USING BIG DATA ANALYTICS AND STATISTICAL METHODS FOR IMPROVING DRUG SAFETY

By

BEHROOZ DAVAZDAHEMAMI

Bachelor of Science in Industrial Engineering
Isfahan University of Technology
Isfahan, Iran
2009

Master of Science in Industrial Engineering
University of Tehran
Tehran, Iran
2012

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2019

USING BIG DATA ANALYTICS AND STATISTICAL METHODS FOR IMPROVING DRUG SAFETY

Dissertation Approved:

Dr. Dursun Delen

Dissertation Adviser

Dr. Rick Wilson

Dr. David Biros

Dr. Bruce Benjamin (Outside member)

Name: BEHROOZ DAVAZDAHEMAMI

Date of Degree: MAY, 2019

Title of Study: USING BIG DATA ANALYTICS AND STATISTICAL METHODS FOR IMPROVING DRUG SAFETY

Major Field: BUSINESS ADMINISTRATION (MSIS)

Abstract:

This dissertation includes three studies, all focusing on utilizing Big Data and statistical methods for improving one of the most important aspects of health care, namely drug safety. In these studies we develop data analytics methodologies to inspect, clean, and model data with the aim of fulfilling the three main goals of drug safety; detection, understanding, and prediction of adverse drug effects.

In the first study, we develop a methodology by combining both analytics and statistical methods with the aim of detecting associations between drugs and adverse events through historical patients' records. Particularly we show applicability of the developed methodology by focusing on investigating potential confounding role of common diabetes drugs on developing acute renal failure in diabetic patients. While traditional methods of signal detection mostly consider one drug and one adverse event at a time for investigation, our proposed methodology takes into account the effect of drug-drug interactions by identifying groups of drugs frequently prescribed together.

In the second study, two independent methodologies are developed to investigate the role of prescription sequence factor on the likelihood of developing adverse events. In fact, this study focuses on using data analytics for understanding drug-event associations. Our analyses on the historical medication records of a group of diabetic patients using the proposed approaches revealed that the sequence in which the drugs are prescribed, and administered, significantly do matter in the development of adverse events associated with those drugs.

The third study uses a chronological approach to develop a network of approved drugs and their known adverse events. It then utilizes a set of network metrics, both similarity- and centrality-based, to build and train machine learning predictive models and predict the likely adverse events for the newly discovered drugs before their approval and introduction to the market. For this purpose, data of known drug-event associations from a large biomedical publication database (i.e., PubMed) is employed to construct the network. The results indicate significant improvements in terms of accuracy of prediction of drug-evet associations compared with similar approaches.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Today every drug goes through a long journey which ,on average, takes 10-15 years (Iizuka, 2007) from the first day it is discovered until its approval by healthcare authorities and introduction to the market. This involves numerous clinical trials aiming at ensuring efficacy and safety of the drug. In other words, the clinical trials are meant to ensure that, first, a given drug efficiently functions as it is intended to treat disease(s), and second, it does not cause any serious side effects to the patients.

Pharmacovigilance, also referred to as drug safety surveillance, has been defined as the science and activities relating to the detection, assessment, understanding, prediction, and prevention of adverse effects or any drug problem (Arthur et al., 2002). In the pharmacovigilance terminology, Adverse Drug Events (ADE) is a general term that refers to any injury caused by a medication. This injury can be an unintended effect of the recommended (i.e. prescribed or labeled) usage of a drug, the off-label usage of a drug, or a medication error (Karimi, Wang, Metke-Jimenez, Gaire, & Paris, 2015). Adverse Drug Reactions (ADR), on the other hand, is a more specific term that only refers to those injuries directly caused by proper usage of medication, and not medication errors (Karimi et al., 2015).

Pharmacovigilance activities are deemed to be important from both health and business perspectives. It is reported that ADEs in each year cause more than 2 million injuries, hospitalizations, and deaths only in the United States (Lazarou, Pomeranz, & Corey, 1998) that incur more than 75 billion dollars to the patients, healthcare system, and insurance agencies (Ahmad, 2003).

1.1. PRE-APPROVAL PHARMACOVIGILANCE

Although pharmacovigilance activities for any drug begin a long time before introducing it to the market through numerous pre-approval clinical trials, such efforts are typically too limited to identify all the potential ADEs that may occur. First, they are often short in time and involve a limited sample size (Zeng, Kogan, Ash, Greenes, & Boxwala, 2002). Moreover, they do not fully represent the target population of the drug as they may exclude patients who receive other medications, focus on a particular age group (e.g. elderly) of patients, and those who have complicated medical conditions (Karimi et al., 2015). Also as Stephens and Talbot (1985) noted, clinical trials may not detect ADEs with very low incident rates.

1.2. POST-APPROVAL PHARMACOVIGILANCE

Due to the mentioned points in the previous section, post-approval ADEs have always been a major global health concern since a considerable proportion of ADEs remain to be revealed in the post-approval stage of the drugs' lifetime. In some cases, those unrevealed ADEs in the pre-approval stage have even caused thousands of deaths; A classic example of such cases is *Rofecoxib*; an NSAID approved in 1999 that became highly welcomed by the physicians in a short time. The drug was originally aimed to treat acute pains and Osteoarthritis, but after a while turned out to cause heart attacks in more than 100,000 patients and ended up being withdrawn by the FDA in 2004.

Unlike pre-approval stage which is highly experiment-based, post-approval pharmacovigilance is highly driven by historical data analysis. Given the critical importance of post-approval pharmacovigilance from both healthcare and business perspectives, and in response to the challenge

posed by large quantities and complexities of data sources that needed to be examined, various data mining algorithms have been developed in the recent decades to bring about improvement in drug safety surveillance. Researchers have used various approaches, all with a heavy reliance on information systems for collection, manipulation, and analysis of data. Four main types of data source have been identified in the literature to be used in ADE studies. The following four sub-sections introduce these resources and mention prior research conducted using each.

1.2.1. SPONTANEOUS REPORTING SYSTEMS

As an effort to rapidly detect and prevent ADEs, many countries and international organizations have run Spontaneous Reporting Systems (SRSs), systems designed to allow patients and professionals to submit their reports of suspected ADEs. This includes the World Health Organization's (WHO) Individual Case Safety Reports (ICSR) database, the TGA Adverse Drug Reaction System (ADRS) in Australia, the yellow card system of Medicines and Healthcare products Regulatory Agency (MHRA) in the UK, and the FDA Adverse Event Reporting System (FAERS) in the US (Karimi et al., 2015).

Although SRSs have been the main source to detect likely ADE cases for years and multiple studies were conducted based on them,(Cai et al., 2017; DuMouchel, 1999; Lin, Xiao, Huang, Chiu, & Soo, 2010; van Puijenbroek et al., 2002) they still have several limitations such as over-reporting, missing and incomplete data, latency, duplicated reporting and voluntary submission (Harpaz et al., 2013). Due to voluntary submission, for instance, it is estimated that these systems in the US and UK reflect less than 10% of the ADE occurrences.(Inman & Pearce, 1993; Yang, Jiang, Yang, & Tang, 2012) Such shortcomings made pharmacovigilance practitioners shift their focus towards resources that are more efficient for post-marketing drug surveillance.

1.2.2. ELECTRONIC HEALTH RECORDS (EHR)

In the past decade and along with extensive adoption of information systems and technologies in the healthcare industry, Electronic Health Records (EHR) have been widely used in this industry to help

practitioners in the collection, storage, and tracking patients' information. The vast amount of data collected by EHRs along with their increasing availability have made them interesting resources for pharmacovigilance researchers and enabled them to detect ADE signals[1] closer to real-time (Trifirò et al., 2009). Although EHR data is generally more complete than SRSs reports and several studies have been conducted recently using EHRs (Friedman, 2009; Haerian et al., 2012; Harpaz et al., 2012; Harpaz, Haerian, Chase, & Friedman, 2010), yet using them for ADE studies involve challenges like complex data preprocessing requirements and various data documentation styles across different providers (Harpaz et al., 2013).

1.2.3. SOCIAL MEDIA

In the recent years, social media has also been considered as a key data source for collecting drugs' post-marketing feedbacks by multiple researchers (Hoang et al., 2016; J. Liu, Zhao, & Zhang, 2016; X. Liu & Chen, 2013; Nikfarjam, Sarker, O'Connor, Ginn, & Gonzalez, 2015a; O'Connor et al., 2014). A Pew internet research by Fox and Jones(2009) found that 61% of Americans look for health information online. This is normally done either through healthcare online forums such as 'DailyStrength' and 'PatientsLikeMe'; or through social networks like Facebook and Twitter. Through the social media, people talk about their concerns, seek advice about their health issues, and discuss their medical experiences. Such information, although noisy, is likely to appear there long before it is reported to any SRS or recorded in any EHR (Benton et al., 2011; Leaman et al., 2010). A novel stream of research using Twitter data focuses on automatic detection of ADEs by constant monitoring of tweets posted by patients using text-mining approaches. Sarker et al. (2015) have done a comprehensive review of the studies conducted in this area.

[1] Signal is defined by the World Health Organization (WHO) as information on a possible causal relationship between a drug and an adverse event, which is unknown or incompletely documented (Trifirò et al., 2009).

1.2.4. BIOMEDICAL LITERATURE

Recently, researchers have realized biomedical literature as well as chemical and biological databases as feature-rich sources for pharmacovigilance studies. Databases such as PubMed, PubChem, KEGG, and DrugBank are rich sources of information about drugs, their chemical and biological characteristics, and their identified ADEs. Several studies have been done by employing data- and text-mining techniques on data from these resources(Avillach et al., 2013; Shetty & Dalal, 2011) or even by combining them with other mentioned resources(Duke et al., 2012) to detect or predict ADEs.

1.3. CLASSIFICATION OF PHARMACOVIGILANCE STUDIES

The type of data source we use for a study determines the class of data mining algorithms that can be applied. The following sections discuss various types of data mining approaches used in the pharmacovigilance literature.

1.3.1. METHODS FOR ASSOCIATION DETECTION AND UNDERSTANDING

The main class of data mining approaches widely used in pharmacovigilance research are those designed to detect meaningful association (i.e. signal) for large sets of drug-event pairs with the aim of identifying and prioritizing risk signals. Of course, the identified signals should then be investigated more carefully to verify the causality between drug and event. This approach is especially applied widely to the SRS data. For example FDA actively uses a data mining engine to compute signal scores indicating statistical associations for millions of drug-event combinations in the AERS (Harpaz et al., 2013). Such DM algorithms are in fact an extension of the Disproportionality Analysis (DPA) methods that, for years, were the main statistical methods to discover drug-event associations based on frequency analysis of 2x2 contingency tables (Bate & Evans, 2009). Given that EHR data, as opposed to SRS, involves information on both ADE and non-ADE cases as well as temporal patients information, modified versions of DPA are typically used to analyzed their data. For instance Schuemie (2011) proposed a DPA-based longitudinal approach to detect ADE signals in the EHR data to take into

account the effect of "length of exposure to a drug" on potential adverse events. However, unlike DPA methods which are only able to detect associations involving one drug and one event, DM-based approaches are capable to handle more complex situations such as drug-drug interactions, drug-induced syndromes, and confounding phenomena (Harpaz et al., 2013).

Apart from DPA extensions, the literature also involves studies that employed methods based on logistic regression and unsupervised machine-learning methods for the same purposes.

Logistic regression-based approaches are especially handy when the goal is to handle multiple potential confounding factors[2]. While the traditional approach to control for confounders is stratification, that approach is not very effective in presence of too many confounders. Jewel (2003) argues that in such cases a more appropriate approach to handle confounding is to incorporate all potential confounders as covariates in a logistic regression model. Nevertheless, even the original logistic regression approach is limited in terms of the number of covariates that can handle. Some newer extensions of logistic regression, namely Bayesian Logistic Regression (BLR) models are even capable to handle millions of covariates in the model. Such models have been used in a number of studies such as (Caster, Norén, Madigan, & Bate, 2010) to detect ADE signals from the WHO spontaneous reporting system data controlling for too many confounding factors.

Unsupervised machine-learning approaches are another class of DM algorithms that are used for signal detection. Many studies have used association rule mining to discover multi-item ADE associations (Harpaz, Chase, & Friedman, 2010; Ji et al., 2011; Reps, Aickelin, & Hubbard, 2016). Nevertheless, these methods typically require substantial computing resources which has limited their application in the past. Clustering methods have also been used in a number of pharmacovigilance studies, primarily as an exploratory tool with the aim of summarizing the complex structures in a macroscopic manner.

[2] A confounder by definition is an extraneous variable that mediates an association between two other variables (i.e. drug and event). (Harpaz et al., 2013)

For example, He et al (2004) applied a KNN clustering algorithm to the drug dispensation sequence data from patients with the disease Angioedema to discover potential relationships between drugs and resulting hospital admissions due to adverse events.

Network analysis is another approach that researchers have recently started to use, mostly to discover interesting multidimensional patterns of ADEs. Applying a network approach to the FDA's spontaneous reporting database, for instance, Ball et al (2011a) revealed that the vaccine HPV4 is associated with syncope and seizures in adolescents.

Given the fact that EHR data contains information about both ADE and non-ADE groups of patients, a popular class of approaches applied to such data are those based on comparison of patterns and frequencies across these groups. That involves studies which employ cohort designs, case-control designs, or self-controlled designs to compare the two groups. In the cohort design studies, the idea is that the patterns of ADE occurrence over time must be different among patients who were exposed to a suspicious drug and those who were not; if so, it is likely to say that there is an association between the drug and the event of interest. On the case-control designs, on the other hand, comparison is made between patients who experienced a particular adverse event and those who did not. Different patterns of drug taking between them, then implies likely association. In the self-controlled design, each patient who has experienced the ADE, is treated as both the case and control subject in the study (i.e. during the drug exposure vs non-exposure periods) and the patterns of ADE across the two periods are compared against each other.

Apart from the medication information, electronic medical records also contain a variety of other structured and unstructured data that have been used in the pharmacovigilance research to detect ADE signals. Park et al (2011) for instance used lab reports in an EHR database to identify abnormal lab results and compare their patterns before and after the use of a medication.

1.3.2. METHODS FOR PREDICTING ASSOCIATIONS

Even though most of the research conducted based on SRS and EHR data are focused on detecting signals of associations between drugs and adverse events (either directly or due to interactions) as well as understanding the factors moderating them, recently emergence of some new data sources (such as chemical and biological information of drugs, biomedical literature, patients' online forums, and social networks) has led to efforts to predict ADEs at the early stages of drug's lifecycle and before it affects too many people.

Quantitative Structure-Activity Relationship (QSAR) is a regression-based method widely used in the chemical and biological sciences that primarily aims at predicting biological activity of chemicals (i.e. the response) based on their chemical and molecular structure. Relying on QSAR, and using historical causal drug-event associations, some research is conducted to identify chemical properties of molecules that may correlate with ADEs and thereby to predict potential ADEs of new drugs on the basis of their chemical properties (Matthews et al., 2009; Pouliot, Chiang, & Butte, 2011). Such QSAR models are now being used internally by the FDA to provide decision support information for a variety of purposes (Harpaz et al., 2013).

In another group of studies, text-mining techniques have been applied to the unstructured data collected from biomedical literature (e.g. Shetty & Dalal, 2011) as well as patients' online communities and social networks (e.g. Leaman et al., 2010) with the aim of identifying drugs' potential adverse events earlier than they are reported to the spontaneous reporting systems. Some prominent ADE cases such as the Rofecoxib case have been used in these studies as benchmark to show how the prediction methods were able to predict that case much earlier than it causes more than 100,000 myocardial infarctions and collected from market due to numerous reports filed for it to the FDA's SAERS in 2004.

Another approach used in the literature to predict ADEs is network analysis. Cami et al (2011) used historical drug-event associations to construct a network having both drugs and events as nodes and

their associations represented by edges. They used topological network measures along with drugs' molecular descriptors to train a logistic regression model to predict the likelihood of existence of an edge (i.e. association) for each drug-event pair.

Combining canonical correlation analysis and network-based diffusion, Atias and Sharan (2011) proposed a novel prediction approach and applied that to a public database of drug side effects called SIDER, to predict ADEs of the new drugs. They validated their model by testing it on a set of 692 drugs with known side effects and showed that for 34% of the drugs the top scoring side effect identified by the algorithm matches a known side effect of the drug.

1.4. AN OVERVIEW OF THE CURRENT WORK

The present dissertation work involves three independent studies. To highlight the role of various information systems as well as data analysis methods in pharmacovigilance, in these studies multiple data sources, each relied on an IS artifact, have been utilized and a handful of data mining and statistical methods has been used to analyze that data with the aim of improving drug safety.

In terms of approach and data source, the first study aims at *detecting ADE signals* by applying data mining and statistical methods to EHR transactional data. Specifically, the goal of the first study is to investigate potential adverse reactions of common diabetic drugs in developing acute renal failure.

The second study employs data mining and statistical techniques along with EHR data with the aim of *understanding drug-ADE associations.* The goal in that study is to investigate the role of prescription sequence factor in changing the likelihood of development of adverse events for the already known drug-ADE associations. As a case study, we have focused on acute renal failure as a common and high-risk adverse event to address the research question of the second study.

The third essay deals with another aspect of pharmacovigilance studies, namely *ADE prediction.* In that study a chronological network approach along with multiple machine learning techniques have been

employed with the aim of identifying similarities among the already-approved drugs and the new drugs and then using those similarities to predict potential ADEs of new drugs before their approval. To this end, we have used reported drug-ADE associations mentioned in the biomedical literature (MEDLINE database) as the main data source and have enriched that with data on the target proteins of drugs in the human body (i.e. a biological property of the drugs).

Overall, in the three studies conducted, we have tried to highlight the potential of using IS artifacts (i.e. databases and computer-based data analysis techniques) to contribute to various aspects of drug safety (i.e. detection, understanding, and prediction of signals).

The three essays are presented in chapters 2, 3, and 4, respectively. Also, the last chapter contains summary and conclusion.

CHAPTER II

ESSAY I: THE CONFOUNDING ROLE OF COMMON DIABETES MEDICATIONS IN DEVELOPING ACUTE RENAL FAILURE: A DATA MINING APPROACH WITH EMPHASIS ON DRUG-DRUG INTERACTIONS

ABSTRACT

Longstanding diabetes mellitus is today known as the primary reason for kidney failure in the patients having that condition. While the prior research has studied the confounding role of some frequently prescribed diabetes medications in developing acute renal failure, some rarely prescribed medications are still under-studied in this regard. In addition, even for those drugs studied in the past, inconsistent findings have been reported. In the present study, by extending a data mining framework from the prior research and equipping that with some standard statistical metric from the medical literature we investigate the general confounding role of the common diabetes medications in developing acute renal failure in a large group of patients with diabetes mellitus (Type II). In addition, we assess the stability of the identified confounding roles by taking into account the potential drug-drug interactions between those diabetes medications with a group of drugs already known to have negative effect on the kidney function. Our results suggest the general dominant confounding role for each of the diabetes medications, but also suggests that these roles are unstable across various prescription combinations due to potential

drug-drug interactions, thereby provide an explanation for the inconsistent findings in the literature.

Keywords: *Adverse Drug Reactions; Itemset Mining; Diabetes; Acute Renal Failure; Drug-Drug Interactions*

2.1. INTRODUCTION

Today almost every drug produced and marketed by pharmaceutical companies has a list of likely side effects printed on its label to warn patients about possible harms they may undergo by taking it. Such known side effects are usually the result of several years of research and clinical trials conducted on the drug by the manufacturer after discovery and before introducing it to the market.

In the pharmacovigilance[3] terminology, Adverse Drug Event (ADE) is a general term that refers to any injury caused by a medication. This injury can be an unintended effect of the recommended (i.e. prescribed or labeled) usage of a drug, the off-label usage of a drug, or a medication error (Karimi et al., 2015). Adverse Drug Reactions (ADR) are a subset of ADEs referring to an unexpected harm caused by the normal use of medication at the normal dosage (Karimi et al., 2015). Therefore, ADRs does not have to do with non-prescribed or off-label usage of a drug or medication errors. In the United States, according to the Office of Disease Prevention and Health Promotion (ODPHP), ADRs account for about 2 million hospital stays as well as 3.5 million physician office visits in each year[4]. Also, the cost incurred by each ADR case in community hospitals in the United States is estimated at around $3,000 (Classen, Pestotnik, Evans, Lloyd, & Burke, 1997; Hug, Keohane, Seger, Yoon, & Bates, 2012).

Such considerable costs to the patients, insurance agencies, and the healthcare industry have caused researchers to seek effective ways for detection, prediction, and prevention of ADRs during the past years. The development of Electronic Health Records (EHR) systems in the past decade has provided pharmacovigilance researchers with great opportunities to detect, predict, and understand adverse drug reactions by analyzing real medical transactions.

---

[3] Pharmacovigilance (a.k.a. drug safety surveillance) is a field of science that tries to detect, assess, understand, and prevent harms and injuries caused by medications in all stages of drugs' lifetime (i.e. discovery, clinical trials, pre-marketing, and post-marketing). (World Health Organization)

[4] https://health.gov/hcq/ade.asp

Acute Renal Failure is one of the most common ADRs due to taking medications identified in the literature (Trifirò et al., 2009). The literature has mentioned several drugs with renal failure as one of their main side effects (Ashley, 2018; Cavalieri et al., 2018; Härmark, Van Der Wiel, De Groot, & Van Grootheest, 2007; Perazella, 2003; Singh, Ganguli, & Prakash, 2003). Also, diabetes mellitus is known as the leading cause of chronic and end-stage kidney disease as Loh and Cohen (2009) and Afkarian and colleagues (2016) note that diabetes mellitus accounts for most of the cases of kidney disease in the United States and other developed countries. Whereas common diabetes medications are not known as major causes of renal failure in the literature, there are studies which suggest some confounding roles for these medications in increasing or decreasing the chances of renal failure development. However, those analyses are mostly focused on frequently prescribed diabetes medications (e.g., insulin and metformin); moreover, in some cases, inconsistent confounding roles have been suggested for the same drug by different researchers.

Although various data-driven methods such as disproportionality analysis (Baksh, McAdams-DeMarco, Segal, & Alexander, 2018; Cohen, Houdeau, & Khromava, 2018; Trippe, Brendani, Meier, & Lewis, 2017), text analysis (Harpaz et al., 2013; Nikfarjam et al., 2015a), and network analysis (Cami et al., 2011; Davazdahemami & Delen, 2018) are proposed in the literature to identify drug's potential adverse events, little research has been done on the potential confounding role of drugs in development of adverse events in the presence of other drugs. In fact, the drug-ADR associations are mostly studied in isolation, whereas prior research suggests that unintended drug-drug interactions (DDIs) may help developing an adverse event in these patients.

Almost all the DDI studies in the literature involve investigation of potential reactions between pairs of drugs whereas many interactions might be the result of taking three or more drugs in a time period. In the present study, we extend the framework proposed by Reps et al.

(2016) and apply it to the prescription records of a large set of diabetic patients to: 1) investigate the general confounding role of common diabetes medications, including those infrequently prescribed drugs, controlling for the effect of kidney-damaging drugs; and 2) assess the stability of those confounding roles across various prescription combinations of the same drug (i.e., assessing the potential DDIs) with the aim of explaining inconsistent confounding roles reported in the prior studies.

