Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis

HAODONG YANG, College of Computing and Informatics, Drexel University
CHRISTOPHER C. YANG, College of Computing and Informatics, Drexel University

Since Adverse Drug Reactions (ADRs) represent a significant health problem all over the world, ADRs detection has become an important research topic in drug safety surveillance. As many potential ADRs cannot be detected through pre-marketing review, drug safety currently depends heavily on post-marketing surveillance. Particularly, current post-marketing surveillance in the United States primarily relies on the FDA Adverse Event Reporting System (FAERS). However, the effectiveness of such spontaneous reporting systems for ADRs detection is not as good as expected because of their extremely high underreporting ratio. Moreover, it often takes FDA years to complete the whole process of collecting reports, investigating cases and releasing alerts. Given the prosperity of social media, many online health communities are publicly available for health consumers to share and discuss any healthcare experience such as ADRs they are suffering. Such health consumer contributed contents are timely and informative, but this data source still remains untapped for post-marketing drug safety surveillance. In this study, we propose to use (1) associations mining to identify the relations between a drug and an ADR and (2) temporal analysis to detect drug safety signals at the early stage. We collect data from MedHelp, and use FDA's alerts and information of drug labeling revision as gold standard to evaluate the effectiveness of our approach. The experiment results show that health-related social media is a promising source for ADRs detection, and our proposed techniques are effective to identify early ADRs signals.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database applications – Data mining; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – Linguistic processing; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia

General Terms: Algorithms; Experimentation

Additional Key Words and Phrases: Drug Safety Signal Detection; Adverse Drug Reactions; Social Media; Health Consumer Contributed Content; Post-marketing Surveillance; Associations Mining; Temporal Analysis

1. INTRODUCTION

It has been long recognized that Adverse Drug Reactions (ADRs) represent a significant world-wide health problem. ADRs are considered to be a leading cause of death in the United States [Chee et al. 2011; Lazarou et al. 1998]. Therefore, ADRs detection has become a significant research area in drug safety surveillance. In 2000, Edwards and Aronson [Edwards and Aronson 2000] gave a comprehensive definition to ADR: Adverse Drug Reaction 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. ADRs would complicate patients’ medical conditions, increase hospital admission, and contribute to more morbidity, even death [Ji et al. 2010]. For example, a systematic review showed that approximately 5.3% of hospital admissions were associated with ADRs [Kongkaew et al. 2008]. Another study, by assessing 3,190 medical records of all newly admitted internal ward patients in an internal hospital over 6 months, found that 304 ADRs were identified in 242 patients, with 60% directly leading to admission, and older age and female gender are significantly associated with ADR related hospital admissions [Hofer-Dueckelmann et al. 2011]. Furthermore, it is estimated that approximately 2 million patients in the United States are affected each year by ADRs [Liu and Chen 2013]. All the statistics that were shown above vividly depicted the seriousness of ADRs with respect to drug utilization.

Currently, there are two major approaches to detect ADRs: pre-marketing review and post-marketing surveillance. Before any pharmaceutical new drugs are approved

A manuscript submitted to ACM Transactions on Intelligent Systems and Technology
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-marketing review process is too constrained in scale and time span to possibly identify all potential adverse effects. Therefore, drug safety currently depends heavily on post-marketing surveillance – the systematic detection and evaluation of medicines once they have been marketed – to detect latent ADRs.

Current post-marketing surveillance in United States primarily depends on FDA Adverse Event Reporting System (FAERS). Healthcare professionals (such as physicians, pharmacists, nurses and others) and consumers (such as patients, family members, lawyers and others) voluntarily report to the system when they experience adverse events or medication errors. Although such systems could cover a large number of patients and drugs, they, to a great extent, are restricted by the nature of passiveness – people report adverse reactions spontaneously and voluntarily. For example, Wood demonstrated 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 [Wood 2000]. Hazell et al. pointed out that the median under-reporting rate of ADRs to spontaneous reporting systems across the thirty seven studies was 94% [Hazell and Shakir 2006]. The surprisingly high under-reporting ratio makes ADRs detection much less effective. Moreover, it often takes FDA a lot of time to complete the whole process of collecting reports, investigating cases and releasing alerts. Therefore, it is usually impossible to detect ADRs timely by only relying on spontaneous reporting systems, thus endangering public health decisions. To sum up, it is urgent to find an alternative data source and techniques for ADRs detection.

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

(1) A cyber-based empowering of patients and healthy individuals that enables them to play a substantial role in their own health and treatment; and
(2) Utilizing diverse data to provide automated and augmented insight, discovery, and evidence-based health and wellness decision support.

The development of Web 2.0 and Health 2.0 technologies makes the first thrust promising. The advancement of Internet not only breeds the various online social media sites such as Facebook, Twitter, LinkedIn, and so on, but also leads to the flourishing of online health communities (OHCs) such as MedHelp, PatientsLikeMe, DailyStrength, and so forth. A recent survey conducted in September 2012 by Pew Internet & American Life Project showed that 72% of internet users say they looked online for health information within the year of 2012 [Fox and Duggan 2013]. We can easily imagine that uncountable health consumers as well as health professionals go to those health-related social media sites frequently to either seek or offer healthcare information. For example, since its introduction in 1994, MedHelp is the pioneer in OHCs. Today, MedHelp empowers over 12 million people each month to take control over their health and find answers to their medical questions [MedHelp 1994]: “Every day, members come to MedHelp to receive the support they need from other patients like them, to research information on drugs and health topics, to document their medical history, and to share their knowledge with others in need.” Such OHCs
Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis

provide a great platform for health consumers to play a substantial role in their own health and treatment.

The development of Web 2.0 and Health 2.0 technologies also makes the second thrust promising. Frequent visits on OHCs would inevitably generate a huge collection of health-related contents that might be even more informative and often up-to-date than some administrative databases. Consumer contributed contents of their experience with drug usage and ADRs suffering are of particular interests to us. It is highly possible that many patients choose OHCs over some available reporting systems such as FAERS because of ignorance of these channels, embarrassment, perceptions of negative provider attitude, etc. [Benton et al. 2011] Instead, they often resort to informal networks such as OHCs to discuss adverse events. Therefore, we anticipate ADR data from online health communities can complement that in spontaneous reporting systems. If we can take good advantage of these consumer contributed contents, we may be able to discover interesting knowledge, insights and patterns that cannot be obtained from other data sources. Therefore, online health-related social media would serve as alternative data source to provide automated and augmented insight, discovery, and evidence-based health and wellness decision support.

Our preliminary research has proved that it is feasible to reveal knowledge from online health communities – OHCs have become a reliable data source for effective ADRs detection [Yang, Jiang, et al. 2012; Yang, Yang, et al. 2012]. However, we have not systematically tested the hypothesis that OHCs data can be used to detect early ADR signals through temporal analysis. In this work, we focus on harnessing OHCs to detect signals of drug-ADR associations. We combined techniques of associations mining and temporal analysis for the purpose of identifying ADR signals as soon as possible.

The remainder of this paper is organized as follows. In section 2, we systematically review related research on ADR detection and temporal data mining. Section 3 presents our techniques for ADR detection: associations mining and temporal analysis. In section 4, we conducted an experiment on a MedHelp dataset and demonstrated the effectiveness of our approach by using FDA’s alerts as gold standard. Section 5 concludes our study and discusses our future research directions.

