Artificial Intelligence In Medicine xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

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

Christopher C. Yang*, Haodong Yang

College of Computing and Informatics, Drexel University, United States

ARTICLE INFO

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

ABSTRACT

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

1. Introduction

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

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

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

* Corresponding author.

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

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

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

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

2. Literature review

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

2.1. Pharmacovigilance using patient-contributed content

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

2.1.1. ADR detection

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

2.1.2. DDI detection

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

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