
<thead>
<tr>
<th>Interactive Drugs (Db)</th>
<th>Quinidine</th>
<th>Ticlopidine</th>
<th>Gemfibrozil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ten Drugs ($D_b$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biaxin</td>
<td>Yes</td>
<td>Arrhythmias</td>
<td>NO NO NO NO</td>
</tr>
<tr>
<td>Lansoprazole</td>
<td>NO NO NO NO NO NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luvox</td>
<td>NO NO NO NO NO NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prozac</td>
<td>NO NO NO NO NO NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gadolinium</td>
<td>NO NO NO NO NO NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heparin</td>
<td>NO NO Yes Bleeding NO NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simvastatin</td>
<td>NO NO NO Yes Myopathy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tacrolimus</td>
<td>Yes Arrhythmias NO NO NO NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zocor</td>
<td>NO NO NO Yes Myopathy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epogen</td>
<td>NO NO NO NO NO NO</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C. Results

In this work, we use exact matching technique for finding the interactive drugs and corresponding adverse reactions for the ten drugs. We set the maximum number $m$ for n-gram as three because the longest term of ADR we obtained in CHV consisted of three words after pre-processing. For lift and $R_c$, formulas, we have to deal with the problem of undefined values because both of their denominators contain item $count(D_1 \cup D_2)$ and there is a chance that the number of threads that contain both $D_1 \cup D_2$ is zero. When this item is equal to zero, another item $count(D_1 \cup D_2 \cup R)$ is obviously equal to zero too. Under this circumstance, we assign zero to the values of lift and $R_c$ instead of undefined values.

In this work, we identify the strongest signals between a pair of drugs and an ADR and compare the performance of different measures. Each ADR is considered independently because each ADR has its own set of vocabularies for detection. Table 3 shows the experiment results in which the value of each cell in the table is the corresponding leverage, and lift of each association. The associations that we are supposed to identify have been highlighted. The number in the parentheses indicates the ranking of that value of leverage or lift among all the ten drugs ($D_b$) under each association ($D_1 \cup D_2$) $\Rightarrow R$. A higher ranking represents a stronger signal for the corresponding DDI. As we can see, most of the DDIs reported in DrugBank have higher ranking than other DDIs in terms of either leverage or lift. Drug biaxin and tacrolimus which could interact with quinidine to produce ADR arrhythmias rank forth and third in leverage and third and first in lift respectively. For ADR bleeding, drug heparin that would increase the risk of bleeding when used with ticlopidine ranks second in both leverage and lift. With regard to the last ADR myopathy, the two drugs simvastatin and zocor that are supposed to be identified to be able to interact with gemfibrozil rank among the first two places respectively in both leverage and lift.

In this study, we also use confidence ratio to detect DDIs signals. Table 4 demonstrates the experiment results in which the associations reported by DrugBank have been highlighted. As we can see, it’s obvious that all the DDIs that we are supposed to detect rank among the first two places under each ($D_1 \cup D_2$) $\Rightarrow R$.

We also notice that there is no global threshold that we can set to filter out DDIs with high-valued leverage, lift or $R_c$ across all the adverse drug reactions. The reason is that our method is highly depends on an external vocabulary CHV which provides diverse sets of terms for different ADRs. Therefore, we used ranking to better discern the experiment results. As we can see in Table 3 and Table 4, all the DDI signals that we need to detect rank very high among all the ten drugs in terms of all three measures. However, we also found that some DDIs that have not been reported by DrugBank also rank very high in the three measures. For example, association (Luvox U Ticlopidine) $\Rightarrow$ Bleeding rank first in all leverage, lift and $R_c$; (Zocor U Quinidine) $\Rightarrow$ Arrhythmias rank first in leverage and lift and third in $R_c$. These could be false positives, but they can also mean potential drug safety signals that have not been identified yet, which may deserve further pharmaceutical investigation.

D. Evaluation

Based on the ranking results, we use top-k precision, sensitivity, and specificity to evaluate the effectiveness of proposed method and better compare the performance of leverage, lift and $R_c$. K equals to 2, 3, and 4 respectively in this study. Precision, sensitivity, and specificity are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Sensitivity} = \frac{TP}{TP + FN}$$
$$\text{Specificity} = \frac{TN}{TN + FP}$$

* http://www.drugbank.ca/interax/drug_lookup
where TP is true positive, FP is false positive, FN is false negative, and TN is true negative. Table 5 presents the performance of different measures under top2, top3, and top4 precision, sensitivity, and specificity. As we can see, $R_c$ achieves the best performance among the three in terms of average value of all the precision, sensitivity, and specificity.