2. LITERATURE REVIEW

2.1 ADR Detection

In the recent years, many endeavors have been dedicated to the detection of drug safety signals for the purpose of discovering relations between a drug and an ADR. These studies could be grouped into five categories in terms of data source they used: (1) spontaneous reporting systems, (2) electronic health records, (3) pharmaceutical databases, and (4) online health-related social media.

2.1.1. Spontaneous Reporting Systems. In the last few decades, the development of computer science makes feasible the accumulation of a large amount of data and many researches developed database-related quantitative methods using the reporting data from spontaneous reporting systems to identify ADRs. The detection of ADRs, as early as possible, from these spontaneous reporting systems forms an important research field for the pharmaceutical industry [Lin et al. 2012].

Current post-marketing surveillance in United States primarily depends on FDA’s FAERS system, and there are also spontaneous reporting centers in other countries. Based on the reporting data, various statistical and data mining methods have been
practically implemented. For example, FDA currently adopts Multi-item Gamma Poisson Shrinker (MGPS) algorithm to detect potential signals from its MedWatch database [Szafman et al. 2004]. The Uppsala Monitoring Center uses Bayesian Confidence Propagation Neural Network (BCPNN) as its signal detection strategy with World Health Organization database [Lindquist et al. 1999].

UK Medicines Control Agency employs Proportional Reporting Ratios (PRR) and chi-squared to recognize adverse reactions, events related to the underlying disease and signals requiring further evaluation by comparing the proportion of all reactions to a drug of interest to the same proportion of all other drugs in UK Yellow Card database [Evans et al. 2001]. The Netherlands Pharmacovigilance Foundation Lareb uses Reporting Odds Ratio (ROR) based upon logistic regression analysis and 95% confidence interval [van Puijenbroek et al. 1999]. Both PRR and ROR belong to the measures of disproportionality that are commonly used for ADR detection.

The performances of these data mining methodologies were compared in a study using a Japanese spontaneous reporting database and the results showed that the ability of detecting a signal varies among these approaches [Kubota et al. 2004]. In the study, they collected 38,731 drug-ADR combinations. The count of drug–ADR pairs was equal to 1 or 2 for 31,230 combinations and BCPNN detected none of them as a possible signal. Also, the GPS detected a possible signal in none of the combinations where the count was equal to 1 but in 7.5% where the count was equal to 2. The ROR and PRR performed better and detected a possible signal in more than half of the combinations where the count was equal to 1 or 2. A number of other data mining methods such as empirical Bayes model [DuMouchel 1999; Hauben 2004; Lanctot and Naranjo 1994], pharmacovigilance map method [Hochberg et al. 2007] and interactive system platform iADRs [Lin et al. 2012] have also been used with spontaneous reporting data.

Although aforementioned data mining methodologies, compared with traditional observational methods, are less time-consuming and less labor intensive, their performance largely depend on the quality of reporting data. Unfortunately, as stated earlier, 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 [Ji et al. 2011]. As a result, the ADR signals cannot be detected until a much later stage when sufficient records are collected. In order to solve this problem, other alternative databases are being chosen to detect ADRs.

2.1.2 Electronic Health Records. Instead of focusing on spontaneous reporting data, many studies investigated electronic health records for ADRs detection. Since electronic health records are primarily generated by health professionals in various healthcare organizations such as hospitals, clinics, etc. when, for example, a patient visits a doctor, this kind of data is more timely and authoritative than spontaneous reports. Ji et al. developed a fuzzy logic-based computational recognition-primed decision (PRD) model to calculate the extent of causality between a drug and some of its adverse effects [Ji et al. 2007]. This PRD model served as a module of a novel intelligent agent software system (called ADRMonitor) which was used for proactively monitoring and detecting potential ADRs of interest with electronic patient records. Also, on the basis of PRD model, another important signal-detection strategy was developed – known as causal association mining algorithm in which a new interestingness measure, causal-leverage, was used to predict potential ADRs from electronic health databases [Ji et al. 2011; Ji et al. 2010]. Wang et al. demonstrated the feasibility of using natural language processing techniques, the
Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis

comprehensive electronic health records, and association statistics for identifying potential ADRs with electronic health records [Wang et al. 2009].

Undoubtedly, data mining techniques based on electronic health records are capable of generating earlier ADR signals than spontaneous reporting data. However, this kind of data is often difficult to access because of privacy issues and it is usually available only to those research groups who have cooperation with hospitals, clinics or any other health organizations and communities [Yang, Jiang, et al. 2012]. The integration of electronic health databases from multiple resources is still a technical and policy challenge, which means depending on their affiliation or collaborating health unit, most researchers may only have a single dataset to work on, which obviously may not cover enough patient records. Therefore, the availability of large scale electronic health records from multiple sources is a limitation for its application on ADRs detection in spite of its usefulness.

2.1.3 Pharmaceutical Databases. A few other works focus on the pharmaceutical databases for detecting ADRs. For example, Pouliot et al. [Pouliot et al. 2011] used a logistic regression model to correlate postmarketing ADRs with screening data from the PubChem BioAssay database. In another work, by extracting taxonomic and intrinsic drug properties from three publicly available databases, World Health Organization Anatomical Therapeutic Chemical (ATC) Classification System, University of Alberta DrugBank, and National Center for Biotechnology Information's PubChem Compound, a predictive pharmacosafety networks (PPNs) was generated to predict likely unknown ADRs [Cami et al. 2011]. However, these kinds of databases are more focused on the chemical or molecular aspect of drugs and their accessibility is also very limited.

2.1.4 Online Health Communities. As stated before, online health communities could serve as alternative data source for insight discovery. However, to the best of our knowledge, there are very limited studies dedicated to ADR detection using health consumer contributed content from online social media.

White et al. have demonstrated that Internet users are able to provide early clues about adverse drug reactions via their search logs that are generated by consumers [White et al. 2013]. This strong evidence showed that consumer contributed content could be used for ADR signal detection. A group from University of Pennsylvania has released a tool – Medpie – that can be used to collect a corpus of medical message board posts, de-identify the corpus, and extract information on potential adverse drug effects discussed by users [Benton et al. 2012; Benton et al. 2011]. 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 [Liu and Chen 2013]. Chee et al. used a machine learning method to classify drugs into FDA's watchlist and non-watchlist based on messages extracted from an online health forum - Health & Wellness Yahoo! Groups [Chee et al. 2011]. However, to obtain a high-performance classifier, it usually requires an enormous amount of human efforts to prepare a training data for detecting the ADR signals and it may not be feasible if we have a large number of drugs and adverse drug reactions. 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 [Leaman et al. 2010]. Further, they developed a system to automatically extract mentions of ADRs from user reviews about drugs by mining a set of language patterns [Nikfarjam and Gonzalez 2011]. In our previous study, using MedHelp as the source of health consumers' posts and comments, we proposed to mine
associations between drugs and adverse reactions and to utilize three measurements – support, confidence and leverage – to identify FDA alerted ADRs [Yang, Jiang, et al. 2012].

2.2 Temporal Data Mining

Time is an important aspect of all real world phenomena, and any systems, approaches or techniques that are concerned with information need to take into account the temporal aspect of data [Chen and Petrounias 1998]. Over the past decades, a great collection of temporal data has been stored in different databases ranging from finance, economics, insurance, communication to other areas such as weather forecast and astronomy because of the increase affordability of storage capacity, resulting in the promise of discovering temporal knowledge from temporal data. In many application domains, temporal data mining has been considered as a precious asset as it has the capability of mining activities rather than just states and, thus, inferring relationships of contextual and temporal proximity, and even helping understand the past and plan the future [Chen and Petrounias 1998; Roddick and Spiliopoulou 2002]. Many endeavors have been put into temporal data mining.