The remainder of the paper is organized as follows. In section 2 through a review of the literature, we discuss the pre- and post-approval ADR research as well as various approaches employed in prior research for this purpose. Finally, we explain the research goals in the last part of that section. Following that, in section 3 we elaborate the proposed approach as well as the settings of the case study conducted to showcase that. Next, the results are presented (section 4) followed by a discussion of theoretical and empirical implications in section 5.

2.2. LITERATURE REVIEW

2.2.1. PRE- AND POST-APPROVAL ADR RESEARCH

It takes ten to fifteen years, on average, for a new drug to pass through the required clinical trials, get approved, and be introduced to the market (Iizuka, 2007). However, even after this long process, it is unlikely that all the risks associated with taking a drug have been identified. It is particularly due to limitations involved in lab experiments. They are often conducted over short timeframes and involve only a limited sample size. In addition, they are focused only on a particular group and usually exclude patients with complicated medical conditions (Karimi et al., 2015; Zeng et al., 2002). Moreover, these trials may not detect drug reactions with very low incidence rates (Stephens & Talbot, 1985). Due to these shortcomings, the side effects of a considerable number of drugs are often only revealed in the post-approval stage.

As a post-approval effort to rapidly detect and take appropriate action to ADRs, many countries and healthcare organizations have run Spontaneous Adverse Drug Reporting Systems (SAERSs); information systems designed to allow patients and professionals to submit their reports of suspected adverse drug events. Some of the examples of such systems are the yellow card system of Medicines and Healthcare products Regulatory Agency (MHRA) in the United Kingdom, and the FDA Adverse Event Reporting System (FAERS) in the United States (Karimi et al., 2015).

Although spontaneous reporting systems have been the main source to detect likely ADR cases for years, they still have several limitations such as over-reporting of highly common ADRs, missing and incomplete data, duplicated reporting and voluntary submission (Harpaz et al., 2013). Due to the voluntary submission of the reports, for instance, it is estimated that these systems in the US and UK reflect less than 10% of the adverse effect occurrences (Inman & Pearce, 1993; Yang et al., 2012). Such shortcomings led pharmacovigilance practitioners to look for resources that are more efficient for post-approval drug surveillance.

In recent years, Electronic Health Records (EHR) have been widely used in the healthcare industry to help practitioners in collection, storage, and tracking patients' information and their treatment progress. The vast amount of data collected by EHRs as well as their increasing availability of low-cost EHR platforms to the healthcare providers have made them interesting resources for pharmacovigilance researchers and presented opportunities to investigate and detect ADR signals[5] closer to real-time (Trifirò et al., 2009). Several data mining approaches have been proposed and applied by data scientists to EHR data in the past few years (Bao, Kuang, Peissig, Page, & Willett, 2017; Friedman, 2009; Polimeni et al., 2009; Santiso, Casillas, & Pérez, 2018; Trifirò et al., 2009). Despite utilizing EHR data for pharmacovigilance purposes have gained

[5] In pharmacovigilance, a signal is defined by the WHO as information on a possible causal relationship between an adverse event and a drug, which is unknown or incompletely documented (Trifirò et al., 2009).

much interest from European and Australian researchers, there is still a lack of sufficient research by the US academics and practitioners on the EHR data from the US healthcare market. Even though EHR data is generally more complete than data collected by spontaneous reporting systems, yet using this data for detection and prediction of ADR cases involves challenges such as complex data preprocessing requirements and various data documentation styles across different healthcare organizations (Harpaz et al., 2013).

Social media has also been considered as a key data source for monitoring drugs' post-marketing feedbacks in the recent few years. This is normally done either through healthcare online forums such as 'Ask a patient', 'Dailystrength', and 'PatientsLikeMe' (Karimi, Kim, & Cavedon, 2011; Leaman et al., 2010; X. Liu & Chen, 2013); or through social networks like Facebook and Twitter and by applying text-mining and sentiment analysis methods (Ginn et al., 2014; Nguyen et al., 2017; Prier, Smith, Giraud-Carrier, & Hanson, 2011).

2.2.2. TAXONOMY OF ADR STUDIES

In terms of research goals, pharmacovigilance studies can be classified into three categories, namely detection, prediction, and understanding studies (Davazdahemami & Delen, 2018).

Detection studies mainly aim at identifying existing associations (not necessarily causal) between drugs and potential adverse reactions, often by analyzing historical usage data obtained from various resources. Of course, additional clinical trials are needed to assess and verify the causality of associations detected by this type of ADR studies, however, it is still valuable to identify potential ADR that might be caused by a medication and focus the expensive and time consuming clinical trial activities on them.

Prediction studies are those that utilize information about already known drug-ADR associations to predict possible ADRs for the newly discovered as well as existing drugs. While detecting and predicting potential associations is a critical task, it is clear that such associations

do not hold all the time and in case of every patient. That is why, for instance, that a particular patient might experience a side effect of a given drug, while that drug may not have any adverse effect in another patient. Hence, it is crucial to investigate and understand the mechanism through which drugs develop side effects in the patients by identifying factors that either intensify or mitigate the strength of a drug-ADR association. This is, in fact, the goal of the understanding group of pharmacovigilance studies.

Many studies have been done in the past with the aim of detecting ADR signals for various drugs. In terms of methodology, some of them (Cai et al., 2017; van Puijenbroek et al., 2002) have used traditional statistical methods for this purpose, whereas many other studies (Friedman, 2009; Harpaz et al., 2013; Harpaz, Haerian, et al., 2010; X. Liu & Chen, 2013; Nikfarjam et al., 2015a; Reps et al., 2016; Trifirò et al., 2009) have employed data mining and analytics techniques to detect ADR signals. Association rule mining techniques have been shown in prior research to be highly efficient in extracting patterns from healthcare data (Borah & Nath, 2018; Harpaz, Chase, et al., 2010; Kuo, Lin, & Shih, 2007; W. H. Lee, Wang, & Chen, 2017; Nahar, Imam, Tickle, & Chen, 2013; Piri, Delen, Liu, & Paiva, 2018). Also in terms of data, various resources have been used in the past ADR detection studies including SAERSs (Cai et al., 2017; DuMouchel, 1999; Harpaz et al., 2013; van Puijenbroek et al., 2002), EHRs (Casillas, Pérez, Oronoz, Gojenola, & Santiso, 2016; Friedman, 2009; Haerian et al., 2012; Harpaz, Haerian, et al., 2010; Reps et al., 2016; Trifirò et al., 2009), and social media (Hoang et al., 2016; J. Liu et al., 2016; X. Liu & Chen, 2013; Nikfarjam et al., 2015a).

While most of the prior ADR detection studies are focused on identifying associations between drug-ADR pairs, many ADRs are actually the outcome of drug-drug interactions (DDIs) among two or more drugs that are prescribed and administered together in a short time window. Compared to the regular ADR detection studies, little research has been done on identifying such

DDIs, especially for studying DDIs involving more than two drugs and their potential role in developing ADRs.

2.2.3. RESEARCH GOALS

Reps et al.(2016) proposed a framework for refining ADR signals including sets of drugs obtained via longitudinal observational (EHR) data. In the present study, we extend their framework by adding extra assumptions and combining it with some standardized statistical metrics and apply that to the prescription records of a group of diabetic patients with the aim of identifying the general confounding roles of common diabetes medications in developing renal failure. In addition, we assess the stability of their roles across various prescription combinations to highlight their potential interactions with other relevant drugs, which leads them to act in an unexpected way with respect to developing renal failure.

The proposed framework in the current study differs from that of Reps et al. in two specific aspects; first, unlike their approach which relies on the "lift" measure to identify frequent itemsets that are more frequent among case patients than among control patients, we use a statistical metric for comparing case and control patients and rely on statistical significance of difference for judging about the confounding effect. Second, in our proposed approach the focus is on identifying the confounding role of single drugs by taking into account their potential interaction with other drugs, as opposed to Reps et al. that mainly investigate the confounding effect of the whole itemset.

In fact, our study is an *event-based* type of data mining analysis as defined by Trifirò et al.(2009), in which the focus is on one or a set of specific events (i.e., renal failure) for their association with possible drugs that may cause them. There is an event-based stream of research focused on investigating the drugs associated with kidney diseases in general and acute renal failure in particular (Cavalieri et al., 2018; Coca & Perazella, 2002; Heerspink et al., 2017; J.

Huang, 2018; Izzedine, Launay-Vacher, & Deray, 2005; Kimura et al., 2017; Markowitz & Perazella, 2005; Naughton, 2008; Perazella, 2003; Singh et al., 2003). Loh and colleagues (2009) mention top ten categories of medications that cause kidney damage involving antibiotics, analgesics, COX-2 inhibitors, proton pump inhibitors, antiviral drugs, high blood pressure drugs, rheumatoid arthritis drugs, lithium, anticonvulsants, and chemotherapy drugs. The same set of drugs is mentioned, more or less, in the other related studies as well. Moreover, it is widely discussed in the medical literature that diabetes is the leading cause of renal failure so that diabetic nephropathy (a.k.a. diabetic kidney disease) is today well known as a progressive kidney disease due to longstanding diabetes type II (Afkarian et al., 2016; Loh & Cohen, 2009). The mechanism through which diabetes leads to the development of diabetic nephropathy is studied by several researchers (Fujita et al., 2014; Lehmann & Schleicher, 2000; Sun, Su, Li, & Wang, 2013). Most of these studies highlight the role of high blood sugar levels as well as high blood pressure in damaging capillaries in the kidneys glomeruli. Given that, the general expectation from common diabetes drugs should be to attenuate damages to kidney through balancing the blood sugar, thereby decreasing the likelihood of developing acute renal failure. Prior research has investigated the confounding role of some of the frequently prescribed diabetes medications and reported inconsistent effects. For instance, while Fatourechi et al. (2009) mention a positive confounding effect for insulin therapy, Thomas et al. (2007) suggest an association between insulin therapy and reduced incidents of renal failure. In addition, infrequently prescribed diabetes medications rarely were studied for their potential confounding roles in developing renal failure.

We apply an extended version of the Reps et al. framework to the longitudinal prescription records of a large group of diabetic patients to investigate the confounding (either attenuating or intensifying) role of common diabetic medications (including those infrequently prescribed ones) in those patients. Furthermore, we investigate whether those confounding roles

are stable across various prescription combinations or they might change due to potential drug-drug interactions, thereby trying to explain inconsistent findings reported in prior research. Therefore, our research question is "*How are the confounding roles of common diabetes medications in developing acute renal failure in diabetic patients? Are these roles stable for each drug across various prescription combinations?"*

2.3. MATERIALS AND METHODS

2.3.1. MATERIALS

In order to address the research question, we obtained data from a longitudinal observational electronic health records database, namely the Cerner HealthFacts data warehouse [6] (http://www.cerner.com). Cerner HealthFacts is the most comprehensive relational database in the U.S. and contains complete medical records of more than 63 million unique patients across the country. The database contains time-stamped entries of patients' visits, physicians' diagnoses, and prescribed drugs (among other patient-event specific characteristics). Prescription and diagnosis records of adult patients (18 or older) diagnosed with diabetes mellitus (ICD9- 250) for the first time during the 4-year period of study (i.e., 2012-2015) were extracted for analysis. The initial data involved 377,910 unique patients. Since the focus of our study was on diabetic drugs as well as kidney-damaging (KD) drugs, we then filtered the prescription records to keep only these types of drugs for analyses.

2.3.2. METHOD

A case-control design was employed to conduct the analyses. In this design, the case group were those diabetic patients who developed acute renal failure (ICD9- 580) during the study period,

[6] Cerner is not a publicly available data source, however, the authors had access to that via their institution, to which the data warehouse is donated for the research purposes.

and the control group involved those diabetic patients who were not diagnosed with renal failure by the end of the study period.

For each patient identified as a *case* subject, we considered two *index dates;* 1) the date he or she was diagnosed with diabetes mellitus for the first time, and 2) the date the patient was diagnosed with renal failure for the first time. Two subjects were matched as *the control* to each case-patient by matching on their age, race, gender, comorbidities and the first index date (i.e., the date he or she was first diagnosed with diabetes mellitus). Moreover, the second index date for each *control* is the same as its matching *case*'s second index date. Matching two controls for each case-patient makes the sample more representative of the population (Reps et al., 2016). An innovative method was used to match the controls to each case patient, in which we coded each patient profile using an ten-character string including two characters for age, one character for gender (male=1, female=2), one character for race (Caucasian=1, African-American=2, Hispanic=3,...), two character for comorbidities, and four characters for a numeric transformation of the first index date (indicating the number of days passed since January 1, 2000). We then used a simple SQL query using the coded patients profile to find all the matches from the potential control patients to each subject in the case group and randomly selected two of them for each case subject. Since there were less than two matches for some case subject profiles, we re-coded the profiles for that particular patients and replaced the four-digit index date with a three-digit one representing the number of weeks passed since January 1, 2000 and ran the queries again. There were also some cases for which we used the month of the index date to find a matching. However, fortunately, there was no problem with regard to finding matches in terms of any of the other factors thanking the large initial dataset employed.

Table 2.1. Profile of D1 and D2 databases

Gender	Race	Age	Comorbidities
Male 55.76%	Caucasian 47.21%	Mean 40.77	Mean 5.38
Female 44.24%	African-American 41.22%	StDev 7.51	StDev 2.15
	Native American 2.66%		
	Hispanic 2.61%		
	Asian 1.04%		
	Other 5.26%		

In the next step, two databases were created. D1 which involved the prescription records pertaining to the last 10 visits prior to the second index date of patients in the case-group (including 1,294 patients); and D2 containing the same information for their corresponding control patients (2,588 patients). In order to construct the databases, if a drug was prescribed two or more consecutive times, we only kept the earliest prescription. Also, all the medications prescribed in a single visit were given the same sequence label. Given these assumptions, we came up with 18,562 and 22,388 prescription records for the case and control patients, respectively. Table 2.1 demonstrates the profile of the two databases in terms of the factors the patients were matched on.

To investigate the potential confounding role of diabetic drugs we then applied frequent itemset mining, using the association rule mining pre-defined procedure in SAS Enterprise Miner, to both D1 and D2 to identify frequent sets of drugs along with their support (i.e. the proportion of transactions in the database that contain that set of drugs). Frequent itemset mining is a branch of frequent pattern mining in which the focus is on identifying sets of items within a transactional database that appear sufficiently often in the whole database. Therefore, a *support threshold* should be determined by the user to specify for the algorithm as to "how often" do we consider "sufficiently often". One of the popular algorithms for extracting frequent itemsets from a

transactional database, used as the main algorithm in the SAS Enterprise Miner platform, is the Apriori algorithm proposed by Agarwal and Srikant (1994). The algorithm begins by identifying frequent single items in a transactional database, and then in each subsequent iteration generates candidate itemsets of size *n* from the itemsets of size *n-1* and then prunes the infrequent candidates with regard to the given support threshold. Even though recently multiple innovative algorithms and metrics have been proposed for effective association rule mining under special circumstances or with different approaches such as incomplete evidence (Galárraga, Teflioudi, Hose, & Suchanek, 2013), in the presence of constraints (Baralis, Cagliero, Cerquitelli, & Garza, 2012), identifying rare rules (Piri et al., 2018), using utility-based (as opposed to frequency-based) mining (D. Lee, Park, & Moon, 2013), and taking into account the weight of items in the rule mining (Vo, Coenen, & Le, 2013), yet Apriori is known as an effective generic association rule mining algorithm.

In order to find more relevant itemsets, we limited the maximum size of itemsets to 5 and the minimum support threshold to 0.5% in both data sets. The reason we limited the maximum size of itemsets to 5 is that the average number of transactions (i.e., distinct medications prescribed) for the patients in the control group was 5.10 with a median of 5; therefore considering itemsets including less than 5 drugs would result in excluding half of the control patients (probably the healthier half) from the analyses and would make the results biased. In addition, given the total number of patients in the case and control groups, considering a support threshold less than 0.5% technically was pointless, since it would result in very small frequencies and the corresponding itemset most probably would not suggest any statistically significant results. It should be noted that we did not take into account the sequence in which the drugs were administered (i.e., a non-sequential itemset mining analysis was done). Hence, for instance two itemsets like *{a,b,c}* and *{c,a,b}* were considered equivalent in calculating supports.

At the end of this process, we come up with a list of itemsets for each group of patients along with their support in their corresponding databases. At this stage, the itemsets containing both types of drugs of interest (i.e. diabetic and KD) were identified to focus on. We call them *Combined Sets (CS)* from now on. For each itemset in the CS, we then find matches from itemsets including all its KD drugs, but no diabetic drugs. We call this second group *Pure Sets (PS)*. Hence, for each itemset in the CS, there are one or more matches in the PS. For example, if *M={acetaminophen, vancomycin, insulin}* is a frequent itemset identified as CS, its matching set in the PS would be *M'={acetaminophen, vancomycin}* which only involves drugs from KD category (note that *M'*⊆ *M*). The CS and PS itemsets were identified in both case and control patients.

In the medical literature, Relative Risk (RR) is a measure used to indicate the risk of developing disease given exposure to its causes (Altman, 1990). Suppose that we expose the case group to a particular factor while keeping the control group unexposed. If we record the number of bad and good outcomes in each group (let's call them *a* and *b* for the case and *c* and *d* for the control group respectively), the RR then would be:

$$RR = \frac{a/(a+b)}{c/(c+d)} \quad \text{(Eq.1)}$$

With the standard error of the log RR being:

$$SE\{\ln(RR)\} = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}} \quad \text{(Eq.2)}$$

In this study, since the outcome of interest (i.e. renal failure) was an ADR, and the goal was to identify the potential confounding effect of diabetic drugs, then each diabetic drug was considered as a potential cause for the outcome. Let us consider the case group in our study first. If we call the number of patients having a particular CS itemset (I_{cs}) in their prescription records

"a", and the number of case patients having the PS itemset (I_{ps}) corresponding to I_{cs} (note that $Ips \subseteq Ics$), *"b"*, then $\frac{a}{a+b}$ would represent the ratio of case patients who were exposed to the diabetic drug involved in I_{cs} (i.e., I_{cs}-I_{ps}) to the case patients who have taken both diabetic and kidney damaging drugs involved in I_{cs} . Similarly, $\frac{c}{c+d}$ can be interpreted as the same ratio in the control group of patients. Therefore, we argue that if this ratio for the case patients is significantly greater than for the control patients, it suggests that controlling for a particular PS itemset (I_{ps}), prescribing a particular diabetic drug (i.e. $I_{cs} - I_{ps}$) generally increases the risk of renal failure (i.e. positive confounder). Conversely, if this ratio for the case patients is significantly smaller than that for the control patients, it suggests the corresponding diabetic drug is a negative confounder in developing renal failure. Finally, if the ratios are not different across two groups it suggests that the corresponding diabetic drug has no confounding role in developing renal failure. To make the risk ratio equation more meaningful for our particular purpose, we call it the confounding coefficient (CC) from this point on and we attribute it to the specific diabetic medication that exists in I_{cs} but not in I_{ps}. In short, if:

$$CC_{drugX} = \frac{a/(a+b)}{c/(c+d)} \quad \text{(Eq.3)}$$

for a specific diabetic drug "X" is significantly (i.e., p-value<0.05) greater than 1, it suggests that the diabetic drug X is a positive confounder of the ADR (i.e., increases the risk of renal failure), since it is taken by the case patients significantly more than the control patients. Similarly, if CC_{drugX} is significantly (i.e., p-value<0.05) less than 1, it suggests that drug X is a negative confounder of the ADR (i.e., decreases the risk of renal failure) since it is taken by the case patients significantly less than by the control patients. In addition, a non-significant CC (i.e., p-value>0.05) suggests no confounding role for the corresponding drug. Moreover, the

larger/smaller significant CC values suggest a potentially stronger positive/negative confounding effect.

Table 2.2. A numerical example of the proposed method

I_{cs}	Case	Ctrl	I_{ps}	Case	Ctrl	CC_{drg050}	p-value
	a	c		b	d		
{drg050,drg230, drg344}	39	57	{drg230,drg344}	0	12	1.20	0.0005

Table 2.2 indicates an example. Suppose that drg050 is a diabetic drug and drg230 and drg344 are two kidney-damaging drugs. The *combined* itemset including these drugs (Ics) is identified as a frequent itemset. Excluding drg050, the diabetic drug, from this itemset we get an itemset purely including kidney-damaging drugs (i.e. Ips). Suppose that 39 case patients had all the three drugs in their prescription records (i.e., *a=39*). Also that the number of case patients having only the two kidney-damaging drugs but not the diabetic drug in their records was 0 (i.e., *b=0*). Similarly, suppose that these numbers in the control group are *c=57* and *d=12*. The Confounding Coefficient (CC) corresponding to the drug drg050 then would be:

$$CC_{drg050} = \frac{39/(39+0)}{57/(57+12)} = 1.20 \qquad \text{(Eq.4)}$$

with a p-value=0.0005. It suggests that the diabetic drug, drg050, is a positive confounder that, if prescribed along with drg230 and drg344, can increase to the risk of developing renal failure mainly caused by the other two drugs.

Similar analyses were performed for all the identified frequent combined sets of drugs and their corresponding pure sets in both case and control groups of patients and the confounding coefficients were calculated multiple times for each common diabetic drug. The results were then integrated, as reported in section 4. Figure 2.1. illustrates the method and procedures in a workflow-type graphical depiction.

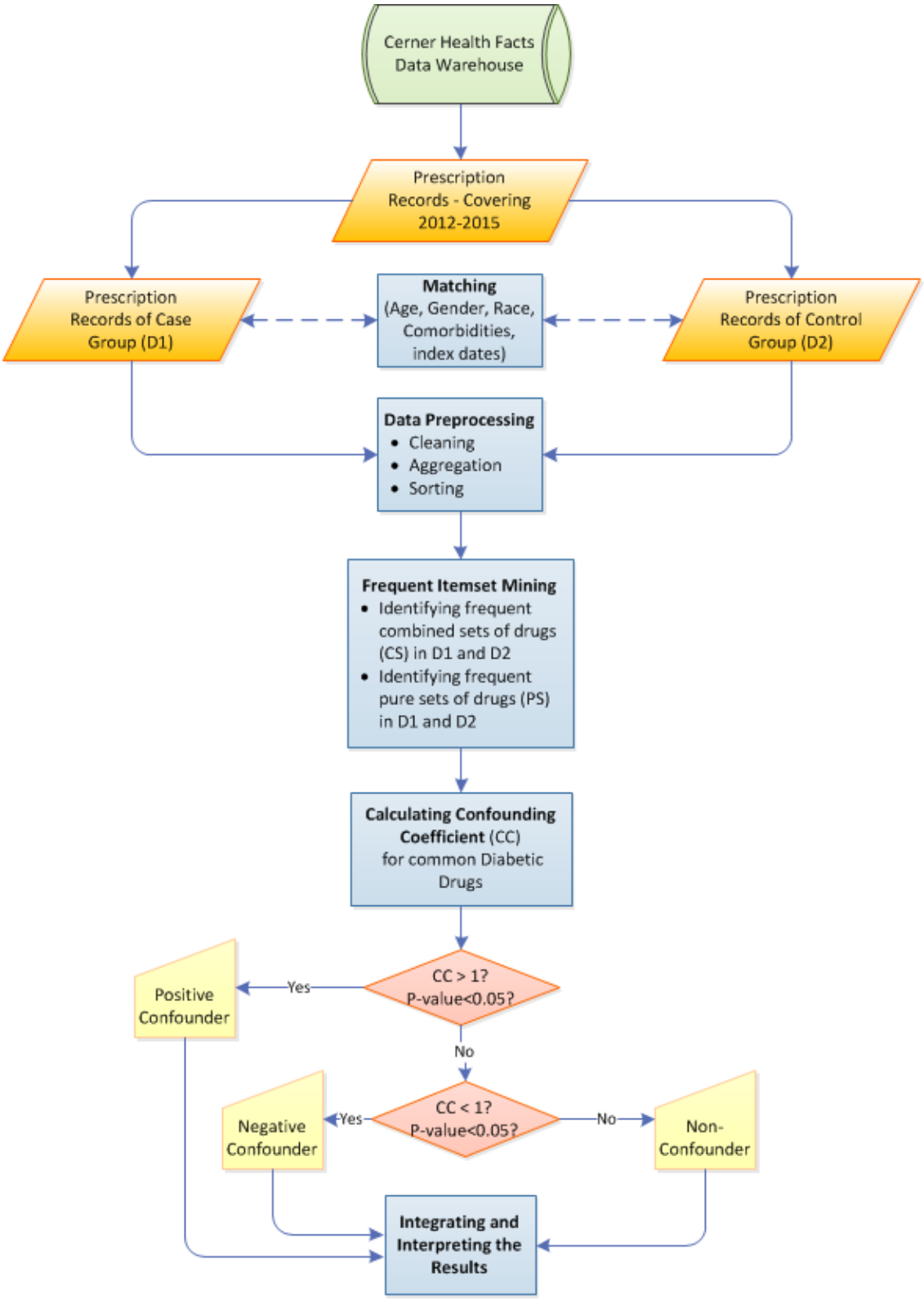

Figure 2.1. A graphical depiction of the data preprocessing and method development procedures

2.4. RESULTS

To conduct this research, we particularly focused on 23 common diabetic as well as 43 kidney-damaging medications that were prescribed at least once for a patient in our initial data set. However, after identifying the case and control groups the numbers decreased to 16 and 31 respectively, as some drugs were not prescribed for the patients in these groups at all.