Table 3 Leverage and Lift

<table>
<thead>
<tr>
<th>Measure</th>
<th>Leverages</th>
<th>Lift</th>
<th>Leverage</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biaxin</td>
<td>1.05E-04(4)</td>
<td>5.74E+00(3)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lansoprazole</td>
<td>-3.70E-06(5)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Luvos</td>
<td>-1.48E-05(6)</td>
<td>0</td>
<td>1.15E-04(1)</td>
<td>0</td>
</tr>
<tr>
<td>Prozac</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gadolinium</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Heparin</td>
<td>-7.77E-05(7)</td>
<td>0</td>
<td>5.37E-05(2)</td>
<td>3.62E-04(1)</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>1.92E-04(2)</td>
<td>4.05E-00(4)</td>
<td>-1.26E-05(3)</td>
<td>0</td>
</tr>
<tr>
<td>Tacrolimus</td>
<td>1.09E-04(3)</td>
<td>6.89E-00(1)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Zocor</td>
<td>2.18E-04(1)</td>
<td>6.89E-00(1)</td>
<td>-2.52E-05(4)</td>
<td>0</td>
</tr>
<tr>
<td>Epogen</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4 Confidence Ratio

<table>
<thead>
<tr>
<th>Associations</th>
<th>(D1 $\cup$ Quinidine) ⇒ Arrhythmias</th>
<th>(D1 $\cup$ Ticlopidene) ⇒ Bleeding</th>
<th>(D1 $\cup$ Gemfibrozil) ⇒ Myopathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biaxin</td>
<td>3.24E+00(2)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lansoprazole</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Luvos</td>
<td>0</td>
<td>1.44E+02(1)</td>
<td>0</td>
</tr>
<tr>
<td>Prozac</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gadolinium</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Heparin</td>
<td>0</td>
<td>1.11E+01(2)</td>
<td>0</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>1.13E+01(4)</td>
<td>0</td>
<td>3.68E+02(1)</td>
</tr>
<tr>
<td>Tacrolimus</td>
<td>1.24E+02(1)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Zocor</td>
<td>1.19E+01(3)</td>
<td>0</td>
<td>2.47E+02(2)</td>
</tr>
<tr>
<td>Epogen</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5 Precision, Sensitivity, and Specificity

<table>
<thead>
<tr>
<th>Measure</th>
<th>Leverage</th>
<th>Lift</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 2</td>
<td>Precision</td>
<td>0.50</td>
<td>0.44</td>
<td>0.42</td>
</tr>
<tr>
<td>Top 3</td>
<td>Sensitivity</td>
<td>0.60</td>
<td>0.80</td>
<td>1.00</td>
</tr>
<tr>
<td>Top 4</td>
<td>Specificity</td>
<td>0.85</td>
<td>0.80</td>
<td>0.72</td>
</tr>
</tbody>
</table>

I. CONCLUSION

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 four kinds of data source, namely clinical trial data, spontaneous reporting systems, electronic medical records, and chemical/pharmacological databases. However, each data source suffers from such limitations as cohort biases, low reporting ratio, access issue, and so on. It is urgent that an alternative data source be found that could supplement extant data sources. In this work, we proposed to harness online healthcare communities, which include large volume of timely health consumer contributed content, to mine DDIs. As social media becomes so prevalent, health consumers are not only searching health information online, but also share their medical experience with peers on such platforms as MedHelp, PatientsLikeMe, etc. Association mining is used in this study to achieve our goal of detecting DDI signals. We collected thread contents of thirteen drugs from MedHelp, selected five reported DDI associations from DrugBank, and computed the values of leverage and lift to identify DDIs. We also proposed another measure, interaction ratio, to capture drug safety signals. Experiment results demonstrated that our method is able to effectively detect DDIs reported by DrugBank. In the future, this work could be further extended in several directions. For example, the dataset, including both number of drugs and number of adverse reactions could be expanded to better verify the effectiveness of our proposed method and compare different measures. Also, it can be seen that the performance of our method largely depend on an external vocabulary CHV, so how to expand the vocabulary to contain as many consumer contributed expressions as possible is an important task to do.

REFERENCES


Informatics, Beijing, August 12, 2012.