2.2.1 Temporal Association Rules. Temporal association rules are an interesting and important extension to association rules by including a temporal factor. Considering the time dimension leads to different forms of association rules such as discovering association rules that may hold during some time intervals but not during others [Li et al. 2003]. In order to discover association rules and their interval time of validity, Ale et al. expanded the notion of association rules by incorporating time to the frequent itemsets, introduced the concept of temporal support, and modified the known algorithm A priori [Ale and Rossi 2000]. Lee et al. proposed a data mining system for discovering interesting temporal patterns from large databases. The mined patterns were expressed in fuzzy temporal association rules which satisfied the temporal requirements specified by the user [W.-J. Lee and Lee 2004]. Lee et al. proposed an innovative algorithm Progressive-Partition-Miner (PPM) to discover general temporal association rules in a publication database [C.-H. Lee et al. 2001]. Winarko and Roddick outlined an algorithm, ARMADA, to discover frequent temporal patterns and to generate richer interval-based temporal association rules [Winarko and Roddick 2007].

2.2.2 Pattern Maintenance and Evolution. Pattern maintenance and evolution address the problem of updating the mined temporal results that themselves have temporal properties. Therefore, when data change happens, its impact on the mined results should be monitored and updated accordingly. For example, temporal databases are continually appended or updated so that the discovered rules need to be updated. In a comprehensive survey of temporal knowledge discovery paradigms and methods conducted by Roddick and Spiliopoulou, changes on association rules caused by updates in the underlying data are referred to as a problem of “rule maintenance” [Roddick and Spiliopoulou 2002]. In order to improve the existing association rule mining algorithms that typically assume that data characteristics are stable over time, Au et al. proposed to use linguistic variables and linguistic terms to represent the changes in the discovered rules [Au and Chan 2005]. Specifically, given a set of database partitions, each of which contains a set of transactions collected in a specific time period, they first discovered a set of association rules in each database partition, and then constructed fuzzy decision trees, which were converted to a set of fuzzy meta-rules to discover the changes of the rules in different time periods. In another study, Gharib et al. proposed an incremental algorithm to maintain the temporal association rules in a transaction database [Gharib et al. 2010]. Their experimental
results on both the synthetic and the real dataset illustrated a significant improvement over the conventional approach of mining the entire updated database. Also, Boettcher provided a thorough overview of recent works on methods for change analysis, focusing on contrast mining and change mining, the two emerging subfields of contemporary data mining research [Boettcher 2011].

2.2.3 Temporal Data Mining in ADRs Detection. A limited number of papers applied techniques of temporal data mining to the healthcare domain, especially the detection of ADRs. For instance, to signal unexpected and infrequent patterns characteristic of ADRs, Jin et al. proposed a domain-driven knowledge representation Unexpected Temporal Association Rule (UTAR), a mining algorithm, MUTARA (Mining UTARs given the Antecedent), and an interestingness measures, residual-leverage [Jin et al. 2010; Jin et al. 2008]. An administrative health database referred to as the Queensland Linked Data Set (QLDS) was used for their ADR detection. Experimental results on real-world data substantiated that MUTARA can signal suspected ADRs while traditional association mining techniques cannot. In another study, Shanmugapriya et al. provided n-wise unexpected temporal association rule to detect ADRs related to drug-drug interactions [Shanmugapriya et al. 2011].

Post-marketing drug safety surveillance could be one of the many important application domains of temporal data mining. For example, the time span between a patient's exposure to a drug and the presence of an ADR could be used for detecting the strength of the adverse reaction. Also, when health consumers discuss adverse reactions of a specific drug, the co-occurrence of the drug and its ADR in the posts or comments of an online health-related social media could be regarded as an association, and its interestingness and impressiveness can be measured by investigating such measures as support, confidence or lift. However, if we ignore the temporal dimension and use the whole dataset to compute those measures, some strong signals during a certain time interval could go undetected as the analysis only concentrates on static association rules over the whole dataset. As stated earlier, it often takes FDA years to achieve the whole process of collecting reports, investigating suspected ADRs, and releasing alerts. However, way earlier than FDA's releasing date, consumers would probably discuss those ADRs intensively through many different channels among which online health community is a crucial one, and temporal analysis could be used to capture the signal as early as possible. Therefore, in the work, we view the temporal factor as a crucial component in ADR detection and monitor the association evolution consistently during the detection process. In other words, our study belongs to the category of pattern maintenance and evolution. Currently, often used techniques in this area are either mining the changes of rules in database partitions of different time periods or maintaining rules incrementally as the transaction databases increase. However, to the best of our knowledge, such techniques have not been used to mine health consumer contributed data for the purpose of early drug safety signal detection.

3. ADR SIGNAL DETECTION WITH SOCIAL MEDIA DATA
3.1 Data Collection and Pre-processing
Online health-related social media, such as MedHelp and PatientsLikeMe, provides a rich collection of consumer contributed contents for detecting ADRs signals. In common online communities, health consumers discuss drugs by starting new posts with other people commenting on the posts. A post and the following comments compose a thread, which could be regarded as an analysis unit. Analysis granularity
should be chosen empirically according to the research question to be answered. In this work, we use a thread as a unit because it is composed of all comments and discussions on a particular issue raised by the user who initiates the thread. A post can be considered as an analysis unit too. However, a post is usually relatively short and the user who made a post may jump into the point he/she is trying to make without describing the concerned issue. As a result, the association between the discussed point (e.g. ADR) and the concerned issue (e.g. a particular drug) will be lost if we take a post as an analysis unit.

As all the threads are composed of free texts, certain natural language processing techniques should be used to pre-process the data. For example, we first remove punctuations and stop words, then tokenize threads and generate n-grams (n=1, 2… m) from all the words in each thread, where the maximum number m is decided empirically, and at last match the n-grams with an external ADR lexicon to identify ADR terms. Stemming is not implemented in order to maintain the original expression of users.

3.2 Consumer Health Vocabulary

As the Internet is characterized by casualness and openness, it is challenging to extract consumers' diverse expressions of adverse reactions from online discussion threads. Moreover, it has long been recognized that health consumers used very different vocabularies to express health concepts from professionals [Q. T. Zeng and Tse 2006]. For example, on the one hand, patients may find it difficult to understand information provided by healthcare professionals such as ADRs of prescribed medications or instructions for taking drugs. On the other hand, sometimes even healthcare professionals have difficulty interpreting expressions used by health consumers. Thus, standard medical lexicon used by professionals like UMLS is not applicable in this case. In other words, in order to extract ADR terms from the threads and then measure the strength of each drug-ADR association, we will have to rely on another kind of ADR lexicon that mainly consists of consumer contributed expressions.