Applying frequent itemset mining to the prescription records of the case and control groups, more than 5,000 frequent itemsets (not necessarily unique) were identified using SAS Enterprise Miner. As noted before, we limited our rule extraction procedure to find itemsets including up to 5 drugs with a minimum support of 0.5% across the whole data set. We then filtered the frequent itemsets to include only the itemsets involving at least one drug of each type. This resulted in 246 unique frequent itemsets, which were, in fact, our Combined Sets (CS) of drugs discussed in the method section. Then for each of the CS itemsets, a subset containing only their corresponding kidney-damaging drugs was created; in our method terminology, we called these subsets the Pure Sets (PS). In the next step for each CS itemset and its corresponding PS, using the frequency of their incidence within the case and control patients' prescription records, we calculated the CC associated with the diabetic drug involved in each CS itemset.

Table 2.3 demonstrates a sample of 10 frequent itemsets along with their corresponding CC calculations.

Table 2.3. A sample of frequent itemsets and CC calculations

I_{cs}	Case	Ctrl	I_{ps}	Case	Ctrl	CC	p-value
	a	c		b	d		
{insulin & acetaminophen & pantoprazole & aspirin}	173	107	{acetaminophen & pantoprazole & aspirin}	3	6	1.03	0.126
{insulin & acetaminophen & ketorolac & aspirin}	76	69	{acetaminophen & ketorolac & aspirin}	1	11	1.14	0.004
{pantoprazole & insulin & ketorolac & aspirin}	46	36	{pantoprazole & ketorolac & aspirin}	0	6	1.16	0.141
{ketorolac & insulin & ibuprofen & acetaminophen}	39	57	{ketorolac & ibuprofen & acetaminophen}	0	12	1.20	0.0005
{pantoprazole & metformin & acetaminophen}	38	64	{pantoprazole & acetaminophen}	330	208	0.44	0.0001
{ketorolac & aspirin & metformin & acetaminophen}	19	33	{ketorolac & aspirin & acetaminophen}	58	47	0.60	0.032
{insulin & ciprofloxacin & esomeprazole}	24	16	{ciprofloxacin & esomeprazole}	1	0	0.96	0.317
{sitagliptin & acetaminophen & esomeprazole}	9	10	{acetaminophen & esomeprazole}	35	106	2.37	0.042
{glyburide & aspirin & acetaminophen}	10	0	{acetaminophen & aspirin}	352	326	18.91	0.042

Table 2.4 shows a summary of results for the diabetes medications emerged in the frequent itemsets. As shown, from 16 common diabetic drugs involved in the case and control patients' records, only nine emerged in frequent itemsets. For each of them, the table indicates the number of times they were present in distinct frequent itemsets as well as the number of times they were recognized as a significant (either positive or negative) confounder (using a 0.05 significance level). For instance, for insulin (and its variations such as insulin aspart, insulin glargine, etc.), the results show that due to high frequency of its prescription for diabetic patients, it was present in 94 out of 246 identified frequent itemsets; from which, our analysis showed that in 53 itemsets insulin plays a significant confounding role. That is its corresponding confounding

coefficient (CC) was significantly different from 1. Among those significant cases, it was revealed that in 48 cases (90.6%) insulin plays a positive confounding role (i.e. CC>1). This suggests that controlling for the KD drugs present in each itemset, diabetic patients who experienced renal failure during the study period (i.e., case group) had been prescribed insulin significantly more frequently than those in the control group who did not experience the adverse outcome. Overall, it suggests that generally, insulin plays a positive confounding role in the development of renal failure in diabetic patients. As shown in Table 2.4, from the 9 drugs analyzed, only Metformin, Linagliptin, and Pioglitazone showed a generally negative confounding behavior and other diabetic medications exhibited a positive role in confounding the issue.

Table 2.4. Common Diabetic Medications and their potential confounding roles

Drug (generic name)	***Total # of itemsets emerged in***	***# identified as significant confounder***	***# identified as positive confounder (%)***	***# identified as negative confounder (%)***	***Average CC***	***Max CC***	***Min CC***	***Conclusion***
Insulin	94	53	48 (90.6%)	5 (9.4%)	1.10	1.23	0.91	Pos. conf.
Metformin	52	38	0 (0%)	38 (100%)	0.52	0.71	0.26	Neg. conf.
Glipizide	21	14	14 (100%)	0 (0%)	2.12	2.73	1.81	Pos. conf.
Sitagliptin	18	11	9 (81.8%)	2 (18.2%)	4.38	6.15	0.93	Pos. conf.
Glimepiride	18	8	8 (100%)	0 (0%)	3.61	5.61	2.10	Pos. conf.
Glyburide	13	7	7 (100%)	0 (0%)	14.36	18.91	12.11	Pos. conf.
Acarbose	11	7	6 (85.7%)	1 (14.3%)	3.09	3.76	0.97	Pos. conf.
Linagliptin	11	6*	1 (16.6%)	5 (83.3%)	0.71	1.09	0.57	Neg. conf.
Pioglitazone	8	4*	0 (0%)	4 (100%)	0.82	0.86	0.76	Neg. conf.

** For these cases, significance was assessed at 0.1 level due to the scarcity of them among records.*

In order to make sure that the 0.5% support threshold used to identify the frequent itemsets was a proper choice, we looked into the number of frequent itemsets identified (for

insulin and metformin) as well as those in which the focal diabetes drug turned out as a significant confounder, considering four different thresholds ranging from 0.2% through 1.5% (see Figure 2.2). As shown in Figure 2.2 while in case of each drug both numbers increase by decreasing the support threshold, the gap between the two lines is considerable going from a 0.5% down to 0.2% threshold. This actually happens since itemsets with a support less than 0.5% are such infrequent that using their frequencies in calculating the CC index for the corresponding diabetes drug does not result in a CC significantly different from 1. Overall, this confirms that 0.5% seems to be a reasonable choice for support threshold, because thresholds lower than that technically does not provide us with considerably more information with regard to the confounding roles.

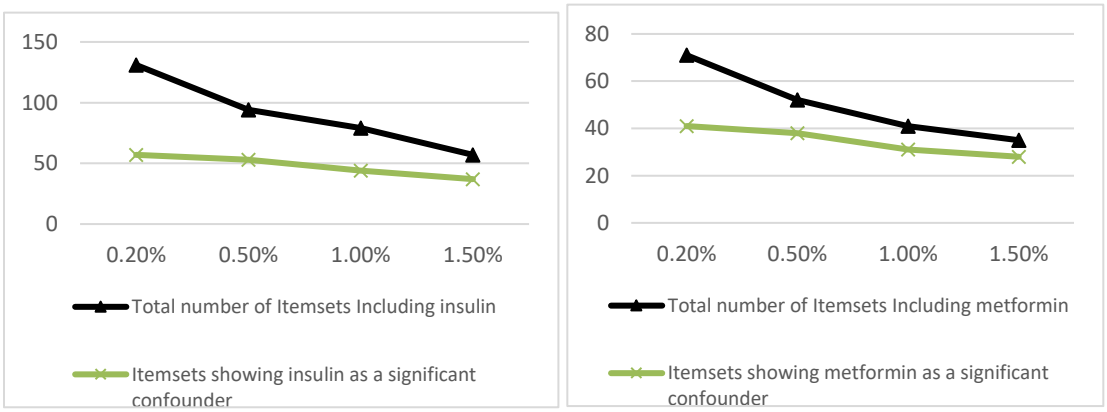

Figure 2.2. Evaluating different support thresholds for identifying frequent itemsets

Even though our analyses identified the dominant type of confounding role for each diabetic drug mentioned in Table 2.4, the results in this table also suggest that these confounding roles are not stable since for most of these diabetic drugs there exist frequent itemsets indicating significant confounding coefficients in the opposite direction as well. For instance, although insulin was said to be a generally positive confounder, our results indicate there were five instances (i.e., 9.4% of itemsets) in which a significant negative confounding role in developing

acute renal failure has been identified. In addition, this drug has not shown any significant confounding role at all in around 44% of the frequent itemsets in which it emerged. This suggests a very important point; that while insulin generally contributes to the development of acute renal failure, its confounding role varies depending on its potential interaction with other drugs administered to the patient. Table 2.5 indicates the five itemsets for which insulin turned out as a significant negative confounder. Our further investigation revealed that ciprofloxacin and sulfamethoxazole-trimethoprim, the two kidney-damaging drugs that are present along with insulin in these itemsets are not present in any of the itemsets that show a significant positive confounding role for insulin. This may suggest DDIs between these drugs leading to lower chances of developing renal failure.

Table 0.5. Itemsets indicating a negative confounding role for insulin

Itemset	$CC_{insulin}$	***p-value***
{sulfamethoxazole-trimethoprim & insulin & aspirin & acetaminophen}	0.93	0.031
{sulfamethoxazole-trimethoprim & insulin & aspirin & acetaminophen}	0.93	0.015
{vancomycin & acetaminophen & sulfamethoxazole-trimethoprim & insulin}	0.91	0.019
{aspirin & insulin & ciprofloxacin}	0.96	0.038
{insulin & ciprofloxacin & acetaminophen & esomeprazole}	0.95	0.025

Additionally, looking into the magnitude of average CC values for the diabetic drugs in Table 2.5, it suggests that Glyburide has the strongest positive effect (avg CC= 14.36) on increasing the likelihood of renal failure in diabetic patients, followed by Sitagliptin (avg CC=4.38). On the other hand, the average CC values for Metformin (0.52) and Linagliptin (0.71) implies that these two common diabetic drugs have the most negative confounding effect on the development of renal failure in these patients.

2.5. DISCUSSION AND CONCLUSIONS

We extended an existing data mining framework in a case-control setting to investigate the potential confounding role of drugs with regard to a given adverse event. The extended framework was applied to the prescription records of a group of diabetic patients to investigate the potential confounding role of common diabetes medications (as well as its stability across various prescription combinations) in the development of acute renal failure (as the adverse event of interest in this study).

The results indicate statistically significant differences between the prescription records of the case and control groups with regard to several common diabetic medications. Particularly, for the two most common medications for diabetes Type II patients (i.e. insulin and metformin), the results suggest potential generally positive and negative confounding roles for them, respectively. That is, controlling for the drugs already known to be associated with renal failure, the proportion of case patients prescribed with insulin was significantly higher than control patients. Similarly, the proportion of case patients prescribed with metformin was significantly lower than those in the control group.

While insulin therapy is today a popular treatment among diabetic Type II patients due to its effectiveness in quickly reducing blood glucose level and positive effects on appetite and letting them have a more regular life, prior research have shown that uncontrolled insulin injection leads to resistance of the body to insulin and ultimately affects kidney due to hypoglycemia resulted (Fatourechi et al., 2009; Iglesias & Diez, 2008). Metformin then can function as a complementary agent to increase the sensitivity of the body to insulin and make a balance in such situations. Also, prior research (Berhanu, Perez, & Yu, 2007; von Websky, Reichetzeder, & Hocher, 2013; Yamanouchi, 2010) suggests that adding linagliptin or pioglitazone to insulin therapy can prevent hypoglycemia due to the accumulation of insulin. Our

results confirm such roles for metformin, linagliptin, and pioglitazone as they came out as negative confounding factors with regard to kidney failure. This actually provides support for the validity of the results obtained by our proposed approach.

While the confounding roles of insulin and metformin have been studied in the prior research, the findings in those studies are somewhat inconsistent. For instance, Thomas et al. (2007) suggest that intensive insulin therapy in critically ill adult patients is associated with reduced incidences of acute renal failure. Also, Hsu et al. (2017) have reported that metformin may have an adverse effect in the renal function in patients with diabetes Type II. Our results provide an explanation for such inconsistencies by highlighting the role of potential drug-drug interactions that may lead a drug to act in an unexpected way with regard to an adverse event. Moreover, this study provides insights with regard to the general confounding effects of some diabetes medications that are under-studied in the literature, due to their lower prescription frequency by the practitioners.

Of course, the present study is essentially a signal detection study and does not imply any causal relationship between the drugs and the ADR under study. This can be considered as the first step in a regular drug safety research which should be followed by assessments from a biological and clinical perspective and in-depth investigation to confirm or reject signals using expert opinions and randomized controlled trials (Shetty & Dalal, 2011). Future research may also validate and extend our approach by employing it for studying confounders of other common ADRs.

Another contribution of this study is expanding a method originally proposed by Reps et al. (2016) for identification of drug-drug interactions (DDIs) involving more than two drugs. While there has been a huge amount of research on detecting DDIs involving pairs of drugs, in practice a typical patient might be prescribed with several drugs in each visit. Hence, taking into

account larger sets of drugs can help to reveal more reliable signals than when only two drugs are analyzed at a time. Again, further research should be conducted in order to assess and confirm the detected DDIs from the biological and clinical points of view.

In short, our results indicate that while a general, either positive or negative, confounding role can be attributed to each of the common diabetic medications that holds in most of the prescription combinations, however, these confounding roles are not stable across various prescription combinations and taking into account drug-drug interactions sometimes a significant positive confounder may act as a negative one, or vice versa. This actually explains the inconsistent confounding roles reported in the literature and highlights the importance of considering DDIs in determining the outcome of a drug prescription.

Of course, this research involves some limitations. Even though we controlled for the total number of comorbidities as a measure of general wellness, we did not control for the specific comorbidities between case and control groups as it would have significantly reduced the sample size and the power of analysis. In fact, we assumed that controlling for the age, gender, total number of comorbidities and the time of being diagnosed with diabetes, we can expect the same level of health between case and control patients regardless of their specific diseases. In addition, assuming that frequency of prescription of a drug has a strong correlation with the number of doses taken and given that this frequency leads to the emergence of the drug in the frequent itemsets, we also did not control for the doses of medications. These issues can be addressed in a randomized controlled trial study aimed at confirming the validity of signals detected here. So future research may expand the proposed approach by employing a larger data set (i.e., a wider study period) which allows for controlling the effect of specific diseases as well. This can be done by including diseases as items, just like the drugs, in the itemset mining analyses.

CHAPTER III

ESSAY II: EXAMINING THE EFFECT OF PRESCRIPTION SEQUENCE ON DEVELOPING ADVERSE DRUG REACTIONS: THE CASE OF RENAL FAILURE IN DIABETIC PATIENTS

ABSTRACT

Objectives: While the effect of medications in development of Adverse Drug Reactions (ADRs) have been widely studied in the past, the literature lacks sufficient coverage in investigating whether the sequence in which [ADR-prone] drugs are prescribed (and administered) can increase the chances of ADR development. The present study investigates this potential effect by applying emergent sequential pattern mining techniques to electronic health records.

Materials and Methods: Using longitudinal medication and diagnosis records from more than 377,000 diabetic patients, in this study, we assessed the possible effect of prescription sequences in developing acute renal failure as a prevalent ADR among this group of patients. Relying on emergent sequential pattern mining, two statistical case-control approaches were designed and employed for this purpose.

Results: The results taken from the two employed approaches (i.e. 76.7% total agreement and 68.4% agreement on the existence of some significant effect) provide evidence for the potential effect of

prescription sequence on ADRs development evidenced by the discovery that certain sequential patterns occurred more frequently in one group of patients than the other.

Conclusion: Given the significant effects shown by our data analyses, we believe that design and implementation of automated clinical decision support systems to constantly monitor patients' medication transactions (and the sequence in which they are administered) and make appropriate alerts to prevent certain possible ADRs, may decrease ADR occurrences and save lives and money.

Keywords: *Adverse Drug Events; Adverse Drug Reactions; Prescriptions Sequence; Emergent Pattern Mining; Electronic Health Records.*

3.1. INTRODUCTION

Today every drug produced and marketed by pharmaceutical companies has a list of likely side effects printed on its label to warn patients about possible harms they may undergo by taking it. Such known side effects are usually the result of several years of research and clinical trials conducted on the drug by the manufacturer after discovery and before introducing it to the market.

There are, however, some limitations involved in these clinical trials. They are often conducted over short timeframes and involve only a limited sample size. Therefore, the sample may not fully represent the population of consumers and may exclude patients who receive other medications. In addition, they are focused only on a particular group and usually exclude patients with complicated medical conditions (Karimi et al., 2015; Zeng et al., 2002). Moreover, these trials may not detect drug reactions with very low incident rates (Stephens & Talbot, 1985). Due to these shortcomings, the side effects of a considerable number of drugs are often only revealed in the post-marketing stage.

In pharmacovigilance[7] terminology, Adverse Drug Event (ADE) is a general term that refers to any injury caused by a medication. This injury can be an unintended effect of the recommended (i.e. prescribed or labeled) usage of a drug, the off-label use of a drug, or a medication error (Karimi et al., 2015). Adverse Drug Reactions (ADRs) are a subset of ADEs referring to an unexpected harm caused by the normal use of medication at the normal dosage (Karimi et al., 2015). Therefore, ADRs do not have to be related to the non-prescribed or off-label usage of a drug or medication errors; instead, they are generally the result of unexpected drug-event or drug-drug interactions. ADRs are reported by Australian Commission on Safety and Quality in Healthcare (ACSQHC, 2012) to cause about 400,000 visits to general practitioners and about 190,000 visits to hospitals in each year in Australia with a population of

[7] Pharmacovigilance (a.k.a. drug safety surveillance) is a field of science that tries to detect, assess, understand, and prevent harms and injuries caused by medications in all stages of drugs' lifetime (i.e. discovery, clinical trials, pre-marketing, and post-marketing). (World Health Organization)

only 23 million people. Also, the cost incurred by each ADR case in community hospitals in the United States is estimated at $3,000 (Classen et al., 1997; Hug et al., 2012).

Such considerable costs to patients, insurance agencies, and the healthcare industry have caused researchers to seek effective ways for detection, prediction, and prevention of ADRs during the past years. Multiple approaches are employed for this purpose and information systems (IS) have been playing a key role in almost all of them so that the main three ones, namely Spontaneous Adverse Drug Reporting Systems (SAERS), analysis of Electronic Health Records (EHR), and analyzing Social Media feedbacks are all heavily relied on information systems.

An important potential factor in the occurrence of adverse drug reactions is the *sequence* by which the drugs are administered. Although this potential factor is mentioned in prior research to be more investigated (Egger, Drewe, & Schlienger, 2003), to the best of our knowledge, no prior research has empirically investigated the effect of this factor on the likelihood of ADR development. Hence, the main aim of our study is to investigate the potential effect of the prescription sequence on the development of adverse drug reactions. One of the five most common ADRs identified in the literature is acute renal failure[8] (Trifirò et al., 2009). The literature has identified several drugs with renal failure as one of their side effects (Härmark et al., 2007; Perazella, 2003; Perneger, Whelton, & Klag, 1994; Singh et al., 2003). Due to its importance and high potential risk, in this study we specifically focus on this particular ADR and investigate the possible effect of the prescription sequence of its corresponding causes on the likelihood of its development.

To this end, we develop two independent approaches both using a case-control study design. First, a sequential emergent pattern mining approach is developed to compare the sequential prescription patterns between the case (those patients who developed a specific ADR) and control patients (those patients, matched to the case patients on various factors, who did not develop the ADR) with the aim of

[8] The other top common ADRs include bullous eruptions, anaphylactic shock, acute myocardial infarction, and rhabdomyolysis.

identifying whether different sequential patterns of the same frequent set of drugs have different effects on the likelihood of developing the ADR. Second, we compared the rank order of various sequential patterns of each frequent set of drugs (identified using a frequent itemset mining algorithm) by the means of the Spearman rank order correlation to specify whether for each non-sequential frequent set of drugs, the frequency of sequential patterns is significantly different among the case and control patients.

The remainder of this paper is organized as follows. In section 2, through a review of the literature, we discuss the role of IS in ADE detection, prediction, and understanding research as well as various approaches employed in prior research for this purpose. Next, in section 3, the research question will be explained in more details. Following that, in section 4, we introduce the data set and the data preparation processes used to investigate the research question. Also, the method of analysis is explained in the same section and it is followed by the results (section 5) and discussion (section 6) of theoretical and empirical implications.

3.2. LITERATURE REVIEW

It takes ten to fifteen years, on average, for a new drug to pass through the required clinical trials, get approved, and be introduced to the market (Iizuka, 2007). However, even after this long process it is unlikely that all the risks associated with taking a drug have been identified. It is particularly due to limitations involved in lab experiments. They are often short time experiments and involve just a limited sample size (Zeng et al., 2002); the samples do not fully represent the target population of the drug as may be focused on particular groups and exclude others (Karimi et al., 2015); and the reactions with very low incidence rates are hard to detect through clinical trials (Stephens & Talbot, 1985).

These shortcomings have caused a considerable number of potential drug-drug and drug-event interactions to remain undetected and it calls for additional investigations in the post-marketing stage of drugs' lifecycle.

Adverse Drug Events is the general term in the Drug Safety Surveillance domain that refers to any injuries caused by a medication. An ADE can be described along several dimensions like the severity of its consequences, the stage of the medical use process in which it occurred, and the type of cause (e.g. medication error, wrong dosage, reaction with other drugs, etc.) (Riccioli, Leroy, & Pelayo, 2009). Since ADE by definition includes every kind of injuries (i.e. either due to a normal or abnormal usage of medications), a more specific definition is proposed in the literature for injuries specifically caused by normal use of medications at the normal, prescribed dosage. These type of unexpected causes are referred to as Adverse Drug Reactions (ADR) in pharmacovigilance terminology (Karimi et al., 2015; Nikfarjam et al., 2015a).

Due to considerable costs and damages incurred by ADRs to the patients, insurance agencies, and healthcare providers, there has been a stream of research on detection, prediction, and understanding of this phenomenon in multiple disciplines including medicine, economics, and IS. The researchers in this area have employed various approaches, but what is shared among them is their heavy reliance on information systems for collection, extraction, and analysis of data required for detection and prediction of ADRs.