To deal with this problem, we resorted to Consumer Health Vocabulary (CHV) Wiki1 to build up our ADR lexicon. Zeng et al. have been devoting themselves to the development of open access, collaborative CHV and the first-generation CHV was developed by them [Q. Zeng et al. 2002; Q Zeng et al. 2001; Q. T. Zeng and Tse 2006; Q. T. Zeng et al. 2005; Q. T. Zeng et al. 2006]. CHV is defined as “a collection of forms used in health-oriented communication for a particular task or need (e.g., information retrieval) by a substantial percentage of consumers from a specific discourse group and the relationship of the forms to professional concepts” [Q. T. Zeng and Tse 2006]. This vocabulary links everyday health-related words and phrases to technical terms or jargon used by healthcare professionals, and the goal of CHV is to help bridge the communication gap between consumers and healthcare professionals [Q. T. Zeng and Tse 2006]. Due to the importance of such vocabularies, some researchers even proposed various techniques for semi-automatically or automatically expanding CHV. For example, Zeng et al. used automated text analysis methods to generate candidate Consumer-Friendly Display (CFD) names from the National Library of Medicine MedlinePlus query logs, and then select CFD names from candidate names by human review [Q. T. Zeng et al. 2005]. Jiang et al. developed a semi-automatic method that employs Principal Components Analysis

1http://consumerhealthvocab.chpc.utah.edu/CHVwiki/
Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis

(PCA) and Logistic Regression for identifying consumer health expressions from consumer-contributed content [Jiang et al. 2013]. They further proposed to use co-occurrence analysis to expand CHV from social media data [Jiang and Yang 2013].

In this study, we searched for all the ADRs in the CHV Wiki to obtain the expressions that are frequently used by consumers, and then include them into our ADR lexicon. Using such vocabularies, technical terms may be "translated" into lay language. In CHV, consumer contributed terms are identified through three stages [Doing-Harris and Zeng 2011]: (1) raw data are crawled from a healthcare social media site, processed by natural language processing techniques, and then n-grams are extracted, (2) candidate terms are identified by filtering the non-medical terms by using UMLS and the US Department of Veteran’s Affairs Medical Records, (3) the identified candidate terms are evaluated by human reviewers for collaborative expert review to determine the CVH terms. By searching for a term, CHV will return a list of CHV preferred names and a list of the corresponding consumer contributed expressions for each CHV preferred names for that term. For example, for the ADR heart disease, we only found one CHV Preferred Name and then discovered 22 different expressions in CHV, such as cardiac disease, cardiac disorder, heart diseases, heart disorders, disease heart, and so on. Different consumer contributed terms are mapped to CHV preferred name and the UMLS preferred name respectively. We added these expressions into the lexicon so as to efficiently identify all the terms related to heart disease in the threads and then measure the signal strength of each drug-ADR association.

3.3 Adverse Drug Reaction Detection

3.3.1 Association Mining. Association rules mining, one of the most important and well researched techniques of data mining, was first introduced by Agrawal et al. when they were trying to identify significant purchasing patterns from a large database of consumer transactions [Agrawal et al. 1993]. 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. [Kotsiantis and Kanellopoulos 2006]

Mathematically, 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 rule 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. An itemset that contains \( k \) items is a \( k \)-itemset. For example, the set \( \{Luvox\} \) is a 1-itemset, and the set \( \{Luvox, diarrhea\} \) is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset [Han et al. 2006].

In this study, we applied the association mining to ADRs signals detection. Accordingly, \( I = \{I_1, I_2, ..., I_m\} \) denote a set of n-grams, \( X \) is the dataset including all threads, each thread represents a transaction \( T \), and we are mining the association \( D \Rightarrow R \) (drug \( \Rightarrow \) ADR) here.

There are two important basic measures for association rules mining, support and confidence. Support is defined as the percentage of transactions in \( X \) that contain \( D \cup R \):

\[
\text{support}(D \Rightarrow R) = \frac{\text{count}(D \cup R)}{\text{total count}}
\]
where, in our case, \( \text{count}(D \cup R) \) is the number of threads that contain both \( D \) and \( R \), and total count is the total number of threads in the whole dataset. A low value of \( \text{support}(D \Rightarrow R) \) suggests that association rule \( D \Rightarrow R \) may occur simply by chance and is not interesting or useful to us. Confidence is defined as the percentage of transactions in \( X \) containing \( D \) that also contain \( R \):

\[
\text{confidence}(D \Rightarrow R) = \frac{\text{support}(D \cup R)}{\text{support}(D)} = \frac{\text{count}(D \cup R)}{\text{count}(D)}
\]

Confidence determines the extent to which the appearance of \( D \) implies the appearance of \( R \). Based on these two measures, an association rule could be identified if both of its support and confidence values exceed a pre-determined threshold that is called minimal support and minimal confidence respectively and often predefined by users. Therefore, we are interested in the occurrence of two kinds of itemsets, one is a 1-itemset containing a certain drug like \{Luvox\}, and the other is a 2-itemset containing a drug and an ADR like \{Luvox, diarrhea\}. Given the total number of threads in the whole dataset, we can obtain \( \text{support}(D \Rightarrow R) \) and \( \text{confidence}(D \Rightarrow R) \) by calculating \( \text{count}(D \cup R) \) (number of threads containing both the drug name and the ADR) and \( \text{count}(D) \) (number of threads containing the drug name).

Nevertheless, one limitation of support and confidence lies in the fact that these two measures work well when ADRs of a particular drug appear frequently in our dataset. However, health consumers discuss about diverse aspects of drugs in the threads, such as drug dosage, drug prescription, concomitant use of different drugs, and so forth. It is very likely that threads that are related to ADR are only a small portion of the total number of threads, especially for those rare ADRs. To effectively identify the rare ADRs for a particular drug, we should also incorporate the support of the ADR in the dataset. To address this problem, we adopted another two measures called leverage and lift that are defined as follows respectively:

\[
\text{leverage}(D \Rightarrow R) = \text{support}(D \cup R) - \text{support}(D) \times \text{support}(R)
\]

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

Both leverage and lift measure the strength of an association not only by looking at its support and confidence but also the correlation between itemsets \( D \) and \( R \). Leverage measures the difference between the proportion of threads containing both \( D \) and \( R \) above those expected if \( D \) and \( R \) were independent of each other whereas lift calculate the ratio of these two. For example, lift can also be written as the following probability expression:

\[
\text{lift}(D \Rightarrow R) = \frac{\text{support}(D \cup R)}{\text{support}(D) \times \text{support}(R)} = \frac{P(D, R)}{P(D) \times P(R)} = \frac{P(D \Rightarrow R)}{P(D)}
\]

Roughly, \( \text{lift}(D \Rightarrow R) = 1 \) denotes that the drug and the ADR are statistically independent with each other, \( \text{lift}(D \Rightarrow R) > 1 \) positively correlated, and \( \text{lift}(D \Rightarrow R) < 1 \) negatively correlated. For both leverage and lift, the higher the values, the stronger the signals.

Association mining techniques are often used on structured databases with explicit transactions and items. However, in our case, we are dealing with semi-structured consumer contributed contents, which means we need to first extract the items (drugs and ADRs in our case) and then perform association mining on them. In
this work, we extract the items using external resources: a list of drugs and ADRs lexicon built from CHV.