As an effort to rapidly detect and take appropriate action to ADRs, many countries and organizations have run Spontaneous Adverse Drug Reporting Systems (SAERSs); information systems designed to allow patients and professionals to submit their reports of suspected adverse drug events. Some of the examples of such systems are the World Health Organization's (WHO) Individual Case Safety Reports (ICSR) database, the Therapeutic Goods Administration's (TGA) Adverse Drug Reaction System (ADRS) in Australia, the yellow card system of Medicines and Healthcare products Regulatory Agency (MHRA) in the United Kingdom, and the FDA Adverse Event Reporting System (FAERS) in the United States (Karimi et al., 2015).

Although spontaneous reporting systems have been the main source to detect likely ADR cases for years, they still have several limitations such as over-reporting of highly common ADRs, missing and incomplete data, duplicated reporting and voluntary submission (Harpaz et al., 2013). Due to voluntary submission of the reports, for instance, it is estimated that these systems in the US and UK reflect less than 10% of the adverse effect occurrences (Inman & Pearce, 1993; Yang et al., 2012). Such shortcomings led pharmacovigilance practitioners to look for resources that are more efficient for post-marketing drug surveillance.

In recent years, Electronic Health Records (EHR) have been widely used in the healthcare industry to help practitioners in the collection, storage, and tracking patients' information and their treatment progress. The vast amount of data collected by EHRs as well as their increasing availability have made them interesting resources for pharmacovigilance researchers and presented opportunities to investigate and detect ADR signals[9] closer to real-time (Trifirò et al., 2009). Several data mining approaches have been proposed and applied by data scientists on EHR data in the past few years. Despite utilizing EHR data for pharmacovigilance purposes have gained much interest from European and Australian researchers, there is still a lack of sufficient research by the US academics and practitioners on the EHR data from the US healthcare market. Even though EHR data is generally more complete than data collected by spontaneous reporting systems, yet using EHR data for detection and prediction of ADR cases involves challenges such as complex data preprocessing requirements and various data documentation styles across different healthcare organizations (Harpaz et al., 2013).

Social media has also been considered as a key data source for monitoring drugs' post-marketing feedbacks in the recent few years by many researchers. A Pew internet research by Fox and Jones (2009) found that 61% of American adults look for health information (i.e. about specific diseases and treatments) online. This is normally done either through healthcare online forums such as 'Ask a patient',

[9] In pharmacovigilance, a signal is defined by the WHO as information on a possible causal relationship between an adverse event and a drug, which is unknown or incompletely documented (Trifirò et al., 2009).

‘Dailystrength’, ‘Yahoo health and wellness’, and ‘PatientsLikeMe'; or through social networks like Facebook and Twitter. Through social media, people talk about their concerns, seek advice about their diseases and health issues, and discuss their experiences with the medications they take. Such information, although noisy, is likely to appear there long before it is reported to any SAERS or detected via EHRs. Most of the time, the topics discussed by patients in social media are the ones which they are reluctant to discuss with their doctor, especially those prescribed for serious conditions like cancer, where the patient can experience high levels of anxiety due to the long-term exposure to the drugs (Benton et al., 2011; Leaman et al., 2010).

Due to these facts, many researchers have started to use social media for ADR detection and prediction purposes. Particularly Twitter as an open-access social network is used in several drug surveillance studies and several text-mining and sentiment analysis approaches were developed to identify patterns and signals of drug-event relationships (Bian, Topaloglu, & Yu, n.d., 2012; Culotta, 2010; Ginn et al., 2014; Nguyen et al., 2017; Prier et al., 2011). Apart from tweets, some researchers have also analyzed people’s comments in public healthcare forums mentioned above for the same purpose (Karimi et al., 2011; Leaman et al., 2010; X. Liu & Chen, 2013). Yet it seems that this field of research is still in its infancy period and calls for a lot more work.

3.3. RESEARCH QUESTION

In terms of the research goals, pharmacovigilance studies can be classified into three categories, namely detection, prediction, and understanding studies (Davazdahemami & Delen, 2018). Detection studies mainly aim at detecting existing associations (i.e., signals) between drugs and potential adverse reactions, often by analyzing historical usage data obtained from various resources. Prediction studies are those that utilize information about already known drug-ADR associations to predict possible ADRs for newly discovered as well as existing drugs. While detecting and predicting potential associations is a critical task, it is clear that such associations do not hold all the time and in the case of every patient. That is why, for instance, a particular patient might experience a side effect of a given drug, while that drug may not

have any adverse effect in another patient. Hence, it is crucial to investigate and understand the mechanism through which drugs develop side effects in the patients by identifying factors that either intensify or mitigate the strength of a drug-ADR association. This is, in fact, the goal of the understanding group of pharmacovigilance studies.

Prescription sequence, the sequence by which the drugs are prescribed and administered, is one of the factors that is suggested in the literature (Egger et al., 2003) to be investigated for its potential effect on the likelihood of ADRs development. To the best of our knowledge, no prior study has empirically investigated this potential effect, though. Hence, the research question we address in this study is:

RQ: *Does the sequence of drug prescription (and consequently drug administration) have any effect on the development of adverse drug reactions?*

3.4. MATERIALS AND METHOD

3.4.1. MATERIALS

In order to address the research question, we used a longitudinal observational electronic health records database, namely the Cerner HealthFacts data warehouse (http://www.cerner.com). Cerner HealthFacts data warehouse is the most comprehensive relational database in the US containing complete medical records of more than 63 million unique patients across the country. Cerner HealthFacts data warehouse contains time-stamped entries of patients' visits, physicians' diagnoses, lab tests, procedures, and prescribed drugs for both primary and secondary care visits. Prescription and diagnosis records of adult patients (18 or older) diagnosed with diabetes mellitus (ICD9- 250) for the first time during the 4-year period of 2012-2015 were extracted for analysis. The reason we limited our dataset to only diabetic patients was first to make the data more homogenous; and second the high rate of development of acute renal failure in this group of patients. The initial data involved prescription and visiting records of 377,910 unique patients.

There is an event-based stream of research focused on investigating the drugs associated with kidney diseases in general, and acute renal failure in particular (Coca & Perazella, 2002; Davazdahemami & Delen, 2019; Izzedine et al., 2005; Markowitz & Perazella, 2005; Naughton, 2008; Perazella, 2003; Singh et al., 2003). Loh and colleagues (2009) mention top ten categories of medications that cause kidney damage involving antibiotics, analgesics, COX-2 inhibitors, proton pump inhibitors, antiviral drugs, high blood pressure drugs, rheumatoid arthritis drugs, lithium, anticonvulsants, and chemotherapy drugs. The same set of drugs is mentioned, more or less, in other related studies as well. Since the focus of our study was on the drugs previously revealed to cause damages to the kidney, we focused on a set of 43 kidney-damaging medications from the top ten categories of drugs mentioned in the literature, and filtered the prescription records to retain only these class of drugs for analysis.

3.4.2. METHOD

In order to address the research question, we employed a case-control study design. In this design, the case group were those diabetic patients who developed acute renal failure (ICD9- 580) during the study period, and the control group involved those diabetic patients who were not diagnosed with renal failure by the end of study period.

Emergent pattern mining is a type of association rule mining that is used to detect differences between databases. The goal of emergent pattern mining is to find itemsets that are more frequent in one database (i.e. the case group in our study) compared to another (i.e. the control group).

For each patient identified as a *case* subject, we considered two *index dates;* 1) the date he or she was diagnosed with diabetes mellitus for the first time, and 2) the date the patient was diagnosed with renal failure for the first time. Two subjects were matched as *control* to each case-patient by matching on their age, race, gender, comorbidities and the first index date[10]. Moreover, the second index date for each *control*

[10] In case we couldn't match the controls who were diagnosed with diabetes on the same month and year, we searched through patients diagnosed in the months before or after the case's first index date to find a match.

is the same as its matching *case*'s second index date. Matching two controls for each case-patient makes the sample more representative of the population and leads to more accurate approximations of the support, as a measure of prevalence of drugs in the patients' prescription records (Reps et al., 2016); it also does not cause any issues in comparison of the two groups as the comparison criterion is the support, which is a percentage in nature. The emergent pattern mining will find sets of drugs that are prescribed more often prior to the second index date for the case subjects compared to the controls.

In short, two databases were created. D1 involves the prescription records pertaining to the last 10 visits prior to the second index date of patients in the case group (including 1,294 patients) whereas D2 contains the same information for their corresponding control patients (2,588 patients).

To investigate the potential effect of prescription sequence, we then applied frequent itemset mining to both D1 and D2 to identify both sequential and non-sequential frequent sets of drugs along with their support. Suppose I_{ns} is an itemset (containing k distinct items) identified as frequent in a database. Consider I_{ns} a non-sequential itemset in that the sequence of items (i.e. the sequence of drug prescription/administration) is not accounted for. Taking into account the sequence of items, then $k!$ itemsets can be driven from I_{ns} each with a unique sequence of items (i.e. sequential itemsets I_{s1}, I_{s2}, I_{s3}, ..., $I_{sk!}$). The non-sequential itemset I_{ns} as well as all its corresponding sequential sets, each would have a *support* index indicating the proportion (and the number) of transactions (i.e. patients) involving them in each database. Given these notations, the two approaches we used to investigate the effect of drug taking sequence follow.

The first approach is based on the Relative Risk notion. In the medical literature, Relative Risk (RR) is a measure used to indicate the risk of developing disease given exposure to its causes (Altman, 1990). Suppose that we expose the case group to a particular factor while keeping the control group unexposed. If we record the number of bad and good outcomes in each group (let's call them a and b for the case and c and d for the control group respectively), the RR then would be:

$$RR = \frac{a/(a+b)}{c/(c+d)} \qquad (Eq.1)$$

With the standard error of the log RR being:

$$SE\{\ln(RR)\} = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}} \qquad (Eq.2)$$

In this study, our goal is to evaluate the confounding role of the sequence factor. For a non-sequential itemset (I_{ns}), if *a* and *c* represent the number of its incidents within the case and control databases respectively; also *b* and *d* represent the number of case and control patients not having the itemset in their records, then Equation 1 would represent the relative risk associated with the itemset I_{ns}. Therefore an RR greater than one would suggest that the itemset I_{ns} is relatively more frequent in case patients than it is in control patients. Then it can be expected that someone having that itemset in his or her prescription records experience the negative outcome (i.e. ADR) more likely than someone who does not have it. The RR can be calculated for each of the possible sequential itemsets in a similar way. Having relative risk values for the non-sequential as well as the corresponding sequential itemsets, we argue that any inconsistency in these values implies the potential effect of the sequence factor. By inconsistency, we mean situations in which either one or more of the following conditions hold:

1- The RR associated with a non-sequential itemset is significant whereas at least one of the corresponding sequential itemsets have non-significant RR value or vice versa.

2- The RR associated with the non-sequential itemset is significant and greater than one (suggesting a *positive* confounding role for that itemset) whereas some of the corresponding sequential itemsets have significant values of RR that is less than one (suggesting a *negative* confounding role for that itemset).

3- The RR measures associated with both the non-sequential and sequential itemsets are significant and greater (less) than one, but they considerably differ in terms of magnitude. We considered a minimum difference of 0.5 in the magnitude of RR as the threshold as it suggests that the incidence of the corresponding sequential pattern is 50% more (or less) frequent than the non-sequential pattern in the case patients compared with controls.

We consider these conditions as inconsistency because they imply that patients experiencing the same set of drugs can have different likelihoods of experiencing the negative outcome depending on the sequence by which those drugs are prescribed/administered.

Table 3.1 indicates an example to clarify this approach. It is shown that the relative risk when the sequence is not taken into account (i.e. RR_{ns}) is significantly greater than 1 whereas accounting for the sequences only the last three itemsets involve a significant relative risk. That is, administering the same drugs in some particular sequences poses a higher risk of developing the negative reaction than other sequences.

The second approach relies on comparing the patterns of incidence of sequential itemsets across patient groups. Here the idea is that if the sequence of drug prescription has nothing to do with the likelihood of the negative outcome, then we should anticipate observing roughly the same pattern of incidence for sequential itemsets across the case and control groups. To compare these patterns, we first sort the sequential itemsets, separately in the case and control groups, according to their support in the corresponding databases and accordingly give a rank order to each itemset in each group. We then apply Spearman Rank Order Correlation to their ranks. A small, non-significant, or negatively significant correlation coefficient between the rank orders suggests that the pattern of incidence of the itemsets is different across the two groups of patients. In other words, some particular prescription sequences that are highly frequent among case patients are not so among the control group and vice versa.

An example is shown in Table 3.2. In this case, the rank order correlation for different sequences of an itemset in the case and control groups is non-significant. It suggests that the case and control patients have experienced different sequential patterns of the same itemset. It can also be realized by looking at the ranks of itemsets in the groups. For instance, sequential itemset {drg303, drg101, drg202} which is the most frequent pattern among case patients (i.e. rank=1), is the second least frequent pattern among control patients.

Figure 3.1 indicates a summary of the methods and procedures.

Table 3.1. Example of the First Approach

Non-sequential	*Case*		*Control*		RR_{ns} *(p-value)*	*Sequential*	*Case*		*Control*		RR_s *(p-value)*
Ins	***a***	***b***	***c***	***d***		*Is*	***a'***	***b'***	***c'***	***d'***	
{drg101,drg202,drg303}	91	1001	88	1462	1.468 (0.008)	{drg101,drg202,drg303}	41	1051	46	1504	1.265 (0.2650)
						{drg101,drg303,drg202}	37	1055	37	1513	1.419 (0.1260)
						{drg202,drg101,drg303}	42	1050	41	1509	1.454 (0.0830)
						{drg202,drg303,drg101}	45	1047	26	1524	2.457 (0.0002)
						{drg303,drg101,drg202}	50	1042	30	1520	2.366 (0.0002)
						{drg303,drg202,drg101}	49	1043	37	1513	1.880 (0.0030)

Table 3.2. Example of the Second Approach

Itemset	*Case*		*Control*		*Spearman Rank Order Correlation*
	Support (%)	*Rank*	*Support (%)*	*Rank*	
{drg101,drg202,drg303}	3.755	5	2.968	1	-0.55 (p=0.257)
{drg101,drg303,drg202}	3.388	6	2.387	3	
{drg202,drg101,drg303}	3.846	4	2.645	2	
{drg202,drg303,drg101}	4.120	3	1.677	6	
{drg303,drg101,drg202}	4.579	1	1.935	5	
{drg303,drg202,drg101}	4.487	2	2.387	3	

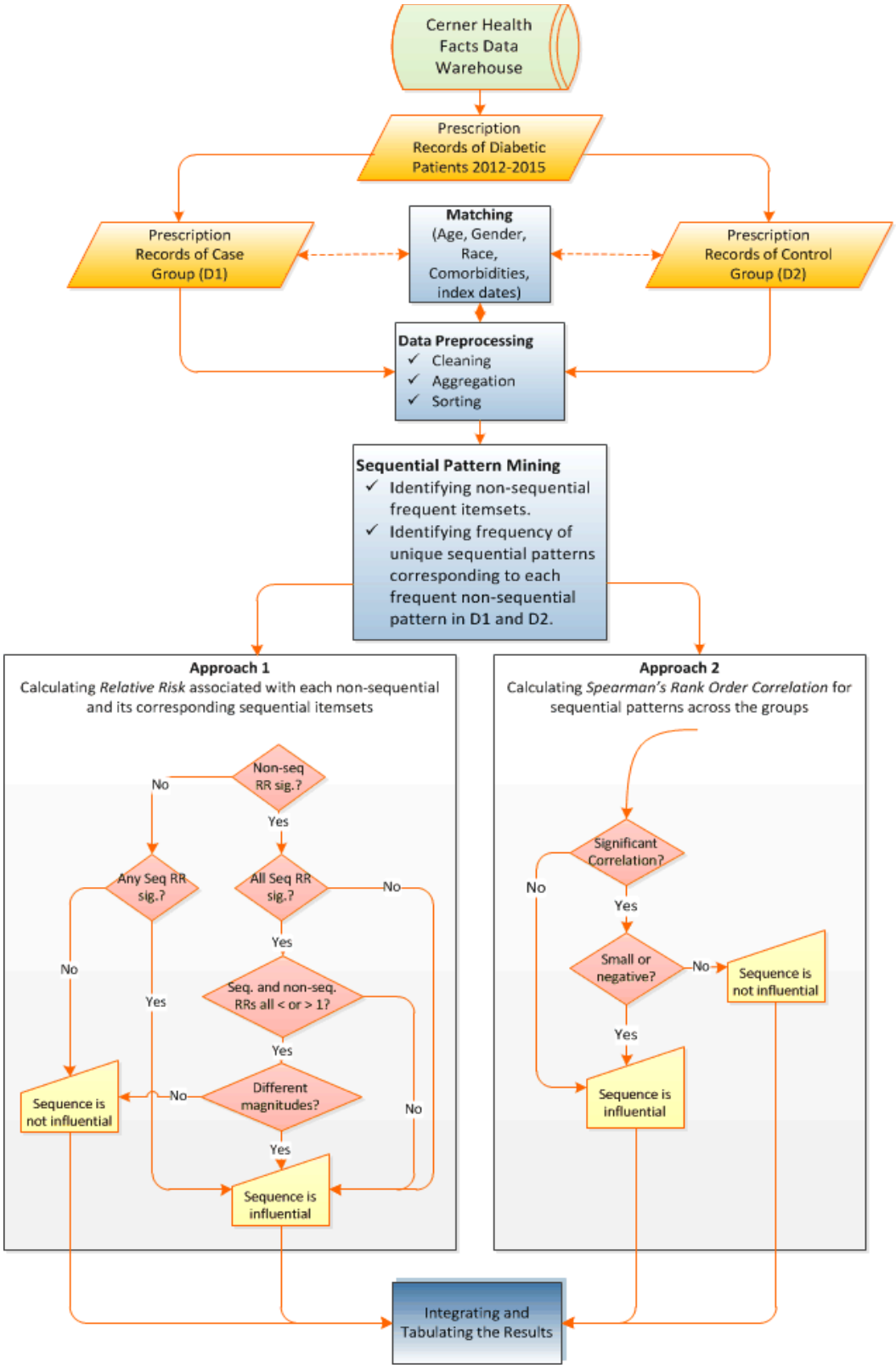

Figure 3.1. Summary of Methods and Procedures for Data Preparation and Analysis

3.5. RESULTS

Two data sets including prescription records of 1,294 case patients and 2,588 control patients were analyzed using the two developed approaches described. Table 3. indicates the age, race, gender, and comorbidities in the cohorts. It should be noted that since the patients in the two groups were matched on all these four factors, the profile shown in this table is indicative of both case and control cohorts of patients.

Table 3.3. The cohorts profile

Gender	*Race*	*Age*	*Comorbidities*
Male 55.76%	Caucasian 47.21%	Mean 40.77	Mean 5.38
Female 44.24%	African-American 41.22%	StDev 7.51	StDev 2.15
	Native American 2.66%		
	Hispanic 2.61%		
	Asian 1.04%		
	Other 5.26%		

To conduct this research, we particularly focused on 43 kidney-damaging medications from the top ten categories of drugs mentioned in the literature that were prescribed at least once for a patient in our initial data set. However, after identifying the case and control groups that number decreased to 31 as some drugs were not prescribed even once for the patients in our data set. Additionally, since the study was focused on diabetic patients we also controlled for 16 common medications that are frequently prescribed for those patients. That is, the identified frequent itemsets were filtered to only include known kidney-damaging and diabetic medications. The list of all medications included in the analyses is provided in Appendix 1.

Also, Table 3. represents the top ten frequent medications in each group of patients along with their relative frequency. As shown, excluding the top four drugs, the frequency patterns of

prescriptions are different among the case and control patients so that, for instance, tacrolimus is among the top ten for the case group while it is ranked 15th in the control group.

For each patient in either group, we then extracted all the prescriptions records of the 47 medications of interest related to the last ten visits prior to the second index date of that patient. Each medication in the data set was labeled with a sequence number indicating the chronological sequence of visit/prescription. Therefore, if two drugs were prescribed in the same visit, they both had the same sequence label in the data set; also, In case a particular drug was prescribed in two or more subsequent visits, we considered the earliest visit for its sequence label.

Table 3.4. Top frequent medication and their frequencies

Drug Generic Name	***Type***	*Rank (Case)*	***Freq** (Case)*	*Relative Freq. (Case)*	*Rank (Ctrl)*	***Freq** (Ctrl)*	*Relative Freq. (Ctrl)*
insulin (variations)	Diabetic	1	9399	35.92	1	8259	34.76
acetaminophen	KD	2	4479	17.12	2	3936	16.57
aspirin	KD	3	2757	10.54	3	2229	9.38
pantoprazole	KD	4	2706	10.34	4	1839	7.74
vancomycin	KD	5	1071	4.09	8	765	3.22
ketorolac	KD	6	948	3.62	6	1272	5.35
esomeprazole	KD	7	777	2.97	9	528	2.22
metformin	Diabetic	8	513	1.96	5	1599	6.73
tacrolimus	KD	9	513	1.96	15	150	0.63
ciprofloxacin	KD	10	501	1.91	11	321	1.35
Ibuprofen	KD	11	498	1.90	7	1107	4.66
glipizide	Diabetic	12	309	1.18	10	381	1.60

First, applying frequent itemset mining to the prescription records of the case and control groups, more than 5,000 frequent itemsets (not necessarily unique) were identified using the *association rule mining* predefined procedure (with non-sequential settings) in SAS Enterprise Miner. We limited our rule extraction procedure to only find itemsets involving up to 4 drugs with a minimum support of 0.5%. This resulted in 193 unique *non-sequential* frequent itemsets of size 4 or less. The reason we limited the size of itemsets to 4 was that for each non-sequential

itemset of size *k* we had to identify the *k!* sequential sets corresponding to that; hence an itemset of 5 would require us to identify 5!=120 sequential itemsets along with their supports whereas due to sample limitations most of those sequential patterns had not emerged in the prescription records whatsoever.

Table 3.5. Top frequent co-occurrences of drugs

Size	Itemset	Support (Case)	Rank (Case)	Support (Ctrl)	Rank (Ctrl)
2	{acetaminophen, insulin}	68.31	1	38.06	2
	{pantoprazole, insulin}	42.14	2	32.64	5
	{aspirin, insulin}	41.39	3	36.78	3
	{acetaminophen, pantoprazole}	33.70	4	35.10	4
	{aspirin, acetaminophen}	33.15	5	42.06	1
	{aspirin, pantoprazole}	19.96	6	19.22	9
	{ketorolac, insulin}	19.14	7	11.35	8
	{vancomycin, insulin}	17.31	8	9.61	9
	{ketorolac, acetaminophen}	16.12	9	13.29	6
	{ketorolac, insulin}	16.12	9	11.35	8
	{acetaminophen, vancomycin}	14.29	10	8.71	10
	{esomeprazole, insulin}	12.36	11	4.71	11
	{ibuprofen, acetaminophen}	9.34	16	13.03	7
3	{insulin, acetaminophen, pantoprazole}	32.97	1	16.32	2
	{insulin, aspirin, acetaminophen}	32.33	2	19.03	1
	{insulin, aspirin, pantoprazole}	19.41	3	8.84	6
	{pantoprazole, aspirin, acetaminophen}	16.12	4	7.29	10
	{acetaminophen, ketorolac, insulin}	15.75	5	11.35	4
	{vancomycin, insulin, acetaminophen}	13.92	6	8.13	7

	{pantoprazole, ketorolac, insulin}	10.35	7	5.41	15
	{insulin, acetaminophen, esomeprazole}	9.52	8	4.71	19
	{insulin, acetaminophen, ibuprofen}	9.16	9	11.03	5
	{vancomycin, pantoprazole, insulin}	8.97	10	<0.5	NA*
	{acetaminophen, metformin, insulin}	7.69	18	13.16	3
	{aspirin, metformin, insulin}	4.58	27	7.74	8
	{aspirin, metformin, acetaminophen}	4.03	32	7.48	9

*This itemset was not detected as a frequent one in the control group due to a support lower than the specified minimum of 0.5%.

Table 3.5 contains the top ten non-sequential sets of medications (size 2 and 3) co-occurred in the prescription records of the case and control groups along with their support within each data set. Again, there are remarkable differences in the frequency patterns as, for instance, while the set including aspirin and acetaminophen is the most frequent set (of size 2) in the prescriptions of the control group, it is ranked 5th among the itemsets corresponding to the case group. Similarly, whereas the set including vancomycin, pantoprazole, and insulin is among the top ten frequent sets (of size 3) in the case group, it was not even detected as frequent for the control group (given the minimum support threshold of 0.5%). These examples demonstrate considerable differences in the patterns of prescriptions between the two groups under study.

The table indicates that while metformin was not within any of the top itemsets of size 2 for either group, it was included in three of the top itemsets of size 3 (ranked 3rd, 8th , and 9th) in the control group. From a clinical viewpoint, this fact suggests that probably metformin has to do with lowering the chances of developing renal failure in diabetic patients, since it was more frequently prescribed for patients in the control group, who ultimately did no develop renal failure during the study period. This confirms findings from prior research regarding metformin and its role in developing renal failure (Davazdahemami & Delen, 2019; von Websky et al., 2013; Yamanouchi, 2010). In addition, the relatively higher support (and rank) of itemsets including

insulin and/or pantoprazole in the case group suggests a potential enhancing role for these medications with regard to the risk of renal failure. For insulin, such a risk-enhancing role has been discussed in prior medical studies (Davazdahemami & Delen, 2019; Fatourechi et al., 2009). Of course, scrutinizing differences between these prescription patterns in more detail could provide us with more clinical insights regarding the role of diabetes and KD drugs in separation or together, in developing renal failure in diabetic patients. Nevertheless such a discussion is beyond the scope of the present study; in addition confirming each of those signals require a vast investigation of the medical literature and possibly conducting randomized clinical trials.

At the next step, running the association rule mining procedure with a sequential rule setting in SAS Enterprise Miner, we obtained the frequency of incidence for the sequential itemsets, corresponding to the non-sequential sets identified earlier, across the case and control patients.

Based on the frequency of each sequential itemset, we also ranked them in a descending order in both case and control databases. The frequency, as well as rank orders, were then used in the calculation of relative risk (RR) as well as the Spearman's rank order correlation (i.e. the two approaches explained in the methods section), respectively, for each of the 193 unique non-sequential frequent itemsets. Finally, we applied the rules discussed in the methods section to determine whether in each case there is a considerable inconsistency between sequential and non-sequential patterns across the two groups of patients.

Table 3.6 illustrates a summary of our analyses using the two approaches. As shown, based on the RR criterion, we found that in 165 out of 193 itemsets (i.e. 85.5%) at least one of the three conditions for the significance of sequence effect was present. Also using the Spearman’s rank correlation, in 144 (74.6%) of the itemsets a considerable effect for the sequence of prescription was inferable. Interestingly, there were only 16 (8.3%) of cases in which none of the

approaches find enough evidence for the influence of prescription sequence in developing the adverse outcome.

The table also shows the number of itemsets in which the conclusion about the influence of sequence was consistent or inconsistent. To test how the two approaches were consistent in terms of their conclusions, we conducted a chi-square test of independence on this 2x2 crosstab (i.e. Table 3.6). The outcome provides support for the consistency of the approaches ($\chi^2 = 17.43$, $p<0.001$).

Table 3.6. Results

Correlation *RR*	*Influential*	*Not influential*	*Subtotal*
Influential	132	33	165
Not influential	12	16	28
Subtotal	144	39	**193**

Overall, the results indicate that with respect to 68.4% of itemsets, both approaches agreed upon the existence of some significant effects that can be attributed to the sequence by which the medications were prescribed. Moreover, in 91.7% of itemsets, at least one of the two approaches revealed such an effect.

3.6. DISCUSSION

In this study, we investigated the potential effect of prescription sequence in the development of adverse drug events. While such potential effect had been mentioned in the literature, it was not empirically investigated prior to this study. To this end, using longitudinal transactional data obtained from the Cerner HealthFacts data warehouse and employing two independent approaches, we looked into the effect of the prescription sequence of 31 known kidney-damaging drugs on the development of acute renal failure in diabetic patients.

The results from each approach suggest a significant effect that can be attributed to the sequence by which the drugs were prescribed along the patients' timeline. Moreover, it was shown that both approaches used to assess this effect are significantly in accordance with one another whereas they were designed independently and using different criteria. This suggests additional proof for the existence of such an effect. In fact, the two proposed approaches are the main theoretical contribution of the present study. Future research may employ the proposed approaches to assess similar sequential effects in other medical contexts.

Also from an empirical viewpoint, we believe that the fact that sequence of prescriptions may result in developing adverse drug effects and intensify their probability suggests designing and implementing of new clinical decision support systems to help physicians in their prescription decisions by taking into account the patients' historical transactions and provide them with appropriate alerts to prevent possible ADRs.

While we believe that our results strongly suggest a nontrivial effect attributable to the prescription sequence, yet of course we agree that our study involves some limitations. Particularly, even though we limited our sample to diabetic patients and controlled for their demographics, diabetes history, and common diabetic medications, still some important factors were not controlled due to sample limitations. Of the highest importance was the effect of patients' exact comorbidities that we did not control for in this study because doing such would greatly affect our sample size. It was not easy to find a control match for each case patient with exactly the same comorbidities. Hence, we limited this control to only a major disease which is highly prevalent among Americans (i.e., diabetes) and also controlled for the total number of comorbidities as a general measure of patients' wellness. We also simply assumed that by controlling for age and other demographics we are also partly controlling for other particular comorbidities that might be attributable to aging. Future research may employ larger samples and

fully control for the effect of comorbidities. A larger sample also provides the possibility to include itemsets involving more than four drugs.

It should be noted here that while the two independently designed approaches in this study consistently and strongly suggest a significant association between prescription sequences and ADR development, yet this association is not necessarily causal. In other words, our results provide a strong signal for the pharmacovigilance researchers to take the prescription sequence factor more seriously in their analyses and also justify conducting further studies in a more controlled environment to assess the causality of the detected association.

In this study, we assumed that all the medications prescribed by doctors were administered by the patients until it was discontinued by their doctor again. In fact, it was not practically possible to monitor whether every drug had been administered as recommended. However, we believe that it is reasonably realistic to assume that medications prescribed in a particular visit were taken before those prescribed in the subsequent visit. Accordingly, instead of taking into account prescription timestamps we considered the timestamps of doctor visits as the base for sequence analysis. In other words, medications prescribed within the same visit were given the same sequence order, which was different from sequences of medications prescribed in the previous or subsequent visits. Future research may possibly address this limitation by using another source of data in which medication administrations were monitored. Finally, future research may improve the proposed approaches by taking into account the effect of drug dosages prescribed for the patients.

In conclusion, it should be said that even though the present study is not perfect (makes certain assumptions and involves some limitations), because of the size and richness of the data used and the methods and measurement metrics developed and administered, its promising results might be considered as a promising baseline for deeper investigations on the effect of prescription

sequence as it possibly can prevent development of ADRs in millions of patients, improving their lives and wellbeing, and saving considerable amounts of money for them as well as for the Government, caregivers, and insurance agencies.

CHAPTER IV

ESSAY III: A CHRONOLOGICAL PHARMACOVIGILANCE NETWORK ANALYTICS APPROACH FOR PREDICTING ADVERSE DRUG EVENTS

ABSTRACT

Objectives: This study extends prior research by combining a chronological pharmacovigilance network approach with machine-learning techniques to predict adverse drug events (ADEs) based on the drugs' similarities in terms of the proteins they target in the human body. The focus of this research, though, is particularly centered on predicting the drug-ADE associations for a set of eight common and high-risk ADEs.

Materials and methods: A large collection of annotated MEDLINE biomedical articles were used to construct a drug-ADE network, and the network was further equipped with information about drugs' target proteins. Several network metrics were extracted and used as predictors in machine-learning algorithms to predict the existence of network edges (i.e., associations or relationships).

***Results*:** Gradient boosted trees (GBT) as an ensemble machine-learning algorithm outperformed other prediction methods in identifying the drug-ADE associations with an overall accuracy of 92.8% on the validation sample. The prediction model was able to predict drug-ADE associations, on average, 3.84 years earlier than they were actually mentioned in the biomedical literature.

Conclusion: While network analysis and machine-learning techniques were used in separation in prior ADE studies, our results showed that they, in combination with each other, can boost the power of one another, and predict better. Moreover, our results highlight the superior capability of ensemble type machine-learning methods in capturing drug-ADE patterns compared to the regular (i.e., singular), machine-learning algorithms.

Keywords: *Adverse Drug Events, Network Analysis, Machine Learning, Prediction, Target Proteins, Ensemble Models.*

4.1. INTRODUCTION

Today every new drug to be approved by the healthcare authorities and marketed by pharmaceutical companies has to pass through numerous clinical trials, which on average take 10-15 years.(Iizuka, 2007) These clinical trials mainly aim at ensuring *efficacy* and *safety* of the drug. A considerable number of drugs fail to get US Food and Drug Administration (FDA) approval due to the potential threats their usage involve even though they might show effectiveness with regard to treating some specific diseases.(Trame, Biliouris, Lesko, & Mettetal, 2016) Nevertheless, even such tough regulations and approval procedures do not 100% guarantee the safety of a drug since those trials themselves involve several limitations and may fail to capture some potential, in some cases serious, safety issues (Karimi et al., 2015; Zeng et al., 2002).

A classic example of such cases is *Rofecoxib*; an NSAID approved in 1999 that became highly welcomed by the physicians in a short time. The drug was originally aimed to treat acute pains and Osteoarthritis, but after a while turned out to cause heart attacks in more than 100,000 patients and ended up being withdrawn by the FDA in 2004. During that time, apart from the lives threatened, this possibly avoidable problem also imposed huge losses to pharmaceutical and insurance companies.

Pharmacovigilance (a.k.a. drug safety surveillance) is a field of science that monitors the drugs during their lifecycle to detect, assess, and understand their potential adverse effects and prevent harms and injuries caused thereof. Although pharmacovigilance activities begin early after drug discovery, its role becomes more critical after drug *approval*, when humans start to take it.

In pharmacovigilance terminology, an Adverse Drug Event (ADE) refers to any injury occurred to a patient caused by administering a drug. It should be noted that there is still no consensus on this terminology across pharmacovigilance and pharmacoepidemiology studies. Some studies(Nikfarjam, Sarker, O'Connor, Ginn, & Gonzalez, 2015b; Reps et al., 2016; Trame et al., 2016; Zeng et al., 2002) define ADE as any injury which not necessarily has a causal relationship with the drug (e.g., injuries due to human errors) and therefore use the more specific term Adverse Drug Reaction (ADR) to refer to the injuries

directly caused by the drug. However, in the present study, we stick with the term ADE while we emphasize that by ADE we mean a drug-induced (i.e. causally related) injury in patients. It is estimated that in the United States, each ADE case in community hospitals on average costs $3,000 (Karimi et al., 2015; Zeng et al., 2002). Also, ADEs are reported by the Australian Commission on Safety and Quality in HealthCare (ACSQHC) to cause about 400,000 admissions to general practitioners in Australia with a population of only 23 million.(Karimi et al., 2015)

Given the great potential health and financial threats mentioned, and considering the fact that today the trend is toward faster approval processes and smaller clinical trials, especially in oncology and rare diseases(Trame et al., 2016), a great amount of research has been done in the past decade to find faster and more effective ways to detect, predict, understand, and prevent ADEs before they affect too many (or ideally any) people.

In this study, we extend the extant literature on ADE prediction by proposing a chronological network analytics approach that can help pharmacovigilance practitioners to save lots of time, money, and more importantly, lives by enabling them to predict potential ADEs *prior to drugs approval*. The proposed approach uses historical information of known drug-ADE relationships in addition to similarities between new and approved drugs, in terms of the proteins they target in human bodies, and tries to predict potential ADEs.

The remainder of this article is organized as follows; the following section reviews the extant literature on detection, prediction, and understanding of ADEs and states the research goals. Then, we explain the materials and methods used to conduct the study followed by the results. Finally, we discuss the contributions of our study and conclude with a few potential future research directions.

4.2. BACKGROUND

4.2.1. RESOURCES FOR ADE STUDIES

Before discussing different approaches used in prior ADE studies, in this section, we discuss various data sources used by researchers to conduct those studies. Four main types of data source have been identified in the literature. The following four sub-sections introduce these resources and mention prior research conducted using each.

4.2.1.1. Spontaneous Reporting Systems

As an effort to rapidly detect and prevent ADEs in the post-marketing phase, many countries and international organizations have run Spontaneous Reporting Systems (SRSs); systems designed to allow patients and professionals to submit their reports of suspected ADEs. This includes the World Health Organization's (WHO) Individual Case Safety Reports (ICSR) database, the yellow card system of Medicines and Healthcare products Regulatory Agency (MHRA) in the UK, and the FDA Adverse Event Reporting System (FAERS) in the US.(Karimi et al., 2015) Although SRSs were the main source for ADE studies for several years, their limitations such as over-reporting and voluntary submissions(Harpaz et al., 2013; Karimi et al., 2015) made pharmacovigilance practitioners look for more efficient alternatives.

4.2.1.2. Electronic Health Records

During the past decade, Electronic Health Records (EHR) have been widely used in the healthcare industry to help practitioners in the collection, storage, and tracking patients' information. The vast amount of data collected by EHRs along with their increasing availability have made them interesting resources for pharmacovigilance researchers and enabled them to detect ADE signals closer to real-time.(Trifirò et al., 2009) Yet, using EHR data involves challenges like complex data preprocessing requirements and multiple standards across different databases(Harpaz et al., 2013).

4.2.1.3. Social Media

Recently social media has been introduced as a novel resource for conducting ADE as well as other healthcare studies. Virtual communities such as health forums (e.g., DailyStrength and PatientsLikeMe) and social networks (e.g., Twitter and Facebook) are places where people discuss their daily health-related experiences and concerns. Such information, although noisy, is likely to appear there long before it is reported to any SRS or recorded in any EHR(Benton et al., 2011; Leaman et al., 2010) and this has made social media a precious resource for early detection of ADEs.

4.2.1.4. Biomedical Literature

Recently, researchers have realized biomedical literature as well as chemical and biological databases as feature-rich sources for ADE studies. Databases such as PubMed, PubChem, KEGG, and DrugBank are rich sources of information about drugs, their chemical and biological characteristics, and their identified ADEs.

4.2.2. ADE STUDIES: DETECTION, PREDICTION, AND UNDERSTANDING

Due to considerable potential costs and damages of ADEs, in the past decades, there has been a great deal of research on this issue in many disciplines including pharmacology, economics, and information systems. While the ultimate goal of all of these studies is to identify drugs' potential ADEs and prevent losses of lives and money thereof, they pursue different tools and strategies to achieve that goal. We believe that ADE studies can be classified into three distinct categories, namely detection, prediction, and understanding.

Detection studies are the largest group of ADE research works focused on finding new and undetected ADE signals (i.e., associations, not necessarily causal) between the existent drugs (already in the market) and adverse events. The signals detected by these studies need to be assessed and verified by clinical trials. ADE detection studies heavily rely on applying statistical (Cai et al., 2017; van Puijenbroek et al., 2002) or data mining (Friedman, 2009; Harpaz et al., 2013; Harpaz, Chase, et al., 2010; X. Liu &

Chen, 2013; Nikfarjam et al., 2015b; Reps et al., 2016; Trifirò et al., 2009; Yang et al., 2012) methods and quasi-experimental settings to the historical data from SRSs (Cai et al., 2017; DuMouchel, 1999; Harpaz et al., 2013; van Puijenbroek et al., 2002), EHRs (Friedman, 2009; Haerian et al., 2012; Harpaz, Chase, et al., 2010; Reps et al., 2016; Trifirò et al., 2009), or social media(Hoang et al., 2016; J. Liu et al., 2016; X. Liu & Chen, 2013; Nikfarjam et al., 2015b; Yang et al., 2012) to extract signals from them.

In the ADE prediction studies, on the other hand, instead of detecting signals for the existent drugs using collected data from their past usage experiences, the focus is on creating signals for the new drugs before they cause any adverse events to the patients. The strategy in this group of studies is mainly to find similarities between the existent and the new drugs and thereby to predict ADEs for the new drugs given the already known relationships between their similar existent drugs with the corresponding ADEs. The statistical regression-based methods(Atias & Sharan, 2011; Cami et al., 2011) as well as machine learning techniques(L.-C. Huang, Wu, & Chen, 2011; L. Huang, Wu, & Chen, 2013; M. Liu et al., 2012) are the dominant methods used by the researchers for this purpose. Also in terms of data sources, prediction studies heavily rely on the biomedical literature as well as drug databases including chemical, physical, and biological information of drugs since such resources enable them to identify drug similarities. Just like ADE detection studies, this group of studies also serve as a signal detector, but the difference is they capture signals for new drugs as well.

The last group of ADE studies in our taxonomy is those focusing on verifying ADE signals and understanding the mechanism through which the drug causes the ADE. Pharmacoepidemiology and pharmacometrics studies fall into this group as they use mathematical and parametric models of biology, pharmacology, and physiology to clarify and understand mechanisms of both beneficial and adverse molecular interactions.(Trame et al., 2016) Several different types of models have been used by the researchers in this group, among which Pharmakokinetics and Pharmacodynamics (Albrecht et al., 2017; Chiang et al., 2018; Lazaar et al., 2016; Vazzana et al., 2015; Wedemeyer & Blume, 2014) are the most popular modeling approaches. The former focuses on modeling how the organism affects the drug, whereas

the focus in the latter is on studying the effect of the drug on the organism; so the researchers usually employ them together, as the complement to each other to determine optimal dosing as well as the beneficial and adverse effect of drugs. In terms of data sources, this group of studies mostly rely on drug databases and EHR historical transactions.

4.2.3. NETWORK ANALYSIS AND PHARMACOVIGILANCE

Although Network Analysis (NA) have been widely used in many areas of science including sociology, communication, biology, economics, and computer science starting from a few decades ago, its application in pharmacovigilance studies is hardly older than 10 years. The main reason for that could be the lack of appropriate information systems and infrastructures for collecting the data required for constructing networks in large scale before the early 2000s.

Networks have been used in pharmacovigilance research with a variety of data sources and for different purposes (not limited to ADE prediction, which is the case in our study). Some researchers, including Ball et al.(Ball & Botsis, 2011b) and Botsis et al.(Botsis & Ball, 2011) used network representations of vaccines and their reported ADEs in the FDA's VAERS to identify the frequent patterns of interactions. Also Zhang et al.(Zhang, Tao, He, Kanjamala, & Liu, 2013) showed that patterns identified in vaccine-vaccine networks can contribute to the vaccine ontology knowledge base. A recent study by Kim et al.(Kim et al., 2018) on hospitalized patients with hematologic malignancies revealed that network centrality metrics can be used to identify the most important causes for drug-related problems (DRPs) by constructing a cause-DRP network using ward pharmacists' documentations in hospital settings.

Apart from the mentioned studies that have used descriptive and qualitative techniques to extract information/knowledge from networks, there are also a few studies focused on using networks of drugs and ADEs for predicting their associations. For instance, Atias and Sharan(Atias & Sharan, 2011) and Cami et al.(Cami et al., 2011) in their studies used a diffusion process and a logistic regression model, respectively, with NA to make ADE predictions. Nevertheless, to the best of our knowledge, NA has not been combined

with ML methods in the literature so far for the prediction purposes and the present study is the first one to do so.

4.2.4. RESEARCH GOALS

While statistical and machine-learning techniques have been widely used with various data sources for pharmacovigilance prediction purposes (Bender et al., 2007; Hammann, Gutmann, Vogt, Helma, & Drewe, 2010; L.-C. Huang et al., 2011; LaBute et al., 2014; J. Liu et al., 2016; Pouliot et al., 2011), we found only a few studies that have utilized the incredible potential of network analysis approaches to explore drug-ADE associations. Specifically, Atias and Sharan(Atias & Sharan, 2011) applied a network-based diffusion process to predict drugs' ADEs. Also, in a later study, Cami et al.(Cami et al., 2011) employed logistic regression (LR) technique in a network approach using data from biomedical literature and chemical databases to predict drug-ADE associations.

We extend the ADE prediction research by employing a Chronological Pharmacovigilance Network (CPN) along with machine-learning techniques to predict drugs' ADEs. For this purpose, we use biomedical literature citations as the main source of data for extracting previously identified drug-ADE associations. Additionally, we incorporate information about the target proteins of drugs into our network structure to make it more informative for training machine-learning algorithms.

A target protein is a chemically definable molecular structure that will undergo a specific interaction with chemicals that we call drugs because they are administered to treat or diagnose a disease(Imming, Sinning, & Meyer, 2006). In other words, drugs act by binding to specific target proteins and changing their biochemical or biophysical activities to treat their indicated diseases.(Yildirim, Goh, Cusick, Barabási, & Vidal, 2007) Given that, we argue that knowledge about the similarity of drugs, in terms of the proteins they target, can contribute to the quality of ADE predictions. Moreover, we believe that complexity of drug-ADE relationships is so much that machine-learning algorithms, and especially ensemble models are more efficient than statistical-based methods (e.g., LR) in capturing that.

4.3. MATERIALS AND METHODS

4.3.1. MATERIALS

We integrated data from two sources, namely, National Library of Medicine's (NLM) MEDLINE and the DrugBank's database of drug-target proteins, in order to operationalize our approach towards modeling of the CPN. MEDLINE, a subset of PubMed database, is a bibliographic database of biomedical information from multiple disciplines that includes more than 29M citations started from 1946. What sets MEDLINE apart from the rest of PubMed is the added-value of using the NLM controlled vocabulary, Medical Subject Headings (MeSH), for indexing, cataloging, and searching for biomedical documents. Also, DrugBank is a freely accessible online drug database including biological, chemical, and genetic information of 10,986 approved and experimental drugs.

First, we selected a sample of eight common and high-risk ADEs reported in the literature(Trifirò et al., 2009) (Acute renal failure, Myocardial infarction, Leukopenia, Agranulocytosis, Rhabdomyolysis, Neutropenia, Thrombocytopenia, and Anemia) and collected all MEDLINE articles mentioning at least one of them as the ADE identified in the article. To this end, we used a search strategy based on NLM's MeSH thesaurus (see the appendix 2). NLM indexers select the most appropriate MeSH indexes to resume the full content of an article after reading the full text.(Avillach et al., 2013)

The initially downloaded dataset involved 10,890 unique publications mentioning associations among 657 drugs with 769 ADEs. However, considering only drugs approved by the FDA by December 2017, we ended up with a dataset including 9,672 publications, 582 drugs, and 732 ADEs.