3.3.2 Temporal Dimension. As stated before, if we ignore the temporal dimension and use the whole data set to compute such measures as confidence, leverage or lift, some strong signals that may only occur during a certain time interval could go undetected as the analysis only concentrates on static associations over the whole dataset. Therefore, in this work, we view temporal dimension an important component in ADR detection. For each drug, we divide all the threads about this drug into a group of subsets, each of which only contains threads happening in a specific year, meaning that each year’s dataset is considered as a process unit. In this work we only consider whole years as analysis unit instead of quarters or months to ensure the data sufficiency. Then association mining will be performed on each time interval of \( p \) year(s) with \( a \) overlapping year(s) between two consecutive intervals, where \( p \) and \( a \) are determined empirically. For instance, given a time span from year 2000 to 2002, parameters \( p = 2 \) and \( a = 1 \), we are going to split the dataset into 2 separate subsets, namely threads happening from 2000 to 2001 and those happening from 2001 to 2002 where year 2001 is overlapped. Then association mining algorithm will be performed on each of the two subsets. Such temporal analysis would help us monitor the changing trend of each drug-ADR association and detect ADRs as early as possible.

Formally, given a set of time-stamped dataset \( X \) over a time domain \( D \), we use \( X(p) \) to denote a subset of \( X \), which contains all data with timestamps belonging to time interval \( p = [t_1, t_2] \). There could be overlapping between two consecutive time intervals. The problem of mining temporal associations in this study is to discover drug-ADR associations during each interval \( p \) and to monitor how these associations change over the time domain \( D \) in terms of different measures.

3.3.3 Algorithm. We developed an algorithm to detect ADR signals from online health communities. We used an association mining technique with temporal analysis and calculated the value of three measures – confidence, leverage, and lift. Below is the pseudo code for the whole process.

**ALGORITHM 1.** Association Mining Algorithm with Temporal Analysis

**Input:** \( p = [t_1, t_2] \) for time interval; \( a \) for overlapping between two consecutive intervals; maximum number of \( m \) for n-gram and \( m \) could be determined by the longest phrase in ADRs lexicon; datasets \( X \); drug list; ADRs lexicon.

**Output:** drug-adverse reaction associations and their confidence, leverage, and lift values in each \( p \).

1: for each drug  
2: for each thread \( i \) such that timestamp \( t \) of \( i \) is in the interval \( [t_1, t_2] \)
3: generate a list of j-gram \((j = 1, 2, \ldots, m)\)
4: for each ADR  
5: compare each j-gram with the ADR
6: if matching then
7: number of threads for this drug-ADR association + 1
8: end if
9: end for
10: end for
11: for each drug-ADR association \( k \)
12: \( \text{support}(k) = \frac{(\text{number of threads containing } k \text{ in } [t_1, t_2])}{\text{(total number of threads for all drugs in } [t_1, t_2])} \)
13: \( \text{support}(\text{drug in } k \text{ in } [t_1, t_2]) = \frac{(\text{number of threads containing } \text{drug in } [t_1, t_2])}{\text{(total number of threads for all drugs in } [t_1, t_2])} \)
14:   support(ADR in k in [t1, t2]) = (number of threads containing ADR in [t1, t2]) / (total number of threads for all drugs in [t1, t2])
15:   confidence(k) = support(k) / support(drug in k in [t1, t2])
16:   leverage(k) = support(k) – support(drug in k in [t1, t2]) × support(ADR in k in [t1, t2])
17:   lift (k) = support (k) / (support (drug in k in [t1, t2]) × support (ADR in k in [t1, t2]))
18: end for
19: If p is not the last time interval then move to next time interval p’ that has overlapping a with p and go to line 2
20: else end if

4. EXPERIMENT
4.1 Dataset Preparation

In this work, we used the discussion threads of drugs in MedHelp.org as our data source. The search engine\(^2\) provided by MedHelp is used to search for all the threads about a certain drug. From the returned search results, 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. The “Posts” section is where health consumers feel free to talk about anything about the drug by either starting a new post or commenting on other people’s posts. To effectively detect ADRs with our proposed techniques, the drug should have active discussion in MedHelp. Therefore, we targeted 20 drugs with more than 500 threads for each of them, and collected all the original posts and comments of these drugs. The reason we aim at drugs with relatively larger number of threads is to ensure we have sufficient data to detect the corresponding ADRs being discussed by consumers. The literature has reported the underreporting problem in drug safety signal detection when the FDA’s spontaneous reporting system is used as the data source [Ji et al. 2011]. The social media data has overcome this limitation to some extent but yet it is not perfect. Detailed information about the drugs will be provided in section 4.2.

4.2 Gold Standard

As mentioned before, current post-marketing surveillance in United States primarily depends on FDA’s FAERS system from which MGPS algorithm is adopted to detect potential signals [Szarfman et al. 2004], and then alert information will be released on FDA’s website\(^3\) along with the alerted timestamp 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 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”\(^4\). In this study, we used 4 other ADRs as another part of our gold standard.

Table I demonstrates the gold standard we used in our experiment, including the names of 10 drugs, number of threads collected from MedHelp, ADR for each drug,

\(^2\)http://www.medhelp.org/search
\(^3\)http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm111085.htm
\(^4\)http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm
Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis

gold standard type and corresponding year in which the alerting or labeling revision happened. The other 10 drugs in our dataset, which do not have any of the above mentioned 10 ADRs, are Adenosine, Biaxin, Epogen, Gadolinium, Geodon, Lantus, Lunesta, Risperdal, Vyvanse and Zyprexa. The time range of the threads is from 1997 to 2012. Therefore, there are 14 Drug ⇒ ADR associations that we are investigating such as Lansoprazole ⇒ Diarrhea, Prozac ⇒ Depression, Zocor ⇒ Kidney Disease etc. Considering that it often takes FDA quite a long time – usually years – from collecting spontaneous reports to releasing an alert, which is also true for a drug labeling revision process, it would be meaningful to detect strong ADR signals before FDA’s alert or labeling revision. Such results could help FDA and drug production companies to start an earlier investigation and shorten the whole time span of post-marketing safety surveillance.

Table I. Experiment Dataset

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Number of Threads</th>
<th>ADR</th>
<th>Gold Standard Type</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lansoprazole</td>
<td>591</td>
<td>Diarrhea</td>
<td>FDA alert: Associated diarrhea can be associated with stomach acid drugs known as proton pump inhibitors (PPIs).</td>
<td>2012</td>
</tr>
<tr>
<td>Luvox</td>
<td>566</td>
<td>Heart Disease; Suicidal</td>
<td>FDA alert: 1. Selective serotonin reuptake inhibitor (SSRI) antidepressant use during pregnancy and reports of a rare heart and lung condition in newborn babies. 2. Suicidal Thoughts or Actions in Children and Adults.</td>
<td>2006; 2005</td>
</tr>
<tr>
<td>Prozac</td>
<td>718</td>
<td>Depression; Suicidal</td>
<td>FDA alert: FDA is highlighting that adults being treated with any type of antidepressant medication should be watched closely for worsening of depression and for increased suicidal thinking or behavior.</td>
<td>2005; 2005</td>
</tr>
<tr>
<td>Heparin</td>
<td>1462</td>
<td>Diarrhea</td>
<td>FDA alert: The adverse events have included allergic or hypersensitivity-type reactions, with symptoms such as low blood pressure, angioedema, shortness of breath, nausea, vomiting, diarrhea, and abdominal pain.</td>
<td>2008</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>794</td>
<td>Kidney Disease</td>
<td>FDA alert: The most serious form of myopathy, called rhabdomyolysis, can damage the kidneys and lead to kidney failure, which can be fatal.</td>
<td>2008</td>
</tr>
<tr>
<td>Zocor</td>
<td>917</td>
<td>Kidney Disease</td>
<td>FDA alert: The most serious form of myopathy, called rhabdomyolysis, can damage the kidneys and lead to kidney failure, which can be fatal.</td>
<td>2008</td>
</tr>
<tr>
<td>Concerta</td>
<td>1058</td>
<td>Hypertension; Blurred vision</td>
<td>Labeling revision: adverse reactions addition.</td>
<td>2006</td>
</tr>
<tr>
<td>Cialis</td>
<td>745</td>
<td>Stroke</td>
<td>Labeling revision: adverse reactions addition.</td>
<td>2005</td>
</tr>
<tr>
<td>Elidel</td>
<td>619</td>
<td>Cancer; Skin discoloration</td>
<td>1. FDA alert: The FDA has issued a public health advisory to inform healthcare professionals and patients about a potential cancer risk from use of Elidel. 2. Labeling revision: adverse reactions addition.</td>
<td>2005; 2007</td>
</tr>
<tr>
<td>Tacrolimus</td>
<td>583</td>
<td>Cancer</td>
<td>FDA alert: The FDA has issued a public health advisory to inform healthcare professionals and</td>
<td>2005</td>
</tr>
</tbody>
</table>
4.3 Experiment Results & Discussion

In our experiment, we used confidence, leverage and lift to calculate the interestingness of each association. We set the maximum number m for n-gram as 3 because the longest term of ADR we obtained from CHV consisted of three words after pre-processing. As for temporal dimension, we compared 6 different types of combinations of p (time interval) and a (overlapping year), as shown in Table II, to see which one has the best performance. Since the last thread we collected from MedHelp was created in 2012, it is highly possible that we did not gather all the data in 2012. Therefore, in the work, we only consider the data from 1997 to 2011.

<table>
<thead>
<tr>
<th>Type</th>
<th>p Value</th>
<th>a Value</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1+0</td>
<td>1</td>
<td>0</td>
<td>Time interval is 1 year, no overlapping.</td>
</tr>
<tr>
<td>2+0</td>
<td>2</td>
<td>0</td>
<td>Time interval is 2 years, no overlapping.</td>
</tr>
<tr>
<td>2+1</td>
<td>2</td>
<td>1</td>
<td>Time interval is 2 years, 1-year overlapping.</td>
</tr>
<tr>
<td>3+0</td>
<td>3</td>
<td>0</td>
<td>Time interval is 3 years, no overlapping.</td>
</tr>
<tr>
<td>3+1</td>
<td>3</td>
<td>1</td>
<td>Time interval is 3 years, 1-year overlapping.</td>
</tr>
<tr>
<td>3+2</td>
<td>3</td>
<td>2</td>
<td>Time interval is 3 years, 2-year overlapping.</td>
</tr>
</tbody>
</table>

4.3.1 Strong Signals for Single Time Interval. First of all, we conducted an experiment to detect statistically strong signals for single time interval. It means that as long as a statistically significant signal appears in any single time interval, we assume that the signal is detected and worth further investigation. The following contents show the experiment results for the 10 ADRs using type 1 parameters (p = 1, a = 0). One important finding is that, for most of the associations that we are supposed to identify, there is an increasing trend after the year of FDA’s alert. This is reasonable because there would be more health consumers discussing the drug after FDA’s alert releases.

Diarrhea. Fig. 1 plots the changing trend of values of confidence, leverage and lift with regard to association Drug ⇒ Diarrhea. Purple line denotes drug lansoprazole whereas blue line denotes heparin, both of which are alerted by FDA to cause diarrhea. Gray line (Avg.) stands for the average value of all 20 drugs. As we can see, for lansoprazole, its value is higher than the average of all drugs in terms of all three measures for all time intervals except 1997, 2001 and 2004. However, only higher value does not mean that the signal is strong. In order to verify the statistical significance of the difference between FDA alerted drug and other ones, we also performed one sample t-test (Confidence Interval Percentage = 95%). It shows that the first time interval that has significant difference in terms of all the three measures between lansoprazole and the average of all the drugs is 1998. Considering that FDA’s “official” alert year for lansoprazole to cause diarrhea is 2012, our approach is able to detect the strong signal much earlier. Unfortunately, in terms of all the three measures, the curve of heparin is below that of average across the whole time span, meaning that the signal of association Heparin ⇒ Diarrhea is too weak to be identified. The reason may lie in the fact that diarrhea is a very common adverse
reaction for many drugs, so the signal of heparin causing diarrhea could be covered (lower value) by other drugs.

Heart Disease. Fig. 2 illustrates the experiment results of heart disease. Similarly, blue line represents FDA alerted drug luvox that could cause heart disease whereas purple line is the average value of all 20 drugs. As we can see, similar trend could be discovered for luvox in terms of all three measures and they fluctuate a lot. Confidence curve reaches its peak in 2005 whereas lift in 2006. Leverage also has quite high values in 2005 and 2006. Considering that FDA released the alert in 2006 that luvox would cause heart disease, it is not surprising to see a lot of consumers talk about it in the two years. One sample t-tests (Confidence Interval Percentage = 95%) demonstrate that the first time interval when significant difference appeared for luvox is 2003, meaning again that strong signals are able to be detected using our proposed approach 3 years earlier than FDA’s alerting time.
Depression. Fig. 3 gives us a clear visual that the value of prozac is always higher than the average value of all drugs in terms of all three measures since 1999. For both confidence and leverage, the values go down gradually after they reach the peak in 1999 whereas for lift, the value ascends with fluctuations and reaches a local maximum in 2004, right before FDA’s alert year 2005. 1999 is the first time interval that all confidence, leverage and lift of prozac are significantly different from the average values of other drugs according to one sample t-test (Confidence Interval Percentage = 95%). Obviously, our method is able to detect strong signals of prozac causing depression much earlier than its “official” time.
Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis

Fig. 3. Experiment Results for Depression

**Suicidal.** Luvox and prozac are two drugs alerted by FDA in 2005 to cause suicidal thoughts. As shown in Fig. 4, confidence, leverage and lift of both drugs are above average curve in most of the time intervals, which means that, in general, the signal of luvox and prozac causing suicidal thoughts is stronger than any other drugs with no alerts. With regard to confidence and leverage, luvox reached the peak in 2002. Based on the curves, we can infer that there were an increasing number of consumers discussing the ADR caused by luvox from 1999 to 2003, way earlier before FDA released the alert in 2005. Similar pattern could be discovered for prozac, which hit the highest confidence and leverage in 2002. T-tests (Confidence Interval Percentage = 95%) show that we can detect strong signals in 1999 for luvox and 2002 for prozac, which precede the FDA alert year quite a lot. We can also observe that there is an increasing trend in 2006, right after FDA released the alert. This is reasonable because there may be more health consumers discussing the drug after FDA's alert.