Second, we used DrugBank(Wishart et al., 2006) database to extract target proteins associated with each FDA-approved drug. While most drugs target only a few proteins in the human body, some have many targets.(Yildirim et al., 2007) In addition to the 582 drugs in initial dataset, we included information about 217 other drugs having at least one common target with one of those 582 drugs. Therefore, the integrated dataset used in the study involved 799 approved drugs and 732 ADEs. The publication years as well as the

drugs approval dates were also imported into our data to be used in constructing training and validation datasets for the model building stage. All of the drugs and ADEs were then mapped to their unique terms from the NLM's Unified Medical Language System (UMLS) for consistency.

4.3.2. METHOD

4.3.2.1. Network Construction

A chronological approach was employed to construct drug-drug and drug-ADE relationships in the network. The ultimate goal in pharmacovigilance is to identify as many as possible ADEs in the pre-marketing phase. Hence, in order to have a valid prediction model, one is only allowed to use drugs information as well as the known drug-ADE associations that are available prior to the time of the drug approval. Given this idea and using the dates of publications and drug approvals, we used all of the information available prior to 2001, to predict drug-ADE associations for the drugs marketed during 2001-2017.

First, a network was constructed in which both drugs and ADEs were considered as vertices. An undirected edge was created between two drugs if they had at least one common target protein. Additionally, a drug was connected to an ADE in the network if there was at least one PubMed article published before 2001 mentioning such association. The network involved all of the 799 drug vertices (regardless of their approval dates) and 10,094 drug-drug edges indicating common target proteins, as well as 5,264 drug-ADE edges representing pre-2001 identified associations. We kept aside drug-ADE relationships recognized (for the first time) during 2001-2017 to validate our prediction model since they were unknown at the time of prediction (i.e., beginning of 2001). Figure 4.1 provides a visualization of the network created.

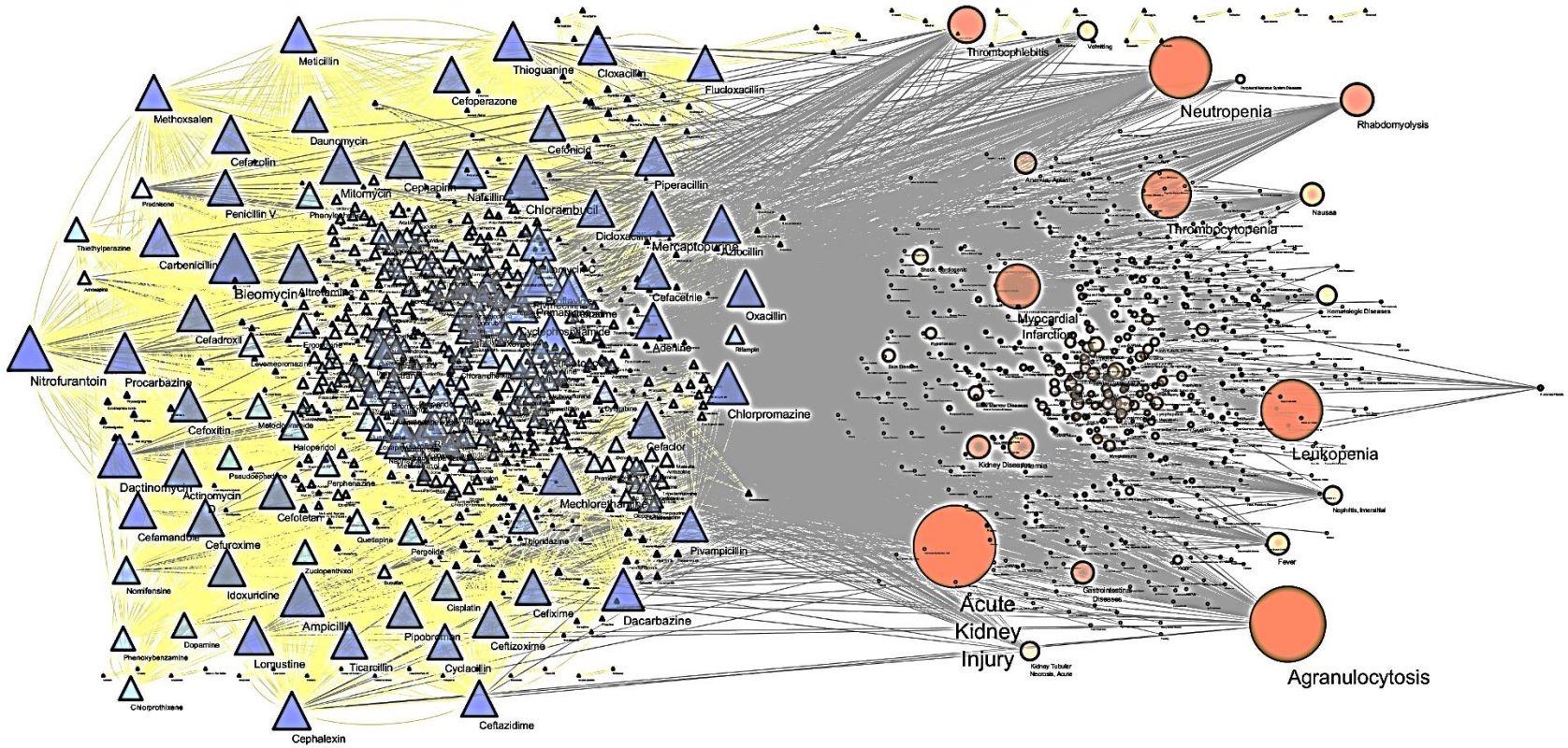

Figure 4.1. The Drug-ADE network created by Cytoscape v3.6. Triangle (blue) nodes represent drugs and circular (orange) nodes represent ADEs. Yellow links between drugs indicate the existence of at least one common target protein by the drugs connected. Also, gray links between drugs and ADEs indicate an association mentioned in at least one PubMed article for the corresponding drug and ADE.

4.3.2.2. Network Metrics

Drug-ADE links were considered as the unit of analysis in this study. Since the focus of our study was on a set of eight common and critical ADEs, we created our dataset by considering all possible combinations of the 799 drugs with those ADEs (i.e., 6,392 records). Once the network was constructed, we extracted seven similarity- as well as three centrality-based metrics for each record to be used as link predictors. The metrics had been proposed in the network analysis literature for link prediction purposes(Cami et al., 2011; Liben-Nowell & Kleinberg, 2007; Zhou, Lü, & Zhang, 2009).

The three centrality-based metrics we used were the absolute difference, product, and sum of degree centralities of corresponding drug and ADE vertices involved in each link. All of these metrics were used in similar studies (Cami et al., 2011; Liben-Nowell & Kleinberg, 2007) to capture assortativity[11] (absolute difference and ratio) and preferential attachment[12] (sum and product).

Table 4.1 indicates the similarity-based predictors extracted from the network along with their definitions. While all of the similarity metrics are defined based on the notion of commonality of neighborhoods between the two nodes of interest, each reflects a different aspect of similarity. In these definitions, Γ(i) and Di denote the set of neighbors and degree of node i, respectively. Also, d and a were used to denote drug and ADE, respectively. Therefore, $\Gamma(d) \cap \Gamma(a)$ refers to the set of common neighbors of a drug and an ADE; similarly, $\Gamma(d) \cup \Gamma(a)$ refers to the set of all of their neighbors.

[11] Assortativity is defined as the extent to which highly central drugs tend to connect more frequently to highly- or low-central ADEs.(Cami et al., 2011)

[12] Preferential attachment denotes that the probability that a new edge has a specific node x as an endpoint, is proportional to the current number of neighbors of x.(Liben-Nowell & Kleinberg, 2007)

The network metrics were obtained with the help of the igraph package in R; a comprehensive package for network analysis.

Table 4.1. Similarity Metrics and their Formulaic Definitions

Similarity Index	Definition/Formula
Jaccard coefficient(Jaccard, 1912)	$\frac{\lvert\Gamma(d) \cap \Gamma(a)\rvert}{\lvert\Gamma(d) \cup \Gamma(a)\rvert}$
Dice index(Dice, 1945)	$\frac{2 \times \lvert\Gamma(d) \cap \Gamma(a)\rvert}{D_d + D_a}$
Adamic/Adar index(Adamic & Adar, 2003)	$\sum_{z\,\epsilon\Gamma(d)\cap\Gamma(a)} \frac{1}{\log\lvert\Gamma(z)\rvert}$
Simpson index(Simpson, 1960)	$\frac{\lvert\Gamma(d) \cap \Gamma(a)\rvert}{Min(D_d, D_a)}$
Geometric index(Bass et al., 2013)	$\frac{\lvert\Gamma(d) \cap \Gamma(a)\rvert^2}{D_d \times D_a}$

Apart from the five mentioned standard similarity metrics, we also incorporated two derived similarity metrics for each drug-ADE pair. First, for each drug-ADE pair, we calculated the average Jaccard similarity of the corresponding drug with all of the drugs connected to the ADE. To calculate this variable we constructed and used a network including only the drugs (and no ADEs) and extracted Jaccard similarities of each drug with all of those connected drugs. We believe that such a variable reflects how a new drug is chemically similar to drugs in general and therefore might cause the same ADE as they do. Based on the same logic and in a similar manner, for each drug-ADE pair in our dataset, we also incorporated average distance from the corresponding drug to all of the drugs connected to the ADE (i.e., the second derived variable). While the first derived variable captures general similarity of each drug with the connected drugs based on their direct neighborhoods, the second one takes into account the indirect links as well.

In the end, a binary target variable was created for each drug-ADE pair to indicate whether that association actually exists according to the MEDLINE citations.

4.3.2.3. Training and Validation Data

Once we formed the dataset using the network, we applied the following rules to divide the dataset into training and validation subsets to train the prediction models and test their efficiency.

Drug-ADE pairs that were actually discovered after 2001, regardless of the drug approval year, placed into the *validation* dataset. All of the remaining pairs including drugs approved after 2001 were also added to the validation set. All other pairs were classified as the *training* dataset. Applying these rules, we ended up with a training dataset containing 5,357 records with 1,087 (i.e., 20.3%) positive responses (target=1). Also, the validation set contained 1,035 records with a response rate of 14.6% (i.e., 151 positives).

4.3.2.4. Prediction Model

We used the *training* dataset to train and build our prediction models. Four different classification algorithms were employed, namely Artificial Neural Network (ANN), Gradient Boosted Trees (GBT), Random Forests (RF), and Logistic Regression (LR).

Due to the unbalanced proportion of positive and negative responses in training data, the Synthetic Minority Oversampling Technique (SMOTE)(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) was applied to make a balanced training (model building dataset), henceforth avoid biases in the training of the models. The KNIME analytics platform version 3.5.1 (a free and open source analytics software platform) was used to build the classification models. Figure 4.2 shows a flow-chart like graphical depiction of the data preparation and model building methods and procedures.

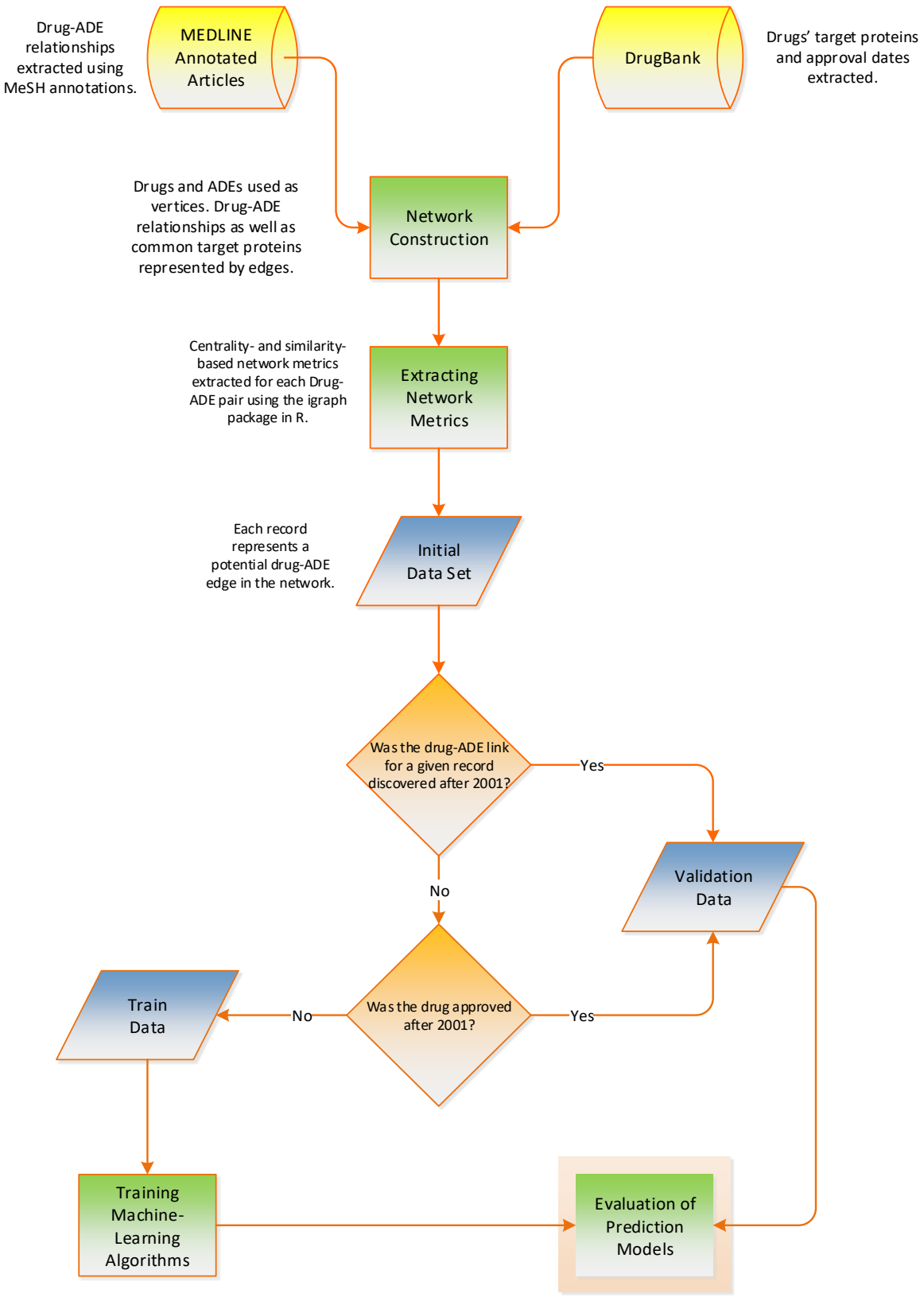

Figure 4.2. A flow-chart like graphical depiction of the methods and procedures

4.4. RESULTS

4.4.1. MODELS ACCURACY

Table 4.2 shows the prediction results of the best models of each algorithm on the validation data. As shown, RF and GBT, the two ensemble-type of algorithms provided more accurate results than ANN and LR[13]. Also overall, GBT turned out to be the best model among all with an overall accuracy of 92.8% and the ability to correctly predict 72.8% of real drug-ADE associations in the validation data (i.e., sensitivity). It suggests that given historical information about drug-ADE associations as well the target proteins of drugs, our best model was able to predict 110/151 (i.e., 72.8%) of drug-ADE associations that were actually discovered during a 17-year period after building the prediction network. In addition, the Positive Predictive Value (PPV) for the GBT model indicates that out of 143 pairs predicted as associations by this model, 110 (i.e. 76.9%) were real associations reported in the MEDLINE. Also overall, the PPV values highlight the superiority of the two ensemble models over the individual models (i.e., ANN and LR) in which only around half of the positive predictions were correct.

Table 4.2. Prediction Models' Accuracy Statistics

Model	Accuracy	Sensitivity	PPV	AUROC
ANN	85.5%	65.6%	50.3%	0.868
RF	92.1%	64.9%	77.2%	0.893
GBT	92.8%	72.8%	76.9%	0.916
LR	85.7%	56.3%	50.9%	0.793

[13] The parameter settings for the best models in KNIME were as follows:

- RF: Split criterion: Gini index; Number of models: 400; no limit on the Tree Depth and Node Size.
- ANN: Number of hidden layers: 2; Number of neurons per layer: 5; Maximum number of iterations: 60.
- GBT: Number of model: 300; Learning rate: 0.3; Tree depth limit: 4; No attribute sampling.

In the only similar study we are aware of in the literature, conducted by Cami et al.,(Cami et al., 2011) historical drug-ADE associations along with drugs' taxonomical and intrinsic properties (e.g., molecular weight, atom count, and so on) from pre-2005 years were used in multiple LR models to predict associations identified during 2005-2010. Comparing to their best model (AUROC=0.869), two of our prediction models (RF and GBT) provide superior results while prediction power of our ANN model is also comparable to theirs.

Our further investigation revealed that from the 110 true positive predictions made by the GBT model, 29 were related to post-2001 marketed drugs, which, given 42 actual positive associations, means a 69% true positive rate for these new drugs. The true positive rate for older drugs was 74.3% (i.e., 81/109 actual associations). Moreover, it turned out that out of 143 positive predictions, 102 were related to pre-2001 marketed drugs, which (given that 81 of the true positive cases were pre-2001 marketed drugs) suggests a PPV of 79.4% (81/102) for this group. Also, 41 positive predictions were related to post-2001 marketed drugs resulted in a PPV of 70.7% (i.e. 29/41). These statistics seem reasonable given the higher number of historical publications about these drugs that makes the model better trained for classifying their associations.

Furthermore, in terms of sensitivity, our approach outperforms Cami et al.'s, where their best-reported model had a sensitivity of 61.2% compared to 72.8% of our model. While this difference might be argued to be due to the narrower focus of our study (i.e., including 8 ADEs), we believe it mostly has to do with the more informative nature of the network we used to train our models as well as the ability of machine-learning techniques to capture complex/non-linear relationships compared to statistical methods like LR. As Table 4.2 shows, our LR model did not perform as good as the other three machine-learning methods. Nevertheless, it is still comparable and complementary to the models provided by Cami et al.(Cami et al., 2011) Even comparing our results to those of the studies that have employed machine-learning techniques (mostly using drugs' structural variables as predictors) with a non-network approach(L.-C. Huang et al., 2011; L. Huang

et al., 2013; M. Liu et al., 2012), our approach outperforms theirs in terms of most of the accuracy statistics. Table 4.3 indicates that especially in terms of sensitivity and PPV, using an ensemble machine-learning model along with the network approach has significantly improved ADE predictions.

Table 4.3. Comparison of model results with the best results reported by similar studies

Article	Network approach	Model	Chem.	Bio.	Other	Acc	Sens	PPV	AUROC
Liu et al. (M. Liu et al., 2012)	No	SVM	Yes	Yes	Yes	0.967	0.631	0.662	0.952
Huang et al.(L. Huang et al., 2013)	No	SVM	Yes	Yes	No	NR	NR	NR	0.760
Cami et al.(Cami et al., 2011)	Yes	LR	Yes	No	Yes	NR	0.608	NR	0.869
Huang et al.(L.-C. Huang et al., 2011)	No	SVM	No	Yes	No	0.675	0.632	NR	0.771
Present study	Yes	GBT	No	Yes	No	0.928	0.728	0.769	0.916

Chem.* indicates whether chemical features of drugs are used for ADE prediction.*Bio. Indicates whether biological features of drugs are used for ADE prediction.Other* indicates whether other features (e.g. taxonomical, phenotypical, etc.) of drugs are used for ADE prediction.

4.4.2. VARIABLES IMPORTANCE

Having the superior prediction model identified, we further investigated how each of the predictors contributed to the model accuracy. To this end, we dropped predictor variables one at a time from our data and ran the best prediction model. Each time we recorded model's AUROC to be compared to that of the original model. Table 4.4 indicates the amount of decrease in AUROC after dropping each predictor along with the relative importance of variables based on normalized AUROC differences.

Table 4.4. Variable Importance Statistics

Dropped Variable	New AUROC	AUROC_diff	Relative Importance
Degree_product	0.86	0.056	1
Degree_ratio	0.864	0.052	0.875
Degree_sum	0.872	0.045	0.656
Geometric index	0.884	0.032	0.250
Avg_Jacc_connected	0.885	0.031	0.219
Adamic/Adar index	0.887	0.029	0.156
Simpson index	0.887	0.029	0.156
Dice index	0.888	0.028	0.125
Jaccard index	0.889	0.027	0.094
Avg_dist_connected	0.891	0.025	0.031
Abs_degree_diff	0.892	0.024	0

As shown in this table *Degree_product*, *Degree_ratio*, and *Degree_sum* representing preferential attachment as well as assortativity of drug-ADE pairs turned out to have the highest contribution to the predictive power of the best (i.e., GBT) model. It suggests that our centrality-based predictors generally played a more important role than similarity-based metrics. Of the three top predictors, two of them *(Degree_product* and *Degree_sum*) were also among the top three in the study performed by Cami et al..(Cami et al., 2011) *Degree_product* was also identified as a strong predictor in the work conducted by Liben-Nowell & Kleinberg.(Liben-Nowell & Kleinberg, 2007) Interestingly, the results show that one of the derived variables, namely *Avg_Jacc_connected*, was the fifth most important predictor with a relative importance of around 22%. Also consistent with prior research,(Cami et al., 2011) *Abs_degree_diff* was the least important predictor of network links.

Finally, by investigating our true positive predictions and considering the actual years that corresponding drug-ADE associations were identified for the first time, we realized that, on average, our model was able to predict ADEs 3.84 years (SD=1.97 years) before they were mentioned in PubMed articles. Table 4.5 indicates a summary of associations predicted by the model for the eight ADEs of interest along with the top associated drug predicted for each. The "average probability" column in this table shows the average across all the real associations, not just those that correctly predicted. The results show that disregarding a few exceptions, the model performance in predicting associations across the ADEs of interest was roughly the same. This suggests generalizability of the proposed approach as it has performed equally well with regard to various ADEs.

Table 4.5. Summary of Predicted Associations by ADE

ADE	Real Associations	Predicted Associations	Average Time Saving	Average Probability	Top Associated Drug
Acute Renal Failure	32	26 (81%)	3.62	0.7655	Ceftazidime (2)*
Agranulocytosis	12	10 (83%)	5.70	0.8293	Albendazole (6)
Anemia	9	6 (67%)	2.67	0.7277	Ribavirin (3)
Leukopenia	4	4 (100%)	3	0.9436	Dexamethasone (1)
Myocardial Infarction	27	18 (67%)	3.72	0.7035	Doxazosin (5)
Neutropenia	17	12 (71%)	3.83	0.7739	Flucytosine (2)
Rhabdomyolysis	40	27 (68%)	3.92	0.6910	Doxylamine (0)
Thrombocytopenia	10	7 (70%)	3.50	0.7517	Tamoxifen (0)

*The numbers in front of drug names in the last column indicate the number of years the model predicted their associations with the corresponding ADE earlier than it was published in PubMed.

4.4.3. ANALYSIS OF PREDICTION ERRORS

Even though our prediction model performed well in terms of common accuracy metrics, it is always insightful to qualitatively analyze the cases that a model fails to accurately predict. Such an undertaking may involve both the drug-ADE pairs that were predicted to be associated while they actually were not (i.e., the false positive cases) and the drug-ADE pairs that were actually associated whereas the model failed to predict their association correctly (i.e., the false negative cases).