![Confidence, Leverage, and Lift Curves for Depression](image)

Fig. 4. Experiment Results for Suicide

**Kidney Disease.** Fig. 5 presents the changing trend of confidence, leverage and lift over time for drugs simvastatin and zocor which are reported by FDA to cause kidney disease. As we can see, for all three measures, both curves fluctuate a lot and reach the peak in 2001 and 2000 for simvastatin and zocor respectively, which presents that there are an extremely large number of health consumers discussing kidney disease caused by the two drugs in the two years. T-tests (Confidence Interval Percentage = 95%) demonstrate that we can detect significantly strong signals for simvastatin and zocor causing kidney disease in 2001 and 2000 respectively. Compared with FDA's alerting year – 2008 – for both drugs, we are capable of identifying the signals much earlier.
Stroke, Skin Discoloration & Blurred Vision. Cialis was originally approved by FDA in 2003, and stroke is a new adverse reaction added on the label of Cialis in 2005. As we can see in Fig. 6, all three measures reach their peak in 2004, go down in 2005 and 2006, and ascend again since 2007. T-test (Confidence Interval Percentage = 95%) shows that statistically significant signal appeared in 2004, a year earlier than the time of labeling revision. Skin discoloration is a new ADR added on the label of elidel in 2007, and Fig. 7 demonstrates the experiment results. Clearly, all three curves ascend and exceed the average since 2007 and the signal is statistically significant in this year according to T-test. However, we cannot detect the signal before the labeling revision. The reason could lie in the fact that skin discoloration is a rare adverse reaction and not too many patients discussed it. However, the labeling revision in 2007 drew people’s attention as well as a lot of discussion, which led the curves to take off since 2007. Blurred vision is a new adverse reaction added on the label of concerta in 2006. As we can see in Fig. 8, the three curves fluctuate a little and arrive at the peak in 2003. T-test (Confidence Interval Percentage = 95%) demonstrate that we are able to detect significantly strong signal in 2003 in terms of all three measures.
Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis

Fig. 6. Experiment Results for Stroke

Fig. 7. Experiment Results for Skin Discoloration
Cancer & Hypertension. In 2005, the FDA issued a public health advisory to inform healthcare professionals and patients about a potential cancer risk from use of both elidel and tacrolimus. Unfortunately, as we can see in Fig. 9, we are not able to detect the signals before FDA’s alert because all the 6 curves are below the average in most of the time intervals before 2005. It may be because very few patients experienced this kind of adverse reaction. Even if they are under the risk of cancer, they may not be aware of it and then rarely discuss it. However, after FDA’s alert in 2005, the curve of tacrolimus took off and reached the peak, which may be because the alert drew a lot of attention and discussion. Hypertension is another new ADR that was added on the label of concerta in 2006. However, we cannot detect strong signal of concerta causing hypertension according to Fig. 10, in which all three measure curves either go below or near the average before 2006. The reason could be that hypertension is a rare ADR that could be caused by concerta and therefore not too many health consumers talked about it. Also, some drugs such as simvastatin and zocor are used to treat hypertension, so the values of the associations simvastatin→hypertension and zocor→hypertension could be relatively high, which is the case according to our experiment, and increase the average values.
According to our experiment results, we found that we cannot simply apply a threshold on confidence, leverage, and lift for all drugs and ADRs to detect the drug-adverse reaction associations. The confidence, leverage, and lift values vary substantially across different drugs and especially across different ADRs. By applying a simple threshold, we can easily miss the true drug-adverse reaction associations or identify many false drug-adverse reaction associations. This can be reflected by the diverse discussions on drugs and the variation of vocabulary usage in describing ADRs. Alternatively, for a specific ADR, we can use top K to extract the drugs that deserve further investigation. In this case, K is a threshold. We can also compute the significance difference between the association metrics of each drug and a particular ADR with the average of all drugs, which means the significance level is a threshold.
To sum up, by combining associations mining and temporal analysis, not only can we detect drug safety signals for almost all the associations that we are supposed to find, we can also identify those signals way earlier than FDA's alert time. By setting parameters $p = 1$ and $a = 0$ and only considering strong signals for single time interval, obviously, 9 out of 14 drug→ADR associations can be successfully detected earlier than official alerts or labeling revision, which makes the total recall up to 0.64. In terms of detected year, all the three measures reached the same performance for all of the drug→ADR associations. Compared with our previous study [Yang, Yang, et al. 2012], the recall decreased slightly. The reason is that our previous study only used 10 drugs and 5 ADRs whereas this work collected 20 drugs and 10 ADRs, and some of the ADRs are not easy to detect such as cancer, stroke, and skin discoloration which were not included in the previous research. However, the previous one was based on static association mining over the whole time span, which did not consider the temporal factor, thus not being able to determine if we can detect signals of ADR before FDA's alerts. Another interesting discovery of our experiment is that we also identified a few number of drug→ADR associations that have relatively higher values during most of the time intervals in terms of confidence, leverage, or lift but haven't either been alerted by FDA or go through labeling change such as biaxin→diarrhea, simvastatin→heart disease, zocor→heart disease, heparin→stoke, simvastatin→stroke, and zocor→stroke. These can be the results of false positives but it can also be potential ADR signals that have not been identified yet, which may deserve further investigation. Compared with the approaches mentioned by Kubota et al. [Kubota et al. 2004], our method has two advantages. First, we have higher recall in terms of detecting possible signal than BCPNN and GPS. Second, we are able to detect early signals, which cannot be achieved by only using metrics such as ROR and PRR.

However, as we can see from Fig. 1 to Fig. 10, most measure curves fluctuate a lot because we only set time interval as one year without overlapping. In this scenario, outliers could dominate the data and make the curve go up and down abruptly. Therefore, based on Table II, we also conducted experiments using other 5 types of combinations of different values of $p$ and $a$.

Fig. 11 shows the curves of confidence value of different parameter settings for associations drug→suicide. As we can see, compared with $1+0$, other curves are relatively smoother, which are more convenient for observation. We also calculated the recall to compare different parameter settings based on if we can detect signals earlier than FDA's alert or labeling revision by using confidence, leverage and lift. As shown in Table III, when we only aim at detecting strong signals for single time interval, parameter setting $1+0$ has the best performance, and under this scenario, confidence, leverage and lift have the same recall. When considering all parameter settings, confidence and lift have the same performance, and leverage has higher recall than both of them.
Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis

Fig. 11. Confidence Curves of Drug→Suicide Using 6 Different Types of Parameters Setting

Table III. Recall of Confidence, Leverage and Lift for All Drug-ADR Associations under Different Parameter Settings for Single Time Interval

<table>
<thead>
<tr>
<th>Parameter Setting</th>
<th>Confidence</th>
<th>Leverage</th>
<th>Lift</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1+0</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>2+0</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>2+1</td>
<td>0.64</td>
<td>0.57</td>
<td>0.64</td>
<td>0.62</td>
</tr>
<tr>
<td>3+0</td>
<td>0.36</td>
<td>0.43</td>
<td>0.36</td>
<td>0.38</td>
</tr>
<tr>
<td>3+1</td>
<td>0.43</td>
<td>0.50</td>
<td>0.43</td>
<td>0.45</td>
</tr>
<tr>
<td>3+2</td>
<td>0.50</td>
<td>0.57</td>
<td>0.50</td>
<td>0.52</td>
</tr>
<tr>
<td>Average</td>
<td>0.51</td>
<td><strong>0.54</strong></td>
<td>0.51</td>
<td></td>
</tr>
</tbody>
</table>

4.3.2 Strong Signals for Consecutive Time Intervals. In the previous section, we only considered detecting statistically strong signals for single time interval. However, in this case, we did not exclude the situation that some strong signals may just appear by chance. For example, with regard to association Luvox→heart disease (Fig. 2), if we depend on single time interval, the first strong signal could be detected in 2003 by using parameter setting 1+0. However, as we can see, the curves dropped substantially in 2004 and the signal was no longer statistically significant, which means the strong signal in 2003 could just happen by chance. In order to better convince people that the detected signals deserve further investigation, they should stay statistically strong for consecutive time intervals instead of just one. Therefore, in this study, we conducted another set of experiment to compare the performance of
confidence, leverage and lift in terms of recall of detecting strong signals for 2 and 3 consecutive time intervals before FDA’s alert or labeling revision.