We found 41 false negative predictions made by the model. Our further investigation revealed that 20 (i.e., around half) of them are related to the drugs approved after 2008. More specifically, we realized that six drugs, all approved after 2008, account for 17 (i.e., 41%) of false negative predictions. We then looked into the known associated ADEs, other than the eight ADEs of interest, for each of those six drugs before 2001 (i.e., when they were experimental drugs yet) which were used to train the prediction model. We found out that, compared to average (i.e., 6.58), the number of known associations for most of those six drugs was considerably low with only one having more than 5 known ADEs. Given these findings, we believe that one main reason for the model making those false negative predictions could be the relatively low number of known ADE associations (i.e., network edges) involving those drugs in the training dataset. Since we only used network metrics as the predictor variables, such lack of sufficient drug-ADE edges may possibly affect all of the predictor variables related to the corresponding drugs. Of course, one way to address this issue is to change the cutoff point for data partitioning (which is currently 2001) so that our training data include more of the known MEDLINE citations involving the drugs approved more recently. In the present study, however, changes in the cutoff year considerably affect the size of validation dataset[14], which could jeopardize the validity of the prediction model.

[14]. For instance, we changed the cutoff year to 2003 and we ended up with only 732 records (i.e. a decrease of 303 records) in the validation dataset.

Our predictions also involved 33 false positive cases. Again, to further investigate the potential causes for those classification errors we looked into the specific drugs and ADEs involved. We realized that around 61% (i.e., 20) of these cases were related to the relatively older drugs, approved in early 90's or even earlier. For such drugs, due to numerous biomedical studies conducted on them over time, the number of known ADE associations and consequently their degree centrality in the network tend to be higher than newer drugs. This directly inflates the centrality-based predictors of drug-ADE pairs, namely degree_ratio, degree_sum, and degree_product. Moreover, it was shown that these were the top three influential predictors of network links in our study. Table 4.6 compares the values of these three predictors, on average, for the false positive versus true positive as well as true negative cases. Clearly, the predictor values in false positive cases are far from those of the true negative cases and are very close to the cases correctly predicted as positive.

Table 4.6. Comparing Top Predictors' Values in False Positive, True Positive, and False Negative Predictions

Predictor	False Positives	True Positives	True Negatives
Degree_Sum	234.82	237.55	203.05
Degree_Product	7131.76	7615.02	3344.65
Degree_Ratio	0.21	0.23	0.11

Overall, our findings suggest that for older drugs the centrality-based predictor values are overly inflated, due to the higher number of citations involving in them, that other predictor variables cannot help the model to discern those cases from actual/real positive cases. Hence, probably incorporating some other network-independent informative covariates suggested in the literature (e.g., molecular or chemical features of drugs) can address this issue to some extent and help the model to better differentiate between positive and negative cases.

4.5. DISCUSSION AND CONCLUSION

In this study, we proposed a new approach to predict ADEs by constructing drug-ADE networks, using biomedical citations as well as drugs target proteins information, and then employing network metrics as predictors of associations in machine-learning algorithms.

While both NA approaches and ML techniques had been employed in the past separately, to the best of our knowledge, the present study is the first one, which employs ML along with an NA approach together in a single study. The promising results we obtained suggest that combining these two powerful tools can enhance the results we may get from each in separation. Our proposed approach outperformed the prior studies (see Table 4.) while the number of predictor variables used in this study is relatively lower than that of the similar studies.

We believe that part of these superior results owes to the incredible power of *ensemble* machine-learning algorithms. As shown in our results, the two ensemble algorithms (i.e., RF and GBT) considerably outperformed the other two approaches. That is simply because of the higher power of ensemble algorithms in capturing sophisticated patterns in the data. While statistical and regular machine-learning techniques train a single model (either linear or non-linear) to reflect the relationship between the variables, ensemble algorithms sample the data hundreds of times and use those samples to build hundreds of prediction models. Then to predict a new case they vote from the created models to specify the final prediction. This way, instead of a single model, which is subject to sample randomization errors, many models are employed to yield predictions.

The results also suggest that *assortativity* and *preferential attachment* (i.e., centrality-based metrics) are better predictors of network edges than similarity-based metrics (e.g., Jaccard coefficient). This is in line with the results from Cami et al.(Cami et al., 2011) and Liben-Nowell & Kleinberg(Liben-Nowell & Kleinberg, 2007). Additionally, we introduced two derived similarity-based network metrics, namely Avg_Jacc_connected and Avg_Distance_connected, for

predicting network edges, and it turned out that the former is among the top five most important predictors. In terms of relative importance, Table 4. shows that this derived variable has contributed to the quality of model around 50% more than the Adamic/Adar index and around 100% more than Jaccard index, two popular similarity-based metrics. It suggests that considering the similarity of a drug with the drugs already associated with an ADE provides more useful information in predicting drug-ADE associations, than considering the similarity of that drug with the ADE itself.

Although the present study is particularly focused on eight highly common and risky ADEs, we argue that the high accuracy of our predictions has nothing to do with that matter because we did not incorporate any information about the ADEs or their relationships in building our prediction models. All of the information used to train our prediction models were historical drug-ADE associations as well as drug-target proteins. Hence, we believe that replicating our approach on a larger scale and with a higher number of ADEs would result in the same quality results, if not better.

Another limitation of this study is that it does not account for the *strength* of drug-ADE associations in the construction of the network. In network analysis, using the strengths of associations as the linkage weights and extracting weighted metrics is a popular and informative approach provided that the weights are assigned to the links in a meaningful way. Considering the frequency of citations mentioning a given association as the strength of that association is not a decent and even meaningful way for weighing the network edges because this frequency does not necessarily reflect the strength of association and might very well be, for instance, due to the high amount of risk involved in the corresponding ADE. Therefore, in this study, we used an unweighted network for the analysis. Future research could extend our approach by developing a way to score drug-ADE associations and use weighted network metrics in building the prediction models.

While our best model performed well in terms of sensitivity, it still made 33 false positive and 41 false negative predictions. Even though we analyzed some potential reasons for these prediction errors, we suspect that a portion of the false positive cases, especially those involved recently approved drugs, might be actually real drug-ADE associations that have not yet been studied and mentioned in biomedical citations. This could be also the case with all the other ADE prediction studies where the models yield a considerable number of false positives. Future research may focus on such cases resulted from ADE predictions and try to investigate them using clinical trials or by analyzing patients transactions from EHR data using methods like prescription sequence symmetry analysis(Pratt et al., 2015; Tsiropoulos, Andersen, & Hallas, 2009).

Given the relatively high accuracy of predictions resulted from employing network approach, both in this study and the other few similar works, we strongly encourage future researchers to utilize the incredible power of networks for prediction purposes in pharmacovigilance. Especially, we believe that incorporating more data sources to construct more informative training networks can lead to even better predictions in the future. Specifically, chemical, physical, and molecular features of drugs (e.g., molecular weight, heavy atom count, melting point, etc.) can be added to the model as covariates to enhance its prediction power. We believe that one big methodological advantage of our study is producing quality results using a considerably lower number of predictors than prior studies (Atias & Sharan, 2011; Cami et al., 2011; L.-C. Huang et al., 2011; M. Liu et al., 2012) and relying mostly on the power of networks and ensemble ML algorithms to identify patterns. Nevertheless, as discussed in the Results section, incorporating some additional covariates can potentially improve the model while maintaining its simplicity. Databases such as DrugBank and PubChem are freely accessible and rich sources of information about drugs that can be used for this purpose.

We used the biomedical literature citations as the only resource for the known drug-ADE associations in constructing the network. There are, however, some other resources such as the side

effect resource (SIDER) database (http://www.sideeffects.embl.de) or some commercial databases like Lexicomp (http://www.lexi.com) that can be used for this purpose as well. Future research may extend our approach by incorporating multiple resources to add as many as possible drug-ADE links to the network since doing so can enhance the information extend of the network and potentially improves the quality and accuracy of the predictions.

Similarly, with regard to the drug targets, we only used a single source (i.e. DrugBank) for this purpose. Even though it was suggested in prior research(Barneh, Jafari, & Mirzaie, 2015) that network-based organization of DrugBank data, particularly the drug similarity network (DSN), can potentially contribute to the prediction of side effects, and we showed that in this study, yet it involves some potential limitations. DrugBank is primarily focused on labeling targets from a pharmacokinetic point of view and possibly includes some determinants of drug disposition labeled as drug targets. We are not sure, though, whether the existence of such instances has improved or limited our model performance since on one hand, they may make the DSN more information-rich, but on the other hand, the nature of drug similarities may not be the same across the network.

Finally, we believe that the chronological settings used in the present study to construct a drug-ADE network based on the chronological drug approvals and known ADE associations may be extended by future researchers to conduct a longitudinal study by constructing multiple drug-ADE networks at different time points and show that evolution of this network over time enriches its informativeness and yields better predictions both in general and with regard to specific associations.

CHAPTER V

SUMMARY AND CONCLUSION

With advances in computer science and data science in the past few decades and given the huge amount of data being accumulated every day on the health care data repositories, computer-based data analytics approaches have been widely developed and applied with the aim of improving health care processes.

One of the areas of health care which has been benefited from such efforts is drug safety. Traditionally, a newly discovered drug had to pass through decades of randomized clinical trials before being approved by the health care authorities and provided to the market. That was basically due to great deal of uncertainty with regard to the various types of risks posed by administering the drug to the real patients. Data analytics and statistical methods have significantly contributed to drug safety by reducing such risks of uncertainty through analyzing data from historical medication usage as well as chemical structures and biomedical characteristics of drugs. These efforts has led to more timely *detection*, more accurate *prediction*, and more effective *prevention* of drugs' adverse events.

5.1. CONTRIBUTIONS

In the first essay, an analytics approach has been extended with the aim of taking into account drug-drug interactions in determining the confounding role of particular medications with regard to developing an adverse event. While each medication, in isolation, may lead to various adverse events in a patient, in a real world patients are usually being prescribed with multiple medications either for a single condition or for multiple conditions diagnosed. So the question is how taking other drugs may intensify or mitigate the already identified effect of a given drug in developing its corresponding adverse events? In other words, how can we realize the confounding role of a given drug with regard to a known and established drug-ADE relationship. By extending an emergent pattern mining method and applying it to the real prescription records of more than 370,000 diabetic patients, in the first study we examined such confounding roles for a group of common diabetic medications on the adverse effect of a group of drugs known to cause acute kidney failure. The results explain the contradictory roles reported in the medical literature for the confounding role of common diabetic medications in absence of other potentially relevant medications.

The second essay provides two independent approaches to examine the effect of prescription sequence on the likelihood of developing adverse drug events. While the sequence by which a given set of drugs are administered was suggested in the literature as a potential factor in developing adverse events, this effect was not empirically examined in the past. The two designed data analytic approaches were applied to the prescription records of a large group of diabetic patients to examine the effect of sequence on developing acute renal failure, as a common adverse event among this type of patients. The results obtained from the two independent approaches consistently revealed a significant effect on the likelihood of developing renal failure, which was attributable to the drugs' prescription sequence.

In the third essay, two freely accessible feature-rich data sources, namely MEDLINE and DrugBank, were employed to construct a network of drugs and their associate adverse events already mentioned in the biomedical literature. The idea in this study was to use the known drug-ADE associations as well as similarities between the newly discovered drugs with already-marketed drugs in terms of their target proteins in the human body to predict potential ADEs of the new drugs. Our results showed that employing network metrics as the predictors of drugs' ADE along with using advanced ensemble machine learning algorithms can significantly improve the accuracy of ADE predictions.

5.2. ASSUMPTIONS AND LIMITATIONS

The present work involves several limitations as discussed below.

In the first and second study, even though we limited our sample to diabetic patients and controlled for their demographics, diabetes history, and common diabetic medications, still some important factors were not controlled due to sample limitations. Of the highest importance was the effect of patients' exact comorbidities that we did not control for in these two study because doing such would greatly affect our sample size. It was not easy to find a control match for each case patient with exactly the same comorbidities. Hence, we limited this control to only a major disease which is highly prevalent among Americans (i.e., diabetes) and also controlled for the total number of comorbidities as a general measure of patients' wellness. We also implicitly assumed that by controlling for age and other demographics we are also partly controlling for other particular comorbidities that might be attributable to aging.

In the same studies we assumed that all the medications prescribed by doctors were administered by the patients until it was discontinued by their doctor again. In fact, it was not practically possible to monitor whether every drug had been administered as recommended. However, we believe that it is reasonably realistic to assume that medications prescribed in a particular visit

were taken before those prescribed in the subsequent visit. Accordingly, instead of taking into account prescription timestamps we considered the timestamps of doctor visits as the base for sequence analysis in the second essay.

Moreover, a limitation in the third study is that it does not account for the *strength* of drug-ADE associations in the construction of the network. In network analysis, using the strengths of associations as the linkage weights and extracting weighted metrics is a popular and informative approach provided that the weights are assigned to the links in a meaningful way. Then future research could extend our approach by developing a way to score drug-ADE associations and use weighted network metrics in building the prediction models. Additionally, with regard to the drug targets, we only used a single source (i.e. DrugBank) for this purpose. Even though it was suggested in prior research (Barneh et al., 2015) that network-based organization of DrugBank data, particularly the drug similarity network (DSN), can potentially contribute to the prediction of side effects, and we showed that in the third study, yet it involves some potential limitations. DrugBank is primarily focused on labeling targets from a pharmacokinetic point of view and possibly includes some determinants of drug disposition labeled as drug targets. We are not sure, though, whether the existence of such instances has improved or limited our model performance since on one hand, they may make the DSN more information-rich, but on the other hand, the nature of drug similarities may not be the same across the network.

5.3. FUTURE RESEARCH DIRECTIONS

This work leads to several areas of future research in drug safety as discussed next.

a) *Studying the confounders of high-risk adverse events:* as discussed in the first study, acute renal failure was studied as the case in that work because it was identified as a high-risk ADR in the medical literature, that can lead to death in case of occurrence. There are, however, several other high-risk ADRs (e.g., myocardial infarction) common

among different groups of patients that need to be studied in terms of their associated drugs and the confounding role of other relevant drugs in decreasing or increasing the likelihood of developing them. Also future research may expand the proposed approach in the first study by employing larger data set (involving a wider time window) which allows for controlling the effect of specific diseases (as opposed to controlling only for the total comorbidities) as well.

b) *Designing an automated decision support system to monitor prescription sequences:* as discussed in the second study, the sequence by which medications are administered can play a significant role in developing ADRs. Future research may employ large sets of historical prescription records to identify the sequential patterns leading to each given ADR and then use those identified patterns to design a clinical decision support system. Such a system can monitor the prescription records of each particular patient and provide the physicians with appropriate alerts when there is some intensified risk of developing a high-risk ADR involved, due to prescribing drugs in certain sequences.

c) *Using network analytics and ensemble machine learning to improve ADR predictions:* in the third study it was shown that how employing network metrics along with ensemble machine learning algorithms can help in identifying sophisticated patterns within drug-ADR associations and apply them effectively in predicting potential ADRs for newly discovered drugs. Future research may extend this idea by constructing more feature-rich networks of drugs and ADRs and extracting and developing new network metrics to be used as predictors of ADRs. In addition, with recent advances in computer hardware and provision of infrastructures for conducting deep learning analyses, future research may employ data sets including a large number of chemical, biomedical, and physical features of the drugs with the aim of predicting their exact ADRs.

REFERENCES

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, *25*(3), 211–230.

Afkarian, M., LR, Z., YN, H., & al, et. (2016). Clinical manifestations of kidney disease among us adults with diabetes, 1988-2014. *JAMA*, *316*(6), 602–610. Retrieved from http://dx.doi.org/10.1001/jama.2016.10924

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487–499).

Ahmad, S. R. (2003). Adverse drug event monitoring at the Food and Drug Administration. *Journal of General Internal Medicine*, *18*(1), 57–60.

Albrecht, D., Ellis, D., Canafax, D. M., Combs, D., Druzgala, P., Milner, P. G., & Midei, M. G. (2017). Pharmacokinetics and pharmacodynamics of tecarfarin, a novel vitamin K antagonist oral anticoagulant. *Thrombosis and Haemostasis*, *117*(04), 706–717.

Altman, D. G. (1990). *Practical statistics for medical research*. CRC press.

Arthur, N., Bentsi-Enchill, A., Couper, R., Duclos, P., Edwards, I., Fushimi, T., … Lazdins-Helds, J. (2002). The Importance of Pharmacovigilance-Safety Monitoring of Medicinal Products. *World Health Organization*.

Ashley, C. (2018). Renal failure-how drugs can damage the kidney. *Pathophysiology*, *14*, 0.

Atias, N., & Sharan, R. (2011). An algorithmic framework for predicting side effects of drugs. *Journal of Computational Biology*, *18*(3), 207–218.

Avillach, P., Dufour, J. C., Diallo, G., Salvo, F., Joubert, M., Thiessard, F., … Fieschi, M. (2013). Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: A contribution from the EU-ADR project. *Journal of the American Medical Informatics Association*, *20*(3), 446–452. http://doi.org/10.1136/amiajnl-2012-001083

Baksh, S. N., McAdams-DeMarco, M., Segal, J. B., & Alexander, G. C. (2018). Cardiovascular safety signals with dipeptidyl peptidase-4 inhibitors: A disproportionality analysis among high-risk patients. *Pharmacoepidemiology and Drug Safety*, *27*(6), 660–667.

Ball, R., & Botsis, T. (2011a). Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS? *Clinical Pharmacology & Therapeutics*, *90*(2), 271–278.

Ball, R., & Botsis, T. (2011b). Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS. *Clinical Pharmacology and Therapeutics*, *90*(2), 271–278. http://doi.org/10.1038/clpt.2011.119

Bao, Y., Kuang, Z., Peissig, P., Page, D., & Willett, R. (2017). Hawkes process modeling of adverse drug reactions with longitudinal observational data. In *Machine Learning for Healthcare Conference* (pp. 177–190).

Baralis, E., Cagliero, L., Cerquitelli, T., & Garza, P. (2012). Generalized association rule mining with constraints. *Information Sciences*, *194*, 68–84.

Barneh, F., Jafari, M., & Mirzaie, M. (2015). Updates on drug–target network; facilitating

polypharmacology and data integration by growth of DrugBank database. *Briefings in Bioinformatics*, *17*(6), 1070–1080.

Bass, J. I. F., Diallo, A., Nelson, J., Soto, J. M., Myers, C. L., & Walhout, A. J. M. (2013). Using networks to measure similarity between genes: association index selection. *Nature Methods*, *10*(12), 1169.

Bate, A., & Evans, S. J. W. (2009). Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology and Drug Safety*, *18*(6), 427–436.

Bender, A., Scheiber, J., Glick, M., Davies, J. W., Azzaoui, K., Hamon, J., … Jenkins, J. L. (2007). Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem*, *2*(6), 861–873.

Benton, A., Ungar, L., Hill, S., Hennessy, S., Mao, J., Chung, A., … Holmes, J. H. (2011). Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of Biomedical Informatics*, *44*(6), 989–996.

Berhanu, P., Perez, A., & Yu, S. (2007). Effect of pioglitazone in combination with insulin therapy on glycaemic control, insulin dose requirement and lipid profile in patients with type 2 diabetes previously poorly controlled with combination therapy. *Diabetes, Obesity & Metabolism*, *9*(4), 512–520. http://doi.org/10.1111/j.1463-1326.2006.00633.x

Bian, J., Topaloglu, U., & Yu, F. (n.d.). Towards Large-scale Twitter Mining for Drug-related Adverse Events. In *SHB'12 : proceedings of the 2012 ACM International Workshop on Smart Health and Wellbeing* (pp. 25–32). Hawaii.

Bian, J., Topaloglu, U., & Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and*

wellbeing (pp. 25–32). ACM.

Borah, A., & Nath, B. (2018). Identifying risk factors for adverse diseases using dynamic rare association rule mining. *Expert Systems with Applications*, *113*, 233–263. http://doi.org/10.1016/J.ESWA.2018.07.010

Botsis, T., & Ball, R. (2011). Network analysis of possible anaphylaxis cases reported to the US vaccine adverse event reporting system after H1N1 influenza vaccine. *Studies in Health Technology and Informatics*, *169*, 564–568. http://doi.org/10.3233/978-1-60750-806-9-564

Cai, R., Liu, M., Hu, Y., Melton, B. L., Matheny, M. E., Xu, H., … Waitman, L. R. (2017). Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. *Artificial Intelligence in Medicine*, *76*, 7–15. http://doi.org/https://doi.org/10.1016/j.artmed.2017.01.004

Cami, A., Arnold, A., Manzi, S., & Reis, B. (2011). Predicting adverse drug events using pharmacological network models. *Science Translational Medicine*, *3*(114), 114ra127-114ra127.

Casillas, A., Pérez, A., Oronoz, M., Gojenola, K., & Santiso, S. (2016). Learning to extract adverse drug reaction events from electronic health records in Spanish. *Expert Systems with Applications*, *61*, 235–245. http://doi.org/10.1016/J.ESWA.2016.05.034

Caster, O., Norén, G. N., Madigan, D., & Bate, A. (2010). Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *3*(4), 197–208.

Cavalieri, S., Cosmai, L., Genderini, A., Nebuloni, M., Tosoni, A., Favales, F., … Licitra, L. (2018). Lenvatinib-induced renal failure: two first-time case reports and review of literature.

Expert Opinion on Drug Metabolism & Toxicology, *14*(4), 379–385.

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chiang, C., Zhang, P., Wang, X., Wang, L., Zhang, S., Ning, X., … Li, L. (2018). Translational high-dimensional drug interaction discovery and validation using health record databases and pharmacokinetics models. *Clinical Pharmacology & Therapeutics*, *103*(2), 287–295.

Classen, D. C., Pestotnik, S. L., Evans, R. S., Lloyd, J. F., & Burke, J. P. (1997). Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality. *Jama*, *277*(4), 301–306.

Coca, S., & Perazella, M. A. (2002). Rapid communication: acute renal failure associated with tenofovir: evidence of drug-induced nephrotoxicity. *The American Journal of the Medical Sciences*, *324*(6), 342–344.

Cohen, C., Houdeau, A., & Khromava, A. (2018). Comment on "Central Demyelinating Diseases After Vaccination Against Hepatitis B Virus: A Disproportionality Analysis Within the VAERS Database." *Drug Safety*, 1–3.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics* (pp. 115–122). ACM.

Davazdahemami, B., & Delen, D. (2018). A chronological pharmacovigilance network analytics approach for predicting adverse drug events. *Journal of the American Medical Informatics Association*, *25*(10), 1311–1321. http://doi.org/10.1093/jamia/ocy097

Davazdahemami, B., & Delen, D. (2019). The confounding role of common diabetes medications in developing acute renal failure: A data mining approach with emphasis on drug-drug

interactions. *Expert Systems with Applications*, *123*, 168–177. http://doi.org/10.1016/j.eswa.2019.01.006

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Duke, J. D., Han, X., Wang, Z., Subhadarshini, A., Karnik, S. D., Li, X., … Li, L. (2012). Literature Based Drug Interaction Prediction with Clinical Assessment Using Electronic Medical Records: Novel Myopathy Associated Drug Interactions. *PLoS Computational Biology*, *8*(8). http://doi.org/10.1371/journal.pcbi.1002614

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician*, *53*(3), 177–190.

Egger, S. S., Drewe, J., & Schlienger, R. G. (2003). Potential drug–drug interactions in the medication of medical patients at hospital discharge. *European Journal of Clinical Pharmacology*, *58*(11), 773–778.

Fatourechi, M. M., Kudva, Y. C., Murad, M. H., Elamin, M. B., Tabini, C. C., & Montori, V. M. (2009). Hypoglycemia with Intensive Insulin Therapy: A Systematic Review and Meta-Analyses of Randomized Trials of Continuous Subcutaneous Insulin Infusion Versus Multiple Daily Injections. *The Journal of Clinical Endocrinology & Metabolism*, *94*(3), 729–740. Retrieved from http://dx.doi.org/10.1210/jc.2008-1415

Fox, S., & Jones, S. (2009). The social life of health information. Pew Research Center.

Friedman, C. (2009). Discovering novel adverse drug events using natural language processing and mining of the electronic health record. In *Conference on Artificial Intelligence in Medicine in Europe* (pp. 1–5). Springer.

Fujita, H., Morii, T., Fujishima, H., Sato, T., Shimizu, T., Hosoba, M., … Drucker, D. J. (2014). The protective roles of GLP-1R signaling in diabetic nephropathy: possible mechanism and therapeutic potential. *Kidney International*, *85*(3), 579–589.

Galárraga, L. A., Teflioudi, C., Hose, K., & Suchanek, F. (2013). AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 413–422). ACM.

Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., Oconnor, K., Sarker, A., … Gonzalez, G. (2014). Mining Twitter for Adverse Drug Reaction Mentions : A Corpus and Classification Benchmark. In *proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM)* (pp. 1–8).

Haerian, K., Varn, D., Vaidya, S., Ena, L., Chase, H. S., & Friedman, C. (2012). Detection of Pharmacovigilance-Related Adverse Events Using Electronic Health Records and Automated Methods. *Clinical Pharmacology & Therapeutics*, *92*(2), 228–234.

Hammann, F., Gutmann, H., Vogt, N., Helma, C., & Drewe, J. (2010). Prediction of adverse drug reactions using decision tree modeling. *Clinical Pharmacology & Therapeutics*, *88*(1), 52–59.

Härmark, L., Van Der Wiel, H. E., De Groot, M. C. H., & Van Grootheest, A. C. (2007). Proton pump inhibitor-induced acute interstitial nephritis. *British Journal of Clinical Pharmacology*, *64*(6), 819–823.

Harpaz, R., Chase, H. S., & Friedman, C. (2010). Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics*, *11*(9), S7.

Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel

Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics*, *91*(6), 1010–1021.

Harpaz, R., Haerian, K., Chase, H. S., & Friedman, C. (2010). Mining electronic health records for adverse drug effects using regression based methods. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 100–107). ACM.

Harpaz, R., Vilar, S., DuMouchel, W., Salmasian, H., Haerian, K., Shah, N. H., … Friedman, C. (2013). Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association*, *20*(3), 413–419.

He, H., Williams, G., Chen, J., Hawkins, S., & Kelman, C. (2004). Exploring possible adverse drug reactions by clustering event sequences. *Data Warehousing and Knowledge Discovery*, 199–208.

Heerspink, H. J. L., Desai, M., Jardine, M., Balis, D., Meininger, G., & Perkovic, V. (2017). Canagliflozin slows progression of renal function decline independently of glycemic effects. *Journal of the American Society of Nephrology*, *28*(1), 368–375.

Hoang, T., Liu, J., Pratt, N., Zheng, V. W., Chang, K. C., Roughead, E., & Li, J. (2016). Detecting signals of detrimental prescribing cascades from social media. *Artificial Intelligence in Medicine*, *71*, 43–56. http://doi.org/https://doi.org/10.1016/j.artmed.2016.06.002

Hsu, W.-H., Hsiao, P.-J., Lin, P.-C., Chen, S.-C., Lee, M.-Y., & Shin, S.-J. (2017). Effect of metformin on kidney function in patients with type 2 diabetes mellitus and moderate chronic kidney disease. *Oncotarget*, *9*(4), 5416–5423. http://doi.org/10.18632/oncotarget.23387

Huang, J. (2018). Drug-Induced Nephrotoxicity and Drug Metabolism in Renal Failure. *Current Drug Metabolism*, *19*(7), 558.

Huang, L.-C., Wu, X., & Chen, J. Y. (2011). Predicting adverse side effects of drugs. *BMC Genomics*, *12*(5), S11.

Huang, L., Wu, X., & Chen, J. Y. (2013). Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. *Proteomics*, *13*(2), 313–324.

Hug, B. L., Keohane, C., Seger, D. L., Yoon, C., & Bates, D. W. (2012). The costs of adverse drug events in community hospitals. *The Joint Commission Journal on Quality and Patient Safety*, *38*(3), 120–126.

Iglesias, P., & Diez, J. J. (2008). Insulin therapy in renal disease. *Diabetes, Obesity & Metabolism*, *10*(10), 811–823. http://doi.org/10.1111/j.1463-1326.2007.00802.x

Iizuka, T. (2007). Experts' agency problems: evidence from the prescription drug market in Japan. *The Rand Journal of Economics*, *38*(3), 844–862.

Imming, P., Sinning, C., & Meyer, A. (2006). Drugs, their targets and the nature and number of drug targets. *Nature Reviews Drug Discovery*, *5*(10), 821.

Inman, W., & Pearce, G. (1993). Prescriber profile and post-marketing surveillance. *The Lancet*, *342*(8872), 658–661.

Izzedine, H., Launay-Vacher, V., & Deray, G. (2005). Antiviral drug-induced nephrotoxicity. *American Journal of Kidney Diseases*, *45*(5), 804–817.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 37–50.

Jewell, N. P. (2003). *Statistics for epidemiology*. CRC Press.

Ji, Y., Ying, H., Dews, P., Mansour, A., Tran, J., Miller, R. E., & Massanari, R. M. (2011). A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Transactions on Information Technology in Biomedicine, 15*(3), 428–437.

Karimi, S., Kim, S., & Cavedon, L. (2011). Drug side-effects: What do patient forums reveal. In *The second international workshop on Web science and information exchange in the medical Web* (pp. 10–11). ACM.

Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., & Paris, C. (2015). Text and Data Mining Techniques in Adverse Drug Reaction Detection. *ACM Computing Surveys, 1*(March). http://doi.org/10.1145/2719920

Kim, M. G., Jeong, C. R., Kim, H. J., Kim, J. H., Song, Y., Kim, K. I., … Oh, J. M. (2018). Network analysis of drug-related problems in hospitalized patients with hematologic malignancies.

Kimura, G., Kasahara, M., Ueshima, K., Tanaka, S., Yasuno, S., Fujimoto, A., … Nakao, K. (2017). Effects of atorvastatin on renal function in patients with dyslipidemia and chronic kidney disease: assessment of clinical usefulness in CKD patients with atorvastatin (ASUCA) trial. *Clinical and Experimental Nephrology*, *21*(3), 417–424.

Kuo, R. J., Lin, S. Y., & Shih, C. W. (2007). Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan. *Expert Systems with Applications*, *33*(3), 794–808. http://doi.org/10.1016/J.ESWA.2006.08.035

LaBute, M. X., Zhang, X., Lenderman, J., Bennion, B. J., Wong, S. E., & Lightstone, F. C. (2014). Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. *PloS One, 9*(9), e106298.

Lazaar, A. L., Yang, L., Boardley, R. L., Goyal, N. S., Robertson, J., Baldwin, S. J., … Mayer, R. J. (2016). Pharmacokinetics, pharmacodynamics and adverse event profile of GSK2256294, a novel soluble epoxide hydrolase inhibitor. *British Journal of Clinical Pharmacology*, *81*(5), 971–979.

Lazarou, J., Pomeranz, B. H., & Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama*, *279*(15), 1200–1205.

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., & Gonzalez, G. (2010). Towards Internet-Age Pharmacovigilance : Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. In *BioNLP '10 Proceedings of the 2010 Workshop on Biomedical Natural Language Processing* (pp. 117–125).

Lee, D., Park, S.-H., & Moon, S. (2013). Utility-based association rule mining: A marketing solution for cross-selling. *Expert Systems with Applications*, *40*(7), 2715–2725.

Lee, W. H., Wang, E. T., & Chen, A. L. P. (2017). Mining accompanying relationships between diseases from patient records. In *Big Data (Big Data), 2017 IEEE International Conference on* (pp. 3861–3868). IEEE.

Lehmann, R., & Schleicher, E. D. (2000). Molecular mechanism of diabetic nephropathy. *Clinica Chimica Acta*, *297*(1), 135–144.

Liben-Nowell, D., & Kleinberg, J. (2007). The Link-Prediction Problem for Social Networks. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, *58*(7), 1019–1031. http://doi.org/10.1002/asi

Lin, S.-F., Xiao, K.-T., Huang, Y.-T., Chiu, C.-C., & Soo, V.-W. (2010). Analysis of adverse drug reactions using drug and drug target interactions and graph-based methods. *Artificial*

Intelligence in Medicine, *48*(2), 161–166. http://doi.org/https://doi.org/10.1016/j.artmed.2009.11.002

Liu, J., Zhao, S., & Zhang, X. (2016). An ensemble method for extracting adverse drug events from social media. *Artificial Intelligence in Medicine*, *70*, 62–76. http://doi.org/https://doi.org/10.1016/j.artmed.2016.05.004

Liu, M., Wu, Y., Chen, Y., Sun, J., Zhao, Z., Chen, X., … Xu, H. (2012). Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *Journal of the American Medical Informatics Association*, *19*(e1), e28–e35.

Liu, X., & Chen, H. (2013). AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. In *International Conference on Smart Health* (pp. 134–150). Springer.

Loh, A. H., & Cohen, A. H. (2009). Drug-induced kidney disease-pathology and current concepts. *Ann Acad Med Singapore*, *38*(3), 240–250.

Markowitz, G. S., & Perazella, M. A. (2005). Drug-induced renal failure: a focus on tubulointerstitial disease. *Clinica Chimica Acta*, *351*(1), 31–47.

Matthews, E. J., Ursem, C. J., Kruhlak, N. L., Benz, R. D., Sabaté, D. A., Yang, C., … Contrera, J. F. (2009). Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part B. Use of (Q) SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities. *Regulatory Toxicology and Pharmacology*, *54*(1), 23–42.

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y.-P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with*

Applications, *40*(4), 1086–1093.

Naughton, C. A. (2008). Drug-induced nephrotoxicity. *American Family Physician*, *78*(6).

Nguyen, T., Larsen, M. E., O'Dea, B., Phung, D., Venkatesh, S., & Christensen, H. (2017). Estimation of the prevalence of adverse drug reactions from social media. *International Journal of Medical Informatics*, *102*, 130–137. http://doi.org/10.1016/J.IJMEDINF.2017.03.013

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., & Gonzalez, G. (2015a). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, *22*(3), 671–681.

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., & Gonzalez, G. (2015b). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, *22*(3), 671–681. http://doi.org/10.1093/jamia/ocu041

O'Connor, K., Pimpalkhute, P., Nikfarjam, A., Ginn, R., Smith, K. L., & Gonzalez, G. (2014). Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. In *AMIA annual symposium proceedings* (Vol. 2014, p. 924). American Medical Informatics Association.

Park, M. Y., Yoon, D., Lee, K., Kang, S. Y., Park, I., Lee, S., … Kim, J. H. (2011). A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. *Pharmacoepidemiology and Drug Safety*, *20*(6), 598–607.

Perazella, M. A. (2003). Drug-induced renal failure: update on new medications and unique mechanisms of nephrotoxicity. *The American Journal of the Medical Sciences*, *325*(6), 349–

362.

Perneger, T. V, Whelton, P. K., & Klag, M. J. (1994). Risk of kidney failure associated with the use of acetaminophen, aspirin, and nonsteroidal antiinflammatory drugs. *New England Journal of Medicine*, *331*(25), 1675–1679.

Piri, S., Delen, D., Liu, T., & Paiva, W. (2018). Development of a new metric to identify rare patterns in association analysis: The case of analyzing diabetes complications. *Expert Systems with Applications*, *94*, 112–125. http://doi.org/10.1016/J.ESWA.2017.09.061

Polimeni, G., Ospedaliera, A., Martino, U. G., Moore, N. D., Fourrier-réglat, A., Victor, U., & Bordeaux, S. (2009). Data mining on electronic health record databases for signal detection in pharmacovigilance : Which events to monitor ? Data mining on electronic health record databases for signal detection in pharmacovigilance : which events to monitor ?, (January 2014). http://doi.org/10.1002/pds.1836

Pouliot, Y., Chiang, A. P., & Butte, A. J. (2011). Predicting adverse drug reactions using publicly available PubChem BioAssay data. *Clinical Pharmacology & Therapeutics*, *90*(1), 90–99.

Pratt, N., Chan, E. W., Choi, N., Kimura, M., Kimura, T., Kubota, K., … Park, B. (2015). Prescription sequence symmetry analysis: assessing risk, temporality, and consistency for adverse drug reactions across datasets in five countries. *Pharmacoepidemiology and Drug Safety*, *24*(8), 858–864.

Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011). Identifying health-related topics on twitter. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 18–25). Springer.

Reps, J. M., Aickelin, U., & Hubbard, R. B. (2016). Refining adverse drug reaction signals by

incorporating interaction variables identified using emergent pattern mining. *Computers in Biology and Medicine*, *69*, 61–70.

Riccioli, C., Leroy, N., & Pelayo, S. (2009). The PSIP approach to account for human factors in Adverse Drug Events: Preliminary field studies. *Stud Health Technol Inform*, *148*, 197–205.

Santiso, S., Casillas, A., & Pérez, A. (2018). The class imbalance problem detecting adverse drug reactions in electronic health records. *Health Informatics Journal*, 1460458218799470.

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., … Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, *54*, 202–212. http://doi.org/10.1016/j.jbi.2015.02.004

Schuemie, M. J. (2011). Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Pharmacoepidemiology and Drug Safety*, *20*(3), 292–299.

Shetty, K. D., & Dalal, S. R. (2011). Using information mining of the medical literature to improve drug safety. *Journal of the American Medical Informatics Association*, *18*(5), 668–674.

Simpson, G. G. (1960). Notes on the measurement of faunal resemblance. *American Journal of Science*, *258*(2), 300–311.

Singh, N. P., Ganguli, A., & Prakash, A. (2003). Drug-induced kidney diseases. *Journal of Association of Physicians of India*, *51*, 970–979.

Stephens, M. D. B., & Talbot, J. C. C. (1985). *The detection of new adverse drug reactions*. Springer.

Sun, Y.-M., Su, Y., Li, J., & Wang, L.-F. (2013). Recent advances in understanding the

biochemical and molecular mechanism of diabetic nephropathy. *Biochemical and Biophysical Research Communications*, *433*(4), 359–361.

Thomas, G., Rojas, M. C., Epstein, S. K., Balk, E. M., Liangos, O., & Jaber, B. L. (2007). Insulin therapy and acute kidney injury in critically ill patients—a systematic review. *Nephrology Dialysis Transplantation*, *22*(10), 2849–2855.

Trame, M. N., Biliouris, K., Lesko, L. J., & Mettetal, J. T. (2016). Systems pharmacology to predict drug safety in drug development. *European Journal of Pharmaceutical Sciences*, *94*, 93–95. http://doi.org/10.1016/j.ejps.2016.05.027

Trifirò, G., Pariente, A., Coloma, P. M., Kors, J. A., Polimeni, G., Miremont-Salamé, G., … Moore, N. (2009). Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiology and Drug Safety*, *18*(12), 1176–1184.

Trippe, Z. A., Brendani, B., Meier, C., & Lewis, D. (2017). Identification of substandard medicines via disproportionality analysis of individual case safety reports. *Drug Safety*, *40*(4), 293–303.

Tsiropoulos, I., Andersen, M., & Hallas, J. (2009). Adverse events with use of antiepileptic drugs: a prescription and event symmetry analysis. *Pharmacoepidemiology and Drug Safety*, *18*(6), 483–491.

van Puijenbroek, E. P., Bate, A., Leufkens, H. G. M., Lindquist, M., Orre, R., & Egberts, A. C. G. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and Drug Safety*, *11*(1), 3–10.

Vazzana, M., Andreani, T., Fangueiro, J., Faggio, C., Silva, C., Santini, A., … Souto, E. B. (2015). Tramadol hydrochloride: pharmacokinetics, pharmacodynamics, adverse side effects, co-administration of drugs and new drug delivery systems. *Biomedicine & Pharmacotherapy, 70*, 234–238.

Vo, B., Coenen, F., & Le, B. (2013). A new method for mining Frequent Weighted Itemsets based on WIT-trees. *Expert Systems with Applications*, *40*(4), 1256–1264.

von Websky, K., Reichetzeder, C., & Hocher, B. (2013). Linagliptin as add-on therapy to insulin for patients with type 2 diabetes. *Vascular Health and Risk Management*, *9*, 681–694. http://doi.org/10.2147/VHRM.S40035

Wedemeyer, R.-S., & Blume, H. (2014). Pharmacokinetic drug interaction profiles of proton pump inhibitors: an update. *Drug Safety*, *37*(4), 201–211.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., … Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, *34*(suppl_1), D668–D672.

Yamanouchi, T. (2010). Concomitant therapy with pioglitazone and insulin for the treatment of type 2 diabetes. *Vascular Health and Risk Management*, *6*, 189–197. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2856574/

Yang, C. C., Jiang, L., Yang, H., & Tang, X. (2012). Detecting signals of adverse drug reactions from health consumer contributed content in social media. In *Proceedings of ACM SIGKDD Workshop on Health Informatics*.

Yildirim, M. A., Goh, K. Il, Cusick, M. E., Barabási, A. L., & Vidal, M. (2007). Drug-target network. *Nature Biotechnology*, *25*(10), 1119–1126. http://doi.org/10.1038/nbt1338

Zeng, Q., Kogan, S., Ash, N., Greenes, R. A., & Boxwala, A. A. (2002). Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine-Methodik Der Information in Der Medizin*, *41*(4), 289–298.

Zhang, Y., Tao, C., He, Y., Kanjamala, P., & Liu, H. (2013). Network-based analysis of vaccine-related associations reveals consistent knowledge with the vaccine ontology. *Journal of Biomedical Semantics*, *4*(1), 1–8. http://doi.org/10.1186/2041-1480-4-33

Zhou, T., Lü, L., & Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B*, *71*(4), 623–630.

APPENDICES

APPENDIX 1

The list of kidney-damaging (KD) as well as common diabetic medications included in the study follows.

GENERIC_NAME	*Type*	*GENERIC_NAME*	*Type*	*GENERIC_NAME*	*Type*
acetaminophen	KD	bevacizumab	KD	insulin (variations)	Diabetic
aspirin	KD	indomethacin	KD	metformin	Diabetic
pantoprazole	KD	hydroxychloroquine	KD	glipizide	Diabetic
vancomycin	KD	pamidronate	KD	sitagliptin	Diabetic
ketorolac	KD	doxycycline	KD	glyburide	Diabetic
esomeprazole	KD	azithromycin	KD	glimepiride	Diabetic
tacrolimus	KD	clindamycin	KD	pioglitazone	Diabetic
ciprofloxacin	KD	tenofovir	KD	linagliptin	Diabetic
ibuprofen	KD	ketoprofen	KD	repaglinide	Diabetic
sulfamethoxazole-trimethoprim	KD	mitomycin	KD	saxagliptin	Diabetic
omeprazole	KD	sulindac	KD	acarbose	Diabetic
lansoprazole	KD	captopril	KD	liraglutide	Diabetic
cyclosporine	KD			nateglinide	Diabetic
phenytoin	KD			canagliflozin	Diabetic
cephalexin	KD			exenatide	Diabetic
acyclovir	KD			bromocriptine	Diabetic
naproxen	KD			saxagliptin	Diabetic
diclofenac	KD			acarbose	Diabetic
celecoxib	KD			liraglutide	Diabetic

APPENDIX 2

This appendix describes the search strategy used for extracting data from the National Library of Medicine's (NLM) MEDLINE database of biomedical citations. MEDLINE is a subset of PubMed database and includes more than 26 million biomedical citations started from 1946 onward. Each article is carefully read and annotated by a group of trained indexers using a vocabulary system called Medical Subject Headings (MeSH). The MeSH thesaurus is a controlled vocabulary system produced by NLM to be used for indexing, cataloging, and searching for biomedical citations and health-related documents. After carefully reading an article, the NLM indexing experts select the most appropriate *descriptors* and *subheadings (a.k.a. qualifiers)* that best describe the content.

To extract required data for the present study we downloaded all the MEDLINE citations from PubMed (https://www.ncbi.nlm.nih.gov/pubmed) with the "AE" MeSH subheading, which is used to indicate mentions of Adverse Effects, and containing at least one of the 8 high-risk ADEs of interest indexed as "Chemically induced". These two MeSH indexes, together, specify the drug and the adverse event mentioned as an association in a given article. For instance, the combination of "acetaminophen/AE" and "Acute kidney failure/chemically induced" for a given article indicates

that a drug-ADE association suggesting the potential adverse effect of acetaminophen on developing kidney failure is mentioned in that study.

The search was done multiple times, each time for one of the ADEs of interest; however, at the end, we removed duplicated citations from our records. For each article, the following information was collected: article PubMed ID (PMID), MeSH descriptors, subheadings, substances, and date of publication.

Since drugs' target protein and date of approval information were to be extracted from another resource (i.e., DrugBank), we then used the Unified Medical Language System (UMLS) to map the drug and ADE terms. UMLS is a biomedical terminology integration system handling more than 150 terminologies including MeSH. It integrates various alternatives of the same biomedical concepts and assigns each concept a unique identifier (CUI) across the whole database. All the drug and ADE terms in the collected dataset were mapped to their corresponding UMLS terms and the CUI associated with each was queried and added to the dataset.

The list of approved FDA drugs along with their target proteins was downloaded from DrugBank's (https://www.drugbank.ca) Therapeutic Target Database (TTD) ver. 6.1.01 and mapped to UMLS terms as well. Then the list was used to filter the articles collected from MEDLINE so that we only kept articles including approved drugs and put away studies focusing on experimental drugs or chemical compounds.

Finally, drug-ADE pairs were created by matching mentions of the "AE" and "Chemically induced" tags in the same publications and the corresponding publication dates were assigned to the created pairs. Repeated pairs were then identified and redundancies were removed by maintaining only the earliest drug-ADE mention (based on dates). Also using the DrugBank data, drug-drug pairs were created by matching the drugs sharing at least one target protein.

The created pairs were then used as the input to both Cytoscape v3.6.0 and the *igraph* package in R to create network visualization and metrics, respectively.

VITA

Behrooz Davazdahemami

Candidate for the Degree of

Doctor of Philosophy

Dissertation: USING BIG DATA ANALYTICS AND STATISTICAL METHODS FOR IMPROVING DRUG SAFETY

Major Field: Business Administration (MSIS)

Biographical:

Educational:

Completed the requirements for the Doctor of Philosophy in Business Administration (MSIS) at Oklahoma State University, Stillwater, Oklahoma in May 2019

Completed the requirements for the Master of Science in Industrial Engineering at University of Tehran, Tehran, Iran in 2012

Completed the requirements for the Bachelor of Science in Industrial Engineering at Isfahan University of Technology, Isfahan, Iran in 2009

Experience:

Graduate Teaching Associate at the MSIS Department, Oklahoma State University, Stillwater, Oklahoma, 2014-2019.

Professional Memberships:

Association for Information Systems (AIS), 2016-2019.

Decision Sciences Institute (DSI), 2016-2019.

Institute of Operations Research and Management Sciences (INFORMS), 2016-2019.