According to Table IV, parameter settings 2+1 and 3+2 have the best performance for detecting strong signals for 2 and 3 consecutive time intervals respectively. In both of the two scenarios, leverage outperformed confidence and lift that have the same performance. The reason that confidence and lift have the same recall is that the lift value of an association equals to the product of the corresponding confidence value and a constant \( \frac{1}{\text{support}(R)} \) – according to their formulas, which makes them have the same T-test results. Compared with lift, leverage uses subtraction instead of division between the two items (refer to their formulas in section 3.3.1), which could generate less outliers that would to a great extent affect the mean value. Also, we can observe that the performance with overlapping is better than that without overlapping (2+1 is better than 2+0; 3+1 and 3+2 are better than 3+0).

**Table IV.** Recall of Confidence, Leverage and Lift in Detecting Strong Signals for 2 and 3 Consecutive Time Intervals

<table>
<thead>
<tr>
<th>Parameter Setting</th>
<th>Strong Signal for 2 Consecutive Time Intervals</th>
<th>Strong Signal for 3 Consecutive Time Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Confidence</td>
<td>Leverage</td>
</tr>
<tr>
<td>1+0</td>
<td>0.29</td>
<td>0.36</td>
</tr>
<tr>
<td>2+0</td>
<td>0.21</td>
<td>0.29</td>
</tr>
<tr>
<td>2+1</td>
<td>0.43</td>
<td>0.50</td>
</tr>
<tr>
<td>3+0</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>3+1</td>
<td>0.29</td>
<td>0.36</td>
</tr>
<tr>
<td>3+2</td>
<td>0.36</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Overall speaking, taking into consideration either 2 or 3 consecutive time intervals, parameter setting 2+1 combined with leverage has the highest recall in terms of confidence, leverage and lift in the first scenario. Therefore, we set up the criterion as follows for deciding at which point in time we would raise an alarm that a potential novel drug-ADR association is found: when a signal of drug-ADR association stays statistically strong for 2 consecutive time intervals under the parameter setting 2+1 (2-year time interval plus 1-year overlapping), this signal should be alarmed for further investigation. Table V shows the detected time for all the drug⇒ADR associations that are supposed to be identified using 2+1 for 2 consecutive time intervals.
Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis

Table V. Detected Year for 14 Associations Using 2+1 for 2 Consecutive Time Intervals

<table>
<thead>
<tr>
<th>Drug→ADR</th>
<th>FDA Alert or Labeling Revision Year</th>
<th>Detected Year (2 Consecutive Time Intervals)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Confidence</td>
</tr>
<tr>
<td>Lansoprazole→Diarrhea</td>
<td>2012</td>
<td>1999</td>
</tr>
<tr>
<td>Heparin→Diarrhea</td>
<td>2008</td>
<td>NA</td>
</tr>
<tr>
<td>Luvox→Heart Disease</td>
<td>2006</td>
<td>NA</td>
</tr>
<tr>
<td>Prozac→Suicidal</td>
<td>2005</td>
<td>2004</td>
</tr>
<tr>
<td>Zocor→Kidney Disease</td>
<td>2008</td>
<td>NA</td>
</tr>
<tr>
<td>Cialis→Stroke</td>
<td>2005</td>
<td>NA</td>
</tr>
<tr>
<td>Elidel→Skin Discoloration</td>
<td>2007</td>
<td>NA</td>
</tr>
<tr>
<td>Elidel→Cancer</td>
<td>2005</td>
<td>NA</td>
</tr>
<tr>
<td>Tacrolimus→Cancer</td>
<td>2005</td>
<td>NA</td>
</tr>
<tr>
<td>Concerta→Hypertension</td>
<td>2006</td>
<td>NA</td>
</tr>
</tbody>
</table>

5. CONCLUSION

Since adverse drug reactions represent a serious health problem all over the world, how to detect ADRs in an early stage has drawn many researchers' attention and efforts. There are an increasing number of studies focused on this area and many techniques have been proposed to detect ADRs based on various data sources such as spontaneous reporting data, electronic health record or pharmaceutical databases. However, these data sources are limited by either high cost or under-reporting ratio or privacy issues. With the development of Web 2.0 and Health 2.0 technology, online health communities are flourishing and providing a great platform for health consumers to conveniently share their health-related experience. Every day, an enormous amount of up-to-date information is generated from those websites that are publicly available, making it highly possible and promising to detect ADRs as soon as possible from them. In this study, we proposed to use associations mining and temporal analysis to detect ADRs. We collected posts and comments of 20 drugs from MedHelp, a popular online health community, extracted 14 adverse reactions, either alerted by FDA or added on drug labels, with their alert releasing or labeling revision time as gold standard, and utilized measures confidence, leverage and lift to identify ADR signals. We conducted two sets of experiments, namely strong signals for single time interval and strong signals for consecutive time intervals. The first set of experiments showed that our approach detected 9 out of the 14 association earlier than FDA’s alerts or labeling revision. We also discovered a few number of associations that have not been alerted by FDA but carried high signal values. In the second set of experiments, we detected statistically strong signals for consecutive
time intervals and determined the criterion for raising an alarm: when a signal of drug-ADR association stays statistically strong for 2 consecutive time intervals under the parameter setting $2+1$ (2-year time interval plus 1-year overlapping), this signal should be alarmed for further investigation.

In the future, our study could be further improved in several directions. First, we intend to extend our database by increasing both the number of drugs and adverse reactions to better verify the effectiveness of the proposed techniques. Second, early detection of drug-drug interactions could be another direction of our future research. Third, in this study, our current method largely depends on the quality of CHV lexicon which includes the common misspellings and variations of terms because CHV is constructed by extracting consumer contributed expressions of healthcare content. However, similar to NLP and stemming, CHV may also miss some misspellings or errors that are not commonly used, so how to expand the vocabulary to include as many health consumers’ ADR-related expression as possible remains a challenging task. Fourth, this work only considers whole years as analysis unit instead of quarters or months to ensure the data sufficiency. However, it is possible that a burst of discussions on a particular ADR for a drug may arise when a large number of health consumers experience the same adverse reaction, especially when the drug is first released. Delaying the detection may not be wise in this case. In our future work, we shall use one month as the window initially. If insufficient data is collected for detection, the window will be changed to one quarter or one year. Last but not the least, our current work mainly focused on content analysis of online health community data, how to incorporate network structure analysis such as link mining is another important direction for future study.

6. REFERENCE


Brant W. Chee, Richard Berlin and Bruce Schatz. Year. Predicting adverse drug events from personal health messages. In Proceeding of the AMIA Annual Symposium. 217.


Jiawei Han, Micheline Kamber and Jian Pei. 2006. Data mining: concepts and techniques: Elsevier.


Huidong Jin, Jie Chen, Hongxing He, C. Kelman, D. McAullay and O'Keefe C. M. 2010. Signaling potential adverse drug reactions from administrative health databases. IEEE Transactions on Knowledge and Data Engineering, 22, 6 (2010), 839.


Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis


Xiao Liu and Hsinchun Chen. 2013. AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums *Smart Health* (pp. 134-150); Springer.


